Department of Mathematical Statistics
Royal Institute of Technology, Stockholm, Sweden
Degree Project in Engineering Physics, First Level

A Comparison of Regression Models for Count Data in Third Party Automobile Insurance

Author:
Annelie Johansson
annelieg@kth.se

Supervisor:
Boualem Djehiche

May 19, 2014

Abstract

In the area of pricing insurances, many statistical tools are used. The number of tools that exist are overwhelming and it is difficult to know which one to choose. In this thesis five regression models are compared on how good they fit the number of reported claims in third party automobile insurance. The models considered are OLS, Poisson, Negative Binomial and two Hurdle models. The Hurdle models are based on Poisson regression and Negative Binomial regression respectively, but with additional number of zeros. The AIC and BIC statistics are considered for all the models and the predicted number of claims are calculated and compared to the observed number of claims. Also, a hypothesis test for the null hypothesis that the Hurdle models are not needed is performed. The OLS regression is not suitable for this kind of data. This can be explained by the fact that the number of claims are not normally distributed. This is the case because many policyholders never report any claims and the data therefore includes an excess number of zeros. Also, the number of claims can never be negative. The other four models are considerably better and all of them fit the data satisfactory. The one of them that performs best in one test is inadequate in another. The Negative Binomial model is a bit better than the other models, but the model choice is not obvious. The conclusion is not that a specific model is preferable, but that one need to choose model critically.

Keywords: Regression Models, Insurance, Count Data, Regression Analysis, Hurdle Regression

Sammanfattning

Det finns många statistiska hjälpmedel att använda inom prissättning av försäkringar. Antalet hjälpmedel är överväldigande och det är svårt att veta vad man ska använda sig av. I den här studentuppsatsen kommer fem regressionsmodeller att jämföras i hur bra de passar för antalet anmälda skador inom trafikförsäkring. De fem modellerna är OLS, Poisson, Negative Binomial och två Hurdle-modeller. Hurdle-modellerna är baserade på Poisson och Negative Binomial respektive, men de är anpassade för ett större antal nollor. Värdena av AIC och BIC jämförs för de olika modellerna och antalet predikterade anmälda skador jämförs med antalet observerade skador. Ett hypotestest är också utfört för att testa nollhypotesen att Hurdle-modellerna inte behövs. Den modell som passar sämst till datan är OLS. Det kan förklaras av att antalet anmälda skador inte är normalfördelade som man antar vid OLS. Det kan bero på att många försäkringstagare aldrig anmäler någon skada och att det därför finns ett överskott av nollor i datan. Det kan även bero på att antalet anmälda skador aldrig kan bli negativt. De andra fyra modellerna är mycket bättre och de kan anpassas till datan på ett acceptabelt sätt. Den av modellerna som är bäst i ett test är sämre i ett annat. Negative Binomial-modellen är aningen bättre än övriga modeller, men det är inte självklart vilken modell som bör användas. Slutsatsen är att det inte finns någon perfekt modell, utan att man måste välja genom att vara kritisk.

Nyckelord: Regressionsmodeller, Försäkring, Räknedata, Regressionsanalys, Hurdleregression

Acknowledgements

The author would like to thank the supervisor Boualem Djehiche for an interesting input into the insurance business and for valuable opinions.

Contents

1 Introduction

In many cases it is of interest to find how different variables affect another variable. With statistical methods one can describe mathematically how the variables of interest (the explanatory variables) affect the resulting variable (the outcome of interest). By omitting the variables that do not have a statistically significant affect on the outcome of interest the final model will be given and it can be used to predict how the outcome of interest will behave in the future. This is called regression analysis. The main problem with these kind of analysis is that there exist several different models and it is hard to choose the best one. It is also of importance to be critical to the result of the prediction, the variables included may not be sufficient to make a good prediction and the number of observations may perhaps not be enough.

In actuarial automobile insurance applications, the outcome of interest often is the number of claims or the amount of money that the claims cost. These two variables are then modeled to depend on different characteristics of the car and the owner of the car. These kind of characteristics can be the car make, the age of the car, the age of the driver, the area where the driver lives and so on. The result of the model is then used to set a fair price on the insurance.

The insurance market is very competing and the companies must develop there insurances in time with the market. In contrast to many other markets the insurance market does not supply any goods, the policyholder only get an insurance when subscribing to it and an insurance is often not used. Thus the policyholder in many cases pays for nothing. This implies that the price is an important issue for the insurance companies. To look at it even more straight forward, consider the following example. One insurance company price an insurance without taking care of any explanatory variables. Those with lower expected number of claims find a better price with another company and those with a larger number of expected claims stay. The company's expenditures will increase which leads to the result that the price also must increase (Frees, 2010). The same situation occurs if the company would price the insurance based on a model that is far away from reality. This explains why the model choice is so important and in this thesis models for predicting the number of claims, count data models, will be compared.

With respect to how important it is to choose the best model and how many models and model specifications that exist one could presume that there are many comparisons of models in the literature, but this is not the case (Lord et al., 2004).

However, there are several analysis of this kind that compares models for count data. One area that uses count data models is the prediction of motor vehicle crashes and this area has been developed during the last 20 years (Lord et al., 2004). Also in the actuarial area model comparisons for count data models have been practiced (Boucher et al., 2007). In most of these studies the models are parametric, which means that they contain the information of the probability that a policyholder report k number of claims during one year (Boucher et al., 2007).

Originally the linear regression model OLS was used when modeling count data (Frees, 2010). The underlying assumption is that the outcome of interest is normally distributed (Frees, 2010). The problem with this is that the outcome of interest often is zero but never negative. In later time the Poisson regression model and the Negative Binomial model has been used (Frees, 2010). These models exist in many variations, for example the Generalized Poisson-Pascal distribution and the Poisson-Goncharov distribution (Boucher et al., 2007). If the data includes more zeros than the mentioned models can predict, the models can be modificated (Frees, 2010). One such modification is the Hurdle model (Frees, 2010).

The focus of this thesis is to compare five regression models, OLS, Poisson, Negative Binomial, Hurdle based on Poisson and Hurdle based on Negative Binomial. The question is which one of these models is the best choice for predicting the number of claims that an insurer will have in one year from the third-party automobile insurance. They are described mathematically in chapter 3. The OLS is included to illustrate how much better fit one can get by using an appropriate model. The five models are the only models that will be considered and only one dataset will be used. The resulting explanatory variables that comes out from the regressions will not be considered, apart from omitting those that are not statistical significant at level 99 percents.

The result shows that the OLS is a bad choice for modeling claims. The other models do better and all of them fit the data well. The Negative Binomial model is a bit better than the other models but not much. The Hurdle models do not give statistically significant better results than the Poisson and Negative Binomial models but they are the best ones in predicting the number of zeros.

2 Method

In the regressions performed in this thesis, the data is fitted to a the specific models by a method called Maximum Likelihood. This means that the unknown coefficients are estimated such as the likelihood of getting the given data is as large as possible. When including a new explanatory variable into a model, the likelihood will increase even if the new variable does not have anything to do with the outcome of interest, so comparing the different likelihood is not always good (Frees, 2010). Instead the AIC and BIC test statistics is used. These are two common ways of comparing likelihoods between different models with respect to the number of estimated parameters. These are described mathematically in chapter 3.

The expected number of claims according to the different models are also be compared to the observed number of claims, both graphically and in table format. This is a good way of getting an idea of how the different models works and visualize there estimates. As can be seen in A1, some of the explanatory variables are not statistically significant. When pricing an insurance those explanatory variables could lead to a price that is unfair. Therefore the models are compared with the insignificant variables omitted. The question if the Hurdle models are contributing anything to the prediction is answered by a hypothesis test with the null hypothesis that no Hurdle is needed.

At last, some test groups are chosen and the predicted number of claims for these test groups are compared with the observed number of claims for the different models. This is important since a model that has good test statistics when looking at the whole dataset, but fails when it comes to a specific group, may not be useful. The test groups are not chosen after any specific formula.

2.1 Data Description

The dataset is found on the website statsci.org and was compiled by a Swedish Committee on the Analysis of Risk Premium in Motor Insurance in 1977. It contains observations of the number of reported claims and the number of policyholders from 2182 number of categories. The number of reported claims and the number of policyholders are given in policy-years. Each category is a collection of all policyholders that live in one of 7 areas of Sweden, have one of 8 car models, have the same number of bonus and have a car that is driven a specific length each year. The year category is divided into 5 parts. The bonus variable is a bonus that one can get by not reporting any claims in one or more years. For example, one category includes all policyholders living in Stockholm, with bonus of zero and with car model 1 that is driven less than 1000 km per year. The number of policyholders that fit into this category is the observation of the number of policyholders and the observation of the number of reported claims is the sum of the claims reported by these policyholders.

2.2 R Software

The regressions are made in R software where all models used could be fitted by already existing functions. The Linear model is done by the function `lm()`.The Poisson model is fitted by the function `glm()`that can be found in the package **stats**. The function `glm.nb()`is found in the package **MASS** and is needed to perform Negative Binomial regression. Hurdle regression is performed by the function `hurdle()`in package **pscl** (Zeileis et al., 2008). The R code can be delivered if requested.

3 Theory

In this chapter the anticipatory statistical theory is explained mathematically. The first section will define common existent matrices and terms. The following sections will explain the regression models that are used in this thesis. After that, some methods that enable us to compare the different models are explained.

3.1 Basic Definitions

Here three matrices are defined. The first one is $\mathbf{y}$ which include all the observations of the outcome of interest. The second one is the matrix containing the observations of the explanatory variables, denoted by $\mathbf{X}$. The third one is a vector with the unknown intercept and the regression coefficients and is denoted by $\boldsymbol{\beta}$.

$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} \qquad \mathbf{X} = \begin{pmatrix} 1 & x_{11} & \dots & x_{1k} \\ 1 & x_{21} & \dots & x_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \dots & x_{n,k} \end{pmatrix} \qquad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{pmatrix}$$

When a symbol is written with a hat it means that the symbol is estimated, $\hat{\boldsymbol{\beta}}$ is for example the estimation of the intercept and the regression coefficients. The number of explanatory variables is k and the number of observations is n.

3.2 Multiple Linear Regression

The simplest of the regression models considered is the Multiple Linear regression model, here called OLS. One assumes that the outcome of interest y depends linearly on a number of explanatory variables $x_1, x_2, ...x_k$, as the following equation shows.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k + \epsilon, \tag{3.1}$$

where ϵ is the error term. The result of interest is to obtain an estimation of the unknown intercept β_0 and the regression coefficients $\beta_1, \beta_2, ..., \beta_k$. To implement this, several observations of y and $x_1, x_2, ...x_k$ are made. The number of observations normally exceeds the number of explanatory variables, so the method of least squares can be used to obtain the estimators $\hat{\beta}_0, \hat{\beta}_1, ..., \hat{\beta}_k$. The method of least squares means finding the values of $\hat{\beta}_0, \hat{\beta}_1, ..., \hat{\beta}_k$ that minimizes the Sum of Squares defined in equation 3.2.

$$SS(\hat{\beta}_0, \hat{\beta}_1, ..., \hat{\beta}_k) = \sum_{i=1}^{n} (y_i - (\hat{\beta}_0 + \hat{\beta}_{i1} x_1 + ... + \hat{\beta}_k x_{ik}))^2. \tag{3.2}$$

So let us calculate the partial derivatives of the Sum of Squares and find the point were they equal to zero.

$$\begin{gathered}\frac{\partial}{\partial \hat{\beta}_j} SS(\hat{\beta}_0, \hat{\beta}_1, ..., \hat{\beta}_k) = \sum_{i=1}^{n} -2x_{ij}(y_i - (\hat{\beta}_0 + \hat{\beta}_{i1} x_1 + ... + \hat{\beta}_k x_{ik})) = 0, \\ j = 0, 1, ..., k.\end{gathered} \tag{3.3}$$

Using the matrix notation introduced in previous section makes it possible to rewrite equation 3.3 in a more compact way and solve it to obtain the unknown parameters.

$$\mathbf{X}^\top \mathbf{X} \hat{\boldsymbol{\beta}} = \mathbf{X}^\top \mathbf{y} \implies \hat{\boldsymbol{\beta}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}. \tag{3.4}$$

(Frees, 2010).

3.3 Poisson Regression

Consider the Poisson distribution in equation 3.5.

$$Pr(y = j) = \frac{\mu^j}{j!} e^{-\mu}. \tag{3.5}$$

In this distribution the mean equals the variance and is represented by μ. When performing a regression with several explanatory variables one assumes that μ differs with the different values of the explanatory variables. Usually this relationship is calculated as follows.

$$\mu_i = e^{\boldsymbol{x_i} \boldsymbol{\beta}}. \tag{3.6}$$

Here, $\boldsymbol{x_i}$ means the ith row in $\mathbf{X}$. The exponential is used as a link function, it reassure us that μ_i will always be positive. Other link functions can also be used but in this case the log link function is enough. Another thing to take into account is that sometimes a difference between the observations is already known in such a way that one can assume that one observation should have a greater outcome of interest than another. For the third party insurance claims data there is one column called *Insured* that include the number of policyholders in policy-years and it is fair to presume that an observation with a greater number of insured also will have a greater number of claims. This is solved mathematically by including the *exposure* E_i to equation 3.6.

$$\mu_i = E_i e^{\boldsymbol{x_i}\boldsymbol{\beta}}. \tag{3.7}$$

The *exposure* here equals the column *Insured* (Frees, 2010).

Now the parameters $\boldsymbol{\beta}$ can be estimated by the maximum likelihood estimator. This is done by maximizing the log-likelihood $L(\boldsymbol{\beta})$.

$$L(\boldsymbol{\beta}) = \sum_{i=1}^{n} (-E_i e^{\boldsymbol{x_i}\boldsymbol{\beta}} + y_i(lnE_i + \boldsymbol{x_i}\boldsymbol{\beta}) - lny_i!). \tag{3.8}$$

In order to do so the values of the parameters in $\boldsymbol{\beta}$ that makes the derivative of 3.8 equal to zero are found numerically.

$$\frac{\partial}{\partial\boldsymbol{\beta}} L(\boldsymbol{\beta})\bigg|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}} = \sum_{i=1}^{n} -E_i \boldsymbol{x_i} e^{\boldsymbol{x_i}\hat{\boldsymbol{\beta}}} + y_i \boldsymbol{x_i} = \mathbf{0}. \tag{3.9}$$

The vector $\hat{\boldsymbol{\beta}}$ is the maximum likelihood estimator of $\boldsymbol{\beta}$ (Frees, 2010).

3.4 Negative Binomial Regression

The Negative Binomial regression model is viewed in many varieties, Hilbe lists 22 Negative Binomial regression models in his book *Negative Binomial Regression* (Hilbe, 2011). Among all these models the standard formula belongs to the NB2 model (Cameron and Trivedi, 1998). The NB2 model is the one that is used in this thesis.

The Negative Binomial probability mass function looks like

$$Pr(y = j) = \frac{\Gamma(j+\theta)}{\Gamma(j+1)\Gamma(\theta)} (\frac{\theta}{\theta+\mu})^{\theta} (\frac{\mu}{\theta+\mu})^{j} \tag{3.10}$$

(Cameron and Trivedi, 1998). The gamma function is specified by $\Gamma(a) = \int_0^{\inf} e^{-t} t^{a-1} dt$ (Boucher et al., 2007). The parameters that are to be estimated are the mean μ and the shape parameter θ. The mean is assumed to be dependent of the categorical variables the same way as before, recall equation 3.7. Also as before, the unknown parameters will be estimated by the maximum log-likelihood estimator. The log-likelihood is calculated as follows.

$$L(\boldsymbol{\beta},\theta) = \sum_{i=1}^{n} lnf(y_i,\mu_i,\theta) = \sum_{i=1}^{n} ln(\frac{\Gamma(y_i+\theta)}{\Gamma(y_i+1)\Gamma(\theta)}) + ln((\frac{\theta}{\theta+E_i e^{\boldsymbol{x_i\beta}}})^{\theta}(\frac{E_i e^{\boldsymbol{x_i\beta}}}{\theta+E_i e^{\boldsymbol{x_i\beta}}})^{y_i})$$

$$= \sum_{i=1}^{n} ln(\frac{\Gamma(y_i+\theta)}{\Gamma(y_i+1)\Gamma(\theta)}) + y_i ln(E_i) + y_i\boldsymbol{x_i\beta} - y_i ln(\theta) - (\theta+y_i)ln(1+\theta^{-1}E_i e^{\boldsymbol{x_i\beta}}). \tag{3.11}$$

The first term, $ln(\frac{\Gamma(y_i+\theta)}{\Gamma(y_i+1)\Gamma(\theta)})$, is equal to $\sum_{k=0}^{y_1-1} ln(k+\theta) - ln(y_i!)$ (Cameron and Trivedi, 1998).

As before, the maximum is found by by derivate the log-likelihood. Since there are two unknown parameters the log-likelihood needs to be derived with respect to both $\boldsymbol{\beta}$ and θ. Calculations give the derivatives shown in the following two equations.

$$\frac{\partial}{\partial\boldsymbol{\beta}} L(\boldsymbol{\beta},\theta)\Big|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}} = \sum_{i=1}^{n} \boldsymbol{x_i} \frac{y_i - E_i e^{\boldsymbol{x_i}\hat{\boldsymbol{\beta}}}}{1+\theta^{-1}e^{\boldsymbol{x_i}\hat{\boldsymbol{\beta}}}} = \mathbf{0}. \tag{3.12}$$

$$\frac{\partial}{\partial\theta} L(\boldsymbol{\beta},\theta)\Big|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}} = \sum_{i=1}^{n}\sum_{k=0}^{y_1-1} \frac{1}{k+\theta} - ln(1+\theta^{-1}E_i e^{\boldsymbol{x_i}\hat{\boldsymbol{\beta}}}) + \frac{E_i e^{\boldsymbol{x_i}\hat{\boldsymbol{\beta}}} - y_i}{\theta + E_i e^{\boldsymbol{x_i}\hat{\boldsymbol{\beta}}}} = 0. \tag{3.13}$$

Equations 3.12 and 3.13 are solved by fitting $\hat{\boldsymbol{\beta}}$ for different values of θ and vice versa (Zeileis et al., 2008).

3.5 Hurdle regression

In third party automobile insurance two decisions must be passed before a claim is made. First, the owner of the damaged car must decide to claim for payment from the one that caused the damage. Then, the one that caused the damage, the policyholder, must decide to report the accident to the insurance company. This leads to more zeros than if the result was only a response of one decision. One model that can deal with this is the Hurdle model (Frees, 2010).

The Hurdle model implies to divide the probability mass function into two parts, the first part is the probability of a zero and the second part adjusts the non-zero values to a probability distribution like Poisson or Negative Binomial. Let $f_1(0)$ represent the probability of a zero and $f_2(j)$ be the probability mass function for the distribution that is used. Then the Hurdle mass function can be stated.

$$Pr(y=j) = \begin{cases} f_1(0), & j=0, \\ \frac{1-f_1(0)}{1-f_2(0)} f_2(j), & j>0. \end{cases} \tag{3.14}$$

If using the Binomial distribution for the zeros, the probability of a zero is related to the explanatory variables through a link function. In this thesis the logit link function, $f_{i1}(0) = \frac{e^{\boldsymbol{x_i \beta_1}}}{1+e^{\boldsymbol{x_i \beta_1}}}$, is mostly used. The mean in the probability distribution used in $f_2()$ is calculated as before with the log link function, $\mu_i = E_1 e^{\boldsymbol{x_i \beta_2}}$. The same explanatory variables $\boldsymbol{x}_i$ are used in both cases, but the fitted parameters β must not equal each other, why the indexes 1 and 2 are used. If the $\boldsymbol{\beta}$ for the function $f_0(0)$ equals the coefficients for the count part and the both functions equals each other, the model reduces to the underlying regression model. When testing if the Hurdle is necessary or not, the functions and the link functions must be equal. Therefore the log link function is used for the zero and count part in section 4.2 and the function for zeros will equal the function for counts. In the other sections the logit link function and Binomial distribution will be used for the probability of zero counts.

The log likelihood for the Hurde model is defined below.

$$\begin{aligned} L(\boldsymbol{\beta_1}, \boldsymbol{\beta}_2) &= \sum_{i=1}^{n} lnf(y_i, \mu_i, f_{i1}(0)) = \sum_{i=1}^{n} ln(1(y_i) f_{i1}(0) + (1-1(y_i)) \frac{1-f_1(0)}{1-f_2(0)} f_2(j)) \\ &= \sum_{i=1}^{n} (1(y_i) ln(f_{i1}(0)) + (1-1(y_i))[ln(1-f_{i1}(0)) + ln(\frac{f_{i2}(y_i)}{1-f_{i2}(y_i)})]. \end{aligned} \tag{3.15}$$

Here, $1(y_i)$ equals 1 if $y_i = 0$ and zero otherwise. If the Negative Binomial function is used, the parameter θ also should be included in f. Since $f_2()$ can be either Poisson or Negative Binomial in this thesis the derivatives will not be written, but the unknown parameters are found in the same way as before (Frees, 2010).

3.6 Test Statistics

When estimating the unknown parameters, the maximum likelihood is what to reach for. Thus, the models with greater likelihood seem to be a better fit. But to compare the log likelihood with each other can be problematic, because when including a new parameter to the model the log likelihood will always increase. In the Negative Binomial model for example, there is one more parameter, θ, than in the Poisson model. One way of comparing models with respect to this is to use the Akaike Information Criteria, AIC, or the Bayesian Information Criteria, BIC. They are defined as AIC=-2log(L) + 2p and BIC=-2log(L) + plog(n), where L is the log likelihood, p is the number of estimated parameters and n is the number of observations. The model with the lowest value of AIC and BIC is the one with the best fit(Boucher et al., 2007).

To test whether the Hurdle is needed or not, one can use the Wald test or the Likelihood Ratio test LR. The null hypothesis for the test is $H0 : \beta_1 = \beta_2$. The LR is a measurement of the ratio between the likelihood under the null hypothesis and the likelihood without the null hypothesis. It is calculated as follows:

$$LR = -2log(L_{H_0}/L) = 2(log(L) - log(L_{H_0})). \tag{3.16}$$

Thus, it is the doubled difference between the log likelihood functions. For sufficient many observations this statistics has a χ^2-distribution with the degrees number of freedom r. The Wald test is the difference between the Maximum Likelihood Estimator without hypotheses and with hypotheses, normalized by the standard deviation.

$$\frac{(MLE - MLE_{H_0})^2}{MLE(stand.dev.)}. \tag{3.17}$$

This statistics is distributed as the LR statistics (Harrell Jr., 2001). The degrees of freedom r for these tests is the number of restrictions under the null hypothesis (Oya and Oya, 1997).

4 Results

The estimated coefficients are not presented here since the aim of this study is to compare regression models and not to investigate which explanatory variables are useful and which are not. The interested reader can find the regression coefficients with standard deviation and significance level in A1. The focus here is to compare the number of predicted counts for the different models with the observed number of counts. This is done for the original models in the first section, for the Hurdle models in the second section, and for ten test groups in the last section. The AIC and BIC statistics are also compared for the original models, the Hurdle models and the modified models where insignificant explanatory variables are omitted. A test that investigate whether a Hurdle is needed or not is presented in section 4.2 together with the other results for the Hurdle models.

4.1 Original Models

The AIC and BIC test statistics are shown in Table 4.1. The OLS Regression extends, it has an AIC and BIC value about twice as much as the other models. With respect to these test statistics the NB2 model is the best one because it has the lowest of both test statistics. Both Poisson and NB2 do better than the corresponding models with Hurdles.

	OLS	Poisson	NB2	Hurdle-Poisson	Hurdle-NB2
AIC	20474	10654	10380	10928	10655
BIC	20764	10796	10527	11213	10945

Table 4.1: The AIC and BIC for the different models.

In Table 4.2 the original models are compared with respect to how many number of counts they predict. The number of predicted zeros for the OLS model is a sum of the predictions that is less than one, between one and two etc. If using less than 0.5 instead of one, the number is still over 500, so the linear model is not very close, as expected. The Poisson and NB2 models are very similar to each other in this view, their predicted number of one claims, two claims, four claims, five claims etc. are equal. The only models that could predict exactly 385 number of zero claims as observed was the Hurdle models. This is not surprisingly since the focus of the Hurdle regression is the number of zeros. For the other number of claims the Hurdle models make good predictions. The Hurdle model with the Poisson distribution for the non-zero counts is closer to the observed number of claims than the Hurdle model with the NB2. Both Hurdle models in this section has the logit link function and Binomial distribution for the zeros. The results from Table 4.2 are shown graphically in Figure 4.1 where the densities of the number of predicted and observed counts are shown up to 20 number of claims. In this figure the Poisson and NB2 models looks closest to the observed densities. Inspect for example the start value for the different graphs, the Hurdle models have too large start values, but the Poisson and NB2 have better ones. The OLS is not similar at all to the observed values, but all other four models follow the observed curve quite well.

	Observed	OLS	Poisson	NB2	Hurdle-Poisson	Hurdle-NB2
0	385	594	371	374	385	385
1	214	91	223	223	247	249
2	169	94	167	167	167	167
3	131	70	132	131	127	126
4	100	73	106	106	100	100
5	98	74	87	87	82	82
6	66	54	73	72	69	69
7	65	66	62	62	59	59
8	51	56	54	54	52	51
9	46	48	47	47	46	45
10	38	47	42	42	41	41
100	1	2	2	2	2	2

Table 4.2: A comparison between the predicted number of zeros, ones etc. from the different models.

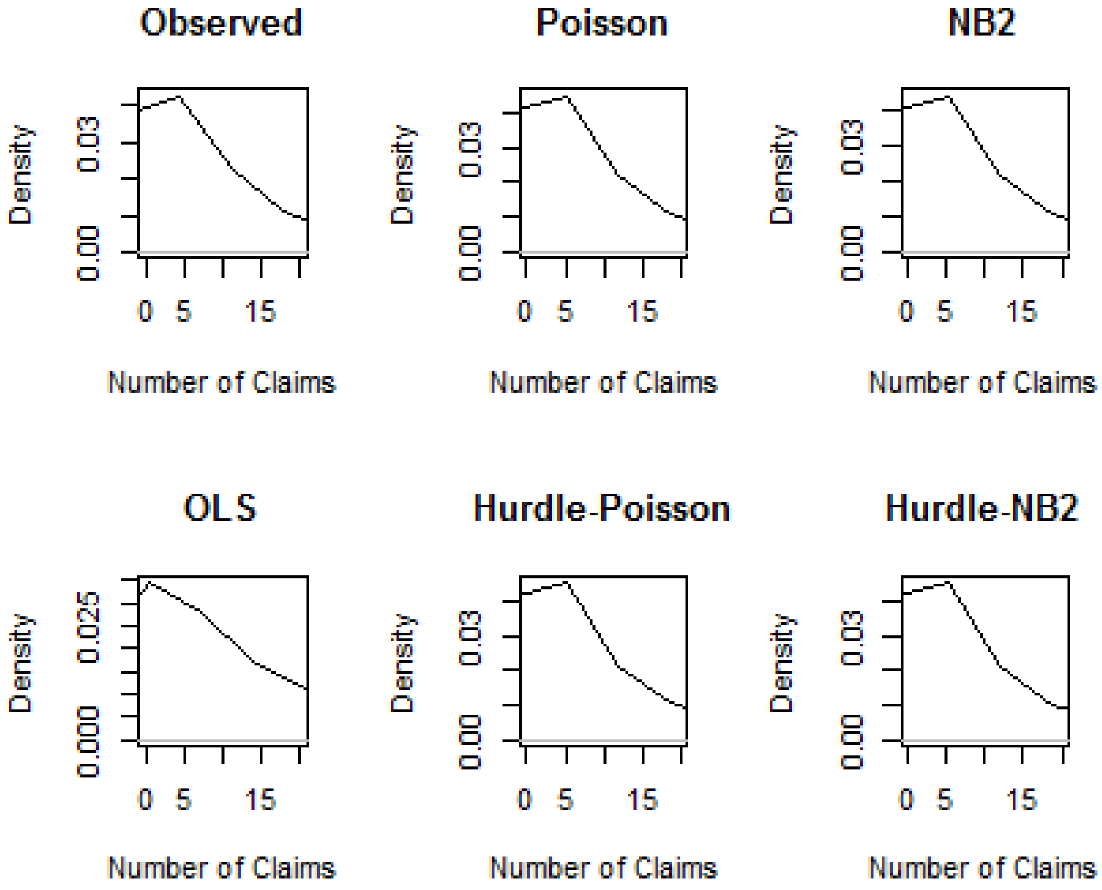

Figure 4.1: The densities of the number of claims for the observed values and for the predicted values from the different models.

4.2 Hurdle Models

When considering the Hurdle regression, the first question to answer is if the Hurdle is needed or not. As described in section 3.6 the null hypothesis that no hurdle is needed can be tested with the LR test and the Wald test. The resulting test values and p-values for these tests are shown in Table 4.3, both for the model with Poisson distribution and NB2 distribution. The test values are not very large in any of the tests. This means there is not a big difference between the models with a Hurdle compared to the models without a Hurdle. The test statistic of the χ^2-distribution with 25 degrees of freedom and significance level 0.99 is 44.3 and the statistic for 26 degrees of freedom is 45.6. All test values are lower than these statistics and the null hypothesis that no Hurdle is needed can not be rejected. All four p-values are large, so the probability of getting the observed data if the null hypothesis is true is very large. This is a hint that the Hurdle not is useful in this case.

Model	Test	value	p
Poisson	LR	16.661	0.8937
Poisson	Wald	16.201	0.9087
NB2	LR	15.536	0.9467
NB2	Wald	15.182	0.937

Table 4.3: The test statistics for null hypothesis that no hurdle is needed for the two Hurdle models with log link function for the zeros.

The AIC and BIC statistics for the Hurdle models are shown in Table 4.4. Here Binomial/Poisson means that the zero counts are fitted to the Binomial distribution and that the counts larger than zero are fitted to the Poisson distribution. The difference between the test statistics in the models with Binomial distribution for the zero counts and those without and the difference between using Poisson and Negative Binomial distribution for the counts larger than zero is small compared to the OLS value. But for both cases the models with Poisson or NB2 for the zero counts produces lower test statistics and is therefore preferred. It is also clear that the NB2 has a lower value than the Poisson.

In Table 4.5 the densities for predicted number of claims can be seen for the four Hurdle models. Here, the models with Binomial distribution for the zero counts predict the right number of zeros but the other two models do not. This is opposite to what the AIC and BIC test statistics tells us. Over all, the Poisson models seems to be closer to the observed claim densities than the NB2 models.

	Binomial/Poisson	Poisson/Poisson	Binomial/NB2	NB2/NB2
AIC	10928	10687	10655	10416
BIC	11213	10972	10945	10712

Table 4.4: The AIC and BIC for the different Hurdle models, where the zero distribution and count distribution are noted as zero distribution/count distribution.

Count	Observed	Binomial/Poisson	Poisson/Poisson	Binomial/NB2	NB2/NB2
0	385	385	380	385	380
1	214	247	219	249	221
2	169	167	165	167	165
3	131	127	130	126	130
4	100	100	105	100	105
5	98	82	87	82	86
6	66	69	73	69	72
7	65	59	62	59	62
8	51	52	54	51	54
9	46	46	47	45	47
10	38	41	42	41	42
100	1	2	2	2	2

Table 4.5: A comparison between the predicted number of zeros, ones etc. from the different Hurdle models.

4.3 Modified Models

The modified models are the same as before but with the variables with less than 99 percent significance omitted. The significance level is found in A1. The AIC and BIC statistics are shown in Table 4.6. These statistics show a large improvement, especially for the Hurdle models. It is though harder to compare these models predicted number of claims with the observed number of claims. This is the case because different variables are omitted in the different models. A better way would be to divide the policyholders into other groups than those in the data. This is not considered in this thesis, since more data is unavailable. Hurdle1 means Binomial/Poisson, Hurdle2 means Poisson/Poisson, Hurdle3 means Binomial/NB2 and Hurdle4 means NB2/NB2.

	Poisson	NB2	Hurdle1	Hurdle2	Hurdle3	Hurdle4
AIC	8869	9585	1758	2917	2224	3159
BIC	9000	9727	1928	3065	2412	3324

Table 4.6: The AIC and BIC for the modified models

4.4 Predicted Number of Claims

Table 4.7 shows a description of the ten test groups. The groups are not chosen after any formula and can not be used to choose the best model. Instead it is a good way to see how the different models work and to see things that are missed in more general statistics as the AIC etc. The different models are used to predict number of claims per test group and the result is shown in Table 4.8. The original models are used, because many variables are omitted in the modified models. Many of the predictions corresponds to the observed values. This is also true for the OLS except for the last test group for which one the OLS predict a quite large value. Another disadvantage with the OLS is that it can take negative values. The Poisson model and NB2 model agree to each other, but the NB2 is a bit closer to the observed value for some of the test groups. The Hurdle models do not perform better than the other models in this test, for these test groups they are even worse. Especially the Hurdle with NB2 for the count has a couple of outliers.

Group	Kilometers	Zone	Bonus	Make
1	1	2	5	1
2	3	2	7	3
3	1	1	2	1
4	2	1	7	7
5	1	6	3	3
6	2	5	2	1
7	2	7	3	8
8	1	6	6	2
9	2	5	5	3
10	3	1	1	9

Table 4.7: An explanation of the 10 test groups.

Group	Observed	OLS	Poisson	NB2	Hurdle-Poisson	Hurdle-NB2
1	27	27	24	25	24	19
2	60	58	58	59	58	98
3	45	37	32	34	32	28
4	74	87	76	75	75	86
5	0	-4	1	1	1	1
6	20	24	19	20	19	21
7	0	1	0	0	0	0
8	0	-0	3	3	3	2
9	0	4	1	1	2	1
10	456	343	499	481	499	617

Table 4.8: The observed and predicted number of claims for the 10 test groups and with the original models. The Hurdle models have the Binomial distribution for the zeros.

5 Discussion

The result shows that it is difficult, if not impossible, to find the perfect model. The model that is the best one in one test is worse in another. To be able to choose model it is important to know which properties one searches for in the model. In the insurance case, it is not only important to have a large likelihood for the whole dataset, but it should also make fair predictions for all groups. Otherwise some groups will get an unfair price of the insurance. Even if it is hard to choose between the models, the NB2 seems to do well in most of the tests. This agrees with the result that (Boucher et al., 2007) came up to, that NBx is the best one.

One drawback of this thesis is that the data is from 1977. The claim behavior could have changed since then in such a way that other models would suit better today. Perhaps people are more enlightened today and report claims more often. It is also possible that people has better cars and that there are fewer accidents. One thing that is sure is that the explanatory variables are out of interest. Firstly, because they are to few. They do not include important information about the driver for example. Secondly, because the car makes that exist today are very different from the ones that existed in 1977.

Another drawback is that the same dataset is used in both prediction and estimation of the unknown coefficients. This could lead us to choose a model with many explanatory variables that fit the available data well, but do not tell us anything about the outcome of interest. So when starting to predict outside the observed dataset the results would be wrong. To make a better comparison more data are needed. It is also possible to omit some parts of the data from the regressions and to use those parts in the valuation of the models, but then maybe important information is excluded and it is not clear which part that should be omitted.

A large problem with these kind of analysis is the lack of available data. But if possible, a comparison of models based on newer datasets could be useful. More explanatory variables to compare could be included and the comparisons could also be done in other areas than third-party automobile insurance. Other models could also be included in the comparison and instead of compare the number of predicted claims for only ten test groups, this could be done for all test groups.

With respect to the spread results and the drawbacks of this thesis the results can not be used to choose model when pricing insurance and the resulting explanatory variables can not be used to say anything about claim behavior. This study should be seen as a comparison of models that illuminate how hard it is to choose the right model. Hopefully this raises the interest of comparing models of the reader so that more comparisons of models will be done in the future and that all model choices will be treated critically. The conclusion is that the model choice is a part of pricing the third party automobile insurance that must take some time and extra thoughts. After all, the model choice decides what results you will get.

Bibliography

Boucher, J.-P., Denuit, M., and Guillén, M. (2007). Risk Classification for Claim Counts: A Comparative Analysis of Various Zero-Inflated Mixed Poisson and Hurdle Models. *North American Actuarial Journal*.

Cameron, A. C. and Trivedi, P. K. (1998). *Regression Analysis of Count Data*. Cambridge University Press.

Frees, E. W. (2010). *Regression Modeling with Actuarial and Financial Applications*. Cambridge University Press.

Harrell Jr., F. E. (2001). *Regression Modeling Strategies*. Springer.

Hilbe, J. M. (2011). *Negative Binomial Regression*. Cambridge University Press.

Lord, D., Washington, S. P., and Ivan, J. N. (2004). Poisson, Poisson-gamma and zero-inflated regression models of motor vehicle crashes: balancing statistical fit and theory. *Accident Analysis and Prevention*.

Oya, K. and Oya, K. (1997). Wald,LM and LR test statistics of linear hypothese in a strutural equation model, Econometric Reviews. *Faculty of Economics*.

Zeileis, A., Kleiber, C., and Jackman, S. (2008). Regression Models for Count Data in R. *Journal of Statistical Software*.

6 Appendices

6.1 A1

Here two tables with the regression coefficients and standard errors in parentheses are shown. The first table includes the original models and the second table includes the Hurdle models. Then there are three tables with the modified models, the last two of them are for Hurdle models. The number of stars states the significant level of the estimated parameter, $^{***}p < 0.001$, $^{**}p < 0.01$ and $^{*}p < 0.05$. The parameter *theta* in the Hurdle models is omitted since it got the value Na in the model with NB2 distribution for the zero counts.

	OLS	Poisson	NB2
(Intercept)	14.01^{***}	-1.81^{***}	-1.78^{***}
	(2.96)	(0.01)	(0.02)
Insured	0.12^{***}		
	(0.00)		
Kilometres2	6.75^{***}	0.21^{***}	0.19^{***}
	(1.82)	(0.01)	(0.02)
Kilometres3	2.90	0.32^{***}	0.28^{***}
	(1.83)	(0.01)	(0.02)
Kilometres4	-0.32	0.40^{***}	0.35^{***}
	(1.88)	(0.01)	(0.02)
Kilometres5	0.50	0.58^{***}	0.52^{***}
	(1.90)	(0.01)	(0.02)
Zone2	-4.56^{*}	-0.24^{***}	-0.22^{***}
	(2.14)	(0.01)	(0.02)
Zone3	-8.13^{***}	-0.39^{***}	-0.38^{***}
	(2.14)	(0.01)	(0.02)
Zone4	-9.87^{***}	-0.58^{***}	-0.56^{***}
	(2.22)	(0.01)	(0.02)
Zone5	-10.12^{***}	-0.33^{***}	-0.34^{***}
	(2.15)	(0.01)	(0.02)

Zone6	-10.40^{***}	-0.53^{***}	-0.52^{***}
	(2.14)	(0.01)	(0.02)
Zone7	-11.17^{***}	-0.73^{***}	-0.73^{***}
	(2.21)	(0.04)	(0.05)
Bonus2	-1.41	-0.48^{***}	-0.44^{***}
	(2.18)	(0.01)	(0.02)
Bonus3	-2.89	-0.69^{***}	-0.68^{***}
	(2.19)	(0.01)	(0.02)
Bonus4	-3.37	-0.83^{***}	-0.82^{***}
	(2.19)	(0.01)	(0.02)
Bonus5	-3.01	-0.93^{***}	-0.92^{***}
	(2.19)	(0.01)	(0.02)
Bonus6	-2.60	-0.99^{***}	-0.99^{***}
	(2.20)	(0.01)	(0.02)
Bonus7	-2.46	-1.33^{***}	-1.33^{***}
	(2.66)	(0.01)	(0.02)
Make2	-4.50	0.08^{***}	0.07^{**}
	(2.56)	(0.02)	(0.03)
Make3	-5.05^{*}	-0.25^{***}	-0.24^{***}
	(2.55)	(0.03)	(0.03)
Make4	-7.36^{**}	-0.65^{***}	-0.68^{***}
	(2.58)	(0.02)	(0.03)
Make5	-3.29	0.15^{***}	0.15^{***}
	(2.56)	(0.02)	(0.02)
Make6	-7.22^{**}	-0.34^{***}	-0.36^{***}
	(2.56)	(0.02)	(0.02)
Make7	-5.73^{*}	-0.06^{*}	-0.08^{**}
	(2.55)	(0.02)	(0.03)
Make8	-5.62^{*}	-0.04	-0.04
	(2.58)	(0.03)	(0.04)
Make9	17.81^{***}	-0.07^{***}	-0.09^{***}
	(3.11)	(0.01)	(0.02)
Insured:Kilometres2	0.01^{***}		
	(0.00)		
Insured:Kilometres3	0.01^{***}		
	(0.00)		
Insured:Kilometres4	0.02^{***}		
	(0.00)		

Insured:Kilometres5	0.02^{***}
	(0.00)
Insured:Zone2	-0.01^{***}
	(0.00)
Insured:Zone3	-0.01^{***}
	(0.00)
Insured:Zone4	-0.02^{***}
	(0.00)
Insured:Zone5	-0.01^{***}
	(0.00)
Insured:Zone6	-0.02^{***}
	(0.00)
Insured:Zone7	-0.03^{***}
	(0.01)
Insured:Bonus2	-0.04^{***}
	(0.00)
Insured:Bonus3	-0.05^{***}
	(0.00)
Insured:Bonus4	-0.06^{***}
	(0.00)
Insured:Bonus5	-0.07^{***}
	(0.00)
Insured:Bonus6	-0.07^{***}
	(0.00)
Insured:Bonus7	-0.08^{***}
	(0.00)
Insured:Make2	0.00
	(0.00)
Insured:Make3	-0.01^{**}
	(0.00)
Insured:Make4	-0.02^{***}
	(0.00)
Insured:Make5	0.00
	(0.00)
Insured:Make6	-0.01^{***}
	(0.00)
Insured:Make7	0.00
	(0.00)

Insured:Make8	0.00		
	(0.01)		
Insured:Make9	0.00		
	(0.00)		
R^2	0.98		
Adj. R^2	0.98		
Num. obs.	2182	2182	2182
AIC		10654.00	10379.60
BIC		10796.20	10527.49
Log Likelihood		-5302.00	-5163.80
Deviance		2966.12	2229.91

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

Table 6.1: Statistical models

	Binomial/Poisson	Poisson/Poisson	Binomial/NB2	NB2/NB2
Count model: (Intercept)	-1.81^{***}	-1.81^{***}	-1.78^{***}	-1.78^{***}
	(0.01)	(0.01)	(0.02)	(0.02)
Count model: Kilometres2	0.21^{***}	0.21^{***}	0.19^{***}	0.19^{***}
	(0.01)	(0.01)	(0.02)	(0.02)
Count model: Kilometres3	0.32^{***}	0.32^{***}	0.28^{***}	0.28^{***}
	(0.01)	(0.01)	(0.02)	(0.02)
Count model: Kilometres4	0.41^{***}	0.41^{***}	0.35^{***}	0.35^{***}
	(0.01)	(0.01)	(0.02)	(0.02)
Count model: Kilometres5	0.58^{***}	0.58^{***}	0.52^{***}	0.52^{***}
	(0.01)	(0.01)	(0.02)	(0.02)
Count model: Zone2	-0.24^{***}	-0.24^{***}	-0.22^{***}	-0.22^{***}
	(0.01)	(0.01)	(0.02)	(0.02)
Count model: Zone3	-0.39^{***}	-0.39^{***}	-0.38^{***}	-0.38^{***}
	(0.01)	(0.01)	(0.02)	(0.02)

Count model: Zone4	-0.58^{***}	-0.58^{***}	-0.56^{***}	-0.56^{***}
	(0.01)	(0.01)	(0.02)	(0.02)
Count model: Zone5	-0.32^{***}	-0.32^{***}	-0.34^{***}	-0.34^{***}
	(0.01)	(0.01)	(0.02)	(0.02)
Count model: Zone6	-0.52^{***}	-0.52^{***}	-0.52^{***}	-0.52^{***}
	(0.01)	(0.01)	(0.02)	(0.02)
Count model: Zone7	-0.74^{***}	-0.74^{***}	-0.74^{***}	-0.74^{***}
	(0.04)	(0.04)	(0.05)	(0.05)
Count model: Bonus2	-0.48^{***}	-0.48^{***}	-0.44^{***}	-0.44^{***}
	(0.01)	(0.01)	(0.02)	(0.02)
Count model: Bonus3	-0.69^{***}	-0.69^{***}	-0.68^{***}	-0.68^{***}
	(0.01)	(0.01)	(0.02)	(0.02)
Count model: Bonus4	-0.83^{***}	-0.83^{***}	-0.82^{***}	-0.82^{***}
	(0.01)	(0.01)	(0.02)	(0.02)
Count model: Bonus5	-0.93^{***}	-0.93^{***}	-0.92^{***}	-0.92^{***}
	(0.01)	(0.01)	(0.02)	(0.02)
Count model: Bonus6	-0.99^{***}	-0.99^{***}	-1.00^{***}	-1.00^{***}
	(0.01)	(0.01)	(0.02)	(0.02)
Count model: Bonus7	-1.33^{***}	-1.33^{***}	-1.33^{***}	-1.33^{***}
	(0.01)	(0.01)	(0.02)	(0.02)
Count model: Make2	0.08^{***}	0.08^{***}	0.07^{**}	0.07^{**}
	(0.02)	(0.02)	(0.03)	(0.03)
Count model: Make3	-0.24^{***}	-0.24^{***}	-0.23^{***}	-0.23^{***}
	(0.03)	(0.03)	(0.03)	(0.03)
Count model: Make4	-0.66^{***}	-0.66^{***}	-0.69^{***}	-0.69^{***}
	(0.02)	(0.02)	(0.03)	(0.03)
Count model: Make5	0.15^{***}	0.15^{***}	0.15^{***}	0.15^{***}
	(0.02)	(0.02)	(0.02)	(0.02)
Count model: Make6	-0.33^{***}	-0.33^{***}	-0.36^{***}	-0.36^{***}
	(0.02)	(0.02)	(0.02)	(0.02)
Count model: Make7	-0.06^{*}	-0.06^{*}	-0.08^{**}	-0.08^{**}
	(0.02)	(0.02)	(0.03)	(0.03)
Count model: Make8	-0.05	-0.05	-0.04	-0.04
	(0.03)	(0.03)	(0.04)	(0.04)
Count model: Make9	-0.07^{***}	-0.07^{***}	-0.09^{***}	-0.09^{***}
	(0.01)	(0.01)	(0.02)	(0.02)
Zero model: (Intercept)	6.07^{***}	-1.88^{***}	6.07^{***}	-1.88^{***}
	(0.53)	(0.34)	(0.53)	(0.34)

Zero model: Kilometres2	0.80^{**}	0.18	0.80^{**}	0.18
	(0.31)	(0.19)	(0.31)	(0.19)
Zero model: Kilometres3	-0.12	0.14	-0.12	0.14
	(0.28)	(0.17)	(0.28)	(0.17)
Zero model: Kilometres4	-1.60^{***}	0.25	-1.60^{***}	0.25
	(0.26)	(0.16)	(0.26)	(0.16)
Zero model: Kilometres5	-1.98^{***}	0.32^{*}	-1.98^{***}	0.32^{*}
	(0.26)	(0.16)	(0.26)	(0.16)
Zero model: Zone2	0.00	-0.12	0.00	-0.12
	(0.36)	(0.20)	(0.36)	(0.20)
Zero model: Zone3	-0.13	-0.40^{*}	-0.13	-0.40^{*}
	(0.35)	(0.19)	(0.35)	(0.19)
Zero model: Zone4	0.97^{*}	-0.36	0.97^{*}	-0.36
	(0.44)	(0.22)	(0.44)	(0.22)
Zero model: Zone5	-2.20^{***}	-0.37^{*}	-2.20^{***}	-0.37^{*}
	(0.31)	(0.18)	(0.31)	(0.18)
Zero model: Zone6	-1.38^{***}	-0.63^{***}	-1.38^{***}	-0.63^{***}
	(0.32)	(0.18)	(0.32)	(0.18)
Zero model: Zone7	-5.28^{***}	-0.74^{**}	-5.28^{***}	-0.74^{**}
	(0.36)	(0.23)	(0.36)	(0.23)
Zero model: Bonus2	-0.25	-0.35^{*}	-0.25	-0.35^{*}
	(0.28)	(0.17)	(0.28)	(0.17)
Zero model: Bonus3	-0.25	-0.53^{**}	-0.25	-0.53^{**}
	(0.28)	(0.17)	(0.28)	(0.17)
Zero model: Bonus4	-0.47	-0.79^{***}	-0.47	-0.79^{***}
	(0.28)	(0.17)	(0.28)	(0.17)
Zero model: Bonus5	-0.21	-0.90^{***}	-0.21	-0.90^{***}
	(0.28)	(0.17)	(0.28)	(0.17)
Zero model: Bonus6	0.82^{**}	-0.76^{***}	0.82^{**}	-0.76^{***}
	(0.31)	(0.20)	(0.31)	(0.20)
Zero model: Bonus7	2.66^{***}	-0.99^{***}	2.66^{***}	-0.99^{***}
	(0.39)	(0.26)	(0.39)	(0.26)
Zero model: Make2	-1.72^{***}	0.04	-1.72^{***}	0.04
	(0.42)	(0.27)	(0.42)	(0.27)
Zero model: Make3	-3.02^{***}	-0.29	-3.02^{***}	-0.29
	(0.41)	(0.27)	(0.41)	(0.27)
Zero model: Make4	-3.85^{***}	-0.50	-3.85^{***}	-0.50
	(0.41)	(0.28)	(0.41)	(0.28)

Zero model: Make5	-1.66^{***}	0.28	-1.66^{***}	0.28
	(0.42)	(0.27)	(0.42)	(0.27)
Zero model: Make6	-1.27^{**}	-0.36	-1.27^{**}	-0.36
	(0.42)	(0.27)	(0.42)	(0.27)
Zero model: Make7	-2.16^{***}	0.16	-2.16^{***}	0.16
	(0.41)	(0.27)	(0.41)	(0.27)
Zero model: Make8	-3.22^{***}	0.06	-3.22^{***}	0.06
	(0.41)	(0.27)	(0.41)	(0.27)
Zero model: Make9	3.85^{***}	0.47	3.85^{***}	0.47
	(1.08)	(0.50)	(1.08)	(0.50)
AIC	10928.18	10687.34	10654.92	10416.07
Log Likelihood	-5414.09	-5293.67	-5276.46	-5156.03
Num. obs.	2182	2182	2182	2182

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

Table 6.2: Hurdle models

	OLS	Poisson	NB2
(Intercept)	56.22^{**}	-1.81^{***}	-1.79^{***}
	(20.16)	(0.01)	(0.02)
Insured	0.04^{***}		
	(0.00)		
factor(Kilometres, exclude = c(”3”, ”4”, ”5”))2	6.02		
	(11.81)		
factor(Zone, exclude = c(”2”))3	-16.98		
	(20.86)		
factor(Zone, exclude = c(”2”))4	10.53		
	(21.31)		
factor(Zone, exclude = c(”2”))5	-52.96^{*}		
	(20.95)		
factor(Zone, exclude = c(”2”))6	-43.18^{*}		
	(20.85)		

factor(Zone, exclude = c(”2”))7	−70.05***		
	(20.92)		
factor(Make, exclude = c(”2”, ”3”, ”7”, ”8”))4	−17.36		
	(20.33)		
factor(Make, exclude = c(”2”, ”3”, ”7”, ”8”))5	−17.27		
	(19.62)		
factor(Make, exclude = c(”2”, ”3”, ”7”, ”8”))6	−13.31		
	(19.82)		
factor(Make, exclude = c(”2”, ”3”, ”7”, ”8”))9	184.47***		
	(20.92)		
Insured:factor(Kilometres, exclude = c(”3”, ”4”, ”5”))2	0.00***		
	(0.00)		
Insured:factor(Zone, exclude = c(”2”))3	−0.01***		
	(0.00)		
Insured:factor(Zone, exclude = c(”2”))4	−0.02***		
	(0.00)		
Insured:factor(Zone, exclude = c(”2”))5	−0.02***		
	(0.01)		
Insured:factor(Zone, exclude = c(”2”))6	−0.02***		
	(0.00)		
Insured:factor(Zone, exclude = c(”2”))7	−0.10***		
	(0.03)		
Insured:factor(Make, exclude = c(”2”, ”3”, ”7”, ”8”))4	−0.02		
	(0.01)		
Insured:factor(Make, exclude = c(”2”, ”3”, ”7”, ”8”))5	−0.01		
	(0.02)		
Insured:factor(Make, exclude = c(”2”, ”3”, ”7”, ”8”))6	−0.01		
	(0.01)		
Insured:factor(Make, exclude = c(”2”, ”3”, ”7”, ”8”))9	0.00		
	(0.00)		
Kilometres2		0.21***	0.19***
		(0.01)	(0.02)
Kilometres3		0.32***	0.28***
		(0.01)	(0.02)
Kilometres4		0.41***	0.36***
		(0.01)	(0.02)
Kilometres5		0.58***	0.52***
		(0.01)	(0.02)

Zone2	−0.24***	−0.22***
	(0.01)	(0.02)
Zone3	−0.39***	−0.38***
	(0.01)	(0.02)
Zone4	−0.58***	−0.55***
	(0.01)	(0.02)
Zone5	−0.33***	−0.34***
	(0.01)	(0.02)
Zone6	−0.53***	−0.52***
	(0.01)	(0.02)
Zone7	−0.73***	−0.72***
	(0.04)	(0.05)
Bonus2	−0.48***	−0.44***
	(0.01)	(0.02)
Bonus3	−0.70***	−0.68***
	(0.01)	(0.02)
Bonus4	−0.83***	−0.82***
	(0.01)	(0.02)
Bonus5	−0.93***	−0.91***
	(0.01)	(0.02)
Bonus6	−0.99***	−1.00***
	(0.01)	(0.02)
Bonus7	−1.33***	−1.32***
	(0.01)	(0.02)
factor(Make, exclude = c(”7”, ”8”))2	0.08***	
	(0.02)	
factor(Make, exclude = c(”7”, ”8”))3	−0.25***	
	(0.03)	
factor(Make, exclude = c(”7”, ”8”))4	−0.65***	
	(0.02)	
factor(Make, exclude = c(”7”, ”8”))5	0.16***	
	(0.02)	
factor(Make, exclude = c(”7”, ”8”))6	−0.34***	
	(0.02)	
factor(Make, exclude = c(”7”, ”8”))9	−0.07***	
	(0.01)	
factor(Make, exclude = c(”8”))2		0.07**
		(0.03)

factor(Make, exclude = c("8"))3			-0.24^{***}
			(0.03)
factor(Make, exclude = c("8"))4			-0.68^{***}
			(0.03)
factor(Make, exclude = c("8"))5			0.15^{***}
			(0.02)
factor(Make, exclude = c("8"))6			-0.36^{***}
			(0.02)
factor(Make, exclude = c("8"))7			-0.08^{**}
			(0.03)
factor(Make, exclude = c("8"))9			-0.09^{***}
			(0.02)
R^2	0.90		
Adj. R^2	0.89		
Num. obs.	420	1703	1945
AIC		8869.47	9585.23
BIC		8994.59	9724.56
Log Likelihood		-4411.73	-4767.62
Deviance		2441.97	1995.99

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

Table 6.3: Modified Statistical models

	Binomial/Poisson	Poisson/Poisson
Count model: (Intercept)	-1.76^{***}	-1.79^{***}
	(0.03)	(0.02)
Count model: Kilometres2	0.17^{***}	0.18^{***}
	(0.02)	(0.01)
Count model: Kilometres4	0.34^{***}	0.35^{***}
	(0.03)	(0.02)
Count model: Kilometres5	0.53^{***}	0.58^{***}

	(0.03)	(0.03)
Count model: Zone5	-0.30^{***}	
	(0.02)	
Count model: Zone6	-0.52^{***}	-0.52^{***}
	(0.02)	(0.01)
Count model: Zone7	-0.75^{***}	-0.74^{***}
	(0.06)	(0.05)
Count model: Bonus6	-1.08^{***}	-1.03^{***}
	(0.02)	(0.02)
Count model: Bonus7	-1.37^{***}	-1.38^{***}
	(0.02)	(0.02)
Count model: factor(Make, exclude = c(”7”, ”8”))2	0.10^{*}	0.10^{*}
	(0.05)	(0.04)
Count model: factor(Make, exclude = c(”7”, ”8”))3	-0.21^{***}	-0.26^{***}
	(0.06)	(0.05)
Count model: factor(Make, exclude = c(”7”, ”8”))4	-0.65^{***}	-0.56^{***}
	(0.05)	(0.05)
Count model: factor(Make, exclude = c(”7”, ”8”))5	0.17^{***}	0.16^{***}
	(0.05)	(0.04)
Count model: factor(Make, exclude = c(”7”, ”8”))6	-0.32^{***}	-0.33^{***}
	(0.04)	(0.04)
Count model: factor(Make, exclude = c(”7”, ”8”))9	-0.06^{*}	-0.04
	(0.02)	(0.02)
Zero model: (Intercept)	6.02^{***}	-1.81^{***}
	(1.14)	(0.29)
Zero model: factor(Kilometres, exclude = c(”3”))2	1.16	
	(0.72)	
Zero model: factor(Kilometres, exclude = c(”3”))4	-1.86^{**}	
	(0.63)	
Zero model: factor(Kilometres, exclude = c(”3”))5	-2.50^{***}	
	(0.64)	
Zero model: factor(Zone, exclude = c(”2”, ”3”, ”4”))5	-1.72^{*}	
	(0.72)	
Zero model: factor(Zone, exclude = c(”2”, ”3”, ”4”))6	-1.02	
	(0.74)	
Zero model: factor(Zone, exclude = c(”2”, ”3”, ”4”))7	-5.47^{***}	
	(0.84)	
Zero model: factor(Bonus, exclude = c(”2”, ”3”, ”4”, ”5”))6	0.84	

	(0.48)	
Zero model: factor(Bonus, exclude = c(”2”, ”3”, ”4”, ”5”))7	2.75^{***}	
	(0.62)	
Zero model: Make2	-1.40	
	(0.87)	
Zero model: Make3	-2.36^{**}	
	(0.86)	
Zero model: Make4	-3.98^{***}	
	(0.89)	
Zero model: Make5	-1.93^{*}	
	(0.86)	
Zero model: Make6	-1.92^{*}	
	(0.86)	
Zero model: Make9	18.00	
	(1918.46)	
Count model: Kilometres3		0.30^{***}
		(0.02)
Count model: Bonus3		-0.71^{***}
		(0.02)
Count model: Bonus4		-0.87^{***}
		(0.03)
Count model: Bonus5		-0.97^{***}
		(0.03)
Zero model: factor(Zone, exclude = c(”2”, ”3”, ”4”, ”5”))6		-0.76^{**}
		(0.25)
Zero model: factor(Zone, exclude = c(”2”, ”3”, ”4”, ”5”))7		-0.71^{**}
		(0.26)
Zero model: factor(Bonus, exclude = c(”2”))3		-0.44
		(0.31)
Zero model: factor(Bonus, exclude = c(”2”))4		-0.92^{**}
		(0.29)
Zero model: factor(Bonus, exclude = c(”2”))5		-0.73^{*}
		(0.31)
Zero model: factor(Bonus, exclude = c(”2”))6		-0.88^{**}
		(0.31)
Zero model: factor(Bonus, exclude = c(”2”))7		-0.86^{*}
		(0.36)
AIC	1757.72	2916.91

Log Likelihood	-848.86	-1432.46
Num. obs.	331	621

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

Table 6.4: Modified Hurdle models

	Binomial/NB2	NB2/NB2
Count model: (Intercept)	-7.94^{***}	-1.75^{***}
	(0.11)	(0.04)
Count model: Kilometres2	0.03	0.15^{***}
	(0.08)	(0.03)
Count model: Kilometres4	1.68^{***}	0.30^{***}
	(0.10)	(0.04)
Count model: Kilometres5	2.16^{***}	0.51^{***}
	(0.10)	(0.04)
Count model: Zone5	0.56^{***}	
	(0.08)	
Count model: Zone6	-0.35^{***}	-0.51^{***}
	(0.08)	(0.02)
Count model: Zone7	1.81^{***}	-0.74^{***}
	(0.14)	(0.06)
Count model: Bonus6	-1.84^{***}	-1.04^{***}
	(0.09)	(0.04)
Count model: Bonus7	-3.68^{***}	-1.38^{***}
	(0.08)	(0.03)
Count model: factor(Make, exclude = c(”8”))2	1.52^{***}	0.11^{*}
	(0.13)	(0.05)
Count model: factor(Make, exclude = c(”8”))3	1.13^{***}	-0.19^{***}
	(0.13)	(0.06)
Count model: factor(Make, exclude = c(”8”))4	0.19	-0.58^{***}
	(0.16)	(0.06)
Count model: factor(Make, exclude = c(”8”))5	1.52^{***}	0.15^{**}

	(0.12)	(0.05)
Count model: factor(Make, exclude = c(”8”))6	0.26^{*}	-0.35^{***}
	(0.12)	(0.05)
Count model: factor(Make, exclude = c(”8”))7	1.43^{***}	-0.07
	(0.13)	(0.06)
Count model: factor(Make, exclude = c(”8”))9	-2.03^{***}	-0.06
	(0.10)	(0.03)
Zero model: (Intercept)	5.92^{***}	-1.69^{***}
	(1.07)	(0.32)
Zero model: factor(Kilometres, exclude = c(”3”))2	0.99	
	(0.66)	
Zero model: factor(Kilometres, exclude = c(”3”))4	-2.01^{***}	
	(0.59)	
Zero model: factor(Kilometres, exclude = c(”3”))5	-2.44^{***}	
	(0.60)	
Zero model: factor(Zone, exclude = c(”2”, ”3”, ”4”))5	-1.59^{*}	
	(0.64)	
Zero model: factor(Zone, exclude = c(”2”, ”3”, ”4”))6	-0.85	
	(0.67)	
Zero model: factor(Zone, exclude = c(”2”, ”3”, ”4”))7	-5.42^{***}	
	(0.76)	
Zero model: factor(Bonus, exclude = c(”2”, ”3”, ”4”, ”5”))6	0.98^{*}	
	(0.44)	
Zero model: factor(Bonus, exclude = c(”2”, ”3”, ”4”, ”5”))7	2.89^{***}	
	(0.58)	
Zero model: Make2	-1.41	
	(0.87)	
Zero model: Make3	-2.39^{**}	
	(0.86)	
Zero model: Make4	-4.01^{***}	
	(0.88)	
Zero model: Make5	-1.94^{*}	
	(0.86)	
Zero model: Make6	-1.94^{*}	
	(0.86)	
Zero model: Make7	-2.42^{**}	
	(0.85)	
Zero model: Make9	18.02	

	(1909.22)	
Count model: Kilometres3		0.27^{***}
		(0.03)
Count model: Bonus3		-0.71^{***}
		(0.04)
Count model: Bonus4		-0.86^{***}
		(0.05)
Count model: Bonus5		-0.96^{***}
		(0.04)
Zero model: factor(Zone, exclude = c(”2”, ”3”, ”4”, ”5”))6		-0.78^{***}
		(0.24)
Zero model: factor(Zone, exclude = c(”2”, ”3”, ”4”, ”5”))7		-0.80^{**}
		(0.28)
Zero model: factor(Bonus, exclude = c(”2”))3		-0.46
		(0.29)
Zero model: factor(Bonus, exclude = c(”2”))4		-0.86^{**}
		(0.28)
Zero model: factor(Bonus, exclude = c(”2”))5		-0.68^{*}
		(0.29)
Zero model: factor(Bonus, exclude = c(”2”))6		-0.89^{**}
		(0.30)
Zero model: factor(Bonus, exclude = c(”2”))7		-0.88^{*}
		(0.35)
AIC	2224.20	3158.79
Log Likelihood	-1079.10	-1550.39
Num. obs.	378	708

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

Table 6.5: Modified Hurdle models