Stockholms universitet

Prediction of heavier vehicle type for traffic accidents using multinomial logistic regression

David Block

Kandidatuppsats i matematisk statistik
Bachelor Thesis in Mathematical Statistics

Kandidatuppsats 2018:8
Matematisk statistik
Juni 2018

www.math.su.se

Matematisk statistik
Matematiska institutionen
Stockholms universitet
106 91 Stockholm

Stockholm University

Mathematical Statistics
Stockholm University
Bachelor Thesis **2018:8**
http://www.math.su.se

Prediction of heavier vehicle type for traffic accidents using multinomial logistic regression

David Block*

June 2018

Abstract

In this bachelor thesis we use data on traffic accidents in Sweden during the period 2003-2016, gathered from police reports. We specifically look at accidents involving heavier vehicles; the five types chosen are heavy motorcycle, car, light truck, heavy truck and bus. The goal is to predict the vehicle type of traffic accidents, using several categorical predictors. We fit a multinomial logistic regression model for this purpose, with outcome variable Vehicle type. The predictors used are mainly those with external effect on accident risk for different vehicle types, such as weather, road surface, traffic situation and road type. Predictors such as year and weekday are also used to account for driving patterns in different time intervals. After fitting the model, we use it to predict vehicle type of accidents on test data, and compare the model prediction to the observed vehicle types of accidents. The model performs to some degree; in future research it could be used to further investigate differences in accident causes between primarily cars and heavy motorcycles.

*Postal address: Mathematical Statistics, Stockholm University, SE-106 91, Sweden. E-mail: dabl2733@student.su.se. Supervisor: Kristoffer Lindensjö, Felix Wahl.

Sammanfattning

I denna kandidatuppsats används data från trafikolyckor i Sverige under perioden 2003-2016, hämtade från polisreporter. Vi tittar specifikt på olyckor som involverar tyngre fordon; de fem valda fordonstyperna är tung motorcykel, personbil, lätt lastbil, tung lastbil och buss. Målet är att prediktera fordonstyp för trafikolyckor, när vi använder flera kategoriska förklarande variabler. Vi anpassar en multinomial logistisk regressionsmodell för det här syftet, med utfallsvariabel Fordonstyp. De förklarande variablerna som används är huvudsakligen de med extern effekt på olycksrisk för olika fordonstyper, t.ex. väder, väglag, trafiksituation och vägtyp. Förklarande variabler som år och veckodag används också för att räkna med körmönster i olika tidsintervall. Efter att vi har anpassat modellen, använder vi den för att prediktera fordonstyp i olyckor på test data, och jämför modellens prediktion med de observerade fordonstyperna i olyckorna. Modellen presterar till viss grad; i framtida forskning skulle den kunna användas för att ytterligare undersöka skillnader i olycksorsaker mellan huvudsakligen personbilar och tyngre motorcyklar.

Acknowledgements

Firstly, i would like to give thanks to my family and my friends for their help and support. I would also like to thank Transportstyrelsen for their great service and providing me with data in a suitable form. Finally, i want to thank my supervisors Felix Wahl and Kristoffer Lindensjö for their advice and guidance.

Contents

1 Introduction

The swedish Riksdag decided 1997 that Nollvisionen, a vision of a minimum amount of fatalities and injuries in traffic accidents, would be the main guideline for traffic safety development i Sweden. During the period 2003 - 2016 which this study encompasses, the number of fatalities on swedish roads has steadily decreased. In 2017, 253 people were killed and 2347 severely injured. This is the lowest count since data on this subject began to be recorded, and the lowest per million people in the EU. Measured from 2000 to 2017, the number of fatalities have dropped by 57%. The change is big enough that it cannot be explained by natural variation of deaths over time alone.

In september 2016 the swedish government decided to further intensify the work on traffic safety (Regerinskansliet, 2016), with Transportstyrelsen and Trafikverket as the two main governing bodies. The decision was taken to make further strides against reaching the specific goal set in 2009, to reduce traffic fatalities by 50% by 2020 (Lindberg et al., 2016), to a level of 220 fatalities per year. At the time of writing, this goal is set to fail (Trafikverket, 2017). Annual studies have shown possible discrepancies in certain aspects of traffic accident data, which perhaps can be highlighted even more. One of these aspects are differences between vehicle types; for example, fatalities involving the vehicle type Car have decreased by 67% during 2000-2017, while fatality count involving vehicle type Heavy motorcycle remain unchanged. Motorcycle drivers are much more exposed than drivers of a car, and cars are also increasingly built with focus on safety. The relative decrease in total accident count (fatal or non fatal) is also bigger for cars however, which could mean that other factors unrelated to the seriousness of the accident also have an effect (Trafikverket, 2017). Improvement of traffic safety is complex, since the goal is to modify interactions of technical, environmental and behavioural factors. *Environmental* factors as Haddon (1970) would call it, are the primary focus in this study. We use data from police reports during period 2003-2016, which focuses on external factors of a traffic accident, such as weather, road surface and road type. We then try to find a model for prediction of different vehicle types in traffic accidents. In doing so, we can possibly highlight connections between certain vehicle types and external factors of accident, which would give direction to further studies.

1.1 Study aim

The aim of this thesis is to find multinomial logit models for the prediction of heavy vehicle types in traffic accidents. By the category heavy vehicle types, we mean in this study five specific types of motor vehicles; Car, Heavy motorcycle, Heavy truck, Light truck and Bus. These are the values that create our nominal outcome variable with multinomial distribution. Heavy vehicle types are chosen because they represent the biggest proportion of vehicles in traffic, and also causes most deaths (Trafikverket, 2017). When interpreting results of the model, the most common vehicle type Car will be used as reference against which the other vehicle types are compared against. The best model will hopefully be able to predict correct vehicle type in traffic accident at high accuracy compared to observed data, given a set of predictor variables. To reiterate: the predictor variables used is primarily those with external relation to the accident, such as weather, road surface and

road type. Other variables represent regional differences as well as differences over time. The majority of these predictor variables used are recorded in close connection to accident site and time of accident. Ideally these predictors help point to differences between vehicle types at similar traffic circumstances, and otherwise the results can be a reference for further study on the subject. The study does not encompass severity of accident, which means that a less serious collision is identical to one with fatal outcome. Data on fatality is available in our dataset, but will not be included.

1.2 Method of using data

The entire dataset will initially be divided into a training set, which contains 70% of data selected at random, and a test set, using the remaining 30% of data. The model selection in Chapter 4, using statistics presented in section 2.5, will be made using only the training set. In Chapter 5, the training set will then be divided into k subsets, as presented in section 2.5.7. A k - fold Cross validation is used to test prediction performance on each of the k subsets of the training set. In this way, we can see if the model is equally suitable over the entire training data. In the same chapter, we finally use the selected model for prediction on the test data.

1.3 Disposition

Chapter 2 of this thesis focuses on the theory behind the study, primarily regarding the logistic and the multinomial logistic regression models that will be used. It is divided into five parts; the first part will focus on generalized linear models and what types of results we can expect from our models. The second and third part will focus on building the binary and multinomial logistic regression models and explain their properties. The fourth part will cover assumptions that is made to validate the multinomial logistic regression model. The fifth part goes through theory on how we measure model fit, compare models and use them for prediction.

In Chapter 3 discusses and looks closer at the data used in the study. We examine our outcome variable, our predictor variables and look shortly at how software handles nominal predictors with multiple values. We also address the possible problem of missing data, as well as possible correlation between our predictor variables. Chapter 4 includes our model selection process, and Chapter 5 is where we use the best model and test it on new data. We interpret the results of our model in Chapter 6, and Chapter 7 is for a discussion and conclusion. Appendix and references are found in two separate chapters at the end.

2 Theory

In this chapter, the theory of use in this study is presented shortly. The multinomial logistic regression that we plan to use is a type of generalized linear model (GLM). We will generalize the use of such models from the binary to the multinomial case, focusing on modeling data.

2.1 Generalized linear model

GLM:s have three components (Agresti, 2002, p. 116); First a random component, which is the distribution of the outcome variable of the model. The outcome variable can be of different forms, such as numerical, binary or categorical (nominal or ordinal), with suited distribution. In this study, the outcome variable is categorical with five values, and has a multinomial distribution.

Secondly, a systematic component. For a vector of predictor variables $x = x_1, ..., x_n$, and a vector of coefficients $\beta^T = \beta_1, ..., \beta_n$ we can write the systematic component as $\beta_0 + x_1\beta_1 + \ldots + x_n\beta_n$. Here β_0 (or α), is referred to as the intercept, i.e the value of the model when $x = 0$. β, also a constant, describes the relationship of the predictor variable to the model. The coefficients of this component have to be estimated, and for logistic regression models this is typically done with maximum likelihood estimation.

Thirdly, a link function between the systematic component and the the expected value of the random component. The generalized model can therefore be written as $h(\mu) = \beta_0 + x_1\beta_1 + \ldots + x_n\beta_n$ with link function h and expected value of the random component as μ. In our model for logistic regression, h is the logit transform, that is the *log odds* of some event occurring.

2.2 Logistic regression

We begin with describing a predictive model where the stochastic outcome variable Z can take two values, often referred to success or failure, or simply 1 and 0. The probability of success given a set of values for the predictor variables $x = x_1, ..., x_n$ can be expressed as $\pi(x) = P(Y = 1|X = x) = 1$ - $P(Y = 0|X = x)$. This type of regression model is called binary and variable Z has a binomial distribution.

2.2.1 Properties

Given a set of values for the predictor variables $x = x_1, ..., x_n$, we can together with the link function log odds, write the logistic regression model as

$$\log\frac{\pi(x)}{1-\pi(x)} \quad = \quad \beta_0 + x_1\beta_1 + ... + x_n\beta_n \quad = \quad \beta_0 + \beta^T x \qquad (1)$$

as formulated in Agresti (2002, p. 182). If we wish to look at the odds for a specific event, we must first pick a reference value of the outcome variable Z. For the values 0 and 1, the reference is often picked as 0. This means that we judge the probability of 1 relative to this reference value. As presented by Agresti (2002, p.44), the odds is obtained by taking the exponential of (1). It can be written as

$$\frac{P(Z=1|x)}{P(Z=0|x)} \quad = \quad \frac{\pi(x)}{1-\pi(x)} \quad = \quad e^{\beta_0+\beta^T x} \qquad (2)$$

where Z is our outcome variable and x is a set of predictor values. The odds are always non-negative, and in this case a value bigger than 1.0 signifies that Z is more probable to take value 1 than 0, given a set of predictor values x. We can also compare odds using (2). Suppose that we have two different specific sets of predictor values, x and $\tilde{x}$. For the set $\tilde{x}$

the previous odds in (2) will change. The rate of change is called the odds ratio, and is defined as

$$\frac{P(Z=1|\tilde{x})/P(Z=0|\tilde{x})}{P(Z=1|x)/P(Z=0|x)} = e^{\beta^T(\tilde{x}-x)}.$$

To obtain the new odds we can then multiply the old odds with the change as

$$\frac{\pi(x)}{1-\pi(x)}e^{\beta^T(\tilde{x}-x)}.$$

This result is very useful for tracking and quantifying results for different sets of predictor values. The model can also be expressed as success probability. If we use (2) and solve for $\pi(x)$ we obtain

$$\pi(x) = \frac{e^{\beta_0+\beta^T x}}{1+e^{\beta_0+\beta^T x}}.$$

For our purpose though, we will use odds as indicator of relative change. We will now look at the case where the response variable is of more than two levels.

2.3 Multinomial logistic regression

So far, the outcome variable Z has had the values of either 1 (success) or 0 (not success). For a multinomial logistic regression model, we instead have J number of discrete values which the outcome variable can take, where $J \geq 2$. This means that the default reference value for a binary outcome variable, 0, is to be set to a chosen value. Often it is picked as the most frequent outcome value in data.

2.3.1 Properties

We again have a set of predictor variables $x = x_1, ..., x_n$, together with $j = 1, ..., J$ possible values of the outcome variable Z. If the J:th and last value of the outcome variable is picked as the reference value, the multinomial logit model can be written as

$$\log\frac{\pi_j(x)}{\pi_J(x)} = \beta_{0j} + \beta_j^T x, \text{ for } j = 1, ..., J-1, \qquad (3)$$

as presented in Agresti (2002, p. 268). As in the previous section, the odds for a specific event is obtained by taking the exponential of (3) and can be written as

$$\frac{\pi_j(x)}{\pi_J(x)} = e^{\beta_{0j}+\beta_j^T x} \text{ , for } j = 1, ..., J-1. \qquad (4)$$

The odds ratio of two different sets of predictor values x and $\tilde{x}$ is then defined as

$$\frac{P(Z=j|\tilde{x})/P(Z=J|\tilde{x})}{P(Z=j|x)/P(Z=J|x)} = e^{\beta_j^T(\tilde{x}-x)}.$$

If we want to express the multinomial logit in terms of the probability of an event, we can begin with the probability that the outcome variable Z takes the reference value J. Using (4), we can see that

$$\sum_{j=1}^{J}\pi_j(x) = \sum_{j=1}^{J-1} e^{\beta_{0j}+\beta_j^T x}\pi_J(x) + \pi_J(x). \qquad (5)$$

Since the sum of all possible outcomes of Z defines the whole sample space, we have that

$$\sum_{j=1}^{J} \pi_j(x) = 1,$$

and therefore we can rewrite (5) as

$$\pi_J(x)\Big(\sum_{j=1}^{J-1} e^{\beta_{0j}+\beta_j^T x} + 1\Big) \Leftrightarrow$$
$$\pi_J(x) = \frac{1}{1+\sum_{j=1}^{J-1} e^{\beta_{0j}+\beta_j^T x}}.$$

Using (4) again we then also see that

$$\pi_j(x) \quad = \quad \frac{e^{\beta_{0j}+\beta_j^T x}}{1+\sum_{j=1}^{J-1} e^{\beta_{0j}+\beta_j^T x}}.$$

This is the probability of Z taking value j given a set of predictor values x.

It is also possible to split a multinomial model into several (in our case two) separate binary models. This have been examined in several studies, among those Hosmer and Lemeshow (2013, p. 282). Begg and Gray (1984) argues that the biggest issue with this is that it can lead to larger standard errors, but also that this problem is minimized if the chosen reference value is very dominant. As we shall see in chapter 3, this is the case for our dataset. Therefore we will use two separate binary logits in certain parts of our later analysis.

2.3.2 Parameter estimation

Fitting a model is done by performing a maximum likelihood estimation, that is finding the parameter values for which the probability of observed data is the greatest. The maximum likelihood equation comes from the probability distribution of Z, our outcome variable.

Following the presentation by Agresti (2002, p. 192), we can record the number of observations n_i in our data that has a *specific* set of predictor values $x_i = x_1, ..., x_n$, where $i = 1, ..., N$ and $\sum_{i=1}^{N} n_i = T$ is the total sample size. If y_i is the success count for all observations with specific predictor values x_i, the joint probability function can be written as

$$f(y|\beta) = \Pi_{i=1}^{N} \frac{n_i!}{y_i!(n_i-y_i)!}\pi(x_i)^{y_i}(1-\pi(x_i))^{n_i-y_i}.$$

If we instead look at the case when our outcome variable has a multinomial distribution, each element y_{ij} is then the observed counts of the j:th outcome value of Z, for a specific set of predictor values $x_i = x_1, ..., x_n$. The joint probability function, as shown in Dobson (2002, p. 141), can in this case be written as

$$f(y|\beta) = \Pi_{i=1}^{N} \frac{n_i!}{\Pi_{j=1}^{J} y_{ij}!}\Pi_{j=1}^{J}\pi_j(x_i)^{y_{ij}}. \qquad (6)$$

These joint probability functions are identical for $J = 2$. In order to find the coefficients (the β:s) of a multinomial logistic regression model we need to maximize the log-likelihood function. The likelihood function has the same appearance as (6), except now y is known to estimate β. Before we take the logarithm of (6) though, we can remove the terms which

do not include $\pi_j(x_i)$. The second derivative needed to find the maxima of (6) will remove these terms anyway, which means we can focus on the kernel

$$\Pi_{i=1}^{N}\Pi_{j=1}^{J}\pi_j(x_i)^{y_{ij}}. \qquad (7)$$

Taking logarithm of (7) we obtain

$$l(\beta|y) = \sum_{i=1}^{N}\Big(\sum_{j=1}^{J-1} y_{ij}(\beta_{0j} + \beta_j^T x_i) - \log(1 + \sum_{j=1}^{J-1} e^{\beta_{0j}+\beta_j^T x_i})\Big),$$

as shown in Agresti (2002, p. 273). If we now wish to find the values for β which maximize the function we calculate the partial derivatives $\frac{\partial\beta;y}{\partial\beta}$ and set to zero. We obtain a set of equations, that can be solved with the Newton- Raphson method presented in Agresti (2002, p. 144).

2.4 Assumptions

We have now seen the important relationships between the distribution of our outcome variable, our systematic component and the link function in order to find odds and probabilities from specific sets of predictor values. We have also looked closer at how the parameters are estimated. Before we start looking at creating a suitable model, we have to look at some assumptions that is done when working with multinomial logits. Although these models do not require normality, linearity or homoscedacity, there are a few other important criteria which have to be checked.

2.4.1 IIA

Independence of Irrelevant Alternatives is an assumption regarding proportion of probabilities when different values of our outcome variable are considered. If we have two values j and J which our outcome variable Z can take, we can for example look at the odds for this pair of values, as shown in Cheng and Long (2007):

$$\frac{\pi_j(x)}{\pi_J(x)} = e^{\beta_{0j}+\beta_j^T x} \text{ , for } j = 1, ..., J-1.$$

This odds is independent of any other value for our outcome variable, and only determined by the vectors of β_j and β_J.

The most common examples of violating this independence comes from discrete choice theory, with the classic being "Red bus, blue bus" exemplified by Tutz (2011, p.228). The tests designed to check the multinomial logit model for this type of violation compares coefficient estimates of a full model $\hat{\beta^f}$ to coefficient estimates of a restricted model $\hat{\beta^r}$ (Cheng & Long, 2007). The difference in $\hat{\beta^r}$ is simply that one value of the outcome variable has been removed. If the coefficient estimates are similar, independence is concluded.

One of these tests is called the McFadden, Train and Tye Test, which is a likelihood ratio test. Setting l_r as the log-likelihood function for estimates $\hat{\beta^r}$ of the restricted model, the test is formulated as

$$MTT = -2(l_r(\hat{\beta^f}) - l_r(\hat{\beta^r}))$$

Comparing results from the log-likelihood equation for $\hat{\beta}^r$ and the log likelihood for estimates $\hat{\beta}^f$, we can test for independence as MTT is χ^2- distributed with degrees of freedom as the rows of $\hat{\beta}^r$. This is equivalent to likelihood ratio test, which we present in section 2.5.2.

2.4.2 Multicolinearity

If two variables (in our case both nominal) are highly associated, this could have an effect when fitting our model. Cramér's V is a statistic of measure of such association (Liebetrau, p. 14-15), with values ranging from 0 to 1. Its based on the χ^2-statistic as presented in Agresti (2002, p. 22). For two nominal variables with s and t possible values respectively, we set $k = \min(s, t)$. With n being the total number of observations in our dataset, Cramér's V is calculated as

$$V_{Cramér} = \sqrt{\frac{\chi^2/n}{k}}.$$

The measure of association is considered high/very high for a value bigger than 0.5, while 0.2-0.5 is moderate and below 0.2 signifies low association to none at all. We will use this to measure all associations between our variables in section 3.4.

2.4.3 Outliers

Observations that differs greatly from the rest of data can also have a big effect when fitting our model. Visualizing residuals can be a great way to find outliers for possible removal. For generalized linear models, there are a couple of different residuals one can look at. We will focus on deviance residuals, which are the signed squared roots of an observation i to the overall deviance, as discussed in Agresti (2002, p. 142) and also Dobson (2002, p. 132).

Following notations from previous chapter, we have first that y_i is the observed counts of successes for the random variable Z, where Z here has a binomial distribution. Secondly, we have that $\pi(x_i) = P(Z = 1|x_i)$, the probability of success for an observation with the specific set of predictor values x_i. Thirdly, we have that n_i signifies the number of observations for every specific set of predictor values x_i, such that $\sum_{i=1}^{N} n_i = T$, that is the total sample size.

The deviance residuals can then be formulated as

$$d_i = sgn(y_i - n_i\pi(x_i))\sqrt{2y_i log(\tfrac{y_i}{n_i\pi(x_i)}) + 2(n_i - y_i)log(\tfrac{n_i - y_i}{n_i - n_i\pi_i})}.$$

We will plot at these residuals for every observation to find outliers when analyzing our selected model in chapter 5.

2.4.4 Perfect separation

A phenomena which is quite common when dealing with exclusively categorical data with many possible predictor variables is the so called Hauck-Donner effect. Hauck and

Donner (1977) discussed this in a much cited article where they pointed out a problem of convergence in the maximum likelihood estimation of logit coefficients. The issue is regarding low cell counts in frequency tables, which means that the probability of an event happening is extremely close to 0 or 1. This is also known as perfect separation, where for example a value of our response variable is not represented in all values of a predictor variable. For our data, this problem occurs for example when predicting motorcycle accidents against car/light truck accidents in different regions. Since all regions in Sweden do not have motorcycle accidents, this can lead to the variance of the coefficient becoming very large.

In our case, this issue would become apparent when looking at standard errors for coefficients in our model fit. Most analytical software also has automatic warnings when perfect separation is apparent.

2.5 Model diagnostics

We will now focus on the procedure to find the best model for our chosen response variable. The methods and statistics discussed are standard ways of measuring model fit, both relatively and absolutely.

2.5.1 Purposeful selection

There are several methods available to compare logit models. From a software standpoint, usually some form of stepwise procedure is implemented. This means going from a full saturated model including all predictors and interaction terms, to a smaller model (or vice versa) without losing the most important information. An alternative to a stepwise procedure is to use Purposeful selection as discussed in Hosmer and Lemeshow (2013, p. 89). The selection process has seven steps and is performed as follows:

1. We first do analysis using the chosen outcome variable together with one predictor variable at a time. Since our data is purely categorical, the recommended method is to create a contingency table of frequency counts. This means that we aggregate the number of observations for every value of our outcome variable and every value of our predictor variable. An important initial notice is when a table cell has zero counts. This can lead to perfect separation (as discussed in section 2.4.4), which causes the maximum likelihood estimate of coefficients to diverge. After checking every contingency table, we use the Pearson chi-squared test for independence (Agresti, 2002, p. 22), checking if the variable has a p-value less than 0.25. If this is the case, we can include it in the next step.

2. We use all the remaining variables in a multivariate model, now checking for p-values of the Wald statistic (Agresti, 2002, p. 11). For further inclusion of a variable in the model the Wald p-value should be les than 0.05. With categorical data, this means that for each predictor variable a reference value is chosen, and then compared to the remaining values of the variable. The Wald-statistic is used to check for significance between these values. This means that one value of the predictor variable can be very significant, while the others are not. Removing the entire variable is probably not a

good idea in this case. A better way is to recode the values if possible, i.e. merging insignificant values. Take for example a predictor $Month$, with has twelve values, one for each month. If several values (months) are insignificant, a way to merge these values would be to combine them into four new values, one for each season. We could then rename the predictor $Month$ to predictor $Season$.

After merging, we then compare the smaller model with the old model with a likelihood ratio test (as presented in section 2.5.2).

3. The likelihood ratio test is also used when removing an entire variable. It is important to compare the estimated coefficients of the smaller model with the old model. Even if a variable is non-significant according to the Wald-test, it may have a stabilizing effect on other variables. A big increase in size of a coefficient (say 20%) could mean that variables need to be added back to the model.

4. After possible merging of different predictor values, and complete removal of insignificant variables, the best model is chosen. The remaining set of predictor variables are now set. This model is called the *main effects* model.

5. We now check for possible interaction terms to improve the model. This perhaps requires more practical consideration than previous steps, and the purpose of the study must also be taken into account; some interactions may be especially interesting to present in our results. Interaction between variables with several possible values will create several interactions between the dummy variables (as discussed in section 3.2). For two variables with five possible values each (four dummy variables), sixteen interaction variables will be created. Needless to say, the model can grow fast. The statistical significance of the interaction terms is again measured by p-value of the Wald statistic. Holding the variables in the *main effects* model set, we can observe the changes of the models by adding and subtracting different interaction terms using step 2.

6. Finally, the model with chosen variables as well as the interaction terms is finally fitted before it can be used for analysis.

2.5.2 Likelihood-ratio test

In evaluating different logit models against each other, using a likelihood ratio test is a common method, presented in Agresti (2002, p. 24). It is basically a comparison between a model M_0 to a larger model M_1. An example of a hypothesis test is

$$H_0 : M_0 \text{ is valid}$$

$$H_1 : M_1 \text{ is valid instead of } M_0$$

where the corresponding likelihood ratio test divides the maximum likelihood functions for the two models L_0 and L_1 as

$$-2log(\tfrac{L_0}{L_1}) = -2(l_0 - l_1).$$

The resulting statistic asymptotically follows the χ^2-distribution under the null hypothesis, where degrees of freedom is the difference in number of parameters of the two models.

2.5.3 AIC

Another way of estimating model fit is using the Akaike information criterion (AIC) (Agresti, 2002, p. 216). The statistic is relative, meaning that one cannot derive meaning from its value alone, but rather in comparisons with other models. It uses the maximum value of the log likelihood function and the number of parameters of the model, and can be expressed as

$$\text{AIC} = 2k\text{- } 2\hat{l}.$$

Here k = number of parameters in model, and $\hat{l}$ is the maximized log likelihood. Ideally, the model chosen has the lowest AIC- value. However in choosing between a more complex model and a simpler model, as Agresti writes (2002, p. 216), a simple model "may be preferred because it tends to provide better estimates of certain characteristics of the true model". As we will see in Chapter 4, the model with lowest AIC is not always suitable.

2.5.4 McFadden R^2

A more absolute value of fit is the R^2 - statistic and it comes in several different modifications. For categorical data, McFadden's R^2 is commonly used. It uses the maximized log-likelihood for a chosen model we wish to evaluate, and divides by the maximized log-likelihood for the null model, i.e. the model with all predictor variables excluded. The statistic is defined as

$$R^2_{McFadden} = 1 - \frac{l_c}{l_0},$$

where l_c and l_0 is the maximized log-likelihood for the chosen model and null model respectively. This statistic has minimum value 0 and never reaches 1. In theory, the higher value of $R^2_{McFadden}$ signifies a better model fit for the chosen model, where 0.2 - 0.4 signifies a very good fit (Hensher & Stopher, 1979, p. 306).

2.5.5 Hosmer- Lemeshow statistic

Another common statistic to measure absolute goodness of fit is the Hosmer - Lemeshow statistic (Hosmer & Lemeshow, 2013, p. 158), which assesses observed event rates and then match them with the expected event rates for the model population.

Following the presentation from Fagerland and Hosmer (2012), since our outcome variable can take more than two values, we need to introduce some new notation. If Z is our outcome variable with multinomial distribution that can take J levels, we set the reference value to J as done in section 2.3. For a set of predictor variables $x_i = x_1, ..., x_n$ and the corresponding outcome variable z_i, we have t independent observations (x_i, z_i) for $i = 1, ..., t$.

We now want to code two new variables, setting $\hat{z}_{ij} = 1$ when $z_i = j$ and $\hat{z}_{ij} = 0$ otherwise, where $i = 1, ..., t$ and $j = 0, ..., J - 1$. Then when the model has been fitted, we set $\hat{\pi}_{ij}$ as the estimated probability for observation i for each outcome j. Sorting all observations by size for $1 - \hat{\pi}_{i0}$, we then split them into h groups, usually 10. The base for the test statistic

then is a table consisting of the sums of observed (O) and estimated (E) frequencies for each value j of our outcome variable,

$$O_{gj} = \sum_{i \in \Omega_g} \hat{z}_{ij}$$
$$E_{gj} = \sum_{i \in \Omega_g} \hat{\pi}_{ij}.$$

where $g = 1, ..., h$. With these sums we can calculate the Hosmer - Lemeshow statistic as

$$\sum_{g=1}^{h} \sum_{j=0}^{J-1} \frac{(O_{gj} - E_{gj})^2}{E_{gj}}.$$

For the null hypothesis that the fitted model is correct, this statistic has a χ^2- distribution, with $(h-2)$ x $(J-1)$ degrees of freedom.

2.5.6 Classification table & ROC- curve

Perhaps the most straightforward way to measure model performance is to create a classification table (Hosmer & Lemeshow, 2013, p. 169). This can be integrated in the process of cross validation, and basically compares the prediction of created model with the actual dataset. If we set c_{jk} as the expected count of observations in our dataset where the outcome value $z_i = j$, and simultaneously the observed count of observations in our dataset where the outcome value $z_i = k$, for $j = 1, .., J$ and $k = 1, .., K$ where $J = K$. When $z_i = j = k$, we have the correct expected value from our model. When $j \neq k$ we have a miss-classification. For a classification table, we want the diagonal cells to have as high frequency as possible:

$j \diagdown k$	1	...	K
1	c_{11}	...	
$\vdots$	...	c_{jk}	...
J		...	c_{JK}

Table 1: Classification table.

The miss-classification error is obtained by summarizing cells where $j \neq k$ and dividing by total number of observations. This value can then be compared to other models. Another example of a classification table can be seen in Hosmer and Lemeshow (2013, p. 171). The classification threshold is usually set to 0.5. This means that for observation i, if $P(z_i = j|i)$ is bigger than 0.5 then the count for $i = 1$ and otherwise 0. This threshold can then be changed to further investigate model performance.

A continuation of this classification method is visualizing the model performance in a plot. Receiver Operating Characteristic Curves (ROC Curves), does this by distinguishing between Sensitivity and Specificity (Tutz, 2011, p. 448). If $k = 1$ and $k = 0$ are two observed values of our outcome variable Z, and j is the expected value of Z from our chosen model, these concepts are defines as:

$$\text{Sensitivity: } P(j = 1|k = 1) \text{ and Specificity: } P(j = 0|k = 0).$$

These values plotted against each other creates a curve. The area under this curve (AUC) is then used as a measure of model performance on a continuous scale of 0 to 1. An AUC below 0.7 is seen as a bad fit and a measure closer to 1 is seen as better. Concrete examples of this can be seen in chapter 5.

2.5.7 k-fold Cross validation

A division of a large dataset into test and training set is often recommended (Hastie et. al, 2009, p. 241). Since the dataset in this study is large and permits this, an initial division of 70% to a training set and 30% to a test set is performed. We use the test data as an independent sample of the full data set, and use it to test our final model. The reason for this division is to detect possible deviations, and also to avoid overfitting.

The division of our training set can then be implemented in the process of cross validation (Hastie et. al, 2009, p. 241-245). One version of this process is called k-fold cross validation, which means that we separate our training set into k subsets. We leave one of the subset as a new test set, and use the other $k-1$ parts together as a new training set. We now train our data k times, using each of the k parts as test set once. This means several separate processes, but can be a thorough method to validate our results, and simultaneously avoiding overfitting. We effectively use our entire original training set as k training subsets and k test subsets.

We will use k-fold cross validation in Chapter 5 to validate our selected model, by comparing prediction performance results over k parts of the training set. This process will not be used in Chapter 4 to select the best model.

3 Data

In this chapter, we look closely at our data, and how it is structured. It is collected from police reports during the period 2003-2016, whose example structure can be seen in Appendix 8.1. The report includes record of the drivers details, along with the details of vehicle type and collision type. There are also a record of different traffic circumstances, e.g. at which type of location (at a crossing, roundabout) the accident happened as well as geographical location. Circumstances such as weather and road surface are also recorded. The entire collection of data has 179968 observations. The training set has 107981 observations and the test set has 71987 observations.

The outcome variable in this study is chosen as vehicle type, which means we focus only on the traffic accidents in the data involving heavier vehicles. Heavier vehicles here is defined as each of Heavy motorcycle, Car, Light truck, Heavy truck and Bus. Accidents involving only pedestrians, bicycles, light motorcycles etc. are excluded from this study.

The predictor variables are primarily selected with focus on possible external effect on an accident, such as weather and season. Other predictors of interest, such as collision type and road type, are also included in the full model.

3.1 Outcome variable

The proportion of accidents in the dataset involving cars is a clear majority. After careful examination, the conclusion is also made that cars and light trucks have very similar accident data. They seem to have similar driving pattern, and also have the same requirements for driving license. Merging these two values to a new value cars/light trucks seems both logical and practical.

We also merge the values heavy trucks and buses; this is partly due to their small proportion in data. Both of these vehicle types are involved in relatively few accidents, because they are more rare in traffic compare to cars. Another reason why we choose to merge these values is because they have similar problems in traffic, their size effecting driving patterns and also maneuverability.

The dominant value cars/light trucks is then chosen as the reference value (marked with r in our table) for the outcome variable in our model:

Variable	Values	Proportion of data
Vehicle type	Car/Light truck (r)	0.91
	Heavy truck/Bus	0.02
	Heavy MC	0.06

Table 2: Values of outcome variable.

3.2 Predictor variables

The predictor variables are exclusively categorical. Their meaning should be straightforward to deduce, however the difference between variables Environment (*Env*) and Region (*Reg*) might need a short clarification. *Env* has local meaning; an accident can happen in a small town in the countryside. In this case the value of *Env* is for a dense populated area.

Conversely, an accident can happen in a sparse populated area in the middle of a big city region. In this case the value *Reg* is for a big city. The value of *Reg* for a big city stands for a large urban area, in Sweden represented by Stockholm, Gothenburg and Malmö. The value for a normal region is all other regions.

After recoding (as discussed in 4.1) the predictors can be listed as follows (with reference value marked with r in our table):

Variable	Values	Proportion of data
Year (Year)	2003 (r)	0.08
	...	
	2016	0.06
Season (Ses)	Spring (r)	0.26
	Summer	0.22
	Fall	0.26
	Winter	0.26
Weekday (Week)	Week (r)	0.75
	Weekend	0.25
Region (Reg)	Big city (r)	0.51
	Normal	0.49
Speed plate (Speed)	50 km/h (r)	0.29
	70 km/h	0.26
	90 km/h	0.16
	Other	0.29
Environment (Env)	Sparse pop. (r)	0.55
	Dense pop.	0.41
	Unknown	0.04
Road type (Road)	Public road (r)	0.59
	Street	0.18
	Express/Motorway	0.18
	Other	0.06
Location type (Loc)	Road section (r)	0.66
	Roundabout	0.03
	Crossing	0.28
	Pavement/Other	0.03
Weather (Wear)	Fair (r)	0.78
	Other	0.07
	Rain	0.11
	Snow	0.05
Road surface (Surf)	Dry (r)	0.52
	Ice/Snow	0.05
	Wet	0.33
	Unknown	0.1
Lighting ratio (Light)	Daylight (r)	0.63
	Darkness	0.26
	Sunset/Sundown	0.11
Accident type (Acc)	Single (r)	0.34
	Crossing/turn	0.23
	Meeting	0.08
	Overtaking	0.24
	Other	0.11

Table 3: Values of predictor variables.

3.3 Dummy variables

When doing analysis with categorical predictors, a common way to account for the different values for each predictor x_k is to introduce dummy variables (Agresti, 2002, p. 178). Most software does this automatically, but using theory from section 2.3, it means that for every predictor x_k with J possible values, a new set of dummy predictors $x^d = x_1^d, ..., x_{J-1}^d$ is created. Notice that the reference value J is omitted from this set. This is because every dummy predictor x_j^d now is binary, with possible values $j = 1$ or $J = 0$. The systematic component of an observation can then be written as

$$\beta_0 + \beta_1 x_1^d + ... + \beta_{J-1} x_{J-1}^d.$$

The reference value J for the predictor variables is usually chosen as the most frequent value for each predictor. For example, the predictor *Loc* have the reference value *Road section* because of its proportion 0.66 of data.

3.4 Missing data

When we first think about missing data points, the ideal situation is if they follow criterion MCAR (Missing completely at random) (Allison, 2002, p. 3). This is a totally unbiased situation, since it assumes that missing data points are independent of both observable and non observable parameters for data.

A more common situation is that missing data points follow a criterion called MAR (Missing at random) (Allison, 2002, p. 3). Formally, MAR holds for two variables X and Y if

$$P(Y \text{ is missing}|Y, X) = P(Y \text{ is missing}|X).$$

The most obvious example of missing data points in our dataset, is where speed plates for our predictor variable *Speed* are not recorded. Many roads do not have frequent speed plates, which may be a reason. However, it might also be an effect of sloppy paperwork. In our dataset there are no clear relation between these missing datapoints and other variables. The missing data points for *Speed* does not seem to have much importance in our estimation of parameters neither, since they are a small proportion of observed data. Allison (2002, p. 5) argues in this case that the missing data mechanism is ignorable.

3.5 Correlation

One way to visualize strength of correlation between categorical variables is by mapping all measures of Cramers's V (discussed in 2.4.2). A heat map with all measures for variable pairs is shown in Figure 1 below:

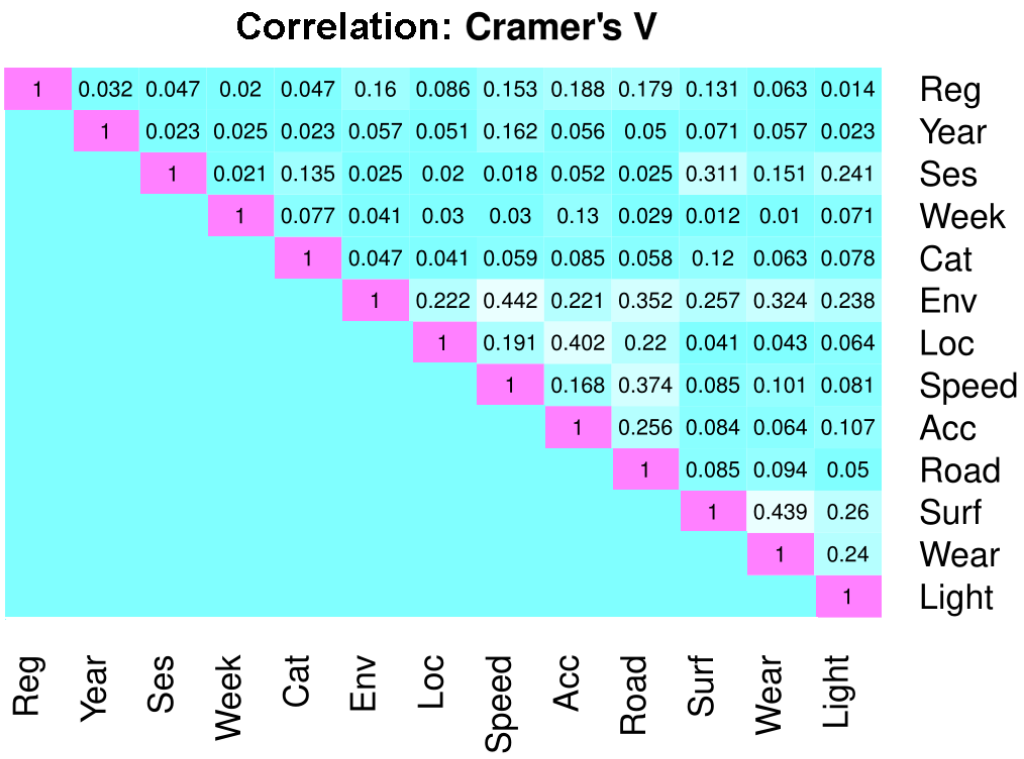


Figure 1: Correlation matrix for Cramer's V

We can see that all correlations are below 0.5, which means that correlation is present but not extremely high (Liebetrau, 1983, p. 3). Some correlations are moderately high, like between variables *Env* and *Speed* which reach 0.442. This does not necessarily pose a problem, but can add to the understanding of the model. The variables are therefore kept intact as we proceed.

4 Model selection

In this chapter, we go through the procedure to obtain the best model for our data. This will be done using the observations in the training set. Using Purposeful selection as outlined in section 2.5.1, we first determine which predictor variables that should be included in the *main effects* model. After that we try to add interaction terms, until the best model can be determined using theory from section 2.5.2 - 2.5.5.

4.1 Finding the *main effects* model

Before we begin fitting our models, we set reference values for outcome variable Z and each of our predictor variables $x = x_1, ..., x_n$ as shown in section 3.1 - 3.2. Having done this, the first step of the Purposeful selection is fitting a separate model for each predictor variable x_k. We do this to investigate which predictor variables can be considered significant when predicting the value of our outcome variable.

We determine significance by measuring the p-value for the Wald statistic. For Wald p-value ≤ 0.05 showing significance, we first notice several non-significant values in multiple predictor variables. This can be seen in Appendix 8.2. For example, the predictor variable *Speed* has originally thirteen values, one for each speed plate and also a value for *unknown plate*. Three of these values are insignificant with reference value 50 km/h. Since the majority of observations are spread over three speed plates (50 km/h , 70 km/h and 90 km/h as seen in section 3.2) we try to merge (recode) the remaining values in a new value called *Other*. At this step, we are also aware of possible perfect separation (as discussed in section 2.4.4) which can be seen by looking at the standard errors of the estimated coefficients.

After merging values we use the likelihood ratio test (as presented in 2.5.2) for the model with merged values (M_{New}) and the one without (M_{Old}). We have the hypothesis test

$$H_0 : M_{New} \text{ is valid}$$

$$H_1 : M_{Old} \text{ is valid instead of } M_{New}. \qquad (6)$$

We reject the null hypothesis if the likelihood ratio test (LTR) p-value is bigger than 0.05. This means we do not merge the values. If we cannot reject the null hypothesis, that is reject validity of the model with merged values, we can keep this recoding. The full table of important merging processes during recoding of the predictor variables is shown in Appendix 8.3. After recoding, we find that all our predictor variables are significant with p-value from the Pearson χ^2 test less than 0.25 as discussed in 2.5.1. The full list of results can be seen in Appendix section 8.4.

The next step is to look closer at a model including all our recoded predictor variables, called the *main effects* model. After fitting we conclude that the *main effects* model shall contain all our predictor variables, since all of them are significant. There are still insignificant values of several predictor variables, but fewer than before. This can be seen in Appendix 8.5.

We can now compare the *main effects* model to the *Null* model, i.e. the model without predictor variables. We use statistics presented in 2.5, where HL signifies the Hosmer-Lemeshow p-value:

Model	*df*	χ^2 p-value	*AIC*	$R^2_{McFadden}$	χ^2	*HL*
Null	0	.	64255	.	.	0
Main effects	88	<2e-16	53625	0.139	10807	5.384e-06

Table 4: Goodness of fit statistics for Null and Main effects models.

$R^2_{McFadden}$ of 0.139 is decent, since a value from 0.2-0.4 is considered a great fit (Hensher & Stopher, 1979, p. 306). The HL- value is minimal though, and the model could probably be extended for better fit. We do this by adding interaction terms.

4.2 Adding interactions

In extending the *main effects* model, we will focus on adding two-way interaction terms. This is because three-way interaction terms, and interaction terms with even higher complexity, are more difficult to interpret. Holding the predictor variables in the *main effects* model set, we can start with checking interaction terms for each predictor variable separately. We first look at the Wald p-value of each separate interaction term. If an interaction term is insignificant, we use the likelihood ratio test for the models with and without the insignificant term as in (6). The main purpose is to find a model that has a good fit, and ideally interactions with significance. However it is also important that our model is readable, i.e. not too complex. Examples of interactions which can be removed are shown in table 5 below. We remove all interaction terms which gives LTR p-values above 0.05 between models:

Model	Dropped interactions terms	AIC	LTR p-value
1	$-Road * Week$	66395	0.93
2	$-Wear * Surf$	66408	0.17
3	$-Week * Speed$	66426	0.98
4	$-Road * Ses$	66422	0.63
5	$-Wear * Week$	66406	0.18
6	$-Surf * Week$	66400	0.22
7	$-Reg * Surf$	66396	0.224
8	$-Loc * Week$	66390	0.26
9	$-Wear * Light$	66385	0.12
10	$-Loc * Surf$	66388	0.08
11	$-Loc * Road$	66396	0.21
12	$-Env * Light$	66390	0.44

Table 5: Removed interactions with LTR p-values > 0.05.

After removal, we obtain a large model, which we call Model A. This model has 26 interaction terms. Fitting this model to our data gives us improved goodness of fit compared to the *main effects* model:

Model	χ^2 p-value	AIC	$R^2_{McFadden}$	χ^2	HL
A	<2e-16	66380	0.15793	12292	0.1798

Table 6: Goodness of fit statistics for model A.

The $R^2_{McFadden}$ of 0.15793 indicates an decent fit and the HL of 0.1798 is a clear improvement. The problem with model A is readability; it is very difficult to read, since it has so many interactions. We therefore want to continue the selection process for interactions. We again remove all interaction terms which gives LTR p-values above 0.05 between models. Some examples are shown here:

Model	Dropped interactions terms	AIC	LTR p-value
14	$-Reg * Acc$	66389	0.13
15	$-Road * Acc$	66391	0.21
16	$-Road * Env$	66390	0.06
17	$-Ses * Env$	66398	0.42
18	$-Road * Light$	66412	0.2
19	$-Ses * Wear$	66422	0.07
20	$-Loc * Reg$	66432	0.05
21	$-Loc * Speed$	66445	0.08
22	$-Env * Speed$	66473	0.11
23	$-Ses * Light$	66499	0.06
24	$-Acc * Week$	66520	0.39

Table 7: Removed interactions with LTR p-values > 0.05.

We now obtain a much smaller model, and call it model B. This model has 13 interaction terms all two-way (as seen in Appendix 8.6). Fitting this model to our data gives us a worsened goodness of fit compared to model A:

Model	χ^2 p-value	AIC	$R^2_{McFadden}$	χ^2	HL
B	<2e-16	68339	0.128	9966	0.15

Table 8: Goodness of fit statistics for model B.

A $R^2_{McFadden}$ of 0.128 is not terrible, and the HL of 0.15 is again decent. Model B seems to have a worse fit than model A, but is much more readable. Simplifying interpretation is important to us when presenting results. Since we will be looking at odds ratios, a three-way interaction term would be intricate to discuss. We have a decision to make though, since the AIC of this model is higher than the previous model. We perhaps need additional measures of model performance. As presented in 2.3.2, we can express the probability $\pi_j(x_i)$ of an event given the outcome variable value j and a set of predictor variables $x_i = x_1, ..., x_n$. For every outcome j, we can then plot the probability of this value for observation i given the set of predictor variables for a chosen model. Looking at both model A and B, the probabilities for each outcome j can be seen in Figure 2 below:

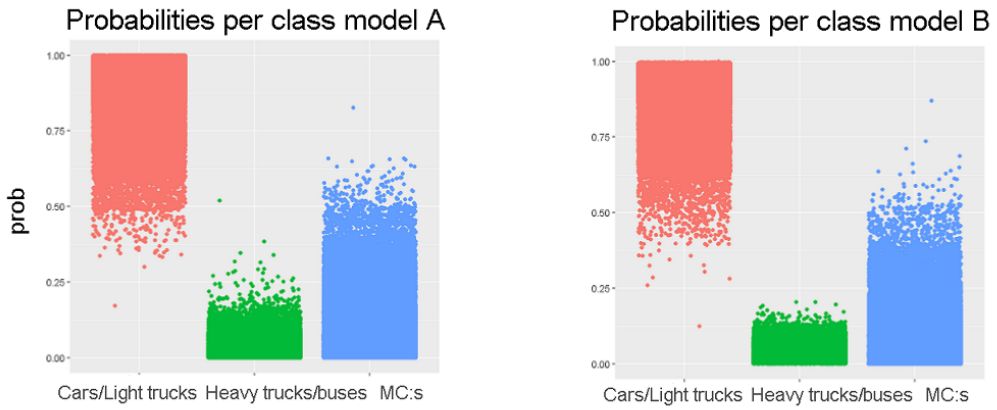


Figure 2: Probability plot for response levels, Model A vs. B

One point on the plot signifies the probability of a particular vehicle type being in an accident for a specific observation in data. We see that the biggest difference between the models is the probabilities for value Heavy truck/bus. The probabilities for this outcome value are visibly reduced using model B, while the changes in probabilities for the other values are less apparent. Since model B is much more readable, but still has a decent fit, it seems to be the best choice. The only other option would be to shrink the model B further. The problem with this is that the LTR p-values we obtain when comparing models with additional interactions removed are all bigger than 0.05. If we for example drop interactions $Env * Loc$ and $Acc * Light$, we get a new model C. Model C has the following values after fitting:

Model	χ^2 p-value	AIC	$R^2_{McFadden}$	χ^2	HL
C	<2e-16	68376	0.127	9872.8	0.05

Table 9: Goodness of fit statistics for model C.

The $R^2_{McFadden}$ drops marginally, however the drop in HL is considerable, down to 0.05. Simplifying model B seems to worsen fit to a high degree. Since model B is readable, removing more interactions seems pointless. Therefore, we have the following examples of fitted models, with goodness of fit statistics for comparison:

Model	χ^2 p-value	AIC	$R^2_{McFadden}$	χ^2	HL
Main	<2e-16	53625	0.139	10807	5.384e-06
A	<2e-16	66380	0.158	12292	0.1798
B	<2e-16	68339	0.128	9966	0.15
C	<2e-16	68376	0.127	9872.8	0.05

Table 10: Goodness of fit statistics for fitted models.

Taking readability into account, we choose to continue our analysis with model B. Before we look closer at the predictive performance, we first check the remaining assumptions for the multinomial model made in section 2.4. We use software to test IIA (Independence of irrelevant alternatives) by removing one value of the outcome variable at a time, and performing the McFadden, Train and Tye Test as discussed in 2.4.1. We obtain likelihood ratio p-values as shown in Table 11 below:

Removed outcome value	LTR p-value
Car/light truck	0.21
Heavy truck/bus	0.38
Motorcycle	0.35

Table 11: LTR test of IIA, model B.

None of the LRT p-values are less than 0.05, which means that IIA is not violated, and that the multinomial logit model can be used for the chosen predictor variables in model B. We also check the assumption of outliers; calculating the deviance residuals of a model, as presented in 2.4.3, is not informative for testing absolute model fit. Before splitting the training data for our k- fold cross validation however, it can be useful for finding outliers for removal.

As mentioned in theory section 2.3, it is possible to divide the multinomial logit model into separate binary logit models. The trade off is marginally higher standard errors for the estimated coefficients. Since our reference value Car/Light truck have proportion 0.91, this problem is minimized (Hosmer & Lemeshow, 2013, p. 282). Our outcome variable Z in this case has a binomial distribution with possible outcome values Car/Light truck and j, where j can take value Motorcycle or Heavy truck/bus depending on which separate binary logit model we fit. The deviance residuals when fitting model B for these separate models on the entire training set can be seen in Figure 3 below:

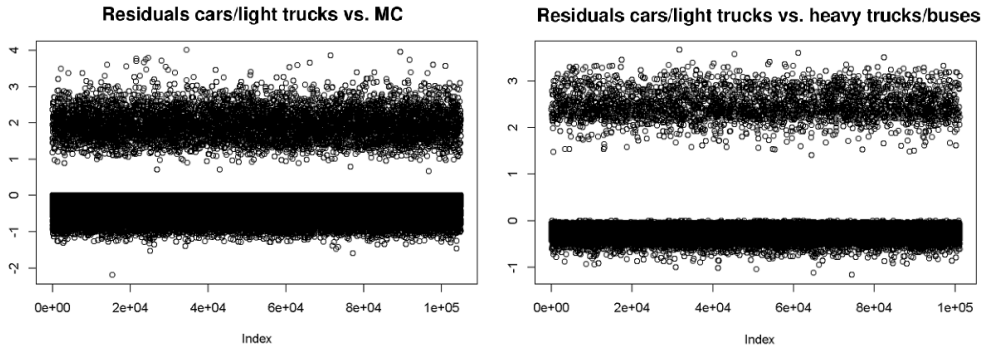


Figure 3: Deviance residuals for response levels, Model B

In general, deviance residuals are bigger for the binary model Car/Light truck vs. Heavy truck/bus. There are no extreme outliers for any of the models however. If there had been clear outliers, these would have been removed and model B could be fitted again. As it is, this will not be necessary. In addition to this, the same is also true for the other models examined in chapter 4; plots of deviance residuals for models A and C can be seen in Appendix 8.9.

5 Prediction

In this chapter we make our prediction based on model B. We first use cross validation to examine the model performance over k subsections of our training set (as discussed in section 2.5.7). This will be done using only training data. Finally we examine the model performance on our test data.

5.1 Prediction performance on training data

We first examine prediction performance of model B over the entire training set using cross validation. Choosing number of partitions $k = 10$, we partition the set of 107981 observations, which means 10797 observations for each of the k subsets. We then fit model B ten separate times, once for each of our new training sets i.e. on training data after omitting subset k. For comparison, we then look at the goodness of fit statistics as presented in section 2.5.2 - 2.5.5. Every model fit is then used to make prediction on the omitted subset k. To evaluate the performance of our prediction, we use classification tables to measure accuracy as well as ROC-curves to measure AUC. These concepts are discussed in section 2.5.6. For ten separate training results of model B, as well as the prediction result of these fitted models on the ten subsets, we obtain the following table:

k	AIC	$R^2_{McFadden}$	Missclassification	Accuracy
1	61336.96	0.12855	0.094	0.906
2	61515.28	0.12851	0.0905	0.9095
3	61342.85	0.12963	0.0926	0.9074
4	61457.44	0.12878	0.09168	0.908
5	61722	0.12927	0.086	0.914
6	61345.59	0.12877	0.093	0.907
7	61443.11	0.12904	0.091	0.909
8	61540.74	0.12721	0.0918	0.908
9	61657.38	0.12886	0.087	0.913
10	61448	0.12864	0.0913	0.9087
Mean	61480.935	0.12873	0.0909	0.909

Table 12: Results from 10-fold cross validation, Model B on training data.

The mean results shows an average accuracy level of 0.909 for prediction on our subsets. The variance of the missclassification error on the subset is 6.430773e-06. These are good values, but do not necessarily mean that the model has a particularly good prediction performance. In our case, a model predicting that all accidents involve cars/light trucks will have similar values. The small variance over the ten subsets is a good sign; the model performs the same over the entire training set.

We can also look at the prediction performance on our subsets in a classification table (as discussed in section 2.5.6) with threshold 0.5. The mean of each ten predictions is shown below:

	Car/Light truck	MC	Heavy truck/Bus
Car/Light truck	9808	664	318
MC	5	5	2
Heavy truck/Bus	0	0	0

Table 13: Mean classification table, Model B on training data.

We see that the model B completely fails to correctly classify the outcome Heavy truck/bus, in fact ignores this outcome completely for threshold 0.5. The prediction results are somewhat better for outcome Motorcycle, while the model correctly classifies the outcome Car/Light truck at a high rate. The mean missclassification of 0.0909 is not bad, but since our data is skewed with the outcome Car/Light truck having proportion 0.91 of our data we can also evaluate prediction performance for different thresholds. We will do this for the prediction on our test data in the next section.

One value missing from our table is the AUC (Area under the curve), determined from plotting the ROC- curve. Since model B seems to have better predictive performance for outcome Motorcycle than outcome Heavy truck/bus, want to visualize this curve for each of these categories separate. Fitting model B on each of our new training sets, we obtain

ten ROC-curves each for outcome values Motorcycle and Heavy truck/bus. Plotting the mean of these curves we see the result in Figure 4 below:

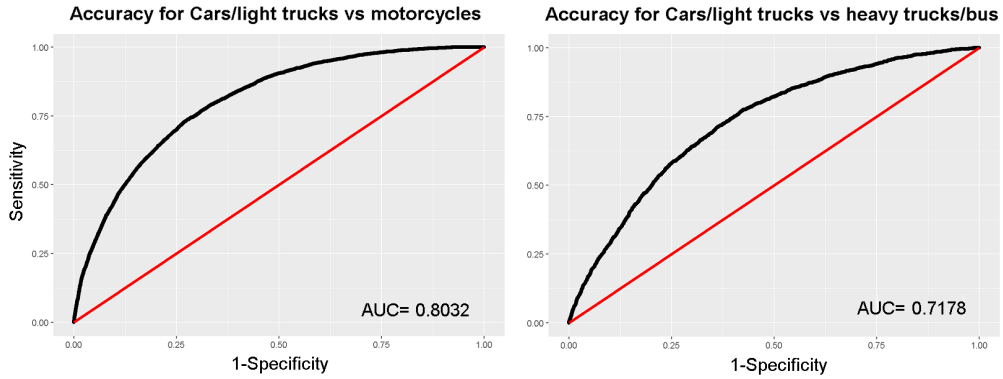


Figure 4: ROC-curves for response levels, Model B

We see that the AUC for cars/light trucks vs. motorcycles is considerably higher than for cars/light trucks vs. heavy trucks/buss. In fact, the AUC for cars/light trucks vs. heavy trucks/buss of 0.7178 is a sign of pretty bad predictive performance. Considering the results shown in the mean classification table, this is not a surprise. Model B clearly fails to distinguish between outcomes Car/Light truck and Heavy truck/bus for our training set. This could be a result of underfitting, i.e. there are predictor variables missing from our dataset which could give a better predictive performance for our chosen outcome variable of vehicle types.

5.2 Prediction on test data

Finally, we also want to examine the performance of model B on a new dataset; the test set. This set has 71987 observations with the same predictor variables $x = x_1, ..., x_n$ as the training set. After fitting the model B on our training set as in the previous section, we predict on the test set. We obtain the following classification table:

	Car/Light truck	MC	Heavy truck/Bus
Car/Light truck	65389	4593	1909
MC	51	41	1
Heavy truck/Bus	0	0	0

Table 14: Classification table, Model B on test data.

The overall predictive performance is slightly worse than the mean result from our 10 subsets of the training data in the previous section. We have a missclassification error of 0.091 for threshold 0.5, an the model again fails to distinguish especially between cars/light trucks and heavy trucks/buses. Before checking predictive performance for

different thresholds, we can look at the probabilities for each outcome j fitted on the training set and predicted on the test set. The results are shown in Figure 5 below:

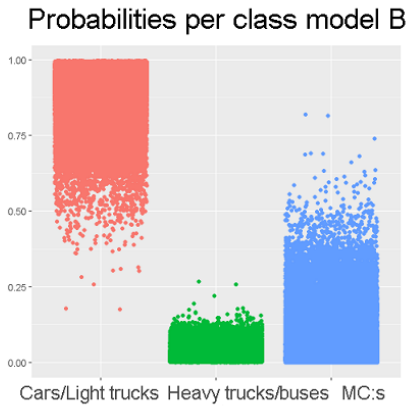


Figure 5: Probability plot for response levels, Model B

Similar to our previous probability plot for the entire training set (Figure 2), the evident missclassification of the outcome value Heavy truck/bus for threshold 0.5 is hardly surprising looking at these probabilities. If we instead examine predictive accuracy for several thresholds we can compare the two outcome values Heavy truck/bus and Motorcycle, again using separate binary logits. The accuracy of Model B on test data over different thresholds is seen here in Figure 6:

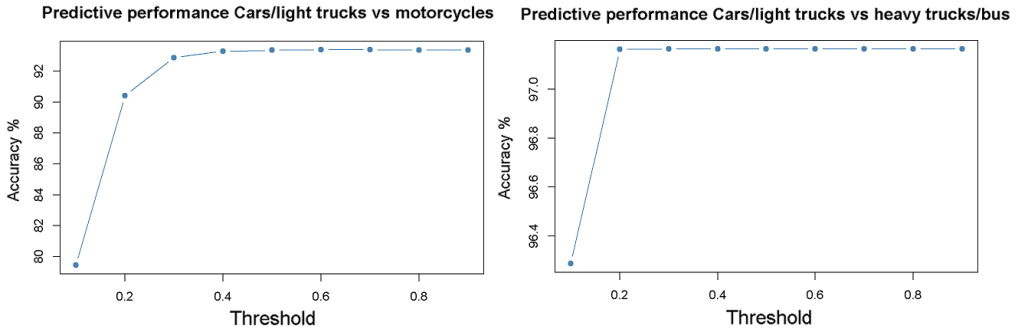


Figure 6: Accuracy vs. threshold, Model B

First, we see that overall predictive accuracy increases sharply as our model starts to classify the vast majority of observations as cars/light trucks. Since our data is skewed, increasing the threshold to above 0.2 will mean that the model for cars/light trucks vs. heavy trucks/buses will not even recognize Heavy trucks/buses as a value. All observations with outcome other than cars/light trucks will be treated as a missclassification. The binary logit for cars/light trucks vs. motorcycles has a similar accuracy development over thresholds. The difference is that model B is better at distinguishing between values

cars/light trucks and motorcycle, which means a higher proportion of missclassification between the observed values and the expected values from the model. In general though, we see that accuracy increases as the threshold rises i.e. the classification of cars/light trucks starts to dominate. In the same way as in previous section, we also want to look at the ROC- curves for the separate binary logits on test data. These are seen in Figure 7 below:

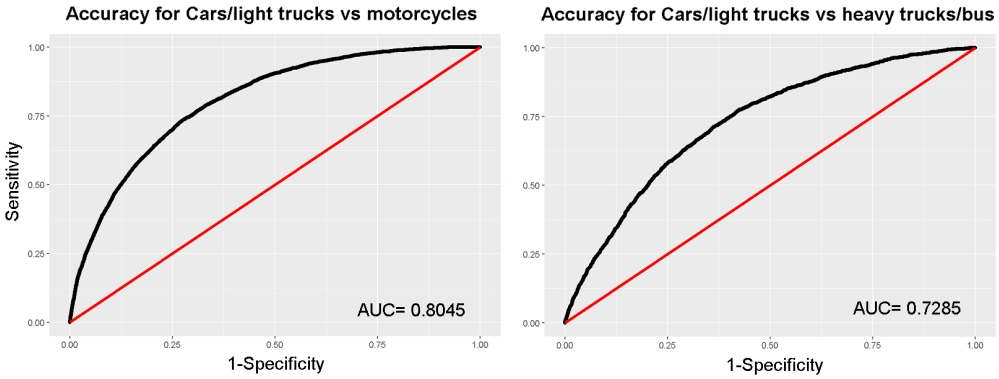


Figure 7: ROC-curves for response levels, Model B

As discussed in section 2.5.6, the bigger area above the diagonal line, the better our model classifies the different vehicle types from data (Tutz, 2011, p.448). In the case of classification of cars/light trucks vs. motorcycles, the model shows a decent accuracy with area 0.8045 under the curve. The classification for cars/light trucks vs. heavy trucks/buses shows a significant drop in accuracy, with AUC = 0.7285. This value is, just as the corresponding value for our training data in figure 4, just above the accepted level considering that an AUC less than 0.7 signifies bad fit.

In summary looking at the results for our prediction on test data, we see very similar results to our prediciton on the training set using cross validation in section 5.1. This is a good sign, since it gives some indication that the model gives consistent results. Among the key findings are that the model B fails to distinguish between outcome values Cars/light trucks and Heavy trucks/buses. This fact is perhaps a result of underfitting, and possibly a result of big difference in proportions between these values in our dataset. Looking at the separate binary logit models, this becomes particularly evident. The multinomial logit does not have a particularly bad fit overall however, and looking at the model B fit statistics on our test data we again see decent goodness of fit:

χ^2 p-value	AIC	$R^2_{McFadden}$	χ^2	HL
<2e-16	45306	0.134	6922	0.328

Table 15: Goodness of fit statistics for Model B on test data.

The $R^2_{McFadden}$ statistic of 0.134 is slightly higher than 0.128 for fitting on our training

data. The HL of 0.328 is also noticeably higher, compare to 0.15 on our training data. This means that we have now fitted model B on all available data, an obtain similar results. If we also plot the deviance residuals for model B on test data in Figure 8 below, we see a similar result as on training data:

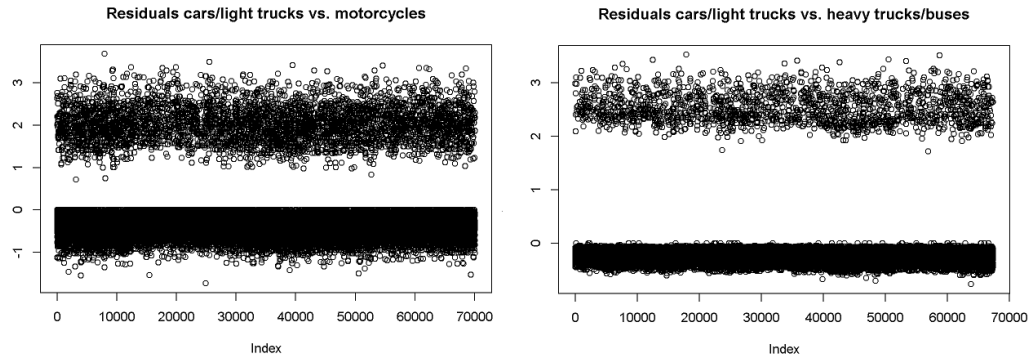


Figure 8: Deviance residuals for response levels, Model B

Just like for the deviance residuals for the model fit on training data in Figure 3, there are no clear outliers to be removed. As our results seems stable, and our prediction performance is decent to some degree, we can now proceed by interpreting the estimated coefficients of the fitted model B.

6 Results

In this chapter, we examine the results obtained by the model B. The coefficients to be interpreted when looking at the estimations from the multinomial logit model are the odds ratios. For two different sets of predictor variable values x and $\tilde{x}$, these are defined in section 2.3.2 as

$$\frac{P(Z=j|\tilde{x})/P(Z=J|\tilde{x})}{P(Z=j|x)/P(Z=J|x)} \quad = \quad e^{\beta_j^T(\tilde{x}-x)}. \quad (8)$$

where Z is a random variable with multinomial distribution and the possible outcome values are $j = 1, ..., J-1$, and J is set as the reference value. For an odds ratio bigger than 1, this simply means that the odds $P(Z = j \mid \tilde{x})/P(Z = J \mid \tilde{x})$ is greater than the odds $P(Z = j \mid x)/P(Z = J \mid x)$. The interaction terms are then ratios of the odds ratios as discussed by for example Norton et al. (2004); for a specific predictor variable x_k with values r and s, we have the ratio

$$\frac{P(Z=j|\tilde{x},x_k=r)/P(Z=J|\tilde{x},x_k=r)}{P(Z=j|x,x_k=r)/P(Z=J|x,x_k=r)} \Big/ \frac{P(Z=j|\tilde{x},x_k=s)/P(Z=J|\tilde{x},x_k=s)}{P(Z=j|x,x_k=s)/P(Z=J|x,x_k=s)}, \quad (9)$$

, that is the ratio of the odds ratios evaluated at different values of x_k, where the value s is set as the reference value of x_k.

Since we have many significant predictor variables and interactions, we present the results for two separate binary models. We then have odds ratios for the outcome values Car/light

truck vs. Motorcycle, and also odds ratios for values Car/light truck vs. Heavy truck/bus. The estimations of the odds ratios are still calculated from the full multinomial logit model, specifically model B. Car/light truck is the reference value for both these models. We choose to present odds ratios in text as Variable(reference vs. non reference). Interaction terms becomes Variable(reference vs. non reference), Variable 2(reference vs. non reference).

For example, say we have a significant coefficient as the odds ratio between the values *Fair* and *Rain* for predictor $Wear$ in the binary model Car/light truck vs. Motorcycle. This odds ratio would be written as Wear(Fair vs. Rain). The interaction with another variable *Light* for values *Daylight* and *Darkness* is then written as Wear(Fair vs. Rain), Light(Daylight vs. Darkness). In effect, this means that the interaction coefficient signifies the ratio between Wear(Fair vs. Rain) at Light(Darkness) and Wear(Fair vs. Rain) at Light(Daylight).

The odds ratios and interactions picked below for further analysis is examples of results after fitting Model B. The full list of significant odds ratios and their 95 % Wald type confidence intervals (Agresti, 2002, p. 13) for model B can be seen in Appendix 8.7-8.8.

6.1 Car/light truck vs. Motorcycle

We begin with looking first at the binary model for cars/light trucks vs. motorcycles , focusing on the statistically significant coefficients. Looking first at the variable Env, environment, we have several coefficients and interactions to consider

Label	OR estimate	Wald CI
Env(Sparse pop. vs Dense pop.)	1.34	1.19- 1.54
Env(Sparse pop. vs Unknown)	1.98	1.51- 2.66
Env(Sparse pop. vs Dense pop.), Acc(Single vs Meeting)	0.249	0.45-0.92
Env(Sparse pop. vs Dense pop.), Week(Week vs Weekend)	0.695	0.618- 0.78

Table 16: Examples of odds ratios, variable Env.

Firstly, we can see that the chance of a motorcycle being in an accident compared to a car/light truck, is higher for value *Dense pop.* compared to *Sparse pop.*, and the same goes for value $Unknown$. This can be seen because these coefficients are above 1.0, namely 1.34 and 1.98 respectively. That motorcycles have relatively higher chance for accident in dense populated areas is perhaps not so shocking. For the interaction term $Env * Acc$, the interpretation would be that the relative chance of motorcycles being in an accident in dense populated area is much higher (about four times as high) for single accidents compared to meetings. The interaction $Env * Week$ then says that the relative chance is also higher for weekdays compared to weekends. This could mean several things; that motorcyclists travel more outside the dense populated areas at weekends, or that the people in the sparse populated areas mostly drive at weekends. Further analysis would be necessary to draw secure conclusions on these types of results. We continue with looking at the predictor variable Acc:

Label	OR estimate	Wald CI
Acc(Single vs Crossing/turn)	0.496	0.39- 0.63
Acc(Single vs Crossing/turn), Loc(Road section vs Roundabout)	0.629	0.44-0.85
Acc(Single vs Overtaking), Light(Daylight vs Darkness)	1.51	1.07-2.2
Acc(Single vs Meeting), Surf(Dry vs Wet)	0.43	0.26-0.72

Table 17: Examples of odds ratios, variable *Acc*.

We see that the chance of a motorcycle being in an accident compared to a car/light truck, is higher for value *Single* compared to *Crossing/turn*. Single vehicle accidents are especially common for motorcycles while accidents at crossings/turns have relatively less chance of happening compared to for cars/light trucks. The interaction $Acc * Loc$ then tells us that this relative chance is even less for motorcycles compared to for cars/light trucks at a roundabout.

The interaction $Acc * Light$ is measuring the odds ratio for accident on a motorcycle when overtaking compared to a single vehicle accident, at different lighting. Evidently, the odds ratio for overtaking accident vs. single vehicle accident is higher when its dark outside. This doesn't necessarily mean that you should avoid overtaking with motorcycle when its dark, but may be an indicator of the risk of reduced vision.

$Acc * Surf$ indicates that the odds for a motorcycle accident in meeting is bigger when the road is dry. This is maybe no surprise, since the vast majority of accidents happen on dry roads. Add also that fewer motorcycles are in traffic when its raining.

If we instead look closer at variable *Speed*, we see that the odds for accident with a motorcycle at 90 km/h, is considerably lower than with cars/light trucks, with odds ratio at 0.244:

Label	OR estimate	Wald CI
Speed(50 km/h vs 90 km/h)	0.244	0.2- 0.297
Speed(50 km/h vs 90 km/h), Light(Daylight vs Darkness)	0.549	0.363-0.83
Speed(50 km/h vs 90 km/h), Light(Daylight vs Sunset/Sundown)	1.52	1.098- 2.14

Table 18: Examples of odds ratios, variable *Speed*.

The interaction term then $Speed * Light$ shows that this odds ratio is higher for accidents in daylight than in the dark, but also higher at dusk/dawn than daylight. In other worlds, there is relatively higher odds for accident with motorcycles at 90 km/h at dusk/dawn, than in both daylight or darkness compared cars/light trucks. Finally we also look at the variables $Week$ and *Reg*:

Label	OR estimate	Wald CI
Weekday(Week vs Weekend)	1.85	1.67-2.02
Reg(Big city vs Normal)	0.81	0.68- 1.01

Table 19: Examples of odds ratios, variables $Weekday$, Reg.

We see that motorcycles have relatively higher chance of accident than cars/light trucks on the weekend. This seems logical, since most people do not take motorcycles to work, but rather drive for fun on the weekend. We also see that motorcycles have relatively lower chance for accident outside the big city regions. In general, the motorcycle accidents are focused around the big cites. Keeping in mind that the reference category cars/light trucks makes up for 91% of total accidents, the odds ratio we get for cars/light trucks vs. motorcycles gives a decent view of how motorcycle accidents compare in different situations.

6.2 Cars/light truck vs. Heavy truck/bus

If we instead look at the binary model for cars/light trucks vs. heavy trucks/buses (value Car/light truck is the reference value which value Heavy truck/bus is compared against), we begin again with the odds ratios for predictor Env:

Label	OR estimate	Wald CI
Env(Sparse pop. vs Dense pop.)	0.724	0.576- 0.89
Acc(Single vs Meeting)	0.5	0.3- 0.79
Env(Sparse pop. vs Dense pop.), Acc(Single vs Meeting)	1.12	0.76- 1.7
Env(Sparse pop. vs Dense pop.), Loc(Road section vs Crossing)	1.42	0.4-1.88

Table 20: Examples of odds ratios, variables Env.

Conversely to our previous model results in table 15, heavy trucks/buses have a relatively higher chance of accident in sparsely populated areas. This result is most likely a consequence of driving patterns for bigger vehicles. The interaction term $Env * Acc$ then tells us that this odds ratio is relatively higher for meeting accidents than for single accidents. Single vehicle accidents are twice common among Heavy trucks/ buses than meeting accidents, but the results suggest that meeting accident at least are more common in sparsely populated areas than in dense populated ones.

The interaction term $Env * Loc$ also shows that accidents at crossings are more common than accidents at road sections in sparsely populated areas than in dense populated ones. The coefficient is 1.42 which indicates a rather large difference; one reason could be more dangerous crossing locations outside the dense populated areas, but this needs to be investigated further. As seen for odds ratios with variable Acc, the relative chance for heavy trucks/buses being in an accident at crossings/turns is much lower than for single accidents:

Label	OR estimate	Wald CI
Acc(Single vs Crossing/turn)	0.274	0.175-0.42
Acc(Single vs Overtaking), Light(Daylight vs Darkness)	1.36	1.0-1.84
Acc(Single vs Crossing/turn), Surf(Dry vs Ice/Snow)	1.72	1.14- 2.6
Acc(Single vs Crossing/turn), Week(Week vs Weekend)	1.73	1.19- 2.5

Table 21: Examples of odds ratios, variables *Acc*.

A contribution factor to this should be that certain roads often are designed for large vehicle transportation, especially for trucks, to simplify access to industries etc. An interesting interaction coefficient is $Acc * Light$ at 1.36. The odds ratio for accidents in single vehicle accidents compared to accidents in overtaking is higher for dark lighting for heavy trucks/buses. This can have many explanations, one of which can be that heavy trucks/buses in general add more risk in traffic situations trying to overtake other vehicles, and in dark lighting this risk is increased. Reading the interaction $Acc * Surf$ wee also see that odds ratio for accidents in single vehicle accident compared to crossing the road or turning is higher for Icy/Snow road surface. A relatively higher risk for heavy vehicles to have accident in crossing the road/ turning with icy surface seems logical. In regards to variable $Acc * Week$, the fact that the chance for an accident crossing the road or turning are relatively higher at weekends is somewhat of a surprise. This is also difficult to explain without further studies.

In some final examples, we see that the interaction $Speed * Light$ gives that the odds ratio for accidents at 90 km/h roads compared to 50 km/h roads is bigger at dark lighting:

Label	OR estimate	Wald CI
Speed(50 km/h vs 90 km/h), Light(Daylight vs Darkness)	2.32	1.69-3.23
Weekday(Week vs Weekend)	0.24	0.194- 0.29
Weekday(Week vs Weekend), Light(Daylight vs Darkness)	1.32	1.0- 1.72
Reg(Big city vs Normal)	1.11	0.99- 1.23
Reg(Big city vs Normal), Light(Daylight vs Darkness)	1.3	1.08- 1.58

Table 22: Examples of odds ratios, variables *Speed*, $Weekday$, *Reg*.

A difference in 2.52 suggests that bigger roads are significantly more dangerous for Heavy trucks/buses in dark lighting. Looking at the variable $Weekday$, we see that the relative chance for accidents with heavy trucks/buses are much lower on weekends. This is probably because these vehicle types are less in traffic during the weekend, the inverse to the situation for motorcycles. This odds ratio is somewhat bigger in dark lighting, which could reflect driving patterns at the start or end of the week.

Finally, the coefficients for variable *Reg* first show that the odds for accident in big city regions lower than outside big city regions. This odds ratio is bigger for dark lighting, perhaps following the same pattern as seen for other interactions. The odds for accident seems to increase somewhat at dark lighting for heavy trucks/buses, compared to cars/light trucks.

7 Discussion & conclusion

Based on model B, we have reached some results of limited significance. In this chapter we will have a quick discussion before concluding the results of this report.

7.1 The model

The multinomial logit model can be used to fit our data and obtain somewhat significant results; the assumptions made in section 2.4 holds. Finding a *well* fitting readable model was difficult for the predictors used. Even the less readable bigger models (model A and bigger) have limited predicitve performance. Concerning our tests of goodness of fit, the bigger model A showed better fit than the smaller models; the Hosmer Lemeshow p-value and the $R^2_{McFadden}$ were both higher, while AIC was lower than for smaller models. The actual predictive performance on our training set was also slightly more accurate. The performance was not dramatically better though, and taking into account the readability of the model, the smaller model B was the best choice. This model had decent values for goodness of fit, while also being more readable. When starting to shrink this model even further, we immediately saw a increase in AIC, as well as a decrease in Hosmer Lemeshow p-value and the $R^2_{McFadden}$-statistic. The predictive performance also became less accurate.

All models failed in prediciton to distinguish between the dominant outcome value Car/light truck and the two other outcome values. This was most apparent for the value Heavy truck/Bus; presumably this is an effect of underfitting, i.e. the predictors used were not able to capture vital differences between the outcome values. In other words, the model may have been to simple. This is reflected in both the classification tables and the ROC-curves. However, this seemed to be less evident when distinguishing between outcome values Car/light truck and Motorcycle. Aspects of these two vehicle types, such as driving pattern in regards to time and place as well as accident type, probably distinguished them more clearly from each other. Focusing on the binary logit Car/light truck vs. Motorcycle would be best for this particular dataset.

One alternative to using the multinomial logit model for our dataset and topic is using a nested logit model; this model is particularly useful if the test for IIA fails, i.e. independence of irrelevant alternatives is rejected. The theory for this model will not be presented here, but is discussed at length by Agresti (2002, p. 361-366).

7.2 The data

Overall, traffic safety is a thankful topic for data analysis, since it is prioritized both in the public and private sector. There is also a clear overriding goal in all studies on traffic safety; preventing deaths. Currently, the used police reports are one of two main open sources of data; the other one being hospital reports. The hospital reports records in more detail severity of accident, time of arrival and other circumstances important to the health of a patient. Together, these data sources are best suited to studies on whether an accident is fatal or not, rather than what caused the accident. They also perhaps highlight

the human aspect of the accident, i.e. important factors such as alochol level and age of individual. For more in depth analysis of accident causality and technical aspects of the accident, records of the kind available to insurance companies and car manufacturers would be extremely helpful. The police reports from period 2003-2016 possibly gave us an underfitted model with vehicle type as outcome variable. To create a more complex model, access to data including more predictor variables would be necessary. Our data provides some insight in differences between vehicle types in accidents, but an addition of hypothetical predictors such as *vehicle brand*, *year of manufacture* and *tire condition* would provide for an even more interesting analysis with greater insights. Other studies on the same dataset could also be designed with accident type (i.e. if the accident happened at meeting, while crossing the road or in overtaking another vehicle) as outcome variable. This study would probably also give interesting results. Because the topic is popular, possible findings could easily be compared both to previous studies and studies from other countries.

7.3 The results

The model gives several interesting results, some expected and others less expected. Since the cause of accident often is complex, the conclusions made from these results are speculative, and further studies are necessary to infer definite statements from data. Keeping in mind that all results are relative to the value Car/light truck, we can see tendencies and connections however. If we begin with motorcycles, they are involved in accidents at a greater rate in dense populated areas, while also having much lower risk of accident in meetings and overtaking other vehicles than in single vehicle accidents. No surprise there, perhaps. Motorcycle traffic is heavier in and around bigger urban areas; the fact that chances of accident in meeting and overtaking is lower could be an effect of not driving as frequently on specific roads with heavy traffic. Single vehicle accidents are the most common accident type, but this is particularly true for motorcycles. The point of driving a motorcycle is somewhat lost on a crowded highway perhaps. Motorcycles are also involved in accidents on 50 km/h roads at much higher rate than on 90 km/h roads, and are more accident prone on weekends compared to weekdays. This is also expected, since motorcycles generally are more of a hobby vehicle, and something you drive in your free time. This is also connected to the weather and road surface aspects, the clear majority of motorcycle accidents happen on dry roads.

Motorcycle accidents happen predominantly in big city areas, and the driver crashes by himself/herself. This is true also for cars/light trucks, but not as predominant. An interesting result in relation to this though, is that motorcycle accidents that are *not* single vehicle accidents have a much higher chance of happening at 90 km/h roads than 50 km/h roads; if a motorcyclist decides to overtake another vehicle, the chance for accident is higher on a 90 km/h road than 50 km/h road. This is true also for meetings, crossing the road and other types of accidents. Why is this? Roads with speed plate 90 km/h certainly see heavier traffic in some cases. The big roads in Sweden, E4, E18 and E20 also have most accidents. Not only road type, but the specific road number could be important information in this case. The human factor can not be underestimated neither; motorcyclists could be driving more recklessly on 90 km/h roads.

What about time of day? What relative effect does lighting have on accident odds? For motorcycles, the chance of accident in meetings, crossing the road and other types of accidents are higher when it is dark outside. The same is true when it is sunset/sundown. The most reasonable explanation would be decreased visibility for other vehicles on the road. At sunset/sundown, the relative chance is even bigger on a 90 km/h road compared to a 50 km/h road. This is not true for dark lighting however; the relative chance of accident at dark lighting is greatly reduced on 90 km/h roads compared to 50 km/h roads. What can one learn from this? It is difficult to examine without looking closer into driving patterns. Higher relative chance of accident on 90 km/h roads at sunset/sundown, may have less to do with visibility and more to do with what type of road and what time of day it is. For roads with heavy traffic during rush hour, chance of accident will probably increase for all vehicle types. In this case, specific time of day would be interesting to have recorded. Regarding the relative increase of motorcycle accident risk on weekends, this is primarily true in daylight. Both accidents in darkness and sunset/sundown are half the risk on weekends, again showing the importance of studying driving patterns both geographically and in time.

If we switch focus to accidents involving heavy trucks/buses relative to cars/light trucks, we have a couple of similarities with above discussed results for motorcycles. For example, single vehicle accidents are the most common accident type by a large margin. Most accidents involving heavy trucks/buses also happen in road sections. Lighting also seem to have an effect in similar circumstances; as with motorcycles, accidents on 90 km/h roads have higher relative risk in darkness than in daylight. There are however several big differences as well. To begin with, accidents with heavy trucks/buses are relatively more common in sparse populated areas. This is especially true with accidents at meeting, and at crossings. Vehicle size is most probably a factor here, as is driving pattern. Heavy trucks/buses are much more likely to be in a meeting accident, because of size. Accidents at crossings are also more common among these vehicles, because of impaired speed of maneuvers. The issue may not only be connected to vehicle size however; heavy trucks/buses are often forced to drive on smaller roads in order to reach their destination. Better planning of roads could probably prevent some of the accidents of these vehicles. Heavy trucks/buses are also involved in accidents at higher rate outside the big city areas, something that could be added to the same discussion.

Heavy trucks/buses drive less on weekends, since they are most commonly used in a profession. Therefore this result is almost inverse to those for motorcycle accidents; heavy trucks/buses have much less relative chance to be in an accident on the weekend. There are some interesting results to add to this fact though. The relative risks of accident in meetings, crossings and overtakings are higher at the weekend compared to in the week. Single vehicle accidents are then relatively less common. Does this have to do with other road-users, or truck drivers/bus drivers themselves? Additional hypothetical predictors could prove very useful in this case, for example *alcohol level of driver* and *time of accident*. The relative risks of accident in meetings, crossings and overtakings are also higher at snowy/icy road surface, as well as on wet road surface. This is a big difference compared to accident data for motorcycles; very few MC:s are driving when the road surface is not dry. Heavy trucks/buses certainly seems to be effected by change in road surface, especially when performing certain maneuvers. Heavy trucks/buses can be harder to control during a slide for example, because of its weight. The weather may not be the only reason for

the accident however. Although professional drivers should in theory be better drivers, the human factor can not be dismissed. There are several cases of truck drivers getting caught without driving license every year.

In summary, we have some expected results and some less expected. For the general case, a motorcycle accident happens on a dry road, in a big city area, and often on the weekend. The accident does not involve any other vehicles, and the speed plate is 50 km/h. On the other hand accidents involving heavy trucks/buses happen outside big city areas, on a dry road during the week. These results are confirmed by the model. Among the less expected is the tendency that motorcycle accidents that are *not* single vehicle accidents have a much higher chance of happening at 90 km/h roads than 50 km/h roads. This could be investigated further. For heavy trucks/buses, the relative increase in accident risk at meetings, crossings and overtakings on certain road surfaces could also be interesting to study closer.

7.4 Conclusion

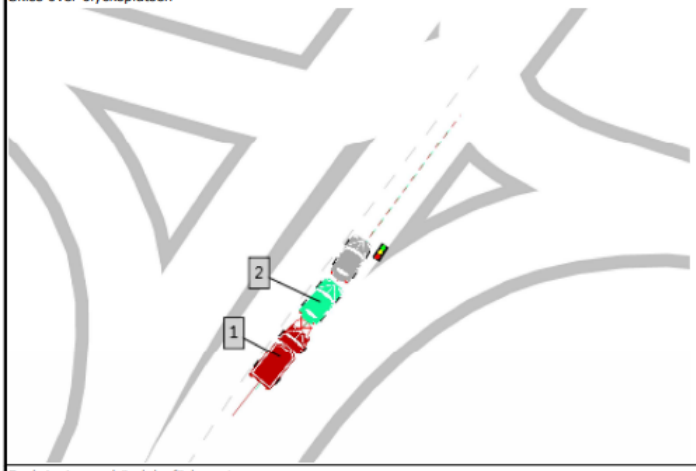
The results primarily confirm general assumptions made on circumstances of accident for the different vehicle types. But the model also gives some results which are worth closer examination. The multinomial logit can be used to model our data, although more or perhaps more relevant predictors are necessary to create a very well fitting model. Specifically predictors focused on technical data, such as *Brand*, *Year of manufacture*, *Tire condition* etc. would be very interesting to have access to. The data from police reports is not primarily intended to reflect differences in vehicle type for accidents; the purpose is instead very general with focus on general information. That said it still provides useful direction when studying vehicle type in traffic accidents, and can possibly serve as a test sample when fitting models on more comprehensive data.

8 Appendix

8.1 Police report template

Polisrapport Vägtrafikolycka		Olycks-ID 11	Polisens diarienummer K-2 /15
Län: Västra Götalands län	Kommun: Göteborg	Olyckstillfälle: 2015- 11:50 ☐ Osäker tid	Olyckstyp: U (upphinnande-motorfordon)
Olycksplats: Sankt Sigfridsgatan, trafikpl. Kallebäcksmotet.			Position i kartan: Säker position

Skiss över olycksplatsen



Beskrivning av händelseförloppet

Förare av pb 2 väntar på grönt ljus. Vid grönt ljus kör inte framförvarande pb. Pb 2 blir då påkörd bakifrån av Lb 1.

Väderleksförhållanden	Väglag	Bebyggelsetyp
Okänt	Okänt	Tättbebyggt område
Ljusförhållanden	**Platstyp**	**Attribut**
Dagsljus	Gatu-/Vägsträcka	
Vägnummer/Gatunamn	Väg A: Sankt Sigfridsgatan, 6	Väg B:
Högsta tillåtna hastighet	50 km/h	
Vägtyp	Annan allmän väg	
Trafikanvisning	Okänt	
Trafikreglering		
Trafiksignal	I funktion	
Gatu-/vägbelysning	Uppgift saknas	

Te nr	Trafikantkategori (antal personer totalt i fordonet)	Refnr	Ålder och kön	Förare	Passagerare			Personskada					Misstänkt påverkad	Övnings körning	Rapporterad av sjv.
					Fram	Bak	Okänt	Död	Svår	Lindrig	Oskadad	Okänd			
1	Lastbil (lätt) (1)	75	-K	X							X		O		N
2	Personbil (1)	75	-K	X						X			O		J

Figure 9: Template of official police report for traffic accident.

8.2 Insignificant values of predictor variables before merging

Vehicle type	Variable	Values (reference vs. non reference)	p-value
Cars/light trucks vs. Mc:s	Loc	Road section vs. Pavement	0.96
	Wear	Fair vs. Snowy rain	0.52
	Surf	Dry vs. Thin ice/visible	0.26
	Month	January vs. February	0.3
	Month	January vs. November	0.09
	Month	January vs. December	0.46
	Län	Stockholm vs. Blekinge	0.05
	Län	Stockholm vs. Dalarna	0.23
	Län	Stockholm vs. Gotland	0.14
	Year	2003 vs. 2004	0.36
	Year	2003 vs. 2005	0.74
	Year	2003 vs. 2006	0.17
	Year	2003 vs. 2007	0.41
	Year	2003 vs. 2009	0.37
	Year	2003 vs. 2010	0.27
	Year	2003 vs. 2015	0.53
	Year	2003 vs. 2016	0.91
Cars/light trucks vs. Heavy trucks/buses	Weekday	Monday vs. Tuesday	0.1
	Weekday	Monday vs. Wednesday	0.24
	Speed	50 km/h vs. 70 km/h	0.54
	Speed	50 km/h vs. 40 km/h	0.14
	Speed	50 km/h vs. 20 km/h	0.99
	Loc	Road section vs. crossing	0.7
	Loc	Road section vs. Bicycle road	0.83
	Län	Stockholm vs. Kalmar	0.55
	Län	Stockholm vs. Södermanland	0.07
	Län	Stockholm vs. Uppsala	0.8
	Län	Stockholm vs. Örebro	0.3
	Year	2003 vs. 2005	0.37
	Year	2003 vs. 2007	0.41
	Year	2003 vs. 2008	0.97
	Year	2003 vs. 2009	0.36
	Year	2003 vs. 2010	0.18
	Year	2003 vs. 2011	0.37
	Year	2003 vs. 2012	0.18
	Year	2003 vs. 2013	0.78
	Year	2003 vs. 2014	0.54
	Year	2003 vs. 2015	0.52
	Year	2003 vs. 2016	0.67

Table 23: Insignificant values of predictor variables before merging.

8.3 Merging of variables

Old variable	Old values	New variable	Merged values
Weekday	Monday	Weekday	Weekday
	...		Weekend
	Sunday		
Month	January	Season	Spring
	...		Summer
	December		Fall
			Winter
Län	Blekinge	Region	Big city
	...		Normal
	Östergötland		
Speed plate	20	Speed plate	50 km/h
	...		70 km/h
	120 km/h		90 km/h
			Other
Location	Pavement	Loc	Other
	Bicycle road		
Road type	Street	Road	Other
	Other road		
Weather	Rain	Weather	Rain
	Snowy rain		
Road surface	Wet/moist	Road surface	Wet
	Thin ice/visible		

Table 24: All mergers of predictor variable values from original dataset.

8.4 Univariate models for predictor variables

Variable	χ^2 p-value	AIC	$R^2_{McFadden}$	χ^2
$Year$	5.6387e-13	77778.98	0.0014634	113.91
Ses	$<$ 2.22e-16	73045.42	0.061763	4807.5
$Week$	$<$ 2.22e-16	77159.87	0.0088006	685.01
Reg	$<$ 2.22e-16	77601.55	0.0031262	243.34
$Speed$	$<$ 2.22e-16	77044.41	0.010387	808.48
Env	$<$ 2.22e-16	77354.64	0.0063497	494.24
$Road$	$<$ 2.22e-16	77005.18	0.010839	843.7
Loc	$<$ 2.22e-16	77502.07	0.004507	350.81
$Wear$	$<$ 2.22e-16	76664.88	0.015263	1188
$Surf$	$<$ 2.22e-16	74272.15	0.046003	3580.7
$Light$	$<$ 2.22e-16	76245.67	0.020597	1603.2
Acc	$<$ 2.22e-16	76199.18	0.021297	1657.7

Table 25: Results for fitting univariate models.

8.5 Insignificant values of predictor variables for main effects model

Vehicle type	Variable	Values (reference vs. non reference)	p-value
Cars/light trucks vs. Mc:s	Loc	Road section vs. crossing	0.26
	Wear	Fair vs. Rain	0.41
	Year	2003 vs. 2004	0.88
	Year	2003 vs. 2005	0.63
	Year	2003 vs. 2010	0.33
	Year	2003 vs. 2015	0.14
	Year	2003 vs. 2016	0.14
Cars/light trucks vs. Heavy trucks/buses	Acc	Single vs. Meeting	0.95
	Speed	50 km/h vs. 70 km/h	0.9
	Loc	Road section vs. crossing	0.7
	Road	Public road vs. street	0.13
	Road	Public road vs. express/motorway	0.11
	Ses	Spring vs. summer	0.13
	Year	2003 vs. 2005	0.16
	Year	2003 vs. 2006	0.36
	Year	2003 vs. 2007	0.98
	Year	2003 vs. 2008	0.39
	Year	2003 vs. 2010	0.8
	Year	2003 vs. 2013	0.28
	Year	2003 vs. 2015	0.63

Table 26: Results for fitting main effects model, only insignificant levels with p-value from Wald test.

8.6 Variables of Model B

Main effects	Interaction terms
Env	$Env * Acc$
Loc	$Acc * Loc$
$Road$	$Env * Loc$
$Speed$	$Env * Reg$
$Wear$	$Acc * Surf$
$Year$	$Acc * Speed$
$Week$	$Acc * Light$
Acc	$Light * Week$
Reg	$Acc * Week$
$Surf$	$Speed * Road$
Ses	$Speed * Light$
$Light$	$Env * Week$
	$Light * Reg$

Table 27: All main effects and interactions of Model B.

8.7 Significant coefficients Cars/light trucks vs. MC:s

Label	OR estimate	Wald CI
Env(Sparse pop. vs Dense pop.)	1.34	1.19- 1.54
Env(Sparse pop. vs Unknown)	1.98	1.51- 2.66
Env(Sparse pop. vs Dense pop.), Acc(Single vs Meeting)	0.249	0.45-0.92
Env(Sparse pop. vs Dense pop.), Acc(Single vs Overtaking)	0.673	0.47- 0.71
Env(Sparse pop. vs Unknown), Acc(Single vs Overtaking)	0.414	0.26-0.65
Env(Sparse pop. vs Dense pop.), Acc(Single vs Other)	0.675	0.54-0.84
Env(Sparse pop. vs Unknown), Acc(Single vs Other)	0.42	0.26- 0.68
Env(Sparse pop. vs Dense pop.), Week(Week vs Weekend)	0.695	0.618- 0.78
Env(Sparse pop. vs Unknown), Week(Week vs Weekend)	1.36	1.02- 1.8
Env(Sparse pop. vs Dense pop.), Loc(Road section vs Roundabout)	1.49	1.1-1.94
Env(Sparse pop. vs Unknown), Loc(Road section vs Roundabout)	2.04	1.13-3.26
Env(Sparse pop. vs Dense pop.), Loc(Road section vs Pavement/Other)	0.62	0.45- 0.84
Env(Sparse pop. vs Dense pop.), Reg(Big city vs Normal)	0.87	0.78- 0.97
Acc(Single vs Crossing/turn)	0.496	0.39- 0.63
Acc(Single vs Meeting)	0.38	0.25- 0.55
Acc(Single vs Overtaking)	0.25	0.195- 0.31
Acc(Single vs Other)	0.673	0.52- 0.87
Acc(Single vs Crossing/turn), Loc(Road section vs Roundabout)	0.629	0.44-0.85
Acc(Single vs Overtaking), Loc(Road section vs Roundabout)	0.499	0.35- 0.68
Acc(Single vs Other), Loc(Road section vs Roundabout)	0.32	0.19- 0.52
Acc(Single vs Crossing/turn), Loc(Road section vs Crossing)	0.5	0.41-0.62
Acc(Single vs Overtaking), Loc(Road section vs Crossing)	0.64	0.51- 0.81
Acc(Single vs Other), Loc(Road section vs Crossing)	0.52	0.42- 0.65
Acc(Single vs Crossing/turn), Loc(Road section vs Pavement/Other)	0.497	0.32- 0.76
Acc(Single vs Overtaking), Loc(Road section vs Pavement/Other)	0.44	0.265- 0.715
Acc(Single vs Other), Loc(Road section vs Pavement/Other)	0.37	0.24- 0.57
Acc(Single vs Crossing/turn), Speed(50 km/h vs 90 km/h)	2.65	1.96- 3.59
Acc(Single vs Overtaking), Speed(50 km/h vs 90 km/h)	2.34	1.66- 3.3
Acc(Single vs Other), Speed(50 km/h vs 90 km/h)	2.04	1.46- 2.8
Acc(Single vs Crossing/turn), Speed(50 km/h vs Other)	1.297	1.1-1.57
Acc(Single vs Meeting), Speed(50 km/h vs Other)	1.49	1.0- 2.23
Acc(Single vs Overtaking), Speed(50 km/h vs Other)	1.36	1.1- 1.72
Acc(Single vs Overtaking), Light(Daylight vs Darkness)	1.51	1.07-2.2
Acc(Single vs Other), Light(Daylight vs Darkness)	2.32	1.8- 3.0
Acc(Single vs Overtaking), Light(Daylight vs Sunset/Sundown)	1.42	1.1- 1.94
Acc(Single vs Other), Light(Daylight vs Sunset/Sundown)	1.48	1.14- 1.97
Acc(Single vs Meeting), Surf(Dry vs Wet)	0.43	0.26-0.72
Acc(Single vs Meeting), Surf(Dry vs Unknown)	0.497	0.28- 0.887
Acc(Single vs Meeting), Week(Week vs Weekend)	1.47	1.14- 1.94
Acc(Single vs Overtaking), Week(Week vs Weekend)	1.44	1.2- 1.74
Speed(50 km/h vs 90 km/h)	0.244	0.2- 0.297
Speed(50 km/h vs 90 km/h), Light(Daylight vs Darkness)	0.549	0.363-0.83
Speed(50 km/h vs 90 km/h), Light(Daylight vs Sunset/Sundown)	1.52	1.098- 2.14
Speed(50 km/h vs Other), Light(Daylight vs Sunset/Sundown)	1.51	1.2-1.9
Speed(50 km/h vs Other), Road(Public road vs Street)	1.67	1.4- 1.9
Speed(50 km/h vs 70 km/h), Road(Public road vs Express/Motorway)	0.49	0.35- 0.7
Speed(50 km/h vs Other), Road(Public road vs Express/Motorway)	0.33	0.24- 0.47
Weekday(Week vs Weekend)	1.85	1.67-2.02
Weekday(Week vs Weekend), Light(Daylight vs Darkness)	0.524	0.43-0.648
Weekday(Week vs Weekend), Light(Daylight vs Sunset/Sundown)	0.512	0.42- 0.634
Reg(Big city vs Normal)	0.81	0.68- 1.01
Reg(Big city vs Normal), Light(Daylight vs Sunset/Sundown)	0.71	0.587- 0.858

Table 28: Odds ratios with 95

8.8 Significant coefficients Cars/light trucks vs. Heavy trucks/buses

Label	OR estimate	Wald CI
Env(Sparse pop. vs Dense pop.)	0.724	0.576- 0.89
Env(Sparse pop. vs Dense pop.), Acc(Single vs Meeting)	1.12	0.76- 1.7
Env(Sparse pop. vs Dense pop.), Acc(Single vs Overtaking)	0.78	0.578- 1.1
Env(Sparse pop. vs Unknown), Acc(Single vs Overtaking)	0.76	0.394- 1.41
Env(Sparse pop. vs Dense pop.), Acc(Single vs Other)	1.23	0.89-1.76
Env(Sparse pop. vs Unknown), Acc(Single vs Other)	0.78	0.4-1.53
Env(Sparse pop. vs Dense pop.), Week(Week vs Weekend)	0.97	0.72- 1.3
Env(Sparse pop. vs Unknown), Week(Week vs Weekend)	0.99	0.39- 2.0
Env(Sparse pop. vs Dense pop.), Loc(Road section vs Crossing)	1.42	0.4-1.88
Env(Sparse pop. vs Dense pop.), Reg(Big city vs Normal)	0.715	0.599- 0.86
Acc(Single vs Crossing/turn)	0.274	0.175-0.42
Acc(Single vs Meeting)	0.5	0.3- 0.79
Acc(Single vs Overtaking)	0.23	0.16- 0.34
Acc(Single vs Other)	0.57	0.37- 0.85
Acc(Single vs Crossing/turn), Loc(Road section vs Roundabout)	0.43	0.2-0.877
Acc(Single vs Overtaking), Loc(Road section vs Roundabout)	0.28	0.135- 0.55
Acc(Single vs Other), Loc(Road section vs Roundabout)	0.28	0.11-0.73
Acc(Single vs Crossing/turn), Loc(Road section vs Crossing)	0.58	0.4-0.84
Acc(Single vs Meeting), Loc(Road section vs Crossing)	0.57	0.32-0.999
Acc(Single vs Other), Loc(Road section vs Crossing)	0.58	0.4-0.858
Acc(Single vs Crossing/turn), Speed(50 km/h vs 70 km/h)	1.8	1.26- 2.6
Acc(Single vs Crossing/turn), Speed(50 km/h vs 90 km/h)	1.85	1.2-2.89
Acc(Single vs Other), Speed(50 km/h vs 90 km/h)	0.71	0.45- 1.17
Acc(Single vs Meeting), Speed(50 km/h vs 90 km/h)	2.0	1.24- 3.3
Acc(Single vs Meeting), Speed(50 km/h vs Other)	1.75	1.1- 2.86
Acc(Single vs Overtaking), Speed(50 km/h vs Other)	2.04	1.42- 3.07
Acc(Single vs Overtaking), Light(Daylight vs Darkness)	1.36	1.0-1.84
Acc(Single vs Crossing/turn), Surf(Dry vs Ice/Snow)	1.72	1.14- 2.6
Acc(Single vs Overtaking), Surf(Dry vs Ice/Snow)	1.75	1.1- 2.85
Acc(Single vs Crossing/turn), Surf(Dry vs Wet)	1.43	1.1 - 1.87
Acc(Single vs Meeting), Surf(Dry vs Wet)	1.54	1.16- 2.02
Acc(Single vs Crossing/turn), Surf(Dry vs Unknown)	1.58	1.03- 2.36
Acc(Single vs Crossing/turn), Week(Week vs Weekend)	1.73	1.19- 2.5
Acc(Single vs Meeting), Week(Week vs Weekend)	1.64	1.15- 2.35
Acc(Single vs Overtaking), Week(Week vs Weekend)	1.51	1.02- 2.26
Speed(50 km/h vs 90 km/h), Light(Daylight vs Darkness)	2.32	1.69-3.23
Speed(50 km/h vs Other), Light(Daylight vs Darkness)	2.31	1.72-3.16
Speed(50 km/h vs Other), Light(Daylight vs Sunset/Sundown)	1.51	0.69-1.51
Week(Week vs Weekend)	0.24	0.194- 0.29
Week(Week vs Weekend), Light(Daylight vs Darkness)	1.32	1.0- 1.72
Reg(Big city vs Normal)	1.11	0.99- 1.23
Reg(Big city vs Normal), Light(Daylight vs Darkness)	1.3	1.08- 1.58

Table 29: Odds ratios with 95

8.9 Plots & classification tables

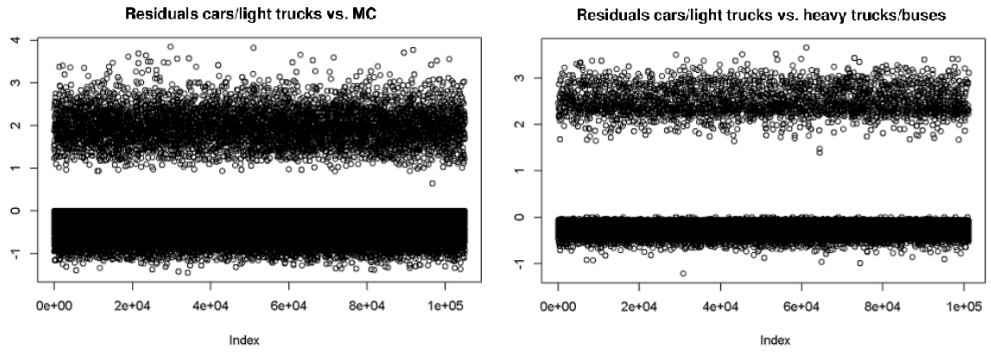


Figure 10: Deviance residuals for response levels, Model A

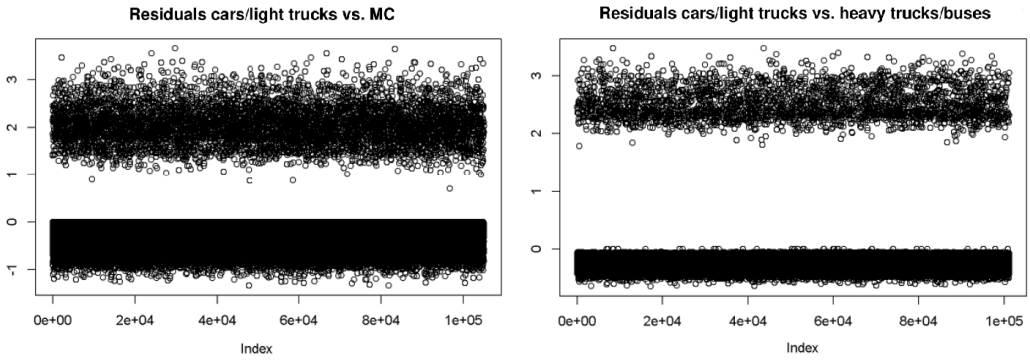


Figure 11: Deviance residuals for response levels, Model C

	Car/Light truck	MC	Heavy truck/Bus
Car/Light truck	98106	6712	3009
MC	69	83	2
Heavy truck/Bus	0	0	0

Table 30: Classification table, Model A on training data.

	Car/Light truck	MC	Heavy truck/Bus
Car/Light truck	98139	6758	3008
MC	36	37	3
Heavy truck/Bus	0	0	0

Table 31: Classification table, Model C on training data.

9 References

Agresti, A. (2002): Categorical Data Analysis. Wiley Series in Probability and Statistics

Allison, P.D. (2002): Missing data. Thousand Oaks: SAGE Publications, Inc. doi: http://dx.doi.org/10.4135/9781412985079

Begg, C.B., Gray, R. (1984): Calculation of polychotomous logistic regression parameters using individualized regressions, Biometrika, vol. 71, pp. 11-18

Cheng, S., Long, J.S. (May 2007): Testing for IIA in the Multinomial Logit Model, Sociological Methods & Research, Volume 35, Number 4, SAGE Publications

Dobson, A.J. (2002): An introduction to generalized linear models, 2nd ed, Chapman & Hall/CRC

Eliason, S.R. (1993): Maximum likelihood estimation, logic and practice, SAGE Publications

Elvik, R. (July 2006): Laws of accident causation, Accident Analysis & Prevention, Volume 38, Issue 4, pp. 742-747

Fagerland, M.W., Hosmer, D.W. (2012): A generalized Hosmer–Lemeshow goodness-of-fit test for multinomial logistic regression models, The Stata Journal, Volume 12, Number 3, pp. 447-453

Haddon, W. Jr. (1970). On the escape of tigers: an ecologic note, Am J Public Health Nations Health, doi: www.ncbi.nlm.nih.gov/pmc/articles/PMC1349282

Hastie, T., Tibshirani, R.,Friedman, J. (2009): The Elements of Statistical Learning, Second edition, Springer

Hauck, W.W., Donner, A. (December 1977): Wald's Test as Applied to Hypotheses in Logit Analysis, Journal of the American Statistical Association, Volume 72, Number 360

Hensher,D., Stopher, P. (1979): Behavioural Travel Modelling.

Held, L., Bové, D.S. (2014): Applied Statistical Inference

Hosmer, D.W., Lemeshow, S., Sturdivant, R.X. (2013): Applied Logistic Regression. Wiley Series in Probability and Statistics

James, G., Witten, D., Hastie,T. , Tibshirani, R. (2013): An Introduction to Statistical Learning with Applications in R, Springer

Larose, D.T. (2015): Data Mining and Predictive Analytics, Wiley

Liebetrau, A.M. (1983): Measures of Association, SAGE Publications

Lindberg, J., Strandroth, J., Ekman, L., Persson, S., Malmström, T. (Dec 2016): Översyn av etappmål för säkerhet på väg till 2020 och 2030, med en utblick mot 2050.

Long, J.S. (1997): Regression models for categorical and limited dependant variables Thousand Oaks, SAGE Publications

Norton, E.C., Wang, H., Ai, C. (2004): Computing interaction effects and standard errors in logit and probit models, The Stata Journal, Volume 4, Number 2, pp. 154–167

Trafikanalys, (2018): Vägtrafikskador 2017. [online] Available at: https://www.trafa.se/globalassets/statistik/vagtrafik/vagtrafikskador/2017/vagtrafikskador-2017-blad.pdf [Accessed 12 May. 2018]

Trafikverket, (2017): Analys av trafiksäkerhetsutvecklingen 2016. [online] Available at: https://trafikverket.ineko.se/Files/en-US/25320/Ineko.Product.RelatedFiles/2017_098_analys_av_trafiksakerhetsutvecklingen_2016.pdf [Accessed 6 May. 2018]

Tutz,G. (2011): Regression for categorical data, Cambridge University Press

Regerinskansliet, Näringsdepartmentet, (2016): Nystart för Nollvisionen - ett intensifierat arbete för trafiksäkerheten i Sverige. [online] Available at: http://www.regeringen.se/4a509c/contentassets/00c9b57223d74e1fa0fe4da50e1e4e83/trafiksakerhet_160905_webb.pdf [Accessed 6 May. 2018]

Svenska Kommunförbundet (1999): Olycksboken, Kommentus Förlag