# Selection of Smoothing Parameters with Application in Causal Inference

Jenny Häggström

**Department of Statistics**
Umeå University 2011

Doctoral Dissertation
Department of Statistics
Umeå University
SE-901 87 Umeå

ISBN: 978-91-7459-147-7
ISSN: 1100-8989
Statistical Studies No. 44
Electronic version available at http://umu.diva-portal.org/
Printed by: Print & Media
Umeå, Sweden 2011

*Once in a lifetime, water flowing underground.*

# Contents

## List of Papers

The thesis is based on the following papers:

I. de Luna, X., and J. Häggström (2010). Estimating prediction error: cross-validation vs. accumulated prediction error. *Communications in Statistics-Simulation and Computation 39*, 880-898.

II. Häggström, J. (2010). Bandwidth selection for backfitting estimation of semiparametric additive models. Under revision. Resubmission invited by *Statistics and Computing.*

III. Häggström, J., and X. de Luna (2011). Data-driven smoothing parameter selection for estimating average treatment effects. Research Report, Department of Statistics, Umeå University, Sweden.

IV. Häggström, J., O. Westerlund, M. Norberg and X. de Luna (2011). Divorcing in middle age and its effects on BMI. Research Report, Department of Statistics, Umeå University, Sweden.

Paper I is reprinted with the kind permission of Taylor & Francis.

# Abstract

This thesis is a contribution to the research area concerned with selection of smoothing parameters in the framework of nonparametric and semiparametric regression. Selection of smoothing parameters is one of the most important issues in this framework and the choice can heavily influence subsequent results. A nonparametric or semiparametric approach is often desirable when large datasets are available since this allow us to make fewer and weaker assumptions as opposed to what is needed in a parametric approach.

In the first paper we consider smoothing parameter selection in nonparametric regression when the purpose is to accurately predict future or unobserved data. We study the use of accumulated prediction errors and make comparisons to leave-one-out cross-validation which is widely used by practitioners.

In the second paper a general semiparametric additive model is considered and the focus is on selection of smoothing parameters when optimal estimation of some specific parameter is of interest. We introduce a double smoothing estimator of a mean squared error and propose to select smoothing parameters by minimizing this estimator. Our approach is compared with existing methods.

The third paper is concerned with the selection of smoothing parameters optimal for estimating average treatment effects defined within the potential outcome framework. For this estimation problem we propose double smoothing methods similar to the method proposed in the second paper. Theoretical properties of the proposed methods are derived and comparisons with existing methods are made by simulations.

In the last paper we apply our results from the third paper by using a double smoothing method for selecting smoothing parameters when estimating average treatment effects on the treated. We estimate the effect on BMI of divorcing in middle age. Rich data on socioeconomic conditions, health and lifestyle from Swedish longitudinal registers is used.

KEYWORDS: Smoothing parameter selection; Nonparametric regression; Semiparametric additive model; Double smoothing; Causal inference; BMI; Divorce

# Preface

Do you remember the childhood pastime of filling out questionnaires in “My Friends”-books? If you do you know that you were supposed to write something in the blank space after “What do you want to be when you grow up?”. As a child I seldom had a ready answer (except for “A millionaire” which does not really count). I never wanted to be an astronaut, hairdresser or veterinarian. And I certainly did not want to become a doctor. Of any kind. The fact that I never had a dream job has possible made me even more grateful that I ended up were I am now. I consider it a luxury to be able to spend your working hours doing something you deeply enjoy and I count myself among the ones lucky enough to have been granted this. I will look back at these years with fondness and this is in many ways thanks to all the nice people that work and have worked at the Department of Statistics. Thanks to all of you!

Writing this thesis has been a surprisingly pleasant experience. Of course, there have been moments of anguish and paralyzing fear but not at all in the amounts I expected. If truth be told more than not it has been pure joy and that is more than anything due to my superb supervisor Xavier de Luna. Always encouraging, optimistic and full of ideas. Thank you so very much! I could not have done this without you.

Thanks also to my co-supervisor Marie Wiberg and to Priyantha Wijayatunga, Eva Cantoni, Sara de Luna and everyone in the Causal Inference research group for reading and commenting on earlier drafts of the papers in this thesis.

Thanks to my co-authors Olle Westerlund and Margareta Norberg. Working with real data was a challenging but valuable experience!

I especially want to thank Henrik, Mathias, Maria and Emma for sharing good and not so good moments in and outside the work place. Hopefully there will be lots of laughter (and wine) in the future too!

Andreas, thank you for all your support and for always believing that I could do this! Mattias and Björn, thanks for swapping gossip and geeky stories about computer stuff in return for statistics stuff! Josefin and Mari, thanks for all the laughter, love and friendship!

Thanks to my family, Sven-Erik, Carina, Jimmy and Jessica, for always being there and feeling happy for me (even though I suspect you sometimes doubted my sanity).

To everyone else that have been part of this chapter of my life, even though I don't mention you by name it does not imply I'm not grateful for your contribution. Thank you too!

I'm also grateful for the financial support received from the Swedish Research Council (Dnr:2007-2465 and the Graduate School in Population Dynamics and Public Policy) and the possibility to use the resources of the High Performance Computing Center North (HPC2N).

Umeå, February 2011
Jenny Häggström

# 1 Introduction

Statistical methods are used to draw scientifically based conclusions about the world. For this to be possible the phenomenon in which we are interested in must be described in terms of a mathematical model. This model should preferably be, at the same time, simple enough to work with and complex enough to capture all the relevant aspects of the phenomenon.

As an illustrative example, consider the Onions dataset available in the R package SemiPar (R Development Core Team, 2010). The dataset contains observations from an experiment involving the production of white Spanish onions in South Australia. Denote the yield (in grams per onion plant) and the area density (in plants per square meter) of the plants by $y$ and $x$, respectively. Suppose we would like to estimate the relationship between area density and yield. A simple and common model when describing a relationship between variables such as these assumes that the relationship between them is linear. For the assumption of linearity to be plausible the data points must cluster along a straight line. In Figure 1 yield is plotted against density for 42 Spanish onion plants in Virginia, Australia. The straight line is the estimated relationship between the two under the assumption of linearity.

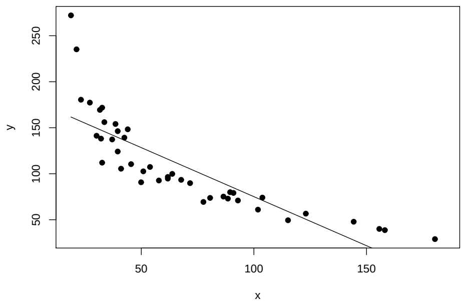

*Figure 1: Simple linear regression applied to the Onions data. Area density is on the $x$-axis and yield is on the $y$-axis. The straight line is the least-squares regression fit.*

If a straight line does not seem to accurately describe the relationship we may come up with another functional form that fits the data better, such as: a u-shaped or some other nonlinear function. Specifying the assumed relationship as a mathematical function with unknown parameters (the intercept and slope in this case), is called having a *parametric* model of the relationship. If your model agrees to a large extent with reality, the methods for fitting the model (that is, for estimating the unknown parameters) will allow us to draw conclusions about the process in question. However, if the model is misspecified (that is if it does not describe reality well) the conclusions drawn will most likely be incorrect or, at best, less precise. One way to avoid the need to specify parametric functional forms is to use *nonparametric* models in which the relationship is only assumed to obey an unspecified smooth mathematical function. In this case, the data is allowed to decide which shape gives the best fit.

Broadly speaking, there are three main purposes of using *smoothing* techniques: *explorative data analysis*, *prediction* and *causal inference.* When smoothing is used for explorative purposes it is an aid to discovering and highlighting underlying patterns in the data. This information can then be used for specifying a parametric model for further analysis. If prediction is of interest sample data can be used to estimate a model aimed at foreseeing the outcomes of new, unseen data (e.g., the yield of a new onion plant). In causal inference, smoothing techniques often serve as building blocks when estimating causal effects of non-randomized treatments on an outcome of interest.

This thesis is concerned with selection of smoothing parameters in different settings. In Paper I, smoothing is used for prediction, and a simple nonparametric model, requiring the selection of one smoothing parameter, is assumed. In Paper II, more complex models are considered, requiring the selection of several smoothing parameters. The selection of more than one smoothing parameter is also considered in Paper III, in which smoothing is used in the context of causal inference. The thesis concludes with Paper IV, in which a selection method proposed in Paper III is implemented when investigating the causal effects of divorce on BMI.

The structure of this summary is as follows: Section 2 describes smoothing (in particular local polynomial regression); Section 3 focuses on causal inference within the potential outcome framework and on smoothing in this setting; Section 4 summarizes the four papers of the thesis; and the last section provides some final remarks and comments on further research.

## 2 Smoothing

Suppose that we have a random sample of $n$ observations consisting of a response $y$ and a covariate $x$ and that the relationship between the two can be described as

$$y_i = \beta(x_i) + \epsilon_i, \quad i = 1, \ldots, n \tag{1}$$

where $\beta(\cdot)$ is a smooth function and $\epsilon$ is a random error with expected value zero. Smoothing techniques aims at retrieving $\beta(\cdot)$ from the observed data on $y$ and $x$. A simple form of smoothing is moving average, in which data is smoothed out by averaging the observations of $y$ that have similar values of $x$. The procedure and the results of applying a moving average (with averages based on 25 observations) to the Onions data are depicted in Figure 2. The estimate of the expected yield for a given density, $\hat{\beta}_h(x_i)$, is the average of $y_i$ and its 24 nearest neighbours (indicated by circles). These averages are interpolated to obtain a final estimate of the smooth function $\beta(\cdot)$. The third plot in Figure 2 shows the moving average fits with averages based on 25 (solid) and 5 (dashed) observations. The moving average based on 25 observations results in a smoother fit than the moving average based on 5 observations. This feature holds in general; using many observations will result in a smoother fit with smaller variance than using fewer observations. Using fewer observations will, however, result in a fit with less bias. Formally, we can write the moving average fit at $x_i$ as

$$\hat{\beta}_h(x_i) = \sum_{j=1}^{n} S_{ij}^h y_j = \sum_{j=1}^{n} \frac{1}{h} K\left(\frac{x_j - x_i}{b_i}\right) y_j \tag{2}$$

where $S_{ij}^h$ are weights (which sum to one), $h \in [1, n]$ is the *smoothing parameter* (which, for a moving average, is the number of observations used in a single average) and $b_i = |x_h - x_i|$; $x_h$, the $h$th nearest (in Euclidean distance) to $x_i$ among the $x_j$:s for $x_j \neq x_i$, defines the size of the window containing the observations used.

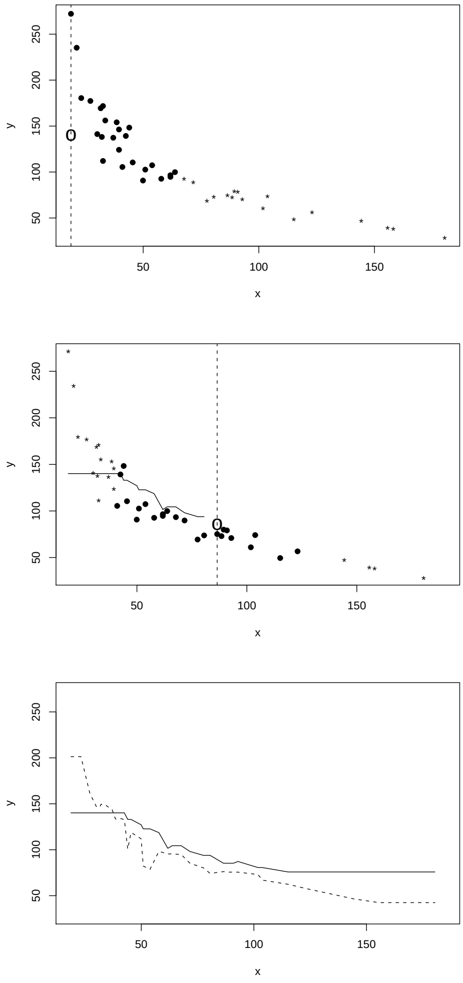

*Figure 2: Moving average applied to the Onions data. In the first plot the dotted line indicates the target point and the 25 observations used to compute the first average are displayed in bold. The circle marks the resulting local average. In the second plot, the curve is the moving average fit that results after computing 29 local averages. Also shown are the 30th local average and the observations used. The final moving average smooth using 25 observations in the local averages is displayed in the third plot (solid), along with a moving average fit based on 5 observations (dashed).*

The *uniform kernel function* is defined as

$$K\left(\frac{x_j - x_i}{b_i}\right) = \begin{cases} 1, & \text{if } |\frac{x_j - x_i}{b_i}| \leq 1 \\ 0, & \text{if } |\frac{x_j - x_i}{b_i}| > 1. \end{cases}$$

Some other common, but somewhat more sophisticated, smoothing methods are *splines*, *kernel regression* and *local polynomial regression* (Fan and Gijbels, 1996, p. 14–45). All of these methods, including the moving average (a special case of kernel regression), are examples of *linear smoothers* in which the fit at $x_i$ can be written as in (2) with the weights not depending on the response. Throughout this thesis local polynomial regression is often used and thus deserves a more thorough exposition.

## 2.1 Local Polynomial Regression

Local polynomial regression (Cleveland, 1979; Fan and Gijbels, 1996) consists of fitting a polynomial of degree $p$ at every $x_i$, $i = 1, \ldots, n$, using only the elements of the data that are deemed to be sufficiently close to the target point $x_i$. Figure 3 is the analogue of Figure 2 when fitting a polynomial of degree 1 (a straight line) at every $x_i$. Instead of taking the mean of the 25 observations here, we use them to compute a weighted linear regression and take $\hat{\beta}_h(x_i)$ as the intercept of this regression fit (indicated by circles in Figure 3).

The local polynomial regression fit at $x_i$, using a polynomial of degree $p$, is

$$\hat{\beta}_h(x_i) = \sum_{j=1}^{n} S_{ij}^h y_j = \mathbf{e}_1^T (\mathbf{X}_i^T \mathbf{W}_{i,h} \mathbf{X}_i)^{-1} \mathbf{X}_i^T \mathbf{W}_{i,h}$$

where $\mathbf{e}_1 = (1, 0, \ldots, 0)^T$, a $(p+1)$-length vector,

$$\mathbf{X}_i = \begin{pmatrix} 1 & (x_1 - x_i) & (x_1 - x_i)^2 & \ldots & (x_1 - x_i)^p \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & (x_n - x_i) & (x_n - x_i)^2 & \ldots & (x_n - x_i)^p \end{pmatrix},$$

and

$$\mathbf{W}_{i,h} = \text{diag}\{K\big((x_1 - x_i)/b_i\big)/b_i \ \ldots \ K\big((x_n - x_i)/b_i\big)/b_i\}.$$

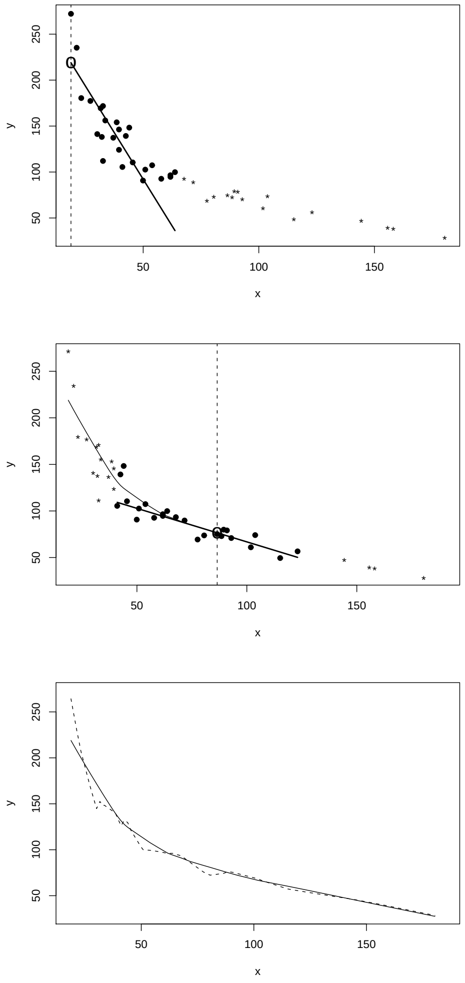

*Figure 3: Local linear regression applied to the Onions data. In the first plot the dotted line indicates the target point, and the 25 observations used to compute the first local regression are displayed in bold. The straight line is the local linear regression fit and the circle marks the resulting estimate. In the second plot the curve is the resulting local linear fit after computing 29 local linear regressions. Also shown is the 30th local estimate and regression fit and the observations used. The final local linear smooth when using 25 observations in the local regressions is displayed in the third plot (solid), along with a local linear fit based on 5 observations (dashed).*

$K(\cdot)$ is a kernel function, that is, a real-valued function assigning weights, usually a symmetric probability density function (Fan and Gijbels, 1996, p. 14–15). The kernel function used in the above example is the tricube kernel, which is defined as

$$K\left(\frac{x_j - x_i}{b_i}\right) = \left\{ \begin{array}{ll} \frac{70}{81}(1 - |\frac{x_j - x_i}{b_i}|^3)^3, & \text{if } |\frac{x_j - x_i}{b_i}| < 1 \\ 0, & \text{if } |\frac{x_j - x_i}{b_i}| \geq 1 \end{array} \right\}.$$

## 2.2 Smoothing Parameter Selection

Regardless of which nonparametric regression method is used, some kind of smoothing parameter has to be selected. For local polynomial regression, it can be the number of observations used to compute a local fit (as in the onion example) or a constant window size defining the local neighbourhood. A common feature of all these scenarios is that a small value will lead to an erratic, wiggly fit (the extreme being interpolation of the data points) that does not distinguish signal from noise, whereas a large value will lead to a smooth fit that might not pick up on important features of the underlying curve. As mentioned above there is a trade-off between bias and variance. A small value for the smoothing parameter lead to a fit with small bias but large variance and vice versa for a large value.

An intuitive way of selecting the smoothing parameter when estimating $\beta(\cdot)$ in (1) is to plot $y$ against $x$, superimpose the fitted curve for different smoothing parameter values and then select the one that seems to produce the best fit. In addition to the subjectivity of this method, which lead to different results depending on who is choosing the curve, making the decision by graphical inspection might not always be feasible.

Now, consider estimating a *semiparametric additive model* (SAM) (Hastie and Tibshirani, 1990), which is a more complex problem than estimating (1). Suppose that instead of having one variable, that we want to model nonparametrically, we have $D$ such variables and $Q$ variables that we would like to model parametrically, either because they are categorical variables or because we feel confident that they relate in a certain way (e.g., linearly) to $y$. The simplest SAM, with $D = Q = 1$, is often referred to as a *partially linear model.* A SAM is specified by

$$y_i = \sum_{d=1}^{D} \beta_d(x_{di}) + \sum_{q=1}^{Q} \gamma_q u_{qi} + \epsilon_i. \tag{3}$$

The unknown smooth functions $\beta_d(\cdot)$ and the parameters $\gamma_q$ in (3) are often estimated by *backfitting*, an iterative estimation method using linear smoothers as building blocks, (Buja, Hastie, and Tibshirani, 1989). Note that we have to select $D$ smoothing parameters in this situation. Furthermore, smoothing parameters that are optimal for estimating the nonparametric part will not be optimal for estimating the unknown parameters in (3), which may cause problems when the main focus is on estimating the parametric portion.

In somewhat complex situations, such as the one above, and when smoothing techniques are used repeatedly (possibly as one step in a multifaceted analysis), it is thus preferable to have a more objective and automatic selection procedure than graphical inspection.

A popular smoothing parameter selection criterion is *leave-one-out cross-validation* (CV), which date back to Stone (1974). To estimate $\beta(\cdot)$ in (1), CV selects the smoothing parameter so that it minimize

$$\frac{1}{n} \sum_{i=1}^{n} \big(y_i - \hat{\beta}_h^{-i}(x_i)\big)^2$$

where $\hat{\beta}_h^{-i}(x_i)$ is the fit at $x_i$ computed without $(x_i, y_i)$. CV targets the *predictive squared error*,

$$Var(y_i|x_i) + \frac{1}{n} \sum_{i=1}^{n} MSE\big(\hat{\beta}_h(x_i)|x_1, \ldots, x_n\big),$$

which can be viewed as a suitable criterion if prediction of the response is the main interest. In many situations (e.g., when estimating $E\big(\beta(x)\big)$) theoretical results (e.g., Speckman, 1988, Opsomer and Ruppert, 1999, Imbens, 2004) indicate that CV and other commonly used criteria are not able to select optimal smoothing parameters. Thus, there is a need for new smoothing parameter selection methods.

# 3 Causal Inference

Suppose that we have a random sample of $n$ units from a population of interest and that each unit $i$ receive one of two treatments. Let $z_i = 1$ if unit $i$ receives treatment 1 and $z_i = 0$ if unit $i$ receives treatment 0 (that is, $z$ indicates the treatment). In the Onions example, we have $n$ onion plants, with some of these planted at location $z = 1$ (Purnong Landing) and the rest planted at location $z = 0$ (Virginia). We define the causal effect of treatment $z_i = 1$ versus treatment $z_i = 0$ on a response variable $y$ for unit $i$ as $y_i(1) - y_i(0)$, where $y_i(1)$ and $y_i(0)$ are the *potential outcomes* (Neyman, 1923, Rubin, 1974) for unit $i$ (i.e., the outcome under treatment $z = 1$ and treatment $z = 0$, respectively). In the onion plants example, $y_i(1)$ is the yield that plant $i$ would produce if it were planted at location 1, and $y_i(0)$ is the yield of the same plant if it were planted at location 0. As a plant can only be at one of these locations, the treatment effect, $y_i(1) - y_i(0)$, on yield for a specific plant is unobservable. The fact that only one of the potential outcomes is observed for each unit is referred to as the *fundamental problem of causal inference* (Holland, 1986). The observed response for unit $i$ is

$$y_i = y_i(0)(1 - z_i) + y_i(1)z_i$$

and we are often interested in estimating the population *average treatment effect*,

$$\tau = E\big(y_i(1) - y_i(0)\big), \tag{4}$$

For randomized treatments we can estimate (4) without bias by taking the difference of the mean responses in the two treatment groups. This is typically not the case in *observational studies* with non-randomized treatment assignment. Even if the treatment assignment is not randomized, however, $\tau$ can be identified if we have available a set of covariates $\mathbf{x}_i = (x_{i1}, \dots, x_{id})^T$ that are observed before the treatment assignment and for which the following assumptions hold;

**Assumption 1 (Unconfoundedness)**

$$y_i(1), y_i(0) \perp\!\!\!\perp z_i|\mathbf{x}_i,$$

and

**Assumption 2 (Overlap)**

$$0 < \Pr(z_i = 1|\mathbf{x}_i) < 1.$$

The covariates in $\mathbf{x}_i$ that affect both the outcome and the treatment are called *confounders.* If all confounders are included in $\mathbf{x}_i$, then we have unconfoundedness. In observational studies, nonparametric estimators based on smoothing regression methods are often used to estimate treatment effects. The use of such methods in combination with a large number of covariates does however pose a problem (known as the *curse of dimensionality* (Bellman, 1961)). Fortunately, in this situation we can replace the covariate vector $\mathbf{x}_i$ with the scalar $p(\mathbf{x}_i) = \Pr(z_i = 1|\mathbf{x}_i)$ which is called the *propensity score.* Indeed, Rosenbaum and Rubin (1983) showed that if Assumptions 1 and 2 hold, then

$$y_i(1), y_i(0) \perp\!\!\!\perp z_i|p(\mathbf{x}_i),$$

and

$$0 < \Pr(z_i = 1|p(\mathbf{x}_i)) < 1.$$

## 3.1 Smoothing in Causal Inference

Average treatment effects can be estimated nonparametrically in different ways (see, e.g., Imbens, 2004, and references therein). This thesis considers, an estimator of $\tau$ called the *imputation estimator* by Imbens, Newey, and Ridder (2005). They show that the imputation estimator is, when using series estimation, consistent, asymptotically normal and efficient. Again, the Onions data is used as an illustration. In addition to having observations on onions from Virginia the dataset contains observations on onions produced at another location, Purnong Landing. Figure 4 shows the complete data with 84 observations from both locations.

Now, suppose that we want to estimate the average effect of the treatment "planted in Purnong Landing" versus "planted in Virginia" on the onion yield and that density is the only confounder. An intuitive way to estimate the effect utilizes the following procedure: 1) estimate the expected yield in Virginia using the data from Virginia and then make predictions for the densities present in Purnong Landing and similarly for the expected yield in Purnong Landing; and 2) estimate the average

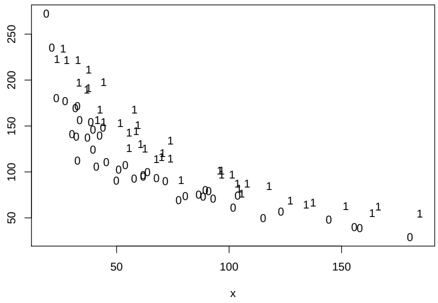

*Figure 4: Complete Onions data. Area density (x-axis) is plotted against yield (y-axis). Observations from onion plants produced in Virginia are represented by 0s, while observations from plants produced in Purnong Landing are represented by 1s.*

treatment effect by taking the mean difference between the estimates from 1. In Figure 5, the yield-functions are estimated with local linear regression (the number of neighbours used is 24 in the first plot and 4 in the second plot). The differences between these, estimates of the conditional treatment effect given area density, are displayed in the third plot in Figure 5. An estimate of $\tau$ is then the average of the conditional treatment effect.

The procedure described above can be formalized in the following manner. The model for the response is

$$y_i = \beta_0(x_i) + \tau(x_i)z_i + \epsilon_i = \beta_0(x_i) + \big(\beta_1(x_i) - \beta_0(x_i)\big)z_i + \epsilon_i$$

with $E(\epsilon_i|x_i, z_i) = 0$. We now have that the average treatment effect equals

$$\tau = E\big(E(y_i|z_i = 1, x_i) - E(y_i|z_i = 0, x_i)\big) = E(\beta_1(x_i)) - E(\beta_0(x_i)).$$

Denote the observed response and covariate vectors for the $n_0$ units with treatment $z_i = 0$ by $\mathbf{y}^0 = (y_1^0, \dots, y_{n_0}^0)^T$ and $\mathbf{x}^0 = (x_1^0, \dots, x_{n_0}^0)^T$, respectively. Similarly, we have $\mathbf{y}^1 = (y_1^1, \dots, y_{n_1}^1)^T$ and $\mathbf{x}^1 = (x_1^1, \dots, x_{n_1}^1)^T$ for the $n_1$ units with treatment $z_i = 1$. Then, we can write the estimator

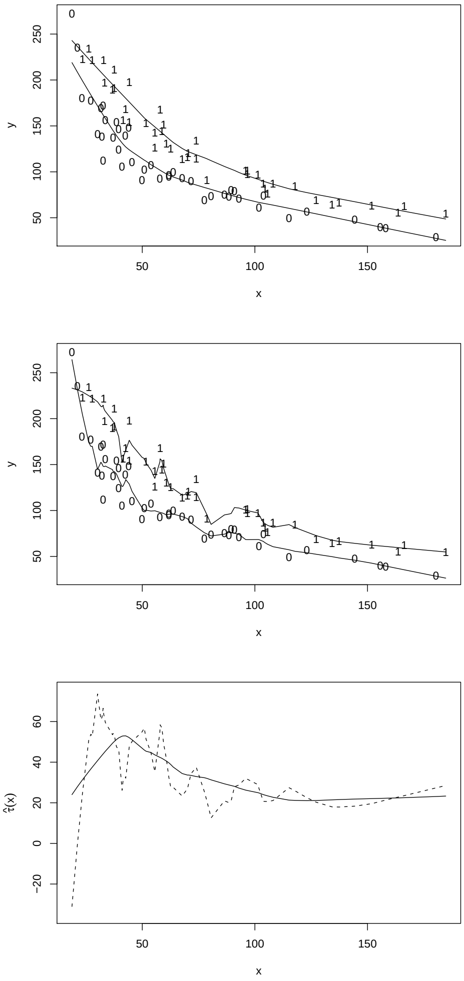

*Figure 5: The estimated yield and treatment effect conditional on the area density using local linear regression. The estimated yield conditional on the area density for the different locations using local linear regression based on 25 and 5 observations is shown in the first and second plot, respectively. In the third plot, estimated treatment effects conditional on area density, the difference between the functions in the first (solid) and second (dahsed) plot is shown.*

of $\beta_j(\mathbf{x}) = (\beta_j(x_1), \ldots, \beta_j(x_n))^T$, $j = 1, 0$, as

$$\hat{\beta}_j^{h_j}(\mathbf{x}) = S_j^{h_j}[\mathbf{x}]\mathbf{y}^j,$$

where $\mathbf{x} = (\mathbf{x}^{0T}, \mathbf{x}^{1T})^T$, and $S_j^{h_j}[\mathbf{x}]$ is the smoothing matrix regressing $\mathbf{y}^j$ on $\mathbf{x}^j$ using smoothing parameter $h_j$. The imputation estimator of $\tau$ is

$$\hat{\tau}^{imp} = \frac{1}{n}\sum_{i=1}^{n} \hat{\tau}^{imp}(x_i) = \frac{1}{n}\sum_{i=1}^{n} \big(\hat{\beta}_1^{h_1}(x_i) - \hat{\beta}_0^{h_0}(x_i)\big). \qquad (5)$$

As can be seen in (5), computing $\hat{\tau}^{imp}$ requires the selection of two smoothing parameters; as was mentioned in Section 2, common methods (such as CV) are not suitable for smoothing parameter selection in this case, as we are estimating a functional of regression curves .

# 4 Summary of Papers

## 4.1 Paper I: Estimating Prediction Error: Cross-Validation vs. Accumulated Prediction Error

Paper I deals with the validation of general prediction rules such as the regression models obtained using a sample of size $n$ for a variable of interest $y$. We consider out-of-sample methods that, in contrast to covariance penalty methods, are nonparametric and allow for the comparison of prediction rules obtained with different inferential frameworks and/or different modelling strategies. The focus is on the accumulated prediction error (APE). The APE measures the generalizability of a prediction rule by the mean of $n - m$ prediction errors, where the $i$th prediction is made using a subsample of size $i - 1$, and $m$ is the size of the subsample on which the first prediction rule is obtained, $i = m + 1, \ldots, n$. Comparisons are made with CV. The two methods can be viewed as estimating different measures of prediction error that answer different questions: “How good is a prediction rule at predicting $n$ observations arriving sequentially?” and “How well will a prediction rule predict a future dataset of size $n$ independently generated by the same mechanism that produced the original dataset?”, respectively. We consider prediction rules which are linear smoothers and illustrate the use of APE and

CV by focusing on smoothing parameter selection in local polynomial regression. Using simulations, we show that APE and CV lead to different optimal smoothing parameters and that this difference does not vanish with increasing sample size.

### 4.2 Paper II: Bandwidth Selection for Backfitting Estimation of Semiparametric Additive Models

Paper II considers optimal smoothing parameter selection for estimating the parametric part of SAMs, (3) above, by backfitting. Asymptotic results have suggested that selecting the smoothing parameters by CV will not lead to $\sqrt{n}-$consistent estimation of the parametric part (Speckman, 1988, Opsomer and Ruppert, 1999). We propose selecting the smoothing parameters by a double smoothing estimator, which is a two-step estimator of the mean squared error that target the parameter of interest as opposed to predicting the response. Comparisons with CV and empirical bias bandwidth selection (EBBS) (Opsomer and Ruppert, 1999) are made using simulation experiments. For the partially linear model, we also make comparisons with a $\sqrt{n}$-consistent estimator consisting of CV and an estimator other than backfitting (Speckman, 1988). The results show that the double smoothing procedure is preferable to CV and in most cases to EBBS. For the partially linear model, DS and EBBS were found to be comparable to the $\sqrt{n}-$consistent procedure for large samples sizes.

### 4.3 Paper III: Data-Driven Smoothing Parameter Selection for Estimating Average Treatment Effects

In this paper we consider estimating the average treatment effect of a binary treatment $z$ on an outcome of interest $y$ using the imputation estimator $\hat{\tau}^{imp}$ described above. To estimate $\tau$, we have to select two smoothing parameters. It is known that smoothing parameters that are asymptotically optimal for estimating the regression functions $\beta_j(\cdot)$ do not lead to $\sqrt{n}$-consistent estimation of functionals of these regression functions (e.g., the average treatment effect). We show that $h_j \propto n^r$ for $-1 < r < -1/4$ results in $\hat{\tau}^{imp}$ being a $\sqrt{n}$-consistent estimator of $\tau$. This result implies that the smoothing parameters optimal for estimating $\tau$ are asymptotically smaller than what is considered optimal for estimating $\beta_j(\cdot)$. We propose double smoothing estimators of the mean squared errors that target the estimation of $\tau$ and, using simulation, compare the performance of these with a method proposed by Imbens et al. (2005) and with CV.

### 4.4 Paper IV: Divorcing in Middle Age and Its Effects on BMI

In the last paper, we study the effect of divorce on change in body mass index (BMI) in individuals divorcing between 40 and 60 years of age. We use data from the Linnaeus database (Malmberg, Nilsson, and Weinehall, 2010) to estimate the average treatment effect on the treated. The Linnaeus database contains information from nationwide registers on socioeconomic and demographic conditions that are individually linked to health and lifestyle information from participants in the Västerbotten Intervention Programme (VIP) (Norberg, Wall, Boman, and Weinehall, 2010). We apply a method proposed in Paper III to select smoothing parameters. We find that women divorcing between 40 and 50 have a lower increase in BMI compared to those not divorcing. A similar effect is found for normal weight women divorcing between 50 and 60. For men no significant results are found.

## 5 Final Remarks and Further Research

In this thesis we have proposed double smoothing estimators of mean squared errors and minimization of these estimators for selecting smoothing parameters in different settings. Using simulation studies that compared these double smoothing methods to common selection methods, such as CV, we show that double smoothing often improves on other methods (resulting in lower mean squared errors).

A natural step in continuing this line of research is to derive large sample properties of smoothing parameters selected by double smoothing and properties of the resulting estimators that are functionals of these smoothing parameters.

Implementing our procedures in R, which will make selection of smoothing parameters using double smoothing estimators easier for practitioners, is also a major aim.

## References

Bellman, R. (1961). *Adaptive Control Processes.* Princeton University Press.

Buja, A., T. Hastie, and R. Tibshirani (1989). Linear smoothers and additive models. *The Annals of Statistics 17*, 453–510.

Cleveland, W. S. (1979). Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association 74*, 829–836.

Fan, J. and I. Gijbels (1996). *Local Polynomial Modelling and Its Applications.* Chapman and Hall, London.

Hastie, T. and R. Tibshirani (1990). *Generalized Additive Models.* Chapman and Hall, Boca Raton.

Holland, P. W. (1986). Statistics and causal inference. *Journal of the American Statistical Association 81*, 945–960.

Imbens, G. W. (2004). Nonparametric estimation of average treatment effects under exogeneity: A review. *The Review of Economics and Statistics 86*, 4–29.

Imbens, G. W., W. Newey, and G. Ridder (2005). Mean-squared-error calculations for average treatment effects. IEPR Working Papers 05.34, Institute of Economic Policy Research (IEPR).

Malmberg, G., L.-G. Nilsson, and L. Weinehall (2010). Longitudinal data for interdisciplinary ageing research : Design of the linnaeus database. *Scandinavian Journal of Public Health*, 1–7.

Neyman, J. (1923). On the application of probability theory to agricultural experiments. essay on principles. Section 9.(1990), translated (with discussion). *Statistical Science 5*, 465–480.

Norberg, M., S. Wall, K. Boman, and L. Weinehall (2010). The Västerbotten intervention programme : background, design and implications. *Global health action 3.*

Opsomer, J. D. and D. Ruppert (1999). A root-n consistent backfitting estimator for semiparametric additive modeling. *Journal of Computational and Graphical Statistics 8*, 715–732.

R Development Core Team (2010). *R: A Language and Environment for Statistical Computing.* Vienna, Austria: R Foundation for Statistical Computing. ISBN 3-900051-07-0.

Rosenbaum, P. and D. Rubin (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika 70*, 41–55.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology 66*, 688–701.

Speckman, P. (1988). Kernel smoothing in partial linear models. *Journal of the Royal Statistical Society Series B 50*, 413–436.

Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions (with discussion). *Journal of the Royal Statistical Society Series B 36*, 111–147.