Work, wealth, and well-being

Essays in macroeconomics

Jonna Olsson
Work, wealth, and well-being
Essays in macroeconomics
Jonna Olsson

Abstract

Structural transformation of the labor market and the aggregate economy

Women's increased involvement in the economy has been the most significant change in labor markets during the past century. In this paper, I account for this period of structural change of the labor market in a macroeconomic model, and study how the increase in female labor force participation has affected the economy's response to aggregate shocks. I explicitly model heterogeneity in gender and household composition as well as the historical decrease of the gender wage gap. The model captures the salient features of historical data, including a strong increase in employment among married women, low crowding-out of married men, and relatively stable employment over time for single women. I then study how the changing labor force composition affects the economy's aggregate employment dynamics. The underlying trend in employment, driven by growth in female labor force participation, contributed to the perceived quick employment recovery after recessions before 1990. In general, incorporating both one- and two-person households matters for employment dynamics, with single households reacting more strongly to shocks and employment responses by subgroups changing over time.

Labor supply in a quantitative heterogeneous-agent model

Since long, the labor-supply channel has played a central role in macroeconomic analysis. Nevertheless, it has almost exclusively focused on representative-agent behavior. The aim of this paper is to examine frameworks that are significantly richer in terms of heterogeneity and uncertainty, and assess whether the predictions yielded by the starker frameworks are robust to these extensions.

Subjective life expectancies, time preference heterogeneity and wealth inequality

There is substantial heterogeneity in statistical and perceived life expectancy in the population. In this paper we document a systematic bias in survival beliefs: individuals with low survival probability relative to their peers underestimate their life expectancies, while individuals with high survival probability overestimate. To gauge the effect of heterogeneity in life expectancy on savings rates and ultimately wealth inequality, we introduce shocks to survival beliefs into an otherwise standard overlapping-generations model. We show that such a model exhibits counter-factual savings behavior as individuals increase their savings when their life expectancy drops. Nevertheless, overall wealth inequality in the economy is virtually unaffected by heterogeneity in survival beliefs, contrary to previous literature.

Health dynamics and heterogeneous life expectancies

In this paper, we provide improved estimates for age-dependent health transitions and survival probabilities for different subsamples of the US population. The estimated yearly transition matrices can be used in any life-cycle model where health and survival dynamics is of interest. The results show substantial heterogeneity in life expectancy in the population. For a 70-year-old man in excellent health, the probability of reaching his 80th birthday is around 75%, while the corresponding probability for a man in poor health is just below 40%.

Stockholm 2019

http://urn.kb.se/resolve?urn=urn:nbn:se:su:diva-172106

ISBN 978-91-7797-808-4
ISSN 0346-6892

Department of Economics

Stockholm University, 106 91 Stockholm
WORK, WEALTH, AND WELL-BEING

Jonna Olsson
Abstracts

**Structural transformation of the labor market and the aggregate economy**

Women’s increased involvement in the economy has been the most significant change in labor markets during the past century. In this paper, I account for this period of structural change of the labor market in a macroeconomic model, and study how the increase in female labor force participation has affected the economy’s response to aggregate shocks. I explicitly model heterogeneity in gender and household composition as well as the historical decrease of the gender wage gap. The model captures the salient features of historical data, including a strong increase in employment among married women, low crowding-out of married men, and relatively stable employment over time for single women. I then study how the changing labor force composition affects the economy’s aggregate employment dynamics. The underlying trend in employment, driven by growth in female labor force participation, contributed to the perceived quick employment recovery after recessions before 1990, and the absence of growth thereafter consequently explains the more recent slower employment recoveries. In general, incorporating both one- and two-person households matters for employment dynamics, with single households reacting more strongly to shocks and employment responses by subgroups changing over time. Despite relatively large changes by subgroup, the aggregate effect is unchanged between the 1970s and the present time due to multiple counteracting forces.

**Labor supply in a quantitative heterogeneous-agent model**

*(with Timo Boppart and Per Krusell)*

In this paper, we examine a class of macroeconomic models with endogenous labor supply in a context where households are heterogeneous in wealth and wages, face earnings uncertainty, and suffer from the fact that they cannot fully insure against this uncertainty. We systematically study a number of issues, including the quantitative importance of lack of aggregation and the role of the extensive vs. intensive margin in the labor supply choice. We address the normative question about who “should” work in the economy, by comparing the unconstrained optimum to the outcome in the incomplete-markets model. We systematically contrast results from using the most commonly used utility function in the literature—one that imposes that the income and substitution effects cancel—with results from using a utility function that assumes a stronger income effect, since aggregate data over time and across countries suggest that the income effects are stronger.
Health dynamics and heterogeneous life expectancies
(with Richard Foltyn)
In this paper, we provide improved estimates for age-dependent health transitions and survival probabilities for different subsamples of the US population. The estimated yearly transition matrices can be used in any life-cycle model where health and survival dynamics is of interest. The results show substantial heterogeneity in life expectancy in the population. For a 70-year-old man in excellent health, the probability of reaching his 80th birthday is around 75%, while the corresponding probability for a man in poor health is just below 40%. There is also substantial inequality in life expectancy between different educational groups. In the group with less than a high school degree, the life expectancy at the age of 50 is 75 years, while the average for those with some college education or more is 80 years. This difference is due to two factors. First, at the age of 50, overall health is worse in the group with lower education. Second, even conditional on health status, the health dynamics and survival probabilities for this group are worse also from the age of 50 and onwards. We estimate that the difference in life expectancy across education groups mainly stems from the worse health and survival dynamics after the age of 50.

Subjective life expectancies, time preference heterogeneity and wealth inequality
(with Richard Foltyn)
There is substantial heterogeneity in life expectancy in the population. However, an individual’s consumption-savings decision is not necessarily guided by the objective statistical life expectancy, but rather by the individual’s beliefs about survival. In this paper we document a systematic bias in survival beliefs: individuals with low survival probability relative to their peers underestimate their life expectancies, while individuals with high survival probability overestimate. To gauge the effect of heterogeneity in life expectancy (objective and subjective) on savings rates and ultimately wealth inequality, we introduce shocks to survival beliefs into an otherwise standard overlapping-generations model. We show that with a bequest motive calibrated to match asset decumulation in old age, such a model exhibits counter-factual savings behavior as individuals increase their savings when their life expectancy drops. Nevertheless, overall wealth inequality in the economy is virtually unaffected by heterogeneity in survival beliefs, contrary to previous literature which finds stronger effects of heterogeneous discount factors.
Acknowledgements

Among many other things, my advisors have taught me that the giraffe is the spirit animal of modern macro. Their argument was something along the following lines: it should be an animal that has a great overview and sees the big picture from above. But why then not a bird, one might ask? A majestic golden eagle? Or a griffin vulture, the highest flying bird in the world, recorded to fly at an altitude of 11,300 meters? Because a macro animal should also be firmly rooted in the ground, with all its feet deep in the micro foundations of economics. Hence the giraffe, the tallest animal on earth.

However, one thing they forgot to mention was that the giraffe is also a mammal with an unusually large, according to some even the biggest, heart relative to body mass. I am not in a position to judge the macro field as a whole, but a big heart is certainly a trait that also defines my two advisors, Per Krusell and Timo Boppart.

I first met Per, my main advisor, in the first-year macro course. After taking his course, the small doubts I had about which field to enter within economics were gone: both the questions asked and the methods applied to try to answer them made sense to me. Per has been extremely generous with both time and knowledge throughout these years, and has been a constant source of advice, encouragement, and nudging. Per’s curiosity and commitment to high quality have helped me immensely. At many points when I have thought I have finally (almost) understood something, his follow-up questions have pushed me to dig deeper, go further, and become a better researcher.

In my third year I started working on a joint project with Per and Timo. Sitting in meetings with these two scholars has often felt like sitting in the research version of a world-class tennis match: ideas bouncing around, often at lightning speed (and sometimes flying far outside the court), and if you lose focus for a second you’ll miss the point. Only a very small fraction of these discussions ended up in this dissertation, but for an aspiring researcher to be, it was an invaluable experience. Timo eventually also became my second supervisor, and has been a solid rock to lean on when it comes to scientific judgement and advice, about academia in general and the job market in particular. His high standards and sometimes brutal honesty combined with his caring personality have been refreshing and improved this dissertation substantially.

I also want to thank the rest of the faculty in the Macro Group: John Hassler, Kurt Mitman, Tobi Broer, Alex Kohlhas, and Yimei Zou have all helped tremendously with feedback, questions, and comments. Kurt, together with Kathrin Schlafmann,
was in some way responsible for me ending up with an interest in quantitative methods, by giving a fun and instructive second-year course and patiently answering all types of questions throughout the years.

The last four years of my studies I have spent at the IIES, which has been the best place one can imagine for a PhD student. A vibrant research community, seminars in all fields, and animated discussions in the kitchen on every possible subject ensure that one never becomes narrow-minded or bored. Thanks to Ingvild Almås (who also served as a great gym buddy), Konrad Burchardi, Jon de Quidt, Arash Nekoei, Peter Nilsson, Mats Persson, Torsten Persson, Jakob Svensson, and Robert Östling for contributing to this dynamic environment which I am grateful to have been part of.

A special thanks to Tessa Bold and Mitch Downey who served as placement officers during my year on the job market and made sure we understood the process and were not unnecessarily surprised at any stage.

The IIES would not work as well as it does if it was not for Christina Lönnblad. I am always positively surprised how administrative things just work out smoothly at the Institute, and suspect that Christina’s magic is the answer. Christina also improved the language of this dissertation considerably. Many thanks also to Tove Happonen (who can do infinitely many more pull-ups than I can) for making this workplace work so well. Without the practical guidance from Ulrika Gålñander (who also brings a lot of joy to the corridor) this dissertation wouldn’t have materialized.

My journey as a PhD student would have been very different, had I not had such exceptional colleagues. I have had the great fortune of sharing an office with Magnus Åhl all the time (except one scary year when he was away). One could not have asked for a better office mate: our discussions about everything from very practical research problems, current events, economic insights, grammar and language, and life in general have benefited me tremendously.

Magnus, Matti Mitrunen, Jaakko Meriläinen, Matilda Kilström, Josef Sigurdsson, Selene Ghisolfi, Serena Cocciolo, Elisabet Olme, Erik Lindgren, and I all started at the same time in the PhD program. I regard myself very lucky to have had this awesome group of people around me, especially during the first two years of course work.

Karin Kinnerud has been a great office mate for the last year, and before that Miriam Hurtado-Bodell. I never shared an office with the two Finns, Matti and Jaakko, but I always felt welcome to crash on their and Divya Dev’s couch whenever I needed. Divya has also been a somewhat too good badminton player for me, but I will keep on practicing.

All students in the Macro Group, including Saman Darougheh, Markus Karlman, Kasper Kragh-Sørensen, Has Van Vlokhoven, and John Kramer have contributed to the fun discussions. From the older IIES cohorts, I am thankful to Niels-Jakob Harbo Hansen, Karl Harmenberg, Mathias Iwanowsky, Hannes Malmberg, and
Erik Öberg. Thanks also to Benedetta Lerva for all hilarious stories, and to Xueping Sun, Philipp Hochmuth, Fabian Sinn, and Sreyashi Sen for being great colleagues over the years.

At the Department of Economics I would like to thank Roine Vestman for being instrumental in admitting me to the PhD program after quite a few gap years outside academia, and for his great support during my first shaky steps as a PhD student. Karolina Ekholm was Erika Öhlund’s and my supervisor (and role model) many many years ago while writing our master’s thesis, and her clever guidance and encouragement were crucial for the fact that both Erika and I finally ended up as PhD students, albeit after some detours. Among many others, I am indebted to Peter Fredriksson and David Strömberg for patiently answering questions about data-generating processes and estimation strategies.

Mentally very far from the University, but geographically pretty close, is the Stora Skuggans 4H-gård, which has been the goal of numerous lunch walks. In stressful periods, the best remedy is to pet the sheep and goats there. Thanks to the farm and especially to Rosalinda and Willmer.

Outside academia I want to thank everyone who has kept on reminding me that of course life is much more than research. You’re too many and too important to be mentioned in the way that you deserve. Still, special shout-outs to the wine tasters who have made sure that I once in a while eat some really good food, and to my book club who has ensured that I read more than just economic papers or novels that I would find myself.

My parents have never questioned my sometimes somewhat unusual career choices, quite the opposite, and I am forever grateful for the unconditional support of my father Karl Axel and my mother Kerstin. My sister Julia and her family have served as a harbour of love and rest and distraction from work. The support from my brother Märten has meant the world to me, and I am also grateful to have my amazing nieces in my life.

Last but not least, two of the papers in this dissertation are jointly written with Richard Foltyn. To co-author with Richard is frustrating, painful, and absolutely amazing. Richard’s no-nonsense approach—not even on a preliminary stage anything less than crystal-clear is permitted regarding assumptions, methodology, or execution—leads to endless discussions and arguments, but it also ensures a journey full of new insights and has made me a much better economist. He never ceases to amaze me. Another effect of working with Richard is that by sheer osmosis my technical skills have improved substantially.

However, Richard is of course not only my partner-in-crime academically but also in life in general. He makes me laugh, expands my thinking, and listens patiently to my endless stream of insightful comments (or should it be called nonsense?). Richard, I am immensely grateful to have you by my side.

Jonna Olsson
Stockholm, Sweden
August 2019
Contents

Introduction .................................................................................................................. 1

Chapter 1: Structural transformation of the labor market and the aggregate economy ................................................................. 5

Chapter 2: Labor supply in a quantitative heterogeneous-agent model ................................................................................................. 67

Chapter 3: Health dynamics and heterogeneous life expectancies ........................................................................................................ 153

Chapter 4: Subjective life expectancies, time preference heterogeneity, and wealth inequality .............................................................. 193

Sammanfattning (Swedish summary) ........................................................................ 241
Introduction

This dissertation consists of four self-contained papers. Chapters 1 and 2 focus on heterogeneity in labor supply while Chapters 3 and 4 deal with heterogeneity in health and life expectancy. Even though the chapters are self-contained, they jointly reflect my interest in the branch of macroeconomics referred to as heterogeneous agents models and my belief that we need to analyze individual behavior and not only aggregates when trying to understand many economic issues.

In the first chapter, Structural transformation of the labor market and the aggregate economy, the focus is on the dramatic change in female labor force participation among married women in the US over the last five decades, from below 40% in the 1960s to almost 70% now. At the beginning of the 1960s, the most common type of household in the economy was a household with two adults of which one, the husband, was the sole breadwinner. Today, this is no longer the case. Women’s increased involvement in the economy has perhaps been the most significant change in the structure of the labor markets during the past century.

I account for this period of structural change of the labor market in a macroeconomic model, and study how the increase in female labor force participation has affected the economy’s response to aggregate shocks. For this purpose I construct a heterogeneous agent model in which I explicitly include the gender and household composition dimensions (alongside heterogeneity in assets and productivity). Otherwise the model is a straightforward business cycle model. The sole exogenous driver of change in the model is the shrinking gender wage gap, and hence, the focus is on the changing economic incentives within the household.

The model is able to capture the salient features of historical data, including a strong increase in employment among married women, low crowding-out of married men, and relatively stable employment over time for single women. Model results show that the underlying trend in employment, driven by growth in female labor force participation, contributed to the perceived quick employment recovery after recessions before 1990, and the absence of growth thereafter consequently explains the more recent slower employment recoveries.

In general, I show that incorporating both one- and two-person households in the model matters for aggregate employment dynamics, with single households...
reacting more strongly to shocks. Furthermore, employment responses by subgroup have changed over time. For instance, among married women, the response to a shock has become more muted as women on average have moved further away from their work margin.

In Chapter 2, Labor supply in a quantitative heterogeneous-agent model, which is joint with Timo Boppart and Per Krusell, I examine labor supply in a framework that is closely related to that in my first chapter, namely one that features wealth and wage heterogeneity across households and has an incomplete savings markets. Here, I only look at one-member households in order to focus on other issues, some of which are present in the literature but have not been subjected to systematic study, and some which are new to the literature.

The first issue is aggregation: in virtually all of the papers in the macroeconomic literature, the utility functions over consumption and hours that are used do not aggregate in wealth; moreover, none of the frameworks allow aggregation over heterogeneous productivity levels. Given that both wealth and individual wages differ markedly in the data, I in this chapter look at the quantitative departures from aggregation. The second issue is how incomplete- and complete-markets models compare. A third concerns the normative question of who “should” work, and how much. I thus adopt the most commonly used social welfare function—the utilitarian function—and examine this issue; in particular I compare the unconstrained optimum to the outcome in the incomplete-markets models. The fourth issue I address is the role of the extensive vs. the intensive margin; in my first chapter I focus only on the former. This chapter thus compares the two and, in addition, develops a framework where both are present, i.e., where each household has a nontrivial extensive-margin choice but also can, conditional on working, choose hours. Fifth, I look at individual and aggregate Frisch elasticities, both under the intensive- and extensive-margin assumptions. Sixth, and finally, there is a systematic comparison between the most commonly used utility function in the literature—one that imposes that income and substitution effects cancel under balanced growth, thus making hours worked independent of technology (wage) growth—and one that assumes a stronger income effect. The reason for this comparison is that aggregate data over time and across countries strongly suggest that income effects are stronger: hours worked fall as countries develop.

On all these issues I find interesting results, some of large quantitative significance. Moreover, the resulting incomplete-markets model, which features strong income effects, performs quite well relative to existing data: inequality is better explained, as is the riskfree rate (the basic model generates much too little wealth heterogeneity and too high a riskfree interest rate).

In Chapters 3 and 4 I turn to the issue about well-being: heterogeneity in health and life expectancy in the population, more specifically within the US. Health inequality in itself is important, but to understand the underlying causes is far beyond the scope of a dissertation in economics. I therefore focus on the potential economic consequences of inequality in health and longevity. Numerous studies have identified health dynamics and health shocks as a major source of risk over the life cycle.
A negative health shock can result in large medical expenditures, which affects the incentives to accumulate assets, and could also affect the earnings potential. The survival probability directly affects the effective discount factor, a mechanism present in any life-cycle model with an uncertain life span. Some also argue that the health state directly influences the marginal utility from consumption. Hence, in order to quantify the risk an individual faces and model the choices and actions the individual takes, a correct health and survival process is crucial. We document the health process in a structured way that is possible to use in macroeconomic models, and take a small step in thinking about how inequality in life expectancy affects economic outcomes, in particular savings.

In the third chapter, **Health dynamics and heterogeneous life expectancies**, co-authored with Richard Foltyn, we provide improved estimates for age-dependent health transitions and survival probabilities for different subsamples of the US population. The estimated yearly transition matrices can be used in any life-cycle model where health and survival dynamics is of interest. The results show substantial heterogeneity in life expectancy in the population. For a 70-year-old man in excellent health, the probability of reaching his 80th birthday is around 75%, while the corresponding probability for a man in poor health is just below 40%. There is also substantial inequality in life expectancy between different educational groups. In the group with less than a high school degree, the life expectancy at the age of 50 is 75 years, while the average for those with some college education or more is 80 years. This difference is due to two factors. First, at the age of 50, overall health is worse in the group with lower education. Second, even conditional on health status, the health dynamics and survival probabilities for this group are worse also from the age of 50 and onwards. We estimate that the difference in life expectancy across education groups mainly stems from the worse health and survival dynamics after the age of 50.

In the fourth chapter, **Subjective life expectancies, time preference heterogeneity and wealth inequality**, also co-authored with Richard Foltyn, we use the results from the previous paper and ask a natural follow-up question: what are the implications of the heterogeneity in life expectancy on savings rates and ultimately wealth inequality?

According to standard economic theory a healthy person should save more for the future, all else equal, given the higher probability of living a longer life. However, an individual’s consumption/savings decision is not necessarily guided by the objective statistical life expectancy, but rather by the individual’s beliefs about survival. We document new facts about systematic bias in these beliefs: individuals with low survival probability relative to their peers underestimate their life expectancies, while individuals with high survival probability overestimate theirs. This systematic bias exacerbates the survival expectancy heterogeneity in the population.

To gauge the effect of survival heterogeneity, objective as well as subjective, on inequality, we use an overlapping-generations general-equilibrium model with uninsurable idiosyncratic shocks. Agents face heterogeneous survival risk that
INTRODUCTION

depends on their current health state, and are subject to health shocks that follow a process estimated from data. Besides this uncertainty, we also include standard persistent and transitory shocks to labor productivity. Since we are interested in savings behavior in late life, it is important to capture other incomes during this period. Therefore we carefully model retirement benefits, closely mimicking the US social security system.

We show that the standard life-cycle model gives rise to counter-factual implications when introducing survival heterogeneity. In an environment without bequests, agents with longer life expectancy save more, as expected. This is in line with the data, where individuals in better health have higher asset holdings. However, as is well known, this standard model without a bequest motive gives rise to counterfactually low savings among the elderly since agents in the model draw down their assets to virtually zero late in life. In the data, on the other hand, individuals on average have substantial asset holdings even beyond the age of 80. We therefore add a warm-glow bequest motive of the type most commonly used in the macroeconomic literature.

The effect of introducing survival heterogeneity into this environment is counter-intuitive and perhaps also unexpected: agents in poor health now save more than their healthy counterparts. The reason is as follows: since agents in poor health are more likely to die soon, they put an increased weight on bequest utility, thus raising their incentive to save. Hence, there are two effects from lower life expectancy that work in opposite directions: a shorter expected life span makes the agent save less for own consumption, but a stronger bequest motive makes the agent save more. The net effect varies depending on calibration of bequest parameters, but the second mechanism is always present with a bequest formulation of this type: a shorter life span makes agents want to save more to leave bequests. This creates a health-wealth gradient that is counter-factual and we argue that this mechanism is implausible.

We conclude that none of the standard models are adequate for investigating the effect of survival heterogeneity on savings rates and wealth inequality. We discuss possible extensions and reformulations and point out directions for further research.
Structural transformation of the labor market and the aggregate economy∗

Jonna Olsson

1. Introduction

Figure 1 shows the dramatic change in female labor force participation among married women in the U.S. over the last five decades. At the beginning of the 1960s, the mode household in the economy was a household with two adults of which one, the husband, was the sole breadwinner. Today, this is no longer the case. In Claudia Goldin’s words, “women’s increased involvement in the economy was the most significant change in labor markets during the past century” (Goldin 2006).

In this paper, I account for this period of structural change of the labor market in a macroeconomic model, and study how the increase in female labor force participation has affected the economy’s response to aggregate shocks. For this purpose, I construct a heterogeneous agent model in which I explicitly include the gender and household composition dimensions (alongside heterogeneity in assets and productivity). Otherwise the model is a straightforward business cycle model, leaving the incorporation of monetary and labor-market frictions to future research.

The model is able to capture the salient features of historical data, including a strong increase in employment among married women, a low crowding-out of married men, and relatively stable employment over time for single women. The model results show that the underlying trend in employment, driven by the growth in female labor force participation, contributed to the perceived quick employment recovery after recessions before 1990, and the absence of growth thereafter consequently explains the more recent slower employment recoveries.

In general, incorporating both one- and two-person households in the model matters for aggregate employment dynamics, with single households reacting more strongly to shocks. Furthermore, the employment responses by subgroup have changed over

∗I am grateful for helpful discussions with Timo Boppart, Richard Foltyn, Karl Harmenberg, John Hassler, My Hedlin, Karin Kinnerud, Per Krusell, Kurt Mitman, Magnus Åhl, and other participants in the IIES Macro Group.
time. For instance, among married women, the response to a shock has become more muted as women on average have moved further away from their work margin. Despite relatively large changes by subgroup over time, the aggregate effect is unchanged between the 1970s and the present time due to multiple counteracting forces.

On a more general level, my paper departs from the standard macroeconomic model assumption of a representative agent and introduces heterogeneity in the gender and household composition dimensions to be able to study how the increase in labor force participation among women has affected the economy’s employment dynamics. The answer to that question has implications for how we understand historical data, but also for what we can expect for the future, now that the labor force participation increase among married women has slowed down (or even stopped).

Why should we believe that the increased labor force participation among married women matters for the economy’s aggregate response? A first answer is purely compositional: women now make up a larger share of the workforce. If labor supply elasticities differ depending on gender and marital status, the changing composition of the workforce should have an effect on the aggregate elasticity in the economy.

A second answer is that even within subgroups, the employment responses to fluctuations in wages might have changed over time. It seems likely that with increasing participation, women have on average moved further away from their reservation wage and hence, would respond less to price changes, while the opposite might be true for married men.
A third answer is that the strong underlying growth in labor supply among married women up until the end of the 1990s had an impact on total employment growth, both in normal times and after the economy was hit by a negative shock. From a mechanical viewpoint, the underlying trend growth in female labor supply sped up the employment recoveries after a recession up until the 1990s. However, whether the trend growth in female employment has changed the employment reaction to negative shocks in any fundamental way or whether it is just a question of a changing trend in employment growth depends on the effect of the growth in employment among married women on the other subgroups in the economy (Albanesi 2018; Fukui et al. 2018).

To evaluate these plausible mechanisms and give a quantitative answer, we need a structural model capturing both the individual household responses and general equilibrium effects. Traditionally, macro models analyzing aggregate shocks and labor supply have used a representative agent setting, i.e., a single household representing all households in the economy. Heterogeneous-agent models analyzing aggregate shocks and labor supply standardly assume a sole-breadwinner household (even though there are exceptions). The few recent papers looking at the trend of an increasing labor supply among women and the effect on aggregate shocks use a representative household structure, and hence cannot speak to questions about intra-household interaction.

In this paper, I use a heterogeneous-agent model of the Bewley/Huggett/Aiyagari type that explicitly incorporates heterogeneity in gender and household composition (alongside heterogeneity in assets and productivity). There are three types of households: single men, single women, and couples (consisting of one man and one woman), and they are all subject to idiosyncratic shocks in terms of productivity and unemployment. Each period, households choose how much to consume and to save, and if they want to supply labor on the market or not. Within couples none, one or both individuals can choose to work. The labor choice is only on the extensive margin. The reason for this is that the most important changes over time have happened on the extensive margin: I argue that it is more important to capture the decision to actually start working than a potential decision to increase the hours from say 36 to 42 hours per week. Hours worked per worker by subgroup have been stable since the 1960s. Moreover, between two thirds and three quarters of the hours fluctuations over the business cycle happen on the extensive margin.

Using this framework, I model a closing of the wage gap between men and women as a result of a decreasing productivity wedge facing women. The increase in labor supply among married women is then an equilibrium outcome driven by the shrinking wage gap and other equilibrium variables such as wealth accumulation etc. I incorporate this closing of the productivity gap as an exogenous trend and do not take a stance on why it arises, but only note that it is consistent with a theory of female-biased technological change, a theory of discrimination and misallocation of female talent, or a combination of both. I model the closing of the productivity
gaps so that the observed wage gap in the model, taking into account any selection effect, is consistent with observed historical data.

A first test of whether the model captures the correct labor supply patterns in the long run is if an increase in female employment is primarily driven by married women. In my model, this turns out to be the case. The closing of the wage gap can explain two thirds of the increase in married women’s labor supply, while single women are hardly affected.

A second test is if the increase in labor supply among married women crowds out their husbands’ labor supply, something we have not seen in the data. In my model, married men decrease their labor supply by approximately five percentage points, which is in line with the data, and far from the increase in labor supply by married women. This result arises despite using a unitary household model with fixed bargaining weights (as opposed to, e.g., Knowles (2012) who uses endogenous bargaining) and no explicit modelling of home production (as opposed to, e.g., Jones et al. (2015)).

It turns out that explicitly incorporating both two-person and one-person households in a model with endogenous labor choice on the extensive margin matters for employment responses. As in the data, single households’ hours fluctuate more than couple households’ hours. The effect arises since on average the mass of single households is closer to the working margin than is the mass of agents in couples. The economic reason is that the single household’s working and savings decisions only depend on the agent’s own productivity. When the individual chooses to work, he/she also accumulates assets for future unproductive periods. When the individual chooses not to work, he/she is also deaccumulating assets saved up from earlier working periods. Hence, the savings decision is tightly linked to the working decision for a single household. In the couple household, on the other hand, the savings decision is not only linked to one individual’s current working decision, but to both individuals’ current productivities and continuation value.

I then use the model to analyze the employment response to a TFP shock at different points in time. I study the economy’s response to “MIT shocks” (as proposed by Boppart et al. (2018)) and the resulting total response to a TFP shock compared to the underlying deterministic trend is unchanged between the 1970s and the present time. This seeming null-result is the sum of multiple counteracting forces. Most importantly, among married women, the employment response from a TFP shock has become more muted over time, due to more and more women moving further away from the working margin. However, since women in absolute terms respond more strongly to aggregate shocks, their increasing share of the labor force drives up the aggregate response and the net result is close to zero.

A connected finding is that if the model economy is hit by a strong negative shock during a period with strong female labor force participation growth, it returns to its pre-recession level of employment after four years. However, if the shock happens later, when the labor force participation is no longer increasing, it takes longer for the employment figure to return to its pre-recession level. This pattern replicates
what we have seen in the data: from the 1990s and onwards, it has taken longer after a recession for the employment figures to climb back to their pre-recession level. In other words, the underlying trend in employment, driven by the increase in female labor force participation, was historically driving up the employment figures after a recession. However, if we measure the employment deviation compared to the underlying deterministic path, the response and speed of recovery have not changed over time.

1.1. Relation to previous literature

Standard macroeconomic models in the tradition of Kydland and Prescott (1982) analyzing the economy’s response to aggregate shocks are populated by an infinitely lived representative household, which derives utility from consumption and leisure and receives income from supplying labor and accumulating savings.

A first step to move away from the representative-agent framework came with the introduction of heterogeneous-agent models (Aiyagari 1994; Bewley 1986; Huggett 1993); however, “the lion’s share of work on quantitative heterogeneous-agent models has focused on the bachelor household – one breadwinner per household” (Heathcote, Storesletten, and Violante 2009, page 339). Naturally, there were early exceptions, for example explicitly modelling intra-family insurance (Attanasio, Low, and Sánchez-Marcos 2005) or the family formation as an idiosyncratic risk (Cubeddu and Ríos-Rull 2003). However, in the ten years since this quote was written, it has become more common to explicitly include the family and its members in dynamic macro models. Notable examples include Heathcote, Storesletten, and Violante (2010), modelling each household as consisting of a husband and a wife when analyzing increasing wage inequality, Guner, Kaygusuz, and Ventura (2011), incorporating both couples and singles in a life-cycle model when evaluating labor supply responses to a tax reform, Chakraborty, Holter, and Stepanchuk (2015), explicitly taking divorce rates and tax mechanisms into account in a framework with singles and couples when evaluating cross-country differences in hours worked, and Mankart and Oikonomou (2016), modelling couples with two potential earners when analyzing the cyclical properties of aggregate employment and participation.

Especially in the literature about female related issues (such as the increase in female labor force participation, discussed below) many life-cycle models explicitly incorporate the family structure. A more general point is made by Borella, De Nardi, and Yang (2018) who, using a life-cycle model, show that including gender and family formation can substantially improve a model’s fit to aggregate data in terms of savings and labor profiles over the life cycle.

The exercise in this paper starts from the assumption of a decrease in the gender productivity gap, and evaluates its impact on the employment figures for both men and women, both in couple households and single households. There is a
large literature on the underlying causes of the increased labor force participation among women in the post-war U.S. The papers which, like the current one, focus on the importance of the closing of the gender wage gap put forward different explanations for why this has happened. Jones, Manuelli, and McGrattan (2015) avoid taking a strong stance, but note that their approach is consistent both with the view that the wage gap is a consequence of discrimination (either directly in wages or through the presence of a glass ceiling) and with the view that the change in the gender wage gap arises from sex-specific productivity changes. More specifically, Galor and Weil (1996) and Ngai and Petrongolo (2017) stress the rise of the service sector, in which they say that women have a comparative advantage. Olivetti (2006) investigates a sex-specific increase in the return to experience, while Attanasio, Low, and Sánchez-Marcos (2008) find that the closing of the wage gap must be combined with a reduction in the cost of child-care to generate the changes we have seen.

A non-exhaustive list of other suggested reasons for the increase in female labor force participation includes the increased availability of household appliances (Greenwood, Seshadri, and Yorukoglu 2005), increased access to birth-control measures (Goldin and Katz 2002), improvement in maternal health (Albanesi and Olivetti 2016), evolving beliefs about payoffs from working among women (Fernández 2013), a higher divorce probability (Fernández and Wong 2014), and changes in culture and social norms (Fernández and Fogli 2009; Fernández, Fogli, and Olivetti 2004).

The implications of the increased labor force participation among women on the economic performance in general and the economy’s response to an aggregate shock have been studied in a few recent papers. Heathcote, Storesletten, and Violante (2017) study the impact of the rise in female labor supply on growth and income inequality, but do not include aggregate shocks. Fukui, Nakamura, and Steinsson (2018) focus on jobless recoveries and how much this phenomenon is a result of a slower (or even zero) increase in the labor force participation of women. Their model uses a representative couple formulation in which the inclusion of home production makes the “crowding-out” effect small enough to match the empirical estimates. Albanesi (2018) estimates a DSGE model allowing for female-biased shocks using aggregate data, and finds that the dynamics of these shocks have changed in recent years, suggesting that the gender convergence has played an important role both to explain jobless recoveries and the Great Moderation. Both of the two latter papers look at aggregate hours for men versus women in a representative-couple sense, and do not distinguish between married and single individuals.

The combination of heterogeneous agents, aggregate shocks and explicitly modelled families with potentially more than one wage earner is scarce. One (rare) model that includes heterogeneous families, each family consisting of a husband and a wife, and analyzes aggregate shocks of the Krusell-Smith type (Krusell and Smith 1998) is Chang and Kim (2006). However, that paper does not consider the effects of an increase in employment among married women.
2. Structural change of the labor market

In this section, I describe the overall labor supply patterns over time and over the business cycle, by subgroup.

For calculations regarding employment and hours by subgroup over time, I use the CPS March Annual Social and Economic supplement (ASEC). This annual survey goes back to 1962 when approximately 30,000 households were surveyed. In the most recent years, more than 90,000 households have been included in the survey. I exclude all observations for individuals younger than 25 or older than 64,¹ and all individuals who are employed in the armed forces.²

2.1. Increased employment rate for married women

First, I describe the broad trends in labor supply across subgroups over time. Figure 2a shows employment by gender and civil status from 1962 up until now. The one observation that stands out is the strong increase in employment among married women: from 38% at the beginning of the 1960s up to 68% in the mid-1990s. This can be thought of as a structural change of the labor market: from a structure with a majority of two-person households with one bread-winner to a structure in which the most common type of household is a two-person household with two earners.

The most important changes over the last 60 years in terms of hours worked by different subgroups happened along the extensive margin, i.e., the number of persons participating in the labor force. Hours worked per worker have been remarkably stable, as Figure 2b illustrates. Married working men work more hours than single working men, and married working women work the least hours. However, the ordering and approximate levels have remained constant since the early 1960s.

Figure 3 shows the year-on-year fluctuations in average weekly hours and the contribution by subgroup. One can immediately notice how married women are contributing to the increase in hours, especially during the early period, and how this increase is even counteracting the fall in hours among the other groups in some years. In the next section, we will look more closely at this dynamic.

¹The main reason for not including individuals younger than 25 is that I thereby exclude the part of the population where the main trade-off is between work and school and hence, I do not have to take a stance on the categorization of schooling.
²For all analyses, I weight the observations using ASECWT.
(a) Employment-to-population ratio  

(b) Average weekly hours per worker  

**Figure 2:** Work by subgroup (age 25-64). Source: CPS.

**Figure 3:** Year-on-year change in average weekly hours per capita (age 25-64), contribution by subgroup. Source: CPS.
2.2. Jobless recoveries: nothing new for men

Figure 4 shows the employment-to-population figures after a recession, normalized to one in the year leading up to the recession. The thicker black line shows the employment-to-population ratio for the prime-aged working population (defined as age 25-64). As can be seen from the figures, after the recessions in 1970, 1974 and 1980, employment was back at or even above its pre-recession level four years later. After the 1990 recession, it was not yet back four years later, and even less so for the 2000 and 2007 recessions, giving rise to the idea of “jobless” recoveries.

However, if we look at the four demographic subgroups separately, there are striking differences. As Figure 4 shows, neither the single nor the married men had returned to their pre-recession level four years later even before 1990. The one group driving up the employment figures in the 1970s and 1980s was married women.

Can we then draw the conclusion that the increased labor force participation among married women is what made the recoveries before 1990 “non-jobless”? At least from a purely mechanical perspective this seems to be the case, but as pointed out by Fukui, Nakamura, and Steinsson (2018), it depends on what we believe in terms of crowding-out. An increase in employment among married women has an income effect on households. In a standard macro model, this would crowd out at least some of the work supplied by married men. Hence, in equilibrium, the effect of an increase in employment among married women on aggregate employment is unclear. Using regional variation in the gender wage gap, Fukui et al. (2018) show that empirically, the crowding-out effect seems to be small.

2.2.1. International perspective

If we look at the rest of the world outside the U.S., the phenomenon of jobless recoveries in more recent years is not as prevalent. Graetz and Michaels (2017) look at evidence from a broader set of countries, and compile data from 17 countries for the years 1970 to 2011. Their conclusion is that recent recoveries have generally not involved any significantly slower recovery of employment. In this respect, the jobless U.S. recoveries seem to be the exception.

Figure 5 shows the ratio of female to male labor force participation for the G7 countries, normalized to the 2016 level. As can be seen, the U.S. was the first country to reach its current level, while all other countries have been on an upward trajectory up until now.

---

3I use the year with peak employment just before the recession hits as year zero, which could coincide with the beginning of the recession according to the NBER (the survey is conducted in March every year).

4For a broader set of countries, and for absolute levels of female labor force participation, see the appendix, section A.2.
Figure 4: Employment figures after the economy has been hit by a recession. Source: CPS.
2.3. Hours volatility and responses to aggregate shocks

We now turn to how much hours worked by subgroup fluctuate over the business cycle. Figure 6 shows two measures of volatility: total volatility and the volatility related to aggregate fluctuations by demographic subgroup, both measured over the years 1962 to 2015. Total volatility for a given demographic subgroup is the percentage standard deviation of the cyclical component of log average hours per capita within the group after removing the trend with an HP filter (using a smoothing parameter of 6.25, following Ravn and Uhlig (2002) for annual data).

I define the volatility related to aggregate fluctuations in the following way. I compute the cyclical component of the labor series for each demographic subgroup as the residual after applying an HP filter (using a smoothing parameter of 6.25). Then, I regress this cyclical component on the cyclical component of GDP (which is also the residual after applying an HP filter with a smoothing parameter of 6.25).\(^5\)

Hence, the equation I estimate is:

\[
\hat{x}_{c,t,i} = \beta \hat{Y}_{c,t} + \epsilon_t
\]

with

\[
\hat{x}_{c,t,i} \quad \text{cyclical component from HP filtering of yearly series of } \text{log}(x_i)
\]

with \(i \in \{\text{married men, married women, single men, single women}\}\)

\(^5\)This method follows Jaimovich and Siu (2009). The results are robust to including additional RHS variables such as lagged GDP.
\( \hat{Y}_{c,t} \) cyclical component from HP filtering of yearly series of log GDP.

where \( x_i \) are annual hours per capita for each demographic subgroup. The coefficient of interest, \( \beta \), captures how responsive aggregate hours for this particular subgroup are to fluctuations in GDP. After estimating this equation, I create a series of predicted hours by subgroup, and then finally measure the percentage standard deviation of these series. Hence, one can think of the cyclical volatility as capturing the component of hours volatility that is related to aggregate economic fluctuations.\(^6\)

As Figure 6 shows, men have a higher volatility than women, and singles have a higher volatility than individuals in couples, confirming the findings in, e.g., Doepke and Tertilt (2016). These two facts hold for both total volatility and volatility related to aggregate fluctuations. The first observation, that men have a higher volatility than women, seems at odds with the conventional wisdom that women have a higher labor supply elasticity than men. Therefore, I now look at sector representation. Figure 7 shows the relation between, on the one hand, how responsive a sector is to aggregate fluctuations and, on the other hand, sector choice by demographic subgroup. The x-axis in the graphs indicates the sector responsiveness to aggregate fluctuations, measured as described above (running regression (1) with \( \hat{x}_{c,t,i} \) denoting a cyclical component of log hours in sector \( i \) at time \( t \)). The y-axis indicates the share of workers in that sector from a specific demographic subgroup. For instance, in the construction sector, 28% of the employed are single

\(^6\)As a robustness check, I rerun the analysis using different filtering techniques and the conclusions remain unchanged, see section A.5.
Agriculture Mining Construction Manuf durables Manuf nondurables Transport Telecomm Utilities Wholesale Retail Finance Business serv Personal serv Entertain Medical serv Prof serv Public

fraction of workers type 10 0.05 0.1 0.15 0.2 0.25 0.3

sector responsiveness to aggregate fluct

Figure 7: The horizontal axis indicates the sector responsiveness to aggregate fluctuations for the period 1962-2017. The vertical axis indicates the fraction of workers being of the particular demographic subgroup (measured as the average for the time period 1962-2017). The size of the marker indicates share of subgroup working in the sector (measured as the average for the time period 1962-2017). Population aged 25-64. Source: CPS.

men, compared to medical services where the corresponding figure is only 7%. The size of each bubble in the diagram indicates the fraction of the subgroup working in that particular sector (in other words: the sizes of the bubbles can be said to sum to one).

One thing that immediately stands out is that men are over-represented in more volatile sectors while the opposite is true for women, and this observation is true for both married individuals and singles. Hence, one could argue that the fact that women in the aggregate are less volatile than men is driven more by sector-specific labor demand than by individual labor supply considerations (which is the primary focus of this paper).

A final observation related to aggregate fluctuations is that approximately two thirds of the fluctuations in total hours worked are due to changes in the extensive
Figure 8: Responsiveness of hours to aggregate fluctuations, time period 1962-2017. Population aged 25-64. Source: CPS.

margin. Figure 8 shows the shares by subgroup, and for all subgroups, the extensive margin is more important than the intensive margin in explaining the total change in hours related to aggregate economic fluctuations.

2.4. Summary of empirical observations

It is time to take stock of the empirical observations up to this point. First of all, since the 1960s, married women have increased their employment rate by roughly 30 percentage points: from 38% in 1962 to almost 70% at the beginning of the 1990s and thereafter. Second, we have not seen the same strong increase among single females. Third, the crowding-out within couples is small. These three observations should be captured by a model of increasing female labor supply.

In terms of volatility and responsiveness to aggregate shocks, there are two observations that should be highlighted. First of all, singles’ hours worked are more volatile than married individuals’ hours worked. This is true comparing married men to single men, and comparing married women to single women. For this cut, there are no apparent sector differences that explain the pattern.

The next observation is that at an aggregate level, men’s hours are more volatile than those of women. This is true both within couples, in which married men’s hours are more volatile than those of married women, and comparing single men to single women. However, the higher aggregate volatility among men is highly correlated with sectors. Men are more likely to work in sectors that are highly business-cycle sensitive. In this paper, I ignore any sectoral differences in volatility.
A last observation from the data is that the extensive margin is of first-order importance. Between two thirds and three quarters of the fluctuations in hours are due to extensive-margin fluctuations, and this is true for all subgroups. Moreover, the main changes over time have happened on the extensive margin. Hours worked per worker have not changed nearly as much. Therefore, in the choice between the intensive and extensive margin, I choose to include the extensive margin in the model.

3. Model

In this section, I lay out the benchmark steady-state model. The model is an infinite horizon general equilibrium model of the Bewley (1986)/Huggett (1993)/Aiyyagari (1994) type. Time is discrete and every time period is assumed to be one year.

Individuals derive utility from consumption and disutility from supplying labor, which is endogenously chosen. There is only an extensive labor margin choice. Individuals face two types of idiosyncratic risks: shocks to productivity and unemployment risk. There is no possibility to insure against these risks. Households can only save in a riskless bond, and they face an exogenous borrowing constraint.

There are three types of ex-ante heterogeneous households: single males, single females, and couple households (consisting of one male and one female). I start with the problem facing a single man or woman, which is identical except for some gender-specific parameters.

3.1. The single household problem

Working-age households have a constant probability of entering a retirement phase $\pi_r \in [0, 1]$ and retired households have a constant probability $\pi_w \in [0, 1]$ of being re-born as a working-age household. With gender denoted by $i \in \{m, f\}$, working-age single households solve the following recursive problem:

$$V^i_w(a, \omega_i, u_i) = \max_{c, \ell, a'} \left\{ u(c, \ell) + \beta_s \left[ (1 - \pi_r) EV^i_w(a', \omega_i', u_i') + \pi_r V^i_r(a') \right] \right\} \quad (2)$$

s.t. \[ c + a' \leq (1 - u_i)(1 - \tau_i)\omega_i\ell + u_i\phi(1 - \tau_i)\omega_i - T + a(1 + r) \quad (3) \]
\[ \ell \in \{0, 1\} \quad (4) \]
\[ a' \geq a \quad (5) \]

Unemployment is denoted by $u_i \in \{0, 1\}$. The budget constraint, (3), says that consumption ($c$) plus savings for the next period ($a'$) cannot exceed labor income if employed ($w(1 - \tau_i)\omega_i\ell$) (which is positive if the household chooses to work) or
unemployment benefits if unemployed \((\varphi w(1 - \tau_i)\omega_i)\), minus the lump-sum tax, plus gross capital income.

The \(\tau_i\) (which later in the calibration will be set to zero for men and to a positive value for women) enters as a wedge between the agent’s underlying productivity and the realized market productivity.\(^7\)

As can be seen from the constraint (4), labor is restricted to be zero or one, in other words, the model only includes the extensive margin labor choice.\(^8\)

The instantaneous utility function \(u(c, \ell)\) has a constant relative risk aversion (CRRA) and is separable in consumption and participation in the labor market:

\[
\begin{align*}
\u(c, \ell) = \frac{c^{1-\sigma} - 1}{1-\sigma} - \psi_i \ell.
\end{align*}
\]

After working-age single households stochastically enter retirement, they solve the following problem:

\[
\begin{align*}
V^i_r(a) &= \max_{c, a'} \left\{ u(c, 0) + \beta_s \left[ (1 - \pi w) V^i_r(a') + \pi w E V^i_w(a', \omega'_i, u'_i) \right] \right\} \\
\text{s.t.} & \quad c + a' \leq (1 - \tau^i_r) \kappa + a (1 + r) \quad (7) \\
& \quad a' \geq a \quad (8)
\end{align*}
\]

where \(\kappa\) denotes retirement benefits. When a retired household is reborn as a working-age household, its productivity and unemployment state \((\omega'_i, u'_i)\) are drawn from the ergodic distribution. Hence, there is no intergenerational transfer in terms of productivity.

### 3.2. The couple household problem

The couple household consists of a husband and a wife, who derive utility from consumption and disutility from supplying labor.

---

\(^7\) Another way of modelling the wage gap is to assume that there is a wedge, or tax, between the marginal product and the wage the individual receives (as in, e.g., Hsieh et al. (2013) or Jones et al. (2015)). However, in a GE model, the question arises of what to do with the tax revenue.

\(^8\) It would be conceptually straightforward to include a choice on the intensive margin as well. For example, the labor choice could be restricted to \(h \in \{0, [h, 1]\}\) where \(h\) represents a lower bound on the hours possible to supply to the market, or one could introduce a fixed hours cost of working, as done in, e.g., Chang et al. (2018), use a part-time penalty, as in, e.g., French (2005), or a combination of fixed cost and non-linear earnings, as in Erosa et al. (2016). However, it would be non-trivial in terms of computational burden. Since most of the changes in the relative numbers of hours worked between groups happened along the extensive margin and between two thirds and three quarters of the business fluctuations are on the extensive margin, I choose to focus only on that margin in this model.
When modelling households that consist of more than one person, one has to assume what type of decision process the household uses. Intra-household decision processes can be categorized into three broad classes: the unitary model, non-cooperative processes or cooperative processes. The unitary model, by far the most commonly used in macro models, assumes the existence of a common household utility function. One important implication of the unitary assumption is that decisions only depend on prices and total household income, and are independent of the distribution of income; in other words, they display income pooling, and decisions do not depend on who brought in the money. However, many micro studies have questioned the plausibility of the unitary model in favor of non-cooperative or cooperative models (see, e.g., Chiappori et al. (2002)).

Non-cooperative models assume that no binding agreements between members exist and that the optimal decision need not be Pareto efficient. Both individuals optimize their own utility (which can include caring/altruism towards the other partner). Typically, the actions constitute a Nash equilibrium. One rare example of a macro model using a non-cooperative decision process is Aiigari et al. (2000). In this study, the household problem is static in nature, which greatly simplifies the analysis; there is no borrowing or lending on a capital market.

Cooperative models assume that negotiations taking place within households result in a Pareto efficient outcome, in the usual sense that no other feasible choice would have been preferred by all household members. There are few macro models using a cooperative decision process, one of them being Knowles (2012) (who also rules out savings).

In this paper, couple households are modelled as unitary households, despite the potential shortcomings of that assumption. The main reason is that a unitary household assumption is a natural starting point for an analysis of the effect of including both singles and couples in a model, due to its simplicity. Second, since savings and intertemporal decisions are essential for the questions at hand, deviations from the unitary assumption would substantially complicate the model.

Working households have a constant probability of retiring $\pi_r \in [0,1]$ and retired households have a constant probability $\pi_w \in [0,1]$ of becoming a working-age household again. Hence, I am assuming that both individuals in the couple enter the retirement phase at the same time.

With the state vector $x_c$ given by $x_c = (a, \omega_m, \omega_f, u_m, u_f)$, working-age couple households solve the following recursive problem:

$$V_c^w(x_c) = \max_{c, \ell_m, \ell_f, a'} \left\{ u(c, \ell_m, \ell_f) + \beta_c \left[ (1 - \pi_r)E V^c_w(x'_c) + \pi_r V^c_r(a') \right] \right\}$$  \hspace{1cm} (9)

$$c + a' = w(\omega_m(1 - \tau_m)(1 - u_m)\ell_m + \omega_f(1 - \tau_f)(1 - u_f)\ell_f) + \phi w((1 - \tau_m)\omega_m u_m + (1 - \tau_f)\omega_f u_f) - 2T + a(1 + r)$$  \hspace{1cm} (10)

$\phi$ is the wage rate of the household head. The first term on the right hand side of equation (10) is the wage that the household head would earn if they worked alone, and the second term is the household head’s wage if they worked, after paying the wage to the other partner. With a non-cooperative decision process, the household head’s wage depends on the work effort of the other partner, which in this model is the same work effort that the household head would make if they were the only one working. The wage rate in equation (10) is the wage rate of the household head. The first term on the right hand side of equation (10) is the wage that the household head would earn if they worked alone, and the second term is the household head’s wage if they worked, after paying the wage to the other partner. With a non-cooperative decision process, the household head’s wage depends on the work effort of the other partner, which in this model is the same work effort that the household head would make if they were the only one working. The wage rate in equation (10) is the wage rate of the household head.
$\ell_i \in \{0, 1\} \; \forall i \in \{m, f\}$  \hspace{1cm} (12)

$a' \geq a$  \hspace{1cm} (13)

Hence, the couple household’s problem is analogous to that of the single household except that a couple household consists of two persons, each with a labor supply decision. Similar to the retired single-person problem, the retired couple households solve the following problem:

$$V_r^c(a) = \max_{c, a'} \left\{ u(c, 0, 0) + \beta_c \left[ (1 - \pi_w) V_r^c(a') + \pi_w E V_w^c(x_{c'}') \right] \right\}$$  \hspace{1cm} (14)

s.t.  \hspace{1cm} $c + a' = (1 - \tau_r m) \kappa + (1 - \tau_r f) \kappa + a(1 + r)$  \hspace{1cm} (15)

As for single households, when a retired household is reborn as a working-age household, the productivity and unemployment states in the next period for the husband and the wife ($\omega'_{m}, \omega'_{f}, u'_{m}$ and $u'_{f}$) are drawn from the ergodic distribution.

The instantaneous utility function is given by:

$$u(c, \ell_m, \ell_f) = \zeta_1 \frac{(c/\zeta_2)^{1-\sigma} - 1}{1 - \sigma} - \psi_m \ell_m - \psi_f \ell_f$$

$\zeta_1$ and $\zeta_2$ together parametrize how the couple derives utility from consumption compared to the single household. In the simplest case, the two individuals in the couple household could split their money equally, buy their own consumption good on the market just as a single individual, and the household utility would be the sum of the two individuals’ utilities. This would mean $\zeta_1 = \zeta_2 = 2$. However, if there are economies of scale in the household and some goods are common (think about, e.g., broadband access or heating), we have $\zeta_2 < 2$. If the couple household has a different relative valuation of consumption vs. disutility of labor compared to singles (due to, e.g., a higher probability of having children who need to consume as well) it would show up in a different $\zeta_1$.

### 3.3. Technology

The production side of the model is completely standard. Competitive firms employ labor and capital hired from households to produce a homogeneous final good, which is used for both consumption and investment. The aggregate production function is assumed to be Cobb-Douglas:

$$F(K, L) = z K^\alpha L^{1-\alpha}.$$  \hspace{1cm} (17)

In the steady-state version of the model we have $z = 1$ at all times. Capital is assumed to depreciate at the rate $\delta$. 
3.4. Government

The government pays out unemployment and retirement benefits every period and finances itself via lump-sum taxes. It runs a balanced budget in every period.

3.5. Equilibrium definition

We can now define a stationary equilibrium. Denote a household’s state vector by \( (a, b) \) where assets \( a \in A \) and type \( b \in B \), with \( A \) and \( B \) given by:

\[
A = \{a, \bar{a}\}
\]

\[
B = \{\left(c, (\omega_m, \omega_f), (u_m, u_f)\right), \left(m, \omega_m, u_m\right), \left(f, \omega_f, u_f\right)\}
\]

using \( c \) to denote couple households, \( m \) single males, and \( f \) single females. Further, define \( C, M, \) and \( F \) as:

\[
C = \{A \times (x, \cdot, (u, \cdot)) | (x, \cdot, (u, \cdot)) \in B, x = c\}
\]

\[
M = \{A \times (x, \cdot, (u, \cdot)) | (x, \cdot, (u, \cdot)) \in B, x = m\}
\]

\[
F = \{A \times (x, \cdot, (u, \cdot)) | (x, \cdot, (u, \cdot)) \in B, x = f\}
\]

and

\[
\mu_c = \int_C d\Gamma, \quad \mu_m = \int_M d\Gamma, \quad \mu_f = \int_F d\Gamma.
\]

Let \( \tilde{\omega}_i \) define the realized market productivity: \( \tilde{\omega}_i = (1 - \tau_i)\omega_i \quad \forall i \in \{m, f\} \).

A recursive competitive equilibrium is given by a set of prices \( \{r, w\} \), decision rules \( C(a, b), E_m(a, b), E_f(a, b) \) and \( A(a, b) \), and a stationary distribution \( \Gamma \) such that:

1. The decision rules solve the households’ problem for all \( (a, b) \).
2. Firms optimize, i.e., factor prices are given by:

\[
r = F_1(K, L) - \delta \quad \text{and} \quad w = F_2(K, L)
\]

3. The government budget balances:

\[
\frac{\pi_w}{\pi_r + \pi_w} \left( \mu_c 2 + \mu_m + \mu_f \right) = \omega \varphi \left( \int_{C_{um}} \tilde{\omega}_m d\Gamma + \int_{C_{uf}} \tilde{\omega}_f d\Gamma + \int_{M_u} \tilde{\omega}_m d\Gamma + \int_{F_u} \tilde{\omega}_f d\Gamma \right) + \left( \frac{\pi_r}{\pi_r + \pi_w} \right) \kappa \left( \mu_c \left( 2 - \tau_f - \tau_m^r \right) + \mu_m \left( 1 - \tau_m^r \right) + \mu_f \left( 1 - \tau_f^r \right) \right)
\]

where \( C_{um} = \{A \times (x, \cdot, (u, \cdot)) | (x, \cdot, (u, \cdot)) \in B, x = c \) and \( u = 1 \} \) etc.
CHAPTER 1

4. Capital and labor markets clear:

\[ K' = \int_{A \times B} A(a, b) \, d\Gamma \]

\[ L = \int_C \left( \bar{\omega}_m E_m(a, b) + \bar{\omega}_f E_f(a, b) \right) \, d\Gamma + \int_M \bar{\omega}_m E_m(a, b) \, d\Gamma + \int_F \bar{\omega}_f E_f(a, b) \, d\Gamma \]

5. For all relevant Borel sets B

\[ \Gamma(B, b) = \sum_B \pi(b|\bar{b}) \int_{a: A(a, \bar{b}) \in B} \Gamma(da, \bar{b}) \]

Note that when I later move on to the deterministic transition path, I will not use the stationary equilibrium but a sequential equilibrium definition in which the equilibrium objects depend on time. To facilitate the understanding, I nevertheless spell out the stationary version here.

4. Baseline calibration

In this section I describe how I calibrate the baseline steady-state version of the economy. Since I primarily want to address long-term trends, I calibrate the model to annual data.

Households face two types of idiosyncratic risks: an earnings risk determined by the transition matrix \( \pi(\omega'|\omega) \) and an unemployment risk determined by the transition matrix \( \pi(u'|u) \). I describe both components in turn. Thereafter I describe the calibration of the remaining parameters.

4.1. Idiosyncratic productivity risk and offered wages

Individuals face a labor productivity risk. I assume that log labor productivity follows an AR(1) process specified as:

\[ \log(\omega'_i) = \rho_i \log(\omega_i) + \varepsilon_{\omega_i}, \quad i \in \{m, f\} \]  \hspace{1cm} (18)

with persistence \( \rho_i \) and innovation \( \varepsilon_{\omega_i} \sim N(0, \sigma_{\varepsilon_i}^2) \).

I use values from Borella et al. (2018) who estimate the process separately for men and women using data from the PSID. As can be seen in Table 1, the males’ productivity process is more risky, with a slightly higher persistence and variance.
CHAPTER 1

For single households, I use the Rouwenhorst procedure\(^9\) to discretize the process into a 15-state Markov chain: \(\omega_i \in \{\omega_1^i, \ldots, \omega_{15}^i\}\).

For couple households, the productivity process for the husband is assumed to follow the same process as for single men, and the productivity process for the wife is assumed to be the same as the one for single women. However, the innovations of the processes for the husband and wife are assumed to be correlated with \(\text{Corr}(\epsilon_m, \epsilon_f) = \rho_e\), in other words:

\[
\begin{bmatrix}
\epsilon_m \\
\epsilon_f
\end{bmatrix}
\overset{\text{iid}}{\sim} N(0, \Sigma)
\]

\[
\Sigma = \begin{bmatrix}
\sigma_m^2 & \rho_e\sigma_m\sigma_f \\
\rho_e\sigma_m\sigma_f & \sigma_f^2
\end{bmatrix}
\]

Since an individual has 15 productivity states, the couple has \(15^2\) potential combinations of husband and wife productivities: \((\omega_m, \omega_f) \in \{\omega_1^m, \omega_2^m, \ldots, \omega_{15}^m\} \times \{\omega_1^f, \omega_2^f, \ldots, \omega_{15}^f\}\). To estimate the transition matrix for the couple productivities, I use a simulation procedure which is a generalization of the method used in De Nardi et al. (2018). Given the AR(1) formulation and the relevant parameters \((\rho_m, \rho_f, \sigma_m^2, \sigma_f^2, \rho_e)\), I simulate a very long sequence (> \(10^9\) periods) of husband and wife productivities given random draws of the correlated innovations. For every draw, I bin the husband and wife productivity in the corresponding bin among the \(15^2\) combinations. Then, I calculate the observed transition probabilities for each state\(^{10}\).

The observed gender wage gap in the model is made up of three distinct pieces, two exogenous and one endogenous. First, there is a small gender productivity gap due to the lower variance of the innovation term to log productivity. Second, there is an exogenous productivity gap, denoted by \(\tau_f\) in the model. Third, there is a selection effect in the model, so that the observed wage gap does not equal the wage gap of offered wages.

Given the higher variance of men’s log productivity, the average offered wage to women is 5% lower than the average offered wage to men, before applying any productivity wedge. However, the selection effect is stronger for women: relatively more women with low wages choose not to work. Therefore, in the absence of

\(^9\)I prefer the Rouwenhorst procedure to, e.g., the Tauchen method since the productivity processes are highly persistent; see Kopecky and Suen (2010) for a detailed description and evaluation of the method.

\(^{10}\)With \(\rho_e = 0\), the joint transition matrix is simply given by the Kronecker product of the transition matrices for males and females, respectively. If one were to accept that the discretized productivity states for couple individuals differ from those for single individuals, one could define \(x^A_t = \log(\omega_m) + \log(\omega_f)\) and \(x^B_t = \log(\omega_m) - \log(\omega_f)\). We would then have \(x^A_{t+1} = \rho x^A_t + (\epsilon_m + \epsilon_f)\) and \(x^B_{t+1} = \rho x^B_t + (\epsilon_m - \epsilon_f)\) where \(\text{Cov}(\epsilon_m + \epsilon_f, \epsilon_m - \epsilon_f) = 0\). Using the Kronecker product for the joint transition matrix, we could then convert \(x^A\) and \(x^B\) back to \(\log(\omega_m)\) and \(\log(\omega_f)\). However, then the productivity states for a single individual and a married individual would differ in levels and therefore I prefer to use the simulation based approach.
a female productivity wedge, the observed wage gap between men and women would be much smaller than 5%. I choose $\tau_f$, the female productivity wedge, so that the observed median wage gap is 20% and the observed mean wage gap is 21%.

In the same way, the observed correlation of working husband wages and working wife wages in the model differs from the chosen $\rho_\varepsilon$ due to the selection effect. It is more likely that a wife with a high productivity draw works than one with a low-productivity draw, and therefore the resulting observed correlation is higher. However, in practice, this difference is small, and I therefore ignore it.

4.2. Idiosyncratic unemployment risk

The unemployment risk is taken from Krueger et al. (2016). Since they use quarterly data and I have an annual model, I have to convert their transition probabilities, taking into account all possible routes from e.g. employed first quarter to employed first quarter one year later (e.g. being employed all four consecutive quarters, or losing one’s job in the first quarter, being unemployed in the second quarter, finding a job in the third quarter again, etc.). The resulting transition matrix is:

$$
\begin{bmatrix}
\pi_{uu} & \pi_{ue} \\
\pi_{eu} & \pi_{ee}
\end{bmatrix} =
\begin{bmatrix}
0.0538 & 0.9462 \\
0.0533 & 0.9467
\end{bmatrix}.
$$

(19)

There is no difference in the model between men and women in terms of separation rates or job-finding rates. In the U.S., there are no statistically significant differences between men and women in terms of search intensity, neither on the extensive nor

---

11The model in Krueger et al. (2016) is a model with aggregate uncertainty in which employment/unemployment transitions depend on the transition of an aggregate two-state TFP process. When adopting their employment/unemployment transitions, I condition on an aggregate transition from a good to a good state.

However, there are clear gender differences between men and women in terms of unemployment level. Albanesi and Şahin (2018) document a gender unemployment gap: women have had a higher unemployment rate than men up until the 1980s, and thereafter the gap virtually disappeared, except in recessions, when it now is reversed (men have a higher unemployment). However, looking at subgroups, single men are more unemployed than any other group, see Figure 9 (for the population aged 35-64, the pattern is the same, see section A.4). The second most unemployed group is single women. According to Albanesi and Şahin (2018), the convergence in unemployment rates between men and women is mainly driven by an increased labor force attachment among women. Hence, in the absence of a three-state model with frictions on the labor market, this convergence is therefore better captured by the labor supply choice than an exogenous unemployment shock.

To summarize, the equal unemployment probabilities used in this model constitute a reasonable starting point. Moreover, I assume that a husband’s and a wife’s unemployment risks are uncorrelated.

4.3. Discount factors and savings targets

The utility function used in the model aggregates, which means that if the couple households had perfectly correlated productivity shocks, unemployment shocks and labor choices, they would behave as a twice as large single household (assuming $\zeta_1$, the consumption utility scale parameter, to be 2). Consequently, couple
households would on average save twice as much as single households. However, shocks within households are not perfectly correlated, as discussed above and, importantly, the husband and the wife do not perfectly correlate their labor supply. Hence, the couple has access to an intra-household insurance mechanism, which makes them less willing to save for precautionary reasons.

Of the three insurance mechanisms mentioned above – the productivity risk, the unemployment risk, and the labor supply mechanism – the lower joint productivity risk is the most important one, given any reasonable calibration of the model. If we assume that the husband’s and the wife’s productivity processes were uncorrelated within the couple, the couple’s need for precautionary savings would decrease substantially. The reason is simply that the variance of their joint productivity would decrease by half compared to the perfectly correlated case. The opportunity to supply labor in an uncorrelated way is approximately one third as important in terms of how much less the couple wants to save, while the uncorrelated risk of unemployment affects the precautionary savings very little (with the calibration in this model, assuming a reasonably high replacement rate in case of unemployment).

In general equilibrium in an economy consisting of only couples, the interest rate would adjust to the couples’ lower demand for savings compared to single households. To induce enough savings the interest rate would go up, and the resulting savings per capita would be similar to the savings in a singles-only economy. However, in an economy populated by both singles and couples, the interest rate cannot adjust all the way.

Table 2 illustrates the effect. In the “singles economy”, there are only single households, and the equilibrium interest rate is 1.54%. The “couples economy” consists of only couples, and the interest rate is more than one percentage point higher. If we compare the per-capita savings between those economies, the difference is not that big: singles save 5.9 per capita, while couples save 4.6 per capita, i.e., 77% of what the singles save in their singles-only economy.

12Assume that the individual perceives a productivity process with the following first two moments: \( \mathbb{E}[\omega_i] = \mu \) and \( \text{Var}(\omega_i) = \sigma^2 \). Then a couple with perfectly correlated productivity processes would have the following two first moments: \( \mathbb{E}[\omega_i + \omega_j] = 2\mu \) and \( \text{Var}(\omega_i + \omega_j) = 4\sigma^2 \). However, if the productivity processes are uncorrelated, the first moment is unchanged, while the second moment is \( \text{Var}(\omega_i + \omega_j) = 2\sigma^2 \).

13Compared to a canonical Aiyagari (1994) model like the one reported in the original paper with the calibration closest to mine (log utility, the resulting interest rate 3.3%), the interest rate in my model is lower. One reason is the more risky productivity process, with a higher persistence (0.97 instead of 0.9) and a higher variance (0.45 instead of 0.4 unconditional). The other reason is the endogenous labor supply. Being able to choose leisure instead of labor in periods with relatively low productivity increases the demand for savings to be able to smooth consumption from good to bad states, and hence the equilibrium interest rate is lower.

14To clearly illustrate the effect, the couples consist of two individuals who are in every aspect like the individuals in the singles economy, the productivity risk is uncorrelated within the couple, and \( \zeta_1 \), the consumption scale factor, is assumed to be 2. In other words: the only thing that differs between the “singles economy” and the “couples economy” is that the couples are optimizing in households consisting of two persons.
However, if we mix those two types of households (assuming, for simplicity, an equal number of single and couple households), we get the third “singles and couples economy”. The equilibrium interest rate in this mixed economy is, as expected, in between the interest rates in the first two economies. The resulting effect on singles is, not surprisingly, that the higher interest rate makes them save more than in the singles-only economy, and work slightly less. The effect on couples is reversed: the lower interest rate makes them want to save less and work slightly more. What might be more surprising is the magnitude of the effects. In per-capita terms, the couples now save only 28% of what the singles do in per-capita terms.

The example given above assumed uncorrelated productivity processes within the couple, while in the baseline calibration in the model, the correlation is assumed to be 0.36. This correlation is far from enough to overturn the result of lower savings per capita for couples than singles in the model. However, turning to the data, in reality couple households save approximately twice as much as single households (in other words, the per-capita savings level is similar).\(^{15}\) To capture this fact, I assign a permanently higher discount factor to couple households. It could be thought of as capturing underlying permanent differences in how households discount the future, but it could also be a reduced form of modelling some other underlying difference between those types of households.\(^{16}\)

\(^{15}\)This fact is not driven by age differences between single and couple households, and it is not driven by any specific type of asset class. See section A.3 for asset holdings by household type and age.  
\(^{16}\)The same increase in couple savings could also have been achieved by increasing their risk aversion, or by introducing a large negative shock that would be (close to) perfectly correlated within a couple. For the latter option, one thing that is easy to think of is retirement, i.e., a long period without income that is close to perfectly correlated within couples. However, a retirement phase with a reasonable replacement rate is not sufficient to drive up the couple household savings enough, as can be seen from the results. To drive up the savings for couple households enough, the retirement phase must be both counterfactually long and have counterfactually low benefits. Results from models with differences in risk aversion and with a retirement phase with zero benefits are available upon request.
4.3.1. The absence of divorce

The model used in this paper does not incorporate any divorce risk, something one might suspect would affect the asset accumulation. However, it is not entirely clear how divorce should be modelled and what the implications for savings would be. First of all, one might argue that divorce is an endogenous decision, probably correlated with e.g. unemployment and bad productivity draws. Or one could argue that it is endogenous, but driven by some other unobserved “shock”, so that from a modelling perspective it can be treated as random.

However, even if one models divorce as a stochastic shock, the implications for savings differ depending on the modelling choices. Fernández and Wong (2014) argue that there are conflicting interests within the family. Assuming that the wife has a lower income in the state of divorce, the married woman would prefer to increase her savings to transfer more assets to the divorced state, while the married man would, on the other hand, prefer to increase consumption in the married state as this is what allows him to smooth consumption. In a partial equilibrium life-cycle model, they show that the net effect is a decrease in savings, assuming an equal split of assets in the case of a divorce and constant Pareto weights.

Cubeddu and Ríos-Rull (2003) also model divorce as a stochastic shock. In a general equilibrium model with family type following a stochastic process (and no further idiosyncratic shocks), they show how the size of savings differs dramatically depending on the details governing the marital type process. The net effect of divorce depends on, e.g., relative decision weights, rules governing the splitting of assets after separation, time horizons and subsequent marriage patterns.

Voena (2015) stresses that the difference in savings depends on the precise legal framework of the divorce process. Using the panel variation in U.S. laws, she shows how the introduction of unilateral divorce is associated with higher household savings as compared to a regime with mutual divorce decisions.

To summarize, it is not entirely clear how the absence of divorce affects the relative savings in the model, and there is even evidence suggesting that couple savings would go down if divorce was introduced.

4.4. Remaining parameters

Table 3 lists the remaining parameters used in the model. I use log utility for consumption so that the preferences are of the “balanced growth” type. Hence, income
and substitution effects cancel and the labor supply would be constant with balanced growth in the model, which allows me to abstract from it.\(^\text{17}\)

The disutility of work for men and women, \(\psi_m\) and \(\psi_f\) respectively, together with the consumption scale for couple households (\(\zeta_1\)), are jointly chosen so that the employment rates for the four demographic subgroups in the model approximately match the employment rates in the data from 2017. As can be seen, the disutility of females is assumed to be 7.6\% higher than the disutility of men in order to hit the correct employment rates by subgroup. This higher disutility of work for females can be seen as a reduced form of modelling various plausible mechanisms distorting the work-leisure choice for women compared to men: discrimination in the workplace, unevenly distributed household chores, or social norms, just to name a few.

Note that \(\zeta_2\), the second parameter guiding the consumption scale for couple households, is indeterminate when we have \(\sigma = 1\), i.e., log utility from consumption.

I assume an equal number of single female households and single male households (\(\mu_m = \mu_f\)) and that the total number of households sums up to one (\(\mu_c = 1 - \mu_m - \mu_f\)).

The retirement benefits for men are assumed to be 30\% of the male median income (which is normalized to 1), since the standard replacement rate is around 40\%, and in the period 1960 up until now, approximately three quarters of the old-age households received social security income.\(^\text{18}\) Women are assumed to get \(\tau_f^*\) less, and in line with the data, I set \(\tau_f^* = 23\%\).

The probability for a working-age household to retire is chosen so that the expected work life is 40 years. The probability for a retired household to be reborn as a working-age household again is chosen so that the retirement phase is 15 years in expectation.

5. Steady-state results

I now turn to the steady-state results of the model. The upper half of Table 4 shows employment, consumption and asset holdings in the data. The employment figures

\(^{17}\text{Boppart and Krusell (2016) show that a utility function where the income effect slightly dominates the substitution effect is more in line with cross-country evidence over a long time horizon. Allowing for such preferences would be an interesting but non-trivial extension.}\)

\(^{18}\text{In 1962, 69\% of the aged units (a married couple living together or a nonmarried person) received social security, in 2015 the corresponding figure was 84\% (Fast Facts and Figures about Social Security, 2017). However, taking into account this change over time does not change any results; hence, I use the approximate average.}\)
<table>
<thead>
<tr>
<th>Parameter</th>
<th>Description</th>
<th>Value</th>
<th>Source</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Preference parameters</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\beta_s$</td>
<td>Discount factor singles</td>
<td>0.949</td>
<td>See section 4.3</td>
</tr>
<tr>
<td>$\beta_c$</td>
<td>Discount factor couples</td>
<td>0.96</td>
<td>Standard value</td>
</tr>
<tr>
<td>$\sigma$</td>
<td>Relative risk aversion</td>
<td>1.0</td>
<td>Standard value</td>
</tr>
<tr>
<td>$\psi_m$</td>
<td>Disutility of work men</td>
<td>1.406</td>
<td>See text</td>
</tr>
<tr>
<td>$\psi_f$</td>
<td>Disutility of work women</td>
<td>1.513</td>
<td>See text</td>
</tr>
<tr>
<td>$\zeta_1$</td>
<td>Consumption scale for couple hh</td>
<td>3.286</td>
<td>See text</td>
</tr>
<tr>
<td>$\zeta_2$</td>
<td>Consumption scale for couple hh</td>
<td>–</td>
<td>N/A</td>
</tr>
<tr>
<td><strong>Production technology parameters</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\alpha$</td>
<td>Capital share</td>
<td>1/3</td>
<td>Standard value</td>
</tr>
<tr>
<td>$\delta$</td>
<td>Depreciation rate</td>
<td>10%</td>
<td>Standard value</td>
</tr>
<tr>
<td><strong>Social security</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\varphi$</td>
<td>Replacement rate unemployment</td>
<td>50%</td>
<td>Krueger et al. (2016)</td>
</tr>
<tr>
<td>$\kappa$</td>
<td>Replacement rate retirement</td>
<td>0.3</td>
<td>See text</td>
</tr>
<tr>
<td>$\tau_f^r$</td>
<td>Retirement “tax” on women</td>
<td>23%</td>
<td>SSA Fact Sheet</td>
</tr>
<tr>
<td>$\tau_m^r$</td>
<td>Retirement “tax” on men</td>
<td>0%</td>
<td>Normalization</td>
</tr>
<tr>
<td><strong>Other parameters</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\tilde{a}$</td>
<td>Borrowing constraint</td>
<td>0</td>
<td>Standard value</td>
</tr>
<tr>
<td>$\mu_m, \mu_f$</td>
<td>Share of hh single males (females)</td>
<td>22%</td>
<td>CPS</td>
</tr>
<tr>
<td>$\pi_r$</td>
<td>Prob. working age hh $\rightarrow$ retired</td>
<td>1/40</td>
<td>Krueger et al. (2016)</td>
</tr>
<tr>
<td>$\pi_w$</td>
<td>Prob. retired hh $\rightarrow$ working age</td>
<td>1/15</td>
<td>Krueger et al. (2016)</td>
</tr>
</tbody>
</table>

Table 3: Calibrated parameters
CHAPTER 1

refer to CPS data from 2017 for the age group 25-64. Consumption and asset data refer to the PSID wave 2008 for the whole population.\textsuperscript{19}

The second half of Table 4 shows the model results. Employment refers to the employment-to-population ratio for the working-age population in the model, to make the figures comparable. As can be seen, the employment figures line up in the right order: married men work the most, thereafter single men and single women, while married women work the least.

Consumption and asset holdings are for all households (both working-age and retired). As can be seen, couples consume more than single men both in the data and in the model. However, in the model, single women consume less than single men, which is not what we see in the data. One reason could be that the model does not include any social benefits and transfers of any kind (besides unemployment benefits), which is an income type that is more important for single women than for other groups.\textsuperscript{20}

The asset levels by group are roughly in line with the data: couples have slightly more than twice as much assets as single men, which is due to the higher permanent discount factor for couples.

There is a selection effect in terms of observed wages compared to offered wages, especially among married women. A married woman offered a high wage is more likely to work than a married woman offered a low wage and therefore, the observed average wage among working married women is higher than their average offered wage. The selection effect in the model is 10% for married women, but very small for single women.

A selection effect is also present among men, albeit not as pronounced. The effect is stronger for married men, since they can, at times of low productivity, choose to stay at home if their wife has a relatively high productivity, while this insurance mechanism does not exist for singles. The result is that married men have a 6% higher observed wage than single men, which might explain a (small) part of the marriage premium for men (typically thought of in the 10-40% range; see Korenman and Neumark (1991)).

\textsuperscript{19}Since 1999, when the PSID was redesigned, the consumption data cover over 70 percent of all consumption items available in other data sources, such as the Consumer Expenditure Survey (CEX). The PSID consumption data cover, besides food, many other non-durable and services consumption categories, including health expenditures, utilities, gasoline, car maintenance, transportation, education, and child care. For a detailed description of the consumption data in the PSID, see Blundell et al. (2016). I also follow their method of imputing rent expenditures for homeowners using the self-reported house value. Asset data refer to cash, bonds, stocks, real estate, business property, cars and vehicles, and pension funds, deducting mortgages and other debts.

\textsuperscript{20}See Low et al. (2018) for an overview of the development of welfare programs over time, in particular how the PRWORA reform replaced AFDC with TANF, and the impact on women’s labor supply in a life-cycle model setting.
Table 4: Data moments and model results. The employment figures refer to employment among the working-age population while consumption and asset holdings are for the whole population. Consumption and asset holdings are normalized to single men values to facilitate the comparison. The employment figures for couple households refer to husband and wife, respectively, while the asset and consumption figures are at the household level. The employment figures in the data refer to CPS data from 2017 for the age group 25-64. Consumption and asset data refer to the PSID wave 2008 for the whole population.

5.1. Comparing steady-state models

In this paper, I focus on a shrinking productivity gap between men and women as the driver for the increase in labor supply among married women. However, an alternative explanation often mentioned is that the disutility of work has decreased for women, for instance as a result of less discrimination and harassment in the workplace, or more permitting social norms in society.

To shed some light on the effect of these two parameters – the gender productivity gap ($\tau_f$) and the female disutility of labor ($\psi_f$) – in this section, I compare steady-state results from models with different calibrations of those two parameters. Table 5 describes four models. Model 1 refers to the baseline model, calibrated so that labor force participation by subgroup and the observed female wage gap are what we observe in the data today, as described in section 4. Model 2 refers to a model where $\psi_f$ is the same as in the baseline model, but the observed female wage gap is what we observed at the beginning of the 1960s.

Model 3 is an alternative specification in which the productivity gap, $\tau_f$, is the same as in the baseline model, but $\psi_f$, the disutility of labor for females, is adjusted upwards by 20%. Model 4, finally, is a model in which I change both $\tau_f$ and $\psi_f$ as compared to the baseline model.

Figure 10 shows the resulting employment rates. There are a couple of insights to be had. First of all, in model 2 a higher productivity wedge, $\tau_f$, makes the married
women work substantially less. What might at a first glance be more surprising is the small (and even positive) effect on single women’s employment rate. However, for single women, a change in the productivity wedge is similar to a change in the overall wage level, and such a change has a small impact on labor supply with balanced growth preferences. The income effect and the substitution effect cancel, and therefore the net effect on single women is small (and in this case even positive). Hence, the model delivers a theory of why married and single women display different responses to changes in wages in the long run.

A second observation is that the increase in married men’s employment is not nearly as large as the decrease in married women’s employment. There is very little “crowding-out” within the couple, in line with what we see in the data.

A third observation is the negligible effect on single men, which is not surprising. The only way in which this higher productivity wedge affects single men is via the general equilibrium channel with changes in the interest rate, the wage and the lump-sum tax.

In model 3, an increase in women’s disutility of labor by 20% leads to decreasing employment figures for both married women and single women, while again single men are, of course, unaffected. The upward adjustment of married men’s labor supply to compensate for their wives’ lower labor supply is negligible.

In model 4, finally, both the productivity gap and the disutility of labor are increased as compared to the benchmark. In this model, single women work slightly less due to the increased disutility, while married women work substantially less, due to both the increased disutility and the increased productivity gap. Once more, married men work slightly more to compensate for their wives’ lower earnings, and single men are hardly affected. Model 4 is an example of a parametrization of the model for which the employment rates by subgroup in 1962 are hit very closely.

The comparison between these steady-state models shows that the increase in labor supply among married women is unlikely to be mainly driven by a general decrease in the disutility of labor for women. If this had been the case, single women would also have been strongly affected. A change in the gender wage gap (in line with what we have seen in the data), on the other hand, affects the labor supply of married women, but not that of single women. The effect on married women is approximately two thirds of the increase we have seen historically. The remaining third could be modelled as an additional change in the disutility of work, which would affect both single and married women.

\[21\text{ A related observation is that the selection effect is stronger in this model: for married women, the average wage among working individuals is 16\% higher than the offered wage.}\]
CHAPTER 1

<table>
<thead>
<tr>
<th></th>
<th>Model 1 baseline</th>
<th>Model 2 $\tau_f \uparrow$</th>
<th>Model 3 $\psi_f \uparrow$</th>
<th>Model 4 $\tau_f \uparrow, \psi_f \uparrow$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\tau_f$</td>
<td>20%</td>
<td>45%</td>
<td>20%</td>
<td>45%</td>
</tr>
<tr>
<td>$\psi_f$</td>
<td>1.513</td>
<td>1.513</td>
<td>1.816</td>
<td>1.816</td>
</tr>
<tr>
<td>Observed wage gap (mean)</td>
<td>21%</td>
<td>43%</td>
<td>19%</td>
<td>42%</td>
</tr>
<tr>
<td>Observed wage gap (median)</td>
<td>20%</td>
<td>45%</td>
<td>20%</td>
<td>45%</td>
</tr>
</tbody>
</table>

**Table 5:** Comparison of four models with different calibrations of the female disutility of labor ($\psi_f$) and the productivity wedge ($\tau_f$) and the resulting observed wage gap.

**Figure 10:** Steady-state comparison. Employment rates by subgroup from four steady-state models with different calibrations as described in Table 5.
6. Transition from low to higher female labor force participation

In this section, I describe the economy during a deterministic transition. The variable that changes over time is $\tau_f$, the female productivity wedge. Hence, I choose to only include the effects directly driven by the shrinking female productivity wedge. Clearly, the approach does not rule out the possibility that other changes (such as, for example, access to contraception, changes in divorce laws, cheaper child-care, and/or changes in social norms) have played an important part. However, the decreasing wage gap is more easily observed in aggregate data, and I deem it an important first step to focus on its effect in isolation.

The decrease in $\tau_f$ may be interpreted in several ways. One is that the decrease of $\tau_f$ is driven by a female-biased technical change (as suggested by, e.g., Galor and Weil (1996) or Ngai and Petrongolo (2017)). Another is that it could be due to a decrease in discrimination. The discrimination against women, leading to the misallocation of talent and lost productivity, was probably worse in the 1960s than now, and one can easily come up with several examples. For instance, as told by Jones (2016), Sandra Day O’Connor, who later became the first female Supreme Court Justice, graduated third in her class from Stanford Law School in 1952. However, the only private sector job she could get upon graduation was as a legal secretary. It is easy to imagine that the same type of discrimination and misallocation of talent takes place also at lower levels, but that this discrimination has decreased over the years.

6.1. Modelling the transition

We now turn to the transition path. The transition is assumed to happen with perfect foresight: in 1960, people learn that $\tau_f$, the female productivity wedge, will go down over time. Hence, agents are surprised only once but thereafter enjoy perfect foresight about the future.\footnote{An alternative would be to model the agents as myopic: at each date, they believe that the current wage structure will prevail forever, being repeatedly surprised every year, as e.g. Heathcote et al. (2010) do as a robustness check.}

The end of the transition, the final steady-state, is the baseline model calibrated as described in section 4, while model 2 is used as the starting point of the transition. In other words, the only difference between the start and the end of the transition is that in 1960, the female productivity wedge was larger. Otherwise, the two economies are exactly the same. It is obviously a stretch to assume that the economy was in a steady state in 1960, and we will keep that in mind when interpreting the results later.

$\tau_{1960}^f$ is calibrated so that the observed median gender wage gap is 45% and the observed average female wage gap is 43% in the initial steady-state. I model the shrinking productivity wedge as a smooth decline from 45% to 20% over a period of 50 years, as Figure 11 illustrates. However, in reality the wage gap has had an
uneven trajectory, with a strong decrease historically up until the 1930s, then being almost stable around 40% between 1950 and the end of the 1970s, and thereafter continuing to decrease (see Goldin (1990) for an overview of the history of the gender wage gap and its underlying forces). Instead of trying to match the observed average wage gap year by year, I choose a crude straight line approximation to facilitate the interpretation of the results. In a model with discretized productivity states, the observed median wage gap is the same as the offered median wage gap (despite differences in the underlying AR(1) parameters), since even with 15 productivity states, 15% of the individuals are in the middle bin. Hence, the selection effect would have to be extremely strong to overturn this.

6.2. Transition results

The resulting employment paths for the different subgroups are shown in Figure 12. A first thing to note is that the increase in married women’s employment is 19.5 percentage points, roughly two thirds of the total change between the years 1960 and 2000.

Second, married men decrease their labor supply by about 5.5 percentage points, but not nearly as much as the increase among married women, which corresponds well to what we observe in the data.

Third, except for the immediate jump at the beginning of the transition (which we will discuss later), single women hardly change their labor supply due to the change in the female productivity wedge. Changes in the female productivity wedge are for them similar to changes in the overall level of wages, and these
changes have a small impact on labor supply if the preferences are of the balanced growth type. The income effect and the substitution effects roughly cancel and the labor supply is approximately the same. This is also in line with what we observe in the data.

Lastly, single men are only affected by general equilibrium effects. Therefore, their labor supply is, not surprisingly, hardly affected except for the strong jump in the first half of the 1960s.

Thus, changes in the female productivity wedge, leading to changes in the gender wage gap corresponding to what we have seen in the data, have a strong effect on married women, and explain roughly two thirds of the total change in employment we have seen for this group during the last 60 years. The effect on single women is very small, on the other hand. Nevertheless, in the short run, single women react a great deal more to a sudden wage change, because their intertemporal elasticity is higher.

Naturally, there are other things that have been going on at the same time in the economy that affect the question at hand. One thing is a potential decrease of the female disutility of work, $\psi_f$. It is easy to imagine that the disutility of work for females has gone down since the 1960s, due to workplace related factors (e.g. less discrimination and harassment in the workplace), factors more generally related to the cost of working (e.g., child-care options and social norms), and couple-specific factors (e.g. more evenly distributed household chores). Some combination of those factors explains the remaining third of the labor supply increase for women that is not captured by the model in this paper, and the fact that single women have increased their labor supply somewhat more than predicted by this model. However, for the purpose of this paper, I choose to focus on the closing of the wage gap and its effects, and note that this gives a lower bound for

Figure 12: Transition path for employment by subgroup.
the overall effects of the increase in the labor force participation among (married) women.\textsuperscript{23}

6.3. Short-run labor supply responses by subgroup

There is a strong immediate response by single women (and single men) at the beginning of the transition. Compared to the steady-state labor supply, the single women’s labor supply immediately drops at the beginning of the transition and climbs back later. When they learn about their higher future wage, they prefer to intertemporally shift some of their labor supply from today to tomorrow. This effect exists but is much smaller for married women.

The difference in the immediate response is a direct result of how the mass of individuals in the model is distributed in relation to the working decision. Figure 13 is helpful to understand the underlying mechanisms. It shows the policy functions for a single woman (not hit by an unemployment shock) in a model with only five productivity states. For a given productivity level, a woman will work if she has little wealth, and choose not to work if she is rich. At the asset level for which the woman decides to stop working, there is also a kink in the savings policy function, and that is where the savings policy function crosses the 45 degree line. In other words, for asset levels below the point where the woman stops working, she is saving up assets, and for the asset levels above the point where the woman stops working, she is decumulating assets. Hence, with a persistent productivity process, there will be bunching of individuals around this exact point. This is seen in Figure 14, which shows the resulting distribution of mass among single women. The thick red line indicates the “working frontier”. In the region north-west of the line, i.e., in the region where individuals have low assets and/or high productivity, the decision is to work. In the region south-east of the line, i.e., in the region with high assets and/or low productivity, the decision is to not work. As can be seen from the illustration, there is a clear excess mass along the “working frontier”. Individuals move randomly between productivity levels and hence, the mass is still spread out, but with a persistence of the productivity process such as the one used in this model, the bunching behavior can clearly be seen.

To facilitate the intuition, Figure 15 shows three stylized examples. 15a shows the case of two single women, with the same median productivity, but one starting with very high asset holdings, the other starting with very low asset holdings. The

\textsuperscript{23}The increase in single households is also a trend that could be taken into account when analyzing changing household composition and the effect on an economy’s response to aggregate shocks. In 1962, one fifth of all women between 25 and 64 were single. By 2017, the corresponding figure had increased to two fifths. However, it is not entirely clear how this increase should be captured in a model with infinitely lived households. One could imagine that there is an influx of new households to the closed economy, that all those new households are singles, and that the distribution of immigrating single households corresponds to the ergodic distribution of existing households in terms of assets and productivity. That would correspond to changing the weights $\mu_m$ and $\mu_f$ over time. None of the main insights change, and the results from such an exercise are available upon request.
(a) Consumption  
(b) Work  
(c) Savings

**Figure 13:** Illustrative policy functions for a single woman (not unemployed) in a model with only five productivity states. The x-axis denotes beginning-of-period assets. Pink color refers to the highest productivity state, while red is the lowest.

**Figure 14:** Distribution of mass for single women over productivity and assets (blue dots) and the working frontier (red line). North-west is working region, south-east is where the women do not work. Mass not drawn to scale, the differences are exaggerated. The x-axis does not show the full state space, but covers where > 99% of the mass are.
woman with a lot of assets is rich enough to afford leisure. She has more than her target wealth (hence impatience dominates precautionary motives), so she consumes from her assets and slowly moves down in the asset space. The woman who starts with very little assets – less than her target wealth – works and accumulates assets (precautionary motives dominates patience). They will both move towards the same target wealth, which coincides with the work frontier: the point where the work decision switches from work to leisure.

Next, consider 15b. I assume that the woman has the same median productivity, but now she is married to a very low-productive husband. If the couple starts out with a lot of assets, it can afford that even the wife enjoys leisure (it can still consume enough given its continuation value since it can consume out of its wealth). However, at some point, but before the couple reaches its target wealth level, the wife starts working (in order to maintain the consumption level). The couple’s continuation value does not only depend on the wife’s productivity, but also on that of the husband, and it is likely that the husband will become more productive at some point in the future and therefore, the impatience motive dominates a bit longer and the couple continues to consume from its assets until it reaches its target wealth. If the couple starts out with very low assets, the wife simply works and the couple saves until it reaches its target wealth. Hence, in this case, the working frontier does not coincide with the target wealth level.

Last, consider 15c. I still assume that the woman has the same median productivity, but now she is married to a reasonably productive husband, productive enough so that he will always work. The couple starting out with very high assets is rich enough to afford leisure for the woman, and the couple consumes a great deal so that it slowly decumulates its assets until it reaches its target wealth level. The couple starting out with very low asset holdings first wants to accumulate assets quickly, and both the husband and the wife will work. At some point, the couple is rich enough to at least let the wife enjoy leisure, but it will still accumulate assets until it reaches its target wealth. Again, the working frontier does not coincide with the target wealth level.

Hence, single households bunch around their working frontier to a larger extent than individuals in couples. For another illustration of the same behavior, first consider Figure 16, which shows the asset level for which the single woman is rich enough to stop working, conditioning on the productivity level, indicated by dots. The grey line shows the points where the savings policy functions cross the 45 degree line. As can be seen, the two decisions coincide.

Now we can compare this to Figure 17a, which shows the same type of graph but for a married woman, assuming that her husband is in the lowest productivity state. As can be seen, for a married woman, the working frontier does not coincide with the wealth level where the switch from accumulation/decumulation of assets happens in the same way. Therefore, there is not the same type of bunching around the working decision and, consequently, the labor response to a temporary shock is smaller.
Figure 15: Stylized illustrations of work decision and asset accumulation. See the text for the description.
Figure 16: Single woman: Illustration of work and savings behavior in the asset/productivity space. The gray line indicates where the savings policy function crosses the 45% line, dots indicate the asset level where the person is rich enough to choose not to work.

Figure 17b shows the case of a married women when the husband has a relatively high productivity (state 11 out of 15). For the states in which the woman has low productivity, she does not work while the husband does. However, higher productivity, even though leading to no change in the working decision, leads to a lower asset threshold for accumulating vs decreasing assets. The reason is the change in continuation value: if the woman has a higher productivity, there is an increasing probability that she will get an even higher productivity in the future and in that future state start working. Therefore, the continuation value increases with higher productivity, and the couple can “eat from its savings”, despite the actual working decision being constant.\textsuperscript{24}

Given the work-save behavior described in this section, the model has predictions for the asset accumulation for different types of households. In the appendix section A.7, I show that the model predictions for asset holdings relative to the observed productivity for single and couple women fit the observed data.

7. Aggregate shocks during the transition path

It is now time to turn to the question about the economy’s response to aggregate shocks during the transition path. I study the economy’s response to “MIT shocks” and regard the impulse response paths as numerical derivatives of the economy’s

\textsuperscript{24}The stylized illustrations in Figure 15 correspond graphically to the case of female productivity of 10 in the two figures in Figure 17.
(a) Married woman, conditioning on having a husband in the lowest productivity state.

(b) Married woman, conditioning on having a husband with productivity state 11 (out of 15).

Figure 17: Illustration of work and savings behavior in the asset/productivity space. The gray line indicates where the savings policy function crosses the 45% line, the dots indicate the asset level where the person is rich enough to choose not to work.
aggregate response. For the TFP shock, I use standard parameters as in Boppart et al. (2018), translated into annual values, i.e., I assume a standard deviation of $\sigma_z = 0.0260$. After impact, a shock decays geometrically at a rate of $1 - \rho_z$, with $\rho_z = 0.815$.25

As an illustrative example, I first shock the model in the year 1970 with a very large shock (three standard deviation) to get results that are visually clear. Figure 18 shows what happens in the model: during the deterministic transition path, the economy is shocked by an increase in TFP, and then transitions back to its underlying transition path.26 As can be seen, the employment figures show that single men and single women react more strongly than married women. Married men react the least. The reason is what was described in the previous section: the mass of single households are concentrated closer to the working margin than the mass of married individuals. Therefore, more single households are affected when the working frontier is moved. I will now analyze the responses more formally.

25Boppart et al. (2018) use a quarterly serial correlation of 0.95, hence the annual value 0.815. A quarterly standard deviation of 0.007 gives an annual value of $0.007 \cdot \frac{1 - 0.95^4}{1 - 0.95} = 0.0260$.

26For a discussion about the appropriateness of MIT shocks during a transition path, section A.6 in the appendix gives a comparison between a model with “true” uncertainty and BKM shocks during a transition path in a representative agent setting. The conclusion is that the two models give the same results, and that the underlying transition matters in that particular example for the economy’s response to aggregate shocks.
7.1. Impact of an aggregate shock

Figure 19 shows the TFP shock and the impulse response function for output at four different points in time during the transition path. The impulse response function for output is measured as \( \text{the difference compared to the underlying deterministic trend} \). As can be seen in the figure, the response does not change in any meaningful way between years. However, what on the aggregate might appear as no changes over time is the sum of multiple counteracting forces.

I now turn to the detailed employment results for different subgroups. The first line of Table 6 shows the immediate impact of an aggregate shock to employment by subgroup in the steady-state model (calibrated to “today”, i.e., the end of the transition as described above). The first observation is how the size of the employment effect in steady state differs between groups. If we first focus on married couples, we see that married men are least impacted (their employment rate goes up by 0.41% as a consequence of a TFP shock of one standard deviation), while married women react a great deal more strongly (1.43% impact). The substantially stronger response among married women than among their husbands is in line with labor supply elasticity estimates from a number of studies (see Blundell and MaCurdy (1999) for an early overview).

In line with the data, singles, and in particular single men, exhibit the highest employment volatility. From a modelling perspective, this is not surprising, given the discussion in section 6.3 about labor responses by subgroup. To reconcile the fact that married women and single women react more than married men, remember that we do not take into account any sector specificities in this
Turning to changes over time, Table 6 shows that married men as a group have increased their response over time. In 1970, the effect of an aggregate shock to their employment level was only half of what is estimated for steady-state “today”. However, the absolute changes for married men are small as compared to the changes estimated for married women, whose response has decreased almost as much in relative terms, and more in absolute terms. In 1970, their response was 32% stronger than today. This finding is in line with the estimates by Blau and Kahn (2007) and Heim (2007), who both estimate a falling wage elasticity for women over the time period 1980 to 2000.

For single men, the impact of an aggregate shock was stronger historically, and for single women the opposite is true.

To summarize, the fact that the impulse-response function of output, compared to the underlying trend, is remarkably stable between years is the sum of counteracting forces. The aggregate employment effect is a weighted sum of the effects by subgroup. The effect on employment among married women has decreased over time, while their share of the labor force has increased. In sum, their contribution to the aggregate employment response is actually slightly smaller now than in 1970. The effect on employment among married men, on the other hand, has gone up, while their share of the work force has not changed that much. Hence, this group is now contributing more to the aggregate employment response. The responsiveness of single men and single women has changed over time in opposite directions, and these changes are off-setting. The net effect is therefore close to zero.

Figure 20 compares the aggregate employment effect in 1970 to the aggregate employment effect in the end steady state, and decomposes the contribution from differences in employment shares and differences in the response. The first bar is the actual employment response in 1970, and the contribution by subgroup. The second bar shows a counterfactual in which the fraction working by subgroup was held at the 1970 level, but the employment responses by subgroup are the

### Table 6: Immediate impact on employment from a TFP shock of one standard deviation. Impact measured as the deviation from the deterministic underlying trend.

<table>
<thead>
<tr>
<th></th>
<th>Married men</th>
<th>Married women</th>
<th>Single men</th>
<th>Single women</th>
<th>Total employment</th>
</tr>
</thead>
<tbody>
<tr>
<td>End steady-state</td>
<td>0.4%</td>
<td>1.4%</td>
<td>4.2%</td>
<td>3.4%</td>
<td>1.7%</td>
</tr>
<tr>
<td><strong>Response during the transition:</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1970s</td>
<td>0.2%</td>
<td>1.8%</td>
<td>4.6%</td>
<td>2.5%</td>
<td>1.6%</td>
</tr>
<tr>
<td>1990s</td>
<td>0.3%</td>
<td>1.6%</td>
<td>5.4%</td>
<td>2.8%</td>
<td>1.7%</td>
</tr>
<tr>
<td>2010s</td>
<td>0.4%</td>
<td>1.4%</td>
<td>4.2%</td>
<td>3.4%</td>
<td>1.7%</td>
</tr>
</tbody>
</table>
end steady state responses. As can be seen, in such a counterfactual experiment, married men contribute more to the aggregate response relative to the 1970s, while married women contribute less, as can directly be concluded from Table 6. The third bar in the figure shows a counterfactual measuring the importance of the changing composition in the workforce, assuming that the employment response by subgroup is kept at the 1970 level, while the employment fractions are those from the end-steady-state. As expected, the largest difference is for married women, which is the group that has substantially increased its employment rate.

7.2. Jobless recoveries

In the previous section, I analyzed the economy’s response to a TFP shock by looking at employment and output compared to the underlying deterministic trend. Now I turn to actual employment levels. If the economy is hit by a negative TFP shock, how does employment react, and how does the reaction depend on where the economy is in the transition path? To see if the model captures the historical facts of employment effects after a recession, I make the following experiment: The economy is shocked with a negative TFP shock of two standard deviations during two points in time. Those two points are chosen to be 1970, to capture the economy’s response in the middle of the transition, and 2010, to capture the response when the economy has almost finalized the transition. Figure 21 shows the resulting employment figures, normalized to one in the year before the shock hits.
As can be seen from Figure 21a, when the economy is hit by a negative TFP shock in 1970, total employment falls. Single men are most affected immediately on impact. Four years later, the economy has returned to its initial total employment level (even slightly above). The subgroup driving this recovery is married women, and to some extent single women. After four years, neither single men nor married men have returned to their pre-recession employment level.

Figure 21b shows the same type of TFP shock, but in the year 2010, i.e., when the employment figure for married women has stabilized at its current level. Now, four years after the shock, the employment rates have still not returned to their pre-shock level. The main difference is that married women are no longer driving up the employment figures. Again, as in the data, single men are the group most severely hit by the negative shock in terms of employment.

These results are not surprising, given the previous discussion about labor supply responses compared to the underlying trend. Even though the model output and the data cannot be directly compared for any given recession (e.g., since the model does not have any sectors), it is reassuring that the model captures the most salient features of the phenomenon sometimes referred to as jobless recoveries.

8. Conclusions

Women’s increased involvement in the economy has been the most significant change in the labor markets during the past century. In this paper, I study how this increase has affected the economy’s response to aggregate shocks. I explicitly model households as being single men, single women, or couple households. The model is able to capture the salient features of historical data in terms of employment by subgroup over time. Incorporating both one- and two-person households is also
shown to matter for the employment dynamics in response to shocks, with single households reacting more strongly.

In this paper, I have modelled the couple household as a unitary household, despite the potential shortcomings of that assumption. A richer model of the intra-household decision process would be an interesting extension; however, allowing for strategic interaction between individuals within a household is not trivial, especially when the problem is dynamic. In that sense, the model in this paper serves as a useful benchmark to start thinking about a more careful modelling of the household that could include cooperative or non-cooperative processes with endogenous bargaining weights.

Except for household heterogeneity, which does not only include the standard dimensions of wealth and productivity but also gender and household composition, the proposed model is a straightforward model of the real business cycle type. Features that could be important for the analysis of aggregate employment dynamics that are not included are nominal price and wage stickiness and frictions on the labor market. Including these features, keeping the household heterogeneity, would enrich the framework. However, explicitly modelling the household is a necessary first step.

References


A. Appendix

A.1. More about sectors

Figure 22 shows the fraction of individuals in the labor force working in the manufacturing sector, while Figure 23 shows the responsiveness of employment to fluctuations in GDP by sector and subgroup.

**Figure 22:** The fraction working in manufacturing (age 25-64). Source: CPS. The graph includes everyone with a defined sector, i.e., also unemployed with sector definition.

**Figure 23:** Responsiveness by sector and subgroup (age 25-64). Result from regressing the cyclical component of hours by type in sector on the GDP cyclical component. Source: CPS.
A.2. International comparison

Figure 24: Ratio of the female to male labor force participation rate for all countries for which data is available from OECD. Figures are normalized to the 2016 ratio. Population aged 25-54. Source: OECD.

Figure 25: Labor force participation by gender, population aged 25-54. Source: OECD.
A.3. Asset holding couple and single households

Figure 26: Asset holdings by household type. Source: PSID 2008, referring to the previous year. All values measured in 2008 USD. House value is the value of first residence minus mortgages.

Figure 27: Asset holdings over the life cycle. Source: PSID 2008, referring to the previous year. All values measured in 2008 USD. Robust standard errors.
A.4. Unemployment for a more narrowly defined age group

As can be seen in Figure 28, which shows the unemployment rates by subgroup for individuals 35-65, the pattern is extremely similar to the pattern for the population 25-64. Hence, the high unemployment rate among single men is not purely driven by them being younger.

Figure 28: Unemployment by subgroups, population 35-64. Source: CPS.
A.5. Different filter options

As a robustness check, I rerun equation (1) using four different filtering techniques, Figure 29 shows the resulting coefficient of interest for the four different subgroups. As can be seen, all conclusions remain regardless of filter choice.

The four filters used are the following:

**Hodrick-Prescott**: High-pass filter, using 6.25 as the smoothing parameter.

**Baxter-King**: Bandpass filter, filtering out stochastic cycles at periods smaller than 2 years and larger than 8 years.

**Christiano-Fitzgerald**: Bandpass filter (based on the generally false assumption that data are generated by a random walk, but shown to still be nearly optimal), filtering out stochastic cycles at periods smaller than 2 years and larger than 8 years.

**Butterworth**: Rational square-wave filter, filtering out stochastic cycles at periods larger than 8 years, and using order 2.

![Figure 29](image_url)

**Figure 29**: The resulting coefficient from regressing the cyclical component of hours per capita on the cyclical component of gdp from four different filters.
A.6. Aggregate shocks during a transition path in an RA model

A natural question to ask is if the MIT/BKM method reproduces the results from a model with true uncertainty, if the shocks take place during a transition (as is the case in this paper). To investigate this, I make a simple example in a representative agent setting with savings and endogenous labor. The utility function is supposed to be of the McCurdy type. I do the following three steps:

**Step 1:** True uncertainty model
- Set up an RA model with $z$ (TFP) as a state variable, solve non-linearly
- Solve for a transition path (letting $\theta$, the Frisch elasticity, go from 2 to 1 during 40 years, and thereafter remain)
- Simulate the model with a shock sequence for $z$

**Step 2:** BKM method
- Set up an RA model w/o uncertainty (otherwise the same as above)
- For each year during the transition: solve for IRFs from a shock
- Simulate the model with the same shock sequence as above

**Step 3:** Compare the results
It turns out that the BKM/MIT method works very well and close to perfectly replicates the results from the model with “true” uncertainty. In this particular example, the IRFs vary along the transition. The effect on the impact on labor supply is approximately 40\% stronger if the shock hits exactly when the transition starts, than if it happens once the economy has reached its new steady state.
CHAPTER 1

Figure 30: Compare results

Figure 31: Comparing IRFs over the transition
(a) Using correct IRFs by year  (b) Using final steady-state IRFs

**Figure 32:** Difference if one uses final steady-state IRFs for the whole transition. Note the scale on the x-axis (otherwise one cannot see any difference). If the final steady-state IRFs are used for the transition path, the impact of the shock early during the transition path is underestimated.

(a) Using correct IRFs by year  (b) Using beginning-of-trans IRFs

**Figure 33:** Difference if one uses beginning-of-transition IRFs for the whole transition. If the beginning-of-transition IRFs are used for the whole time period, the impact of shocks is overestimated later.
A.7. Asset holdings for single vs married women, model vs data

Figure 34 shows model results for asset holdings against the productivity level for single and married women, respectively, conditional on working. The reason for selecting only working women is that this is what we can compare with the data: for non-workers, we do not observe productivity. Looking at single women, it is clear that we observe women of all productivity levels working, but there is a strong connection between wealth, productivity, and the work decision: single women with high wealth and low productivity do not work.

Figure 34b shows the corresponding graph for married women from the model (unconditional on the spouse’s productivity). As can be seen, married women of all productivity levels are in the workforce. However, for a particular (low) productivity level, a married woman can still be in the workforce, despite having a higher household asset level (compare, e.g., a single woman with log(productivity) of -0.5 to a married woman with this same productivity level).

Hence, we have three predictions from the model. First, we should not see single women with high assets and low productivity working. Second, there should be relatively more single women close to the borrowing constraint. Third, there should be a higher correlation between the single woman’s productivity and her asset level (conditional on not being on the borrowing constraint).

Figure 35 shows the corresponding figures from the data. I use PSID data from the years between 1998 and 2008, and pool all observations. As can be seen, there are hardly any single women with high assets and low productivity working, compared to the same asset level for married women. I interpret “being close to the borrowing limit” as having negative assets. 23% of the single working women have negative assets. The corresponding figure for married women is 10%.

Formally, to verify if the correlation between (positive) asset holdings and observed productivity is higher for single women than for married women, I run the following regression:

\[
\log(a_{ij}) = \alpha_j + \beta_j \log(\omega_i) + \gamma X_i + \epsilon_{ij} \tag{20}
\]

where \(i\) indicates the individual observation and \(j\) is the civil status: single or couple \((j \in \{s, c\})\). \(a\) is the asset holdings,\(^{27}\) \(\omega\) is the observed hourly wage, and \(X_i\) is a set of control variables: age, age squared, and education level.\(^{28}\) \(\beta_j\) informs us about the correlation between the observed wage and the asset level for singles and married, respectively. Table 7 shows the results. As can be seen, when controlling for the age profile and education, the correlation between asset holdings and productivity becomes lower. However, the correlation is higher for single women, and the difference between the coefficients for singles and married is statistically significant for both specifications.

\(^{27}\) Assets include cash, bonds, stocks, real estate, cars, and is net of mortgages and other debt.

\(^{28}\) I divide the education level into three distinct groups: below 12 years, exactly 12 years, or above 12 years of schooling.
Table 7: Source: PSID 1998-2008, pooling all years. The regression only includes women.

<table>
<thead>
<tr>
<th>Assets</th>
<th>(1)</th>
<th>(2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>log(ω) × singles</td>
<td>1.321</td>
<td>0.971</td>
</tr>
<tr>
<td></td>
<td>(0.038)</td>
<td>(0.037)</td>
</tr>
<tr>
<td>log(ω) × couple</td>
<td>1.015</td>
<td>0.686</td>
</tr>
<tr>
<td></td>
<td>(0.014)</td>
<td>(0.014)</td>
</tr>
<tr>
<td>Controls</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>N</td>
<td>34,276</td>
<td>34,031</td>
</tr>
<tr>
<td>βs = βc (p-value)</td>
<td>0.00</td>
<td>0.00</td>
</tr>
</tbody>
</table>

(a) Single women  
(b) Married women

Figure 34: Model results for asset holdings vs productivity level.
Figure 35: Asset holdings vs productivity level for single vs married women. Source: PSID 1998-2008, pooling all years. Assets include cash, bonds, stocks, real estate, cars, and are net of mortgages and other debt. I leave out asset holdings larger than USD 7 million (0.2% of the observations) and asset holdings smaller than USD -300.000 (0.1% of the observations).
Labor supply in a quantitative heterogeneous-agent model

Timo Boppart, Per Krusell, and Jonna Olsson

1. Introduction

Since long, the labor-supply channel has played a central role in macroeconomic analysis and, in particular, in our core theories of how the business cycle works. Nevertheless, the microeconomics behind households’ decisions on whether and how much to work has typically been modeled in a stark and rather unrealistic way. It has almost exclusively focused on representative-agent behavior, thus taking for granted that the behavior of the population of all households can be summarized accurately by that of one. Moreover, this one agent’s choice environment has abstracted from some of the arguably most central features of households’ economic life: large and uncertain fluctuations in their economic labor-market outcomes. The aim of the present research is to examine frameworks that are significantly richer in these respects and that allow us to assess whether or not the predictions yielded by the starker frameworks are robust to these extensions.

To this end, we examine a class of macroeconomic models based on household labor supply in a context where households are heterogeneous in wealth and wages, face earnings uncertainty, and suffer from the fact that they cannot fully insure against this uncertainty. Our framework still abstracts from features such as age and gender heterogeneity, which potentially also are of great importance, but we argue that it nevertheless offers a microeconomic structure that rather adequately captures the essence of households’ most important economic decision making. As a consequence, we think it ought to be useful for studying a host of macroeconomic phenomena, involving how total hours worked and employment are determined in the long run as well as how they vary over the business cycle. The models also ought to be useful for understanding how labor supply varies across households and, more broadly, for understanding inequality both as far as its determinants and its effects.

The specific issues we speak to include, first of all, aggregation. We consider heterogeneity in wealth and individual productivity (wages). As far as aggregation in wealth, with a very small number of (notable) exceptions, standard macroeconomic frameworks use frameworks that do not aggregate in wealth. Aggregation obtains, under complete markets, only for very special utility functions, and such utility functions are typically not used in
the literature.\footnote{The most commonly used utility function that admits aggregation in wealth is the GHH form, due to Greenwood et al. (1988), but it does not admit wealth effects, which we think are essential in quantitative work. The Cobb-Douglas formulation also admits aggregation, but it does not have a sufficient number of parameters to allow us to separately target both how much households work and how they respond to wage changes.} In particular, by far the most commonly used utility function over consumption and hours in the macroeconomic literature is the MaCurdy function (MaCurdy 1981), and it does not admit aggregation. Moreover, under incomplete markets, an increasingly commonly used assumption and one we use here, aggregation does not obtain even for the utility functions that do aggregate under complete markets. Turning to heterogeneity in productivity, aggregation also does not obtain in our standard models. Given that both wealth and individual wages differ very markedly across households in the data, we study how large the resulting aggregation bias is. We use the obvious approach here: we compare a framework with no heterogeneity to one with heterogeneity that is similar to the one observed.

A second issue that is in focus here is uncertainty and incomplete insurance against labor-market risk. Here we simply follow the tradition established by Huggett (1993) and Aiyagari (1994); in particular, we add labor choice to the Aiyagari setting. This is by no means the first paper with this feature, but the systematic comparison we offer between complete and incomplete markets is new.

Third, in a framework with heterogeneity in productivity across households, an important question is the normative one of who “should” work, and how much. Our models are regularly used to make normative statements, and we therefore discuss the who-should-work question in the present setting using the most commonly used social welfare function and compare it to alternatives.

Fourth, an important part of the analysis herein goes beyond the typical focus on exactly how many hours each household will work and focuses instead on the extensive-margin labor choice: will they work at all? If we look at people’s working hours, many people actually work zero hours. In the U.S., the civilian labor force participation in the beginning of 2019 was 63%. Even in the age group 25–54, i.e., among prime-aged individuals, 17% of the population is out of the labor force. For these reasons, a realistic description of labor supply should, in our view, take the extensive margin into account. Our analysis is based on two kinds of frameworks, one with an intensive-margin choice only and the other with an extensive-margin choice only, but we also develop a new setting which combines the two.

Fifth, we systematically examine Frisch elasticities in our heterogeneous-agent model, both on the individual and the aggregate level. On the individual level, we look at departures from the MaCurdy function, which imposes a constant Frisch elasticity, and we consider the extensive margin, implying an aggregate elasticity that depends on the joint distribution of wealth and wages.

Sixth, we take strong income effects to be an important feature of household behavior. Aggregate data, both over time and across countries, supports this position and, in fact, is very hard to reconcile with preference formulations without strong income effects. In particular, Bick, Fuchs-Schundeln, and Lagakos (2018) and Boppart and Krusell (2016)
suggest that the income effect outweighs the substitution effect: if both wages and independent income (such as capital income) are scaled up by the same amount, hours worked falls. The latter paper suggests a new preference class—we refer to it as BK preferences—that extends the most commonly used preference class, the KPR class, due to King, Plosser, and Rebelo (1988), in a way that still is consistent with balanced growth. Throughout the paper, we examine BK preferences and compare their features to the KPR case.\textsuperscript{2}

In sum, our study is both backward- and forward-looking: we critically re-examine models that are commonly used, from the perspective of the rich heterogeneity in the data we have on households, but we also examine new models with features that we hope will improve our understanding of how macroeconomic outcomes come about as aggregate expressions of a complex microeconomic reality.

A very brief summary of our findings is: (i) the aggregation bias in total hours worked is with less than 6\% quantitatively quite limited for the intensive-margin models, but it can be large, e.g., larger than 15\% for extensive-margin models; (ii) many quantitative features of our incomplete-markets economies are shared by the corresponding complete-markets economies; (iii) the incomplete-markets model leads to a rather large departure from the utilitarian, frictionless optimum, in which agents should work more, the higher is their productivity, and this departure is significantly larger for BK than for KPR preferences; and (iv) the framework with incomplete markets and an active labor-supply channel leads to lower real interest rates and higher wealth inequality, thus helping us explain the data relative to models without an active labor supply; the BK framework is somewhat more powerful in these regards than the KPR setting.

We are obviously not the first to study labor supply in heterogeneous-agent macroeconomic frameworks, but there are still relatively few frameworks of this sort. To our knowledge, the first appearance of a model of this sort is Krusell and Smith (1998), and a set of closely related frameworks have been proposed since then, for instance Pijoan-Mas (2006), Marcet, Obiols-Homs, and Weil (2007), Chang and Kim (2006, 2007), and Chang, Kim, Kwon, and Rogerson (2018). Some of these papers are explicit about modeling labor-market frictions, or at least interpret the process for wages or productivity from a frictional perspective; see Krusell, Mukoyama, Rogerson, and Şahin (2008) and Krusell, Mukoyama, Rogerson, and Şahin (2017). At the same time, and despite the many insights in the pre-existing literature, there has not been any systematic study of the list of issues we study here.

More broadly speaking, labor supply has a long tradition in macroeconomics, especially as regards its role in the transmission of business cycles. Lucas and Rapping (1969) suggested that employment fluctuations could be understood as (individually optimal) household responses in their desired work hours to fluctuations in wages. Kydland and Prescott (1982) then created the complete real business cycle theory, where labor supply played the integral part Lucas and Rapping had pointed to. Hansen (1985) and Rogerson (1988) emphasized the extensive margin of labor supply as enabling more powerful labor-supply

\footnote{On the microeconomic level, one could perhaps hope to have causal evidence on this point but it is hard to find fully convincing studies. One recent paper, Giupponi (2018), finds support of strong income effects.}
responses to shocks. Almost all of the newkeynesian models that were to follow, and that have become the mainstream applied models used in the policy sphere, build straight on these real models where an active labor-supply channel is at the core of business-cycle propagation. In sum, labor supply permeates modern business-cycle theory. At the same time, almost all of these models rely on the representative-agent construct. Recently, a new strand of models is emerging—the so-called HANK, or Aiyagali, literature—introducing (nominal) price-stickiness features into heterogeneous-agent models. It is our contention, however, that the question of how households supply labor has not been in focus here. Rather, these new frameworks have focused on consumption (in particular on the heterogeneity in the marginal propensities to consume), while the microeconomics of labor supply is just an add-on: the labor-supply modeling used for the representative agent is just imported into the model with heterogeneity. There is much to do here and we hope the present paper will be useful in this endeavor; an interesting recent paper exploring wage-setting issues from the household perspective is Åhl (2019).

Long-run macroeconomic analysis has focused less on labor supply; the neoclassical growth model and the most common extensions to it simply abstract from labor supply. However, as Boppart and Krusell (2016) argue, if the income effects are strong enough, we should see systematic (downward) trends in how much households desire to work, affecting growth more generally, so long as aggregate productivity keeps growing. The nature of technical change and its importance for the labor markets has been more in focus, especially recently; see, e.g., Acemoglu and Restrepo (2017, 2018) and Aghion et al. (2017), which emphasize effects of AI, automation and other technological developments on unemployment and inequality through labor demand. Our perspective here is simply that in the longer run, labor supply will still surely play a key role, and it has been underemphasized in this literature and therefore deserves more attention.

The paper is organized as follows. In the initial part, section 2, we set out the benchmark model we aim to study: one with household heterogeneity, an active labor-supply channel, incomplete markets, and aggregate shocks. In this section, the model is set up to provide a direction for the rest of the paper and the research as a whole. The sections that follow will both consist of much simpler models and extensions. In particular, all of section 3 studies complete markets, though some of the material consider long-run growth, which is not described in the initial section. This section also proposes the model(s) with an extensive margin, a model which is somewhat different than that used in the literature; here is also where we discuss a model with both an extensive and an intensive hours choice. Specific highlight are: aggregation, in section 3.2.2 and section 3.3.2; the extensive margin, in section 3.3; the who-should-work question, in section 3.4; and the long-run growth behavior, in section 3.5.

In section 4, we define our incomplete-markets equilibria, now with long-run technology growth and with an extensive margin. In this section we briefly discuss computation, after which we calibrate the model and then display all the key results and model comparisons. Section 5 concludes.
2. The quantitative model

In this section, we describe a simple version of the HRBC (the Heterogeneous-agent Real-Business Cycle) model we aim to analyze. It is a simple version in that it coincides with the setting in Krusell and Smith (1998) and the purpose of the remaining sections of the paper is to build toward an extension of this model that allows for growth, an extensive-margin labor supply, and a more general class of utility function. The steady state corresponding to the model described here is the standard Aiyagari (1994) setting, augmented to allow for endogenous labor supply.

2.1. The benchmark model

There is a unit mass of households, each with some asset level \( a \) and some idiosyncratic productivity state \( \omega \). We denote the joint distribution of assets and productivity across people by \( \Gamma \). The remainder of the variables will be described as the definition of equilibrium is laid out. The benchmark model—defined as a recursive competitive equilibrium (RCE)—can thus be described as follows.

**Definition 1.** An RCE consists of pricing functions \( r_k \) and \( w \), a value function \( V \), decision rules \( f_a \) and \( f_h \), an aggregate labor supply function \( H^n \), and a law of motion for the distribution, \( H^k \), such that:

1. \( V \) solves the household’s problem: for all \((a, \omega, \Gamma, z)\),

\[
V(a, \omega, \Gamma, z) = \max_{a', h} \left\{ u(a(1 - \delta + r_k(\Gamma, z)) + h\omega(\Gamma, z) - a', h) + \beta E[V(a', \omega', H^k(\Gamma, z), z') | \omega] \right\}
\]

s.t. \( a' \geq a, h \in [0, \infty) \).

2. \( f^a(a, \omega, \Gamma, z) \) and \( f^h(a, \omega, \Gamma, z) \) solve the maximization problem on the right-hand side of the dynamic-programming equation above for all \((a, \omega, \Gamma, z)\).

3. \( r_k \) and \( w \) satisfy \( r_k(\Gamma, z) = F_1(\bar{k}, \bar{h}, z) \) and \( w(\Gamma, z) = F_2(\bar{k}, \bar{h}, z) \), where

\[
\bar{k} \equiv \sum_{\omega} \int_a a\Gamma(da, \omega)
\]

and

\[
\bar{h} = H^n(\Gamma, z).
\]

4. \( H^n \) satisfies

\[
H^n(\Gamma, z) = \sum_{\omega} \int_a \omega f^h(a, \omega, \Gamma, z)\Gamma(da, \omega)
\]

for all \((\Gamma, z)\).
5. $H^k$ satisfies
\[
H^k(\Gamma, z)(B, \omega) = \sum \pi_{\omega|\hat{\omega}} \int_{a: f^a(a,\hat{\omega},\Gamma,z)\in B} \Gamma(da, \hat{\omega})
\]
for all $(\Gamma, z)$, all Borel sets $B$, and all $\omega$.

The individual productivity process we assume is exogenous and discrete, i.e., $\omega \in \{\omega_1, \omega_2, \ldots, \omega_I\}$. The aggregate shock, $z$, should be thought of as an AR(1), independent of $\omega$. We assume a standard constant-returns-to-scale production function $F$ and a time-independent borrowing constraint. The model described here is rather standard and is identical to that discussed in the appendix of Krusell and Smith (1998), where $u(c, h)$ was assumed to be a power function of a Cobb-Douglas aggregate of $c$ and $h$. In Krusell and Smith (1998) the individual and aggregate shocks are correlated and both $\omega$ and $z$ have two-point supports.

3. The complete-markets version of the model

The quantitative model just described has been studied in the literature for special cases. Thus, some of its properties are certainly known. Yet the model is rather complex and an important part of the present paper aims to examine some of its basic features, many of which are only touched on, at best, in the literature. We particularly look for the determinants of aggregate consumption and hours worked and the distribution of consumption, hours, and wealth across people. Given that the model must be solved numerically, we can only hope for incomplete insights, but based on studying special cases it is still possible to understand many of the model characteristics in significant depth.

The key special cases we can study in some depth involve (i) complete-markets and (ii) static versions of the model. As for the first case, it is well known that the model just described does not allow Gorman aggregation. However, it is important to point out that there is (at least potentially) more than one departure from aggregation in this model. First, even when abstracting from labor supply as in the basic Aiyagari (1994) model, the savings market is incomplete: households cannot save in state-contingent assets and therefore cannot insure themselves against their idiosyncratic productivity shocks. Thus, the wealth distribution matters for savings and consumption choices. Conversely, the distribution does not matter in a case without labor supply—because individual decisions are affine in asset wealth—if there are no shocks, or there is a full set of contingent claims, but only if the utility function of consumption is in the HARA class; the standard assumption in quantitative macroeconomic analyses is to assume that utility is a power function of consumption, which is part of the HARA class. A closely related, but not necessarily less
important, point is that our models do not aggregate with respect to the heterogeneity in productivity. This issue is an obvious one but at the same time not one that is discussed in the literature. As a result, because there is significant heterogeneity in wealth and wages across people in the data, one of our goals is to assess the quantitative departure from aggregation.

Second, with endogenous labor the utility function does not necessarily aggregate; in fact, most preferences over consumption and leisure assumed in quantitative analysis are not part of the Gorman class (which would allow aggregation in terms of the consumption and leisure choices, again because consumption and hours would be affine in wealth). In particular, the utility function allowing a constant Frisch elasticity—MacCurdy (1981)’s formulation—is arguably the most commonly used formulation in the business-cycle literature and it does not allow aggregation. This is also the function that we will use in the quantitative analysis below. Two important exceptions that do allow aggregation are the Cobb-Douglas function and GHH preferences (Greenwood, Hercowitz, and Huffman 1988). Hence, in a quantitative version of the present model, intra- as well as intertemporal substitution of labor, and hence consumption, depend nontrivially on the distribution of wealth across households. Below, we will pay particular attention to this second source of lack of aggregation in detail, and analyze systematically how the inclusion of heterogeneous agents and endogenous labor affects the economy’s aggregate behavior. Here, preferences that allow a balanced growth path where hours fall over time, as Boppart and Krusell (2016) argue is the most relevant case empirically for developed economies, require income effects that exceed substitution effect along the balanced path and such preferences have not been examined in detail in the literature, even in static frameworks.

A third complication that will be considered in this model is an extensive-margin labor supply. Under complete markets—in this case often expressed using lottery equilibria, as in Hansen (1985) and Rogerson (1988)—the extensive margin should not be viewed as a friction, but under incomplete markets it is a non-trivial source of non-convexity. We will therefore pay some attention to the role of the extensive margin also in simple versions of the present model.

Finally, since we consider heterogeneity in labor productivity across individuals, households face differences in the relative prices (in this case the wage rate) and our framework does not admit aggregation even if markets are complete and preferences are part of the Gorman class.

The outline of the rest of this section is as follows. In section 3.1 we describe the class of utility functions to which we will restrict the analysis.

In section 3.2, we look at a static model with intensive-margin labor supply. We first analyze the household problem in some detail, showing how the Frisch elasticity of the individual household—how labor supply depends on the wage—depends on wealth. Thereafter we turn to aggregation issues, and compare aggregate outcomes, including the aggregate Frisch elasticity, in a model with heterogeneous agents compared to one with a representative agent. In short, how large aggregate labor supply is, and how it responds to
shocks, depends on the distribution of wealth in this model, even under complete markets, so a key question is how quantitatively important the wealth distribution is in these regards.

In section 3.3 we add an operating extensive margin. By restricting the choice set of hours worked we introduce yet another reason why the overall distribution matters for the aggregates: it determines the fraction of constrained households. We show, in particular, how aggregate hours worked and the aggregate Frisch elasticity differ between a (complete-markets) heterogeneous agent model and a representative agent model: these differences can be large.

We then, in section 3.4, briefly turn to a normative question: who should work in our heterogeneous-agent model? To answer this question, of course, one needs to endow the planner with weights applied to all the individuals. We pay specific attention to utilitarian planners—who apply equal weights on individuals—since they are often used in the literature, arguably because agents are identical ex ante in most quantitative incomplete-markets models. It is natural, in this context, to discuss welfare theorems, so we also point to the link between the equilibrium wealth distribution and the planner weights.

In section 3.4.3, finally, we add a time dimension to the analysis. In the absence of capital, an analysis of a long-run social planner problem can be divided up into a sequence of static problems, so we begin with this case. We also discuss the connection between social optima and competitive equilibria here, an issue that is a little less trivial in the dynamic model, though it of course follows the same general logic as in the static model. We finally discuss conditions under which there exists an exact balanced growth path in our economy and we characterize some of its properties.

### 3.1. Restriction to a class of utility functions

We first of all restrict attention to utility functions that are consistent with balanced growth. In Boppart and Krusell (2016) it is shown that an hours path with constant negative hours growth is consistent with a balanced growth path, where the other main economic aggregates—output, consumption, investment, and the stock of capital—all grow at constant rates, if and only if the per-period preferences fall into the “BK class”, i.e., if the utility function is of the form:

\[
\begin{align*}
  u(c, h) &= \frac{\left( c \cdot v(he^{\frac{\nu}{1-\nu}}) \right)^{1-\sigma} - 1}{1 - \sigma} \quad (1) \\
  u(c, h) &= \log(c) + \log \left( v \left( he^{\frac{\nu}{1-\nu}} \right) \right) \quad (2)
\end{align*}
\]

for \( \sigma \neq 1 \), or

where \( v \) is an arbitrary, twice continuously differentiable function. The parameter \( \nu \) is key in that it regulates the relative forces of the income and substitution effects (of wages on labor supply) along a balanced growth path; when \( \nu > 0 \) (< 0), the former (latter) is stronger. The formulation nests the classic balanced-growth utility function with zero
growth in hours as proposed by King, Plosser, and Rebelo (1988): by setting $\nu = 0$ we obtain the standard “KPR class”.

The familiar MaCurdy (1981) formulation,

$$\frac{c^{1-\sigma} - 1}{1 - \sigma} - \psi \frac{h^{1+\frac{1}{\beta}}}{1 + \frac{1}{\beta}},$$

is a special case of this class and it is straightforward to show how it is obtained by the choice of a particular functional form for $v$:

$$v(x) = \left(1 - \frac{\psi(1 - \sigma)}{1 + \frac{1}{\beta}} x^{\frac{1}{\beta} + \frac{1}{\gamma}}\right)^{\frac{1}{1-\rho}},$$

with $x \equiv hc^{\frac{\nu}{1-\nu}}$, and the following parameter restriction:

$$\nu = \frac{\sigma - 1}{\sigma + \frac{1}{\beta}}. \quad (3)$$

We will focus on the MaCurdy function in the quantitative section of this paper, but in the section on the static complete-markets model we also briefly consider a slight extension to this class by setting

$$v(x) = \left(1 - \frac{\psi(1 - \sigma)}{1 + \frac{1}{\beta}} x^{\frac{1}{\beta} + \frac{1}{\gamma}}\right)^{\kappa}, \quad (4)$$

again with $x \equiv hc^{\frac{\nu}{1-\nu}}$, but using a $\nu$ that does not necessarily satisfy (3). This is the “parametric class” suggested by Boppart and Krusell (2016) that nests most frequently used functional forms (like Cobb-Douglas, MaCurdy, and a case of GHH preferences). The additional new parameter $\kappa$ is useful in that it helps regulating the extent of complementarity between $c$ and $h$. In the case when $\nu$ satisfies the restriction (3), we have that if $\kappa = 1/(1-\sigma)$, hours and consumption are additively separable; if $\kappa(\sigma - 1)(\kappa(1 - \sigma) - 1) > 0 (< 0)$, the cross-derivative between hours and consumption is positive (negative). Later, when we introduce both an intensive and an extensive margin, we will focus on the MaCurdy class in order to simplify the exposition and hence use a zero cross-derivative.

3.2. Features of the static model: the intensive margin

We start by assuming that households can choose labor freely along the intensive margin. In other words: there are no restrictions on how much or little the household can choose to work, and we can really think of hours as “effort” in this sense rather than a time restriction.

First, we analyze the household problem in some detail, and contrast KPR preferences with the more general BK preference formulation. Thereafter, we will turn to the question about aggregation.
3.2.1. The household problem

The household maximizes \( u(c, h) \) by choice of \( c \) and \( h \) subject to a budget

\[
c = \omega h + a,
\]

where \( \omega \) is the household’s productivity (wage) and \( a \) is the asset (wealth) level.

3.2.1.1. The effect of wages on hours worked

Under KPR preferences, the optimal amount of hours worked does not change if the wage changes, as long as the wealth of the household changes proportionally; on a balanced growth path, KPR preferences guarantee that wealth and wages grow at the same rate. On the other hand, if wealth is not scaled with the wage change, wages affect hours: with constant, positive (negative) wealth, hours worked are increasing (decreasing) in the wage. Figure 1 illustrates by plotting isohours curves. At zero wealth, the isocurve is vertical: a given level of hours is consistent with any productivity level. For a positive (negative) wealth level, a productivity increase increases (decreases) hours worked; the linearity reflects the cancellation of income and substitution effect defining KPR preferences.

When we go outside the KPR class to the more general BK class, these statements need to be altered. To illustrate this simple point, we use the MaCurdy (1981) utility function. It is given by

\[
u(c, h) = \frac{c^{1-\sigma} - 1}{1 - \sigma} - \psi \left( \frac{h^{1+\frac{1}{\beta}}}{1 + \frac{1}{\beta}} \right).
\] (5)

When the income effect exceeds the substitution effect (along a balanced growth path), so that hours fall when a country gets richer, we need to have \( \sigma > 1 \). In contrast, the case of \( \sigma = 1 \) corresponds to KPR preferences.

For the case with \( \sigma > 1 \), Figure 2 illustrates how, for a given asset level, the optimal choice of hours and consumption depends on the productivity level in a non-monotonic
way (the example has $\sigma = 2.5$). Given a positive level of wealth, and for sufficiently low wage levels, hours increase as wages increase, just like under KPR preferences. Intuitively, the wage bill is small enough here that wage changes do not affect the household’s wealth much and hence the substitution effect dominates. As wages grow further, however, eventually their effect on work is always negative; now wages dominate the household’s income and the income effect then dominates in this preference class. Under KPR and a fixed positive wealth level, a wage increase raises hours by less and less as the wage grows; under BK and $\sigma > 1$, it goes further and eventually leads to falling hours.

We summarize these insights compactly as follows:

Finding 1. Under KPR preferences, for a given positive (negative) level of asset wealth, hours worked increase (decrease) as wages grow. Under BK preferences, wages have a non-monotone effect on hours; under MaCurdy preferences with $\sigma > 1$, wages first increase and then decrease in hours.

The non-monotonicity feature of BK preferences will be important for understanding hours and participation in the long run, which we study later in the paper.

3.2.1.2. The Frisch elasticity of the individual household We now turn to the Frisch elasticity of the individual household: the elasticity of hours with respect to the wage rate, given a constant marginal utility of wealth. This elasticity is relevant for two reasons. First, it helps us understand the nature of consumption smoothing for a given individual; in the incomplete-markets model below, the extent of such consumption smoothing is a key determinant of inequality as well as macroeconomic aggregates. If households vary their hours in response to wages, we will see larger earnings heterogeneity and higher overall output than if they do not. Second, the individual Frisch elasticity
Figure 3: Household Frisch elasticity as a function of assets for different parameters of the utility function. The upper row has \( \nu = 0.23 \), the lower row has \( \nu = 0 \). The three columns represent three different values of \( \theta \). The lines in each subgraph represents three different values of \( \kappa \).

is, of course, a determinant of the aggregate Frisch elasticity, which is of great importance for how the aggregate economy responds to shocks (such as aggregate productivity shocks).

We will first analyze how the elasticity of the individual household depends on its asset holdings, assuming a constant productivity.

The Frisch elasticity is given by

\[
\frac{u_2 u_{11}}{h (u_{22} u_{11} - (u_{12})^2)}
\]

at the household’s optimal choice of \( c \) and \( h \).

Figure 3 shows the resulting Frisch elasticity as a function of the household’s asset holdings (on the x-axis) for a number of different parameter combinations in the utility function.\(^4\) All the six subplots maintain the same value of \( \sigma \) and the same value of \( \psi \). The three red lines in any given plot correspond to different values of \( \kappa \). Different combinations of \( \theta \) and \( \nu \) then appear in the different subplots. This somewhat arbitrary way of illustrating

\(^4\) We restrict the parameters so that the utility function satisfies standard regularity conditions: \( u_1 > 0, u_2 < 0 \), and the determinant of the Hessian \( u_{11} u_{22} - (u_{21})^2 > 0 \).
the different roles of different parameters and of wealth (or productivity, as in Figure 4) is used to illustrate some specific findings, each of which is then summarized in a verbal statement.

The first row of graphs in the figure correspond to a BK formulation (with \( \nu > 0 \)). From left to right, the value of \( \theta \) increases, and the middle value of \( \theta \) is such that, given the maintained value for \( \sigma \), the resulting utility function is MaCurdy for the middle level of \( \kappa \), and hence has a Frisch elasticity that is wealth-independent, as illustrated by the horizontal dashed line. We see that \( \kappa \) affects the Frisch elasticity, and more so for larger values of \( \theta \). More importantly, we see that as wealth increases, the Frisch elasticity converges to \( \theta \), independently of all the other parameter values, so \( \kappa \)'s importance is only played out for low wealth levels.

The second row of the figure displays cases within the KPR class (i.e., where \( \nu = 0 \)). Here we again see the same qualitative effects of \( \kappa \) and of wealth. In this row of graphs, no case involves a constant Frisch elasticity, since \( \sigma \) is still 1.7 (since with \( \nu = 0 \), only \( \sigma = 1 \) gives the MaCurdy, constant-Frisch case). We summarize our finding as follows.

**Finding 2.** As wealth rises, the Frisch elasticity approaches \( \theta \) for any utility function within the class of utility functions given by (4). The level of \( \kappa \) also matters, and potentially significantly so, for the Frisch elasticity, so long as asset wealth is not too high.

Clearly, with many consumers with low wealth, there can be significant heterogeneity in Frisch elasticities, even in the complete-markets economy.

We now turn to heterogeneity in productivity. A first observation is that if the household has zero assets, the Frisch elasticity does not depend on productivity: productivity is just a scaling variable in this case.

**Finding 3.** With zero wealth, the Frisch elasticity is a constant that does not depend on productivity, but only on \( \sigma, \theta, \nu, \) and \( \kappa \).

With positive assets holdings and for very low productivity levels the asset holdings are important in relative terms, so here the Frisch elasticity approaches \( \theta \). As productivity increases, the Frisch elasticity approaches the level that it delivers at zero wealth. This is all illustrated in Figure 4. We display the same preference specifications, with a BK case in the first row, the dashed line in the middle subplot being a MaCurdy function. Here, the departures from a constant Frisch elasticity are smaller. In the KPR case in the second row, we also see modest departures from a constant Frisch elasticity.

For the sake of completeness, Figure 26 in the appendix shows how the Frisch elasticity varies for different levels of productivity, given that the household has negative assets. For productivity levels large enough, the asset holdings are of less importance, and the Frisch elasticity approaches the “zero-wealth constant”. For lower productivity levels, the Frisch elasticity diverges from that constant.
Figure 4: Household Frisch elasticity as a function of productivity for different parameters of the utility function. Asset holdings are assumed to be positive. The upper row has $\nu = 0.23$, the lower row has $\nu = 0$. The three columns represent three different values of $\theta$. The lines in each subgraph represents three different values of $\kappa$.

Clearly, the model’s predictions for how households respond to temporary wage changes, as measured by the Frisch elasticity, are nontrivial within the considered preference class, which includes KPR preferences. In particular, the Frisch elasticity depends on the household’s assets and specific productivity level. How these differences across households add up to an aggregate response of a temporary wage change is considered in section 3.2.3.

3.2.2. Quantitative departures from aggregation

The preference classes we study here do not admit Gorman aggregation. As already pointed out above, this is a general feature of most applied aggregate models of endogenous labor supply, and most of these models are representative-agent models. The question addressed in what follows is how much of a difference this non-aggregation makes in practice. To give a rough quantitative answer to this question we will use the static model just studied in a version that corresponds to a snapshot from a general-equilibrium model with exact balanced growth, but with a nontrivial distribution of agents over wealth and productivity levels. In particular, we will impose an exogenous bivariate distribution over assets and productivity that is similar to that observed in the data.
Figure 5 shows both the asset distribution and the distribution of assets vs. productivity.\(^5\)

In the incomplete-markets models that are the ultimate aim of this study, the wealth-productivity distribution is endogenous. In the present section, it is not, due to the assumption of complete markets: the distribution can be seen as a “parameter” that we are free to choose; hence we choose it to approximately mimic the data.

In our general-equilibrium model here, we assume a unit mass of households, each with an individual productivity \(\omega_i\) and an individual asset holdings \(a_i\). The households rent out their labor services and capital services to a production sector that produces with a standard Cobb-Douglas technology.

Hence, the household wants to maximize utility, subject to the budget constraint (omitting the \(i\) subscripts for readability), which reads

\[
\max_{c,h} u(c, h) \quad \text{s.t.} \quad c = h\omega w + (1 + r - \gamma)a \tag{6}
\]

where \(r\) is the rental rate of capital and \(w\) is the economy-wide wage rate, in equilibrium given by \(F_1(K, L) - \delta\) and \(F_2(K, L)\) respectively. The gross growth rate \(\gamma\) appears in the budget because there is growth at that rate; hence the equation is expressed in terms of variables that are transformed into stationary form. The resulting static model is thus our

---

\(^5\)Assets are the sum of cash, bonds, stocks, business assets, pension assets and real estate, net of mortgages and other debt. For productivity, we use the average of the observed wages for the last 10 years in the case of the individual having at least one recorded wage during this time period. To calculate hourly wage we take annual labor income (sum of regular labor income, labor income from business, and farm income equally split between husband and wife) and divide by annual hours. Wage observations below half the state minimum wage for that particular year are set to that number. If the individual has no wage observations in the period 1998–2008, we impute a productivity based on observables (age, race, education, marital status, presence of children in the household, if the individual supports a child living outside the home).
way of capturing the economy’s behavior along a balanced growth path, where productivity grows at rate \( \gamma \).

Along a balanced growth path the interest rate is given by \( r = \gamma / \beta - 1 \), and we assume \( \gamma = 1.02 \) and \( \beta = 0.98 \). The capital share in the production function is assumed to be one third and the depreciation rate is 5% yearly. We use the utility function as defined in (4), vary the parameters in it, and analyze how heterogeneity in assets and/or productivity affects aggregation quantitatively.

### 3.2.2.1. Heterogeneity in assets only

First, we ignore the heterogeneity in productivity and assume that everyone in the economy has equal productivity, \( \omega_i = 1 \forall i \). Asset holdings are heterogeneous and correspond to the asset distribution we observe in the data. With heterogeneous assets and preferences that are not in the Gorman class, the marginal propensity to decrease hours worked out of wealth differs between rich and poor households and therefore the wealth distribution affects total hours worked. Hence, we should expect a difference between the heterogeneous-agent case and the representative agent case; the question is just how large it is.

We select \( \psi \) (this parameter guides the level of labor supplied by the households) so that the labor market and the capital market clear at the balanced growth interest rate level. Thereafter we compare the outcome to one where the assets are redistributed so that all households have the same amount of assets. In other words, we give each household the average asset holdings, and thereafter find the new corresponding equilibrium. Thus, in this “representative-agent” economy individual households are literally homogeneous in every respect: preferences, productivity, and assets. We then measure the difference in outcomes between those two economies: how much more (or less) labor is supplied in the economy with heterogeneous asset holdings?

The results for a number of parameter combinations—close to calibrations considered in the literature—in the utility function (adjusting \( \theta, \nu, \kappa \)) are shown in Figure 6. As can be seen, the departure from aggregation depends on the parameters in the utility function, but is limited to around 1%. As previously pointed out, there are departures from aggregation also for utility functions within the KPR class (the case of \( \nu = 0 \)); if anything, as seen in the figure, the KPR class gives larger departures. The departures are also increasing in the parameter \( \kappa \).

**Finding 4.** The departures from aggregation in total hours worked due to wealth heterogeneity alone are between 0 and 1% for the parameter constellations considered, including KPR preferences. Work is higher under wealth heterogeneity.

### 3.2.2.2. Heterogeneity in productivity only

Next, we assume that everyone in the economy has an equal amount of assets but that productivity is heterogeneous and corresponds to what we observe in the data. That this economy does not aggregate is also clear; heterogeneous productivity can be viewed as heterogeneity in the relative price between consumption and leisure such that preferences do not even aggregate if they are part of the Gorman class.
Figure 6: Difference in total hours worked between model with heterogeneity in assets (using actual asset distribution) and the corresponding representative agent solution. The x-axis indicates different levels of \( \kappa \). The three panels indicate three different levels of \( \theta \). The lines indicate different values of \( \nu \).

Figure 7: Difference in total hours worked between model with heterogeneity in productivity (using actual productivity distribution) and the corresponding representative agent solution. The x-axis indicates different levels of \( \kappa \). The three panels indicate three different levels of \( \theta \). The lines indicate different values of \( \nu \).

The experiment is exactly the same as in the previous section: for a number of different configurations of parameters in the utility function we first solve for the equilibrium in the economy with heterogeneous productivity, and thereafter compare to an economy where productivity is distributed evenly \((\omega_i = 1 \ \forall i)\). The results are shown in Figure 7. The deviation between the representative-agent economy and the heterogeneous-productivity economy now ranges between hours worked being 3% more to being 3% less in the case of heterogeneity in productivity.

Finding 5. The departures from aggregation in total hours worked due to productivity heterogeneity alone are between -3% and 3% for the parameter constellations considered.

3.2.2.3. Heterogeneity in both assets and productivity Finally, we introduce heterogeneity in both the asset and productivity dimensions, with a correlation structure like
that in the data. Again, we compare the labor supply in the economy with heterogeneity to the labor supply in the representative-agent economy. The results are shown in Figure 8 and, as can be seen, the difference between hours worked in the economy with heterogeneous agents and those in the representative-agent economy is now between 0 and 6%. Thus, the interaction of heterogeneity in wealth and productivity raises hours worked.

Finding 6. The departures from aggregation in total hours worked due to wealth and productivity heterogeneity together are up to 6% for the parameter constellations considered. Work is higher under heterogeneity.

3.2.3. The Frisch elasticity in the aggregate

Since the Frisch elasticity of the individual household differs depending on the asset holdings and productivity, it should be clear that the average Frisch elasticity in the economy differs depending on what we assume about heterogeneity.

To examine this difference quantitatively, we calculate an average Frisch elasticity by simply taking the weighted average across households in the economy with heterogeneity (in either assets, productivity, or both). This measure we can then compare to the Frisch elasticity of the representative agent.

The results are shown in Figure 9. As can be seen, the difference between the representative agent’s Frisch elasticity and the average Frisch elasticity in the heterogeneous agent economy does not exceed 4% in any of the parameter combinations. Of course the difference is zero for the parameters that correspond to the MaCurdy formulation (i.e., $\theta = 0.5$, $\kappa = 1.43$ and $\nu = 0.19$), since then the Frisch elasticity is constant and equal to the representative-agent case for all households regardless of productivity and asset holdings. From the figure, the following broad conclusions can be drawn.
Finding 7. The aggregate Frisch elasticity in the economy with heterogeneity in wealth and productivity is up to four percent higher than in the corresponding representative-agent economy. Wealth heterogeneity accounts for a higher elasticity and productivity heterogeneity for a lower one.

3.3. The static model with an extensive (and intensive) margin

So far we have only dealt with labor choice along the intensive margin. However, if we look at people’s working hours, many people work zero hours. In the U.S., the civilian labor force participation in the beginning of 2019 was 63%. Even in the age group 25–54, i.e., among prime-aged individuals, 17% of the population is out of the labor force. It thus seems hard to ignore the extensive margin in a model of labor supply.

As argued by Hansen (1985) and Rogerson (1988), models with an extensive-margin labor choice can behave very differently than models with an intensive-margin choice only. Their settings, however, are frameworks with full insurance and, in essence, the representative-agent feature is then kept to a large extent. In contrast, in incomplete-markets economies households are not fully insured and as a result can be in very different situations. Therefore, their responses to shocks can differ very widely; this point is also clear from previous work, such as Chang and Kim (2007) and Krusell, Mukoyama, Rogerson, and Şahin (2008) in the context of a steady state, and Krusell, Mukoyama, Rogerson, and Şahin (2017) in the context of aggregate fluctuations. What we point out here is that some of these heterogeneous effects can be studied also under complete markets: i.e., we simply explore how agents with different wealth and productivity positions differ in their hours worked and in their responses to shocks.

In this section, we will thus look at a framework where there is only an extensive margin but we will also consider the case with both an extensive and an intensive margin. We introduce a tractable way of modelling these features and discuss the implications for aggregation and for the aggregate Frisch elasticity. We will keep the static setting for the remainder of this section and we will now limit attention to the MaCurdy utility function—thus setting $\alpha_{12} = 0$—in order to simplify the exposition.

3.3.1. The household problem with both margins

To introduce an extensive margin into the labor supply choice we assume that the hours choice is constrained as follows:

$$h_i \in \{0\} \cup [\bar{h}, 1], \forall i.$$  \hspace{1cm} (7)

If $h = 0$, the problem is the same as in the previous section, with intensive margin choice only (we impose an upper bound for hours worked of 1 here for simplicity only; in general we can use any $\bar{h}$). When $h > 0$ this is a non-convex set; if it equals 1, the constraint implies that the household can only choose between working and not working, i.e., only has an extensive margin choice. We do not consider the deeper sources of the non-convexity,
Figure 9: Difference between the average Frisch elasticity in the economy with heterogeneous agents and the corresponding representative agent solution. The x-axis indicates different levels of $\kappa$. In each row, the three panels indicate three different levels of $\theta$. The lines indicate different values of $\nu$. 

(a) Heterogeneity in assets.

(b) Heterogeneity in productivity.

(c) Heterogeneity in assets and productivity.
which could be various forms of fixed costs, such as commuting costs or the need to purchase specific clothes for work, or involve ways in which productivity per time unit falls when less time is spent on the job. The main advantage of our formulation is its simplicity.\footnote{Alternative formulations could be to use a part-time penalty, as in, e.g., French (2005), or a combination of fixed cost and non-linear earnings, as in Erosa et al. (2016).}

We will now look at the household problem in detail in the presence of this non-convexity before turning to the implications for the household’s labor-supply elasticity.

3.3.1.1. The household problem and basic results  We start by writing down the household problem to gain some intuition about the non-convex hours choice set. Consider a static problem where an agent is endowed with a certain asset level, and choose labor and consumption. Hence, the agent’s problem is to maximize the one-period utility \( u(c, h) \) with respect to \( c \) and \( h \) given the budget constraint \( c = \omega h + a \). We use the MaCurdy formulation for utility:

\[
u(c, h) = \frac{c^{1-\sigma} - 1}{1 - \sigma} - \psi \frac{h^{1+\frac{1}{\sigma}}}{1 + \frac{1}{\sigma}}.\]

Taking the first-order conditions we obtain the familiar relationship for marginal utility of consumption and marginal disutility of work:

\[
c^{-\sigma} = \frac{\psi h^{\frac{1}{\sigma}}}{\omega},\]

which, combined with the budget constraint, gives us the preferred hours and consumption choice as functions of the asset level \( a \) and the productivity level \( \omega \). We call these preferred choices from the unconstrained problem \( h^*(a, \omega) \) and \( c^*(a, \omega) \) respectively.

Now assume that the hours choice is constrained as described above: \( h \in \{0\} \cup [h, 1] \). If the household’s preferred hours choice from the unconstrained problem, \( h^*(a, \omega) \), falls within \([h, 1]\), the solution to the constrained problem is the same as in the unconstrained problem. But what if \( h^*(a, \omega) \in (0, h) \)? Then the choice of working \( h \) or 0 is determined by whether or not \( u(a + \omega h, h) \) exceeds \( u(a, 0) \). With our choice of utility function, the decision to work \( h \) or not work at all is hence given by

\[
\frac{(a + h\omega)^{1-\sigma}}{1 - \sigma} - \psi \frac{h^{1+\frac{1}{\sigma}}}{1 + \frac{1}{\sigma}} \leq \frac{a^{1-\sigma}}{1 - \sigma}.
\]

We call the choices from a restricted problem \( h^R(a, \omega) \) and \( c^R(a, \omega) \).

For a given productivity level the household will choose to work if the asset level is low. However, if the asset level is large enough, the household will switch to not working.
3.3.1.2. Introducing lotteries

Given the presence of the non-convexity in the choice set for \( h \), in some cases households would, if they could, choose to randomize between 0 and \( h \). We will consider randomization in some of our analysis below. When we do, we thus assume that a household can assign a probability \( e \in [0, 1] \) to be employed and work. Hence, the household’s problem is now given by:

\[
\max_{c,h,e} \left[ c^{1-\sigma} - 1 - e\psi \frac{h^{1+\frac{1}{\beta}}}{1 + \frac{1}{\beta}} \right]
\]

subject to

\[
c = e\omega h + a \quad \text{(9)}
\]
\[
e \in [0, 1] \quad \text{(10)}
\]
\[
h \in [h, 1]. \quad \text{(11)}
\]

We will assume for now, just to simplify the notation and discussion, that the combination of assets, productivity, and preference parameters is such that the constraint \( h \leq 1 \) is not binding. We then have the following first-order conditions:

\[
c^{-\sigma} - \mu_1 = 0
\]
\[-e\psi h^{\frac{1}{\beta}} + \mu_1 \omega + \mu_4 = 0
\]
\[-\psi \frac{h^{1+\frac{1}{\beta}}}{1 + \frac{1}{\beta}} + \mu_1 \omega h + \mu_2 + \mu_3 = 0
\]
\[\mu_2 e = 0
\]
\[\mu_3 (1 - e) = 0
\]
\[\mu_4 (h - h) = 0,
\]

where the \( \mu \)s denote Lagrangian multiplier with the Kuhn-Tucker conditions \( \mu s \geq 0 \).

If the solution to the unconstrained problem for a given asset level and productivity level, \( h^*(a, \omega) \), is larger than \( h \), the solution to this problem with lotteries, denoted by the superscript \( L \), is simply given by \( h^L(a, \omega) = h^*(a, \omega) \), \( c^L(a, \omega) = c^*(a, \omega) \) and \( e^L(a, \omega) = 1 \).

When a household choose to randomize, i.e., \( e^L(a, \omega) \in (0, 1) \) we obtain the following expression for consumption:

\[
c = \frac{1}{\omega h} \cdot \frac{h^{1+\frac{1}{\beta}}}{1 + \frac{1}{\beta}}.
\]

Hence, only at this consumption level is the marginal benefit of consumption equal to the marginal disutility of increasing the fraction working. If this consumption level is between \( a \) (which is the consumption the household achieves by not working) and \( \omega h + a \) (which is the consumption the household achieves by letting everyone work \( h \)), the household will choose to randomize.
3.3.1.3. Illustrations  It is informative to compare the household choices for three cases: the unconstrained choice \((h = 0)\), the constrained choice \((h > 0)\), and the constrained choice where we allow for lotteries.

Figure 10 contrasts the optimal choice of hours and consumption for these three cases, and how the choice depends on asset holdings. Hours is (weakly) decreasing for all cases: the richer the household, the less it will choose to work. For the unconstrained problem, the hours will gradually fall towards (but never reach) 0. For the constrained problem, hours will gradually fall until they hit the \(h\) level. Then for a wealth range hours will be constant at this level, until the household is rich enough to choose \(h = 0\). The convexification of this problem makes the drop from \(h = h\) to \(h = 0\) go via assigning less and less probability to work until the probability is 0.

In contrast, for optimal consumption, the non-convexity creates a non-monotonic jump at the point where the household is rich enough to decide to withdraw completely from the labor market. The convexification of the problem, by allowing the household to choose a fraction working, removes the non-monotonicity in the consumption choice. In the region where the household is using lotteries, the consumption is flat.

When the household becomes richer in terms of assets, the constraint in the hours choice becomes less and less binding. Asymptotically, when wealth goes to infinity, the hours choice in the unconstrained problem approaches zero and the constrained and the unconstrained solutions coincide.

Next, Figure 11 shows how the optimal choices of hours and consumption depend on productivity. Since we assume \(\sigma > 1\), i.e., we work outside the KPR class and assume that the (balanced-growth) income effect of a wage increase on hours exceeds that of the substitution effect, optimal hours is not a monotonic function; this is entirely in line with the findings in section 3.2.1, but now the focus is on how the extensive margin affects household outcomes. Recalling the key intuition, for a very low household productivity, there is little point in working, since the wage is just too low. At the same time, because the wage is low in this region relative to assets, the effect of a wage increase is that the substitution effect dominates, so that hours worked rise. For a middle productivity level, the choice of hours is at its maximum. For high productivity, the income effect dominates, and the optimal hours is decreasing in productivity.

In this case, the higher is productivity, the more binding is the hours constraint. When productivity approaches infinity, an unconstrained household would choose to work an infinitesimal amount of hours (still delivering substantial income), while the constrained household has to keep on working at \(h\). The household allowed to randomize, on the other hand, approaches the unconstrained solution as productivity increases.\(^7\)

3.3.1.4. The Frisch elasticity of the individual household  One implication of using the MaCurdy preferences is, as previously mentioned, that the Frisch elasticity is constant and equal to \(\theta\). With non-convexity in the choice set for hours, this is no longer the

\(^7\)In the appendix, section A.3, we formally compute the welfare cost of the non-convexity and how it depends on, e.g., productivity.
Figure 10: Illustration of the choice of hours and consumption as a function of assets, in the unconstrained problem compared to the constrained problem, with and without lotteries. For the lottery problem, the hours choice is defined as \( eh_L(a) \). Productivity is fixed, \( \omega = 1 \) and we have \( h = 0.4 \).

Figure 11: Illustration of the choice of hours and consumption as a function of productivity, in the unconstrained problem compared to the constrained problem, with and without lotteries. For the lottery problem, the hours choice is defined as \( eh_L(a, \omega) \). Assets are fixed, \( a = 1.5 \) and we have \( h = 0.4 \).
CHAPTER 2

case. Again, we can gain insight by looking at Figure 10. For the asset region where the choices of hours coincide for the unconstrained and the constrained problem, the Frisch elasticities also coincide. However, for the asset region where the household is choosing $h$, the lower bound of working, the Frisch elasticity is effectively zero. For the asset region where the household continuously shifts from working to not working, i.e., is randomizing, the objective function becomes

$$u(c, h) = \frac{c^{1-\sigma}}{1 - \sigma} - c\psi \frac{h^{1+\frac{1}{\sigma}}}{1 + \frac{1}{\sigma}},$$

(12)

which is linear in the choice variable $c$, and consequently the Frisch elasticity tends to infinity. For higher asset levels, where the household works zero hours, the Frisch elasticity is effectively zero again. These features that the elasticity is infinity over a range is of course well known from Hansen (1985) and Rogerson (1988) but here the focus is on the regions where it is, instead, zero. How households are placed across 0, $\theta$, and infinity Frisch elasticity regions is crucial in models with significant heterogeneity across households, whether or not markets are incomplete. We turn to this issue next.

3.3.2. Aggregation and the aggregate Frisch elasticity

Just as the simpler economy with only intensive-margin hours choice, the economy described in this section, with non-convexities in the hours choice, does not admit a representative agent. By restricting the choice set of hours worked with a positive $h$ some agents are potentially constrained or forced to randomize between 0 and $h$. This introduces yet another reason why the overall distribution starts to matter for the aggregates: it determines the fraction of constrained households, on top of the reasons mentioned earlier (the fact that the MaCurdy preferences are not in the Gorman class and that we allow for heterogeneity in productivity).

In this case, the difference between the heterogeneous-agent economy and the economy with a representative agent who can choose any number of hours becomes larger. Quantitatively, the difference depends on how many individuals are restricted. Figure 12 shows the difference in total hours worked for two cases: one where the lower limit for work, $h$, is set to 0.4 and one where it is set to 0.43; the difference here is thus small. Note that hours worked for a household with average assets and productivity (or the actual numbers worked for the representative agent) is 0.452, so in both cases we are setting the lower limit for work below the average. As can be seen from the figure, the difference between the aggregate hours worked in the heterogeneous-agent economy now differs substantially from the aggregate hours worked in the representative agent economy and can easily amount to more than 10%. Thus, heterogeneity per se—under the preferences adopted here and when there is an extensive-margin labor choice—can matter significantly for hours worked in the aggregate economy. Thus, in our incomplete-markets economy studied later, this force will be present as well.

Finding 8. Aggregate hours worked is quite sensitive to the precise constraints on hours choice on the individual level for realistic wealth-productivity distributions.
We now turn to the Frisch elasticity. As discussed above, the Frisch elasticity for the individual household can be $\theta$, 0, or $\infty$. Since the fraction of households falling into each category depends on household heterogeneity, the aggregate Frisch elasticity is a non-trivial object.

To give an idea, we solve for equilibrium in the heterogeneous agent economy in the same way as before. We then imagine a wage increase by 1%, and take note of how much the labor supply would increase for each household. For a household with a Frisch elasticity of $\theta$, the labor supply would increase accordingly, for a household which is currently randomizing between 0 and $h$, the labor supply would increase to $h$, and for a household with a Frisch elasticity of 0 there would be no effect. We then sum up the total increase in hours worked and compare that to the previous hours worked in this economy. The result we report as the average Frisch elasticity in the economy. As can be seen in the results reported in Figure 13, the average Frisch elasticity can be both higher and lower in the heterogeneous-agent economy (note that the Frisch elasticity of the representative agent trivially is $\theta$, indicated by the identity line, since we are using MaCurdy preferences).

The discussion based on the static model makes clear that the aggregate Frisch elasticity does not have to remain constant over time, if the distribution of households changes relative to its cutoffs in the decision rule. Over time, there can be significant such changes. In particular, it should be noted that, as we will show later in 3.5.2.1, under some conditions, all the households either randomize or quit work altogether as time goes to infinity. Therefore, we may expect the Frisch elasticity of labor supply to change over time and,

![Figure 12: Difference in total hours worked between model with heterogeneity in assets and productivity and the corresponding representative agent solution. Positive number indicates higher number of hours worked in the heterogeneous agent economy.](image)
3.4. Who should work and when?

In the previous sections we have looked at departures from aggregation and Frisch elasticities individually and in the aggregate; both of these are relevant for positive analysis in the long run and in order to understand fluctuations. We now turn to some normative aspects of our model.

A common approach in heterogeneous-agent macroeconomics is to construct welfare measures. The most commonly used welfare measure is based on the utilitarian social welfare function, which adds utils across the population using equal weights (see Davila, Hong, Krusell, and Rios-Rull (2012) and the references therein). It is either used for simplicity or because, in most models, agents are identical from an ex-ante perspective: they have the same preferences and the same types of constraints, and they are then subject to stochastic processes that make them different ex post. Of course, in studies that consider policy change in a dynamic model for a given initial condition—where there are differences across people due to historical shocks—it may still be relevant to consider other social welfare functions than the utilitarian one.

Unlike in the incomplete-markets model that is the ultimate focus of the present paper, in the complete-markets model we study here there is a one-to-one connection between welfare weights (Negishi weights) and competitive equilibria. In this section, we illustrate

\[ h = 0.40 \] (a)

\[ h = 0.44 \] (b)

Figure 13: The aggregate Frisch elasticity. For description of calculation, please see text.

in general, increase, under BK preferences with income effects exceeding substitution effects.
welfare properties by mostly focusing on two perspectives: a utilitarian social welfare function (given its prominence in the macroeconomic literature) and working backward toward welfare weights that are associated with the decentralized equilibrium under a specific asset distribution. An overriding question for us now is: who should work, and how much? A related question is whether or not the competitive economy under study produces too much work in the aggregate. In a dynamic version of the model we also have to ask about the timing of work. Ultimately, we are interested in these questions in incomplete-markets economies, but it is obviously a first step to analyze them under complete markets.

We begin by looking at the planner’s problem under different assumptions about heterogeneity and planner weights. Thereafter we analyze the correspondence between solutions to the planner’s problem and competitive allocations. Finally, we look at the timing of work, both from a planner’s perspective and for market allocations. Throughout we restrict attention to MaCurdy preferences and we will mostly focus on the case with strong income effects ($\sigma > 1$).

### 3.4.1. Static planning problems

In this section, we show how the answer to the question “who works and how much” in the social planner economy depends on the specific assumptions we make about planner weights and heterogeneity. We will start by examining the set of planner weights most often used in macroeconomics: utilitarian weights.

#### 3.4.1.1. Utilitarian weights

We begin by looking at the role of the productivity distribution for the case of utilitarian planner weights. Hence, we assume that individuals are heterogeneous only in $\omega$. To further simplify the exposition we first assume $h = 1$ so that $h \in \{0, 1\}$, i.e., only the extensive margin is at play. Thus, we simply ask who should work (and not how much).

We can then write the planner’s problem as the task of maximizing

$$\int \left[ \frac{c(\omega)^{1-\sigma} - 1}{1 - \sigma} - \pi(1; \omega) \frac{\psi}{1 + \frac{1}{\sigma}} \right] \Gamma(d\omega),$$  

(13)

with respect to $\{c(\omega), \pi(1; \omega)\}_\omega$, where $\pi(1; \omega)$ thus denotes the working share of the people with productivity $\omega$, subject to the constraints

$$\pi(1; \omega) \in [0, 1],$$  

(14)

and

$$n\lambda = \int c(\omega)\Gamma(d\omega),$$  

(15)

where

$$n = \int \omega \pi(1; \omega)\Gamma(d\omega).$$  

(16)
Here, $\lambda$ denotes the current aggregate productivity level.

Because the preferences are additively separable in $c$ and $\pi$ and concave in $c$, a planner with utilitarian weights will choose the same consumption level for all individuals, i.e., $c(\omega) = \bar{c} \forall \omega$. We thus have $\bar{c}^{-\sigma} = \mu$, where $\mu$ is the Lagrange multiplier on the resource constraint.

The planner chooses $\pi(1; \omega)$ depending on
\[
\frac{\psi}{1 + 1/\theta} \lesssim \mu \omega \lambda. \quad (17)
\]

Since we are assuming heterogeneity between individuals in terms of productivity, this implies a threshold $\omega^* \equiv \frac{\psi}{1 + 1/\theta} \cdot \frac{1}{\mu}$ such that the planner chooses
\[
\begin{align*}
\pi(1; \omega) &= 0 \quad \text{if } \omega < \omega^* \\
\pi(0; \omega) &= 1 \quad \text{if } \omega > \omega^*.
\end{align*}
\]

Hence, the most productive people are assigned to work.

The resource constraint can then be rewritten as
\[
\int_{\omega > \omega^*} \omega \Gamma(d\omega) = \mu^{-\frac{1}{\theta}} \lambda^{-1}, \quad (18)
\]

which, for a given distribution $\Gamma(\omega)$, implicitly defines the unknown marginal cost of resources $\mu$ (note that $\omega^*$ also depends on $\mu$).

To conclude, for every given level of $\lambda$, the planner will choose a cut-off $\omega^*$. Individuals with a productivity below $\omega^*$ do not work, while individuals with higher productivity work. It is also easy to verify that the cut-off $\omega^*$ will be shifting up when $\lambda$ grows, given that we have assumed $\sigma > 1$.

**Finding 9.** With utilitarian weights and an extensive-margin choice only, the most productive people, and only they, work. As aggregate productivity grows, fewer and fewer people work.

This finding obtains generally, i.e., for any productivity distribution, given utilitarian weights. It is an important insight given that many macroeconomic models with heterogeneity use precisely utilitarian weights.

Extending this problem to the case admitting interior hours choices too, i.e., $\mathcal{H} = \{0\} \cup [h, 1]$ with $h < 1$, is straightforward. In this case, the planner maximizes
\[
\int_{\omega} \left[ \frac{c(\omega)^{1-\sigma} - 1}{1 - \sigma} - \int_{h \in \mathcal{H}} \pi(h; \omega) \frac{\psi}{1 + \frac{1}{\theta}} h^{1+\frac{1}{\theta}} dh \right] \Gamma(d\omega) \quad (19)
\]

---

9For some specific distributions $\Gamma(\omega)$ (such as a Pareto distribution) equation (18) can be solved explicitly for $\mu$. 

with respect to \( \{c(\omega), \pi(h; \omega) \forall h \in \mathcal{H} \} \), where \( \pi(h; \omega) \) denotes the share of the population with a given productivity level working \( h \). The maximization is subject to

\[
\pi(h; \omega) \in [0, 1],
\]

\[
\int_{h \in \mathcal{H}} \pi(h; \omega) dh = 1,
\]

and the resource constraint

\[
n \lambda = \int_\omega c(\omega) \Gamma(d\omega),
\]

where

\[
n = \int_\omega \int_{h \in \mathcal{H}} \pi(h; \omega) h dh \Gamma(d\omega).
\]

If the hours choice is interior for an individual with productivity level \( \omega \) then we have

\[
\psi h(\omega) \frac{1}{\theta} = \mu \omega \lambda,
\]

where \( \mu \) again is the Lagrange multiplier on the resource constraint, with \( \mu = \bar{c} - \sigma \). Note that (24) only holds with equality for \( \omega \) with an interior solution. For other productivity levels there might be a mass point for \( h \) at 0, \( \bar{h} \), or 1. We obtain \( h(\omega) = 1 \) for households with \( \omega > \frac{\psi h}{\mu \lambda} \). We have \( h(\omega) = 0 \) or \( h(\omega) = \bar{h} \) for households with \( \psi h \frac{1}{\theta} > \mu \omega \lambda \). So for all individuals with \( \omega < \psi h \frac{1}{\mu \lambda} \) the planner chooses either \( h(\omega) = 0 \) or \( h(\omega) = \bar{h} \) depending on

\[
- \frac{\psi h \frac{1}{1+\theta}}{\frac{1}{1+\theta}} + \mu \omega \lambda h \leq 0.
\]

This implies a threshold \( \omega_1 \equiv \frac{\psi h}{\mu \lambda (1+\theta)} \) such that the planner chooses \( h = 0 \) for all \( \omega \leq \omega_1 \) and \( h = \bar{h} \) for all \( \omega > \omega_1 \).

To summarize, the planner solution will result in three break points:

\[
\omega_1 = \frac{h \frac{1}{1+\theta}}{\frac{1}{1+\theta}} \cdot \frac{\psi}{\mu \lambda},
\]

\[
\omega_2 = \frac{h \frac{1}{1+\theta}}{\frac{1}{1+\theta}} \cdot \frac{\psi}{\mu \lambda},
\]

\[
\omega_3 = \frac{\psi}{\mu \lambda}.
\]

Individuals with \( 0 < \omega < \omega_1 \) will not work at all. Individuals with \( \omega_1 < \omega < \omega_2 \) will work \( \bar{h} \). Individuals with \( \omega_2 < \omega < \omega_3 \) will work \( h = \left( \frac{\mu \omega \lambda}{\psi} \right)^\theta \). Finally, individuals with \( \omega_3 < \omega \) will work highest possible hours, i.e., 1.
In analogy with (18) we can now write the resource constraint as
\[ h \cdot \int_{\omega_1}^{\omega_2} \omega \Gamma(d\omega) + \int_{\omega_2}^{\omega_3} \omega \left( \frac{\mu \omega \lambda}{\psi} \right) \theta \Gamma(d\omega) + \int_{\omega_3}^{\infty} \omega \Gamma(d\omega) = \mu^{-\frac{1}{\sigma}} \lambda^{-1}. \] (26)

For a given distribution \( \Gamma(d\omega) \) this equation implicitly defines the equilibrium \( \mu \) (note that \( \omega_1, \omega_2, \) and \( \omega_3 \) also depend on \( \mu \)).

To conclude, the features we established using the model with only an extensive margin extend to the case with \( h \in \{0\} \cup [h, 1] \). For every given level of \( \lambda \), there will be cut-offs which will guide the working choice, with more productive agents working more or harder, and the cut-offs will be shifting up as everybody becomes more productive, i.e., as \( \lambda \) grows.

**Finding 10.** With utilitarian weights and both an extensive and an intensive margin, work is monotonic in productivity: more productive people work more hours. As productivity grows, people work less and less in terms of the intensive as well as the extensive margin.

### 3.4.1.2. Utilitarian weights and identical individuals

A special case of interest is the one used in many macroeconomic analyses, namely that where everyone in the economy has the same productivity. Hence, we now assume that individuals all have the same \( \omega \equiv \overline{\omega} > 0 \). To simplify the exposition, we restrict attention to an extensive-margin choice only.

The planner still solves the problem defined by equation (13) subject to the constraints (14) and (15). Again, since preferences are additively separable in \( c \) and \( h \) and concave in \( c \), the planner will choose the same consumption level \( \overline{c} \) for all individuals, with \( \overline{c} - c = \mu \).

For the work choice, the planner will now need to use convexification, at least when \( \lambda \) is high enough. Hence, the choice is over the fraction of individuals who will work and it will be made in such a way that the marginal disutility from an additional individual working equals the marginal utility from consumption:
\[ \frac{\psi}{1 + \frac{1}{\sigma}} = \lambda \overline{\omega} \mu, \] (27)

where we assume that \( \lambda \) is large enough for this equation to hold with equality (for low enough values of \( \lambda \), everybody will work). The resource constraint can then be rewritten as
\[ \overline{\omega} \pi(1; \overline{\omega}) = \mu^{-\frac{1}{\sigma}} \lambda^{-1}. \] (28)

Equations (27) and (28) now together define \( \pi(1; \overline{\omega}) \) and \( \mu \). The fraction of individuals working is given by
\[ \pi(1; \overline{\omega}) = (\overline{\omega} \lambda)^{\frac{1}{1-\sigma}} \left( \frac{\psi}{1 + \frac{1}{\sigma}} \right)^{-\frac{1}{\sigma}}. \] (29)

To conclude, when everyone has the same productivity the utilitarian planner is indifferent about who is working in the economy: it only cares about the fraction of individuals working. The fraction working, \( \pi(1; \overline{\omega}) \), is decreasing in \( \lambda \).
CHAPTER 2

Finding 11. With a representative-agent setting and a high enough aggregate productivity, we recover the Rogerson (1988) characterization: convexification is used actively. Interpreted from the perspective of a time sequence of economies with rising aggregate productivity, the fraction of people working every period is strictly decreasing over time.

We will return to the details of growth behavior later, e.g., for determining the asymptotic rate at which the labor force is declining as a function of the rate of productivity growth rate.

3.4.1.3. Heterogeneous weights  Now we turn back to the case where individuals are heterogeneous in \( \omega \). With the utilitarian planner weights we saw above that the most productive individuals were the ones who should be working. How does this conclusion change for other planner weights? The answer, conceptually, is obvious: a higher planner weight will generate more utility for the household, which leads to more consumption and, at least weakly, less work. Let us, however, work out some examples for illustration.

Suppose, to focus the question, that we assume that the weights are identical for all individuals \( i \) with identical \( \omega \), i.e., \( \chi(i, \omega) = \chi(\omega) \forall i, \omega \). We also restrict attention to an extensive-margin choice only. Then we can write the planner’s problem just in \( \omega \) space. The planner thus maximizes

\[
\int_{\omega} \left[ c(\omega) \left( \frac{1}{1-\sigma} - \frac{1}{\pi(1; \omega) + \frac{1}{\theta}} \right) \chi(\omega) \Gamma(d\omega) \right] \tag{30}
\]

subject to the same constraints as before: (14) and (15).

Generally, the work decision is now guided by the following rule:

\[
\omega \lambda \mu \leq \chi(\omega) \frac{\psi}{1 + \frac{\theta}{\mu}}. \tag{31}
\]

It is clear from this equation that what matters for how the working assignment depends on productivity is \( \omega / \chi(\omega) \). Thus, if this object is increasing (decreasing) in \( \omega \), higher-productivity households will work more (less). The planner weight \( \chi(\omega) \) is an abstract one at this point but will soon be given content when it is connected to initial-period asset holdings in the context of the market equilibrium. Turning to consumption, we obtain

\[
c(\omega) = \left( \frac{\chi(\omega)}{\mu} \right)^{\frac{1}{\beta}}, \tag{32}
\]

i.e., whereas hours are increasing or decreasing in \( \omega \) depending on the second derivative of \( \chi(\omega) \), consumption is increasing in \( \omega \) if and only if \( \chi(\omega) \) is increasing, i.e., only the first derivative of \( \chi \) matters.
To take a specific, but interesting example, suppose that the planner weights are $\chi(\omega) = \omega \forall \omega$. Then equation (31) becomes independent of $\omega$:

$$\lambda \mu = \frac{\psi}{1 + \theta}. \quad (33)$$

Assuming that $\lambda$ is high enough that not all households work, this equation must hold with equality, since we must have some individuals working in the economy.

The consumption for any particular productivity level in this case is given by

$$c(\omega) = \left( \frac{\omega \lambda (1 + \frac{1}{\theta})}{\psi} \right)^{\frac{1}{\theta}}. \quad (34)$$

Thus, the consumption plans are pinned down uniquely but there is indifference as to who works, subject to total production equaling total consumption.

**Finding 12.** The planner weights $\chi$ allow much generality in terms of who is assigned to work. The key determinant of whether more productive (higher-$\omega$) households work more or not is whether or not $\omega / \chi(\omega)$ is increasing in $\omega$. If $\chi(\omega)$ is linear, the planner is indifferent as to who works. Regardless, the higher is aggregate productivity, the fewer people work.

Just like above, these results can easily be extended to the case where an intensive margin choice is added ($h < 1$). The insights above then generalize.

### 3.4.2. Static competitive equilibria

We are now ready to turn to the mapping between the planner problem and the decentralized economy. In this section we use the static model to define competitive equilibria, solve for them, and illustrate the correspondence between solutions to the planner’s problem and competitive allocations. Intuitively, the planner weights $\chi(\iota, \omega)$ will correspond, nonlinearly, to the initial asset endowments of households.

A static equilibrium is particularly simple in that there is no equilibrium interaction: all prices are exogenous and it is as if households act in isolation. With more periods, the market for savings is operative and nontrivially determines outcomes through an endogenous interest rate; we look at this issue in detail later in section 3.4.3.

The household’s problem in a decentralized economy without convexification is naturally to maximize

$$u(c, h) = \frac{c^{1-\sigma} - 1}{1 - \sigma} - \psi \frac{h^{1+\frac{1}{\theta}}}{1 + 1/\theta} \quad (35)$$

by choice of $(c, h)$ subject to the constraints

$$c = h \omega \lambda + a(\iota, \omega) \quad (36)$$

$$h \in \mathcal{H} = \{0\} \cup [h, 1], \quad (37)$$
with \( a(\iota, \omega) \) given. However, to obtain a tight link to the set of Pareto optima (as defined by solutions to the planning problem for different weight functions \( \chi(\iota, \omega) \)), we will allow for randomization as described in 3.3.1.2. Hence, the household’s task is to maximize the following convexified problem:

\[
\begin{align*}
\bar{u}(c, h) &= \frac{c^{1-\sigma} - 1}{1 - \sigma} - \frac{\psi}{1 + \frac{\beta}{\sigma}} \int_{h \in \mathcal{H}} \pi(h) h^{1 + \frac{\beta}{\sigma}} dh \\
\end{align*}
\]  

by choice of \((c, (\pi(h))_{h \in \mathcal{H}})\) subject to the constraints

\[
\begin{align*}
\bar{c} &= \lambda \int_{h \in \mathcal{H}} \pi(h) h \omega dh + a(\iota, \omega) \\
\int_{h} \pi(h) dh &= 1, \quad \text{and} \quad \pi(h) \in [0, 1], \quad \forall h \in \mathcal{H},
\end{align*}
\]

with \( a(\iota, \omega) \) given. Regardless of convexification, the household’s choice is independent of the behavior of other households. The market-clearing condition can be stated as requiring that total consumption must be smaller or equal to total production. Trivially, this will be satisfied with equality since utility is increasing in consumption, given that we assume

\[
\int_{\omega} \int_{\iota} a(\iota, \omega) \Gamma(d\iota, d\omega) = 0,
\]

i.e., that the sum of initial endowment claims across people equals zero.\(^{10}\)

3.4.2.1. Comments on the role of convexification We see that, regardless of whether convexification is allowed or not, analysis of the equilibrium of the static economy boils down to analysis of the household’s choice problem. Turning to the role of convexification, let us make some preliminary remarks for a simple case and then consider more general cases. To that end, let us assume that all initial asset endowments are zero, i.e., \( a(\iota, \omega) = 0 \ \forall \iota, \omega \), and that only the extensive is operative (\( \underline{L} = 1 \)). Then we see immediately that a decentralized economy without convexification generates the result that all agents work, no matter how high or low \( \lambda \) is, so long as it is positive. It is easy to show that this case is not generally optimal. In particular, two agents with similar enough endowment levels would benefit from a coin-flip deal where consumption is equalized but work effort is determined by the coin flip. A less trivial question is whether an equilibrium where convexification is not allowed is constrained-optimal, in the sense that there is no alternative allocation (that does not use convexification) in which no agent can be made better off without making another agent worse off. The answer for our simple economy is yes; convexification is necessary in order to improve on the outcome. Constrained suboptimality can arise in models where choices are restricted—such as by not allowing convexification—because these restrictions interact with endogenous equilibrium prices. In the static model here, however, there is no endogenous equilibrium price.

\(^{10}\)This of course is not restrictive; we could assume that there is a net amount of endowments, and then the market-clearing condition is that total consumption must equal total production plus total endowments.
From now on we focus on the cases where convexification is allowed. We will proceed by drawing connections between households’ endowment values and the planner utility weights assigned to them.

### 3.4.2.2. The mapping to planner weights for a general distribution of initial assets

Let us consider a more general initial asset allocation but, for simplicity, still restrict attention to the extensive-margin choice. Let us also assume that there is no asset inequality between households with the same \( \omega \); hence we can write \( a(i, \omega) = a(\omega) \). Each household \( \omega \in [0, \infty) \) now solves

\[
\max_{c(\omega), \pi(1; \omega)} \left[ \frac{c(\omega)^{1-\sigma} - 1}{1 - \sigma} - \pi(1, \omega) \frac{\psi}{1 + \frac{1}{\psi}} \right]
\]

subject to

\[
c = a(\omega) + \pi(1; \omega) \omega \lambda,
\]

\[
\pi(1; \omega) \in [0, 1], \quad \forall \omega,
\]

where \( \pi(1; \omega) \) denotes the fraction of the family with productivity \( \omega \) that is working. We will also assume for now, just to simplify the notation and discussion, that the combination of initial assets, \( \omega \lambda \) and preference parameters are such that the constraint \( \pi(1; \omega) \leq 1 \) is not binding for any household.

The first-order conditions are

\[
\omega \lambda c(\omega)^{-\sigma} = \frac{\psi}{1 + \frac{1}{\psi}} - \eta(\omega)
\]

\[
\eta(\omega) \geq 0,
\]

\[
\eta(\omega) \pi(1; \omega) = 0,
\]

where \( \eta \) is the multiplier on the constraint \( \pi(1; \omega) \geq 0 \): a family with \( \eta > 0 \) is thus in a corner with \( \pi(1; \omega) = 0 \). Of course, the solution to these conditions and the budget will make \( c \) and \( \eta \) be functions of \( \omega \) but we do not need to indicate this by explicit function arguments here. When we refer to the solutions, we will use explicit arguments. So for those households having an interior solution for \( \pi(1; \omega) \), consumption will be given by

\[
c(\omega) = \left( \frac{\omega \lambda (1 + \frac{1}{\psi})}{\psi} \right)^{\frac{1}{\beta}}.
\]

The associated value for the probability of working for such a household will satisfy

\[
\pi(1; \omega) = \left( \frac{(1 + \frac{1}{\beta}) \omega \lambda}{\psi} \right)^{\frac{1}{\beta}} - a(\omega) \right) (\omega \lambda)^{-1},
\]
and the condition for making this a solution to the maximization problem is that, given the parameter values and this household’s $\omega$ (and, hence, its $a(\omega)$), this latter value be between 0 and 1. Looking at this expression, we see that there will be a range of $\omega$ values for which consumption does not directly depend on the asset level.

Now let us try to find a set of planner weights $\chi(\omega)$ that correspond to the asset levels in the equilibrium allocation just derived. Using the planner’s condition for determining consumption levels, (32) in 3.4.1.3, we see that to match consumption levels, $\chi(\omega)$ has to equal (or, really, be proportional to) $\omega$. Recall that, in that planning solution, the planner was indifferent as to who worked. For this range of $\omega$s, the fraction working, $\pi(1; \omega)$, given in equation (49) is trivially consistent with the planner’s solution. For asset/endowment levels such that the household chooses not to work, consumption simply equals asset holdings: $c(\omega) = \mu a(\omega)$, which from the planner’s first-order condition for consumption delivers $\chi(\omega) = a(\omega)^{\sigma}$. This value is also high enough relative to $\omega$—recall that the working decision was determined by $\omega/\chi(\omega)$—that the planner indeed makes this household not work. This completes the mapping between a static competitive equilibrium with an arbitrary asset distribution to a planner allocation with a set of Pareto weights.

**Finding 13.** The equilibrium allocation can be arrived at as a solution to the planner’s problem—an application of the first welfare theorem. For households that choose nontrivial convexification in the equilibrium allocation, the planner is indifferent in terms of what probability to choose, but the household in the equilibrium allocation is not. For convexifying households, higher asset levels do not result in higher consumption, nor in higher planner weights, but merely in lower working probabilities.

Two equilibrium allocations are illustrated in Figure 14: it shows the connection between initial assets and the choice of fraction working for two specific asset distributions and a specific choice of preference parameters. The first distribution has perfect asset equality, and hence work decisions differ by $\omega$. The second equilibrium works out the asset distribution such that everybody makes the same work decision: more productive agents then need to be poorer.

**3.4.2.3. Connection to the planning problem with utilitarian weights** Using the same logic and equations as in the previous section, we can of course start from a planning allocation given some arbitrary weights and work out the associated asset levels. For a particular case, we find the answer instructive: the case of a utilitarian planner, where individuals are heterogeneous in terms of $\omega$. The reason, again, is that the utilitarian planner’s objective is often used in practice, typically motivated with the “behind-the-veil” argument. This argument supposes that there is a pre-stage, at which households are all identical but after which their skills are realized as draws from a common distribution. In the pre-stage, households would then, in an optimal allocation, arrange transfers to be set up ex post to compensate for the risk of receiving low skill draws. The implied optimal allocation is, then, exactly the utilitarian allocation we studied in 3.4.1.1.
1.0 1.5 2.0 2.5 3.0 3.5 4.0 4.5 ... is larger than the increase in utility from the extra
consumption for their productivity at that consumption level.

Figure 14: Choice of $\pi(1; \omega)$ for different productivity levels; in both examples we assume a truncated Pareto distribution of productivity. The orange line shows the asset levels (left axis) and the blue line shows the chosen $\pi(1; \omega)$ (right axis).

(a) The choice of $\pi(1; \omega)$ if everyone is given initial assets $a(\omega) = 0 \forall i, \omega$.

(b) Distribution of initial assets so that every family chooses the same $\pi(1; \omega)$.

So by connecting this particular optimum to initial asset endowments in the decentralized equilibrium we exactly obtain the transfers needed to reach this “macroeconomic” optimum.

To simplify the exposition, we again assume that $h = 1$. As we saw in 3.4.1.1 the planner solution is that everyone consumes $\bar{c}$ and individuals with efficiency above $\omega^*$ work, while those with efficiency below $\omega^*$ do not work.

The asset distribution that gives rise to the same allocations of consumption and work in a decentralized economy as in the planner’s solution is easy to work out. It is given by

$$a(i, \omega) = \bar{c} - h_{SP}(\omega)\omega \lambda,$$

where $h_{SP}(\omega)$ denotes the planner’s choice of work allocation for that particular productivity level.

To give an example, we assume the individuals are distributed along the productivity dimension according to a truncated Pareto distribution with lower and upper bounds $\omega$ and $\bar{\omega}$ and with shape parameter $\kappa$. Then Figure 15a shows the resulting asset distribution as a function of individual productivity. Figure 15b shows the corresponding graph if we choose $h < 1$, so that also the intensive margin is in play.

The intuition behind why the individuals make the same working choice and therefore mechanically the same consumption choice as in the planning optimum is straightforward. For the individuals with high productivity, $\omega > \omega^*$, the initial assets are negative, and hence they are forced to work; moreover, the more productive they are, the higher is the initial debt, so the net of this debt and their labor income cancel to deliver the same consumption allocation as for any other household. For the individuals with low productivity, they have received initial assets to be able to consume $\bar{c}$ without working. The disutility from working for them is larger than the increase in utility from the extra consumption for their productivity at that consumption level.
Finding 14. To decentralize the utilitarian optimum, it is necessary to “tax” high-skill individuals by assigning them large initial debts; the higher are their skills, the larger is the debt. As a result of poverty, they will have to work. Low-skill individuals not assigned to work in the optimum, on the other hand, are recipients of transfers in the form of initial asset endowments.

3.4.2.4. Connection to the planning problem with identical individuals We finally turn back to the case when everyone in the economy has the same productivity. As we saw in 3.4.1.2, the utilitarian planner’s solution is then that everyone consumes $\bar{c}$ and that a fraction of the households work. Moreover, the planner is indifferent as to who works. There are two different ways to think about the optimum in this homogeneous-agent utilitarian case. One is a lottery interpretation: the planner assigns an equal chance (or, rather, risk) of working to all households. The other interpretation is that some households are pre-assigned to work whereas others are not, without involving any lotteries. That is, the utilitarian planning outcome is set up to weight all households’ utilities equally but this does not necessarily imply equal treatment in terms of outcomes, even ex ante.

Formally, the individual’s problem is again to maximize (38) subject to (39) and (40). The first way to decentralize this allocation would be to give a fraction $\pi(1;\omega)$ of the individuals negative assets $a = \bar{c} - \lambda \omega$ and a fraction $(1 - \pi(1;\omega))$ positive assets $a = \bar{c}$. This would give individuals different utility, both in expectation and ex post, but the allocations of consumption and work would correspond to the planner’s solution. This is reminiscent of the result in the dynamic model studied in Krusell, Mukoyama, Rogersson, and Şahin (2008), where at any point in time different households have different assets but some work and some do not, only to later switch; an agent’s asset position rather is an indication of how much an agent will have to work over the remaining lifetime.
The second way to decentralize this economy, one that gives all individuals the same expected utility, is to introduce a lottery. This treatment exactly follows Rogerson (1988) and involves zero initial asset holdings for all agents. The ex-post utility of course will be different: the individuals who “win” the lottery do not have to work and will end up with higher utility than those who “lose” and have to work.

Finding 15. It is possible to decentralize the utilitarian homogeneous-agent optimum in two very different ways: one involves equal ex-ante utilities and lotteries, whereas the other involves unequal ex-ante utilities and no lotteries.

The fact that a given ex-post allocation can be decentralized with and without lotteries—by the reshuffling of initial asset endowments—is an important point that we discuss further in 3.4.3.2 below.

3.4.3. The dynamic model

So far we have only considered static models, though in the absence of capital it is clear that one could interpret a sequence of static economies as a dynamic economy. The connection between social optima and competitive equilibria follows the same logic as in a static model but is a little less trivial to work out with two or more periods because it also involves an intertemporal price: the interest rate. In this section we will focus on a dynamic economy where there is an extensive-margin labor choice only and we will discuss the mapping between competitive equilibria and planning solutions. We will focus on BK preferences with \( \sigma > 1 \). Throughout, we will assume that there is aggregate productivity growth at gross rate \( \gamma > 1 \).

We begin with an illustration of how the solution to a utilitarian planning problem can be decentralized in a multi-period setting: any given person will work for some periods of time and thereafter withdraw from the labor force; the productive people will work more periods. Thereafter, we will look at a two-period model with lotteries to illustrate the mechanisms at work in the clearest possible way. Whether or not lotteries are allowed in the dynamic economy turns out to play an important qualitative role: the timing of work is clearcut with lotteries—we obtain “endogenous retirement”, i.e., people front-load work—but it is not clearcut in an economy where agents cannot convexify.

3.4.3.1. Decentralizing the utilitarian solution Consider a utilitarian planner and heterogeneous productivities among agents. As discussed before, in the absence of capital the planning problem is actually just a sequence of static problems and, hence, no new analysis needs to be considered here. As we saw in section 3.4.1, over time, we simply obtain a sequence of cutoff points in \( \omega \) space, one for each time period. We denote those cutoff points \( \{\omega^*_t\}_{t=0}^T \); each time period \( t \), the individuals with productivity above (below) \( \omega^*_t \) will (not) work. In terms of consumption, all individuals receive the same amount; this amount, which we denote \( \bar{c}_t \), depends on time since work effort, and hence total
CHAPTER 2

output, will change over time both due to productivity growth and changes in total hours worked.

How do we find the initial asset distribution that replicates this planner solution? Assuming that we have $T$ time periods, and restricting attention to $\bar{h} = 1$, the individual now solves the following problem:

$$\max \sum_{t=0}^{T} \beta^t \left[ \frac{c_{t}^{1-\sigma} - 1}{1-\sigma} - \pi_t(1) \frac{\psi}{1+\beta} \right],$$

subject to

$$\sum_{t=0}^{T} p_t c_{t}(\iota, \omega) = a_0(\iota, \omega) + \omega \sum_{t=0}^{T} p_t \gamma^t \pi_t(1),$$

where $\frac{p_t}{p_{t+1}} = 1 + r_{t+1}$. Thus, given a planning allocation, if every household is given initial assets according to

$$a_0(\iota, \omega) = \sum_{t=0}^{T} p_t \bar{c}_t - \omega \sum_{t=0}^{T} p_t \gamma^t \pi_t^{SP}(1; \omega),$$

where $\pi_t^{SP}(1; \omega)$ indicates the planner’s working choice for this particular individual, every individual will choose consumption and working such that the planner’s solution is decentralized. The interest rate for each period we can trivially find from the Euler equation of the individuals, which all look identical, since utility is additively separable in consumption and hours and they all consume the same amount.

We illustrate these points assuming that the individuals are distributed along the productivity dimension according to a truncated Pareto distribution with lower and upper bounds $\omega_l$ and $\omega_h$ and with shape parameter $\kappa$. Figure 16 shows two examples of resulting asset distribution as a function of individual productivity, one where we assume two periods ($t \in \{0, 1\}$), and one where we assume ten time periods.

![Initial asset distribution replicating utilitarian SP (2 periods)](image1)

(a) Initial asset distribution for 2 periods.

![Initial asset distribution replicating utilitarian SP (t=10)](image2)

(b) Initial distribution if $T = 10$.

**Figure 16**: The initial asset distribution needed to replicate the planner work allocations in a dynamic model with utilitarian planner weights.
How can we be sure that the individuals will actually choose the same work sequence as dictated by the planning allocation, given $a_0(i, \omega)$? The answer is found in how interest rates clear asset markets over time: they do so in such a way that there are both borrowers and lenders. This point will reappear in the next section, where we consider market equilibria with other asset structures.

**Finding 16.** The decentralization of the utilitarian optimum in a multi-period model implies an asset distribution where initial assets are inversely related to individual productivity levels, just like in the one-period model. In the case with a continuous productivity distribution and an extensive margin only, the initial asset distribution is a step function, with more steps the more periods there are.

Before moving on, let us point out that the above logic of how utilitarian optima are supported by competitive equilibria straightforwardly extends to the case when $\bar{h} < 1$, as in the one-period illustration earlier in Figure 15b.

### 3.4.3.2. A two-period model with lotteries

The implementation of the utilitarian optimum with a continuous $\omega$ distribution raised the question of how work decisions more generally are made over time in a market economy. The answer depends on the distribution of wealth. In an economy with complete markets—which in the case with an extensive margin means that lotteries must be allowed—one can, however, obtain some insights by looking at the planning economy. In an economy without lotteries, efficiency is not guaranteed and one has to solve for market equilibria directly.

The key focus for our study of the two-period economy will be on an important insight from studying static planning problems that we have so far not emphasized: work will be front-loaded. Recall from 3.4.1.3 that for general planner weights across individuals, the individual is assigned to work if, and only if, $\omega$ is high enough. Hence, as productivity grows (recall that its gross growth rate is $\gamma > 1$), to the extent the individual will transit between working and not working, it must be in the direction from working to not working. This means, from the individual choice perspective, that the agent chooses front-loading, i.e., endogenous retirement.

Front-loading is a general result in the complete-markets economy with BK preferences and productivity growth and the only remaining question is then what market mechanism ensures this outcome. The answer, on some level, must be that the interest rate adjusts appropriately, as a choice about when to work must involve the intertemporal price. This is what we will now illustrate in the context of the two-period model.

We assume that the gross interest rate is given by $1 + r$. To simplify the exposition we now only consider the extensive margin, in other words, we set $\bar{h} = 1$, but the results carry over to a situation with both an intensive and an extensive margin.

The agent’s maximization problem is given by

---

11Constrained efficiency may not materialize here, as there is a nontrivial (intertemporal) price determination and it could interact with the restriction on the individual’s choice.
\[
\max_{\{c_t, e_t\}_{t \in \{0,1\}}} \left[ \sum_{t=0}^{1} \beta^t \left( \frac{c_{t-1}^{\gamma} - 1}{1 - \sigma} - e_t \frac{\psi}{1 + \frac{1}{\tau}} \right) \right]
\]
subject to
\[
\sum_{t=0}^{1} \frac{c_t}{(1 + r)^t} = \omega \sum_{t=0}^{1} \left( \frac{\gamma}{1 + r} \right)^t e_t + a
\]
\[
e_t \in [0,1] \quad t \in \{0,1\}.
\]
Here, \(a\) is the agent’s initial wealth level.

The first-order conditions for \(c_t\) deliver
\[
c_0^{\gamma} = \mu \tag{54}
\]
\[
c_1 = c_0 (\beta (1 + r))^{\frac{1}{\sigma}} \tag{55},
\]
where \(\mu\) is the multiplier on the budget constraint.

Using the budget constraint and equation (55), we can write \(c_0\) as a function of \(\omega, a, e_0,\) and \(e_1:\)
\[
c_0 = \omega \left( e_0 + \frac{\gamma}{1 + r} e_1 \right) p^{-1} + a p^{-1},
\]
where \(p = 1 + \beta \frac{\gamma}{1 + r}^{1 - \sigma}\).

The first-order conditions for \(e_t\) yield (omitting the multipliers for the \(e_t \in [0,1]\) restriction)
\[
\frac{\psi}{1 + \frac{1}{\tau}} \cdot \frac{1}{\omega} = \mu \quad \text{for } e_0 \tag{56}
\]
\[
\frac{\psi}{1 + \frac{1}{\tau}} \cdot \frac{1}{\omega} = \mu \frac{\gamma}{\beta (1 + r)} \quad \text{for } e_1. \tag{57}
\]

These equations are key for the intertemporal work decision and, hence, for the interest-rate determination. Namely, it is clear that there cannot be interior solutions for both \(e_0\) and \(e_1\) at the same time, unless \(\gamma = \beta (1 + r)\). Hence, a household decides on the timing of the work depending on \(\gamma \leq \beta (1 + r)\). This insight, moreover, does not depend on the agent’s asset position or productivity level.

We thus have three cases and the solutions are straightforward in each of them:
\[
\frac{\gamma}{\beta (1 + r)} > 1 : \quad \text{The household back-loads working hours. If we have an interior solution for } e_0, \text{ the household will necessarily set } e_1 = 1. \quad \text{If we have an interior solution for } e_1, \text{ the household will necessarily set } e_0 = 0.
\]
γβ(1+r) < 1: The household front-loads working hours. The household will start reducing its fraction of individuals working in the second period (if the household is productive and/or rich enough so that it does not want to work full time both periods).

γβ(1+r) = 1: The household is indifferent. In this case, either both or none of the first-order conditions for e0 and e1 hold. Hence, the household either work full time in both periods, randomize in both periods, or does not work at all.

Hence, it must be, since we know from the planning solution with general weights, that work cannot be back-loaded, that \( \beta(1 + r) \geq \gamma \). Intuitively, the interest rate cannot be too small—relative to productivity growth—or else people postpone working (because working in the future produces more in present-value terms). It is instructive to go through the logic by which the case \( \beta(1 + r) < \gamma \) can be ruled out: the market for savings and loans would not clear. Section A.6 in the appendix goes through the detail.

If all households in the economy have the same productivity, the equilibrium interest rate will satisfy \( \frac{\gamma}{\beta(1+r)} = 1 \) if the productivity is high enough. However, it is also easy to show that there exist productivity levels for which we can have an equilibrium interest rate satisfying \( \frac{\gamma}{\beta(1+r)} < 1 \), even though all households have the same productivity, due to the constraints. E.g., for the case if all households have very low productivity, the equilibrium interest rate will be such that \( \frac{\gamma}{\beta(1+r)} < 1 \).12

Finding 17. In a multi-period model with zero initial asset holdings, it must be that interest rates are such that working decisions are delayed; in a two-period model the interest rate must be high: \( \gamma / (1 + r) \leq \beta \). This creates “endogenous retirement” in the economy: households endogenously choose to front-load work. In the special case of an equal asset distribution, higher-skilled individuals work less and withdraw faster from the labor force.

3.4.3.3. A two-period model without lotteries Before moving on, let us briefly discuss what outcomes are expected if convexification were not allowed. Will equilibria, in this case, have fundamentally different properties than when convexification is allowed? We now briefly comment on this issue in the context of our two-period model. We restrict attention to the special case where there is only an extensive-margin choice: \( h = 1 \).

When lotteries are no longer available, household behavior is harder to analyze in a multi-period model. Using the two-period model as an illustration, the utility of a household

\(^{12}\)For the intuition, think about a scenario where all agents are so low-productive they want to work full time in both periods. Then the interest rate must be high enough to support the autarky scenario.
with no initial assets can be rewritten as a function of productivity \((\omega)\) and working decisions in period 0 and 1 \((e_0 \in \{0, 1\}, \text{ and } e_1 \in \{0, 1\})\) using the lifetime budget constraint:

\[
\omega^{1-\sigma} \frac{p^\sigma}{1-\sigma} \left( e_0 + e_1 \frac{\gamma}{1+r} \right)^{1-\sigma} - \frac{\psi}{1+\frac{1}{\beta}} (e_0 + \beta e_1),
\]

where we still define \(p\) as

\[
p = 1 + \beta \frac{1}{1+r}.\]

With no initial assets, \(e_0 = e_1 = 0\) is not an option, so three cases remain: work the first, the second, or both periods. In the case with lotteries, the timing of work depended crucially on \(\gamma/\bar{\beta}(1+r)\): if it was low, the household preferred to front-load work, i.e., it preferred working in period 0 over working in period 1. Here, matters are less simple. There are, in general, three cutoff-values for \(\omega\) determining how much and when to work. The qualitatively new case is a cutoff level for productivity such that a household which is rather productive, and therefore only works in one period, chooses to switch from working in the first period to working in the second period instead. This level of productivity is defined by the solution to

\[
\omega^{1-\sigma} \frac{p^\sigma}{1-\sigma} - \frac{\psi}{1+\frac{1}{\beta}} = \omega^{1-\sigma} \frac{p^\sigma}{1-\sigma} \left( \frac{\gamma}{1+r} \right)^{1-\sigma} - \beta \frac{\psi}{1+\frac{1}{\beta}},
\]

which simply sets the utility value of the two options to be equal. The resulting value, \(\omega^*\), is given by:

\[
\omega^* = \left[ p^{-\sigma} (1-\sigma) \frac{(1-\beta) \frac{\psi}{1+\frac{1}{\beta}}}{1- \left( \frac{\gamma}{1+r} \right)^{1-\sigma}} \right]^{\frac{1}{1-\sigma}}.
\]

The key point here is that the timing of work depends not only on the intertemporal price but on individual productivity as well. If a household is productive enough to only need to work in one period, but still has \(\omega < \omega^*\), the household will prefer work in only the first period; if the household is very productive, i.e., \(\omega > \omega^*\), then working in the second period is preferable.

The intuition for this result derives from the fact that the nonconvex choice for hours translates into a non-convex consumption choice: as the timing of work is changed, the marginal utility of consumption changes discontinuously. In particular, the disutility from working in the second rather than the first period is simply a function of \(\beta\), independently of \(\omega\), since utility is linear in hours. The resulting change in the utility of consumption, however, depends on \(\omega\), as utility is nonlinear in consumption. So when a household is productive enough, postponing work gives higher utility, since the marginal loss in present-value income is worth less in utility terms than the discounting gain from postponing work.

In conclusion, more complex patterns for the timing of the individual’s work choice appear when convexification is not allowed. This has implications for intertemporal price determination too and it is possible, in particular, that \(\bar{\beta}(1+r) < \gamma\), which we showed was not possible in the lottery economy.
It also would be interesting to examine constrained optimality in the two-period model, since there is a nontrivially determined interest rate here and there can be an interaction between the individual’s non-convex choice and this price. Our conjecture, that we have not verified yet, is that the equilibrium will not always be constrained-optimal, but we leave it for future inquiry.

3.5. The long run

The present section looks at long-run work determination. We are again interested in who will work, and who should work, and the focus is on the case where income effects exceed substitution effects along a balanced path. In Boppart and Krusell (2016) it is shown that BK preferences, along with a standard neoclassical technology with labor-augmenting productivity growth at gross rate $\gamma$, deliver a balanced growth path where hours worked—modeled as an intensive-margin choice—fall over time at a constant rate $\gamma - \nu$, where again $\nu$ describes the extent to which the income effect exceeds the substitution effect. This finding presumes a representative agent, the absence of risk, and complete markets, and our purpose here is to explore heterogeneity, first in the context of complete markets, and then, in section 4, we will look at risk and incomplete markets. The extensive margin will be explored as well, both with complete and incomplete markets.

The maintained focus in the section will be on asymptotic features (levels and growth rates). There are also interesting implications for how persistent productivity growth affects individual and aggregate Frisch elasticities. Many of the results from the analysis of the static model can be used to look at these elasticities, however; we merely do not have them in focus in this section.

We begin by looking at who should work, how much, and when, based on a utilitarian planner’s objective. This analysis, which allows both an extensive and an intensive margin, computes a full transition path and characterizes its different stages, illustrating the different ways in which work patterns will change over time. We then move our focus to asymptotic features of the economy: can there be balanced growth, and if so what are its features? This answer to this question, which we address first from a planning perspective, turns out to depend crucially on the nature of the distribution of productivities among agents. We finally briefly comment on decentralization, i.e., the market equivalents of the planning outcomes.

3.5.1. Utilitarian optimum: an illustration

In the present section, we extend the horizon to gain insights into the transitional behavior of the economy. For this, we use the utilitarian planning problem and assume skill heterogeneity. As explained above, the planning problem is convenient since it can be studied purely as a sequence of static problems; the utilitarian case is an interesting normative benchmark.
In our example, individuals are distributed along the productivity dimension according to a truncated Pareto distribution with lower and upper bounds $\omega_l$ and $\omega_h$ and with shape parameter $\kappa$. Furthermore, we assume $h < 1$, i.e., the intensive margin is also in play. Then for a certain parameter setting Figure 17 illustrates the resulting development for who is working, the growth of consumption, the cutoff points, and the mass of individuals in each of the four mutually exclusive working modes: working full time, on the intensive margin, at $h$, or not at all.

The top left subgraph illustrates that the economy moves through “regimes” in terms of who is working. Each regime is a collection of the different active working modes of households in the population at that time. A number of such modes are possible and five modes are visited in the particular example displayed. The first regime (labeled -5) is that all households either work at the maximum number of hours (1) or at an interior point. After around 20 years, in the second regime (labeled -3), some households also work at the lowest amount $h$; they are in a corner solution. In the third regime (labeled 0), all the four modes of working coexist. In the last two regimes (1 and 2) the full-time and intensive-hour households withdraw, in order, so that in the last regime there are only households working at the minimum amount or not at all. At any point in time, of course, someone is working, since there is no capital accumulation and work is the only source
of consumption.

Because the economy goes through distinct regimes in this case, the path for aggregates, such as consumption growth in the top right panel of Figure 17, is somewhat uneven (but continuous), because the nature of the economy’s response to aggregate productivity change depends on people’s working modes. This is an illustration of what we saw above in section 3.3.2: there is no aggregation theorem in this class of economies and aggregate intra- as well as intertemporal substitution of labor, and hence consumption, depends nontrivially on the distribution of households.

In the utilitarian planning solution, again, the time at which a household makes its labor-force withdrawal is larger the higher is the household’s level of productivity. The bottom left panel of the figure thus illustrates how the household productivity cutoffs $\omega_1$ (between not working and working at $h$), $\omega_2$ (between $h$ and an interior solution), and $\omega_3$ (between an interior solution and the maximum) evolve over time. Because the aggregate wage path is monotonic in the example, there will not be any reversals, e.g., with households withdrawing and then coming back into the labor force later, which is reflected in the cutoff lines being monotonically increasing.

We also see that, in our example, the transition to the final regime, where a small fraction of the households—the most productive—work and provide output for the rest of the economy, occurs in about 100 years. This transition speed depends on the assumption of a utilitarian objective function. In a quantitative market equilibrium model, the distribution of households over initial asset holdings and productivity should be calibrated to the data. The data reveals, not surprisingly, that the most productive households also have high asset wealth. Translated into planning weights, this means higher weights on agents with high productivity (a case we looked at earlier). Such a formulation would change the nature of the transition path. A fraction of the productive households will withdraw early from the labor force, and this force will imply a slower transition of working patterns. This is because it is more difficult for the low-productivity households to provide consumption for the rest of the economy than it is for the high-productive ones, and in the interest of consumption smoothing this transition therefore becomes slower, given the same primitives.

**Finding 18.** *The utilitarian optimum allows consumption smoothing more easily since the most productive agents will work the longest and hence the transition to an economy where only a few households work can occur fast.*

### 3.5.2. Balanced growth in a planner’s solution

Under what conditions can we have a balanced growth path in this economy? Since aggregate working hours will move toward zero and the extensive margin will be active, the asymptotic behavior does not appear balanced, but we will show that a form of balanced growth can be shown to occur both in the planner’s solution and in a market. The planning problem is easier to analyze and this is where we begin.
3.5.2.1. Finite productivity support  In the static planning problem above we saw that since preferences do not allow for aggregation, aggregate labor supply depends on the joint distribution of planner weights and labor efficiency units. Hence, in general, the dynamics of the economy will depend on the distribution. In this section we examine the asymptotic dynamics of the planner’s solution. Note that the asymptotic behavior of the economy is also of interest when it comes to solving numerically for the decentralized equilibria, which in part motivates this analysis. So we ask whether the heterogeneity “washes out” in the long run or whether the asymptotic dynamics of aggregate prices and quantities even asymptotically depend on the distribution.

As we will see below the answer to that question depends on the distribution of planner weights and labor efficiency units. To prepare the analysis, consider once again the “representative-agent” economy without capital accumulation where all individuals have the same \( \omega = \omega \) and where we use utilitarian planner weights. In this case, the planner sets \( h(i) = 1 \) if \( \omega \lambda < \psi^{1-\sigma} \), an interior solution \( h(i) = (\omega \lambda) - \frac{\psi^{1-\sigma}}{1-\psi} \psi^{1-\sigma} \) if \( \psi^{-1} \leq \omega \lambda < (\psi h^{1+\sigma})^{-\frac{1}{\sigma}} \), and a fraction we define as \( s \) of the population working at \( h \) is set according to

\[
\frac{1}{\psi} \left( 1 + \frac{1}{\psi} h^{-(\sigma+\frac{1}{\sigma})} \right)^{\frac{1}{\sigma}}
\]

if \( \omega \lambda \geq (\psi h^{1+\sigma})^{-\frac{1}{\sigma}} \). Since \( \lambda \) goes to infinity, asymptotically, we will therefore be in the regime where only the extensive margin is active. Hence, asymptotically, as productivity grows at gross rate \( \gamma > 1 \), labor supply will shrink at gross rate \( \gamma^{1-\sigma} \) (because the fraction \( s \) that works \( h \) shrinks at this rate) and consumption of all individual will grow at gross rate \( \gamma^{\frac{1}{\sigma}} \).

Finding 19. In the case where all households are identical and have identical planner weights, the economy will converge to a balanced growth path where the only working agents work at \( h \) and the size of this group goes to zero with the growth rate \( \gamma^{1-\sigma} \); consumption per capita, on the other hand grows at rate \( \gamma^{\frac{1}{\sigma}} \).

We will now show that for a large class of distributions of planner weights and labor efficiency units the economy will converge asymptotically to the same point as just discussed, i.e., to representative-agent behavior. The key feature is that distribution of the ratio \( \frac{\omega}{\chi(i, \omega)} \) is bounded from above. The reason is that in this case, the planner would like to make everyone work less than \( h \) in finite time. Consequently, we know that asymptotically only the extensive margin is changing and we can describe the asymptotic growth rates of the economy in the following proposition, which is proved in the appendix, section A.2.

\[\text{\textsuperscript{13}}\text{This conclusion is true for the MaCurdy preferences we have assumed throughout this section. In the appendix section A.5 we show that these preferences are robust in terms of aggregate observables.}\]
Proposition 1. Suppose $\chi(i, \omega)$ is strictly positive for all individuals and we have a finite $\omega/\chi(i, \omega)$ for all individuals, and we have $\sigma > 1$ and $\gamma > 1$. Then, asymptotically all the individuals either work 0 or $h$. The share working, $s$, is asymptotically zero and is asymptotically shrinking at constant rate, i.e.,

$$\lim_{t \to \infty} \frac{s_{t+1}}{s_t} = \gamma^{-\frac{\sigma}{\sigma-1}} < 1.$$  \(58\)

Consumption of all the individuals and the total labor input in efficiency units, $n$, is growing asymptotically at the constant rate

$$\lim_{t \to \infty} \frac{n_{t+1}}{n_t} = \lim_{t \to \infty} \frac{C_{t+1}}{C_t} = \lim_{t \to \infty} \frac{c_{t+1}(i, \omega)}{c_t(i, \omega)} = \gamma^{\frac{1}{\sigma}} > 1.$$  \(59\)

So as long as the distribution of $\omega/\chi(i, \omega)$ is bounded, the asymptotic dynamics of the aggregate economy look identical to the one of a representative agent economy.

We illustrate the result just obtained with a truncated Pareto distribution in $\omega$ and a utilitarian welfare function. I.e., we suppose that the density in $\omega_i$ is given by the following truncated function:

$$\Gamma(\omega)d\omega = \frac{\omega_i^\kappa}{\omega^{\kappa+1}} \left( \frac{1}{1 - (\omega_0/\bar{\omega})^{\kappa}} \right),$$  \(60\)

for $\omega_0 \leq \omega \leq \bar{\omega}$ (and zero otherwise). We have $\kappa > 0$. (With $\kappa \to -1$, this would nest a uniform distribution.) Figure 17 shows the resulting behavior where it can be seen how the consumption growth rate and the $\omega_3$ threshold converge to their asymptotic level.

The asymptotic growth rates we constructed here can also be reached in finite time provided some conditions hold. This insight is closely related to one that will be used in the section below where we look at market economies. So consider a very simple case: the representative-agent economy. In this case we will reach a productivity level satisfying the following condition:

$$\bar{\omega} \lambda \geq \left( \psi h^{\frac{1}{\sigma+1}} \right)^{-\frac{1}{\sigma-1}}$$

in finite time and, at that point in time and forever after, all the households work either $h$ or 0. Furthermore, it can be shown that the fraction working, $s$, decreases at constant gross growth rate $\frac{s_{t+1}}{s_t} = \gamma^{-\frac{\sigma}{\sigma-1}} < 1$ and consequently income, the marginal utility of consumption, and so on all also grow at constant rates.

Similar results obtain for much more elaborate endowment distributions: when $\lambda$ is high enough, all agents will be either at $h$ or 0 hours of work. To obtain a balanced growth path where some agents are always at interior solutions does, however, relies on an unbounded distribution either of productivities (as in the previous example with a truncated Pareto distribution) or relative planner weights. For distributions that do not have the finite-support property for $\omega/\chi(i, \omega)$, the long-run outcome can thus be qualitatively different and even the long-run growth rates depend on the distributional properties. We now illustrate this point in the next section.
3.5.2.2. Infinite productivity support  We now consider an unbounded \( \omega \chi(\iota,\omega) \) by means of the following example. So consider a planner with utilitarian weights. We assume (an untruncated) Pareto distribution of household productivity levels \( \omega \) such that the marginal distribution in \( \omega \) can be expressed as

\[
\Gamma(\omega) d\omega = \kappa \frac{\omega^\kappa}{\omega^{\kappa+1}},
\]

for \( \omega \geq \omega_0 \) (and zero otherwise). We have \( \kappa > 1 \). The following proposition now obtains.

**Proposition 2.** A utilitarian planning solution given that the individual productivity distribution is Pareto, with the parametric form just described, leads to exact balanced growth where the fraction of people working grows at gross rate \( \gamma^{1+\sigma(\kappa-1)} < 1 \) and per-capita consumption grows at \( \gamma^{1+\sigma(\kappa-1)} > 1 \).

This is a striking example since it shows how in general even the long-run (asymptotic) growth rates of aggregate variables can depend on the distribution of productivities across the household distribution. I.e., this model is an endogenous-growth model. Similarly, planner weights too will matter in general. It also shows that even along an exact balanced growth path the intensive margin of labor supply can change. These results require unbounded support, and the unboundedness is of course unreasonable in a literal sense since it is really about the relative wage distribution. However, the behavior of an economy with very large finite support will be very similar for a very long period of time even if its asymptotic behavior will be qualitatively different.

3.5.3. Market equilibrium

Are there balanced-growth market equilibria for our economy in case of an extensive-margin labor choice? The section above looking at this issue for general planner weights of course provides an answer because of the welfare theorems, which hold in this environment provided lotteries are allowed. We also saw, however, that qualitatively different balanced paths are possible depending on the distribution of productivities and welfare weights, and what is less trivial is precisely to find a mapping between this distribution and the productivity-asset distribution. Another issue is whether balanced growth—with declining aggregate hours—can be made consistent with the production technology in a neoclassical model with capital accumulation. This issue is easier to address because the answer is affirmative and exactly follows the analysis in Boppart and Krusell (2016).

In section A.7 in the appendix, we show that if the productivity-asset distribution is bounded appropriately, then there is a balanced growth path where lotteries are used among a subset of the population (defined by assets and productivity) and where the remainder of the population does not work at all. These probabilities, moreover, go to zero
at the rate $\gamma - \nu = \gamma - \frac{\sigma}{1 - \sigma}$. Given that lotteries are used, we can think of the utility function as displaying an infinite Frisch elasticity. On the transition toward the balanced path, the regime where hours lotteries represent the only working mode will be reached in finite time. Before this time, individual’s Frisch elasticities will not all be 0 or infinity and, hence, the economy’s response to shocks will change in nature as the economy transits toward its long-run state.

It is important to note that a balanced path where lotteries are used can also be supported without lotteries. We now make this point formally; the analysis can be viewed as an extension to Rogerson (2006), who look KPR preferences and do not consider growth.

### 3.5.3.1. The balanced-growth path decentralized without lotteries

Let us consider the household’s problem on a balanced path based on a distribution with finite productivity-asset support. It reads:

$$\max_{\{c_t, h_t\}} \sum_{t=0}^{\infty} \beta^t \left[ \frac{c_t^{1-\sigma}}{1-\sigma} - \psi h_t \right]$$

subject to

$$\sum_{t=0}^{\infty} \frac{c_t}{(1+r)^t} = a_0 + \sum_{t=0}^{\infty} \frac{w_t h_t}{(1+r)^t},$$

where $w_t = w_0 \gamma^t$, and the constraint that $h_t \in \{0, 1\}$ for all $t$.

The standard Euler equation implies that consumption will grow by $(\beta(1+r))^{\frac{1}{\sigma}} \equiv g_c$. We take as given from the previous section that $1 + r = \gamma / \beta$, hence, $g_c = \gamma^{\frac{1}{\sigma}}$. We can then write the left-hand side of the budget constraint as

$$\sum_{t=0}^{\infty} \frac{c_t}{(1+r)^t} = c_0 \sum_{t=0}^{\infty} \left( \frac{g_c}{1+r} \right)^t = c_0 \sum_{t=0}^{\infty} \left( \frac{\beta \gamma^{1-\sigma}}{1+r} \right)^t.$$

The right-hand side of the budget constraint can be written as

$$a_0 + \sum_{t=0}^{\infty} \frac{w_t h_t}{(1+r)^t} = a_0 + w_0 \sum_{t=0}^{\infty} h_t \left( \frac{\gamma}{1+r} \right)^t.$$

We are not sure which periods the household will work, but let us define the following:

$$\sum_{t=0}^{\infty} h_t \left( \frac{\gamma}{1+r} \right)^t = \sum_{t=0}^{\infty} h_t \beta^t = \lambda,$$

with $\lambda \in [0, \frac{1}{1-\beta}]$, where the lower limit corresponds to never working and the upper limit represents working in every period for the rest of eternity. Thus, we have recast the problem in terms of the total remaining time, appropriately discounted, that the household chooses to work. It is also clear that the household will be indifferent as to the timing here, so long as the total satisfies the chosen $\lambda$. 

---

117
We can now write the problem as follows:

$$\max_{c_0, \lambda} \left\{ \sum_{t=0}^{\infty} \beta^t \left( \frac{c_0 \gamma^t}{1 - \sigma} \right)^{1-\sigma} - \psi \lambda \right\} =$$

$$\max_{c_0, \lambda} \left\{ \frac{c_0^{1-\sigma}}{1-\sigma} \cdot \sum_{t=0}^{\infty} \left( \beta \gamma^{1-\sigma} \right)^t - \psi \lambda \right\} =$$

$$\max_{c_0, \lambda} \left\{ \frac{c_0^{1-\sigma}}{(1-\sigma)(1 - \beta \gamma^{1-\sigma})} - \psi \lambda \right\}$$

subject to

$$\frac{c_0}{1 - \beta \gamma^{1-\sigma}} = a_0 + w_0 \lambda.$$ 

Inserting the budget constraint into the maximization problem gives a static maximization problem

$$\max_{\lambda} \left\{ \frac{(a_0 + w_0 \lambda)^{1-\sigma}}{(1-\sigma)(1 - \beta \gamma^{1-\sigma})^\sigma} - \psi \lambda \right\}.$$

The first-order condition with respect to $\lambda$ reads

$$\frac{w_0 (a_0 + w_0 \lambda)^{-\sigma}}{x} - \psi,$$

where we define $x = (1 - \beta \gamma^{1-\sigma})^\sigma$. We then see that if assets are large enough, there is no interior solution to the problem, but the household will choose to work zero periods, i.e., $\lambda = 0$. If the assets are low enough, the household will have to work every period, i.e., $\lambda = \frac{1}{1-\beta}$.

If assets are somewhere in between, the interior solution for how much to work is given by

$$\lambda = \frac{1}{w_0} \left( \frac{\psi x}{w_0} - \frac{1}{\sigma} - a_0 \right).$$

4. **Incomplete markets**

We finally turn to the incomplete markets model in the presence of growth. Unlike in the previous models where we assumed complete markets, the asset distribution is now an endogenous outcome of the model. Moreover, in the incomplete-markets model households can benefit from working longer hours when their productivity is high, and less
hours when their productivity is low, but they cannot do this as effectively here as there is a lower bound on asset holdings. The incomplete insurance, moreover, implies that interest rate is lower than under complete markets, making richer households—who are relatively well insured—less willing to save. As a result of these features, in the event of a sequence of many bad labor productivity shocks, households are forced to work longer hours despite their current low productivity, since they already have drawn down their assets and are borrowing-constrained. This is an argument for average productivity in the economy with incomplete markets to be lower than in the corresponding complete markets economy.\footnote{See Pijoan-Mas (2006) for a full discussion about this topic.} On the other hand, total hours may be higher if the insurance motive for working is strong. All these issues will be discussed below.

We first formulate the model in the presence of growth, and with intensive-margin labor choice only. Thereafter, we turn to the case of extensive-margin labor choice. After that, we present our quantitative results and interpret them in light of the complete-markets analysis.

4.1. The intensive margin

We begin by defining a balanced-growth equilibrium in levels and then show how it can be transformed, focusing first on the case with an intensive-margin choice. Growth is labor-augmenting at rate $\gamma$. This means, given the more general preference class in Boppart and Krusell (2016), that consumption can grow at a different rate than labor productivity; this discrepancy is regulated by the parameter $\nu$ (i.e., the KPR formulation is the special case obtaining for $\nu = 0$).

**Definition 2.** A balanced-growth equilibrium consists of growth rates $g$ and $g_h$, prices $r_k$ and $w_t$, a value function $V_t$, decision rules $f_t^k$ and $f_t^h$, and distributions $\Gamma_t$ such that, for all $t$,

1. $g = \gamma g_h = \gamma^{1-v}$.
2. $V_t(a, \omega) = \max_{a', h} u(a(1-\delta + r_k) + h\omega w_t - a', h) + \beta E[V_{t+1}(a', \omega')|c]$ s.t. $a' \geq ag^{t+1}, h \in [0, \infty)$. Notice, here, that the borrowing constraint changes over time (unless $a = 0$) and gets more and more stringent with $a < 0$.
3. $f_t^a(a, \omega)$ and $f_t^h(a, \omega)$ solve the maximization problem on the right-hand side of the dynamic-programming problem above for all $(a, \omega)$.
4. $r_k$ and $w$ satisfy $r_k = F_1( k_t, \gamma^t h_t)$ and $w_t = \gamma^t F_2( k_t, \gamma^t h_t)$, where $k_t \equiv \sum_\omega \int_a a f_t^k(a, \omega) \Gamma_t(da, \omega)$ and $h_t \equiv \sum_\omega \int_a \omega f_t^h(a, \omega) \Gamma_t(da, \omega)$.
5. $\Gamma_{t+1}(B, \omega) = \sum_\omega \pi_\omega |\omega| \int_{a: f_t^a(a, \omega) \in B} \Gamma_t(da, \omega)$ for all Borel sets $B$ and for all $\omega$. 


6. \( f^t(a^t, \omega) = g^t f^0_g(a, \omega) \) and \( f^h_t(a^h_t, \omega) = g^h_t f^0_h(a, \omega) \), and \( \Gamma_t(Bg^t, \omega) = \Gamma_0(B, \omega) \) for all \( a, B, \text{ and } \omega \).

Note that due to growth, the distribution over \( a \) will not be stationary. However, as we will show below, once \( a \) is detrended by the appropriate growth rate we obtain a stationary distribution.

The level-based definition just defined can be stated in stationary form as follows:

**Claim 1.** The balanced-growth equilibrium defined above is equivalent to a stationary equilibrium defined by prices \( r_k \) and \( w \), a value function \( V \), decision rules \( f^a \) and \( f^h \), and a distribution \( \Gamma \) such that

1. \( V \) solves
   \[
   V(a, \omega) = \max_{a', \omega} u(a(1 - \delta + r_k) + h\omega w - a'g^t_h, h) + \beta g^{1-\sigma} E[V(a', \omega') | \omega]
   \]
   s.t. \( a' \geq a, h \in [0, \infty) \).

2. \( f^a_t(a, \omega) \) and \( f^h_t(a, \omega) \) solve the maximization problem on the right-hand side of the dynamic-programming problem above for all \( (a, \omega) \).

3. \( r_k = F_1(\tilde{k}, \tilde{h}) \) and \( w = F_2(\tilde{k}, \tilde{h}) \), where \( \tilde{k} \equiv \sum_{\omega} \int_a a\Gamma(da, \omega) \) and \( \tilde{h} \equiv \sum_{\omega} \int_a \omega f^h_t(a, \omega) \Gamma(da, \omega) \).

4. \( \Gamma(B, \omega) = \sum_{\omega} \pi_{\omega} |\tilde{\omega}| \int_{a: f^a(a, \omega) \in B} \Gamma(da, \omega) \) for all Borel sets \( B \) and for all \( \omega \).

We now prove the claim. Observe that the notation used in the proof differs from the notation used in the claim.

**Proof.** Using the last condition of the balanced-growth equilibrium, note that in the third condition we can write \( \tilde{k}_t = (\sum_{\omega} \int_a a\Gamma_0(da, \frac{d\tilde{a}}{\tilde{a}^t}, \omega)) \), which is equivalent to \( \tilde{k}_t \equiv \frac{\tilde{k}_t}{\tilde{g}_h} = (\sum_{\omega} \int_a \tilde{a}\Gamma_0(da, \omega)) \), where we have defined \( \tilde{a} = \frac{a}{g^t} \). Notice also that \( \tilde{k}_t = \tilde{k} \), i.e., a constant, in a balanced-growth equilibrium.

Similarly, we obtain \( \tilde{h}_t = \sum_{\omega} \int_a \omega g^{t_h} f^0_h(a, \omega) \Gamma_0(da, \frac{d\tilde{a}}{\tilde{a}^t}, \omega) \), implying that

\[
\tilde{h}_t = \frac{\tilde{h}_t}{\tilde{g}_h} = \sum_{\omega} \int_{\tilde{a}} \omega f^h_0(\tilde{a}, \omega) \Gamma_0(da, \tilde{a}, \omega),
\]

which also is constant under balanced growth: \( \tilde{h}_t = \tilde{h} \).

Given \( \tilde{g} = \gamma \tilde{g}_h \) and that \( F_1 \) and \( F_2 \) are both homogeneous of degree 0, we now see that the two firm first-order conditions can be expressed as

\[
r_k = F_1(\tilde{k}, \tilde{h}) \quad \text{and} \quad w_0 = F_2(\tilde{k}, \tilde{h}).
\] (62)
Turning to the fourth equilibrium condition, using the (very) last condition stating that the distribution is (in an appropriate sense) constant on the balanced growth path, we obtain

\[ \Gamma_0(B/g^{t+1}, \omega) = \sum_{\omega} \pi_{\omega|\bar{\omega}} \int_{\bar{a}:f_0^{\bar{a}}(\tilde{a},\bar{\omega})g^t \in B} \Gamma_0(d\bar{a}, \bar{\omega}), \]

where we used the definition of \( \bar{a} \). Defining \( \bar{B} = B/g^t \) for any Borel set \( B \), we obtain

\[ \Gamma_0(\bar{B}, \omega) = \sum_{\omega} \pi_{\omega|\bar{\omega}} \int_{\bar{a}:g^{t+1}_{B}(\tilde{a},\bar{\omega}) \in \bar{B}} \Gamma_0(d\bar{a}, \bar{\omega}), \]

which can equivalently be stated as

\[ \Gamma_0(\bar{B}, \omega) = \sum_{\omega} \pi_{\omega|\bar{\omega}} \int_{\bar{a}:g^{t+1}_{B}((\tilde{a},\bar{\omega}) \in \bar{B}} \Gamma_0(d\bar{a}, \bar{\omega}). \quad (63) \]

Looking at consumer optimization under balanced growth, finally, we obtain (after using the same kinds of definitions as above),

\[ V_t(\tilde{a}(1-\delta + r_k)g^t, \omega) = \max_{\tilde{a}^t, \tilde{h}} u(\tilde{a}g^t + \tilde{h}g^t_{\omega} \omega w_0 \gamma^t - \tilde{k}'g^{t+1}, \tilde{h}g^t_{\omega}) + \beta E[V_{t+1}(\tilde{a}'g^{t+1}, \omega')|\omega] \]

s.t. \( \tilde{a}'g^{t+1} \geq \tilde{a}g^{t+1}, \tilde{h}g^t_{\omega} \in [0, \infty) \).

Now consider our instantaneous utility function as formulated in equation (1) for \( u \) and let \( g^t_{\omega} = \gamma^{-v} \) and \( g = \gamma^{1-v} \). Then \( g^{t+1} - c \) can be factorized out from \( u \). Dividing both sides of the equation by this quantity and defining \( V_t(\tilde{a}g^t, \omega) \equiv g^{t+1} \tilde{V}(\tilde{a}, \omega) \), we can write

\[ \tilde{V}(\tilde{a}, \omega) = \max_{\tilde{a}^t, \tilde{h}} u(\tilde{a}(1-\delta + r_k) + \tilde{h}w_0 \gamma^t - \tilde{k}'g, \tilde{h}) + \beta g^{t+1} E[\tilde{V}(\tilde{a}', \omega')|\omega] \quad (64) \]

s.t. \( \tilde{a}' \geq \tilde{a}, \tilde{h} \in [0, \infty) \), with associated policy functions \( \tilde{f}_t^a(\tilde{a}, \omega) \) and \( \tilde{f}_t^h(\tilde{a}, \omega) \).

Now \( r_k, w_0, \tilde{V}, \tilde{f}_t^a, \tilde{f}_t^h \), and \( \Gamma_0 \), determined by equations (62), (63), and (64), define a stationary equilibrium. \( \square \)

Four items differ compared to the formulation of our quantitative baseline model in definition 1 in section 2.1. First, we restrict attention to balanced paths here. Second, the discount factor in the consumer’s problem is multiplied by \( g^{1-c} \). Third, an additional gross “cost” of saving, \( g \), appears in the consumer’s budget. Fourth, and finally, \( g \) appears in the updating on the right-hand side of the equation determining the stationary distribution.
4.2. The extensive margin: version I

When the labor choice is in a two-point set \( \{0, h\} \), the equilibrium definition needs to be altered only slightly: \( f^h(a, \omega) \) is now an indicator function, taking on the value 1 if the agent chooses to work and 0 if the agent chooses not to work. For convenience, we restrict attention here to utility functions \( u \) that are additively separable in consumption and leisure (the MaCurdy family).

A complication that arises when there is an extensive-margin choice is that in the case where \( \nu > 0 \)—the case where ongoing labor-productivity growth would call for lower and lower hours worked—workers cannot choose to work less and less per time unit, since the set of available hours choices only contains 0 and \( h \). Instead, they would work more and more rarely, as detailed above in the complete-markets formulation in section 4.3: there would be a continually decreasing participation rate. Individual choice would entail an \( f^h \) that would be equal to 1 on an increasingly small (relative) part of its domain for cash on hand \( \omega \) as wages keep rising; when the worker works one period there is instead a very large addition to assets. These features may or may not be consistent with an exact balanced-growth equilibrium; we discuss this issue further in section 4.4.6 below. Because of this challenge, let us consider a slightly different model formulation.

Consider a two-point set for hours worked that changes with productivity: if productivity grows, the working option involves less hours (in the case with stronger income effects than for KPR). This formulation is motivated by what workers would like, given their preferences: it would allow working regularly while at the same time enjoying more and more leisure. In particular, assume that the labor choice set is \( \{0, h_{\gamma - \nu t}\} \). The implied setup is consistent with a (transformed) stationary equilibrium.

**Claim 2.** The balanced-growth equilibrium with a labor choice set \( \{0, h_{\gamma - \nu t}\} \) is equivalent to a stationary equilibrium defined by prices \( r \) and \( w \), a value function \( V \), decision rules \( f^a \) and \( f^h \), and a distribution \( \Gamma \) such that

1. \( V \) solves
   \[
   V(a, \omega) = \max_{a', h} u(a(1 - \delta + r) + hh\omega - a'g, h) + \beta g^{1 - \sigma} E[V(a', \omega')|\epsilon] \tag{65}
   \]
   s.t. \( a' \geq g, h \in \{0, 1\} \).

2. \( f^a(a, \omega) \) and \( f^h(a, \omega) \) solve the maximization problem on the right-hand side of the dynamic-programming problem for all \( (a, \omega) \).

3. \( r \) and \( w \) satisfy \( r = F_1(\bar{k}, \bar{h}) \) and \( w = F_2(\bar{k}, \bar{h}) \), where \( \bar{k} \equiv (\sum \omega \int_a a \Gamma(da, \omega)) \) and \( \bar{h} \equiv \sum \omega \int_a \omega h f^h(a, \omega) \Gamma(da, \omega) \).

4. \( \Gamma(B, \omega) = \sum \pi(\omega) \int_{a: g f^a(a, \omega) \in B} \Gamma(da, \omega) \) for all Borel sets \( B \) and for all \( \omega \).

The proof of this claim follows the proof of the previous claim line by line.
4.3. The extensive margin: version II

Let us briefly return to the model with an extensive margin that does not change with productivity. The purpose here is to discuss our conjecture as to the asymptotic behavior of this economy. We have not, as of yet, managed to verify this conjecture and have to leave it for future work.

First, let us simply note that the kind of transformation used for the economy with an intensive margin does not work if $\bar{h}$ does not shrink over time. Second, we noted that individuals will want to work less and less in this economy but that it is not obvious whether an exact balanced growth path exists. Under complete markets, as seen in section 4.3, such a balanced path exists, and it is associated with a total commitment to work going forward that is shrinking at a constant rate. It is also, however, associated with indifference as to the timing of the work, and under incomplete markets shocks and the inability to fully transfer resources over time will in general prevent this.

One possibility is that the asymptotic path has all agents, or all agents except a vanishingly small set of agents, effectively fully insured due to individual capital accumulation. However, what is key in this economy is assets relative to an appropriate transformation of wages rather than the absolute asset level and if people keep withdrawing from the labor force, wages will grow faster than will output, by implication, and assets. This would contradict effectively full insurance.

Given incomplete insurance we know, from the individual’s Euler equation, that the real interest rate will be somewhat depressed. In particular, if an individual would save in a permanent-income manner, thus consuming the return and saving the rest, assets would grow less fast than the growth rate of consumption and output. As a result, assets would decline relative to consumption so long as the consumer does not work. On the other hand, any consumer who works in the current period would, in the limit, accumulate an enormous amount of assets. Thus, the domain for assets will contain the whole real line and not lend itself easily to a transformation: to the right of the asset level at which the consumer is indifferent between working or not assets will jump further and further to the right; to the left they will shrink.

4.4. Results

We begin with the calibration and then outline how we use numerical methods to find equilibria; these methods go beyond what is used in the literature in that they also offer insights into convergence to balanced growth paths. We then go over the nature of the policy rules, with maintained emphasis on comparisons with the static/complete-markets models discussed earlier. After this, we discuss the aggregate results, first focusing on interest rates and aggregate hours worked and then on various measures of inequality as well as efficiency.
CHAPTER 2

4.4.1. Calibration

For the purpose of this section we again use the MaCurdy utility function:

\[ u(c, h) = \frac{c^{1-\sigma} - 1}{1 - \sigma} - \psi \frac{h^{1+\frac{1}{\sigma}}}{1 + \frac{1}{\sigma}} \]  

We will both use the KPR version, \( \sigma = 1 \), and a BK version, where we set \( \sigma = 1.7 \), a calibration that is also entertained in Boppart and Krusell (2016); with this value follows a value for \( \theta \) that is equal to 1.5, ensuring that hours fall at the rate observed in data in a cross-section of countries, a rate that corresponds to \( \gamma^{-\nu} \) in the model, where \( \nu = (\sigma - 1)/(\sigma + \frac{1}{\sigma}) \). Different values of \( \psi \) are then considered, including \( \psi = 0 \) (no valued leisure), delivering different levels of aggregate hours. We examine both a model with an intensive margin only and a model with an extensive margin only.\(^{15}\)

For all models we set the yearly growth rate \( \gamma = 1.01 \) and we use a(n annual) discount factor \( \beta = 0.96 \). The rest of the model is parameterized in a standard way, with idiosyncratic productivity shocks following an AR(1) process in logs with persistence 0.9 and conditional standard deviation of 0.1744, discretized into a 7-state Markov chain. The firm side produces with Cobb-Douglas technology with capital share of 1/3. Depreciation, \( \delta \), is assumed to be 10%.

4.4.2. Finding a steady state: the use of turnpikes

The existence of steady states in the Aiyagari model has been established in the literature; see, e.g., Acemoglu and Jensen (2015). There are results on uniqueness—see, e.g., Light (2018)—but there are also examples of multiplicity of steady states. We know of no proofs of global convergence to steady state for this class of models. For this reason, caution has to be exercised when trying to find a steady state with numerical methods.

We follow two procedures here. One is standard: the model is transformed into stationary form—by use of the growth rates that we conjecture will characterize a balanced path; see the previous sections—and then an iterative procedure is used to find the steady state. This iteration is particularly simple here because it only involves one equilibrium price: the real interest rate.\(^{16}\) So one guesses on the interest rate, solves the dynamic programming problem of the household given this interest rate, which implies choices for hours and capital savings, and then finds the implied stationary distribution of these variables given the stochastic process assumed. That stationary distribution in turn gives the economy-wide capital-hours ratio, which returns the real interest rate. We find this iterative procedure to be fast and stable.

\(^{15}\)The case with both an intensive and an extensive margin would also be interesting to study; here, the two extreme cases are focused on for ease of comparison.

\(^{16}\)The interest rate maps uniquely into the aggregate wage rate, given a constant-returns-to-scale production function.
The second procedure is to instead solve for a long transition path for the untransformed model. That is, one fixes a final-period interest rate—arbitrarily chosen—and then solves for a path of equilibrium interest rates. This numerical task is more challenging, because it involves guessing and iterating on a whole sequence of interest rates. However, we find that also this procedure is stable and fast. We use 200 time periods and our end-period guess is an interest rate that is a steady state for the equivalent economy without further growth. I.e., we know that the guess is wrong, and in some cases the guess is quite far from the steady state sought. The key now is that once an equilibrium is reached here, one can assess whether it appears to converge to a steady state, and whether that steady state coincides with the steady state found using the first procedure.

We find that this "turnpike" approach works very well and strongly indicate global convergence. Figure 18 shows the resulting interest rate path and average hours worked for the KPR model with labor choice on the intensive margin. As can be seen, the results confirm the results from the transformed stationary equilibrium: in the long run, the real interest rate as well as hours worked settle down to constant after roughly 20 years and then remain at those constants until about 10 years from the endpoint. I.e., the economy gets on the turnpike rather quickly, stays on it for a long time, and then exits at the end, just like a car would.

Figure 18: Results from a turnpike model, KPR preferences and intensive margin labor choice.

Figure 19: Results from a turnpike model, BK preferences and intensive margin labor choice.
Figure 20: Growth of capital and labor measured in efficiency units. Results from the turnpike model with intensive margin labor supply.

Figure 19 shows the corresponding figures for the model with BK preferences. Here, as can be seen, average hours are falling toward zero at a constant rate in the long run, as expected.

Figure 20 illustrates the resulting growth rates for capital and labor, measured in efficiency units from the turnpike model (for KPR preferences and BK preferences, respectively). As can be seen, with KPR preferences, the growth rate converges to $\gamma$, while in the model with BK preferences, the growth rate converges to $\gamma^{1-\nu}$.

Very similar results are found for the extensive-margin versions of the model. We omit them for the sake of brevity.

In sum, our balanced growth paths are straightforwardly computed and there are strong indications of global convergence to these paths. The long-run growth rates coincide with those of the corresponding complete-markets economies.

4.4.3. Policy rules

We now discuss the numerically computed policy rules for our economies and compare them to the results under complete markets. We begin by showing the decision rules in the stationary state and then show how growth in wages affect these graphs. Throughout we focus on labor supply, i.e., we omit consumption decision rules, as hours worked are our focus here.

First, Figure 21 shows policy functions for labor from the stationary growth model with an intensive-margin labor choice. The figure illustrates two important features. One is rather obvious: for a given productivity level (on a given colored line in the graph), higher assets lowers labor supply. The second is a feature of intertemporal substitution in the
incomplete-markets economy: the higher the productivity, the more the agent chooses to work, given a fixed asset level. That is, the agent takes the opportunity work when the wage is high and, at that point, saves a large portion of the earnings for the future.

Second, and relatedly, it is instructive to look at Figure 22, which repeats exactly the same information as Figure 1 and Figure 2 in the static model. In this figure, we contrast hours isoquants in two stylized static models: one with KPR preferences and one with BK preferences. For the household, a productivity increase is perceived as an increase in wages. In the (static) KPR model we see that, given a level of positive assets, an increase in wages always leads to higher optimal hours. However, in the case of BK preferences, increased productivity can actually lead to lower optimal hours. As for the static model, the reason is that, when the wage is low relative to assets, the substitution effect dominates, and we find an effect similar to that in the KPR case, whereas when the wage is high relative to assets, the income effect will dominate under BK and, hence, hours fall when productivity rises. We thus see, in our figure, that the non-monotonicity appears sooner for low asset levels.

Third, we now turn to how decisions change over time when wages increase. So consider Figure 23, which results from the intensive-margin model. For three points in time, it shows the point in the asset/productivity space where the household chooses to work exactly 0.5. As can be seen from the graph in the left panel of the figure, in the model with KPR preferences, the isohours line shifts out monotonically to the right. I.e., as aggregate wages grow, it takes higher and higher asset levels to reproduce the same hours choice: wages induce higher work, and the income effect of higher assets are needed to balance the wage increase.

17To facilitate intuition, we omit the case of negative wealth here.
Figure 22: Hours choice. Illustration of the combination of assets (x-axis) and productivity (y-axis) that yield the same hours choice. Each line represents an hours isoquant.

Figure 23: Isohours lines at three points in time in a model with intensive margin labor choice. Results from a turnpike model. The line for each year indicates where the household chooses to work exactly 0.5. The y-axis depicts indices on the discretized productivity grid (note that the corresponding productivity levels are not equidistant).

With BK preferences, depicted in the graph on the right-hand panel of the figure, the same monotonicity is not present. At low asset levels, the indifference curve is moving left over time, as aggregate wages increase, but at high levels it is moving right. This feature, to which we will return when we discuss the extensive margin in the BK model, derives from the features uncovered in the static model and just discussed above, namely the backward-bending nature of isohours curves in asset-productivity space under BK preferences. At low asset levels, higher wages, or productivity, will make hours worked decrease in the BK model: here, earnings are high relative to other wealth, and the income effect dominates the substitution effect. Hence, for a given, low asset level (say, 5) in the graph, the green-dotted isohours curve (where wages are the highest) is above the other lines, indicating a smaller area of high work effort. At high asset levels, in contrast, the substitution effect dominates and higher wages increase the area of high work effort.
The exact same patterns as just discussed emerge when we consider the extensive-margin models. The policy functions for work/leisure at three points in time for the model with KPR preferences are shown in Figure 24a. If a household is in the south-east region of this graph, the household has a relatively low productivity and is rich enough to be able to afford leisure and postpone the working decision. As the figure shows, the breakpoints are shifting out over time: as the economy grows, the household needs more assets to consider itself rich enough to enjoy leisure, given a fixed productivity level.

Figure 24b shows the same type of information as Figure 24a, but for the model with BK preferences and an upper bound $\overline{h}$ for the feasible hours choice set that falls at the appropriate rate ($\gamma^{-\nu}$). As can be seen, as for the KPR case, the breakpoints in asset space for the indifference between hours and leisure move monotonically. However, this monotonic movement is due to the fact that the interpretation of the indifference curves changes as wages grow: they depict indifference between not working and working at a level of hours that falls over time.$^{18}$

In the next section, we turn to the equilibrium determination of interest rate and aggregate hours worked. There, we will only look at variables transformed by their respective growth rates, i.e., all variables will be stationary. Interest rates do not have a trend, so do not need to be transformed, but in this model hours need to be transformed to the extent preferences depart from KPR. Hence, aggregate hours will be reported relative to a trend that equals $\gamma^{-\nu t}$ (where $\nu = 0$ is the KPR case). Hence, a high value should be reported as high relative to this trend.

### 4.4.4. Interest rates and aggregate hours

Table 1 summarizes our main results for the aggregates. As is well known, the interest rate in the incomplete-markets model is lower than in the complete-markets counterpart,$^{129}$

---

$^{18}$As we shall see below, monotonicity appears to fail if $\overline{h}$ is fixed.
due to the precautionary savings motive. This is illustrated by the interest rates in the first rows of the table, corresponding to the complete-markets case, being significantly higher than the values in the rows below, which all report results for incomplete-markets economies.

The model without valued leisure ($\psi = 0$) is represented by numbers in the second row. Here, the numbers in parenthesis refer to hours worked and we see that they are 100%. We see that the interest rate in equilibrium is very similar in the KPR and BK economies; although the higher consumption curvature in utility for the BK case induces higher precautionary saving, the effect of the increased capital stock on the interest rate is small.

We then introduce valued leisure ($\psi = 1$) and an intensive-margin labor choice: $h \in [0, \infty)$. As can be seen in the table, the interest rate now falls further, and rather significantly. There are, in principle, two effects here. With an active labor-supply channel, the de-facto insurance is higher, and households with a bad past string of productivity realizations can improve their asset position, and protect against further bad luck, by working harder. One might then expect that the effects of the frictions would be lower, and hence that interest rates would be higher. They are, however, lower. It turns out, namely, that another channel dominates, and it precisely builds on the fact that households now are able to intertemporally substitute and work more in times of high productivity and work less in times of low productivity. The mechanism works as follows: households now shift more assets from good periods of high productivity, because they now work harder too, than under fixed hours. On average, this increases the amount of capital households carry, because of the asymmetry given by the lower bound on assets: they increase assets more in good times than they decrease it in bad times. As a result of these mechanisms, there is a higher aggregate capital stock under valued leisure the interest rate has to fall in order to clear the capital market.

We see that the interest rate, as expected is lower under BK than under KPR, since there is a higher consumption curvature/risk aversion in the former case, inducing more saving. We also see that hours rise somewhat as under BK preferences relative to KPR, but the rise is small: from 0.89 to 0.90. We interpret this as an effect of higher saving under higher consumption curvature/risk aversion, implying lower consumption and hence higher hours worked.

Next, we turn to the case of a labor supply choice on the extensive margin: $h \in \{0, 1\}$. We contrast two models, one with a lower disutility of work than the other ($\psi = 0.8$ vs. $\psi = 1$). Under the former, i.e., the lower disutility of work, the KPR preference case

---

19Note that in the complete-markets case the interest rate is increasing in $\sigma$ for the case of no valued leisure or intensive margin labor choice. However, in the case of extensive margin, the interest rate in the complete-market model with KPR preferences and BK preferences coincide. In a complete markets economy with preferences given by (66), the long run interest rate is given by $(1 + r) = \gamma^\theta / \beta$, with $g$, the growth rate of consumption, given by $g = \gamma^{1-\nu}$ (with $\nu$ given by (3)). In the case of KPR preferences, $\nu = 0$. In the case of BK preferences and an extensive margin labor choice, we have effectively a $\theta = \infty$, which gives $g = \gamma^{-1/\sigma}$. Hence, the interest rate in a complete-markets setting with extensive margin labor choice with KPR preferences and BK preferences coincide.

20Keeping consumption and wages fixed, the direct effect of an increase in $\sigma$ on hours, i.e., going from KPR to BK, is negative, and hence goes the other way.
gives a labor force participation of 89%. The interest rate in this model is 3.86%, which is higher than in the intensive margin model. However, when the disutility is increased to 1.0 (still maintaining KPR preferences), labor force participation falls to 78% and the interest rate falls to 3.30%. The reason is that with lower labor force participation, there are more periods when the household has no income, and therefore the need to shift assets to extended periods of no working increases. Thus, the interest rate needs to fall to clear the capital market.

In the extensive margin labor supply choice model with BK preferences, we observe the same the pattern, but generally the interest rate is slightly lower than in the KPR case, again due to the higher risk aversion. In this model, which is the “version I” case of the extensive margin (where $\bar{h}$ is falling over time), the labor force participation does not fall when productivity grows: instead, the amount of hours conditional on working falls.

### 4.4.5. Inequality

Table 2 displays information about various inequality measures in the different models considered. We look at Gini coefficients for asset inequality, consumption inequality

<table>
<thead>
<tr>
<th></th>
<th>KPR preferences</th>
<th>BK preferences</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\sigma = 1.0$</td>
<td>$\sigma = 1.7$</td>
</tr>
<tr>
<td><strong>Complete markets</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>No valued leisure</td>
<td>5.21%</td>
<td>5.94%</td>
</tr>
<tr>
<td>Valued leisure, $h \in [0, \infty)$</td>
<td>5.21%</td>
<td>5.41%</td>
</tr>
<tr>
<td>Valued leisure, $h \in {0, h\gamma^{-vt}}$</td>
<td>5.21%</td>
<td>5.21%</td>
</tr>
<tr>
<td><strong>Incomplete markets</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>No valued leisure</td>
<td>4.43% (100%)</td>
<td>4.41% (100%)</td>
</tr>
<tr>
<td>Valued leisure, $h \in [0, \infty)$</td>
<td>$\psi = 1.0$</td>
<td>3.77% (0.89)</td>
</tr>
<tr>
<td>Valued leisure, $h \in {0, h\gamma^{-vt}}$</td>
<td>$\psi = 0.8$</td>
<td>3.86% (89%)</td>
</tr>
<tr>
<td>Valued leisure, $h \in {0, h\gamma^{-vt}}$</td>
<td>$\psi = 1.0$</td>
<td>3.30% (78%)</td>
</tr>
</tbody>
</table>

Table 1: Equilibrium interest rates and aggregate hours (in parentheses) for stationary growth models. Aggregate hours are expressed as average hours worked for the intensive margin labor supply models, and as labor force participation for the extensive margin labor supply models. For the model with BK preferences and intensive margin labor choice, hours are reported relative to trend.
Chapter 2

(within parentheses), inequality in hours worked (within brackets), and inequality for earnings (within double brackets).

Before commenting on how different specific models compare, let us make some overall observations about the extent of inequality in our benchmark models with an active labor-supply channel. We see that wealth inequality is high, not as high as in the data (it is around 0.8), but it is much higher than it is in the baseline model (Aiyagari 1994). This is noteworthy, given the discussion in Hubmer et al. (2016), which various extensions to the baseline model aimed at raising wealth inequality: heterogeneity in discount factors, heterogeneity in returns, and superstar earnings processes. Apparently, the model with an active labor-supply channel we study here helps significantly in making wealth inequality match data better. We also observe that earnings inequality is high, though not as high as in the data (in the data it is a little over 0.4). Again, an active labor-supply channel helps: it makes hours go up when the wage is high, reflecting intertemporal substitution. Consumption inequality has the lowest Gini, reflecting a strong need to smooth over time and across states of nature.

In terms of model comparisons, the table shows, first of all, that wealth inequality is higher as a result of the active labor-supply channel: the associated Gini goes up by 7 percentage points (both for KPR and BK preferences) when we compare $\psi = 0$ to $\psi = 1$. The interpretation here is that intertemporal substitution makes hours comove with wages, hence increasing earnings inequality. As a result, savings rise overall, but not sufficiently to outweigh the direct earnings-based effect. We also see that whether we look at a model with an intensive or an extensive margin makes only a small difference for wealth Gini.

Overall, consumption inequality is much lower than wealth inequality. This is a well known, but of course important, point: wealth inequality in large part is a result of consumption smoothing and hence higher wealth inequality helps keep consumption inequality low. As for how consumption inequality differs across models, the most striking feature is that BK preferences give lower Gini. This is, again, not surprising, since higher consumption curvature means that it is costlier to accept higher consumption variation, and hence consumers make sure to decrease it. Consumption inequality differs only slightly depending on whether labor supply is active or not.

Inequality in hours worked is also very low, of course reflecting the fact that it is costly to vary hours: $\theta = 1.5$ is a high enough Frisch elasticity to prevent much movement in hours, given the calibrated process for individual wages. We also note, conditional on an extensive-margin choice, that hours inequality increases significantly if the cost of working rises ($\psi$ goes from 0.8 to 1): fewer people now work and, hence, inequality rises.

Let us, finally, turn to issues touching on the efficient allocation of work across people of different productivities. The full correlation table can be found in the appendix, section A.8. In 3.4.1.1, the point was made that under a utilitarian planner’s objective function, i.e., when all agents utilities receive the same weight, a striking conclusion emerges

---

21The key feature of the superstar earnings process is not only that extremely high productivity can occur but also that extremely large drops in productivity can occur from the highest productivity levels.
under complete markets: the high-productivity agents should be made to work harder (or, simply, work, in a model with an extensive margin only). The utilitarian benchmark, moreover, is commonly used in the incomplete-markets literature since it captures the notion that households are ex-ante identical. Does a strongly positive hours-productivity correlation emerge here as well? For KPR, a strong positive correlation is borne out: it is 0.53 for the intensive-margin model and about half that for the extensive-margin model. For the BK model, however, the correlation is positive but closer to zero; one reason for this is the stronger income effect, making higher wages also lower hours worked. We also note that assets and earnings are positively correlated.

Finally, in Figure 15, we used the static model to illustrate what type of asset distribution would be needed to decentralize a utilitarian social planner solution in the case with extensive margin labor supply: the higher productivity of the agent, the lower the agent’s assets. In the incomplete-markets model, a strong negative correlation is not borne out, of course—in line with the data. The model rather predicts a positive correlation here—in stark contrast to the utilitarian social planner solution.

Table 2: Inequality. Resulting Gini coefficient for wealth, consumption (in parentheses), hours [in brackets], and earnings [[in double-brackets]].
Finally, let us return to version II of the extensive-margin BK model. In section 4.2 we assumed that the labor choice was in a two-point set \( \{ 0, h^{\gamma^{-v_t}} \} \). In other words, households could keep on working with the same regularity as productivity grows, but enjoy more leisure since every time they work, they work less hours, thus mimicking the intensive-margin model. In the present section we briefly continue the discussion begun in section 4.3 regarding the challenging case of a two-point set \( \{ 0, h \} \) for all \( t \). Here, the way to enjoy more leisure is to work less periods, since every period households work, they will still work full time.

In the case of KPR preferences, we saw that the policy functions for when to work and when to enjoy leisure shifted out over time (Figure 24a). In that model, when income and substitution effects cancel, the breakpoints shift out monotonically by productivity state: regardless of productivity level, the asset level needed to be rich enough to enjoy leisure is increasing. However, in the case of BK preferences, and therefore falling number of work periods, the picture appears to become more complicated due to non-monotonic behavior. Figure 25 illustrates the lack of monotonicity: here, we display decision rules based on aggregate wages growing over time—not equilibrium wages exactly, since we have not yet been able to solve the model. As can be seen, as for the intensive choice, the breakpoints in asset space for the indifference between hours and leisure move to the left, as wages grow, if assets are low, but to the right, if assets are high. This is intuitive and, again, follows the logic emphasized in the static analysis: when assets are low enough, raised wages lead to lower desired work hours, since the income effect dominates when the wage is high relative to assets. This qualitative feature flips over as assets increase and wage changes mainly generate substitution effects.
The monotonicity is difficult to manage numerically and it also seems that the entire asset domain will expand as wages grow: for low assets, working becomes more and more rare as wages grow, and hence assets fall further in the absence of earnings income. At the same time higher and higher asset holdings will also materialize when assets are high enough. Our conjecture, still, is that there will be an asymptotic steady state with a stable interest rate. Preliminary calculations indicate that this interest rate will be quite low, but at the very least the details are still quite open.

5. Concluding comments

The ultimate goal behind the development of macroeconomic models is of course to use them for understanding data and as laboratories for asking policy questions. In this paper, however, we instead focus on a narrower task: model comparison, i.e., we compare the framework we develop here with those typically used in modern macroeconomic analysis. Model comparisons are not goals in themselves, but they are an important part of producing more robust and long-lasting insights in the work toward the ultimate goal. Thus, we view our goal here as an intermediate one.

Although we relate informally to microeconomic data in constructing our theoretical frameworks here, we do not systematically evaluate these but rather intend to carry out this task in future work. We look forward to this undertaking: although many views flourish—as to whether people with high wages work harder than those with low wages (many of our colleagues claim they do), wealth is positively correlated with hours (many of our colleagues claim they do), and so on—the fact is that the correlations in the data are not so strong, and the model framework we propose here appears rich enough to at least potentially capture much of what we observe.

The present paper also does not accomplish all of the intermediate tasks that we set out to accomplish in this project. One of the most important next steps is to study aggregate fluctuations explicitly and, then, to complete the comparison between Frisch elasticities in the static models and those in the incomplete-markets model. We plan to accomplish this task using “MIT shocks”, i.e., the methods put forth in Boppart et al. (2018), which are really the same methods as those used to compute the transition paths for the incomplete-markets model here. Thus, no new methods are needed but much systematic study of aggregate shocks remains.

Another important step is to conduct explicit analysis of constrained-efficient allocations, along the lines of Davila et al. (2012): the case of an active labor-supply channel is interesting and has not been in focus. It also becomes increasingly important as technology progresses and we wonder about the labor-force participation of different groups: will the incomplete-markets economy have good efficiency properties or are significant policy interventions called for? The who-should-work question is high on our agenda.
CHAPTER 2

References


CHAPTER 2


CHAPTER 2


A. Appendix

A.1. Household Frisch elasticity with negative assets

Figure 26 shows the individual household’s Frisch elasticity as function of productivity and different parameters of the utility function. For a detailed description of the underlying calculations, please see 3.2.1.2.

A.2. Proofs

Proof of Proposition 1. With productivity growth (i.e., $\gamma > 1$) and dominating income effects (i.e., $\sigma > 1$) the planner would like to decrease hours worked by all individuals. Then, under the assumption that $\lambda^{(i, \omega)}$ is finite, households either work zero or $\bar{h}$ in finite time and the share working $\bar{h}$ is asymptotically zero. An easy way to see this is to consider the planner’s problem without the $\bar{h}$ constraint. In this case the first-order conditions are

$$\mu_1 = \lambda^{(i, \omega)} c_1(i, \omega)^{-\sigma}$$

(67)
and
\[ \gamma^t \omega_t = \chi(i, \omega) \psi \left( h^t(i, \omega)^* \right)^{\frac{1}{\theta}} \]  
(68)

where \( \mu \) is the Lagrangian multiplier and \( h^t(i, \omega)^* \) the optimal hours choice while ignoring the \( h \) constraint. From this we see that in a given point in time consumption grows at the same rate for all individuals and also the desired hours choice \( h^t(i, \omega)^* \) will shrink at the same rate for all individuals (at a given point in time). Finally note that under the assumption of a finite \( \frac{\omega}{\chi(i, \omega)} \) for all individuals, \( h^t(i, \omega)^* \) converges to zero asymptotically. Hence, any positive \( h \) starts eventually binding.

Now let us analyze the regime when all the \( h^t(i, \omega)^* \) are below \( h \). Under this regime we know that individuals with \( \frac{\omega}{\chi(i, \omega)} \geq \frac{\psi h^t}{\mu_t \gamma^t (1 + \frac{1}{\theta})} \) will be working \( h \) and all the other individuals do not work. Let us call the maximum of \( \frac{\omega}{\chi(i, \omega)} \) in the population as \( (\omega / \chi)^{\text{max}} \). Then asymptotically we must have

\[ \lim_{t \to \infty} \mu_t \gamma^t = \frac{\psi h^t}{(\omega / \chi)^{\text{max}} (1 + \frac{1}{\theta})}. \]  
(69)

and \( s \) converges to zero. Why does \( \mu_t \gamma^t \) have to converge to this level? First, note that \( \frac{\psi h^t}{(\omega / \chi)^{\text{max}} (1 + \frac{1}{\theta})} \) can be ruled out since it would imply no labor supply and consequently no consumption. Second, suppose \( \mu_t \gamma^t \) stabilizes at a level strictly smaller than \( \frac{\psi h^t}{(\omega / \chi)^{\text{max}} (1 + \frac{1}{\theta})} \). Then, labor supply would be strictly positive and constant. This however cannot be an equilibrium since it would imply that all consumption levels grow at gross rate \( \gamma \) and we get a contradiction with a constant \( \mu_t \gamma^t \).

Since \( \mu_t \gamma^t \) stabilizes asymptotically we must have \( 1 = \frac{\mu_t + 1}{\mu_t} \gamma = \left( \frac{C_t + 1}{C_t} \right)^{-\sigma} \gamma = \gamma^{1 - \sigma} \left( \frac{s_t + 1}{s_t} \right)^{-\sigma} \),

where \( s_t \) is the share still working and we used the resource constraint for the last equality. Solving this equation for the growth rates in consumption and the participation rate finally proves the statements in the proposition. ■

Proof of Proposition 2. From the static planning analysis we know that all the households with \( \omega > \frac{\psi}{\mu_t \gamma^t} \) work 1, household with \( \frac{\psi}{\mu_t \gamma^t} > \omega > \frac{\psi h^t}{\mu_t \gamma^t} \) work \( \left( \frac{\mu_t \gamma^t \omega}{\psi} \right)^{\theta} \) hours, households with \( \frac{\psi h^t}{\mu_t \gamma^t} > \omega \geq \frac{\psi h^t}{\mu_t \gamma^t (1 + \frac{1}{\theta})} \) work \( h \) hours, and households with \( \omega < \frac{\psi h^t}{\mu_t \gamma^t (1 + \frac{1}{\theta})} \) do not work at all. Then, aggregate consumption and output is given by

\[ C_t = \mu_t^{-\frac{1}{\theta}} = \gamma^t \left[ \int_{\frac{\psi h^t}{\mu_t \gamma^t (1 + \frac{1}{\theta})}}^{\frac{\psi h^t}{\mu_t \gamma^t}} \frac{\kappa \omega^K}{\omega^K h^t} \omega^h d\omega + \int_{\frac{\psi h^t}{\mu_t \gamma^t}}^{\frac{\psi}{\mu_t \gamma^t}} \frac{\kappa \omega^K}{\omega^K - \theta} \left( \frac{\mu_t \gamma^t}{\psi} \right)^{\theta} d\omega + \int_{\frac{\psi}{\mu_t \gamma^t}}^{\infty} \frac{\kappa \omega^K}{\omega^K} d\omega \right]. \]  
(70)
Solving the integrals allows us to express the equation as

\[ \mu_t^{-1} \psi = \mu_t^{\gamma} \kappa \omega_0^\gamma \kappa \omega_0^\gamma \psi \left( \frac{h}{\theta} \right)^{1+\frac{1}{\theta}} - \left( \frac{1+\frac{1}{\theta}}{\kappa - 1} \right) \left( 1 - \frac{1-h^\theta}{1-\kappa + \theta} + \frac{1}{\kappa - 1} \right) \].

(71)

Note that the only endogenous variable in this equation is \( \mu_t \) and we consequently have

\[ \frac{\mu_{t+1}}{\mu_t} = \gamma \frac{1-\kappa}{1+\sigma(\kappa-1)} < 1. \]

We then obtain, for consumption growth,

\[ \frac{c_{t+1}}{c_t} = \frac{c_{t+1}(\omega)}{c_t(\omega)} = \gamma \frac{1+\kappa}{1+\sigma(\kappa-1)} > 1. \]

(72)

Finally, note that the labor market participation rate is given by 1 minus the cdf evaluated at \( \frac{\psi h^{\frac{1}{\theta}}}{\mu \gamma^\theta (1+\theta)} \). Under the Pareto distribution this gives us for the participation rate

\[ s_t = \omega_0^\kappa \left( \frac{\mu_t \gamma^\theta (1+\frac{1}{\theta})}{\psi h^{\frac{1}{\theta}}} \right)^\kappa. \]

(73)

Consequently, the participation rate decreases at gross rate

\[ \frac{s_{t+1}}{s_t} = \gamma \frac{1-\kappa}{1+\sigma(\kappa-1)} < 1. \]

(74)

But note that since we have a fat tail in the \( \omega \) distribution some household will always be constraint at \( h = 1 \) and some households are in an interior solution and decrease their hours along the intensive margin. Hence, even asymptotically the intensive margin will matter.

A.3. Welfare effects of the introduction non-convexities in the hours choice

The fact that is is impossible to supply \( 0 < h < h \) hours implies that—if this constraint is binding—it generates a higher disutility of working for a family in order to earn the same amount of labor income.

Suppose we have a family that chooses to have a lottery which makes a fraction \( \pi \) work \( h \). Then the optimality condition implies

\[ c^{-\sigma} \omega \lambda = \psi \frac{h^{1+\frac{1}{\theta}}}{1+\frac{1}{\theta}}. \]

Note that the total disutility from working is then \( \pi \psi \frac{h^{1+\frac{1}{\theta}}}{1+\frac{1}{\theta}} \) which is strictly larger than the disutility from earning the same amount without the constraint which is \( \psi \left( \frac{\pi h^{1+\frac{1}{\theta}}}{1+\frac{1}{\theta}} \right). \)

Hence, total disutility to earn the same amount is with the constraint higher by a factor
π⁻¹ > 1. As π goes to zero this ratio goes to infinity simply because disutility from working drops faster to zero without constrain compared to with constraint. As π approaches one the gap shrinks to zero.

Now let us express this welfare loss in terms of consumption equivalent. Solving the intratemporal optimality condition for \( c \) gives

\[
c = \left( \psi - \frac{h^{1 + \frac{1}{\theta}}}{(1 + \frac{1}{\theta})\omega \lambda} \right)^{-1/\sigma}.
\]

Hence we can as what consumption scalar \( z > 1 \) makes the family indifferent between

\[
\frac{1}{1 - \sigma} \left( \psi - \frac{h^{1 + \frac{1}{\theta}}}{(1 + \frac{1}{\theta})\omega \lambda} \right)^{-(1-\sigma)/\sigma} z^{1-\sigma} - \pi \psi \frac{h^{1 + \frac{1}{\theta}}}{1 + \frac{1}{\theta}} = \frac{1}{1 - \sigma} \left( \psi - \frac{h^{1 + \frac{1}{\theta}}}{(1 + \frac{1}{\theta})\omega \lambda} \right)^{-(1-\sigma)/\sigma} \psi \frac{(\pi h)^{1 + \frac{1}{\theta}}}{1 + \frac{1}{\theta}}.
\]

Solving this equation for \( z \) gives:

\[
z = \left[ 1 - (\sigma - 1)(\omega \lambda)^{-(1-\sigma)/\sigma} (\pi h)^{\frac{1}{\theta}} \right]^{-\frac{1}{\sigma-1}} > 1.
\]

This is non-monotone in \( \pi \). Furthermore it is monotonically increasing in \( \omega \lambda \). It is also monotonically increasing in \( \psi \) and \( h \).

**A.4. Welfare to supply a fixed amount of labor services**

Consider the following planner that minimizes welfare cost of supplying \( \bar{L} \) units of labor services. This planner solves

\[
\min_{\{h(i)\}} \int \chi \psi \frac{h(i)^{1 + \frac{1}{\theta}}}{1 + \frac{1}{\theta}} \Gamma(di)
\]

subject to

\[
\int \omega h(i) \Gamma(di) = \bar{L}
\]

The first-order conditions can be written as

\[
\chi \psi h(i)^{\frac{1}{\theta}} = \mu \omega
\]
where we call the multiplier $\mu$. Integrating over all households and using the constraint gives

$$\int \chi \psi h(i)\left(1 + \frac{1}{\theta} \Gamma\right) = \mu \bar{L}. \quad (78)$$

Furthermore we can solve for the individual labor supply and plug it into the objective to get

$$\int (\chi \psi)^{-\theta} \omega^{1+\theta} \Gamma\left(\frac{\psi}{\omega}\right) = \mu \theta \bar{L}. \quad (79)$$

Hence, we can express the marginal cost of supplying an additional unit of labor service as

$$\left[\int \frac{(\chi \psi)^{-\theta} \omega^{1+\theta} \Gamma\left(\frac{\psi}{\omega}\right)}{\chi \psi} - \theta \omega\right] - \frac{1}{\theta} \bar{\chi} \bar{\psi} \bar{\omega}^{-1} \left(1 + \frac{1}{\theta}\right) \bar{L} = \mu. \quad (80)$$

where $\bar{\chi}, \bar{\psi},$ and $\bar{\omega}$ are averages.

A.5. Non-separable preferences

In our benchmark setting, we use MaCurdy (1981)’s formulation of the utility function, which features additive separability in consumption and hours. This formulation is convenient because the marginal utility of consumption does not depend on whether or not the household works. However, there are many other non-additive utility functions where the income effect is stronger than the substitution effect. In Boppart and Krusell (2016), MaCurdy’s formulation is a special case but there are other, interesting non-additive cases as well. We now discuss these briefly.

We focus on a homogeneous-agent case for simplicity, i.e., a version of Rogerson (1988), though not the special KPR case he considers. A fundamental difference in the non-additive case with lotteries is that consumption behavior of the working and non-working households are qualitatively different. A point in case is that where lotteries are needed, though we stress—for the same reason as explained in Section 3.4.2.4—any ex-post lottery equilibrium outcome can be generated as a deterministic equilibrium with appropriately chosen initial wealth levels of households. Thus consider a social planner who solves

$$\max_{c_0, c_1, \pi(1)} \pi(1)u(c_1, 1) + (1 - \pi(1))u(c_0, 0) \quad (81)$$

subject to $c_0 \geq 0, c_1 \geq 0, \pi(1) \in [0, 1], and$

$$\pi c_1 + (1 - \pi(1))c_0 = \lambda \pi(1). \quad (82)$$

It is straightforward to solve this problem for a given function $u$ in the desired class and then to examine the result of changing $\lambda$. We will, in particular, take a parametric class of $u$ functions and find the asymptotic (gross) growth rates of the choice variables: $g_\pi, g_1,$ and $g_0$, with the obvious notation. Recall that in the MaCurdy benchmark in this paper, these rates were—in the homogeneous-agent case—given by $\gamma_\pi, \gamma_1,$ and $\gamma_0$, respectively: the consumption growth rates of working and non-working people are the same.
CHAPTER 2

<table>
<thead>
<tr>
<th>Growth rate</th>
<th>Positive</th>
<th>Sign of $(1-\sigma)(1+pb\kappa)$</th>
<th>Negative</th>
<th>-0</th>
</tr>
</thead>
<tbody>
<tr>
<td>$g_\pi$</td>
<td>$\gamma^{1-\sigma}$</td>
<td>$\gamma^{\frac{x}{\sigma}} = \gamma^{\frac{(1-\sigma)pb\kappa}{\sigma}}$</td>
<td>$\gamma^{\frac{1-\sigma}{\sigma}} = \gamma^{\frac{-x}{\sigma}}$</td>
<td>$\gamma^{\frac{1-\sigma}{\sigma}} = \gamma^{\frac{-\theta}{\sigma}}$</td>
</tr>
<tr>
<td>$g_0$</td>
<td>$\gamma^{\frac{1}{\sigma}}$</td>
<td>$\gamma^{\frac{1-\sigma}{\sigma}} = \gamma^{\frac{(1-\sigma)pb\kappa}{\sigma}}$</td>
<td>$\gamma^{\frac{1}{\sigma}} = \gamma^{\frac{1}{\sigma}}$</td>
<td>$\gamma^{\frac{1}{\sigma}} = \gamma^{\frac{1}{\sigma}}$</td>
</tr>
<tr>
<td>$g_1$</td>
<td>0</td>
<td>$\gamma$</td>
<td>$\gamma$</td>
<td>$\gamma^{\frac{1}{1+pb\kappa}} = \gamma^{\frac{x}{1-\sigma}}$</td>
</tr>
</tbody>
</table>

Table 3: Asymptotic growth rates

For the utility function, we use the following parametric form:

$$u(c, h) = \frac{c^{1-\sigma} \left[ 1 - a \left( hc^{\frac{\nu}{1-\nu}} \right)^b \right]^d - 1}{1 - \sigma},$$

(83)

where we have

$$a = \frac{\phi(1-\sigma)}{1 + \frac{1}{\theta'}}, \quad b = 1 + \frac{1}{\theta'}, \quad \text{and} \quad d = (1-\sigma)\kappa. \quad (84)$$

This form nests many well-known cases. We assume throughout that $\sigma > 1$ and $\kappa < 0$ and define $p \equiv \frac{\nu}{1-\nu}$ and $x = (1-\sigma)pb\kappa$ to save on notation. We will simply report the results here; for details, see our online appendix. It turns out that the asymptotic behavior is qualitatively different depending on the sign of $(1-\sigma)(1+pb\kappa)$, so that there are three distinct cases, as shown in Table 3. We see from the table that the benchmark case we consider is knife-edge in that $g_0 = g_1$. However, it is robust—in the sense that there is a large part of the parameter space with this property—in the characterization of $g_\pi$ and $g_0$. For the working group, however, we see in the table that the asymptotic consumption growth rate can take on several values depending on parameters. This, however, does not have aggregate consequences in the case where $(1-\sigma)(1+pb\kappa) \geq 0$ because then the asymptotic share of total consumption going to this vanishing group is zero. We also see, by examining the middle case, $(1-\sigma)(1+pb\kappa) < 0$, that it is possible to have the vanishing group see its consumption growth rate rise at the pace of productivity, i.e., very fast, with a stable total share of overall production.

In sum, although qualitatively different cases are uncovered in the non-additive case, the MaCurdy formulation used as a benchmark in the paper is robust in terms of aggregate observables.

---

22For the MaCurdy case, we have $(1-\sigma)(1+pb\kappa) = 0$ and $v = \frac{1-\sigma}{\sigma-\frac{1}{\theta'}}$. Plugging this in gives us $p = \frac{\sigma-1}{1+\frac{1}{\theta'}}$. We get:

$$g_1 = \gamma^{\frac{1}{1+pb\kappa}} = \gamma^{\frac{1}{1+\left(\frac{x}{1-\sigma}\right)^{\frac{1}{\theta'}}}} = \gamma^{\frac{1}{\theta'}} = g_0$$

Equivalently, we could use $g_1 = \gamma^{\frac{x}{1-\sigma}}$ and that MaCurdy requires $\kappa = \frac{1}{1-\theta'}$ and concluded the same.
Productivity Choice of \((e_0, e_1)\) Is the hh borrowing or saving?

| \(\omega < \omega_B^a\) | (1,1) | Borrowing |
| \(\omega_B^a < \omega < \omega_B^b\) | \((x, 1)\) | Borrowing |
| \(\omega_B^b < \omega < \omega_B^c\) | \((0, 1)\) | Borrowing |
| \(\omega_B^c < \omega\) | \((0, x)\) | Borrowing |

Table 4: Choice of \(e_1\) in each period depending on productivity, when \(\frac{\gamma}{\beta(1+r)} > 1\). Choice \(x\) denotes an interior solution, i.e. \(e_1 \in (0, 1)\), for the period.

A.6. Intertemporal market clearing in the two-period market economy

To keep the problem analytically tractable we only consider how the household’s solution vary as a function of productivity, and assume \(a = 0\) for the remainder of this section. Note that this implies that a household will for sure work at least part of a period, not working at all is not a possible outcome.

We will now go through the three potential cases one by one to point out the most important insights:

**Back-loading of work:** \(\frac{\gamma}{\beta(1+r)} > 1\)  If \(\frac{\gamma}{\beta(1+r)} > 1\), we see that if we have an interior solution for \(e_0\), the household will necessarily set \(e_1 = 1\). Hence, we can structure the solution to the household problem along the productivity dimension according to Table 4 and calculate three breakpoints:

\(\omega_B^a\) When the household is in the interior solution for \(e_0\) (i.e., equation (56) holds with equality), and chooses \(e_0 = 1\)

\(\omega_B^b\) When the household is in the interior solution for \(e_0\) (i.e. equation (56) holds with equality), and chooses \(e_0 = 0\)

\(\omega_B^c\) When the household is in the interior solution for \(e_1\) (i.e. equation (57) holds with equality), and chooses \(e_1 = 1\)

After some algebra we get the following expressions for the breakpoints:

\[
\omega_B^a = \left(\frac{\psi}{1 + \frac{1}{\beta}}\right)^{\frac{1}{1-\sigma}} \left(\frac{1 + \frac{\gamma}{1+r}}{p}\right)^{\frac{\sigma}{1-\sigma}}
\]

\[
\omega_B^b = \left(\frac{\psi}{1 + \frac{1}{\beta}}\right)^{\frac{1}{1-\sigma}} \left(\frac{\gamma}{1+r}\right)^{\frac{\sigma}{1-\sigma}}
\]

\[
\omega_B^c = \left(\frac{\psi}{1 + \frac{1}{\gamma}}\right)^{\frac{1}{1-\sigma}} \left(\frac{\gamma}{1+r}\right)^{\frac{\sigma}{1-\sigma}} \left(\frac{\gamma}{\beta(1+r)}\right)^{\frac{1}{\sigma-1}}
\]
Households with very high productivity, \( \omega > \omega^B_b \), are obviously borrowers, since they produce nothing in the first period. But how do we know that the households with the lowest productivity are always borrowers? If what they consume in the first period is more than what they produce in the first period (called \( y_0 \)) they have the borrow the difference. Hence, they are borrowers if:

\[
c_0 > y_0 \quad \Rightarrow \quad \omega \left( 1 + \frac{\gamma}{1+r} \right) p^{-1} > \omega \quad \Rightarrow \quad \frac{\gamma \sigma}{\beta (1+r)} > 1.
\]

Since we have assumed that \( \frac{\gamma}{\beta (1+r)} > 1 \), the last inequality is always true.

How do we know that the households with \( \omega^B_a < \omega < \omega^B_b \) are always borrowers? Again, we plug in the relevant values, and get:

\[
c_0 > y_0 \quad \Rightarrow \quad \omega \left( e_0 + \frac{\gamma}{1+r} \right) p^{-1} > \omega e_0
\]

We plug in the value for \( e_0 \) (which we can solve for using \( e_1 = 1 \), the budget constraint, equation (55) and equation (56)) and get:

\[
\frac{\gamma}{1+r} > \left( \frac{\psi}{1+\frac{1}{\beta}} \right)^{-\frac{1}{\beta}} \omega \frac{1-\sigma}{\sigma} \frac{1}{\beta (1+r)^{\frac{1-\sigma}{\gamma}}}
\]

If anyone in this groups would like to save, it would be the one working the most in the first period, i.e. the household with productivity \( \omega^a \). Plugging in the expression for \( \omega^a \) in the inequality above, we finally again get that \( c_0 > y_0 \) iff \( \frac{\gamma \sigma}{\beta (1+r)} > 1 \).

Hence, in this case, with \( \frac{\gamma}{\beta (1+r)} > 1 \), all households, regardless of productivity level, always want to borrow.

**Front-loading of work:** \( \frac{\gamma}{\beta (1+r)} < 1 \) In this case, when the productivity is large enough, the household will start reducing its fraction of individuals working in the second period. Again, we can structure the solution to the household problem along the productivity dimension according to Table 5 and calculate three breakpoints:

- \( \omega^F_a \): When the household is in the interior solution for \( e_1 \) (i.e., equation (56) holds with equality), and chooses \( e_1 = 1 \)
- \( \omega^F_b \): When the household is in the interior solution for \( e_1 \) (i.e. equation (56) holds with equality), and chooses \( e_1 = 0 \)
<table>
<thead>
<tr>
<th>Productivity</th>
<th>Choice of ((e_0, e_1))</th>
<th>Is the hh borrowing or saving?</th>
</tr>
</thead>
<tbody>
<tr>
<td>(\omega &lt; \omega_a^F)</td>
<td>((1,1))</td>
<td>Borrowing or saving</td>
</tr>
<tr>
<td>(\omega_a^F &lt; \omega &lt; \omega_b^F)</td>
<td>((1, x))</td>
<td>Borrowing or saving</td>
</tr>
<tr>
<td>(\omega_b^F &lt; \omega &lt; \omega_c^F)</td>
<td>((1, 0))</td>
<td>Saving</td>
</tr>
<tr>
<td>(\omega_c^F &lt; \omega)</td>
<td>((x, 0))</td>
<td>Saving</td>
</tr>
</tbody>
</table>

Table 5: Choice of \(e_t\) in each period depending on productivity, if \(\frac{\gamma}{\beta(1+r)} < 1\). Choice \(x\) denotes an interior solution, i.e. \(e_t \in (0,1)\), for the period.

\(\omega_c^F\) When the household is in the interior solution for \(e_0\) (i.e. equation (57) holds with equality), and chooses \(e_0 = 1\)

After some algebra we get the following expressions for the breakpoints:

\[
\omega_a^F = \left( \frac{\psi}{1 + \frac{1}{\beta}} \right)^{\frac{1}{1-\sigma}} \left( 1 + \frac{\gamma}{1 + r} \right)^{\frac{\sigma}{1-\sigma}} \left( \frac{\gamma}{\beta(1+r)} \right)^{\frac{1}{\sigma - 1}}
\]

\[
\omega_b^F = \left( \frac{\psi}{1 + \frac{1}{\beta}} \right)^{\frac{1}{1-\sigma}} \left( 1 + \frac{1}{p} \right)^{\frac{\sigma}{1-\sigma}} \left( \frac{\gamma}{\beta(1+r)} \right)^{\frac{1}{\sigma - 1}}
\]

\[
\omega_c^F = \left( \frac{\psi}{1 + \frac{1}{\beta}} \right)^{\frac{1}{1-\sigma}} \left( 1 + \frac{1}{p} \right)^{\frac{\sigma}{1-\sigma}}
\]

It is clear that the households with high productivity, \(\omega > \omega_b^F\), must be savers, since they produce nothing in the second period.

For the households with very low productivity, we can calculate that they must be borrowers if \(\frac{\gamma}{\beta(1+r)} > 1\), otherwise they are savers.

For the households with productivity \(\omega_a^F < \omega < \omega_b^F\), we get the following condition for the households which are borrowers:

\[
\omega < \left( \frac{\psi}{1 + \frac{1}{\beta}} \right)^{\frac{1}{1-\sigma}} \left( \frac{\gamma}{\beta(1+r)} \right)^{\frac{-1}{\sigma - 1}}
\]

We can show that for households exactly at the breakpoint \(\omega = \omega_a^F\) we recreate the borrowing condition \(\frac{\gamma}{\beta(1+r)} > 1\). Households exactly at breakpoint \(\omega = \omega_b^F\) are always savers. Hence, in this range the above condition for \(\omega\) is what guides the decision.
CHAPTER 2

Productivity Choice of \((e_0, e_1)\) Is the hh borrowing or saving?

| \(\omega < \omega E\) | (1,1) | Borrowing |
| \(\omega E < \omega\) | \((x, x)\) | Borrowing or saving |

Table 6: Choice of \(e_t\) in each period depending on productivity if \(\frac{\gamma}{p(1+r)} = 1\). Choice \(x\) denotes an interior solution, i.e. \(e_t \in (0,1)\), for the period.

The equality case: \(\frac{\gamma}{p(1+r)} = 1\) If we have the case of \(\frac{\gamma}{p(1+r)} = 1\), then either both or none of the first-order conditions for \(e_0\) and \(e_1\) are binding. We can calculate the breakpoint for when a household wants to work \((e_0, e_1) = (1,1)\) using the budget constraint and the first-order conditions (56) / (57), and we get:

\[
\omega E_a = \left(\frac{\psi}{1 + \frac{1}{\theta}}\right)^{\frac{1}{\sigma}} \left(1 + \frac{\gamma}{p(1+r)}\right)^{\frac{1}{\sigma}}
\]

So in this case we have the choice of working according to Table 6.

A.7. Balanced growth in a market economy with capital

The task in this section is to provide conditions under which the economy admits an exact balanced growth path in a neoclassical economy with capital. So suppose that households are heterogenous in their initial asset level \(a\) and their labor market efficiency \(\omega\). Households solve the following problem:

\[
\max \sum_{t=0}^{\infty} \left[ \sum_{h \in \mathcal{H}} \frac{c_{t+1}}{1 + r_{t+1}} - \frac{\psi}{1 + \frac{1}{\theta}} \int_{h \in \mathcal{H}} \pi_t(h) h^{1 + \frac{1}{\theta}} dh \right]
\]

subject to

\[
\mathcal{H} = \{0\} \cup [h, 1]
\]

\(0 \leq \pi_t(h) \leq 1, \quad \forall h \in \mathcal{H},\)

\(\int_{h \in \mathcal{H}} \pi_t(h) dh = 1,\)

and

\[
a(1 + r_0) + \sum_{t=0}^{\infty} p_t \left[ \omega w_t \int_{h \in \mathcal{H}} \pi_t(h) h dh \right] = \sum_{t=0}^{\infty} p_t c_t.
\]

We have \(\frac{p_t}{p_{t+1}} = 1 + r_{t+1}\) and \(p_0 = 1\). Here \(T\) is a governmental transfer and \(\tau \in [0,1)\) is a proportional tax on labor income.
When we refer to decisions made as a result of the maximization problem just stated, we may use the additional argument \((a, \omega)\). In the following let us also define a household’s wealth level in period \(t\) as follows:

\[
a_{t+1}(a, \omega) = a_t(a, \omega) [1 + r_t] + \omega w_t \int_{h \in H} \pi_t(h; a, \omega) h(a, \omega) dh - c_t(a, \omega),
\]

where

\[
a_0(a, \omega) = a.
\]

In this economy an equilibrium is a sequence

\[
\left\{ \left( c_t(a, \omega), \{ \pi_t(h; a, \omega) \}_{h \in H}, a_{t+1}(a, \omega) \right) \right\}_{t=0}^{\infty},
\]

a sequence of factor prices \(\{r_t, w_t\}_{t=0}^{\infty}\), and an aggregate capital and labor input \(\{k_t, n_t\}_{t=0}^{\infty}\), where

\[
k_t = \int_a \int_\omega a_t(a, \omega) \Gamma(da, d\omega) \quad \text{and} \quad n_t = \int_a \int_\omega \int_{h \in H} \pi_t(h; a, \omega) h(a, \omega) dh \Gamma(da, d\omega),
\]

such that:

1. the sequence \(\left\{ c_t(a, \omega), \{ \pi_t(h; a, \omega) \}_{h \in H} \right\}_{t=0}^{\infty}\) solves the household problem for all \(a, \omega\) and for given factor prices;
2. and the factor prices are consistent with perfect competition on the firm side, i.e.,

\[
r_t = F_1(k_t, \gamma^t n_t) - \delta \quad \text{and} \quad w_t = \gamma^t F_2(k_t, \gamma^t n_t), \quad \forall t.
\]

Let us define a balance-growth equilibrium as follows:

**Definition 3.** A balanced-growth equilibrium is an equilibrium along which all the variables grow at constant rates.

Note, of course, that different variables can grow at different rates, which may involve both increasing, decreasing, and constant variables.

In this economy there exists an exact balanced-growth equilibrium if \(\sigma > 1, \gamma > 1, \omega > 0, a\) is bounded from below, and an appropriate initial condition is met. We have:

**Proposition 3.** Suppose that \(a\) and \(\omega\) are distributed such that

\[
c_t(a, \omega) > \left( \frac{\omega w_t (1 - \tau) (1 + \frac{1}{\beta})}{\psi \lambda^\beta} \right)^{\frac{1}{\beta}}, \quad \forall a, \omega.
\]

Then, there exists a balanced growth path along which for all \(a, \omega\):

\[
\frac{c_{t+1}(a, \omega)}{c_t(a, \omega)} = \frac{a_{t+1}(a, \omega)}{a_t(a, \omega)} = \frac{k_{t+1}}{k_t} = \frac{1}{\gamma^\sigma} > 1,
\]

\[149]
The fraction working $h$ is for all $a, \omega$ either zero or $0 < s_t(a, \omega) < 1$, where
\begin{equation}
\frac{s_{t+1}(a, \omega)}{s_t(a, \omega)} = \frac{n_{t+1}}{n_t} = \gamma^{\frac{\sigma - 1}{\sigma}} < 1. \tag{94}
\end{equation}

The interest and wage rates are given by
\begin{equation}
r_t = \frac{\gamma}{\beta} - 1 = F_1 \left[ \left( \frac{k}{n} \right)^{\ast}, 1 \right] - \delta, \forall t, \tag{95}
\end{equation}
where $\left( \frac{k}{n} \right)^{\ast}$ is the balanced-growth capital-labor ratio, and
\begin{equation}
w_t = \gamma^t \left[ F \left[ \left( \frac{k}{n} \right)^{\ast}, 1 \right] - \left( \frac{\gamma}{\beta} - 1 + \delta \right) \left( \frac{k}{n} \right)^{\ast} \right]. \tag{96}
\end{equation}

Finally, the individual consumption level is given by
\begin{equation}
c_t(a, \omega) = s_t(a, \omega) \omega w_t + a_t(a, \omega) \left[ \frac{\gamma}{\beta} - \gamma^{\frac{1}{\sigma}} \right], \tag{97}
\end{equation}
where the fraction working $h$ of a particular family, $s_t(a, \omega)$, is given by
\begin{equation}
s_t(a, \omega) = \max \left[ (\omega w_t)^{-\left( 1 - \frac{1}{\sigma} \right)} - a_t(a, \omega) \frac{\gamma}{\beta} - \gamma^{\frac{1}{\sigma}} \frac{1}{\omega w_t}, 0 \right]. \tag{98}
\end{equation}

The intuition why balanced growth obtains follows that presented in 3.5.2.1.\textsuperscript{23} Note that we can use this proposition to find out what the long-run interest rate will be, as it gives us long-run consumption growth.

**Proof of Proposition 3.** We prove it by guessing and verifying the solution.

Condition (92) makes sure that all households are “rich” enough such that it is optimal to randomize for each family member to work either $h$ or $0$, where the share working $s_t(a, \omega)$ is strictly small than one. Then, the household’s problem can be simplified as
\begin{equation}
\max_{\{c_t, s_t\}} \sum_{t=0}^{\infty} \left[ \frac{c_t^{1-\sigma}}{1-\sigma} - \frac{\psi}{1 + \frac{1}{\beta} h^{1+\frac{1}{\sigma}}} s_t h^{1+\frac{1}{\sigma}} \right] \tag{99}
\end{equation}
subject to
\begin{equation}
0 \leq s_t \leq 1. \tag{100}
\end{equation}

\textsuperscript{23}Note that (92) is stated in terms of endogenous variables, i.e., $c_t(a, \omega)$. It is possible to state sufficient conditions on primitives such that this condition is met but for brevity we omit this statement. It is rather complex because of the lack of aggregation: the condition has to be stated indirectly as no closed-form solutions are available for factor prices in terms of primitives. The key conditions is fulfilled as long as the minimum $a$ and $\omega$ values are “large enough”.

150
and
\[ a(1 + r_0) + \sum_{t=0}^{\infty} p_t \omega_t s_t h = \sum_{t=0}^{\infty} p_t c_t, \quad (101) \]

where \( s \) is the share working \( h \) for any given \( \omega \). We know that by assuming (92) the constraint \( s_t \leq 1 \) is never binding. Then, the associated Lagrangian can be written as
\[ \mathcal{L} = \sum_{t=0}^{\infty} \left[ c_t(a, \omega)^{1-\sigma} - \frac{\psi}{1+\frac{1}{\beta}} s_t(a, \omega) h^{1+\frac{1}{\beta}} \right] + \lambda(a, \omega) \left[ a(1 + r_0) + \sum_{t=0}^{\infty} p_t \omega_t s_t h - \sum_{t=0}^{\infty} p_t c_t \right] + \sum_{t=0}^{\infty} \kappa_t(a, \omega) [s_t(a, \omega)] \]
The first-order conditions are:
\[ \omega \omega_t h c_t(a, \omega)^{-\sigma} = \frac{\psi h^{1+\frac{1}{\beta}}}{(1+\frac{1}{\beta})} - \kappa_t(a, \omega) \quad (102) \]
\[ \kappa_t(a, \omega) \geq 0, \quad (103) \]
\[ s_t(a, \omega) \geq 0, \quad (104) \]
\[ \kappa_t(a, \omega)s_t(a, \omega) = 0, \quad (105) \]
\[ \frac{c_{t+1}(a, \omega)}{c_t(a, \omega)} = \left[ \beta(1 + r_{t+1}) \right]^{\frac{1}{\beta}}, \quad (106) \]
and the period budget constraint
\[ a_{t+1}(a, \omega) = a_t(a, \omega) [1 + r_t] + \omega \omega_t s_t(a, \omega) h - c_t(a, \omega), \quad (107) \]
where \( \kappa_t \) is the multiplier attached to the non-negativity constraint on \( s_t \).

From the firm’s optimization problem we get the standard first-order conditions that can be expressed as \( r_t = F_1(k_t, n_t) - \delta \) and \( \omega_t = \gamma^t F_2(k_t, n_t) \). The conjectured consumption growth rate implies \( r_t = \frac{\gamma}{\beta} - 1 \) and pins down the balanced growth capital-labor ratio \( \left( \frac{k}{n} \right)^* \) according to (95). For the wage rate (96) then follows immediately from the firm’s first-order condition and the fact that the production function exhibits constant returns to scale. Furthermore, note that since \( \omega_t \) grows at constant rate \( \gamma \), the first-order condition (102) implies that \( \kappa_t(a, \omega) \) is constant along the balanced growth path for all households. Hence, the number of households that is not working at all, i.e., \( s_t(a, \omega) = 0 \) is constant over time. Furthermore, note that (107) is consistent with both \( s \) and \( a \) growing at the conjectured balanced growth rates. Note also that this is consistent with \( k \) and \( n \) growing at identical rates. Finally, for \( \kappa_t(a, \omega) = 0 \), the first-order condition (102) is consistent with (98). ■

A.8. Results incomplete markets models: correlations
# Chapter 2

<table>
<thead>
<tr>
<th></th>
<th>KPR preferences</th>
<th>BK preferences</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\sigma = 1.0$</td>
<td>$\sigma = 1.7$</td>
</tr>
<tr>
<td>No valued leisure</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Assets/consumption</td>
<td>0.87</td>
<td>0.87</td>
</tr>
<tr>
<td>Consumption/hours</td>
<td>$-$</td>
<td>$-$</td>
</tr>
<tr>
<td>Assets/hours</td>
<td>$-$</td>
<td>$-$</td>
</tr>
<tr>
<td>Assets/earnings</td>
<td>0.38</td>
<td>0.40</td>
</tr>
<tr>
<td>Productivity/hours</td>
<td>$-$</td>
<td>$-$</td>
</tr>
<tr>
<td>Earnings/hours</td>
<td>$-$</td>
<td>$-$</td>
</tr>
<tr>
<td>Assets/productivity</td>
<td>0.38</td>
<td>0.40</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Valued leisure, $h \in [0, \infty)$</th>
<th>$\psi = 1.0$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Assets/consumption</td>
<td>0.81</td>
</tr>
<tr>
<td>Consumption/hours</td>
<td>0.07</td>
</tr>
<tr>
<td>Assets/hours</td>
<td>$-0.34$</td>
</tr>
<tr>
<td>Assets/earnings</td>
<td>0.24</td>
</tr>
<tr>
<td>Productivity/hours</td>
<td>0.53</td>
</tr>
<tr>
<td>Earnings/hours</td>
<td>0.75</td>
</tr>
<tr>
<td>Assets/productivity</td>
<td>0.51</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Valued leisure, $h \in {0, h_{\gamma^{-vt}}}$</th>
<th>$\psi = 0.8$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Assets/consumption</td>
<td>0.76</td>
</tr>
<tr>
<td>Consumption/hours</td>
<td>$-0.01$</td>
</tr>
<tr>
<td>Assets/hours</td>
<td>$-0.33$</td>
</tr>
<tr>
<td>Assets/earnings</td>
<td>0.17</td>
</tr>
<tr>
<td>Productivity/hours</td>
<td>0.25</td>
</tr>
<tr>
<td>Earnings/hours</td>
<td>0.64</td>
</tr>
<tr>
<td>Assets/productivity</td>
<td>0.44</td>
</tr>
</tbody>
</table>

| $\psi = 1.0$                                        |
|--------------------------------------------------|--------------|
| Assets/consumption                              | 0.78         | 0.72           |
| Consumption/hours                               | 0.06         | $-0.13$        |
| Assets/hours                                    | $-0.29$      | $-0.43$        |
| Assets/earnings                                 | 0.15         | 0.03           |
| Productivity/hours                              | 0.24         | 0.13           |
| Earnings/hours                                  | 0.75         | 0.74           |
| Assets/productivity                             | 0.55         | 0.55           |

**Table 7**: Resulting Pearson correlation coefficients between different model moments. Note that the hours measure in the models with extensive margin labor choice is a binary variable (hence the point-biserial correlation coefficient is reported).
Health dynamics and heterogeneous life expectancies

Richard Foltyn Jonna Olsson

1. Introduction

In this paper, we provide improved estimates for age-dependent health transitions and survival probabilities for different subsamples of the US population. The estimated yearly transition matrices for health and death can be used in any life-cycle model where the evolution of health and survival probability is of interest. The implied life expectancy, which does not only depend on the standard dimensions of gender, race, and age, but also on health, can be used to assess inequality in longevity in the population based on health and potentially also other characteristics of the individual.

Numerous studies have identified health dynamics and health shocks as a major source of risk over the life cycle. A negative health shock can result in large medical expenditures (De Nardí et al. 2010; Kopecky and Koreshkova 2014), which affects the incentives to accumulate assets, and could also affect the earnings potential (Coile et al. 2016; French 2005). The survival probability directly affects the effective discount factor, a mechanism present in any life-cycle model with an uncertain life span (see De Nardí et al. (2017) for estimates of this effect as well as a comprehensive estimate of the cumulative effects of bad health). According to Finkelstein et al. (2013), the health state directly influences the marginal utility from consumption. Hence, in order to quantify the risk an individual faces and model the choices and actions the individual takes, a correct health and survival process is crucial.

We use the Health and Retirement Study (HRS), a representative panel of elderly US households, to investigate the development of health and longevity in the later stages of life. The survey includes, among other things, information on self-reported health and the date of death, if applicable. The survey started in 1992, and many of the respondents have died over the sample period, making it the preferred data set for studying survival and health dynamics.

However, the data set also has some peculiarities. For most cohorts and time periods, the survey has been conducted on a biennial basis. In practice, due to variation in interview dates, we effectively observe individual time spans of one, two or
three years for each interview wave.\footnote{Moreover, even though attrition is low, there are individuals who are not observed in one or more waves, so the time span between two observations can exceed three years.} Second, the death dates are coded exactly, and do not follow the interview wave structure. Third, even if all observations were perfectly biennial, the observations are overlapping: we observe transitions for one person at ages $a$ and $a + 2$ and for another person at ages $a + 1$ and $a + 3$. Both observations should contribute to estimating the one-year health and survival transitions between ages $a + 1$ and $a + 2$.

In this paper, we estimate a yearly five-state Markov chain that can be used in a life-cycle model. Our methodology is similar in spirit to that of Pijoan-Mas and Ríos-Rull (2014), but takes into account the irregular and overlapping observations described above and estimates one-year health transition probabilities and survival probabilities that depend on sex, race, age and health status (as opposed to previous studies using the HRS, which estimate two-year probabilities).

Conceptually, our method is a straightforward maximum likelihood estimator, where we maximize the probability of observing the transition paths in data. Each transition path consists of an initial health state, a number of periods, and an end point, which is either a health state (conditional on survival) or death. The estimation includes rolling forward each start observation the appropriate number of years with the Markov chain that is being estimated.

To put structure on the Markov chain we use a nested logit, where survival and health transitions conditional on survival are modeled as functions of the current health state and age. The probability of survival follows the usual binary-outcome logit model, while, conditional on survival, health transitions are modeled using multinomial logit.\footnote{We also implemented an alternative specification, estimating the health and survival probabilities separately in a two-step procedure, which yields very similar results.}

The resulting health gradient for longevity is strong. For a 70-year-old nonblack man in excellent health, the probability of reaching his 80th birthday is around 75\%, while the corresponding probability for a nonblack man in poor health is just below 40\%. Another way of expressing the same health gradient is that the expected longevity for a 50-year-old nonblack man in excellent health is 79 years, while it is only 73 years for a man in poor health.

The relationship between socio-economic status and life expectancy is well established but remains poorly understood (see, e.g., Chetty et al. (2016) and the references therein). With our methodology we can include time-invariant characteristics when estimating the health and survival dynamics. We show that there is substantial inequality in life expectancy between different educational groups. The average nonblack man with less than a high school degree has a life expectancy of 75 years at the age of 50, while the average for nonblack men with some college education or more is 80 years. This difference is due to two factors. First, at the age of 50, overall health is worse in the group with lower education. Second, even conditional on health status, the health dynamics and survival probabilities for
this group are worse also from the age of 50 and onwards. We estimate that of the
differences in life expectancy between the low and high educated, approximately
one fifth is due to worse overall health at the age of 50, while the lion’s share is due
to worse health and survival dynamics after that age.

Most estimates of health processes used for life-cycle models collapse the state
space for health into two groups: good or bad health (De Nardi et al. 2017; French
2005; French and Jones 2011). There are two benefits from using the full state space
as reported in the HRS: First, trivially, a larger state space captures more of the
heterogeneity in the population. Second, a richer state space allows for more correct
dynamics and persistence of the process. The drawback of using more states is, of
course, the additional computational burden. However, we think that it is key to
properly capture the persistence and duration dependence of staying in bad health
when modeling the health risk that individuals face (see De Nardi et al. (2017) for a
discussion).

Our estimated process is highly persistent, especially for the worst health state.
Once there, the probability of remaining in the worst health state another period is
above 75%. The importance of health persistence is stressed by, e.g., Contoyannis
et al. (2004), and the persistence we estimate is well in line with the persistence for
the five-state frailty index that Hosseini et al. (2018) develop.

If we use our five-state process, but interpret the results according to a two-state
classification (with the two worst health states classified as “bad”, in line with
previous literature), the result is a negative duration dependence in the probability
of recovering from bad health. For a 70-year-old nonblack man who has been in
bad health only one year, the probability of recovering to good health is 20%, while
if he has been in bad health five years, the probability is five percentage points
lower.

In the next section we describe the structure of the HRS data, and thereafter we
describe the estimation in detail. Section four shows the baseline results while
section five analyzes the health and life expectancy inequality between different
education groups. The last section concludes.

2. Data

We use the Health and Retirement Study (HRS), a representative panel of US
households in older ages, to investigate the development of health and longevity in

---

3 One reason is that to estimate yearly transitions, authors have resorted to using PSID, which until
1997 was a yearly survey. However, the number of individuals there is relatively small and therefore
it is necessary to combine data into coarser health states.

4 It is also substantially higher than the persistence which Hosseini et al. (2018) calculate for self-reported
health based on PSID. This points both at the strength of HRS (which includes more individuals in
older ages), and at the importance of estimating age-dependent transition probabilities.

5 Another alternative is to, like De Nardi et al. (2017), use a two-state health process with a second-order
Markov process and fixed ex-ante heterogeneity to capture the duration dependence.
the later stages of life. The survey includes, among other things, questions about self-reported health state and date of death, if applicable.

Figure 1: Structure of the Health and Retirement Study. The y-axis shows the age of the respondents in each cohort and wave. Graph taken from https://hrs.isr.umich.edu/documentation/survey-design, where additional information about the survey design is available.

The survey started in 1992 and in this paper, we use HRS data up to and including the eleventh wave in 2012.\(^6\) The first cohort included in the survey was between 51 and 61 years old in 1992, and thereafter new cohorts have been added. Many of the respondents have died over the sample period, making it an appropriate data set for studying survival. Figure 1 shows the panel structure of the HRS and how new cohorts have been incorporated into the survey over time.

As can be seen from Figure 1, the survey has been conducted biennially for most cohorts and time periods. However, in practice, there is a substantial variation in the time elapsed between interviews. Each survey round is conducted over a period of time, and the actual time elapsed between interviews for a respondent for two consecutive waves varies between one and three years. For respondents missing one or more interviews, the time interval between two interviews or the time elapsed between the last interview and the death date is more than three years. Figure 2 shows the distribution of the actual time elapsed from one observation to

\(^6\)The RAND version O, covering waves up until 2012, is the most recent RAND release that includes data from the National Death Index (NDI). Since correct death dates are crucial for our analysis, this is our preferred data set. There is one later RAND release, covering the 2014 wave as well, but there NDI data is lacking. An analysis shows that there are discrepancies in death dates between the exit interview information and the NDI date of death.
Figure 2: Time elapsed between two observations (second observation being either a new interview or the death date).

Figure 3: Two individuals in the HRS, illustrating the irregular and missing observations. Waves and calendar years indicate the biennial structure, age indicates the actual age at the time of the interview.

The next for the full sample. As can be seen, slightly more than 80% of the transitions are best characterized as two-year transitions, but almost 20% are not. Figure 3 illustrates two typical observations in our sample.

The two key variables we use are self-reported health and date of death. Self-reported health is simply the respondent’s answer to the question “Would you say your health is excellent, very good, good, fair, or poor?”. The first reason to use this variable is the general availability of this information. Very similar questions are asked in many other surveys, both within the US (for example the Panel Study of Income Dynamics (PSID) and the Medical Expenditure Panel Survey (MEPS) include this question) and also globally (for instance the Survey of Health, Aging and Retirement in Europe (SHARE) asks about self-reported health). Hence, the insights into the dynamics of self-reported health and life expectancy conditional on this measure can be used for analyses based on many other data sets.

Second, a number of studies have shown that self-reported health is highly correlated with other subjective and objective measures of health and is also a good predictor for future mortality (see e.g. Idler and Benyamini (1997) and Pijoan-Mas and Ríos-Rull (2014)). Self-reported health can be interpreted as a one-dimensional variable capturing high-dimensional information, letting the respondent aggregate
Figure 4: Distribution of health states by age. Red color indicates worst (“poor”) health state while dark green indicates best (“excellent”) health state. Observations are grouped into two-year age bins.

(a) Nonblack men

(b) Nonblack women

this information him- or herself.\(^7\)

Figure 4 shows the distribution of nonblack individuals by health state for different ages.\(^8\) The overall health is declining in age, but it might be surprising that the health distribution among 50-year-old individuals is not that much better than among 90-year-old individuals. This suggests that the aggregation of underlying health measures done by the respondent also takes into account the relative health within cohort. A 70-year-old respondent who reports “excellent” health can still feel worse than his/her 20-year self, but “excellent” in comparison to what the person perceives could be expected as a 70-year-old. Since all our estimates will be conditional on age this is taken into account.

The second key variable is the date of death. An important feature of the HRS data is that the death date is recorded, regardless of whether the respondent stayed in the survey until his/her death or not. Hence, the attrition for this particular information is virtually zero. The death date is first recorded in the so-called exit interview or by the surviving spouse. The survey is then complemented with information from the National Death Index (NDI). Therefore, a correct date of death is recorded even for those who dropped out of the survey, and the death date does not follow the biennial wave structure but corresponds to the actual death date.

We exclude all observations with missing age, race, gender or self-reported health, and those where we only have one observation for the individual (since then we do not have any transition probabilities to estimate). Further, we restrict the sample

\(^7\)An alternative is to let the researcher do the aggregation into a single index, incorporating different physical and mental conditions. See Amengual et al. (2017) and Hosseini et al. (2018) for other suggested health groupings.

\(^8\)Black individuals are generally slightly worse in terms of health. Corresponding graphs for black men and women can be found in the appendix.
Table 1: Sample description.

<table>
<thead>
<tr>
<th></th>
<th>N. of indiv.</th>
<th>N. obs.</th>
<th>Avg. obs./indiv.</th>
<th>Age</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Min.</td>
</tr>
<tr>
<td>Male</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Non-black</td>
<td>12,765</td>
<td>77,612</td>
<td>6.1</td>
<td>50</td>
</tr>
<tr>
<td>Black</td>
<td>2,397</td>
<td>12,252</td>
<td>5.1</td>
<td>50</td>
</tr>
<tr>
<td>Female</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Non-black</td>
<td>15,485</td>
<td>99,455</td>
<td>6.4</td>
<td>50</td>
</tr>
<tr>
<td>Black</td>
<td>3,556</td>
<td>19,556</td>
<td>5.5</td>
<td>50</td>
</tr>
</tbody>
</table>

to individuals above the age of 50.\(^9\) We estimate the subsamples of men/women and the nonblack/black population separately, since it is well known that the life expectancies for these subgroups follow very different trajectories, and are reported separately by the NVSS.\(^10\) Table 1 shows the number of individuals and observations by subgroup. We show estimates for nonblack men and women in the main document, due to the larger sample size. The results for black men and women are available upon request. For all estimations we weight observations by their person-level analysis weight.

3. Estimation

Our goal is to estimate a Markov chain for annual survival probabilities and health-to-health transitions conditional on survival. The approach takes into account that observations in the HRS can occur at irregular (mostly biennial but also non-biennial) frequencies.\(^11\) Our proposed methodology allows us to address two shortcomings of the HRS data:

(1) We need to estimate annual transition probabilities, but HRS waves are biennial. Due to variation in interview dates, we effectively observe transitions over one, two, three or more years, with about 80% of transitions being best described as two-year transitions.

(2) Additionally, transition intervals are overlapping: we often observe transitions between ages \(a\) and \(a + 2\) for one respondent, while observing a transition from

\(^9\)Even though no one in the core sample should be under that age, the full sample includes spouses that could be younger.

\(^{10}\)The HRS also has information about Hispanic origin, but the Hispanic group was not added to the annual life table program until 2006, while we use earlier data for some comparisons (see Arias (2014) for technical notes describing the life table program). Therefore, we follow the older life table classifications and only use the black/nonblack categories for our estimates (with Hispanics included in the nonblack group). Note also that other groups than white and black (and later Hispanics) are not reported separately by the NVSS, due to concerns about data limitation and misclassification. We keep the label “nonblack” throughout the document, but when comparing results to life tables we make the comparison with the “white” group.

\(^{11}\)This is in contrast to Pijoan-Mas and Rios-Rull (2014), who estimate two-year transition probabilities and force all transitions to be at a biennial frequency regardless of the actual time passed between two consecutive interviews, and furthermore restrict transitions to occur between even ages, i.e. they assume that the observed transitions line up with ages 50, 52, 54, . . . .
These two observations contribute to estimating the transition matrices between ages \( a + 1 \) and \( a + 2 \). Therefore, we cannot simply estimate two-year transition matrices and take a matrix “square root” to obtain one-year transitions.

We implement a custom maximum-likelihood estimator that takes into account varying transition frequencies and overlapping transition intervals. In this section we describe an approach that jointly estimates the parameters governing both health-to-health and survival transitions in a single step.\(^{12}\)

### 3.1. Transition probabilities

While the HRS itself is organized into individual/year observations, for the purpose of estimating transition probabilities we will reinterpret the sample such that one transition constitutes one observation. Each transition consists of an observed starting health state, a number of periods, and an end point, which is the next time the individual is observed. The end point can be either a new health state or death. In what follows, we index transitions by \( i \) but usually skip writing the index explicitly.

We refer to a transition’s starting date as \( t \), to its length (in years) as \( T \), and to its end date as \( t = t + T \). We denote by the tuple \((h_t, x_t)\) the information available at time \( t \), \( t \in \{t, \ldots, t+T\} \), where \( h_t \in \{1, \ldots, H\} \) is an individual’s self-reported health state, with 1 representing the best and \( H \) the worst realization. The vector \( x_t \) contains any other variables of interest, in particular age. We allow for time-invariant characteristics such as birth year, gender, race or education level to be included in \( x_t \), but restrict the time-varying variables to age and potentially calendar year. This restriction is necessary as we need to compute the evolution of \( x_t \) over \( t+1, t+2, \ldots \) for multi-year transitions, which is not possible in general except for variables that follow a deterministic path (such as age and calendar year).

Let \( s_t \) be a binary indicator for whether a person is alive at date \( t \),

\[
    s_t = \begin{cases} 
    1 & \text{if alive at } t \\ 
    0 & \text{else} 
    \end{cases} \quad (1) 
\]

We assume that the one-period-ahead probability of survival is given by the binary-outcome logit model

\[
    p_{t+1}^s \equiv \Pr( s_{t+1} = 1 \mid h_t, x_t ) = \frac{1}{1 + e^{-g(h_t, x_t | \gamma)}} \quad (2)
\]

\(^{12}\)As a robustness test, we also estimate the process in two separate steps: first health-to-health transitions conditional on survival, and then survival probabilities conditional on health. The results are very similar and available upon request. Since the nested version uses more information, this is our preferred method.
Survival probabilities are governed by the parameter vector $\gamma$ which is to be estimated. Similarly, conditional on survival, the probability that health state $j$ is realized next period is given by the multinomial logit formula

$$p_{t+1}^{h,j} \equiv \Pr \left( h_{t+1} = j \mid s_{t+1} = 1, h_t, x_t \right) = \frac{e^{f_j(h_t,x_t|\beta_j)}}{\sum_m e^{f_m(h_t,x_t|\beta_m)}}$$  \hspace{1cm} (3)

with parameter vector $\beta_j$ to be estimated for each outcome $j$.\footnote{We assume that all parameters in $\beta_j$ are specific to outcome $j$ and there are no “common” parameters shared across all outcomes. This is due to the fact that we have no outcome-specific regressors and thus any common parameters would cancel out in (3), leaving these parameters unidentified.} Here we use the notation $p_{\cdot|s}^h$ to indicate that this probability is conditional on survival. We can then compute the unconditional probability of being in health state $j$ in the next period as

$$p_{t+1}^{h,j} = p_{t+1|s}^{h,j} \times p_{t+1}^{s}$$  \hspace{1cm} (4)

Below we will frequently want to emphasize that we condition on a particular health state $h_t = k$, and hence we will use the expressions

$$p_{t+1|k}^{s} \equiv \Pr \left( s_{t+1} = 1 \mid h_t = k, x_t \right)$$  \hspace{1cm} (5)

$$p_{t+1|k,s}^{h,j} \equiv \Pr \left( h_{t+1} = j \mid s_{t+1} = 1, h_t = k, x_t \right)$$  \hspace{1cm} (6)

which are otherwise identical to those established earlier.

In the following sections we lay out the estimation strategy to determine the parameter vectors $\gamma$ and $\beta_j$ for all outcomes $j$. We defer stating the exact functional forms of $g(\cdot)$ and $f(\cdot)$, as these are specific to the particular model to be estimated (e.g. whether cohort fixed effects or education are included, etc.).

### 3.2. Maximum-likelihood approach

In this section we describe the ML approach that jointly estimates the parameters governing both health-to-health and survival transitions in a single step.

**An illustrative example.** Before deriving the probabilities that should be plugged into the log-likelihood function, it is worthwhile to work through an illustrative example for the case of the nested logit estimator. We consider a simplified setup with only two health states and assume a two-year transition, as illustrated in Figure 5.

At $t + 2$ there are three possible outcomes, but seven distinct paths via which these outcomes can be realized. It turns out that the probability distribution that should enter the likelihood function is one over paths, not outcomes. To make
this point, first consider a PMF over outcomes in \( t + 2 \), which is given by the three probabilities

\[
\begin{align*}
\Pr (h_{t+2} = 1 \mid h_t) \\
\Pr (h_{t+2} = 2 \mid h_t) \\
\Pr (s_{t+2} = 0 \mid h_t)
\end{align*}
\]

For either health state \( j = 1, 2 \) these can be computed as follows:

\[
\Pr (h_{t+2} = j \mid h_t) = \sum_{m=1}^{2} \Pr (h_{t+2} = j \mid h_{t+1} = m, h_t) \Pr (h_{t+1} = m \mid h_t)
\]

On the other hand, the probability of observing death in \( t + 2 \) can be written as

\[
\Pr (s_{t+2} = 0 \mid h_t) = \sum_m \Pr (s_{t+2} = 0 \mid h_{t+1} = m, h_t) \Pr (h_{t+1} = m \mid h_t) + \Pr (s_{t+1} = 0 \mid h_t)
\]

The issue with this formulation is that whenever death in \( t + 2 \) is observed, the probability of this outcome includes the case that the individual already died in \( t + 1 \), which corresponds to path 7 in Figure 5. However, due to how the transition data was constructed this is impossible, as a one-period observation would have been recorded if an individual had already died in \( t + 1 \) (remember that the date of death is always recorded correctly and does not follow the wave structure). Hence, the probability associated with path 7 should never enter the likelihood function. This issue becomes even more pronounced for longer transitions, since the probability of ending up in the absorbing death state in the penultimate period is strictly increasing in the transition length.

To properly address this issue, we instead propose to compute the distribution over paths instead of outcomes. Naturally, in the above example we do not know whether path 1 or 2 was realized when we observe the outcome \( h_{t+2} = 1 \), so the probabilities of both will have to be included in that case, and analogously for the remaining outcomes.
To shut down all paths leading to “premature” death before the terminal period (which is only path 7 in the above example), we want to evaluate the probabilities of the events

\[ \begin{align*}
    &\Pr(h_{t+2} = 1 \land s_{t+1} = 1 | h_t) \\
    &\Pr(h_{t+2} = 2 \land s_{t+1} = 1 | h_t) \\
    &\Pr(s_{t+2} = 0 \land s_{t+1} = 1 | h_t)
\end{align*} \]

For either health outcome \( j = 1, 2 \) we find that

\[ \Pr(h_{t+2} = j \land s_{t+1} = 1 | h_t) = \Pr(s_{t+1} = 1 | h_{t+2} = j, h_t) \times \Pr(h_{t+2} | h_t) = \Pr(h_{t+2} | h_t) \]

which follows since

\[ \Pr(s_{t+1} = 1 | h_{t+2} = j, h_t) = 1 \]

An individual who is in health state \( j \) at \( t + 2 \) must have been alive at \( t + 1 \), so the additional restriction that \( s_{t+1} = 1 \) is redundant for health outcomes. However, this is not the case for the probability of being dead in \( t + 2 \):

\[ \Pr(s_{t+2} = 0 \land s_{t+1} = 1 | h_t) = \sum_{m=1}^{2} \Pr(s_{t+2} = 0 \land s_{t+1} = 1 | h_{t+1} = m, h_t) \times \Pr(h_{t+1} = m | h_t) = \sum_{m=1}^{2} \Pr(s_{t+2} = 0 | h_{t+1} = m, h_t) \Pr(h_{t+1} = m | h_t) \]

The second line follows since conditional on \( h_{t+1} = m \), we necessarily have \( s_{t+1} = 1 \). This formulation shuts down any paths with \( s_{t+1} = 0 \). Note that we can still compute the probability of such paths and include them in the log-likelihood formula, but since these are never observed, they will never be selected to actually contribute to the log-likelihood function. The consequence is that we can entirely skip computing the probability of these paths, which also leads to the at first puzzling corollary that for the components that actually do enter the log-likelihood, we have

\[ \Pr(h_{t+2} = 1 | h_t) + \Pr(h_{t+2} = 2 | h_t) + \Pr(s_{t+2} = 0 \land s_{t+1} = 1 | h_t) < 1 \]

in general.

**General setup.** We now return to the issue of estimating the parameter vector \( \theta \) that governs health-to-health and survival transitions, defined as

\[ \theta \equiv (\beta_2, \ldots, \beta_m, \ldots, \beta_H, \gamma) \in \mathbb{R}^K \]

where \( K = (H - 1)K_h + K_s, \beta_m \in \mathbb{R}^{K_h} \) for each \( m \) as before and \( \gamma \in \mathbb{R}^{K_s} \). We omit the normalized base outcome parameter vector \( \beta_1 = 0 \) for health state 1. From any
transition bracketed by the dates \( t \) and \( T \) we obtain one observation, a PMF over health states “augmented” by the state of death. We call this vector \( \mu_t \in \mathbb{R}^{H+1} \). In \( t \) we impose the degenerate initial distribution

\[
\mu_t = \left(0, \ldots, 0, 1, 0, \ldots, 0\right)^\top
\]  

with unity in the position corresponding to the initial health state \( h_t \).

The one-year health-to-health transition matrix conditional on survival is given by

\[
\Pi^h_t(x_t | \beta) = \begin{bmatrix}
p^{h,1}_{t+1|1,s} & \cdots & p^{h,H}_{t+1|1,s} \\
\vdots & \ddots & \vdots \\
p^{h,1}_{t+1|H,s} & \cdots & p^{h,H}_{t+1|H,s}
\end{bmatrix}
\]  

(8)

where the conditional probabilities \( p^{h,j}_{t+1|k,s} \) are defined in the same way as in (6). This transition matrix is a function of the covariate vector \( x_t \) and possibly calendar time \( t \), but not of the current health state \( h_t \) as it contains transitions for all possible current \( h_t \).

Let \( \pi^s_t \) be the vector of survival probabilities between periods \( t \) and \( t+1 \) for each health state \( k \in \{1, \ldots, H\} \) today,

\[
\pi^s_t(x_t | \gamma) = \left(p^{s}_{t+1|1}, \ldots, p^{s}_{t+1|k}, \ldots, p^{s}_{t+1|H}\right)^\top
\]  

(9)

where any element \( p^{s}_{t+1|k} \) is obtained as stated in (2). Given the distribution over health states conditional on being alive in \( t \), \( \mu^h_t \), the probability of being alive in \( t+1 \) is therefore

\[
p^{s}_{t+1}(\gamma) = \pi^s_t(\gamma)^\top \mu^h_t
\]  

(10)

We can now write down the joint health/survival transition matrix, given by

\[
\Pi_t(x_t | \theta) = \begin{bmatrix}
p^{h,1}_{t+1|1,s}p^{s}_{t+1|1} & \cdots & p^{h,H}_{t+1|1,s}p^{s}_{t+1|1} & \left(1 - p^{s}_{t+1|1}\right) \\
\vdots & \ddots & \vdots & \vdots \\
p^{h,1}_{t+1|H,s}p^{s}_{t+1|H} & \cdots & p^{h,H}_{t+1|H,s}p^{s}_{t+1|H} & \left(1 - p^{s}_{t+1|H}\right) \\
0 & \cdots & 0 & 1
\end{bmatrix}
\]  

(11)

We can then generate the distribution \( \mu_t \) over health/death states for any \( t \) by repeatedly applying the transition matrix, starting with the degenerate initial distribution (7). The law of motion for \( \mu_t \) is therefore

\[
\mu_{t+1}(\theta)^\top = \mu_t(\theta)^\top \Pi_t(x_t | \theta)
\]
In line with the initial discussion on computing PMFs over outcomes versus realizations of complete paths, we need to discard any paths that pass through the state \( s_{\tau-1} = 0 \). This can be achieved by computing the PMF \( \mu_{\tau-1} \) according to (11) and then defining the “pseudo” PMF

\[
\tilde{\mu}_{\tau-1} = (\mu_{1,\tau-1}, \ldots, \mu_{H,\tau-1}, 0)
\]

Note that \( \sum_j \tilde{\mu}_{j,\tau-1} = \Pr(s_{\tau-1} = 1 \mid h_t, x_t) \), the probability of being alive in \( T - 1 \). The terminal distribution of interest can then be computed as before, i.e.

\[
\mu_{T}(\theta) = \tilde{\mu}_{T-1}(\theta)\top\Pi_{T-1}(x_{T-1} \mid \theta)
\]

**Log-likelihood.** We are now ready to write down the likelihood function for observation \( i \). Let \( \delta_{t}^{h,j} \) be the indicator variable defined as

\[
\delta_{t}^{h,j} = \begin{cases} 
1 & \text{if } h_{t} = j \\
0 & \text{else}
\end{cases}
\]

and \( s_{\tau} \) be the indicator for being alive in \( \tilde{T} \), analogous to (1). Then the likelihood function for transition \( i \) is given by

\[
L_{i}(\theta) = s_{\tau} \left( \sum_{j=1}^{H} \delta_{t}^{h,j} \log \mu_{j,\tilde{T}}(\theta) \right) + (1 - s_{\tau}) \log \mu_{H+1,\tilde{T}}(\theta)
\]

The estimated parameter vector \( \hat{\theta} \) is hence the vector that maximizes the weighted sum of the log-likelihoods over all observations.\(^{14}\)

### 3.3. Baseline model specification

In the baseline specification of the functions governing the survival and health transitions, we include age polynomials of degree two as the only covariates other than the current health state, i.e. \( x_{t} = (a_{t}) \). We thus have the following functional forms for \( g(\bullet) \) and \( f_{j}(\bullet) \):

\[
g(h_{t}, x_{t} | \gamma) = \frac{2}{n} \sum_{n=0}^{H} \sum_{k=1}^{H} \gamma_{nk} \times \delta_{t}^{h,k} \times a_{t}^{n}
\]

\[
f_{j}(h_{t}, x_{t} | \beta_{j}) = \frac{2}{n} \sum_{n=0}^{H} \sum_{k=1}^{H} \beta_{jnk} \times \delta_{t}^{h,k} \times a_{t}^{n} \quad \forall j \in \{2, \ldots, H\}
\]

\(^{14}\)Even though it is conceptually a standard log-likelihood estimation, the implementation is non-standard and not included in any existing software, but specifically implemented for the problem at hand.
Note that one benefit of using the nested logit formulation is that it is possible to include different covariates in the functions governing the survival process and the health process. However, we find that a second-order polynomial is sufficient in both.

4. Baseline results

In this section, we first show the results from the estimations and thereafter, to verify our estimates, we compare the model-predicted transition probabilities to what we observe in the data. We then compute model-predicted life expectancy conditional on health, taking into account potential future health transitions to evaluate the health gradient for survival. Once more, we verify our results, this time by making a comparison to national statistics from NVSS. We conclude with a short discussion on the duration dependence of being in “bad” health, classifying the worst two health states as “bad” (in line with previous literature). We show that our five-state health process captures some of the duration dependence of bad health.

The resulting transition matrices by age, which can be directly incorporated in a lifecycle model, are downloadable from the authors’ websites.

4.1. Health transitions and survival probabilities

The health-to-health transition probabilities for nonblack males are shown in Figure 6, while those for nonblack women are shown in Figure 7. As can be seen, the health state is persistent: for a 70-year-old nonblack man in poor health, the probability of remaining in the same poor health state next year is around 75%. The probability of improving to anything better than the second worst health state is low, below 5%.

The same pattern holds true for all health states: to remain in the current health state is the most likely outcome, and to improve or deteriorate one step is the second most likely outcome. For a 50-year-old nonblack man in the best health state, the probability of remaining in excellent health is around 70%, but as age increases, it becomes more likely to transition to the second best health state.

Figure 8 shows the survival probabilities for nonblack men and women conditional on health. Unsurprisingly, the survival probability is decreasing in age, but there is also a clear health gradient. The probability of surviving one year ahead for a 50-year-old in the best health state (excellent) is almost 100% while for an otherwise similar individual in the worst health state it is approximately 95%. As is well known, the survival probability conditional on age is higher for women than for men.
Figure 6: One-year health-to-health transition probabilities conditional on survival for non-black men (model estimates). Initial health states in rows, terminal health states in columns. Shaded areas indicate bootstrapped 95% confidence intervals.
Figure 7: One-year health-to-health transition probabilities for nonblack women (model estimates). Initial health states in rows, terminal health states in columns. Shaded areas indicate bootstrapped 95% confidence intervals.
Figure 8: One-year survival probability (model estimates). Shaded areas indicate bootstrapped 95% confidence intervals.
Another way of illustrating the dynamics of the estimated health and survival process is given in Figure 9, which shows the evaluation of probabilities for each health state and for being dead by year, given an initial health state and a starting age. As can be seen, the survival probability differs substantially depending on the initial health state: for a nonblack man aged 70 in excellent health, the predicted probability of surviving an additional 10 years is more than 80%, but if he is in poor health, the probability is just around 40%.

4.1.1. Comparing model predictions and data

In order to compute model predictions to “raw” data moments, we compute the two-year transition probabilities implied by our annual model. Then, we compare these to the fraction of individuals with a particular outcome in a subsample restricted to two-year transitions, which is the large majority of observations, as shown in Figure 2.

Figure 10 shows the model-predicted two-year health-to-health transitions for non-black men compared to the actual observed two-year transitions, while Figure 11 shows the corresponding information for nonblack women. Given the strict functional form assumptions that we impose, the estimated probabilities and the data are remarkably close.

To compare model predictions for survival with observed data, we compare the estimated cumulative probability of having died within two years of the actual observed death within two years since the last observations. Figure 12 shows the results. Again, the model predictions and the data are remarkably close.

To assess how well our model predicts long-run outcomes, we compare actual survival rates as observed in the HRS with model predictions over a time horizon of up to 20 years. Figure 13 and Figure 14 show the model-predicted survival probability against the fraction actually surviving, plotted for eight different time periods, with each period being the time elapsed between a certain survey wave in the period 1992–2006 and the final wave in 2012. Each dot represents a two-year age bin. We discard age bins with less than 200 observations. The top left graph shows the age bins that were in the panel in the first wave and, as shown by Figure 1, there was only one cohort in the sample at that point. Therefore, there are fewer dots in this graph than in the others. As can be seen, the estimated model captures the survival probability very well.

4.2. Life expectancy conditional on health

To calculate the life expectancy conditional on health, we need to take into account all future health-to-health transition probabilities, since the road to death could go via any health transition path. The measure answers the question: what is the
Figure 9: Survival probability conditional on initial health state. X-axis indicates years after initial age (upper row 50, lower row 70 years). The colors indicate probability per health state (dark green being the best health state, red the worst). The hatched area represents the probability of being dead.
Figure 10: Two-year health-to-health transitions conditional on survival for nonblack men: data vs. model. Initial health states in rows, terminal health states in columns. The data includes all two-year observations (i.e., excluding shorter and longer transitions), model predictions are the two-year predictions. Shaded areas indicate bootstrapped 95% confidence intervals.
Figure 11: Two-year health-to-health transitions conditional on survival for nonblack women: data vs. model. Initial health states in rows, terminal health states in columns. The data includes all two-year observations (i.e., excluding shorter and longer transitions), the model predictions are the two-year predictions. Shaded areas indicate bootstrapped 95% confidence intervals.
Figure 12: Two-year survival probability: data vs. model. The data includes all deaths within two years after the previous observation, the model prediction is the cumulative probability after two years. Shaded areas indicate bootstrapped 95% confidence intervals.
Figure 13: Nonblack men: Model-predicted survival probability (on the x-axis) against the fraction actually surviving (on the y-axis), plotted for eight different time periods. The top left graph represents the time period between the first wave (1992) and the last wave (2012). Each dot represents a two-year age bin.
Figure 14: Nonblack women: Model-predicted survival probability (on the x-axis) against the fraction actually surviving (on the y-axis), plotted for eight different time periods. The top left graph represents the time period between the first wave (1992) and the last wave (2012). Each dot represents a two-year age bin.
The life expectancy of a person we observe at age $t$ in health state $h$? It is computed as follows:

$$e_{h,t} = \left[ \sum_{\tau=1}^{T} \sum_{k=1}^{H} \tau \times (1 - p_{\tau+1|k}^s) \times \mu_{k,\tau} \right] + \frac{1}{2}$$

where

$$\mu_{j,\tau+1} = \sum_{k=1}^{H} p_{\tau+1|k,s} \times p_{\tau+1|k}^s \times \mu_{k,\tau}$$

where the position of 1 corresponds to the initial health state. The addition of the half year is to correct for the fact that people do not die exactly on their birthday, but deaths are instead approximately uniformly spread out over the year.

Figure 15 shows the resulting life expectancies for men and women conditional on the initial health state. As can be seen, the health gradient is substantial: the difference in expected life length between a 50-year-old nonblack man in the best and in the worst health state is 6 years.

4.3. Comparing to life tables

The HRS data we use is from the period 1992 to 2012. If we had had many more individuals born each year, we could have computed cohort specific health and survival probabilities by age. However, the sample is not large enough to permit this. Instead, the survival probabilities we calculate should be viewed as period life expectancies for the sample period as a whole, and correspond to a weighted
average of what is reported in the period life tables by the National Vital Statistics System (NVSS) during those years.\textsuperscript{15}

We calculate the model-implied life expectancy at the age of 50 given our estimates and given the observed health distribution at this age. Our model gives 78.2 for men and 81.9 for women. This is well in line with what is reported by the NVSS during this period: for white men the NVSS life expectancy at the age of 50 ranges from 77.0 (in 1993) and 79.9 (in 2012) during the sample period. For white women it ranges from 81.7 (in 1993-1996) to 83.4 (in 2012). Hence, the model predictions are within what is reported by NVSS.\textsuperscript{16}

The same calculation done for 70-year olds confirms the conclusion. Our model predictions, based on our estimates and the observed health distribution, are 83.3 for nonblack men and 85.7 for nonblack women. The corresponding ages reported by NVSS during the period 1992 to 2012 ranges between 82.3 (in 1993) and 84.4 (in 2012) for men. For women the NVSS reports estimates between 85.3 years (in 1993) and 86.5 years (in 2012). Once more, the model predictions are in line with the national averages reported by the NVSS for the white population.

4.4. Duration dependency

It is common in the literature to aggregate the health states into two coarser categories: good (covering excellent, very good, and good health) and bad (covering fair and poor). However, a benefit of actually using the more fine-grained grid is that it is possible to capture the transition dynamics more precisely.

It has been shown that the probability of transitioning from the coarser bad health state into the good health state decreases with time: the longer an individual has been unhealthy, the less likely he/she is to become healthy again (De Nardi et al. 2017). To capture this duration dependency using a health process with only two states, it is necessary to use a higher-order Markov chain. However, using a five-state process, and following the literature by classifying the two worst health states as bad, we partly capture this duration dependency even though the process is a first-order only.

\textsuperscript{15}There are two types of life tables: period (or current) life tables and cohort (or generation) life tables. The period life table, which is what you find in a regular life table, presents what would happen to a hypothetical cohort if it experienced the mortality conditions of a particular period in time throughout its entire life. The cohort life table, on the other hand, presents the mortality experience of a particular birth cohort from the moment of birth through consecutive ages.

\textsuperscript{16}As pointed out by Pijoan-Mas and Ríos-Rull (2014), life expectancies computed from the HRS should differ slightly from the national average, since the HRS does not include institutionalized individuals.
As a stylized example, consider the following health-to-health transition matrix:

\[
\Pi^h = \begin{bmatrix}
3/4 & 1/4 & 0 & 0 & 0 \\
1/4 & 1/2 & 1/4 & 0 & 0 \\
0 & 1/4 & 1/2 & 1/4 & 0 \\
0 & 0 & 1/4 & 1/2 & 1/4 \\
0 & 0 & 0 & 1/4 & 3/4 \\
\end{bmatrix}
\]

Assume that all individuals start out at time \( t = 0 \) in health state 3. We are interested in the individuals in bad health at time \( t = 2 \), and their probabilities of transitioning back to the good health state, depending on whether they were in bad health for one or two periods.

It turns out that after two periods, 31.25\% of individuals are in bad health. 18.75\% have been in bad health for two periods, and two thirds of these are in health state 4 while one third are in health state 5. Hence, the probability of transitioning back to good health, conditional on having been in bad health for two periods, is 16.7\%.

However, the probability of transitioning back to good health for the individuals who have only been in bad health for one period is 25\% (this follows immediately since the unhealthy who were in good health in period \( t = 1 \) can, by construction, only be in health state 4 in this stylized example).

We define a measure of the recovery probability for age \( t \) depending on spell length \( j \) as:

\[
r_t(j) = \Pr(h_{t+1} \in G \mid h_{t-k} \in B \ \forall \ 0 \leq k < j, \ h_{t-j} = 3)
\]

where \( G = \{1, 2, 3\} \) and \( B = \{4, 5\} \). We choose 3 as the starting health state to simplify the interpretation of the results; the probability of transitioning to a bad health state from health state 1 or 2 is very close to zero, as shown in Figure 6 and Figure 7.

The results predicted by the full model for \( t \in \{60, 70, 80\} \) are shown in Figure 16. As can be seen, the probability of recovering from bad health decreases as a function of how many years the individual has spent in bad health. A 60-year old nonblack man who has spent just one year in bad health has a 22.5\% probability of recovering, but if he has spent the last five years in bad health the probability is down to 17\%.

Even though this is not as much duration dependence of bad health as found by De Nardi et al. (2017) in the PSID (they estimate that individuals in the age group 70+ who have spent more than 6 years in bad health have approximately a 12\% probability of recovering), it is a substantial improvement of the dynamics compared to a first-order Markov chain with two states, while being more parsimonious than a second-order specification.
CHAPTER 3

Figure 16: Probability of recovering from a bad health state, as a function of the number of years spent in bad health.

5. Life expectancy and education level

Our model permits the inclusion of time-invariant characteristics to be included in $x_t$, the vector of additional covariates. In this section, we include the level of education in both $g(\cdot)$ and $f(\cdot)$, fully interacted with age and health.\(^{17}\) We define education in three levels: 1) less than high school, 2) high school, or 3) (some) college or more.\(^{18}\)

As is well known, more highly educated individuals are healthier. Figure 17 shows the distribution of self-reported health for nonblack men for the three education groups. As can be seen, the fraction of individuals in bad health is larger the lower the education.

However, even conditional on self-reported health, the health and survival prospects are worse for low-educated individuals. Figure 18 shows the evaluation of probabilities for nonblack men for each health state and for being dead, given an initial health state and a starting age, contrasting the lowest and the highest education groups.\(^{19}\)

As shown in the figure, the less educated group is both more likely to transition to bad health, and bad health is more persistent for the less educated group (visibly shown by the fact that the red areas are larger in the graphs for the lower education level). For instance, a 50-year old man with less than high school education, starting

\(^{17}\)We also experimented with including a respondent’s marital status when first observed, since, e.g., Guner et al. (2018) report a health gap between married and unmarried individuals. However, the combination of education and married status makes each group too small to be able to draw any conclusions. Since marital status is not really time-invariant (due to divorce or death of partner) we choose to focus on the education level.

\(^{18}\)In this section, we present the results for nonblack men. Other demographic groups are available upon request.

\(^{19}\)Results for the second group, with high school education, are shown in the appendix, as well as the full health transition matrices and the survival probabilities by education level.
Figure 17: Distribution by health state among nonblack men for different ages. Red color indicates worst (“poor”) health state while dark green indicates best (“excellent”) health state. Observations are grouped into two-year bins.
Figure 18: Survival probability conditional on initial health state for nonblack men. The X-axis indicates years after initial age (upper row 50, lower row 70 years). The colors indicate probability per health state (dark green being the best health state, red the worst). The hatched area represents the probability of being dead.
out in poor health, has a probability of approximately 30% to survive an additional 30 years. However, for a 50-year old man with (some) college education, but in the same poor initial health, the probability is more than 40%. As another example, a 70-year old in good initial health condition and with less than high school education has a probability of approximately 60% to survive an additional 10 years. For a man of the same age and with the same initial health condition, but with (some) college education, the probability is 70%.

Hence, the life expectancy differs by education level. Figure 19 shows the life expectancy by education level, age and health state and, as can be seen, the life expectancy, conditional on health state, is higher for more highly educated people.

The average life expectancy for 50-year old nonblack men with less than high school education is 74.9 years, while the average for 50-year olds with (some) college education is 80.1 years. This difference is a result of two factors. First, the group with lower education has worse average self-reported health at the initial age of 50. Second, conditional on health, their health dynamics and survival probabilities are
worse from this age and onwards.

To disentangle these two effects we make the following experiment: we take the health and survival process of the lowest education group but use the initial health distribution of the highest education group. The average life expectancy for this hypothetical group of 50-year olds would be 76.0 years. Hence, the difference in life expectancy for high vs low educated 50-year old men is approximately one fifth due to worse overall initial health, and four fifths due to worse health dynamics after that age. Table 2 shows the full set of combinations of health and survival processes and initial health distributions. As can be seen, it is generally true that the health and survival process after the age of 50 has a larger effect on life expectancy than the health distribution at the age of 50.

6. Conclusion

Many studies have identified health dynamics and health shocks as a major source of risk faced by individuals. To incorporate this risk in life-cycle models, we need a health and survival process that captures the main features of the data, while still keeping the formulation parsimonious. In this paper, we provide improved estimates for annual age-dependent health transitions and survival probabilities for different subsamples of the US population.

The resulting Markov process can be used in any life-cycle model where the risks associated with health and survival are of interest. The estimated health process shows high persistence and the survival probability differs substantially depending on health. The estimated difference in expected longevity for a 50-year-old white man in excellent vs. poor health is 6 years.

The estimation method we propose is specifically designed to handle the peculiarities of the health and death data in the HRS. However, besides the specificities of the absorbing state of death, the method for estimating age-dependent one-year transition probabilities from data that is observed at irregular and overlapping intervals could be used outside the health and survival realm. It could be applied to any process where the dependent variable is ordinal (or limited in other ways).
References


CHAPTER 3

A. Health distribution

The health distribution by age for black women and black men is shown in Figure 20.

\textbf{Figure 20}: Distribution by health state for different ages. Red color indicates worst ("poor") health state while dark green indicates best ("excellent") health state. Observations are grouped into two-year bins.
B. Evaluation of probabilities for health state: education group 2

Evaluation of probabilities for each health state and for being dead, given an initial health state and a starting age, for nonblack men with high school education.

![Graph showing survival probability conditional on initial health state for nonblack men.](image)

**Figure 21**: Survival probability conditional on initial health state for nonblack men. The x-axis indicates years after initial age (upper row 50, lower row 70 years). The colors indicate probability per health state (dark green being the best health state, red the worst). The hatched area represents the probability of being dead.
C. Health transition matrices and survival probabilities by education level

Figure 22, Figure 23, and Figure 24 show the health transition matrices for the different education levels. Figure 25 shows the resulting survival probabilities, conditional on health state, age, and education level.

Figure 22: One-year health-to-health transition probabilities for nonblack men with less than high school (model estimates). Initial health states in rows, terminal health states in columns. Shaded areas indicate bootstrapped 95% confidence intervals.
Figure 23: One-year health-to-health transition probabilities for nonblack men with high school (model estimates). Initial health states in rows, terminal health states in columns. Shaded areas indicate bootstrapped 95% confidence intervals.
Figure 24: One-year health-to-health transition probabilities for nonblack men with (some) college (model estimates). Initial health states in rows, terminal health states in columns. Shaded areas indicate bootstrapped 95% confidence intervals.
Figure 25: One-year survival probability (model estimates) for nonblack men, by education level. Shaded areas indicate bootstrapped 95% confidence intervals.
Subjective life expectancies, time preference heterogeneity and wealth inequality

Richard Foltyn Jonna Olsson

1. Introduction

There is a substantial heterogeneity in life expectancy in the population: an average 70-year-old man in excellent health has a 75% chance of reaching his 80th birthday, while the corresponding probability for a man in poor health is just below 40%. According to standard economic theory, the healthy man should save more for the future, given the higher probability of living a longer life. The question we ask in this paper is how heterogeneity in life expectancy affects savings rates and ultimately wealth inequality – within a cohort and in the aggregate economy.

Preference heterogeneity, and especially heterogeneity in time preferences, is one of the potential sources of wealth inequality that have been used in previous literature, but preferences are difficult to measure and quantify.\(^1\) In this paper, we investigate one type of time preference heterogeneity, namely heterogeneity in life expectancy, which we document using micro data.

An individual’s consumption/savings decision is not necessarily guided by the objective statistical life expectancy, but rather by the individual’s beliefs about survival. We document new facts about a systematic bias in these beliefs: individuals with a low survival probability relative to their peers underestimate, while individuals with a high survival probability overestimate, their life expectancies. This systematic bias exacerbates the survival expectancy heterogeneity in the population.

To gauge the effect of survival heterogeneity on inequality, we use an overlapping-generations general-equilibrium model with uninsurable idiosyncratic shocks. Agents face heterogeneous survival risk that depends on their current health state, and are subject to health shocks that follow a process estimated from data. Besides this uncertainty, we also include standard persistent and transitory shocks to labor.

\(^1\)See De Nardi and Fella (2017) for further references.
productivity. Since we are interested in savings behavior in late life, it is important to capture other incomes during this period. Therefore we carefully model retirement benefits, closely mimicking the US social security system.

We compare several scenarios along two main dimensions: first, we vary the model environment according to how agents form expectations about survival, which can be uniform survival beliefs (without health heterogeneity), objective survival beliefs, or subjective survival beliefs. In the latter two, survival beliefs are allowed to depend on an individual’s health. As for the second dimension, we examine economies in which agents either do or do not have a warm-glow bequest motive, since this turns out to be central to household behavior. Together, these dimensions give rise to six different model environments, where the model without survival belief heterogeneity or a bequest motive serves as the benchmark case.

We show that the standard life-cycle model gives rise to counter-factual implications when introducing survival heterogeneity. In an environment without bequests, agents with a longer life expectancy save more, as expected. This is in line with the data, where individuals in better health have larger asset holdings. The effect on within-cohort inequality at higher ages is substantial, but the overall effect on wealth inequality in the economy is small. The reason is that the richest individuals are those of ages 60–64. At those relatively younger ages, the rich, who are also the healthiest, have a life expectancy which is not too far from the average and therefore their savings rate is only marginally affected. Moreover, their subjective beliefs are close to the objective survival probabilities, and therefore the additional effect of subjective beliefs is small.

However, as is well known, this standard model without a bequest motive gives rise to counterfactually low savings among the elderly since agents draw down their assets to virtually zero late in life.\(^2\) In the data, on the other hand, individuals do, on average, have substantial asset holdings even beyond the age of 80. Therefore, we add a warm-glow bequest motive as in De Nardi (2004), which is the most commonly used formulation in the macroeconomic literature.

We calibrate the bequest parameters so that the benchmark scenario without survival heterogeneity approximately matches the median asset holdings at older ages observed in the data. The effect of introducing survival heterogeneity into this environment is counter-intuitive and perhaps also unexpected: agents in poor health now save more than their healthy counterparts. The reason is as follows: since agents in poor health are more likely to die soon, they put an increased weight on bequest utility, thus raising their incentives to save. Hence, there are two effects from lower life expectancy that work in opposite directions: a shorter expected life span makes the agent save less for own consumption, but a stronger bequest motive makes the agent save more. The net effect varies depending on the calibration of

\(^2\)In a model without retirement benefits, agents keep slightly more assets to be able to sustain consumption in the case of an unexpectedly long life. However, even then, asset holdings are far lower than what we observe in the data.
bequest parameters, but the second mechanism is always present with a bequest formulation of this type: a shorter life span makes agents want to save more to leave bequests. This creates a health-wealth gradient that is counter-factual and we argue that this mechanism is implausible.

We conclude that none of the standard models are adequate for investigating the effect of survival heterogeneity on savings rates and wealth inequality. We discuss possible extensions and reformulations and point out directions for further research.

This paper speaks to three broad strands of literature. The first is macroeconomic studies pointing out the importance of heterogeneity in time preferences to explain wealth inequality (Hendricks 2007; Krusell and Smith 1998). Compared to these studies, we analyze the importance of a preference heterogeneity that is micro-founded by health heterogeneity.

The second is the literature about the general impact of health on wealth (Coile and Milligan 2009; De Nardi et al. 2017; Lee and Kim 2008; Poterba et al. 2017; Smith 1999). These studies incorporate multiple links between health and economic outcomes, while we restrict ourselves to a very specific channel, namely survival heterogeneity. We argue that this channel is interesting in itself, and that it is necessary to understand it correctly in order to include it in the broader assessment of the health-wealth gradient.

The third is the literature concerned with subjective survival expectations (Groneck et al. 2016, 2017; Hamermesh 1985; Heimer et al. 2018; Hurd and McGarry 2002; Ludwig and Zimper 2013; Smith et al. 2001). Many studies have documented the existence of an age bias in subjective life expectancies, and a few of the papers within this group are concerned with the implications for the consumption/savings behavior. However, none have documented the within-cohort bias, or analyzed the implications for wealth inequality.

In the next section, we briefly describe how we estimate the health and death process and give details about the systematic bias in survival expectations in our sample. Section three describes the model we use to quantify the importance of the heterogeneity in survival expectations. After that, we discuss the parametrization and present our results. The last section concludes the paper and comments on the implications of these findings, and points out directions for further research.

2. Empirical evidence

2.1. Data

We use the Health and Retirement Study (HRS), a representative panel of elderly US households, to investigate the evolution of health and longevity in the later stages of
life. The survey includes, among other things, questions about self-reported health, expectations about survival, and date of death, if applicable.

The survey started in 1992 and in this paper, we use HRS data up to and including the eleventh wave in the year 2012. The first cohort included in the survey was between 51 and 61 years old in 1992, and thereafter new cohorts have been added. Many of the respondents have died over the sample period, making it an appropriate data set for studying survival.

2.2. The health-wealth gradient

Figure 1 shows net total wealth over the life cycle, by self-reported health status. The health-wealth gradient is well documented, but the underlying causal relationship is debated. One line of argument, especially common in the medical discipline, is that low economic status leads to poor health. The underlying reasons could be many: poor people have access to less or lower quality medical care, do not invest enough in preventive health measures, and/or have more health-deteriorating habits.

---

3 The RAND version O, covering waves up until 2012, is the most recent RAND release that includes data from the National Death Index (NDI). Since correct death dates are crucial for our analysis, this is our preferred data set. There is one later RAND release, covering the 2014 wave as well, but there NDI data is lacking. An analysis shows that there are discrepancies in death dates between the exit interview information and the NDI date of death.
However, there are also many arguments for the reversed causality: poor health has economic consequences in itself. First, poor health may restrict the individual’s earnings potential by making it more costly to work and/or lowering the wage. Second, poor health may lead to large medical expenditures. Third, poor health may lower the savings incentives due to a lower survival expectancy.

In this paper, we focus on the savings incentives channel. Our aim is to answer the question: how much of the difference in wealth accumulation between individuals in good health and bad health is due to their (perceived) difference in longevity? We go beyond the objective survival expectations and also look at how the consumption/savings decision is affected by an individual’s subjective survival expectations.

There are a few empirical studies that corroborate the existence of this channel and suggest a causal link. For instance, Heimer et al. (2018) show that greater survival optimism correlates with higher savings rates, not only controlling for standard demographic characteristics such education and marital status, as well as income, but also characteristics such as financial literacy and risk tolerance. Hurd et al. (2004) show that individuals with very low subjective survival probabilities retire and claim social security benefits earlier.

In this paper, we examine the isolated effect of heterogeneity in survival expectancy on savings behavior and its implications for wealth inequality. Therefore, we need to formulate heterogeneity in survival expectations, both objective and subjective, in a way that can be used in a structural model. This is what we turn to in the next section.

2.3. Objective health and survival probabilities

The HRS contains self-reported health states and a respondent’s death date, if applicable. For our model, we need a yearly Markov process for health transitions and survival as a function of the model’s state variables age and health. We estimate this Markov process as described in Foltyn and Olsson (2019). Conceptually, the method is a straightforward maximum likelihood estimator, where the probability of observing the transition paths in the data is maximized.

To put structure on the Markov process, we follow Pijoan-Mas and Rios-Rull (2014) and use a nested logit model, where survival and health transitions conditional on survival are modeled as functions of the current health state and age. The probability of survival follows the usual binary-outcome logit model while, conditional on survival, health transitions are modeled using multinomial logit. Thus, it is assumed that the one-period-ahead probability of survival is given by

\[
p_{t+1}^s = \frac{1}{1 + e^{-g(h_t, x_t | \gamma)}}
\]  (1)
Figure 2: Survival probability conditional on the initial health state. The X-axis indicates years after initial age (upper row 50, lower row 70 years). The colors indicate probability per health state (dark green being the best health state, red the worst). The hatched area represents the probability of being dead.

where $h_t$ is the current health state and the vector $x_t$ contains any other variable of interest, in particular age.

Survival probabilities are governed by the parameter vector $\gamma$ to be estimated (together with $\beta$, a vector governing the health transition probabilities). The result of the estimation of the objective health and survival probabilities is a first-order Markov process for health and the absorbing state of death. This process can be used to calculate objective life expectancies, conditional on age, health, race, and gender. Similarly, we estimate a new set of “subjective parameters” $\tilde{\gamma}$ that govern subjective beliefs about survival, which we discuss in detail in the next section.

Figure 2 illustrates the dynamics of the estimated health and survival process. The figure shows the evaluation of probabilities of each health state and of being dead along a 30-year forecast horizon for a given initial health and age. As can be seen, the survival probability differs substantially depending on the initial health state: for a nonblack man aged 70 in excellent health, the predicted probability of surviving an additional 10 years is more than 80%, while the probability is just around 40% if instead starting out in poor health.
CHAPTER 4

Figure 3: Objective vs. subjective survival probabilities as a function of age. The number next to the black line indicates target age. The green line shows average expectation among the population. Shaded areas indicate bootstrapped 95% confidence intervals.

2.4. Average expectation errors in survival probabilities

In the expectations survey module of the HRS, respondents are asked about the probability they assign to certain events. One of these questions is about the probability of surviving to a certain age, for example: “Using a number from 0 to 100, what do you think are the chances that you will live to be at least 100 years?”

The target age the respondent is asked about depends on his/her age. For instance, in 1995, respondents below the age of 70 were asked about the probability of living until the age of 80, while respondents above the age of 85 were asked about the target age of 100. In later surveys and for some age intervals, the respondent is asked about several target ages.

Using these elicited beliefs, we compare the average probability that individuals of a certain age assign to survival until a given target age to the probability according to official lifetables. The results are shown in Figure 3. As can be seen, there is a systematic error along the age gradient: younger individuals on average tend to underestimate, while older individuals tend to overestimate, their survival probability as compared to objective lifetable estimates.

This age-dependent error is a stylized fact in the literature about survival expectations (see, e.g., Groneck et al. (2016); Heimer et al. (2018); Ludwig and Zimper (2013) and the references therein). The pattern has been used, e.g., to improve

---

4Before the respondent answers the questions about expectations, the interviewer discusses probabilities and verifies that the respondent understands the concept.
the fit of the asset profile of the canonical life-cycle model with the data: due to underestimation early in life, young agents do not accumulate as much assets, while overestimation in later years dampens the rate at which assets are decumulated.

In the previous section, we did not only estimate the survival probability as a function of age and gender, but also as a function of health and race, which allows us to disaggregate the expectation errors further. Figure 4 shows expectation errors by gender, race, age, and health among respondents who have answered the question about their perceived probability of survival until the age of 75. The first observation is the striking positive correlation between subjective self-reported survival probability and the predicted (objective) survival probability, which means that subjective beliefs are informative and not just random noise. The second observation is the systematic bias in beliefs.

As was shown in Figure 3, on average individuals underestimate their probability of survival until the age of 75. From looking at Figure 4 and focusing on nonblack males, it becomes clear that it is the individuals in bad health that are underestimating their survival probability, while the individuals in excellent health are on average reasonably close to their objective survival probability.

Figure 5 and Figure 6 show the same information for target ages 85 and 95, respectively. Again, we see that on average, individuals in bad health are more pessimistic than those in good health. For example, as we saw in Figure 3, on average individuals overestimate their survival probability when they are asked about a target age of 95. If we look at Figure 6, and again focus on nonblack males, we see that individuals in bad health are not that far off from their objective probability, while individuals in good or excellent health are severely overestimating their survival probability and driving up the average.

To summarize, we want to stress two observations: first, subjective beliefs are informative and correlated with objective probabilities. Second, subjective beliefs are biased. Subjective probabilities overestimate the health/survival gradient, with individuals in bad health underestimating their survival probability relative to individuals in good health. Hence, there is both a systematic error along the age gradient and, within cohort, along the health gradient.

2.5. Estimation of the subjective life expectancy process

In this section, we use the health transitions and survival probabilities estimated in Foltyn and Olsson (2019) as a basis for estimating a different set of survival parameters that govern subjective survival beliefs. We take as given the parameters

---

5This confirms the findings by Smith et al. (2001) who compare subjective expectations to actual longevity on an individual basis (with a much smaller sample) and find that expectations and actual survival are consistent, and moreover that expectations are updated in the event of bad health shocks.
Figure 4: Objective vs. subjective survival probabilities for target age 75. Each dot represents the average for a gender/race/age/health group. The x-axis shows the average self-reported survival probability for that group, the y-axis the predicted (objective) survival probability. The color of the dot indicates health status, with red being poor health and green being excellent health.
Figure 5: Objective vs. subjective survival probabilities for target age 85. Each dot represents the average for a gender/race/age/health group. The x-axis shows the average self-reported survival probability for that group, the y-axis the predicted (objective) survival probability. The color of the dot indicates health status, with red being poor health and green being excellent health.
Figure 6: Objective vs. subjective survival probabilities for target age 95. Each dot represents the average for a gender/race/age/health group. The x-axis shows the average self-reported survival probability for that group, the y-axis the predicted (objective) survival probability. The color of the dot indicates health status, with red being poor health and green being excellent health.
controlling health-to-health transitions conditional on survival since the HRS does not elicit any beliefs about future health states.

As explained above, the underlying data for this exercise takes the following form: HRS respondents are asked at date $t$ to state their beliefs about surviving to a certain target age $\bar{a}$ (for example 75 or 85), which we reinterpret as the probability of being alive in period $T$, with $T = t + (\bar{a} - a_t)$. Thus, an observation $i$ is given by the tuple $(p_i, h_i, x_i, T_i)$ where $h_i$ denotes current health state, $x_i$ is a vector of covariates including age and $p_i$ is the subjective survival belief. We treat multiple observations from one individual independently: say a respondent $\ell$ is surveyed on survival beliefs to horizons $T^1_\ell$ and $T^2_\ell$ in calendar years $t_1$ and $t_2$. This gives rise to the data

$$(p_{\ell t_1}, h_{\ell t_1}, x_{\ell t_1}, T^1_\ell)$$

$$(p_{\ell t_2}, h_{\ell t_2}, x_{\ell t_2}, T^2_\ell)$$

which we treat as four independent observations (except when bootstrapping confidence intervals, which we cluster at the individual level).

We are interested in estimating a parameter vector $\gamma$ that captures these subjective beliefs conditional on $(h, x, T)$. Assume that the $i$-th individual forms $T$-year-ahead survival beliefs based on the model

$$p^*_i = \phi_T(h_{it}, x_{it}, z_{it})$$

(2)

where $\phi_T$ is an unknown nonlinear function that maps $(h, x, z)$ into $[0, 1]$. The respondent’s belief is allowed to depend on a vector of additional covariates $z$ that are either unobserved or not included in our postulated model of survival.

In what follows we partition the sample into groups indexed by $g$, such that each unique combination of $(h, x, T)$ forms a separate group. Denote by $\Gamma_g$ all individual/year observations that satisfy

$$\Gamma_g = \{(i, t) \mid h_{it} = h_g, x_{it} = x_g, T_{it} = T_g\}$$

i.e., all observations where the individuals are of the same age and have the same covariates, are in the same health state, and state their beliefs about survival to the same target age. Denote by $\overline{p}^*_{\ell g}$ the (weighted) sample average of reported survival beliefs conditional on $(h_g, x_g, T_g)$, i.e.

$$\overline{p}^*_{\ell g} = \frac{\sum_{(i,t) \in \Gamma_g} w_{it} \times \phi_T(h_g, x_g, z_{it})}{\sum_{(i,t) \in \Gamma_g} w_{it}}$$

(3)
with \( w_{it} \) denoting sampling weights.

Now consider the nested-logit model counterpart of (3), which we denote by

\[
\tilde{p}_g^s = \Pr \left( s_{T_g} = 1 \mid h = h_g, x_g, T_g, \gamma \right)
\]

i.e. the predicted probability of being alive for group \( g \) which is parametrized by the vector \( \gamma \). The observed sample moment for each group can then be written as

\[
\bar{p}_g^s = \tilde{p}_g^s + u_g
\]

where \( u_g \) is the residual that is not explained by our model. At this point, we are not imposing any restrictions on the residual. In particular, we are not postulating that this is an average of group-\( g \) individuals’ forecast errors, since the individual forms beliefs according to (2). Our aim is to minimize these group-specific residuals using the least-squares objective function

\[
J(\gamma) = \frac{1}{W} \sum_{g=1}^{N_G} W_g \left( \bar{p}_g^s - \tilde{p}_g^s (h_g, x_g, T_g \mid \gamma) \right)^2
\]

(4)

where \( W_g = \sum_{(i,t) \in \Gamma_g} w_{it} \) is the sum of weights in group \( g \). The estimated vector \( \hat{\gamma} \) is hence the arg min of \( J(\gamma) \).

The resulting survival probabilities for nonblack men are shown in Figure 7 on the right, juxtaposing the objective survival probabilities estimated in Foltyn and Olsson (2019) on the left.\(^6\) As can be seen, the subjective belief about survival while in health state “excellent” or “very good” is 100%. This does not mean that individuals in one of those health states believe that they will live forever, but rather that death is preceded by a deterioration in health.

Note that ex ante, it is not obvious that a dynamically consistent belief system exists that can match the elicited subjective beliefs.\(^7\) It is easy to imagine subjective beliefs that are contradictory: for instance, if 50-year-olds say that the probability of surviving until the age of 75 is 80%, and the probability of surviving until the age of 85 is 40%, this implies that the probability of surviving until 85, \textit{conditional on turning 75}, is 50%. If then the 75-year-olds said that their probability of surviving until the age of 85 is 80%, we would have a belief system that is dynamically inconsistent.

However, it turns out that the subjective beliefs can actually be mapped into a dynamically consistent set of survival beliefs conditional on age and health. In

\(^6\)Results for other demographic groups are available upon request. For the model in the next section, we will use the estimates for nonblack men.

\(^7\)As an alternative to dynamically consistent beliefs, one could for instance model survival beliefs as the result from a Bayesian learning process, as in Groneck et al. (2016), who use a model of Choquet Bayesian learning of survival beliefs, allowing for likelihood insensitivity. However, since our estimated subjective belief system matches the elicited beliefs well, we think this is appropriate for the purposes of this paper.
(a) Objective probabilities

(b) Subjective probabilities

Figure 7: One-year survival probability (model estimates). Shaded areas indicate bootstrapped 95% confidence intervals. For each bootstrapped sample we re-estimate the objective health process.
Figure 8 we plot the model-predicted subjective survival against elicited beliefs. As can be seen, the estimated model for subjective beliefs captures the main picture, since the dots, each representing an age/health/target-age group, line up reasonably close to the 45 degree line.

Figure 9 summarizes the results, showing the life expectancy by age and health state using the objective and the subjective survival process. As can be seen, at all ages, the difference in life expectancy between the best and the worst health state is larger when using subjective life expectancies. The divergence between objective and subjective life expectancies is particularly large for individuals in bad health states, who substantially underestimate survival at younger ages. Conversely, individuals in all health states overestimate their chances of survival late in life.

3. Model

In this section, we describe the overlapping generations model we use for our analysis. Time is discrete and every time period is assumed to be one year. Agents derive utility from consumption and face three types of idiosyncratic risks: shocks to persistent productivity, transitory productivity shocks and shocks to health and survival. Agents can only save in a riskless bond, and they face an exogenous borrowing constraint.

3.1. The agent’s problem

There is a unit mass of individuals distributed across $N_t$ cohorts according to the ergodic distribution implied by the transition matrix of survival probabilities. An individual of age $t \in \{1, \ldots, N_t\}$ and health $h \in \{1, \ldots, N_h\}$ has a one-period survival probability to age $t + 1$ given by $\pi_{th}^s$, with $\pi_{th}^s = 0$ for $t = N_t$ regardless of health state.

Individuals are assumed to be in the labor force for the first $T_R - 1$ years of their life and exogenously retire in the period when they attain age $T_R$. While in the labor force, they are hit by persistent and transitory labor productivity shocks. During retirement, individuals receive social-security retirement benefits which depend on their last persistent labor state in working age.

Bequests are distributed to new-born individuals in a dynastic way: one young individual receives the bequests from one dying individual.
Figure 8: Elicited beliefs about survival vs. estimated subjective survival probabilities. Each dot represents the average for an age/health group. The x-axis shows the average self-reported survival probability for that group and the y-axis the predicted survival probability according to the subjective model. The color of each dot indicates health status, with red being poor health and green being excellent health.
Figure 9: Life expectancy by age and health state. The color indicates health state: green is excellent while red is poor health. In the left graph, showing the life expectancy using objective probabilities, the black line indicates the average in the population, weighted by the distribution over health states.

Retired agents. A retired individual with idiosyncratic states \((a, p, h, t)\), where \(a\) is cash-at-hand, \(p\) is the (fixed) persistent component of labor earnings, \(h\) is the current health state, and \(t\) is the age, maximizes utility according to

\[
V_R (a, p, h, t) = \max_{c, b'} \left\{ u(c) + \beta \mathbb{E} \left[ V_R (x') \mid h, t \right] + (1 - \pi^h) V_b (a'_b) \right\} \tag{5}
\]

subject to the constraints

\[
\begin{align*}
a & \geq c + b', \quad c \geq 0, \quad b' \geq 0 \\
a' &= R b' + \iota'_R \\
\iota'_R &= y'_R w - T_R (y'_R w) \\
y'_R &= \omega_T (p) \bar{\epsilon} \\
a'_b &= R b' - T_b (R b')
\end{align*}
\]

where \(x' = (a', p, h', t + 1)\) is the continuation state conditional on survival. Next-period after-tax retirement income is denoted by \(\iota'_R\) and depends on the non-linear tax schedule \(T_R (\bullet)\). \(p_R (\bullet)\) is a function mimicking the regressive replacement rate of the US social security system, \(w\) is the economy-wide wage rate, \(\bar{\epsilon}\) is the average of the transitory earnings shocks hitting the working-age population and \(\omega_T\) is the value of the deterministic age profile of earnings just prior to retirement. Bequests are subject to a potentially non-linear tax schedule \(T_b (\bullet)\). The gross return on savings is given by \(R = 1 + r - \delta_k\).
Specifically, in the terminal period with $t = N_t$, the individual solves
\[
V_R (a, p, h, t) = \max_{c, b'} \left\{ u(c) + \beta V_b \left( a'_b \right) \right\}
\]
\[
\text{s.t. } a \geq c + b', \quad c \geq 0, \quad b' \geq 0
\]
\[
a'_b = R b' - T_b \left( R b' \right)
\]

Working-age households. Non-retired individuals are assumed to be in the labor force, so we use the subscript $LF$ to denote their value and policy functions. The vector of idiosyncratic state variables is $x = (a, p, h, t)$.

A working-age individual draws a persistent and a transitory labor shock component that together with a deterministic age profile of earnings, $\omega_t$, determine his labor productivity in the given period. The persistent component $p$ takes on the values from the set $P$, while the transitory shock realizations $\epsilon$ are drawn from $\mathcal{E}$.

A working-age individual with $t < T_R - 1$ who will continue to be in the labor force next period solves
\[
V_{LF} (x) = \max_{c, b'} \left\{ u(c) + \beta \left( \pi_{th}^s \mathbb{E} \left[ V_R (x') \big| p, h, t \right] + (1 - \pi_{th}^s) V_b \left( a'_b \right) \right) \right\}
\]
\[
\text{s.t. } a \geq c + b', \quad c \geq 0, \quad b' \geq 0
\]
\[
a'_b = R b' - T_b \left( R b' \right)
\]
\[
a' = R b' + \iota'
\]
\[
\iota' = \left[ y' - T_{ss} (y') \right] w - T_y \left( \left[ y' - T_{ss} (y') \right] w \right)
\]
\[
y' = \omega_{t+1} p' \epsilon'
\]
where $x' = (a', p', h', t + 1)$. We denote the earnings of working individuals, net of income taxes $T_y (\bullet)$ and payroll taxes $T_{ss} (\bullet)$, as $\iota'$.

In the final period of their working life, i.e. when $t = T_R - 1$, individuals in the labor force solve
\[
V_{LF} (a, p, h, t) = \max_{c, b'} \left\{ u(c) + \beta \left( \pi_{th}^s \mathbb{E} \left[ V_R (a', p, h', t + 1) \big| h, t \right] + (1 - \pi_{th}^s) V_b \left( a'_b \right) \right) \right\}
\]
subject to
\[
a \geq c + b', \quad c \geq 0, \quad b' \geq 0
\]
\[
a'_b = R b' - T_b \left( R b' \right)
\]
\[
a' = R b' + \iota'_R
\]
\[
\iota'_R = y'_R w - T_y \left( y'_R w \right)
\]
\[
y'_R = \omega_{T_R - 1} p_R (p) \epsilon'
\]
which is identical to the retired individual’s problem.

3.2. Technology

The production side of the model is standard. Competitive firms employ labor and capital hired from households to produce a homogeneous final good, which is used for both consumption and investment. The aggregate production function is assumed to be Cobb-Douglas:

\[ F(K, L) = K^{\alpha_k} L^{1-\alpha_k}. \]

Capital depreciates at the rate \( \delta_k \).

3.3. Government

We assume that the government runs a PAYGO social security system that has to balance in each period, and that remaining (wasteful) government expenditures have to be fully financed by estate and income taxes. We first describe the social security system and thereafter the general government budget.

3.3.1. Social security system

We use a stylized version of the actual retirement income formula used in the US social security system. It captures the main features, such as a regressive replacement rate based on pre-retirement income and a cap for maximum benefits. In the model, we define retirement benefits to be a product of the economy-wide wage rate, the life-cycle profile wage component from the last year before retiring, the average transitory component, and a function that mimics the regressive replacement rate of the US social security system:

\[ \iota_R(p) = w \times y_R(p) = w \times \omega_{T_R-1} \bar{e} p_R(p) \]

where the replacement function \( p_R(\cdot) \) is given by

\[
p_R(p) = \begin{cases} 
\rho_1 p & \text{if } p \leq \bar{p}_1 \\
\rho_1 \bar{p}_1 + \rho_2 (p - \bar{p}_1) & \text{if } \bar{p}_1 < p \leq \bar{p}_2 \\
\rho_1 \bar{p}_1 + \rho_2 (\bar{p}_2 - \bar{p}_1) + \rho_3 \left( \min \left\{ \bar{p}_{\text{max}}, p \right\} - \bar{p}_2 \right) & \text{else}
\end{cases}
\]

where \( \bar{p}_1 \) and \( \bar{p}_2 \) are bend points and \( \bar{p}_{\text{max}} \) the contribution and benefit base (CBB) in the social security income formula, expressed in terms of the individual’s permanent labor state. In the appendix, section A.2, the transformation between actual replacement rate bend points and CBB expressed in USD (\( b^1, b^2, \) and \( e^\text{max} \)) and
CHAPTER 4

the model counterparts are described in detail, as well as the derivation of total government expenditures on retirement, $G_{ss}$.

The government expenditures on retirement are financed by a payroll tax. The payroll tax function is defined as

$$T_{ss}(y) = \tau_{ss} \times \min\{y_{\max}, y\}$$

where $y_{\max}$ expresses maximum taxable earnings in terms of labor productivity, i.e.

$$y_{\max} = \left( \frac{e_{\max}}{e_{med}} \right) y_{med}$$

where $e_{\max}$ is the contribution and benefit base expressed in USD as above, and $e_{med}$ are the median earnings in the reference year.

The derivation for total payroll taxes raised in each period, $T_{ss}$, can be found in the appendix, section A.3. To balance the social security system, we need to find $\tau_{ss}$ such that $G_{ss} = T_{ss}$.

3.3.2. Government budget

We assume that the government raises estate taxes and labor income taxes to finance non-discretionary expenditures that amount to a constant fraction $g$ of output. In the following section, we will first describe the estate tax, thereafter the income tax, and finally how the government budget is balanced.

**Estate taxes.** The estate tax schedule is defined as

$$T_b(b) = \begin{cases} 
0 & \text{if } b \leq \chi_b \\
\frac{\tau_b}{2} \left[ \sin \left( \pi \left[ \frac{b-\chi_b}{B} - 1 \right] \right) \frac{B}{\pi} + b - \chi_b \right] & \text{if } \chi_b < b \leq \chi_b + B \\
\tau_b (b - \chi_b - B) + \frac{\tau_b}{2} B & \text{else}
\end{cases}$$

This formulation is effectively a step function so that estates valued at less than $\chi_b$ are exempt from taxes. For values $b$ in an interval $\chi_b < b \leq \chi_b + B$, the marginal tax rate is increasing, and for $b > \chi_b + B$, the marginal tax rate is $\tau_b$. The tax schedule is twice continuously differentiable, which is required by our solution algorithm. Details about the estate tax are given in the appendix, section A.4.
Income taxes. For income taxes, we adopt the same tax function as in Heathcote et al. (2017), which is defined as

\[ T_y(\iota) = \iota - \lambda \iota^{1-\tau} \tag{7} \]

where \( \iota \) is either earnings (net of payroll taxes) or retirement income. We assume that the progressivity parameter \( \tau \) is fixed, and we pin down \( \lambda \) such that the government budget is balanced in each period.

Total income taxes raised by the government amount to the sum of income taxes raised from the employed and from the retired individuals:

\[ T_{inc} = T_e + T_R \]

The derivation of the total income tax raised by the government in each period can be found in the appendix, section A.5.

Government budget balance. The non-discretionary expenditures amount to a constant fraction \( g \) of output: \( G = gY \). The government budget is balanced by solving for a \( \lambda \) in (7) such that \( G = T_{inc} + T_b \) holds.

3.4. Equilibrium definition

A recursive competitive equilibrium is given by a set of prices \( \{ R, w \} \), tax rates \( \{ \tau_{ss}, \lambda \} \), decision rules \( C(a,p,h,t) \) (for consumption) and \( B(a,p,h,t) \) (for savings), and a stationary distribution \( \Gamma \) such that:

1. The decision rules solve the agents’ problem for all \( (a,p,h,t) \).
2. Factor prices are given by:
   \[ r = F_1(K,L) \quad \text{and} \quad w = F_2(K,L) \]
3. \( \tau_{ss} \) and \( \lambda \) are set so that both the social security system and the general government budget balance.
4. Capital and labor markets clear:
   \[ K' = \int B(a,p,h,t) \, d\Gamma \quad \text{and} \quad L = \sum_{t=1}^{T} \sum_p \sum_{\epsilon} \mu_t(t) \mu_p(p) \mu_\epsilon(\epsilon) \omega_t p \epsilon \]
   where \( \mu_t(t) \), \( \mu_p(p) \) and \( \mu_\epsilon(\epsilon) \) are the ergodic distributions over age, the persistent and the transitory labor shocks, respectively.
5. The distribution \( \Gamma \) is stationary, i.e., for all relevant Borel sets \( B \)
   \[ \Gamma(B,p,h,t) = \sum_p \sum_{\bar{h}} \sum_{\bar{t}} \pi(p|\bar{p})\pi(h|\bar{h})\pi(t|\bar{t}) \int_{a:B(a,p,h,t) \in B} \Gamma(da,\bar{p},\bar{h},\bar{t}) \]
4. Calibration

In this section, we explain our strategy to calibrate the model. First, we parametrize the instantaneous utility function and the bequest motive. Then, we adopt a three-fold strategy. First, some parameters are taken directly from the literature. Second, parameters guiding the health and death process are directly estimated from the data as reported in section 2.3. Lastly, a set of parameters are estimated with the method of simulated moments.

4.1. Preferences

We assume that utility in each period is of the standard CRRA form with a risk-aversion parameter \( \sigma \) set to one.

In the macro literature, it is common to follow De Nardi (2004) and describe the bequest motive as

\[
V_b(a) = \theta_B \frac{(a + \kappa)^{1-\sigma} - 1}{1-\sigma}
\]

where \( \theta_B \) determines the strength of the bequest motive, and \( \kappa \) determines to what extent bequests are a luxury good. In the model environment where bequests are not operative we let \( \theta_B = 0 \). For the alternative scenario with a warm-glow bequest motive, we choose parameter values for \( \theta_B \) and \( \kappa \) to match the life-cycle asset profiles of the elderly as observed in the HRS.

4.2. Externally calibrated parameters

Demographics and the life cycle. Agents are assumed to enter the economy at age 20, and retire at the age of 65, hence \( T_R = 46 \). The maximum age an agent can reach is 99, hence \( N_t = 80 \).

Age-dependent wage profile and idiosyncratic earnings risk. We assume that the log-labor earnings of an individual in the labor force follow a process with transitory and persistent shocks:

\[
\log y_t = \log \omega_t + \log p_t + \log \epsilon_t
\]

where \( \omega_t \) is the age profile part, \( p_t \) is the persistent component and \( \epsilon_t \) is the transitory component of earnings. The persistent component is assumed to follow an AR(1) process specified as:

\[
\log p_t = \rho \log p_{t-1} + \eta_t
\]

with persistence \( \rho \) and innovation \( \eta_t \sim N(0, \sigma^2_{\eta}) \). The transitory shock is given by \( \log \epsilon_t \sim N(0, \sigma^2_{\epsilon}) \).
Hence, the stochastic part of the wage process is characterized by the parameters \((\rho, \eta^2, \epsilon^2)\) which we set to \((0.9695, 0.0384, 0.0522)\), following Krueger et al. (2016).

We use the Rouwenhorst procedure to discretize the persistent part of the process into an eleven-state Markov chain, and we discretize the transitory shock into three states.

We choose the deterministic age profile of earnings estimated for high-school graduates by Cocco et al. (2005), and renormalize it such that the average labor productivity is unity.

**Remaining externally calibrated parameters.** The remaining parameters that are set externally are listed in Table 1. The bend points and the contribution and benefit base are reported in US dollars, to facilitate the interpretation. Details for transforming them into values used in the model are found in the appendix, section A.2.

### Table 1: Calibrated parameters

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Description</th>
<th>Value</th>
<th>Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>(\alpha_k)</td>
<td>Capital share</td>
<td>36%</td>
<td>Standard value</td>
</tr>
<tr>
<td>(\delta_k)</td>
<td>Depreciation rate</td>
<td>9.6%</td>
<td>Standard value</td>
</tr>
<tr>
<td>(\rho_1)</td>
<td>Replacement rate bracket 1</td>
<td>90%</td>
<td>2013 SS rules</td>
</tr>
<tr>
<td>(\rho_2)</td>
<td>Replacement rate bracket 2</td>
<td>32%</td>
<td>2013 SS rules</td>
</tr>
<tr>
<td>(\rho_3)</td>
<td>Replacement rate bracket 3</td>
<td>15%</td>
<td>2013 SS rules</td>
</tr>
<tr>
<td>(b_{1s})</td>
<td>Bendpoint 1 (in USD)</td>
<td>9,492</td>
<td>2013 SS rules</td>
</tr>
<tr>
<td>(b_{2s})</td>
<td>Bendpoint 2 (in USD)</td>
<td>57,216</td>
<td>2013 SS rules</td>
</tr>
<tr>
<td>(\epsilon_{\text{max}})</td>
<td>CBB (in USD)</td>
<td>113,700</td>
<td>2013 SS rules</td>
</tr>
<tr>
<td>(g)</td>
<td>Gov. spending (share of GDP)</td>
<td>6%</td>
<td>Brinca et al. (2016)</td>
</tr>
<tr>
<td>(\tau)</td>
<td>Tax progressivity</td>
<td>0.137</td>
<td>Brinca et al. (2016)</td>
</tr>
<tr>
<td>(\tau_b)</td>
<td>Marginal tax on estates</td>
<td>30%</td>
<td>Authors’ approximation(^9)</td>
</tr>
</tbody>
</table>

4.3. Health and death process

We use the process for health transitions and death probabilities described in section 2.3. However, agents enter the model at the age of 20, but the health and death process are set up such that the mortality rate at age 20 is zero. This is consistent with the assumption that individuals enter the labor market at age 20, and we do not consider the health and death process before that age.

\(^8\)Note that Krueger et al. (2016) remove the age effect before estimating this process and hence, we can use this stochastic process on top of the age-dependent profile.

\(^9\)The top marginal tax rate today is 40% (see https://www.irs.gov/pub/irs-pdf/i706.pdf); however, not all taxable estates fall into the top category. We choose 30% as an approximation.
death process we estimated starts at the age of 50. Therefore, we make the following assumptions: everyone is born in health state “excellent”. Thereafter, we use the health transition matrix for the age of 50 to roll forward the population, assuming certain survival. At the age of 50, agents start facing a positive probability of death according to our estimated process. The resulting cohort sizes and distribution of health states are shown in Figure 10.

4.4. Estimated parameters

Given the parameters and processes described above, we estimate the remaining parameters with the method of simulated moments, i.e., we minimize the weighted sum of squared distances between targeted and simulated moments. For this exercise, we use the benchmark model where we let all agents have the same survival expectations.

**Target moments.** In the first model, with no bequest, the capital-to-output ratio is the only target moment used. We target a level of 3.0, by choosing the appropriate value of $\beta$, the discount factor. In the model with a bequest motive, we still target a capital-to-output ratio of 3.0, but we also match the life-cycle profile of assets. We use the median wealth levels at ages 55, 60, 65, 70, 75, 80 and 85. The capital-to-output ratio and the life-cycle profile are jointly matched by choosing the parameters for the discount factor ($\beta$), the bequest utility weight ($\theta_B$), the bequest utility shifter ($\kappa$), and the estate tax exemption ($\chi_b$).

**Estimation results.** The resulting parameter estimates are shown in Table 2. In the estimation of the model without bequest, the resulting capital-to-output ratio
is 3.0. In the estimation of the model with bequest, the resulting capital-to-output ratio is also 3.0, and the resulting asset holdings by age and their data counterparts are shown in Table 3. As can be seen, the resulting wealth level in the model for the age of 65, i.e., exactly when the model agents enter retirement, is slightly higher than what is observed in the data. The reason is the fixed retirement age. Since all agents retire at the exact same age, they will all accumulate assets until exactly that point, while in reality, people retire more gradually.

To assess the bequest motive, it is informative to look at non-targeted moments. The bequest-to-wealth ratio in the model is 1.9%. It is difficult to precisely measure this figure in the data, but according to Gale and Scholz (1994), using SCF data, it should be closer to 0.9%. However, according to others, 2% is “a conservative estimate”.10 Another measure is the fraction of deceased individuals who pay estate taxes. The model result is 0.74%. This figure has varied during the last 15 years: in 2004 it was 0.8% and in 2013 0.2%.11 Hence, the predictions from our model regarding these two measures seem to be reasonably in line with the data.

A last measure is estate tax revenue as a fraction of GDP. In the model, this figure is 0.02%. In the data, this figure has varied during the last few years, from a high of 0.17% in 2007 to a low of 0.07% in 2011.12 Hence, for this value, our model prediction is at the lower end; however, this figure is to a large extent a result of the upper tail of the asset distribution, something that we know that our model does not fully capture.

5. Results

As mentioned above, we solve the model under three distinct assumptions about the survival beliefs:

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Description</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model without bequest</td>
<td>β Discount factor</td>
<td>0.981</td>
</tr>
<tr>
<td>Model without bequest</td>
<td>β Discount factor</td>
<td>0.961</td>
</tr>
<tr>
<td>θ_B</td>
<td>Bequest weight</td>
<td>10.33</td>
</tr>
<tr>
<td>κ</td>
<td>Bequest shifter</td>
<td>1.98</td>
</tr>
<tr>
<td>χ_b</td>
<td>Estate tax exemption</td>
<td>27.32</td>
</tr>
</tbody>
</table>

Table 2: Estimated parameters

---

11 Calculated as the ratio of the number of estates subject to estate taxes, as reported by the IRS, and the number of deaths taken from CDC records.
12 Calculated as the ratio of net estate taxes paid, as reported by the IRS, and official GDP figures.
Table 3: Life-cycle asset profile in the model with a bequest motive, relative to the median asset holdings at age 50.

1. No survival heterogeneity (NSH). All individuals face the same average survival expectations.
2. Objective survival heterogeneity (OSH). Individuals have heterogeneous objective survival expectations.
3. Subjective survival heterogeneity (SSH). Individuals have heterogeneous subjective survival expectations.

The first scenarios serve as our baseline: this is the standard model where all agents face the same survival risk and hence, all agents have the same effective discount factor, conditional on age. For this scenario, we collapse the health states into only one, and use the average survival rates.

In the second scenario, we use the objective process for health transitions and survival probabilities described in section 2.3. Hence, in this model, people are perfectly informed about their true survival probability conditional on health and age.

In the third scenario, agents believe and act according to the subjective survival process we estimated in section 2.5. However, this subjective process does not correspond to the true survival process, which is what we use when simulating the model.

The section is structured as follows. We will first analyze a model without a bequest motive. We discuss how the solution to the agent’s problem changes between the three scenarios (with no survival heterogeneity, objective survival heterogeneity, and subjective survival heterogeneity), and then aggregate the changed behavior by the individuals into economy-wide general equilibrium effects.

Thereafter, we repeat the analysis, but this time using a model with a bequest motive, and see how that affects the results.

5.1. Model without bequest motive

We start with a model in which the agents have no bequest motive and hence only save for their own consumption needs. The main effect from introducing health and survival heterogeneity is that agents in bad health save less than their healthy counterparts.
The reason why individuals in bad health save less is straightforward: they expect to live a shorter life. The same but opposite effect is present for individuals in good health: their life expectancy is longer, and therefore they save more. This is true regardless of whether agents compute their life expectancy using objective survival probabilities or subjective beliefs. However, the difference in the savings rate between agents in bad and good health increases if we let the agents act according to their subjective beliefs. The reason is the bias we documented in section 2.5: individuals in bad health underestimate their survival probability more (or, for some ages, do not overestimate to the same extent) as compared to agents in good health.

These patterns are documented in Figure 11, which shows the difference in the savings rate for the baseline model without any bequest motive. The bars show, by health state, the difference in the savings rate (in percentage points) for the model with objective heterogeneity (in darker color) compared to the baseline model with no heterogeneity in survival. The lighter color shows the additional effect of adding subjective beliefs. The horizontal axis represents the cash-at-hand percentiles for the age in question in the baseline model with no heterogeneity.\(^{13}\)

As can be seen, for 50-year-olds, the difference is most pronounced for individuals in bad health, as indicated by the red bars. With objective survival heterogeneity, individuals in the 10th cash-at-hand percentile in poor health have a savings rate that is 9 percentage points lower than individuals of the same wealth in the model with no heterogeneity. Adding subjective beliefs, the difference is magnified: the savings rate of an individual in the 10th percentile is now 22 percentage points lower than for an individual with the same wealth with average survival expectations.

The reason for the bigger impact of the subjective belief on the individuals in worst health can be found by looking at Figure 7. At lower ages, the subjective and the objective survival expectations are very similar for individuals in the better health states and therefore the additional effect of adding subjective beliefs is small. However, individuals in worse health are severely underestimating their longevity and thus the additional effect on the savings rate is larger.

The same pattern can be seen in the graphs for 60-year-olds and 70-year-olds, even though for the latter, the additional effect from adding the subjective beliefs is slightly weaker if we continue to focus on the individuals in poor health. The reason for this can be found in Figure 5: at the age of 70, individuals start taking into account the probabilities that are shown in this figure and, as can be seen, the subjective survival beliefs of the individuals in bad health are now closer to the objective probabilities.

\(^{13}\)The reason why we plot results by cash-at-hand percentiles (as opposed to the cash-at-hand level) is that the graphs then show the difference at the wealth levels that matter, i.e., where there is a positive mass of agents in equilibrium. A large difference in policy functions for very high wealth levels for the 90-year-olds is not that informative, since no 90-year-old will be that rich in equilibrium anyhow.
Figure 11: Difference in the savings rate for the model without bequest motive (percentage points/100, i.e., 0.1 indicates a difference of 10 pp). The darker color shows the model with objective survival heterogeneity compared to the baseline model with no survival heterogeneity. The lighter color shows the additional effect of adding subjective survival beliefs. The x-axis depicts cash-at-hand percentiles by age (equilibrium values for the baseline model). The color indicates health state: green is excellent while red is poor health.
Figure 12: Relative change in savings for the model **without bequest motive** (percent/100, i.e., 0.1 indicates an increase by 10 percent). The darker color shows the model with objective survival heterogeneity compared to the baseline model with no survival heterogeneity. The lighter color shows the additional effect of adding subjective survival beliefs. The x-axis depicts cash-at-hand percentiles by age (equilibrium values for the baseline model). The color indicates health state: green is excellent while red is poor health.
However, at the age of 70, the effect of adding subjective beliefs starts showing up for individuals in good health; in other words, there is an additional effect of the subjective beliefs on the savings rate for healthy individuals, as indicated by the lighter green parts of the green bars. As can be seen from the graph of 80-year-olds, adding subjective beliefs at this age mainly affects the individuals in good health.

Figure 11 shows the change in the savings rate in percentage points, but this might be misleading. For instance, at the age of 80, the savings rate is very low for the poorest individuals and hence even a substantial increase in relative terms looks like a small increase in percentage points. To better assess the magnitudes, we plot the relative changes in savings in Figure 12. The graph for 80-year-olds shows that the relative increase in savings for 80-year-olds in excellent health is large, even for the poorest individuals: it is an increase of 1200%. For the 90-year olds, the difference is enormous in relative terms.

Figure 13 shows the resulting life-cycle profiles for the three different scenarios: no survival heterogeneity (NSH), objective survival heterogeneity (OSH), and subjective survival heterogeneity (SSH). In all graphs, we integrate out the productivity dimension.

As expected, and as all three graphs show, the life cycle profile for wealth peaks at the age of 64, which is the last year before retirement. When the agents enter retirement, they start drawing down their wealth and the average individual who survives until the age of 90 has drawn down all of his savings (remember that all individuals receive retirement benefits, so they are not risking zero consumption).

The profile for the OSH model illustrates that individuals in bad health have less wealth than individuals in good or excellent health. It should be noted that the mass of individuals in bad health is not static, but consists of individuals who have been in bad health for many periods, as well as individuals who just recently draw a bad health shock.

Moreover, the profile for the SSH model shows that the difference in wealth between individuals in good and bad health at the age of 64 increases when the agents act according to their subjective beliefs. Note also that the asset holdings at very high ages increase slightly, since old individuals are, on average, over-optimistic about their lifespan. However, this slight increase is still far from what we see in the data, a fact that we will return to in section 5.2.

5.1.1. General equilibrium effects

In the previous section, we looked at the partial equilibrium responses from adding heterogeneity in survival expectations. We now move on to the effects on aggregate variables and the general equilibrium effects. The main takeaway is that overall, the
(a) No survival heterogeneity (NSH)

(b) Objective survival heterogeneity (OSH)

(c) Subjective survival heterogeneity (SSH)

Figure 13: Life-cycle profiles for wealth, model without bequest motive.
general equilibrium effects of introducing survival heterogeneity are small. Table 4 shows three key statistics from the model.

The reason why the interest rate increases slightly is the lower incentive to save on average among those in the ages 50 to 70. As Figure 11 shows, the demand for savings is substantially reduced at those ages among the individuals in bad health as compared to the baseline without heterogeneity. This effect is not counteracted by the slight increase in demand for savings by those in excellent health. Therefore, the interest rate needs to rise slightly to induce enough savings in the aggregate.

When agents act according to their subjective survival beliefs (SSH model), the interest rate increases slightly further, to 2.54%. The reason is that in the age group around 55–65, which is when people save the most, agents do, on average, have a downward bias in expected longevity. Hence, their incentives to save are reduced and consequently, the interest rate must increase to induce enough savings in the economy.

The next observation is that the average bequest is smaller in both the model with objective and the model with subjective survival beliefs, compared to the baseline model with no heterogeneity. Remember that in this model there is no bequest motive, so all bequests are accidental. Hence, with more information about the likelihood of their own demise, people can adjust their savings better and therefore the average accidental bequests are smaller in the model with objective survival heterogeneity.

When adding subjective survival heterogeneity, the accidental bequests are approximately of the same size as in the model with objective beliefs. The really old (aged 70+) are on average richer in the SSH model (due to their average over-optimism about survival), but the somewhat younger individuals (aged between 50 and 70) are somewhat poorer. These two effects cancel out and hence the average size of bequest is very similar in the two models with subjective and objective survival heterogeneity.

The last thing to note is that the wealth Gini is virtually unchanged between the three models. However, if we look at the within-cohort inequality, the picture is different. Figure 14 shows three statistics for wealth inequality, by age group: Gini, P90/P50 ratio, and P50/P10 ratio (the latter two in logs).

<table>
<thead>
<tr>
<th></th>
<th>Baseline: NSH (no heterogeneity)</th>
<th>OSH (objective exp.)</th>
<th>SSH (subjective exp.)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Interest rate</td>
<td>2.37%</td>
<td>2.38%</td>
<td>2.54%</td>
</tr>
<tr>
<td>Average bequest size</td>
<td>2.18</td>
<td>2.04</td>
<td>2.02</td>
</tr>
<tr>
<td>Wealth gini</td>
<td>0.65</td>
<td>0.65</td>
<td>0.65</td>
</tr>
</tbody>
</table>

Table 4: Comparing results from three scenarios (model with no bequest).
Figure 14: Wealth inequality in the model with no bequest motive, by 5-year age groups.
As can be seen, the difference is largest in the group of 90-94 year olds: within this age group, the Gini is 17pp higher in the model with subjective survival beliefs than in the baseline model. However, the two lower panels of the figure give more insight: in the baseline NHS model and in the OHS model, the P90/P50 metric is large, since the P50 is very small (approximately zero). However, in the SSH model, the P50 is small, but significantly larger than zero, so that the resulting P90/P50 ratio is still comparatively small. However, in the SSH model the P50/P10 ratio then becomes very large, since the P10 here is very small (approximately zero). In sum, the large swings in inequality measured for the very old age groups are driven by the fact that a smaller or larger share of the group has virtually zero assets, not that there are any really rich individuals in this age group. The top 1% in the oldest age group in the OHS model owns 22% more assets than the top 1% in the same age group in the baseline NHS model. However, these asset levels are still low compared to the really asset rich in the middle aged groups.

To conclude, the within-cohort inequality within the very oldest cohorts is affected by the survival heterogeneity, as could be expected after seeing the large differences in the policy functions. However, the richest individuals in the economy are found in the age group 60–64, as we saw in the life-cycle profiles. For this age group, the “extra savings” by healthy individuals, due to a longer expected life, are very small, as we saw in Figure 11 – the main effect in this age group is that unhealthy individuals save less. Therefore, the effect of heterogeneity in survival on the top wealthiest individuals in the economy is very small, and therefore the effect on overall inequality is negligible.

However, as previously noted, the asset levels among the oldest individuals in any of the scenarios analyzed in this section are far from the actual asset holding we see among older individuals. This motivates us to make the bequest motive operative, which is what we turn to in the next section.

### 5.2. Model with a bequest motive

Figure 15 shows asset holdings in older ages, as measured by net total asset holdings including housing (net of mortgage), and net assets excluding housing and mortgages. As can be seen, the decumulation of assets in older ages is far from what a standard life-cycle model without a bequest motive predicts. Individuals keep a substantial amount of assets for reasons not present in the simplest life-cycle model. Motivated by this fact we add a bequest motive, as given in (8).

We calibrate the parameters of the bequest motive so that the median asset levels for the oldest age group, the bequests as a fraction of total wealth in the economy, and the fraction of taxed estates are approximately hit. The resulting life-cycle profile for the baseline NHS model is shown in Figure 16 and, as can be seen, the resulting asset holdings in older ages are now more in line with the data, with substantial asset holdings even for the average individual who survives beyond
Figure 15: Mean asset holdings in older ages, by two-year age bins. 95% confidence intervals given by dashed green lines. Source: HRS.
his/her 80th birthday. Note also that the peak wealth at the age of 65 is too pronounced as compared to the data, which is partly due to assuming a fixed retirement age.

We now move on to the effect of adding survival heterogeneity. Figure 17 shows the difference in the savings rate for the three scenarios. As can be seen, the results are counter-intuitive and perhaps also unexpected: individuals at the age of 50 in poor health save more, which is the complete opposite of the behavior we saw in the previous sections without bequest. The reason is as follows: since the agents in poor health are more likely to die soon, their effective discounting on the bequest utility decreases, and therefore the bequest motive for savings increases. Hence, the net effect is that these individuals want to save more.

The mechanism that a decrease in expected longevity increases savings due to the bequest motive is always present as long as we have a model with a bequest motive formulated as in (8). With a low weight on the bequest motive (lower than what we estimated the weight to be in line with the data), the life-expectancy channel could still dominate and the net effect could still be that individuals with a shorter life expectancy save less, but the underlying mechanism – that a decrease in expected longevity makes the agent want to save more for bequest reasons – would still be present.  

Figure 18 shows the resulting life-cycle profiles for the three scenarios. As can be seen, due to the bequest motive, individuals in older ages do not decumulate their wealth, and the resulting profile is more in line with the data. As is also clear, there is hardly any difference between individuals in poor vs. excellent health,

---

14 In the appendix, section A.1 provides some more intuition and characterizes the required bequest weight in order to get a net effect of decreased savings in the event of an increased survival probability in a stylized two-period model.
neither in the model with objective survival heterogeneity, nor in the model with subjective heterogeneity. This is a direct result of what we saw in Figure 17: the savings rates are not that different as compared to the model with no survival heterogeneity.

As can be seen in the life-cycle profile for the SSH model, the median wealth among individuals in their 50s in poor health is slightly higher than among individuals in excellent health. This is the effect of the higher savings due to the bequest motive. However, wealth in the top P95 is slightly higher for 65-year-olds in excellent health than for 65-year-olds in poor health. The reason is that for extremely rich individuals, the savings motive to smooth consumption in the event of a longer life is becoming relatively more important again, after having saved up so much that the marginal utility from the bequest is small.

6. Moving forward

The models used in the two preceding sections are in many ways the two most standard models one can think of: a canonical life-cycle model and a version extended with the standard bequest motive most commonly used in the macroeconomic literature.

The first model, the canonical life-cycle model without a bequest motive, gave rise to lower savings for individuals in bad health, and higher savings for individuals in good health. However, as is well known, the canonical life-cycle model gives rise to a counter-factually fast asset decumulation at older ages, contrary to the substantial asset holdings observed in the data.

This motivated us to include a bequest motive. Again, as is well known, such a model can be calibrated to fit the average asset profile in older ages reasonably well. However, this had counter-intuitive and perhaps unexpected implications for the savings behavior of individuals with different life expectancies. People in bad health, and hence a shorter expected life span, save more due to the lower effective discounting on the bequest motive. We argue that this mechanism is not plausible.

The conclusion is hence that neither of the standard models are really useful for understanding the true implications of survival heterogeneity on savings and wealth accumulation. This conclusion can be interpreted in two ways. Either there is no bequest motive, and the reason for savings in older ages is something else, or the current way of modelling the bequest motive is not the right one, at least not for analyzing the effect of differences in expected longevity. There is probably some truth in both statements.

There is overwhelming evidence that there is a bequest motive. This gives rise to the question of what a better way of formulating it would be. A natural first suggestion would be to try to shut down the “date-of-handover” channel for the
Figure 17: Difference in the savings rate for the model with a **standard bequest motive** (percentage points/100, i.e., 0.1 indicates a difference of 10 pp). The darker color shows the model with objective survival heterogeneity compared to the baseline model with no survival heterogeneity. The lighter color shows the additional effect of adding subjective survival beliefs. The x-axis depicts cash-at-hand percentiles by age (equilibrium values for the baseline model). The color indicates health state: green is excellent while red is poor health.
Figure 18: Resulting life-cycle profiles for wealth, model with a standard bequest motive.
bequest, by always discounting the utility of the bequest to some fixed future date, regardless of when the actual death occurs. However, in practice, this is isomorphic to having a time shifting weight on the bequest, and does not solve the problem.  

Another suggestion would be to not discount the bequest at all. One could argue that for the realized utility of the bequest, it does not matter at all exactly when the wealth is handed over, rather it is only the sum that matters. However, a simple formulation with no discounting of the bequest term in the utility function gives rise to dynamically inconsistent preferences.

One might think that the inclusion of other economic effects related to the current health status would remedy the problem. For instance, if unhealthy individuals were subject to lower wages (as they are in the data), expensive medical shocks and/or had some inherent low-savings characteristics, they would have lower asset holdings on average. Hence, since leaving a bequest is a luxury good, they would not, on average, have a very strong bequest motive, and therefore they would on average save less. This argument is true in the sense that it would be a remedy for some of the counter-factual cross-sectional implications: we would most certainly get a model where the unhealthy individuals again have lower asset holdings, as in the data. However, the implausible behavioral mechanism would still exist. Imagine for instance a rich person who, thanks to good income and no large negative shocks, saved up a lot. For this asset rich person, the bequest motive is operational. Now imagine that this person receives a bad health shock (think about a cancer diagnosis), which shortens his/her expected life span. Then, the implication for this person’s savings behavior would be that he/she saves more, despite the shorter expected life span, due to the bequest motive. Hence, we argue that even though more health-related shocks could be a remedy for the cross-sectional counterfactual health-wealth gradient, it would not solve the underlying implausible connection between survival expectancy and savings rate.

A more promising path is to explicitly incorporate the parent-child connection, allow for inter-vivo bequests, and maximize over the dynasty utility. In such a model, an increase in the expected life span would translate into an expectation of a longer period with two agents alive in parallel in the dynasty. However, then it would be important to also have a motive for asset holdings in older ages for childless individuals.  

With a warm-glow motive, the exact definition of who eventually gets the bequest is not as important.

---

15For instance, one could model the bequest motive as $\beta^T - \theta_R (\kappa + \alpha)^{1-\sigma}$, which means that the agent always discounts the bequest utility as if it were handed over in time $T$. However, the formulation is isomorphic to a bequest motive of the standard type, but with a weight that is increasing over time. This gives rise to counter-factually large savings late in life, and does not remove the expected-date-of-handover effect.

16According to Livingston (2015), the number of childless women between 40 and 44 has varied between 15 and 20 percent during the last 30 years.
References


CHAPTER 4


A. Appendix

A.1. More on the bequest motive

With the conventional way of formulating a bequest motive, the effect of an increase in survival has unexpected and non-intuitive effects on savings. To illustrate the argument, we write down a simple two-period model where the only uncertainty is the survival probability, $\pi$, between the first and the second period. After the second period, the agent dies with certainty. We assume that the agent has some initial assets $a_0$, the discount factor is $\beta < 1$, and the gross interest rate is $R$. Hence, the agent solves

\[
\max_{c_0, c_1} \left\{ u(c_0) + \beta \left( \pi \left[ u(c_1) + \beta V_b(a_2) \right] + (1 - \pi) V_b(a_1) \right) \right\}
\]

subject to the constraints

\[
c_t > 0 \quad \forall t, \quad a_1 = a_0 - c_0, \quad a_2 = Ra_1 - c_1.
\]

Taking first-order conditions with respect to the choice variables gives the following optimality conditions:

\[
\frac{\partial u}{\partial c_0} = \beta^2 \pi R \frac{\partial V_b}{\partial a_2} \bigg|_{a_2^*} + \beta (1 - \pi) \frac{\partial V_b}{\partial a_1} \bigg|_{a_1^*},
\]

\[
\frac{\partial u}{\partial c_1} = \beta \frac{\partial V_b}{\partial a_2} \bigg|_{a_2^*},
\]

where $a_1^*$ and $a_2^*$ denote the optimal choice of savings in each period. Hence, the impact of an increase in survival probability on $c_0^*$, optimal first period consumption, is ambiguous and depends on how the bequest motive is parametrized.

The utility function $u$ has the usual CRRA functional form:

\[
u(c) = \frac{c^{1-\sigma} - 1}{1-\sigma}
\]

and we use the standard bequest motive:

\[
V_b(a) = \theta_B \frac{(a + \kappa)^{1-\sigma} - 1}{1-\sigma}
\]

where $\theta_B$ determines the strength of the bequest motive, and $\kappa$ determines to what extent bequests are a luxury good. For simplicity, we assume $\kappa = 0$ and solve for the agent’s problem as given in (9). Some algebra gives that if an increase in the
survival probability makes the agent consume more or less in the first period, i.e.,
the sign of \( \frac{\partial c^*_0}{\partial \pi} \), is determined by the ratio

\[
\theta_B \lesssim \left( \frac{R^{1/\sigma}}{1 - R^{1/\sigma} \theta B} \right) ^\sigma \equiv \hat{\theta}_B
\]

If \( \theta_B < \hat{\theta}_B \), an increase in the survival probability has the expected effect: the agent
consumes less in the current period and saves more, given that it is more likely to
survive to the next period.

On the other hand, if \( \theta_B > \hat{\theta}_B \), an increase in the survival probability leads to
decreased savings and more consumption in the first period. There are two mecha-
nisms behind this. First, when the probability of surviving increases, the effective
discounting of the next-period bequest utility increases, and hence the incentive
to save decreases. We call this the expected-date-of-handover channel. Second, an
increase in the survival probability leads to a higher expected interest rate income
over the remaining life, and therefore the agent can afford more consumption also
in the first period. We call this channel the income effect. If the weight on the
bequest motive is high, these two effects dominate the effect of wanting to save for
a longer expected life.

In the above example, we assumed \( \kappa = 0 \). It can be shown that if we assume \( R = 1 \),
the sign of \( \frac{\partial c^*_0}{\partial \pi} \) is actually independent of \( \kappa \) (although it still affects the level of
savings). However, if we allow for a positive interest rate, the extent to which the
bequest is a luxury good in combination with the level of initial assets, \( a_0 \), matters.
Moreover, we assumed no second-period income in our above example. If we allow
for income in the second period, the level of that income compared to the initial
assets affects the incentives. For a given \( a_0 \), the higher the second-period income,
the less bequest weight is needed to get \( \frac{\partial c^*_0}{\partial \pi} > 0 \) – since then the bequest becomes
relatively more important as a reason to save.

### A.2. Retirement benefits

First, consider the following stylized version of the actual retirement income for-
mula used in the US social security system, where \( \bar{e} \) is an (annualized) measure of
historical average monthly earnings, \( b^1_s \) and \( b^2_s \) are bend points in USD for some
reference year, and \( e^s_{max} \) is the contribution and benefit base (CBB), i.e. the maximum
earnings subject to payroll taxes. Retirement income measured in USD, \( i^s_R \), is then
approximately given by

\[
i^s_R(\bar{e}) = \begin{cases} 
\rho_1 \bar{e} & \text{if } \bar{e} \leq b^1_s \\
\rho_1 b^1_s + \rho_2 (\bar{e} - b^1_s) & \text{if } b^1_s < \bar{e} \leq b^2_s \\
\rho_1 b^1_s + \rho_2 (b^2_s - b^1_s) + \rho_3 \left( \min \{ e^s_{max}, \bar{e} \} - b^2_s \right) & \text{else}
\end{cases}
\]

236
CHAPTER 4

where $\rho_1$, $\rho_2$ and $\rho_3$ are decreasing replacement rates applied to earnings ranges bracketed by the bend points $b_i^S$.

In the model, we define retirement benefits to be a product of the following components:

$$ i_R(p) = w \times y_R(p) = w \times \omega_{T_R-1}p_R(p) $$

where the function $p_R(\cdot)$ is constructed below analogously to $i_R^S$.

To this end, denote by $e_{med}^S$ the USD median earnings in the reference year. To express the bend points in terms of the persistent labor state in retirement, we implicitly define $\tilde{p}_i$ corresponding to $b_i^S$ as

$$ b_i^S = \frac{w \times \omega_{T_R-1}p_i \bar{e}}{w \times y_{med}} $$

where we have normalized the bend point $b_i^S$ by real-world median earnings and the model counterpart with the median earnings in the model. We find that

$$ \tilde{p}_i = \left( \frac{b_i^S}{e_{med}^S} \right) y_{med} \omega_{T_R-1} \bar{e} $$

Analogously, the CBB in terms of persistent labor productivity is

$$ \tilde{p}_{max} = \left( \frac{e_{max}^S}{e_{med}^S} \right) y_{med} \omega_{T_R-1} \bar{e} $$

By factoring out the common term $\omega_{T_R-1} \bar{e}w$ that is independent of a retired individual’s idiosyncratic state vector, we can write the replacement formula purely in terms of the permanent labor state as follows:

$$ p_R(p) = \begin{cases} 
\rho_1 p & \text{if } p \leq \tilde{p}_1 \\
\rho_1 \tilde{p}_1 + \rho_2 (p - \tilde{p}_1) & \text{if } \tilde{p}_1 < p \leq \tilde{p}_2 \\
\rho_1 \tilde{p}_1 + \rho_2 (\tilde{p}_2 - \tilde{p}_1) + \rho_3 \left( \min \{p_{max}, p\} - \tilde{p}_2 \right) & \text{else}
\end{cases} $$

Note that this implies that the bend points and the CBB are proportional to the wage level $w$.

The government expenditures on retirement benefits in a period are given by

$$ G_{ss} = \sum_{t=T_R}^{N_t} \sum_p \mu_t(p) \mu_p(p) i_R(p) $$

which is a weighted sum over the retirement incomes received by all retired cohorts, with weights $\mu_t$ and $\mu_p$ denoting the PMFs of the ergodic distribution of age.
and persistent labor states, respectively. We denote the mass of retired individuals by

$$\Pi_R = \sum_{t=T_R}^{N_t} \mu_t(t) \mu_p(p) \mu_e(\epsilon) T_{ss}(y_e) w$$

and the average permanent component of retirement income as

$$\overline{p}_R = \sum_{p} \mu_t(p) p_R(p).$$

### A.3. Social security budget balance

The payroll taxes raised each period are

$$T_{ss} = \sum_{t=1}^{T_R-1} \sum_{p} \sum_{\epsilon} \mu_t(t) \mu_p(p) \mu_e(\epsilon) T_{ss}(y_e) w$$

where $\mu_e(\epsilon)$ is the ergodic distribution over transitory labor shocks. The payroll tax function is defined as

$$T_{ss}(y) = \tau_{ss} \times \min\{y_{max}, y\}$$

where $y_{max}$ expresses maximum taxable earnings in terms of labor productivity, i.e.

$$y_{max} = \left(\frac{e_{max}}{e_{med}}\right) y_{med}.$$

To balance the social security system, we need to find $\tau_{ss}$ such that $G_{ss} = T_{ss}$. Equating $G_{ss} = T_{ss}$ implies that

$$\tau_{ss} = \frac{\omega T_{R-1} \overline{p}_R \epsilon \Pi_R}{\sum_{t=1}^{T_R-1} \sum_{p} \sum_{\epsilon} \mu_t(t) \mu_p(p) \mu_e(\epsilon) \min\{y_{max}, \omega t \epsilon \}}.$$

### A.4. Estate taxes

Our solution algorithm requires a continuously differentiable tax rate on estates. Additionally, we want to impose a tax exemption to make sure that only a small fraction of estates is subject to the estate tax.

To this end, we use the cosine function to create a marginal tax rate that is defined as follows:

$$\frac{\partial T_b(b)}{\partial b} = \begin{cases} 
0 & \text{if } b \leq \chi_b \\
\frac{\beta_b}{\tau_b} \left[ \cos \left( \frac{\pi}{2} \left( \frac{b-\chi_b}{B} - 1 \right) \right) + 1 \right] & \text{if } \chi_b < b \leq \chi_b + B \\
0 & \text{else}
\end{cases}$$

(15)
We assume that the marginal tax rate is increasing on the interval \([\chi_b, \chi_b + B]\) and constant everywhere else. The tax function itself is obtained by integrating (15), which gives

\[
T_b(b) = \begin{cases} 
0 & \text{if } b \leq \chi_b \\
\frac{\tau_b}{\frac{B}{\pi}} \left[ \sin \left( \pi \frac{b - \chi_b - 1}{B} \right) \frac{B}{\pi} + b - \chi_b \right] & \text{if } \chi_b < b \leq \chi_b + B \\
\tau_b b - \chi_b - B + \frac{\tau_b}{2} B & \text{else}
\end{cases}
\]

A.5. Income taxes

In this section we derive an expression for the total amount of income taxes raised by the government. Before proceeding, we state the following useful definitions: We denote by \(\overline{p}\) the average persistent labor shock,

\[
\overline{p} = \sum_p \mu_p(p) p
\]  

and by \(\Pi_{LF}\) the size of the labor force,

\[
\Pi_{LF} = \sum_{t=1}^{T_R-1} \mu_t(t) = 1 - \Pi_R
\]

Additionally, average labor productivity can be defined as

\[
\overline{y} = \Pi_{LF}^{-1} \left[ \sum_{t=1}^{T_R-1} \sum_p \sum_{\epsilon} \mu_t(t) \mu_p(p) \mu_e(\epsilon) \omega_t p \epsilon \right] = \Pi_{LF}^{-1} \left[ \overline{p} \cdot \overline{\epsilon} \sum_{h} \mu_t(t) \omega_t \right].
\]

Now, consider the aggregate tax revenues raised from working individuals, which are given by

\[
T_e = \sum_{t=1}^{T_R-1} \sum_p \sum_{\epsilon} \mu_t(t) \mu_p(p) \mu_e(\epsilon) \left[ (\omega_t p \epsilon - T_{ss} (\omega_t p \epsilon)) w \right] \\
- \lambda \left( (\omega_t p \epsilon - T_{ss} (\omega_t p \epsilon)) w \right)^{1-\tau}
\]

\[
= \left[ w \Pi_{LF} \overline{y} - \lambda w^{1-\tau} \overline{y}_{ss,\tau} \right] - T_{ss}
\]

where we define

\[
\overline{y}_{ss,\tau} = \sum_{t=1}^{T_R-1} \sum_p \sum_{\epsilon} \mu_t(t) \mu_p(p) \mu_e(\epsilon) \left( \omega_t p \epsilon - T_{ss} (\omega_t p \epsilon) \right)^{1-\tau}
\]

to simplify the notation.
CHAPTER 4

Income taxes raised from retired individuals amount to

\[ T_R = \sum_{t=I_R}^{N_t} \sum_p \mu_t(t) \mu_p(p) \left[ y_R(p)w - \lambda \left( y_R(p)w \right)^{1-\tau} \right] \]

\[ = \sum_{t=I_R}^{N_t} \sum_p \mu_t(t) \mu_p(p) \left[ \omega_{T_R-1} p_R(p) \bar{e}w - \lambda \left( \omega_{T_R-1} p_R(p) \bar{e}w \right)^{1-\tau} \right] \]

\[ = \Pi_R \left[ w \omega_{T_R-1} p_R(\bar{e}) - \lambda w^{1-\tau} \omega_{T_R-1} p_R(\bar{e})^{1-\tau} \right] \]

\[ = T_{ss} - \lambda \Pi_R w^{1-\tau} \omega_{T_R-1} p_R(\bar{e})^{1-\tau} \]

with

\[ \bar{p}_{R,\tau} = \sum_p \mu_p(p) p_R(p)^{1-\tau} \]

Thus, the total revenue from income taxes is

\[ T_{inc} = T_e + T_R \]

\[ = w \Pi_L \bar{y} - \lambda w^{1-\tau} \left[ \bar{y}_{ss,\tau} + \Pi_R \omega_{T_R-1} p_R(\bar{e})^{1-\tau} \right] \]
Den här avhandlingen består av fyra fristående kapitel. Kapitel 1 och 2 fokuserar på heterogenitet i arbetsutbud, medan Kapitel 3 och 4 handlar om heterogeniteten i hälsa och förväntad livslängd. Även om kapitlen är fristående, speglar de tillsammans mitt intresse för den gren av makroekonomin som använder modeller med heterogena agenter, och min åsikt att vi bör studera individuellt beteende och inte bara aggregerade serier när vi försöker förstå många ekonomiska frågor.

I det första kapitlet, *Structural transformation of the labor market and the aggregate economy*, fokuserar jag på den dramatiska ökningen i arbetskraftsdeltagande bland gifta kvinnor i USA under de senaste 50 åren, från under 40% på 1960-talet till nästan 70% nu. I början av 1960-talet var den vanligaste typen av hushåll i ekonomin ett hushåll med två vuxna varav en ensam familjeförsörjare. Så är inte längre fallet. Kvinnors ökade deltagande i ekonomin har varit den kanske viktigaste strukturella förändringen av arbetsmarknaden under det förra århundradet.

Jag redogör för denna period av strukturell förändring av arbetsmarknaden i en makroekonomisk modell, och studerar hur det ökade kvinnliga arbetskraftsdeltagandet har påverkat ekonomins respons på aggregerade chocker. För detta ändamål konstruerar jag en modell med heterogena agenter där jag explicit inkluderar heterogenitet i kön och sammansättning av hushåll (jämte heterogenitet i förmögenhet och produktivitet). I övrigt är modellen en rättfram konjunkturcykelmodell. Den enda exogena förändringsfaktorn i modellen är en krympande löneskillnad mellan män och kvinnor, följaktligen är mitt fokus de förändrade ekonomiska incitamenten inom hushållet.

Modellen fängar de framträdande dragen i historiska data: den kraftiga sysselsättningsökningen bland gifta kvinnor, den låga utträffningen av gifta män från arbetsmarknaden, och den relativt stabila sysselsättningen bland ensamstående kvinnor. Resultat från modellen visar att den underliggande trenden i sysselsättning, driven av den kraftiga ökningen i kvinnligt arbetskraftsdeltagande, bidrog till vad som uppfattades som en snabb återhämtning av sysselsättningen efter recessioner före 1990, och avsaknaden av underliggande tillväxt följaktligen förklarar den längsammare återhämtningen av sysselsättning efter recessioner under senare tid.
SAMMANFATTNING

Generellt visar jag hur inkorporering av en- och två-personshushåll i modellen påverkar den aggregerade sysselsättningsdynamiken, då en-personshushåll reagerar kraftigare på chocker. Dessutom har sysselsättningsresponsen inom olika subgrupper förändrats över tid. Exempelvis har gifta kvinnors respons på en chock dämpats över tid, eftersom de i genomsnitt har förflyttat sig längre bort från sin indifferenspunkt för arbete.


Kapitel 3 och 4 handlar om välbehinnande: heterogenitet i hälsa och förväntad livslängd i befolkningen i USA. Ojämlikhet i hälsa är viktigt i sig, men att förstå de underliggande orsakerna är långt utanför vad som kan omfattas av en avhän- dling i ekonomi. Vad vi gör är att dokumentera hälsoprocessen på ett strukturerat sätt så att resultatet kan användas i en makroekonomisk modell, och tar ett litet steg på så sätt att vi undersöker hur ojämlikhet i förväntad livslängd påverkar ekonomiska utfall, speciellt sparande. Många studier har identifierat hälsochcker som en av de största riskerna i livet. En negativ hälsochock kan föranleda stora medicinska utgifter, vilket påverkar incitamenten att spara, och kan också påverka inkomsterna direkt. Den förväntade livslängden har en direkt påverkan på den
effektiva diskonteringsfaktorn, en mekanism inbäddad i alla livscykelmodeller med osäker livslängd. Vissa argumenterar vidare att en individs hälsostatus har en direkt påverkan på marginalnyttan från konsumtion. För att kvantifiera risken en individ står inför och modellera de val och beslut individen tar är således en korrekt modellerad hälsoprocess avgörande.


I det fjärde kapitlet, Subjective life expectancies, time preferences heterogeneity and wealth inequality, även det samförfattat med Richard Foltyn, använder vi resultaten från det föregående kapitlet och ställer oss den naturliga följdfrågan: vilka konsekvenser har heterogenitet i förväntad livslängd på sparkvoter och i slutändan förmögenhetsojämlikhet? 

I enlighet med standardantaganden i ekonomisk teori bör en individ i god hälsa spara mer för framtiden, givet att denna person har en högre sannolikhet att leva ett långt liv. Dock är inte en individs konsumtions- och sparandebeslut nödvändigtvis styrda av objektiv statistisk förväntad livslängd, utan snarare av vad individen uppfattar som sin överlevnadssannolikhet. Vi dokumenterar nya fakta gällande systematisk felskattning i dessa uppfattningar: individer med låg överlevnadssannolikhet relativt sin omgivning underskattar sin överlevnadssannolikhet, medan individer med hög överlevnadssannolikhet överskattar. Denna systematiska felskattning intensifierar heterogeniteten i förväntad livslängd i populationen.

För att uppskatta effekten av heterogenitet i förväntad livslängd, objektiv såväl som subjektiv, på förmögenhetsojämlikhet använder vi oss av en allmän jämväktsmodell med överlappande generationer där individerna inte kan försäkra sig mot risken. Individerna har heterogena överlevnadssannolikheter som beror på deras aktuella hälsostatus, och deras hälsa förändras beroende på chocker. Utöver denna osäkerhet inkluderar vi även persistenta och övergående chocker till arbetsproduktivitet.
SAMMANFATTNING

Vi visar att en standardmodell av livscykeltyp har kontrafaktiska implikationer när vi introducerar heterogenitet i förväntad livslängd. I en modell utan arv sparar individer med längre förväntad livslängd mer, som förväntat. Detta stämmer överens med data, där individer med bättre hälsa har större förmögenheter. Men som vi redan känner till, så har en sådan modell utan arv kontrafaktiskt låga förmögenheter bland äldre individer.

Därför inkorporerar vi nytta från att lämna arv i individernas nyttofunktion. Effekten av detta är dock också kontraintuitiv och möjligtvis också oväntad: individer i dålig hälsa sparar nu mer än sina hälsosammare motsvarigheter. Anledningen är följande: eftersom individider i dålig hälsa har en högre sannolikhet att dö snart, lägger de större vikt på den potentiella nytton från att lämna efter sig ett arv, och däremot har de högre incitament att spara.

Vår slutsats är därmed att ingen av standardmodellerna är lämplig för att studera effekten av heterogenitet i förväntad livslängd på sparkvoter och förmögenhetsöjämlikhet. Vi diskuterar potentiella förändringar av standardmodellerna och pekar ut riktningar för framtida forskning.
IIES Monograph Series

22. Lundvik, Petter, Business Cycles and Growth, 1992
25. Daltung, Sonja, Risk, Efficiency, and Regulation of Banks, 1994
27. Stennek, Johan, Essays on Information-Processing and Competition, 1994
29. Dahlquist, Magnus, Essays on the Term Structure of Interest Rates and Monetary Policy, 1995
32. Lagerlöf, Nils-Petter, Intergenerational Transfers and Altruism, 1997
37. Domeij, David, Essays on Optimal Taxation and Indeterminacy, 1998
38. Flodén, Martin, Essays on Dynamic Macroeconomics, 1999
41. Vestin, David, Essays on Monetary Policy, 2001
43. Johansson, Åsa, Essays on Macroeconomic Fluctuations and Nominal Wage Rigidity, 2002
44. Groth, Charlotta, Topics on Monetary Policy, 2002
45. Gancia, Gino A., Essays on Growth, Trade and Inequality, 2003
47. Kohlscheen, Emanuel, Essays on Debts and Constitutions, 2004
48. Olovsson, Conny, Essays on Dynamic Macroeconomics, 2004
49. Stavlöt, Ulrika, Essays on Culture and Trade, 2005
50. Herzing, Mathias, Essays on Uncertainty and Escape in Trade Agreements, 2005
52. Pienaar, Natalie, Economic Applications of Product Quality Regulation in WTO Trade Agreements, 2005
53. Song, Zheng, Essays on Dynamic Political Economy, 2005
54. Eisensee, Thomas, Essays on Public Finance: Retirement Behavior and Disaster Relief, 2005
55. Favara, Giovanni, Credit and Finance in the Macroeconomy, 2006
57. Larsson, Anna, Real Effects of Monetary Regimes, 2007
60. Queijo von Heideken, Virginia, Essays on Monetary Policy and Asset Markets, 2007
61. Finocchiaro, Daria, Essays on Macroeconomics, 2007
64. Damsgaard, Erika Färnstrand, Essays on Technological Choice and Spillovers, 2008
66. Folke, Olle, Parties, Power and Patronage, 2010
67. Drott, David Yanagizawa, Information, Markets and Conflict, 2010
68. Meyersson, Erik, Religion, Politics, and Development, 2010
69. Klingelhofer, Jan, Models of Electoral Competition, 2010
70. Perrotta, Maria Carmela, Aid, Education and Development, 2010
72. Mueller, Andreas, Business Cycles, Unemployment and Job Search, 2011
77. Li, Yinan, *Institutions, Political Cycles and Corruption*, 2012
78. Håkanson, Christina, *Changes in Workplaces and Careers*, 2013
86. Rogall, Thorsten, *The Economics of Genocide and War*, 2015
87. Baltrunaite, Audinga, *Political Economics of Special Interests and Gender*, 2016
93. Öberg, Erik, *On Money and Consumption*, 2017
97. Dehdari, Sirus Håfström, *Radical Right, Identity, and Retaliation*, 2018
98. Harmenberg, Karl, *Essays on Income Risk and Inequality*, 2018
99. Kilström, Matilda, *Households’ Responses to Policy in Labor and Credit Markets*, 2018
102. Cocciolo, Serena, *Participatory Governance and Public Service Provision*, 2019
This thesis consists of four self-contained essays.

**Structural transformation of the labor market and the aggregate economy** studies the aggregate economic effects of the increased labor force participation among married women during the last century.

**Labor supply in a quantitative heterogeneous-agent model** systematically examines a class of macroeconomic models with endogenous labor supply.

**Health dynamics and heterogeneous life expectancies** provides improved estimates for health and survival dynamics in the population and documents heterogeneity in longevity across subgroups.

**Subjective life expectancies, time preference heterogeneity, and wealth inequality** studies the effect of survival heterogeneity on savings rates and wealth inequality.