Delft University of Technology

Faculty of Electrical Engineering, Mathematics and Computer Science

A comparison of the frequentist and Bayesian approach to multinomial logistic regression in statistics:

an application to study habits data from PRIME

Submitted in partial fulfilment
of the requirements for the degree of
Bachelor of Science in Technische Wiskunde

Sterre Hart Schriemer (5148626)

Thesis committee:
Dr. A.J. Cabo (supervisor)
Dr. T.W.C. Vroegrijk

January 18, 2023

Abstract

Frequentist statistics and Bayesian statistics are the two main approaches to statistical inference. The frequentist approach is commonly integrated into academic curricula, while the Bayesian approach is less frequently employed. However a comparison of the approaches, further investigating their shortcomings and advantages, might give a better comprehension of statistics and more insight in statistical inference. Therefore the current study applied both the frequentist and Bayesian approach to multinomial logistic regression.

The multinomial logistic regression model can be described as a generalized linear model and as a random utility model, and the current study has shown that these models generate an equivalent probability function. Moreover, the method of estimating coefficients in the frequentist and in the Bayesian approach were described. The multinomial logistic regression model was subsequently applied to data from educational research, conducted by PRIME. Three different R packages were used to perform the multinomial logistic regression: the VGAM package (frequentist, generalized linear model), the mlogit package (frequentist, random utility model) and the UPG package (Bayesian, random utility model). The results of the analysis of one dependent variable were subsequently compared.

The results indicated that the frequentist and Bayesian approach differ in their estimation time and model fit: the Bayesian approach required more computational time, but resulted in a better model fit. The frequentist 95% confidence intervals and Bayesian 95% credible intervals are comparable, but the interpretation of these is considerably different due to the philosophical underpinnings of both approaches. Moreover in the Bayesian approach, existing knowledge and information can be incorporated by choosing the prior distribution. Furthermore the Bayesian approach gives a posterior distribution, which is more informative than only a point estimate. In comparing the three different R packages it is noted that all three have a slightly different theoretical background. Since the packages all have their own shortcomings and advantages, combining them when conducting multinomial logistic regression could be desirable.

Contents

Chapter 1

Introduction

Frequentist statistics and Bayesian statistics are the two main approaches to statistical inference. The key difference between the two approaches is their interpretation of the concept *probability.* In frequentist statistics, probability is defined in connection to countable events, being the limit of the relative frequency of an event after many trials [McElreath, 2020]. These frequencies can be calculated by conducting an experiment. For example if one wants to investigate whether a coin is weighted, you can throw the coin 100 times. If the coin turns heads 80 times out of 100, then a frequentist calculates the probability of such a result occuring if the coin is unweighted.

On the other hand, in Bayesian statistics, probability expresses a degree of belief in an event [McElreath, 2020]. Hence the probability actually expresses the chance of an event happening. In doing so, Bayesian statistics includes prior information, for example whether you have seen something weird about the coin. This allows researchers to incorporate existing knowledge or information when making statistical inferences about the data. Another important difference is that Bayesians view the unknown parameters as random variables, while frequentists see these parameters as fixed values which can be determined by repeated sampling.

Both approaches can be used to conduct statistical analyses. Throughout history these has been a debate surrounding the use of Bayesian or frequentist statistics. In 1925, R.A. Fisher, a statistician who can be considered frequentist and who had much influence on the development of statistics as a field, even announced that "The theory of inverse probability [which is used in Bayesian statistics] is founded upon an error and must be wholly rejected" [Fisher, 1925]. The frequentist approach has mainly been developed in the early 20th century [McElreath, 2020] and Fisher wrote two books that were very influential in these times: *Statistical Methods for Research Workers* (1925) and *The Design of Experiments* (1935). These books led to an enormous increase in the popularity of the use of frequentist statistics and concepts like the $p-$value [Lehmann, 2011]. His influence amongst other factors led to the frequentist approach becoming the dominant statistical paradigm in scientific literature [McElreath, 2020].

The Bayesian approach to statistical inference is much older: versions of it were already used in the 18th and 19th century [McElreath, 2020]. But it was successfully marginalized in the beginning of the 20th century, mostly due to the philosophical debate surrounding statistics. Moreover, the computations that needed to be performed in Bayesian statistics were so difficult that it was extremely challenging or even impossible to do these. However, due to increased use of computers, beginning in the 1990s, the application of Bayesian statistics has grown rapidly [McElreath, 2020]. Moreover, since statisticians recognized shortcomings in frequentist statistics, it has been advocated as an alternative approach of statistics.

However, Bayesian statistics is not commonly integrated into the curriculum of most studies yet: often students first learn to apply frequentist statistics. Furthermore, when scholars are doing statistical analyses, they often apply either the Bayesian or the frequentist approach in their studies, but not both. However, comparing both approaches, looking at their shortcomings and advantages, might give a better comprehension of statistics and more insight in the problem that is currently investigated. As McElreath (2020) mentions in his book *Rethinking statistics*: "Changing the representation of a problem often makes it easier to address or inspire new ideas that were not available in the old representation. (...) Switching teaches us new things about both approaches." Therefore in the current study both a frequentist and Bayesian approach is used to perform multinomial logistic regression. Subsequently the results of these approaches are compared.

1.1 Multinomial logistic regression

Multinomial logistic regression is used to predict the probability of the different possible outcomes of a categorical dependent variable. For example, in countries where there are multiple political parties, such as the Netherlands, logistic regression can be used to predict which of the parties a person is likely to vote for. Several independent variables can be used to make these predictions. These variables can be either continuous or categorical. For example, voting behavior can be predicted by looking at the independent variables age (continuous), gender (categorical) and income (continuous).

Multinomial logistic regression is used in a lot of fields, for example in health science, marketing research and social science. In the current study we are looking at study habits and beliefs and whether we can predict those by looking at grade goals, prior math grade and self-efficacy. Since these study habits and beliefs are categorical dependent variables, insight in this can be provided by multinomial logistic regression.

1.2 Overview of the current study

The main aim of this study is to compare the frequentist and Bayesian approach to multinomial logistic regression. Another aim of this study is to compare several packages in R for doing multinomial logistic regressions, such that it can serve as a guide for researchers or students who need to perform these. Thus it has two research aims:

1. Comparison of frequentist and Bayesian approach in performing multinomial logistic regression
2. Comparison of three different R packages to perform multinomial logistic regression

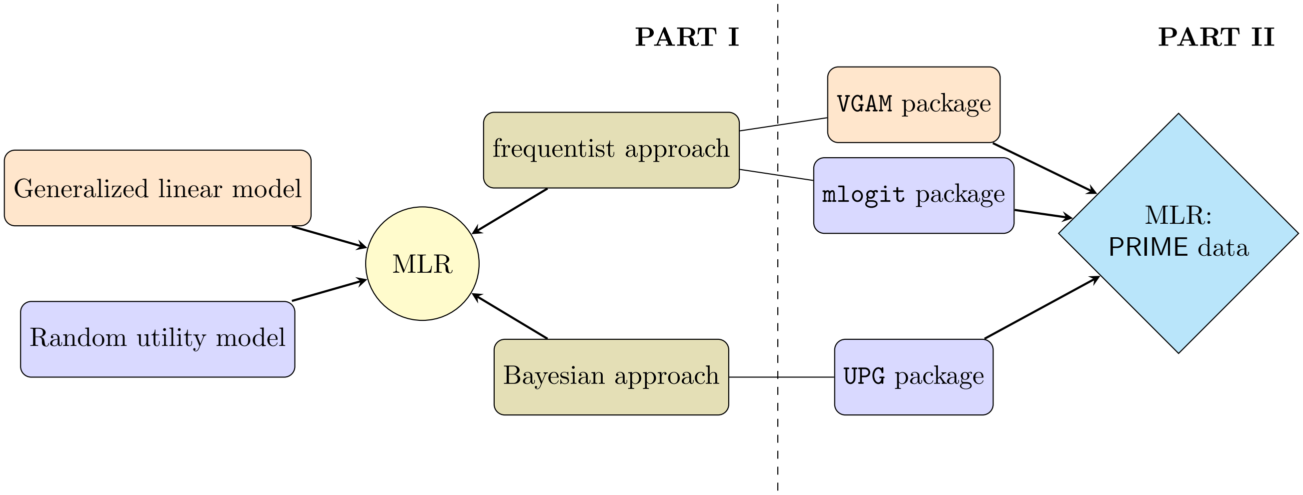

Figure 1.1: Overview of the current study: Part I (theoretical background) and Part II (application to PRIME data.

In order to achieve these aims, the study is split into two parts, which are depicted in the overview in figure 1.1. In part I of the study, the theoretical background of multinomial logistic regression is presented. Firstly in chapter 2 the multinomial logistic regression model is described as a generalized linear model and as a random utility model and in section 2.3 it is shown that these models both generate the same probability function for MLR. The assumptions used in both the derivation of the probability function and the estimation of the coefficients are outlined in section 2.4. Moreover, in chapter 3 the method of estimating the coefficients is outlined for both the frequentist and Bayesian approach, along with some more information about important concepts in both approaches. Statistical guides often outline the interpretation of these coefficients, but do not really explain what happens in the 'black box', which is the statistical program. Therefore, chapter 3 also aims to shed some light on the numerical methods that are used in the statistical programs.

In the second part of the study, multinomial logistic regression is applied to a data set from PRIME. This data set is described in chapter 4. Three different R packages are utilized to perform the multinomial logistic regression and in chapter 5 the results of these are given. These results are subsequently compared in chapter 6, where both attention is given to the differences between the frequentist and Bayesian approach to multinomial logistic regression and also to the differences between the different R packages. Lastly in chapter 7 the findings are summarized, followed by a discussion of the limitations and suggestions for future research.

Chapter 2

Multinomial logistic regression

The model underlying multinomial logistic regression can be described in several equivalent ways. In the literature these models are often intertwined and no clear distinctions are made. This is not very problematic, since it can be shown that the probability function that is generated by both models is equivalent. However, since they both have different backgrounds, understanding them provides a better understanding of the multinomial logistic regression. Hence in this section, two models underlying multinomial logistic regression are presented: the generalized linear model (GLM) and random utility model (RUM). Moreover it will be shown that they generate an equivalent probability function. Lastly the assumptions are outlined in section 2.4, where a distinction is made between the assumptions used in the derivation of the probability function and those used in the estimation of the coefficients.

2.1 Generalized linear model (GLM)

A first method to describe the multinomial logistic regression is as a generalized linear model (GLM). The GLM was proposed in 1972 by Nelder and Wedderburn who provided a unified framework for several regression models [Fox, 2016], such as linear regression and logistic regression. According to Fox (2016), their paper is the *"the "bible" of GLMs, a rich and interesting—if generally difficult—text."* In the original paper of Nelder and Wedderburn, four different distributions of the dependent variable are considered: the normal distribution, binomial distribution, Poisson distribution and gamma distrbution [Nelder and Wedderburn, 1972]. Further work has included other distributions into the generalized linear model, such as the multinomial distribution [Fox, 2016].

2.1.1 General formulation of the model

Suppose we have k independent variables and n different observations. The GLM has three important components that are necessary to estimate the dependent variable [Fox, 2016]:

1. Firstly a **random component**, which specifies the conditional distribution of the dependent variable Y_i (given the values of the k independent variables X_{ik} in the model). The most common distributions used in statistical modelling are members of the exponential family, such as the gamma, normal, Poisson, binomial and multinomial distribution.
2. Moreover a **linear predictor**, which is a linear function of k independent variables and unkown coefficients $\beta_0, \ldots, \beta_k$. This gives the linear predictor $\eta_i = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_k X_{ik}$, for observation $i \in \{1, \ldots, n\}$.
3. Lastly a linearizing **link function** g that specifies the link between the random component and the linear predictor. The link function is smooth and invertible and relates the expectation of the dependent variable, $\pi_i = \mathbb{E}(Y_i)$, to the linear predictor, giving:
$$g(\pi_i) = \eta_i = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_k X_{ik}$$

The coefficients $\beta_0, \ldots, \beta_k$ subsequently need to be estimated.

2.1.2 Derivation of the probability function

The three components of the GLM can be used to describe the multinomial logistic regression model. In this multinomial logistic regression model it is assumed that there are k independent variables, $X_1, \dots, X_k$ and m different categories of the dependent variable Y, which we number $1, 2, \dots, m$. Lastly, n observations are included in the model. Multinomial logistic regression is then described with the following three components:

Firstly the **random component** is chosen as the multinomial distribution, since the dependent variable in multinomial logistic regression has several possible categories.

Secondly the **linear predictor** is defined as $\eta_{ij} = \beta_{0j} + \beta_{1j}X_{i1} + \cdots + \beta_{kj}X_{ik}$, for observation $i \in \{1, \dots, n\}$ and category $j \in \{1, \dots, m\}$.

Lastly the logit function is chosen as the **link function**. Let π_{ij} denote the probability that the ith observation falls into the jth category of the dependent variable. The logit function for binomial logistic regression is defined as the logarithm of the probability of an event happening, divided by the probability of an event not happening:

$$g(\pi_i) = \log \frac{\pi_i}{1 - \pi_i} = \eta_i = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_k X_{ik}$$

(Remark: with two categories of the dependent variable, therefore no subscript j)

This logit function is also the link function in multinomial logistic regression. Suppose category m is chosen as the baseline category. Subsequently, independent binary logistic regressions will be performed for the remaining $m - 1$ categories [Firth, 1991]. The multinomial logit function is then defined as $\boldsymbol{g(\pi_i)} = (g_1(\pi_i), \dots, g_{m-1}(\pi_i))$ [Firth, 1991], with for $j = 1, \dots, m - 1$:

$$g_j(\pi_i) = \log \frac{\pi_{ij}}{\pi_{im}} = \eta_{ij} = \beta_{0j} + \beta_{1j}X_{i1} + \cdots + \beta_{kj}X_{ik} \tag{2.1}$$

The choice of the baseline category is arbitrary.

Lemma 2.1 (GLM) The probability that observation i falls into category j is equal to:

$$\pi_{ij} = \frac{e^{\beta_{0j}+\beta_{1j}X_{i1}+\cdots+\beta_{kj}X_{ik}}}{1 + \sum_{l=1}^{m-1} e^{\beta_{0l}+\beta_{1l}X_{i1}+\cdots+\beta_{kl}X_{ik}}} \qquad \text{for } j = 1 \dots m - 1 \tag{2.2}$$

$$\pi_{im} = \frac{1}{1 + \sum_{l=1}^{m-1} e^{\beta_{0l}+\beta_{1l}X_{i1}+\cdots+\beta_{kl}X_{ik}}} \qquad \text{for } j = m$$

Proof. Firstly exponentiate both sides of the link function (equation 2.1) giving for $j = 1, \dots, m - 1$

$$\frac{\pi_{ij}}{\pi_{im}} = e^{\beta_{0j}+\beta_{1j}X_{i1}+\cdots+\beta_{kj}X_{ik}} \tag{2.3}$$

$$\pi_{ij} = \pi_{im} e^{\beta_{0j}+\beta_{1j}X_{i1}+\cdots+\beta_{kj}X_{ik}} \tag{2.4}$$

The m probabilities must sum to 1, hence

$$\pi_{im} + \sum_{l=1}^{m-1} \pi_{il} = 1$$

$$\pi_{im} + \sum_{l=1}^{m-1} \pi_{im} e^{\beta_{0l}+\beta_{1l}X_{i1}+\cdots+\beta_{kl}X_{ik}} = 1$$

$$\pi_{im}(1 + \sum_{l=1}^{m-1} e^{\beta_{0j}+\beta_{1l}X_{i1}+\cdots+\beta_{kl}X_{ik}}) = 1$$

$$\pi_{im} = \frac{1}{1 + \sum_{l=1}^{m-1} e^{\beta_{0l}+\beta_{1l}X_{i1}+\cdots+\beta_{kl}X_{ik}}} \tag{2.5}$$

And for $j = 1, \dots, m-1$:

$$\begin{aligned}\pi_{ij} &= \pi_{im} e^{\beta_{0j}+\beta_{1j}X_{i1}+\cdots+\beta_{kj}X_{ik}} \\ &= \frac{e^{\beta_{0j}+\beta_{1j}X_{i1}+\cdots+\beta_{kj}X_{ik}}}{1+\sum_{l=1}^{m-1} e^{\beta_{0l}+\beta_{1l}X_{i1}+\cdots+\beta_{kl}X_{ik}}}\end{aligned} \tag{2.6}$$

The last category m acts as a type of baseline, due to which the probabilities of the dependent variable categories for each observation sum to 1. The baseline category has coefficients $\beta_{0m} = \beta_{1m} \dots \beta_{km} = 0$. □

2.2 Random utility model (RUM)

In the current section the random utility model will be outlined. The economist Daniel McFadden (1974) contributed largely to the theory of random utility models: in 2000 he even received a Nobel Prize for his work. In economics, utility refers to the satisfaction or happiness that a consumer derives from consuming a good or service. It is a measure of the usefulness or value that a consumer places on a particular choice. The aim of random utility models is to model the probability of discrete choices, by looking at the utility of the different alternatives. In the model we have a decision maker, labeled $i \in \{1, \dots, n\}$, who faces m different alternatives. For each alternative $j \in \{1, \dots, m\}$, there is a utility U_{ij}, which consists of two parts: V_{ij}, which is the part that is known by the researcher, and an unobservable part ε_{ij}, which is treated by the researcher as random. Hence we have $\forall i, j : U_{ij} = V_{ij} + \varepsilon_{ij}$. The decision maker chooses the alternative with the highest level of utility, hence chooses alternative j if and only if $U_{ij} > U_{il}$ $\forall l \neq j$. The level of utility for one alternative does not matter from the perspective of the researcher, only the difference between utility levels for different alternatives matters. Even though the model is derived with this utility maximization, it can be used to simply model decision-making by describing the relation of explanatory variables to the outcome of a choice [Train, 2009].

2.2.1 Derivation of the model

The random utility model for multinomial logistic regression will be derived in this section. As for the generalized linear model, it is assumed that there are k independent variables, $X_1, \dots, X_k$ and m different categories of the dependent variable Y, which we number $1, 2, \dots, m$. Lastly, n decision makers (or observations) are included in the model. Suppose we have a decision maker i, who faces m different alternatives. For the decision maker i the level of utility is defined for each alternative $j \in \{1, \dots, m\}$:

$$\begin{cases} U_{i1} &= V_{i1} + \varepsilon_{i1} \\ U_{i2} &= V_{i2} + \varepsilon_{i2} \\ & \vdots \\ U_{im} &= V_{im} + \varepsilon_{im} \end{cases}$$

Where

$$V_{ij} = \beta_{0j} + \beta_{1j}X_{i1} + \cdots + \beta_{kj}X_{ik} \tag{2.7}$$

For the sake of simplicity, we use V_{ij} in the derivation but it is important to note that just as in GLM, V_{ij} is a linear function of the different observed variables (see equation 2.7). The decision maker i chooses the alternative with the highest level of utility, hence chooses alternative j if and only if $\forall j \neq l : \quad U_{ij} > U_{il}$ The probability that decision maker i chooses alternative j is again denoted by π_{ij} and thus is:

$$\begin{aligned}\pi_{ij} &= \mathbb{P}(U_{ij} > U_{i1}, U_{ij} > U_{i2} \dots, U_{ij} > U_{il}) \qquad (2.8) \\ &= \mathbb{P}(U_{ij} > U_{il} \quad \forall l \neq j) \\ &= \mathbb{P}(V_{ij} + \varepsilon_{ij} > V_{il} + \varepsilon_{il} \quad \forall l \neq j) \\ &= \mathbb{P}(\varepsilon_{il} < \varepsilon_{ij} + V_{ij} - V_{il} \quad \forall l \neq j)\end{aligned}$$

Now suppose that ε_{ij} is given. Then assumption 3 from section 2.4 can be used, which states that the expression of the cumulative distribution of each ε_{il} evaluated at $\varepsilon_{ij} + V_{ij} - V_{il}$ is given by $e^{-e^{-(\varepsilon_{ij}+V_{ij}-V_{il})}} \quad \forall j = 1, \dots, m$. Moreover it is assumed in section 2.4 that the ε's are independent, hence the cumulative distribution over all $l \neq j$ is the product of the marginal cumulative distributions:

$$\mathbb{P}(\varepsilon_{il} < \varepsilon_{ij} + V_{ij} - V_{il} \quad \forall l \neq j | \epsilon_{ij}) = \prod_{l \neq j} e^{-e^{-(\varepsilon_{ij}+V_{ij}-V_{il})}} \tag{2.9}$$

However ε_{ij} is not given, so in order to compute the choice probability we have to take the integral of equation 2.9 over all values of ε_{ij}, weighted by its density [Train, 2009] giving:

$$\pi_{ij} = \int (\prod_{l \neq j} e^{-e^{-(\varepsilon_{ij}+V_{ij}-V_{il})}}) e^{-\varepsilon_{ij}} e^{-e^{-\varepsilon_{ij}}} d\varepsilon_{ij}$$

.

Lemma 2.2 (RUM) If we assume that the unobserved component ε_{ij} is identically and independently distributed for each alternative, with cumulative distribution $F(\varepsilon_{ij}) = e^{-e^{-\varepsilon_{ij}}}$ and density $f(\varepsilon_{ij}) = e^{-\varepsilon_{ij}} e^{-e^{-\varepsilon_{ij}}}$, then we have that the probability that observation i falls into category j is equal to:

$$\pi_{ij} = \frac{e^{V_{ij}}}{\sum_l e^{V_{il}}} \tag{2.10}$$

Proof.

$$\begin{aligned}
\pi_{ij} &= \int_{\varepsilon_{ij}=-\infty}^{\infty} (\prod_{l \neq j} e^{-e^{-(\varepsilon_{ij}+V_{ij}-V_{il})}}) e^{-\varepsilon_{ij}} e^{-e^{-\varepsilon_{ij}}} d\varepsilon_{ij} && \text{note:} \quad V_{ij} - V_{ij} = 0 \\
&= \int_{\varepsilon_{ij}=-\infty}^{\infty} (\prod_{l} e^{-e^{-(\varepsilon_{ij}+V_{ij}-V_{il})}}) e^{-\varepsilon_{ij}} d\varepsilon_{ij} && \text{sum terms exponent} \\
&= \int_{\varepsilon_{ij}=-\infty}^{\infty} (e^{-(\sum_l e^{-(\varepsilon_{ij}+V_{ij}-V_{il})}}) e^{-\varepsilon_{ij}} d\varepsilon_{ij} && \text{set } t = e^{-\varepsilon_{ij}} \text{ such that } dt = -e^{-\varepsilon_{ij}} d\varepsilon_{ij} \\
&= \int_{\infty}^{0} e^{-t \sum_l e^{-(V_{ij}-V_{il})}} (-dt) \\
&= \int_{0}^{\infty} e^{-t \sum_l e^{-(V_{ij}-V_{il})}} dt \\
&= \frac{e^{-t \sum_l e^{-(V_{ij}-V_{il})}}}{-\sum_l e^{-(V_{ij}-V_{il}}} \Bigg|_0^\infty \\
&= \frac{1}{\sum_l e^{-(V_{ij}-V_{il})}} \\
&= \frac{e^{V_{ij}}}{\sum_l e^{V_{il}}} && (2.11)
\end{aligned}$$

The two characteristics of probabilities are satisfied: $0 \leq \pi_{ij} \leq 1 \quad \forall i = 1, 2, 3, \dots$ and $\sum \pi_{ij} = 1$. □

2.3 Equivalence of generalized linear model and random utility model

Proposition 2.3 (Equivalence GLM and RUM) The probability functions for multinomial logistic regression derived by the generalized linear model and random utility model are equivalent.

Proof. As proven in lemma 2.1 for GLM the probability function is given by:

$$\pi_{ij} = \frac{e^{\beta_{0j}+\beta_{1j}X_{i1}+\cdots+\beta_{kj}X_{ik}}}{1+\sum_{l=1}^{m-1} e^{\beta_{0l}+\beta_{1l}X_{i1}+\cdots+\beta_{kl}X_{ik}}} \qquad \text{for } j = 1,\ldots,m \tag{2.12}$$

$$\pi_{im} = \frac{1}{1+\sum_{l=1}^{m-1} e^{\beta_{0l}+\beta_{1l}X_{i1}+\cdots+\beta_{kl}X_{ik}}} \qquad \text{for } j = m \tag{2.13}$$

We will show that equations 2.12 and 2.13 can be derived from the probability function of MLR derived by RUM. As proven in lemma 2.2 for RUM the probability function is given by:

$$\begin{aligned} \pi_{ij} &= \frac{e^{V_{ij}}}{\sum_{l=1}^{m} e^{V_{il}}} && \text{for } j = 1,\ldots,m \\ &= \frac{e^{\beta_{0j}+\beta_{1j}X_{i1}+\cdots+\beta_{kj}X_{ik}}}{\sum_{l=1}^{m} e^{\beta_{0l}+\beta_{1l}X_{i1}+\cdots+\beta_{kl}X_{ik}}} && \text{Plug in } V_{ij} = \beta_{0j}+\beta_{1j}X_{i1}+\cdots+\beta_{kj}X_{ik} \text{ (equation 2.7)} \end{aligned} \tag{2.14}$$

If we take equation 2.14 and choose category m as the baseline category we can show that equation 2.14 is equivalent to equations 2.12 and 2.13. This means that we set $\beta_{0m} = \beta_{1m} = \cdots = \beta_{km} = 0$. Hence for $j = m$:

$$\begin{aligned} \pi_{im} &= \frac{e^{\beta_{0m}+\beta_{1m}X_{i1}+\cdots+\beta_{km}X_{ik}}}{\sum_{l=1}^{m} e^{\beta_{0l}+\beta_{1l}X_{i1}+\cdots+\beta_{kl}X_{ik}}} \\ &= \frac{e^0}{e^0+\sum_{l=1}^{m-1} e^{\beta_{0l}+\beta_{1l}X_{i1}+\cdots+\beta_{kl}X_{ik}}} \\ &= \frac{1}{1+\sum_{l=1}^{m-1} e^{\beta_{0l}+\beta_{1l}X_{i1}+\cdots+\beta_{kl}X_{ik}}} \end{aligned}$$

And for $j = 1,\ldots,m-1$, we similarily have:

$$\begin{aligned} \pi_{ij} &= \frac{e^{\beta_{0j}+\beta_{1j}X_{i1}+\cdots+\beta_{kj}X_{ik}}}{\sum_{l=1}^{m} e^{\beta_{0l}+\beta_{1l}X_{i1}+\cdots+\beta_{kl}X_{ik}}} \\ &= \frac{e^{\beta_{0j}+\beta_{1j}X_{i1}+\cdots+\beta_{kj}X_{ik}}}{e^0+\sum_{l=1}^{m-1} e^{\beta_{0l}+\beta_{1l}X_{i1}+\cdots+\beta_{kl}X_{ik}}} \\ &= \frac{e^{\beta_{0j}+\beta_{1j}X_{i1}+\cdots+\beta_{kj}X_{ik}}}{1+\sum_{l=1}^{m-1} e^{\beta_{0l}+\beta_{1l}X_{i1}+\cdots+\beta_{kl}X_{ik}}} \end{aligned}$$

□

Hence the random utility model and the generalized linear model both generate the same function for the probability of i choosing j. In the following section the assumptions of the multinomial logistic regression model are discussed.

2.4 Assumptions

In the previous section it is shown that both models generate the same probability function. In most statistics books there is no clear distinction between the derivations of the model and therefore the assumptions are presented jointly in this section. The first three assumptions are used in the derivation of the probability function, while the last three assumptions are important for the estimation of the regression coefficients.

1. **Dependent variable:** Firstly it is assumed that the dependent variable has a multinomial distribution, meaning that it takes on a number of finite values, which are unordered. This is fulfilled if the dependent variable is categorical [Fox, 2016].

2. **Linearity in the logit:** Moreover logistic regression assumes a linear relationship between the continuous independent variables and the logit transformation of the dependent variable, or in equation form:
$$\log \frac{\pi_{ij}}{\pi_{im}} = \eta_{ij} = \beta_{0j} + \beta_{1j} X_{i1} + \cdots + \beta_{kj} X_{ik} \tag{2.15}$$

3. **Independence of errors:** Additionally logistic regression assumes that responses of different cases are independent of each other. That is, it is assumed that each response comes from a different, unrelated case [Field et al., 2012], giving that the error terms in the model are independent of each other.

   In the random utility model, it is specifically noted that each unobserved part ε_{ij} is independent, identically distributed, with the Gumbel distribution [Train, 2009]. This implies that the cumulative distribution function of ε_{ij} is given by
$$F(\varepsilon_{ij}) = e^{-e^{-\varepsilon_{ij}}} \qquad \forall j = 1, \ldots, m \tag{2.16}$$
   and that the density function of ε_{ij} is given by
$$f(\varepsilon_{ij}) = e^{-\varepsilon_{ij}} e^{-e^{-\varepsilon_{ij}}} \qquad \forall j = 1, \ldots, m \tag{2.17}$$
   The variance of this distribution is $\frac{\pi^2}{6}$ [Train, 2009]. Hence the unobserved part of each alternative, ε_{ij}, follows the same distribution, is independent and has the same variance.

4. **Multicollinearity:** Furthermore there should be no multicollinearity among the predictor variables. Multicollinearity occurs when two or more predictor variables are highly correlated, which can lead to unstable and unreliable estimates of the regression coefficients [Fox, 2016].

5. **Separability:** However, problems also occur when the independent variables are very strong predictors of the dependent variable. There should be no perfect separation of the predictor variables, which occurs when one predictor variable can perfectly predict the response variable. This can lead to singularity in the model, resulting in estimates that are not unique. If there are k independent variables, then this assumption is checked by seeing if there is no hyperplane separating one category of the dependent variable. More specifically, if there are k independent variables, then we check if there is no separating linear surface of dimension $k-1$ in the $k-$dimensional X space [Fox, 2016].

6. **Independence of irrelevant alternatives**: Lastly it is assumed that there is independence of irrelevant alternatives. The assumption of independence of irrelevant alternatives is satisfied when the choice between two alternatives does not depend on any other alternatives. More specifically, if we look at the ratio of the probabilities of two alternatives i and s, using lemma 2.2.1 we have:
$$\frac{\pi_{ij}}{\pi_{is}} = \frac{\frac{e^{V_{ij}}}{\sum_l e^{V_{il}}}}{\frac{e^{V_{is}}}{\sum_l e^{V_{il}}}} = e^{V_{ij} - V_{is}} \tag{2.18}$$
   Where we used the formula $\pi_{ij} = \frac{e^{V_{ij}}}{\sum_l e^{V_{il}}}$ which is derived in the previous section. The ratio of the two probabilities does not depend on any other alternatives, hence the relative odds of choosing i over s is independent of other alternatives.

   A famous example of a situation in which this assumption is not satisfied is the red-bus-blue-bus-problem [Train, 2009]. Suppose a traveler can go to work by car or by taking the blue bus and that those options both have probability of $\frac{1}{2}$, hence $\mathbb{P}(\text{car}) = \mathbb{P}(\text{blue bus}) = \frac{1}{2}$. We now see that the ratio of the probabilities is 1: $\frac{\mathbb{P}(\text{car})}{\mathbb{P}(\text{blue bus})} = 1$. When a red bus is also added as a third alternative, the probability of taking the red bus is likely to be the same as the probability of taking the blue bus, hence the ratio of the probabilities is likely to be 1: $\frac{\mathbb{P}(\text{red bus})}{\mathbb{P}(\text{blue bus})} = 1$. Also, by the assumption of irrelevant alternatives we have that the ratio of the probabilities of taking the blue bus and the car remains the same when a third alternative is added, hence we still have $\frac{\mathbb{P}(\text{car})}{\mathbb{P}(\text{blue bus})} = 1$. Thus by combining these two ratios, we see that the only possible distribution of probabilities is $\mathbb{P}(\text{car}) = \mathbb{P}(\text{blue bus}) = \mathbb{P}(\text{red bus}) = \frac{1}{3}$. However: in a real-life situation it would be very unlikely that introducing another colour of a bus would change the probabilities so dramatically. Hence it would be more likely that $\mathbb{P}(\text{car})$ remains $\frac{1}{2}$ and $\mathbb{P}(\text{blue bus}) = \mathbb{P}(\text{red bus}) = \frac{1}{4}$. Thus the assumption of independence of irrelevant alternatives is not satisfied in this example.

All in all, several assumptions are made in the multinomial logistic regression model. The first three are necessary for the derivation of the probability function, while the last three are important for the estimation of coefficients. Failure to meet these assumptions may result in very large coefficients (or standard errors) or in failure of the software to converge to values for the coefficients. In section 4.3 an assumption check is described for the assumptions that enable this. The coefficients $\beta_{0j}, \ldots, \beta_{kj}$ can subsequently be estimated in several ways, either adopting a frequentist or a Bayesian approach.

Chapter 3

Estimating the regression coefficients

In the current section the methods of estimating coefficients will be outlined. The multinomial logistic regression model will firstly be presented in matrix notation. This enables a clear derivation of the likelihood function, which is used in both the frequentist and the Bayesian approach. Subsequently the concept of maximum likelihood estimation is explained, which is used in the frequentist approach. Lastly the Bayesian approach to computing the coefficients is described. For both methods the numerical approximation methods are outlined, so that it is conceptually clear what happens inside the 'black box' of the statistical programs.

3.1 Model in matrix notation

To derive the likelihood function, it is useful to present the model in matrix notation. In the multinomial logistic regression model it is assumed that there are k independent variables and m different categories of the dependent variable. As proved in proposition 2.3 the probability function of i falling into category $j \in \{1, \dots, m\}$ is given by:

$$\pi_{ij} = \frac{e^{\beta_{0j}+\beta_{1j}X_{i1}+\cdots+\beta_{kj}X_{ik}}}{1+\sum_{l=1}^{m-1} e^{\beta_{0l}+\beta_{1l}X_{i1}+\cdots+\beta_{kl}X_{ik}}} \quad \text{for } j = 1, \dots, m-1 \tag{3.1}$$

$$\pi_{im} = 1 - \sum_{j=1}^{m-1} \pi_{ij} \quad \text{for category } m \tag{3.2}$$

The log odds are given by:

$$\log \frac{\pi_{ij}}{\pi_{im}} = \beta_{0j} + \beta_{1j}X_{i1} + \cdots + \beta_{kj}X_{ik} \quad \text{for } j = 1, \dots, m-1 \tag{3.3}$$

Therefore the regression coefficients represent the effect on the log-odds of membership in category j versus the baseline category m. The coefficients form a vector $\boldsymbol{\beta_j} = (\beta_{0j}, \dots, \beta_{kj})^T$. Moreover if we denote the number of observations by n, then we define $\boldsymbol{X_i}' = (1, X_{i1}, \dots, X_{ik})$ for $i = 1, \dots, n$. The model matrix is a $(n \times k+1)$-matrix X, where k is the number of independent variables:

$$X = \begin{bmatrix} \boldsymbol{X_1'} \\ \boldsymbol{X_2'} \\ \vdots \\ \boldsymbol{X_n'} \end{bmatrix} \tag{3.4}$$

Giving as the probability function

$$\pi_{ij} = \frac{e^{\boldsymbol{X_i'\beta}_j}}{1+\sum_{l=1}^{m-1} e^{\boldsymbol{X_i'\beta}_j}} \quad \text{for } j = 1, \dots, m-1 \tag{3.5}$$

$$\pi_{im} = 1 - \sum_{j=1}^{m-1} \pi_{ij} \quad \text{for category } m \tag{3.6}$$

$$\tag{3.7}$$

The aim of multinomial logistic regression is to determine the values of the $\boldsymbol{\beta_j}$ for all $j = 1, \ldots, m-1$. The value of $\boldsymbol{\beta_m}$ is $\mathbf{0}$, since this is the reference category. Each vector $\boldsymbol{\beta_j}$ contains a regression coefficient for all independent variables plus the intercept, hence each vector has $k+1$ entries. Thus in total $(m-1)\cdot(k+1)$ parameters need to be determined, which are contained in the large vector $\boldsymbol{\beta}$, which has $(m-1)\cdot(k+1)$ entries.

3.2 Likelihood function

The likelihood function is used in both the frequentist and Bayesian approach to estimate the parameters. In order to derive the likelihood function we recall that it is assumed that there are m different categories of the independent variable. Let $Y = (Y_1, \ldots, Y_n)$ be the observed data set with sample size n. The probability that Y_i takes on one of the possible values $j \in \{1, \ldots, m\}$ is denoted by π_{i1}, $\pi_{i2}, \ldots, \pi_{im}$. This probability can be rewritten by defining indicator variables $W_{i1}, \ldots, W_{im}$.

Definition 3.1 The indicator variables $W_{i1}, \ldots, W_{im}$ are defined as:

$$W_{ij} = \begin{cases} 1 & \text{if } Y_i = j \\ 0 & \text{if } Y_i \neq j \end{cases} \tag{3.8}$$

with $\sum_{j=1}^{m} W_{ij} = 1$.

Proposition 3.2 The conditional likelihood function of n independently sampled observations of Y conditional on X in multinomial logistic regression is given by:

$$p(y_1, \ldots, y_n | X) = \prod_{i=1}^{n} \prod_{j=1}^{m} \left(\frac{e^{\boldsymbol{X_i'\beta}_j}}{1 + \sum_{l=1}^{m-1} e^{\boldsymbol{X_i'\beta}_j}}\right)^{W_{ij}} \tag{3.9}$$

Proof. The dependent variable Y_i can take on only one of the possible values. This gives

$$\begin{aligned} p(y_i) &= \pi_{i1}^{W_{i1}} \pi_{i2}^{W_{i2}} \cdots \pi_{im}^{W_{im}} \\ &= \prod_{j=1}^{m} \pi_{ij}^{W_{ij}} \end{aligned} \tag{3.10}$$

Since it is assumed that the observations are sampled independently, the joint probability distribution of the observations is given by the product of the marginal probabilities, hence

$$\begin{aligned} p(y_1, \ldots, y_n) &= p(y_1)p(y_2)\ldots p(y_n) \\ &= \prod_{i=1}^{n} \prod_{j=1}^{m} \pi_{ij}^{W_{ij}} \end{aligned} \tag{3.11}$$

The conditional likelihood function of n independently sampled observations of Y conditional on X is therefore:

$$p(y_1, \ldots, y_n | X) = \prod_{i=1}^{n} \prod_{j=1}^{m} \left(\frac{e^{\boldsymbol{X_i'\beta}_j}}{1 + \sum_{l=1}^{m-1} e^{\boldsymbol{X_i'\beta}_j}}\right)^{W_{ij}} \qquad \text{use lemma 2.3} \tag{3.12}$$

□

This likelihood function is used in both the frequentist and the Bayesian approach. The difference between both methods is the way how they infer the parameters. In the frequentist approach it is assumed that the value of these parameters is a fixed value which can be approximated. Numerical methods are often used to approximate the value of these parameters. However, in the Bayesian setting we instead seek a posterior distribution for the parameters, by for example employing Markov Chain Monte Carlo methods. Both approaches will be discussed in the following sections.

3.3 Frequentist: Maximum likelihood estimation

In the frequentist approach the parameters are estimated with maximum likelihood estimation, which will be outlined in this section. This method maximizes the likelihood function such that, under the assumed statistical model, the observed data is most probable. The vector $\boldsymbol{\beta}_j$ that assigns the greatest probability to the observed values of the dependent variables is the maximum likelihood estimate.

The logarithm of the likelihood is taken firstly, since this limits the necessary calculations. The lograrithm is a monotone function, hence the value $\boldsymbol{\beta}_j$ maximizes the likelihood function if and only if this value maximizes the log-likelihood function. Therefore the logarithm of the likelihood function can be used to determine the maximum.

Proposition 3.3 The log-likelihood function of multinomial logistic regression is given by

$$\log(p(y_1,\dots,y_n|X)) = \sum_{i=1}^{n}\sum_{j=1}^{m-1} W_{ij}\boldsymbol{X_i'}\boldsymbol{\beta}_j - \sum_{i=1}^{n}\log(1+\sum_{l=1}^{m-1} e^{\boldsymbol{X_i'}\boldsymbol{\beta}_j})$$

Proof.

$$\begin{aligned}
\log(p(y_1,\dots,y_n|X)) &= \log(\prod_{i=1}^{n}\prod_{j=1}^{m}(\frac{e^{\boldsymbol{X_i'}\boldsymbol{\beta}_j}}{1+\sum_{l=1}^{m-1} e^{\boldsymbol{X_i'}\boldsymbol{\beta}_j}})^{W_{ij}}) \\
&= \sum_{i=1}^{n}\sum_{j=1}^{m} W_{ij}\log(\frac{e^{\boldsymbol{X_i'}\boldsymbol{\beta}_j}}{1+\sum_{l=1}^{m-1} e^{\boldsymbol{X_i'}\boldsymbol{\beta}_j}}) \\
&= \sum_{i=1}^{n}\sum_{j=1}^{m} W_{ij}(\boldsymbol{X_i'}\boldsymbol{\beta}_j - \log(1+\sum_{l=1}^{m-1} e^{\boldsymbol{X_i'}\boldsymbol{\beta}_j})) \\
&= \sum_{i=1}^{n}\sum_{j=1}^{m} W_{ij}\boldsymbol{X_i'}\boldsymbol{\beta}_j - \sum_{i=1}^{n}\sum_{j=1}^{m} W_{ij}\log(1+\sum_{l=1}^{m-1} e^{\boldsymbol{X_i'}\boldsymbol{\beta}_j}) \\
&= \sum_{i=1}^{n}\sum_{j=1}^{m-1} W_{ij}\boldsymbol{X_i'}\boldsymbol{\beta}_j - \sum_{i=1}^{n}\sum_{j=1}^{m} W_{ij}\log(1+\sum_{l=1}^{m-1} e^{\boldsymbol{X_i'}\boldsymbol{\beta}_j}) && \boldsymbol{\beta_m} = 0, \text{ baseline category} \\
&= \sum_{i=1}^{n}\sum_{j=1}^{m-1} W_{ij}\boldsymbol{X_i'}\boldsymbol{\beta}_j - \sum_{i=1}^{n}\log(1+\sum_{l=1}^{m-1} e^{\boldsymbol{X_i'}\boldsymbol{\beta}_j}) && \text{since } \sum_{j=1}^{m} W_{ij} = 1
\end{aligned}$$

□

To find the critical points of the log-likelihood function, the partial derivatives of the log-likelihood with respect to $\boldsymbol{\beta_j}$ need to be determined, giving a system of $m-1$ equations. The vector of partial derivatives is subsequently set to 0, giving a vector $\boldsymbol{\beta}$. The log-likelihood function is globally concave, hence the vector is the unique global maximum [Croissant, 2010]. Helpful in the differentiation is that:

$$\frac{\partial}{\partial \boldsymbol{\beta}_j} \sum_{j=1}^{m-1} \boldsymbol{X_i'}\boldsymbol{\beta}_j = \boldsymbol{X_i'} \tag{3.13}$$

since the other terms in the summation do not depend on $\boldsymbol{\beta_j}$ and can thus be treated as constants. Differentiating the log-likelihood function with respect to each $\boldsymbol{\beta_j}$ gives:

$$\begin{aligned}
\frac{\partial \log(p(y_1,\dots,y_n|))}{\partial \boldsymbol{\beta_j}} &= \sum_{i=1}^{n} W_{ij}\boldsymbol{X_i'} - \sum_{i=1}^{n} \frac{1}{1+\sum_{l=1}^{m-1} e^{\boldsymbol{X_i'}\boldsymbol{\beta}_j}} \frac{\partial}{\partial \boldsymbol{\beta}_j}(1+\sum_{l=1}^{m-1} e^{\boldsymbol{X_i'}\boldsymbol{\beta}_j}) && \text{Use: } \frac{\partial}{\partial x}\log(y) = \frac{1}{y}\frac{\partial y}{\partial x} \\
&= \sum_{i=1}^{n} W_{ij}\boldsymbol{X_i'} - \sum_{i=1}^{n} \frac{e^{\boldsymbol{X_i'}\boldsymbol{\beta}_j}}{1+\sum_{l=1}^{m-1} e^{\boldsymbol{X_i'}\boldsymbol{\beta}_j}} \frac{\partial}{\partial \boldsymbol{\beta}_j}(\sum_{l=1}^{m-1} \boldsymbol{X_i'}\boldsymbol{\beta}_j) && \text{Chain rule} \\
&= \sum_{i=1}^{n} W_{ij}\boldsymbol{X_i'} - \sum_{i=1}^{n} \frac{e^{\boldsymbol{X_i'}\boldsymbol{\beta}_j}}{1+\sum_{l=1}^{m-1} e^{\boldsymbol{X_i'}\boldsymbol{\beta}_j}} \boldsymbol{X_i'} && (3.14)
\end{aligned}$$

Setting these partial derivatives to zero and determining the $\boldsymbol{\beta_j}$'s gives the unique maximum likelihood estimate. However, there is often not an explicit solution for $\boldsymbol{\beta}$ and therefore we need a method for numerical approximation.

3.3.1 Estimating coefficients by the Newton-Raphson method

The numerical Newton-Raphson method is used in the frequentist approach to approximate $\boldsymbol{\beta}$. The Newton-Raphson method is an iterative algorithm for finding the roots of a nonlinear equation. In history, this method was supposedly already used by the ancient Babylonians, but the first well-known use of the method was in 1669 by Newton [Epperson, 2013]. The method will be explained for a general function $f(\boldsymbol{x})$.

The Newton-Raphson method can be used to approximate the solution of an equation of the form $f(x) = 0$ where $f : X \subseteq \mathbb{R} \to \mathbb{R}$ is a differentiable function. The method will firstly be outlined for determining a single estimate, since this can be graphically illustrated. Hence we want to find a number x such that $f(x) = 0$. The following steps are taken:

1. Make an initial guess of the root, namely x_0.
2. To improve the guess, draw the tangent line to the curve $f(x)$ at the point $(x_0, f(x_0))$. This line is denoted by $f'(x_0)$
3. The point where the tangent line crosses the $x-$axis is denoted as $(x_1, 0)$ and x_1 is taken as the revised and improved approximation of the root x.

In order to find x_1, we look at the equation of the tangent line, $y = f(x_0) + f'(x_0)(x_1 - x_0)$. We set $y = 0$ to find where this line crosses the $x-$axis, giving $0 = f(x_0) + f'(x_0)(x_1 - x_0)$. Solving this for x_1 gives $x_1 = x_0 - \frac{f(x_0)}{f'(x_0)}$. This process is repeated to find x_2, which is an even better approximation of the root. This process is repeated, giving $x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}$, until the values $x_1, x_2, \dots$ converge to a fixed value x, which is the root of the equation [Colley, 2012]. The x_n do not necessarily have to converge, but when they do, they converge to the root.

However, in the multinomial logistic regression we want to find the vector $\boldsymbol{\beta}$ that maximizes the log-likelihood function. Since this is a vector, we are not considering only one function, but $m-1$ different functions. In this situation the Newton-Raphson method can also be used in order to approximate the solution of an equation of the form $f(\boldsymbol{x}) = \boldsymbol{0}$, where $\boldsymbol{f} : X \subseteq \mathbb{R}^m \to \mathbb{R}^m$. We have to find $\boldsymbol{x}$ such that $\boldsymbol{f}(\boldsymbol{x}) = \boldsymbol{0}$. We again make an initial guess of the root, namely $\boldsymbol{x_0}$. If $\boldsymbol{f}$ is differentiable, then the tangent plane $\boldsymbol{y} = \boldsymbol{f}(\boldsymbol{x})$ is approximated by the following equation, which is again set to 0:

$$\boldsymbol{y} = \boldsymbol{f}(\boldsymbol{x_0}) + J_{\boldsymbol{f}}(\boldsymbol{x_0})(\boldsymbol{x_1} - \boldsymbol{x_0}) = \boldsymbol{0} \tag{3.15}$$

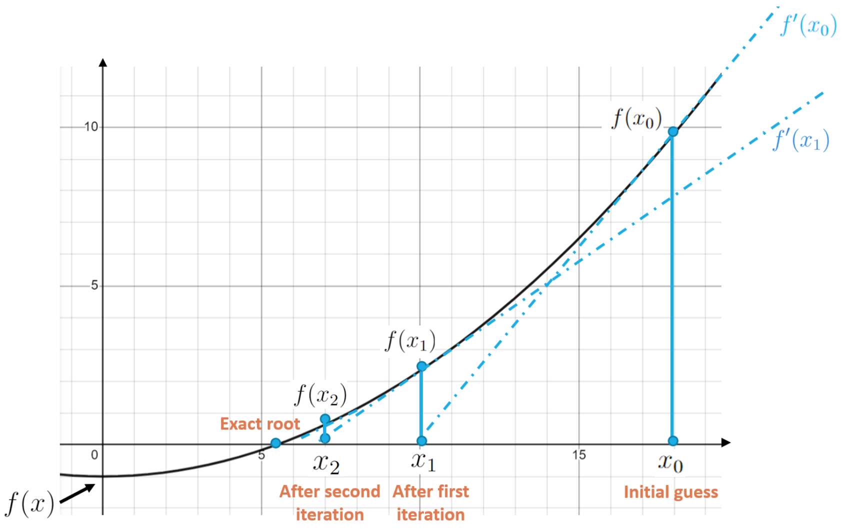

Figure 3.1: Graphical illustration of the Newton-Raphson method for $f : X \subseteq \mathbb{R} \to \mathbb{R}$.

where $J_{\boldsymbol{f}}$ is the Jacobian matrix, which is the matrix of all first-order partial derivatives. We suppose that the Jacobian matrix is invertible and solve this for $\boldsymbol{x_1}$:

$$\begin{aligned}
\boldsymbol{0} &= \boldsymbol{f}(\boldsymbol{x_0}) + J_{\boldsymbol{f}}(\boldsymbol{x_0})(\boldsymbol{x_1} - \boldsymbol{x_0}) \\
J_{\boldsymbol{f}}(\boldsymbol{x_0})(\boldsymbol{x_1} - \boldsymbol{x_0}) &= -\boldsymbol{f}(\boldsymbol{x_0}) \\
(J_{\boldsymbol{f}}(\boldsymbol{x_0}))^{-1} J_{\boldsymbol{f}}(\boldsymbol{x_0})(\boldsymbol{x_1} - \boldsymbol{x_0}) &= (J_{\boldsymbol{f}}(\boldsymbol{x_0}))^{-1}(-\boldsymbol{f}(\boldsymbol{x_0})) \\
I(\boldsymbol{x_1} - \boldsymbol{x_0}) &= -(J_{\boldsymbol{f}}(\boldsymbol{x_0}))^{-1}(\boldsymbol{f}(\boldsymbol{x_0})) \\
\boldsymbol{x_1} &= \boldsymbol{x_0} - (J_{\boldsymbol{f}}(\boldsymbol{x_0}))^{-1}(\boldsymbol{f}(\boldsymbol{x_0}))
\end{aligned}$$

Where invertibility of the matrix $J_{\boldsymbol{f}}(\boldsymbol{x_0})$ is used in the third step. This can again be generalized, giving the following formula:

$$\boldsymbol{x_{n+1}} = \boldsymbol{x_n} - (J_{\boldsymbol{f}}(\boldsymbol{x_n}))^{-1}(\boldsymbol{f}(\boldsymbol{x_n})) \tag{3.16}$$

The sequence of numbers which is generated does not always converge, but when it does it converges to the root of the equation. The iterations continue until the value is a good approximation of the root. Multiple numerical methods exist that are closely related to the Newton-Raphson method and all work in a similar manner. For example Fisher's scoring is used in the R package VGAM (one of the packages which will be used in the current study) [Yee, 2008]. Fisher's scoring is a Newton-like method but uses Fisher's information instead of the first derivative to approximate the root.

3.3.2 Important concepts: NHST, Wald statistic, $p-$value and confidence intervals

In the frequentist approach to statistics, null hypothesis significance testing (NHST) is very common. In NHST a null hypothesis H_0 is formulated, which is a hypothesis of no effect or no relationship [Pernet, 2016]. The statistical analysis can lead to two possible conclusions: rejecting H_0 (and accepting $H_{\text{alternative}}$) or not rejecting H_0 (but not accepting H_0 as correct) [van der Vaart et al., 2017].

In logistic regression, the null hypothesis that is commonly tested for each coefficient is that the value of the coefficient is 0, hence that there is no effect of the independent variable on the dependent variable. For each estimated coefficient the null hypothesis $H_0 : \beta_{kj} = 0$ is tested by calculating the Wald statistic [Fox, 2016], which is the value of the regression coefficient divided by its associated standard error.

Definition 3.4 The Wald statistic for an individual coefficient β_{kj} and the null hypothesis $H_0 : \beta_{kj} = \beta_{kj}^{(0)}$ is defined as:

$$Z_0 = \frac{\hat{\beta}_{kj} - \beta_{kj}^{(0)}}{SE(\hat{\beta}_{kj})}$$

with $SE(\hat{\beta}_{kj})$ the standard error of the estimated coefficient$\hat{\beta}_{kj}$.

Hence when testing $H_0 : \beta_{kj} = 0$, which is commonly used, we look at the Wald statistic $Z_0 = \frac{\hat{\beta}_{kj}}{SE(\hat{\beta}_{kj})}$.

p-value

"A p-value is the probability under a specified statistical model that a statistical summary of the data (e.g., the sample mean difference between two compared groups) would be equal to or more extreme than its observed value." [Wasserstein and Lazar, 2016] The p-value is often combined with Null hypthesis significance testing. According to van der Vaart et al. (2017), under the null hypothesis, the $p-$value is the maximum over all possible values of the probability that an identical experiment gives a more extreme value for the test statistic than the value that has been found in the experiment. It is a common convention to consider values of p that are below 0.05 as (statistically) significant. This originated since Fisher once mentioned using 0.05 as a treshold value. Since the $p-$value does not give very much information, it is recommended to also always report confidence intervals.

Confidence intervals

A confidence interval is a stochastic subset of the parameter space that has a "high probability" of containing the true parameter [van der Vaart et al., 2017]. Often a 95%-confidence interval is reported. In the frequentist interpretation, the true value of the parameter is assumed to be fixed. Therefore the true value is either in the confidence interval, or it is not. The confidence interval can be interpreted in the sense that if we independently repeat an experiment 100 times and compute the confidence interval 100 times, then we may expect that at least 95% of the intervals include the true parameter value.

The computation of the confidence interval depends on the test statistic for which it is computed. The Wald statistics follows an asymptotic standard-normal distribution under the null hypothesis, an approximation that is usually reasonable except if the sample size is small [van der Vaart et al., 2017]. An asymptotic $100(1-\alpha)\%$ confidence interval for β_{kj} is given by

$$\beta_{kj} = \hat{\beta_{kj}} \pm z_{\alpha/2} SE(\hat{\beta_{kj}})$$

where $z_{\alpha/2}$ is the value of $Z \sim N(0,1)$ with probability $\alpha/2$. For a 95% confidence interval, the value of $z_{2.5}$ is obtained from the table of the normal distribution in van der Vaart et al. (2017), which gives:

$$\beta_{kj} = \hat{\beta_{kj}} \pm 1.96 SE(\hat{\beta_{kj}})$$

.

3.4 Bayesian: Infer posterior distribution

As mentioned before, the frequentist approach supposes that the values of $\boldsymbol{\beta}_j$ are fixed and that these can be approximated. However in the Bayesian approach it is supposed that the parameters $\boldsymbol{\beta_j}$ are not fixed values, but rather random variables. The Bayesian approach has three important components, namely [McElreath, 2020]:

1. **Likelihood** The likelihood function, which is the first and most influential component of the Bayesian model. This function is similar to the one used in the frequentist approach and is denoted by $\mathbb{P}(X|\boldsymbol{\beta})$, with X denoting the observed data and $\boldsymbol{\beta}$ denoting the coefficients [Lee, 2012].

2. **Prior distribution** For every coefficient that one aims to estimate, a prior distribution has to be chosen [van der Vaart et al., 2017]. This prior distribution can be chosen as 'uninformative', when not much information is available. However, if there is reason to assume that a parameter has a certain distribution, then this can be reflected in choosing an informative prior distribution, which reflects the information [McElreath, 2020].

3. **Posterior distribution** The posterior distribution is subsequently computed. Given the prior distribution on the parameters we seek the posterior distribution.

   The derivation of the posterior is based on Bayes' theorem [Grimmett and Welsh, 2014]. To prove Bayes' theorem, the definition of conditional probability is required.

Definition 3.5 The conditional probability is a measure of the probability of event A occurring, given that event B has occurred, hence $\mathbb{P}(B) > 0$. It is defined as

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)} \tag{3.17}$$

where $\mathbb{P}(A \cap B)$ denotes the probability of both event A and event B occurring.

Theorem 3.6 (Bayes' theorem) Let A and B be events and $\mathbb{P}(B) > 0$, then

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(B|A)\mathbb{P}(A)}{\mathbb{P}(B)} \tag{3.18}$$

Proof. We use the definition of conditional probability, which gives

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}$$

and also

$$\mathbb{P}(B|A) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(A)}$$

Hence $\mathbb{P}(A \cap B) = \mathbb{P}(B|A)\mathbb{P}(A)$. Substituting this into $\mathbb{P}(A|B)$ gives

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)} = \frac{\mathbb{P}(B|A)\mathbb{P}(A)}{\mathbb{P}(B)} \tag{3.19}$$

□

By conditioning on the observed data X, the posterior distribution of each component of $\boldsymbol{\beta}$ can be computed by Bayes theorem, giving:

$$\text{Posterior} = \mathbb{P}(\boldsymbol{\beta}|X) = \frac{\mathbb{P}(X|\boldsymbol{\beta})\mathbb{P}(\boldsymbol{\beta})}{\mathbb{P}(X)} = \frac{\text{Likelhood} \times \text{Prior}}{\mathbb{P}(X)} \tag{3.20}$$

where $\frac{1}{\mathbb{P}(X)}$ is a standardizing aspect and considered as the average likelihood of the data, which ensures that the posterior acts as a probability function (summing to 1) [McElreath, 2020]. Therefore it can be neglected and stated that the posterior distribution is proportional to the product of the likelihood and the prior distribution:

$$\text{Prior} \times \text{Likelihood} \propto \text{Posterior} \tag{3.21}$$

The posterior distribution represents our updated beliefs about the model parameters, taking into account the data that we have observed. As an example, see figure 3.2. The likelihood function is the same in every row, but the priors differ. For example in the first row a uniform distribution is chosen as the prior distribution, leading to a posterior that is proportional to the likelihood. The second and third row both have an informative prior, representing the information that one assumes to have about the posterior probability. This leads to different posterior distributions.

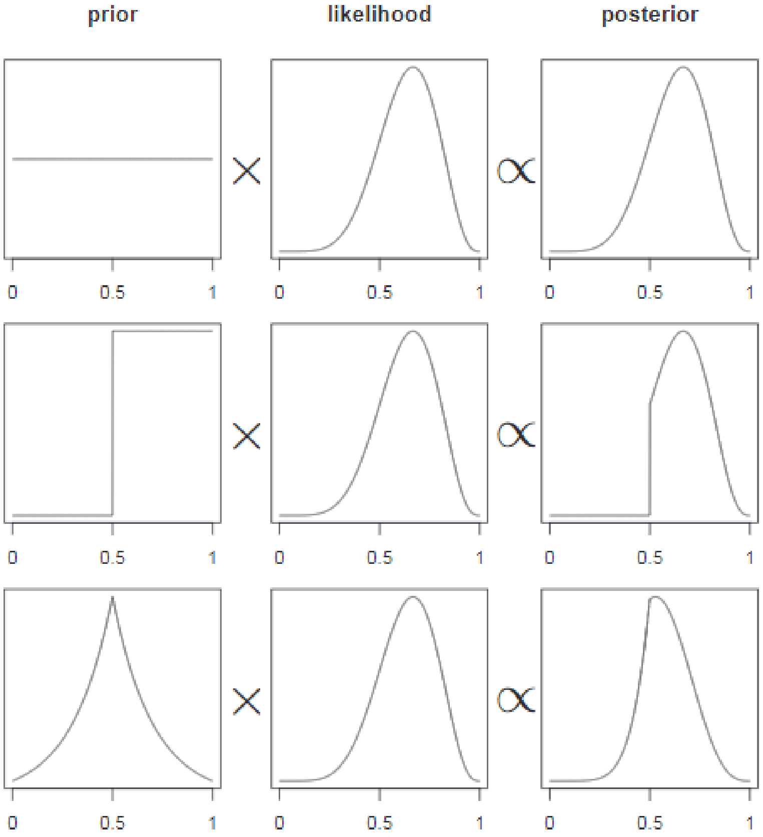

Figure 3.2: The posterior distribution is proportional to the product of the likelihood and the posterior distribution (from McElreath(2020))

In general, the posterior distribution cannot be found in closed form and must be approximated, usually using some type of Markov Chain Monte Carlo method.

3.4.1 Determine posterior distribution by MCMC sampling

The posterior distribution can be computed analytically in only a few cases. In most cases, the posterior distribution is numerically computed. To numerically obtain the posterior distributions of the regression coefficients $\boldsymbol{\beta}$, Markov Chain Monte Carlo (MCMC) sampling is a commonly used technique. The general idea is to start with the probability distribution that is proportional to the posterior distribution. By drawing samples from this distribution, the desired posterior distribution can be approximated. It is often easier to draw a large random sample of a distribution than to calculate properties directly. MCMC sampling combines both Monte Carlo methods and Markov chains, which are shortly introduced.

Monte Carlo methods can be used to estimate the distribution of probability for a given problem by simulating a large number of random samples according to a probability distribution. Subsequently the statistical properties of those samples are used to make predictions about the underlying distribution [Bolstad and Curran, 2016]. The histogram of the drawn random sample approaches the posterior distribution when the sample size is large. It is not necessary to know the exact posterior distribution, only knowing the shape of the posterior distribution is enough to draw samples from it [Bolstad and Curran, 2016].

Furthermore Markov chains are used in MCMC sampling. A Markov chain is a sequence of random variables that follows the Markov property. The Markov property states that the probability of a future state depends only on the current state and not on the past states. In MCMC algorithms, a Markov chain is constructed, where the next sample that is drawn, only depends on the last drawn sample. Furthermore the Markov Chain is designed such that it is ensured that the stationary distribution of the chain is the posterior distribution [Zens et al., 2021]. The quality of the sample increases as more steps are included, resulting in a sample that more closely approximates the actual desired distribution.

MCMC sampling allows the estimation of the posterior distribution that would be difficult or even impossible to approach analytically. MCMC methods can be used for quite complicated models having a large number of parameters [Bolstad and Curran, 2016]. In the last decades of the twentieth century, Bayesian statistics became increasingly popular since sufficient computing power enabled their use.

3.4.2 Credible intervals

In Bayesian statistics, the posterior distribution represents all of the information that can be inferred about the parameters of a model after analyzing the data. Subsequently, point estimates and credible intervals can be obtained from the posterior distribution. A Bayesian credible interval is an interval of the posterior probability which is the Bayesian equivalent of the frequentist confidence interval. A credible interval is defined by two parameter values that contain between them a certain amount of posterior probability [McElreath, 2020]. The amount of probability can for example be chosen as 80%: in figure 3.3 two types of corresponding credible intervals are shown.

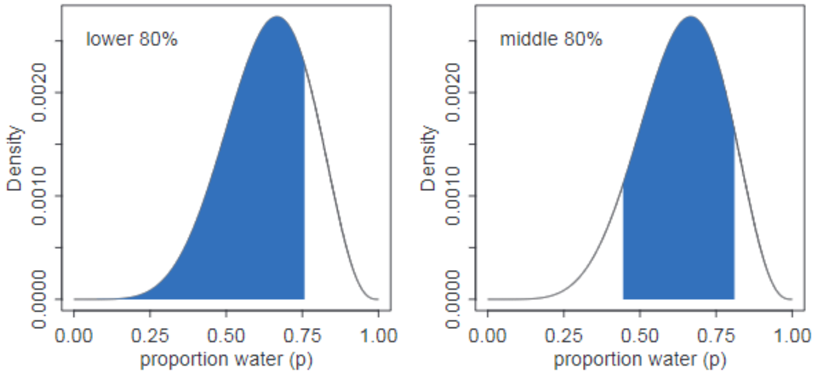

Figure 3.3: Two types of 80% credible intervals: middle and lower (from McElreath(2020))

In the current study a middle 95% credible interval is chosen as the credible interval, since this is most comparable to the 95% confidence interval from the frequentist approach. However, it is important to note that other types of credible intervals are also possible.

Since the Bayesian parameter estimate is the entire posterior distribution, there is not one point estimate. However, since scientists often want to report some sort of point estimate, it is common to include the posterior mean as an point estimate [McElreath, 2020].

Chapter 4

Application to data from PRIME

In order to compare different approaches and R packages, multinomial logistic regressions are performed with three different packages: the `VGAM` package (frequentist, generalized linear model), the `mlogit` package (frequentist, random utility model) and the `UPG` package (Bayesian, random utility model), as depicted in figure 4.1. The data that is analyzed is part of a study of PRIME. In the current chapter this study is described. Moreover it is noted how the data was prepared for analysis. Lastly the assumptions that allow a check, either visually or computationally, are checked and reported in the last section.

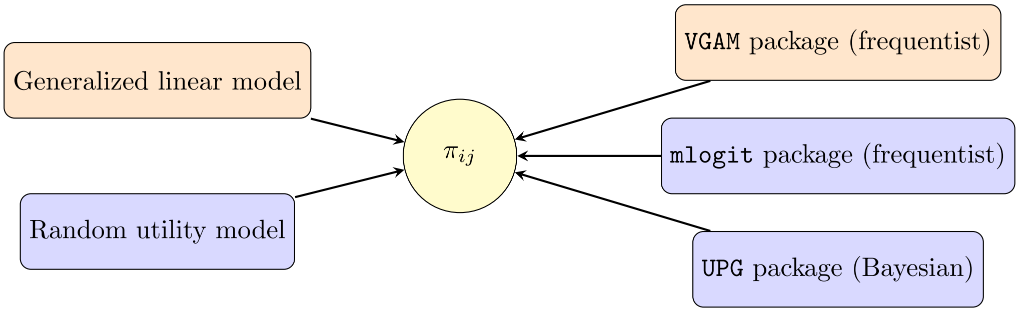

Figure 4.1: The same probability equation π_{ij} is derived with the Generalized linear model and Random Utility model. The analysis will be performed with three different R packages (either frequentist or Bayesian, colours indicating GLM or RUM)

4.1 Description of PRIME study

The PRogramme of Innovation in Mathematics Education (PRIME) is part of the Interfaculty Teaching from the department of Applied Mathematics at the TU Delft and aims to redesign mathematics courses for engineers [PRIME TU Delft, 2022]. Besides developing education, PRIME also investigates the mathematics education through several studies, giving rise to possible improvements for further development of courses. The current data set is from such a study, which was conducted among students who did a mathematics course. The study aimed to examine the influence of prior math grades, mathematics self-efficacy and grade goal on study beliefs and habits.

4.1.1 Participants

The participants in the study were first-year students at the TU Delft, who were enrolled in a mathematics course. In total 286 students filled in the survey. After deletion of 7 cases who did not finish the survey, data from 279 students were available for analysis. The mean age is 19.53 years old (standard deviation: 2.91) and the median is 19. The youngest participant is 17 years old and the oldest 41 years old. Three students did not report their age. Mostly male participants participated in the study, namely 215, and 62 female participants. The percentage of female participants who completed the survey is 22.2%, which seems little, but it quite in line with the percentage of female students at TU Delft, which is 29.8%. Two students identified as non-binary.

4.1.2 Dependent variables: study beliefs and habits

The study beliefs and habits are the dependent variables in the multinomial logistic regressions. The survey included eight questions to assess study habits, adapted from [Kornell and Bjork, 2007]. The following questions were asked, of which the first three assess study beliefs and the other five assess scheduling habits:

1. When you study, do you typically read a textbook/article/other source material more than once? (three possible answers)
2. Imagine that while you are studying a type of problem and you become convinced that you know the answer to the problem (e.g. the solutions). What would you do next? (three possible answers)
3. If you quiz yourself while you study (either using a quiz at the end of a chapter, or a practice quiz, or flashcards, or something else), why do you do so? (four possible answers)
4. How do you usually decide what to study next? (five possible answers)
5. All other things being equal, what type of exams do you study more for? (four possible answers)
6. Which of the following best describes your pattern of study? (three possible answers)
7. What time of day do you most often do your studying? (four possible answers)
8. During what time of the day do you believe your studying is (or would be) most effective? (four possible answers)

Every question corresponds to a separate dependent variable and the possible answers correspond to the different categories of each dependent variable. For each question a separate multinomial logistic regression is performed. A study overview is shown in figure 4.2. The independent variables are mathematics self-efficacy, grade goal and prior math grade and will be discussed shortly in the next section.

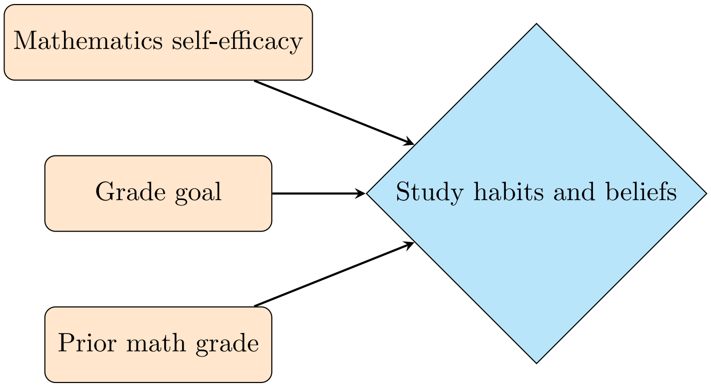

Figure 4.2: The independent variables are mathematics self-efficacy, grade goal & prior math grade. The dependent variable is study habits and beliefs

4.1.3 Independent variables

Mathematics self-efficacy: Self-efficacy is a construct which was introduced by Albert Bandura (1977) and is defined as "an individual's belief in his or her capacity to execute behaviors necessary to produce specific performance attainments" [Bandura, 1977]. Hence mathematics self-efficacy can be seen as a student's judgments about their capabilities to perform specific math-related tasks [Grigg et al., 2018]. Five items were included in the current questionnaire to assess mathematics self-efficacy, adapted from Grigg, Perera, McIlveen, and Svetleff (2018). Students were asked to rate each item on a 5-point scale ranging from 1 (strongly disagree) to 5 (strongly agree). The variable 'Self-efficacy' is constructed by averaging those five items.

Grade goal: Three items were used to measure grade goal, adapted from [Locke and Bryan, 1968]. Students were asked which grade they aimed to achieve, which grade they would be satisfied with and which grade they expected to obtain in the mathematics course. The answer had to be given on a scale of 1 to 10. The variable 'Grade goal' is constructed by averaging those three items.

Prior math grade: In order to estimate prior mathematics achievement, the participants were asked which grade they received in high school mathematics. The answer had to be given on a scale of 1 to 10. The score that was given is the variable 'Prior math grade'.

4.2 Preparing data for analysis

We will treat all three independent variables (self-efficacy, grade goal and prior math grade) as continuous variables. For the `mlogit` and `UPG` package it is required to reshape the data into a specific dataframe in order to be able to run the analysis, Details of the required format can respectively be found in [Croissant, 2010] and [Zens et al., 2021]. The R-code to reshape the data is included in appendix A.

4.2.1 Internal consistency: Cronbach's alpha

To measure self-efficacy and grade goal, different questions were asked in the questionnaire. These items should all measure the same thing, hence they should be correlated with each other. To assess this internal consistency we can use Cronbach's alpha [Cronbach, 1951], which can be computed with the following formula:

$$\alpha = \frac{k}{k-1}(1 - \frac{\sum s_i^2}{s_T^2}) \tag{4.1}$$

Where k is the number of items, s_i^2 is the variance of the ith item and s_T^2 is the variance of the total score, formed by summing all the items. The variance is the standard deviation squared.

To understand the values of Cronbach's alpha, two cases are illustrated. Firstly suppose the items are all different and have no internal consistency. This implies that the items are all independent. For independent variables, the variance of the total score (formed by summing the items) is the sum of their variances [Bland and Altman, 1997], i.e. $s_T^2 = \sum s_i^2$. Thus in this case $\alpha = \frac{k}{k-1}(1 - \frac{\sum s_i^2}{s_T^2}) = \frac{k}{k-1} \cdot (1-1) = 0$. On the other hand, suppose the items are all identical and thus have perfect internal consistency, then the standard deviation, s_i for each item will be identical. This gives $s_T = k \cdot s_i$ and $s_T^2 = k^2 s_i^2$. Resulting in $\frac{k \cdot s_i^2}{s_T^2} = \frac{\sum s_i^2}{s_T^2} = \frac{1}{k}$ and $\alpha = \frac{k}{k-1}(1 - \frac{\sum s_i^2}{s_T^2}) = \frac{k}{k-1} \cdot (1 - \frac{1}{k}) = \frac{k}{k-1} \cdot \frac{k-1}{k} = 1$. All in all, values close to 1 indicate high internal consistency, while values close to 0 indicate that none of the items are related.

Self-efficacy

Cronbach's alpha is computed to assess the internal consistency of the items measuring self-efficacy. In the following table the mean score and the variance is included for all the items. The mean of the total score is 19.129 and $s_T^2 = 15.106$.

Cronbach's alpha is computed using two methods: firstly the in-built R-function from the psych package. And secondly it is computed with formula 4.1. The code of which can be found in appendix B Both methods give approximately the same value for alpha, namely $\alpha = \frac{5}{4}(1 - \frac{4.377}{15.106}) = 0.888$, indicating high internal consistency.

SE_1		SE_2		SE_3		SE_4		SE_5	
mean	s_1^2	mean	s_2^2	mean	s_3^2	mean	s_4^2	mean	s_5^2
3.685	1.008	3.867	0.856	3.857	0.850	3.552	1.054	4.168	0.608

Grade goal

Cronbach's alpha is also computed to assess the internal consistency of the items measuring grade goal. In the following table the mean score and the variances are included for all items.

Grade aim		Grade lowest satisfied		Grade expected		Total score	
mean	s_{aim}^2	mean	s_{lowest}^2	mean	s_{expect}^2	mean	s_T^2
8.466	1.422	6.560	0.823	7.505	1.467	22.530	8.674

Cronbach's alpha is again computed with two methods: the in-built R-function and with the formula. Both methods give approximately the same value for alpha, namely $\alpha = \frac{3}{2}(1 - \frac{3.712}{8.674}) = 0.858$, which indicates a high degree of internal consistency.

Thus the items measuring both grade goal and self-efficacy seem to be highly correlated and therefore it seems appropriate to include the mean scores as the independent variables in the analysis.

4.3 Assumptions

Before doing analyses the assumptions should be checked. Since if the assumptions are not met, the results should be interpreted with caution. However, not all assumptions can be checked, but failure to meet them often exhibits in the analysis itself: for example by failure to converge or wildly large values for coefficients [Field et al., 2012].

4.3.1 Dependent variables

The dependent variables have a multinomial distribution, since there are either 3, 4 or 5 possible categories, which are unordered. Hence this assumption is fulfilled.

4.3.2 Multicollinearity

Moreover multicollinearity is checked, which occurs when two or more predictor variables are highly correlated. Multicollinearity can be checked both visually with a scatterplot and by computing the correlation coefficient between two variables. The correlation coefficient that is computed is Pearson's correlation coefficient, which is computed by the following formula [Schober et al., 2018].

$$r = \frac{\sum(x_i - \bar{x})(z_i - \bar{z})}{\sqrt{\sum(x_i - \bar{x})^2(z_i - \bar{z})^2}} \tag{4.2}$$

where x and z are the two independent variables for which the coefficient is computed. The mean values are denoted by $\bar{x}$ and $\bar{z}$ and the x_i and z_i are the different values over which is summed. The value of Pearson's correlation coefficient can range between -1 and 1, where values close to -1 and 1 indicate high degree of correlation. In this case, the data points are all on one line. Values close to 0 indicate a low degree of correlation. In this case, the data points are scattered around without a pattern.

Self-efficacy and grade goal For self-efficacy and grade goal the value of Pearson's correlation coefficient is given by $r = 0.662$, which indicates that the correlation is quite strong. However, Tabachnick et al. (2007) suggest that as long correlation coefficients among independent variables are less than 0.90, the assumption is met.

Also visually, it can be seen that the data points are quite strongly correlated. As shown in figure 4.3, the data points seem to be close to a line, although also scattered.

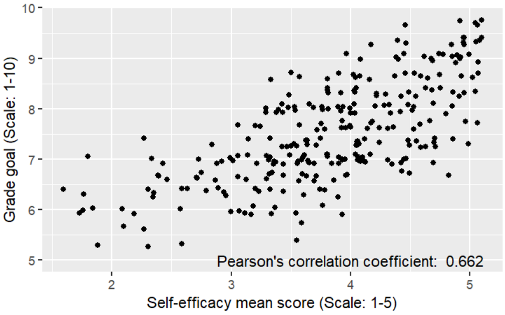

Figure 4.3: Scatter plot of prior math grade and self-efficacy

Prior math grade and grade goal For prior math grade and grade goal the value of the Pearson's correlation is also quite high, namely $r = 0.689$. Also visually it can be seen that the data points seem to be scattered around a line.

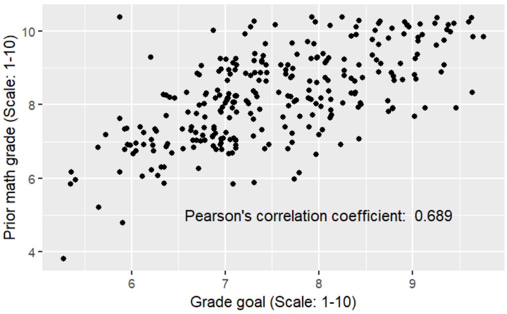

Figure 4.4: Scatter plot of prior math grade and grade goal

Prior math grade and self-efficacy For prior math grade and self-efficacy the value of the Pearson's correlation is also quite high, namely $r = 0.551$. Again the data points seem to follow a line, but this time they appear a little more scattered, which is also reflected in the correlation coefficient.

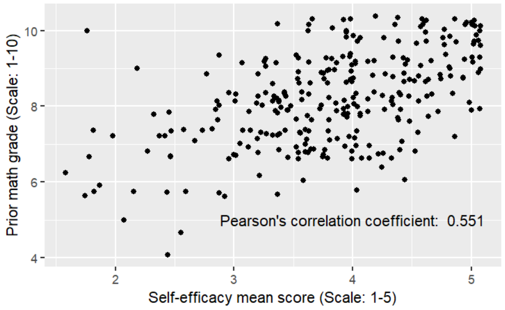

Figure 4.5: Scatter plot of prior math grade and self-efficacy

Hence by looking at these correlation coefficients and plots, it can be concluded that the independent variables are moderately correlated. But according to Tabachnick et al. (2007), since the correlation coefficients are less than 0.90, the assumption is still met.

4.3.3 Separability

Separability is another assumption which can be checked visually. According to Fox (2016), in the case of three independent variables X_1, X_2, X_3 the data are separable if there is a separating plane in the three-dimensional space of the Xs, which separates the outcomes. In order to assess this, a 3D plot can be made, in which the points are coloured according to their outcome category. Hence in figure 4.6, the points are coloured according to the four categories of the dependent variable 'why_quiz'. If the data are separable, the data points of one category would be clustered in a group that can be separated from the other points by a 2-dimensional plane. A visual inspection of all angles of the 3D plot of the dependent variable 'why_quiz' seems to indicate that no such 2-dimensional plane exists and hence that the data are not separable. In figure 4.6 two different angles of the 3D plot are included. With the code included in appendix C, this 3D plot can be generated in R.

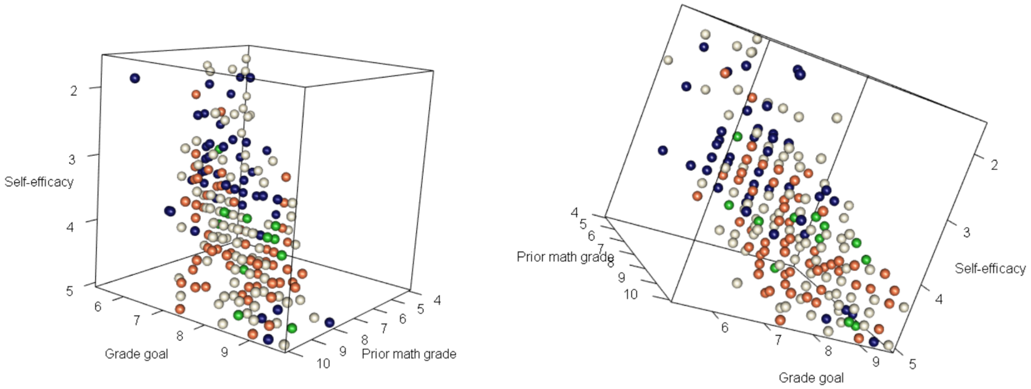

Figure 4.6: Two angles of 3D-plot to assess separability for the dependent variable why_quiz

Chapter 5

Results

To compare the different approaches and packages the results of the multinomial logistic regression with one dependent variable are discussed in this chapter. In the original study eight different variables are included, but discussing the results of all eight, would be very elaborate and not necessary for answering the research questions of the current study. However, the code to perform the analyses is included in Appendix A and all code is made such that one can choose the preferred dependent variable in the first line of the code and then subsequently the analyses and assumption checks are performed for the dependent variable. Hence the results can easily be computed for all other dependent variables.

5.1 Descriptive statistics

Firstly the descriptive statistics of the independent variables are shortly discussed. In the following figures, the histograms are shown with a line indicating the mean value of the independent variables. For grade goal the mean value is given by 7.510 (standard deviation 0.981), with 5.333 as the minimum and 9.677 as the maximum. The histogram indicates that the distribution is positively-skewed, due to the modus being grade 7. For self-efficacy the mean value is given by 3.826 (standard deviation 0.777) with 1.600 as the minimum value and 5.000 as the maximum value. This distribution is negatively-skewed, with the modus being mean score 4. Lastly the mean value of prior math grade is 8.193 (standard deviation 1.208) with 4.000 as the minimum value and 10.000 as the maximum value. This distribution is also negatively-skewed, with the modus being grade 8.

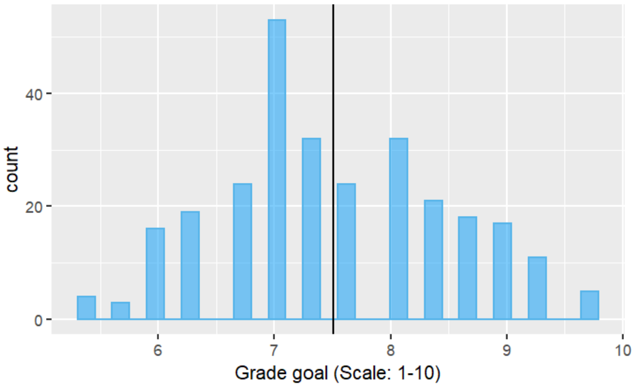

Figure 5.1: Histogram grade goal

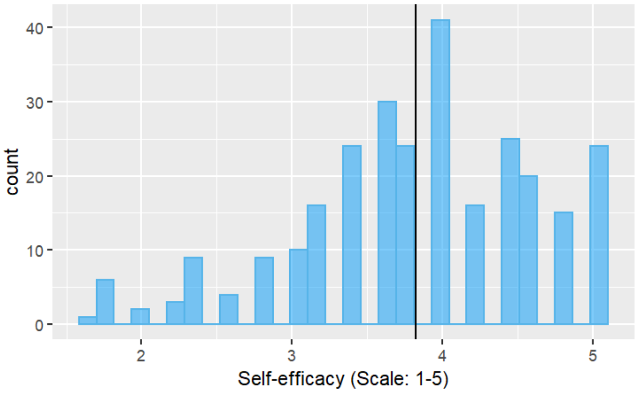

Figure 5.2: Histogram self-efficacy

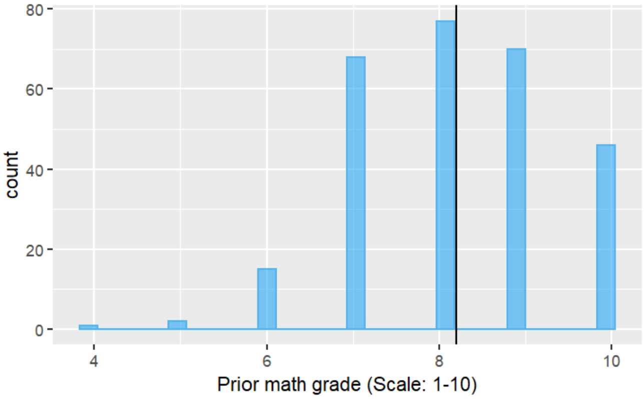

Figure 5.3: Histogram prior math grade

5.2 Elaborate results for one variable

A detailed analyses of the dependent variable 'why_quiz' is included in this section. For 'why_quiz' the following question was included in the survey: "If you quiz yourself while you study (either using a quiz at the end of a chapter, or a practice quiz, or flashcards, or something else), why do you do so?" Four different answers could be chosen, thus there are four different categories of the dependent variable. The percentages of the responses can be found in table 5.1. Most students quiz themselves to figure out how well they have learned the information they are studying. The lowest percentage of students quizzes themselves since they think it is more enjoyable then reading.

I learn more that way than I would through rereading	27.60%
To figure out how well I have learned the information I'm studying	43.40%
I find quizzing more enjoyable than reading	7.53%
I usually do not quiz myself	21.51%

Table 5.1: Possible answers and response percentages

Multinomial logistic regression was used to analyze the relationship between the dependent variable 'why_quiz' and the independent variables self-efficacy, grade goal and prior math grade. The analysis was performed with three different R packages, namely `mlogit`, `VGAM` and `UPG`. The results are discussed in the following sections. The command 'summary' was used in order to get the main results from the packages, which are often firstly obtained when running an analysis.

5.2.1 `mlogit` (frequentist)

The first results are from the `mlogit` package, which is a package that adopts the frequentist approach. Moreover the package is built upon the random utility model. The summary of the output of this analysis can be found in appendix D and will be outlined in this section.

In the summary firstly the frequencies of the alternatives are given in a table, similar to table 5.1. Moreover it is mentioned that the Newton-Raphson method needed 5 iterations to estimate the coefficients, taking less than 1 second. Additionally the gradient of the last iteration of the Newton-Raphson method is given, which is close to zero ($1.63 \cdot 10^{-7}$). Subsequently the value of the log-likelihood function is given, which is the value that is maximized by the Newton-Raphson method, as described in section 3.3.1 The value of the log-likelihood in the current model is -333.65. Furthermore the following estimates are reported:

McFadden R^2

In linear regression, when the dependent variable is continuous, R^2 is called the coefficient of determination. This can be interpreted as the proportion of the variance explained by the model [van der Vaart et al., 2017]. To determine R^2 a numerical estimate of the dependent variable is required (since the residual sum of squares and total sum of squares are needed). However, in logistic regression the dependent variable is categorical and therefore this cannot be calculated. Therefore McFadden (1974), who won the Nobel Prize for his work on random utility models, introduced R^2_{McFadden} to assess the proportion of variance explained by the model (originally denoted by him as ρ^2) [McFadden, 1974]. This value is defined by the following formula:

$$R^2_{\text{McFadden}} = 1 - \frac{\log(L_c)}{\log(L_{null})}$$

where L_c denotes the likelihood value from the current model and L_{null} denotes the likelihood value from the null model with only an intercept. The null model is the baseline model, which only includes a constant value for the intercept. To determine this constant, we look at the frequencies of the outcomes. Our best guess of the outcome will be that the category with the largest frequency is chosen. Therefore, the baseline model includes the outcome that occurs most often as the intercept and further all coefficients equal to zero [Field et al., 2012].

For the original R^2 value, a value close to 0 indicates that the model does not explain much variance, while a value close to 1 indicates that the proportion of variance explained by the model is large. However McFadden states in 1977 that "Those unfamiliar with R^2_{McFadden} should be forewarned that its values tend to be considerably lower than those of the R^2 index and should not be judged by the standards for a "good fit" in ordinary regression analysis. For example, values of 0.2 to 0.4 for R^2_{McFadden} represent an excellent fit." [McFadden, 1977] However, in the current model the value of R^2_{McFadden} is 0.031, indicating that the model does not explain much variance.

Likelihood ratio χ^2-test

In addition the likelihood ratio χ^2 test is reported, which also compares the fit of the fitted model and the null model. This is computed by

$$\chi^2 = -2(\log(L_{null}) - \log(L_c))) = 2\log(L_c) - 2\log(L_{null})$$

where L_c denotes the likelihood value from the current model and L_{null} denotes the likelihood value from the null model with only an intercept. A statistically significant value of the likelihood ratio test indicates that at least one of the regression coefficients in the model is not equal to zero. In the current model $\chi^2 = 26.188$, with $p = 0.002$, indicating that at least one of the regression coefficients in the model is not equal to zero. Hence the model including the predictors is statistically significantly better than the model without those predictors.

Wald statistic and estimated coefficients

For each estimated coefficient the null hypothesis $H_0 : \beta_{kj} = 0$ is tested by calculating the Wald statistic [Fox, 2016], denoted by z (see section 3.3.2 for the definition of this statistic). The resulting $z-$statistic and

$p-$value are reported in the table. These values tell us whether the coefficient for that predictor is significantly different from zero. If it is, we can assume the predictor is making a statistically significant contribution to the prediction of the outcome [Field et al., 2012].

	Estimate β_{kj}	Standard Error	$z-$value	$p-$value	
Intercept(2) β_{02}	1.430	1.246	1.147	0.251	
Intercept(3) β_{03}	−3.656	2.148	−1.702	0.089	
Intercept(4) β_{04}	3.985	1.489	2.676	0.007	*
Self-efficacy(2) β_{12}	−0.747	0.286	−2.609	0.009	*
Self-efficacy(3) β_{13}	−0.366	0.502	−0.729	0.466	
Self-efficacy(4) β_{14}	−0.899	0.328	−2.738	0.006	*
Prior math grade(2) β_{22}	0.011	0.168	0.067	0.947	
Prior math grade(3) β_{23}	−0.159	0.284	−0.558	0.577	
Prior math grade(4) β_{24}	−0.068	0.200	−0.339	0.735	
Grade goal(2) β_{32}	0.246	0.234	1.052	0.293	
Grade goal(3) β_{33}	0.664	0.396	1.678	0.093	
Grade goal(4) β_{34}	−0.037	0.288	−0.127	0.899	

Table 5.2: Coefficients `mlogit` with category 1 as the reference category, * denotes statistical significance

To obtain more insight in the meaning of these coefficients, we look at the probability function and odds ratio:

Interpretation of the results: probability function

Firstly the obtained estimates can be filled in as the coefficients β_{kj} in the probability equation, with j denoting the categories of the dependent variable to which the reference category is compared and k denoting the independent variable. Category 1 is the reference category, hence $\beta_{01} = \beta_{11} = \beta_{21} = \beta_{31} = 0$

$$\pi_{12} = \frac{e^{\beta_{02}+\beta_{12}X_{11}+\beta_{22}X_{12}+\beta_{32}X_{13}}}{1+\sum_{l=1}^{m-1} e^{\beta_{0l}+\beta_{1l}X_{11}+\beta_{2l}X_{12}+\beta_{3l}X_{13}}} \quad \text{for } j = 2 \tag{5.1}$$

$$\pi_{13} = \frac{e^{\beta_{03}+\beta_{13}X_{11}+\beta_{23}X_{12}+\beta_{33}X_{13}}}{1+\sum_{l=1}^{m-1} e^{\beta_{0l}+\beta_{1l}X_{11}+\beta_{2l}X_{12}+\beta_{3l}X_{13}}} \quad \text{for } j = 3 \tag{5.2}$$

$$\pi_{14} = \frac{e^{\beta_{04}+\beta_{14}X_{11}+\beta_{24}X_{12}+\beta_{34}X_{13}}}{1+\sum_{l=1}^{m-1} e^{\beta_{0l}+\beta_{1l}X_{11}+\beta_{2l}X_{12}+\beta_{3l}X_{13}}} \quad \text{for } j = 4 \tag{5.3}$$

$$\pi_{11} = 1 - \sum_{j=1}^{m-1} \pi_{ij} \quad \text{for category 1} \tag{5.4}$$

Subsequently the model can be used to predict probabilities. Suppose that a person has the same scores as the first participant of the PRIME study who chose category 2 of the dependent variable. Now we are going to see how well the model predicts this outcome, by filling in the coefficients. Therefore we look at the probability of choosing each category of the dependent variable, according to the model. The values for the independent variables of the first participant are: $X_{11} = 4$ (self-efficacy), $X_{12} = 8$ (prior math grade) and $X_{13} = 7.333$ (grade goal). The estimated coefficient β_{kj} are used to predict the probabilities.

$$\pi_{12} = \frac{e^{\beta_{02}+\beta_{12}X_{11}+\beta_{22}X_{12}+\beta_{32}X_{13}}}{1+e^{\beta_{01}+\beta_{11}X_{11}+\beta_{21}X_{12}+\beta_{31}X_{13}}+e^{\beta_{02}+\beta_{12}X_{11}+\beta_{22}X_{12}+\beta_{32}X_{13}}+e^{\beta_{03}+\beta_{13}X_{11}+\beta_{23}X_{12}+\beta_{33}X_{13}}} \quad \text{for } j = 2 \tag{5.5}$$

$$
\begin{aligned}
&= 0.426 \\
\pi_{13} &= \frac{e^{\beta_{03}+\beta_{13}X_{11}+\beta_{23}X_{12}+\beta_{33}X_{13}}}{1+e^{\beta_{01}+\beta_{11}X_{11}+\beta_{21}X_{12}+\beta_{31}X_{13}}+e^{\beta_{02}+\beta_{12}X_{11}+\beta_{22}X_{12}+\beta_{32}X_{13}}+e^{\beta_{03}+\beta_{13}X_{11}+\beta_{23}X_{12}+\beta_{33}X_{13}}} \quad \text{for } j=3 \\
&= 0.067 \\
\pi_{14} &= \frac{e^{\beta_{04}+\beta_{14}X_{11}+\beta_{24}X_{12}+\beta_{34}X_{13}}}{1+e^{\beta_{01}+\beta_{11}X_{11}+\beta_{21}X_{12}+\beta_{31}X_{13}}+e^{\beta_{02}+\beta_{12}X_{11}+\beta_{22}X_{12}+\beta_{32}X_{13}}+e^{\beta_{03}+\beta_{13}X_{11}+\beta_{23}X_{12}+\beta_{33}X_{13}}} \quad \text{for } j=4 \\
&= 0.201 \\
\pi_{11} &= 1-0.426-0.067-0.201 \quad \text{for } j=1 \\
&= 0.306
\end{aligned}
$$

Hence the model would also estimate that the probability of choosing category 2 is the biggest, namely 0.426. When comparing the first six participants in the same way, we find the following probabilities:

	π_{i1}	π_{i2}	π_{i3}	π_{i4}	Actual chosen category	Highest probability for chosen category?
$i=1$	0.306	0.426	0.067	0.201	2	✓
$i=2$	0.285	0.426	0.091	0.198	2	✓
$i=3$	0.291	0.441	0.079	0.189	2	✓
$i=4$	0.130	0.470	0.026	0.373	4	
$i=5$	0.308	0.393	0.107	0.191	2	✓
$i=6$	0.274	0.398	0.042	0.286	1	

Table 5.3: Predicted probabilities of choosing each category for the first six participants of the PRIME study, where $i \in 1, \dots n$ denotes the particpant

For the participants that choose category 2, the probabilities of choosing this category is also the highest. However, even though this is the highest probability, choosing the other categories also still have considerably high probabilities. Moreover, for the participants who did not choose category 2, this category also has the highest predicted probability, which is not in line with their actual behavior. These imperfect predictions can be explained by the fact that the model is not very well fitted to the data, since only three coefficients are statistically significant and the R^2_{McFadden} is not very high.

Interpretation of the results: odds ratio

The odds ratio is commonly only interpreted for the values that are significant [Tabachnick et al., 2007]. The value of the odds ratio, is the exponential of the coefficient and is an indicator of the change in odds resulting from a unit change in the predictor. In order to understand the odds ratio, we look again at the log odds, which are given by:

$$\log \frac{\pi_{ij}}{\pi_{im}} = \beta_{0j} + \beta_{1j}X_{i1} + \cdots + \beta_{kj}X_{ik} \quad \text{for } j = 1, \dots, m-1 \tag{5.6}$$

Hence regression coefficients represent the effect on the log-odds of membership in category j versus the baseline category m. It should be noted that the log-odds can be determined for any pair of categories. The log-odds can be exponentiated, giving

$$
\begin{aligned}
\frac{\pi_{ij}}{\pi_{im}} &= \exp(\beta_{0j} + \beta_{1j}X_{i1} + \cdots + \beta_{kj}X_{ik}) \quad \text{for } j = 1, \dots, m-1 && (5.7) \\
&= \exp(\beta_{0j})\exp(\beta_{1j})X_{i1} \cdots \exp(\beta_{kj})X_{ik} \quad \text{for } j = 1, \dots, m-1 && (5.8)
\end{aligned}
$$

The exponentiated coefficients $\exp(\beta_{kj})$ are called the ‘odds ratios’ since they represent the ratio of the odds of a response at two $X-$values: if X_{ik} is one unit larger, this multiplies the odds of the outcome category j occurring by $\exp(\beta_{kj})$, compared to the baseline category. If the odds ratio is greater than 1 it indicates that as the predictor increases, the odds of the outcome occurring increase. Conversely, a value less than 1 indicates

	Odds ratio		Odds ratio
Intercept(2) β_{02}	4.177	**Prior math grade(2)**β_{22}	1.011
Intercept(3) β_{03}	0.026	**Prior math grade(3)**β_{23}	0.853
Intercept(4) β_{04}	53.797	**Prior math grade(4)**β_{24}	0.934
Self-efficacy(2) β_{12}	0.474	**Grade goal(2)**β_{32}	1.278
Self-efficacy(3 β_{13}	0.693	**Grade goal(3)**β_{33}	1.942
Self-efficacy(4) β_{14}	0.407	**Grade goal(4)**β_{34}	0.964

Table 5.4: Odds ratio for the coefficients computed with `mlogit`, category 1 as the reference category

that as the predictor increases, the odds of the outcome occurring decreases. The closer the odds ratio is to 1, the smaller the effect. The odds ratios are as follows:

Hence for example, keeping all other coefficients equal, then a one unit increase in self-efficacy decreases the odds of choosing category 2 (compared to category 1) by 0.474.

5.2.2 `VGAM` (frequentist)

The results for `mlogit` have been discussed quite elaborately. Another package in line with the frequentist apprach is the `VGAM` package. This package contrarily has the generalized linear model as the theoretical background. The results of `VGAM` and `mlogit` are actually very similar, so these will only be discussed generally. The results are included in appendix E.

It is firstly reported that 5 Fisher scoring iterations are performed. The log-likelihood $\log L_c$ is -333.655, the same as in the `mlogit` package. Besides this the residual deviance is given, instead of R^2_{McFadden} in `mlogit`. Residual deviance is computed as [Fox, 2016]

$$\text{residual deviance} = -2\log L_c = 667.309$$

	Estimate β_{kj}	Standard Error	$z-$value	$p-$value	
Intercept(2) β_{02}	1.430	1.246	1.147	0.251	
Intercept(3) β_{03}	−3.656	2.148	NA	NA	
Intercept(4) β_{04}	3.985	1.489	2.676	0.007	*
Self-efficacy(2) β_{12}	−0.747	0.286	−2.609	0.009	*
Self-efficacy(3) β_{13}	−0.366	0.502	−0.729	0.466	
Self-efficacy(4) β_{14}	−0.899	0.328	−2.738	0.006	*
Prior math grade(2) β_{22}	0.011	0.168	0.066	0.947	
Prior math grade(3) β_{23}	−0.159	0.284	−0.558	0.577	
Prior math grade(4) β_{24}	−0.068	0.200	−0.339	0.735	
Grade goal(2) β_{32}	0.246	0.234	1.052	0.293	
Grade goal(3) β_{33}	0.664	0.396	1.678	0.093	
Grade goal(4) β_{34}	−0.037	0.288	−0.127	0.899	

Table 5.5: Coefficients `VGAM` with category 1 as the reference category, * denotes statistical significance

The coefficients are presented again. These are displayed in table 5.5 and it can be noted that these are approximately the same as the coefficients and estimates obtained by `mlogit`. However, the value for the Wald statistic and $p-$value is NA due to the Warning: Hauck-Donner effect detected in the following estimate(s): Intercept(2). The Hauck Donner-effect is a shortcoming of the Wald test which occurs when a Wald test statistic is no longer monotone increasing as a function of increasing distance between the parameter estimate and the null value [Yee, 2021]. When this effect occurs, it might indicate that the data is separable [Yee, 2015], violating one of the assumptions of multinomial logistic regression. This was not directly visible in the 3D plots assessing this assumption.

The interpretation of these coefficients is equivalent to the interpretation of the coefficients in mlogit.

5.2.3 UPG (Bayesian)

The third package that is used to conduct the multinomial logistic regression is the UPG package, which is built upon the random utility model. UPG is in line with the Bayesian approach and therefore a prior distribution can be specified for each coefficient β_{kj}. However, in the official documentation of the UPG package, only one sentence is devoted to the prior distribution and not much emphasis is put on this aspect of Bayesian analysis. Moreover, no guidelines of choosing a prior are given. Instead, a default prior is given and since no real information or knowledge exists to motivate another choice, the default option is also used in the current study. This default prior is a normal distribution, for which the default parameter values are $A_0 = 4$ and $B_0 = 4$.

Subsequently the MCMC algorithms compute the posterior distribution of each coefficient. Every run of the Bayes model gives different values. One can specify how many draws and burn-in one wants in the UPG package. The burn-in draws from the posterior distribution are meant as a ‘training phase’, aimed to already let the chain converge to the posterior distibution. In the package manual of UPG it is stated that for a model run it is good to use 10000 as the draws and 2000 as the burn-in [Zens et al., 2021]. This is done in the current study. In the summary of the output it is noted that the MCMC sampling took a total of 35.95 seconds. The results of one MCMC sampling instance are given, but these do not vary a lot when doing the analysis again. To support this claim, in appendix G the outcomes of five model runs are given and it can be seen that the results are approximately the same. Therefore it can be concluded that the Markov chains indeed converge to the posterior distributions.

	Mean estimate β_{kj}	SD	Q2.5	Q97.5	95% CI excl. 0
Intercept(2) β_{02}	1.089	0.983	−0.833	3.206	
Intercept(3) β_{03}	−1.470	1.371	−4.235	1.123	
Intercept(4) β_{04}	2.754	1.100	0.626	4.942	*
Self-efficacy(2) β_{12}	−0.753	0.255	−1.261	−0.258	*
Self-efficacy(3) β_{13}	−0.360	0.440	−1.223	0.499	
Self-efficacy(4) β_{14}	−0.946	0.302	−1.551	−0.363	*
Prior math grade(2) β_{22}	0.022	0.152	−0.281	0.317	
Prior math grade(3) β_{23}	−0.270	0.267	−0.810	0.249	
Prior math grade(4) β_{24}	−0.054	0.174	−0.400	0.282	
Grade goal(2) β_{32}	0.282	0.211	−0.127	0.698	
Grade goal(3) β_{33}	0.495	0.350	−0.175	1.190	
Grade goal(4) β_{34}	0.137	0.248	−0.348	0.625	

Table 5.6: Coefficients UPG with category 1 as the reference category

The summary output of the UPG package provides estimates by giving the mean of the posterior distribution and its corresponding standard error, credible regions and even a asterix when a credible region does not contain the value 0. This output hence seems very similar to the output of frequentist logistic regression, especially since the coefficients with an asteriks are also the ones that are statistically significant in the frequentist method. The interpretation of the mean estimates is hence also similar to the interpretation of the coefficients in the frequentist

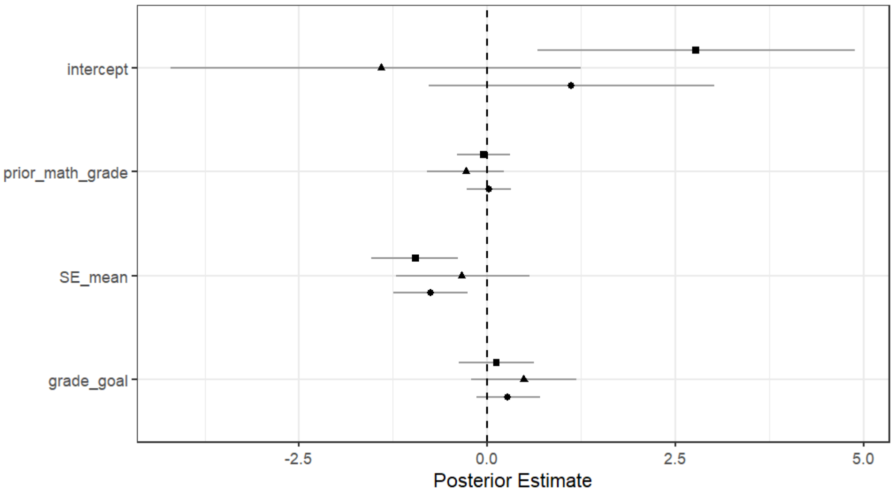

Figure 5.4: Posterior estimates of `UPG`

method.

A plot of the posterior estimates and credible regions is also provided to visualize the results, see figure 5.4. No other statistics or values are given in the default summary output of `UPG`.

Chapter 6

Comparison

In the current chapter the Bayesian and the frequentist approach to multinomial logistic regression will be compared. Moreover in the second part of this chapter the specific differences in the use of the three R packages will be outlined.

6.1 Comparison Bayesian and Frequentist

A first difference is the estimation time in the frequentist and the Bayesian approach. Both frequentist packages (`mlogit` and `VGAM`) estimate the model in less than 1 second, while the MCMC estimation of the Bayesian `UPG` package takes approximately 35 seconds. Moreover the parameters in the frequentist package are fixed and hence the same every time a model is computed, while the posterior distributions computed in the Bayesian approach are different in every run (although the differences get neglectable when enough samples are drawn, as is done in the current study).

Another difference is that in the Bayesian approach a prior distribution can be chosen for each coefficient. However, not much information is available about choosing priors in multinomial logistic regression and therefore the default prior is chosen in this study. In the following sections the model fit will be discussed and also the interpretation of the results of both approaches.

6.1.1 Model fit comparison

To compare the models two criterions are used. Firstly the Akaike information criterion (AIC) and secondly the Bayesian information criterion (BIC). Both are commonly used to compare models. They balance the goodness of fit by the number of parameters. A lower AIC or BIC value indicates a better fit. They are computed in the following way [Fox, 2016]:

$$\text{AIC} = -2\log(L) + 2k$$

$$\text{BIC} = -2\log(L) + 2\log(n)k$$

where $\log(L)$ is the maximized log-likelihood under the model, n is the number of recorded measurements and k is the number of estimated parameters.

	AIC	**BIC**
`mlogit` (frequentist)	703.309	734.884
`VGAM` (frequentist)	703.309	734.884
`UPG` (Bayesian)	468.399	499.974

As can be seen in the table 6.1.1, both `mlogit` & `VGAM` generate the same values for the AIC and BIC, which comes as no surprise since the estimated coefficients are also approximately the same. However, the `UPG` generates much lower AIC and BIC, indicating that the model is a better fit to the data.

6.1.2 Confidence and credible intervals

The frequentist packages do not provide confidence intervals in the summary, however these can easily be extracted. The Bayesian `UPG` package provides the credible intervals. The credible intervals are very much in line with the confidence intervals, as shown in the following plot:

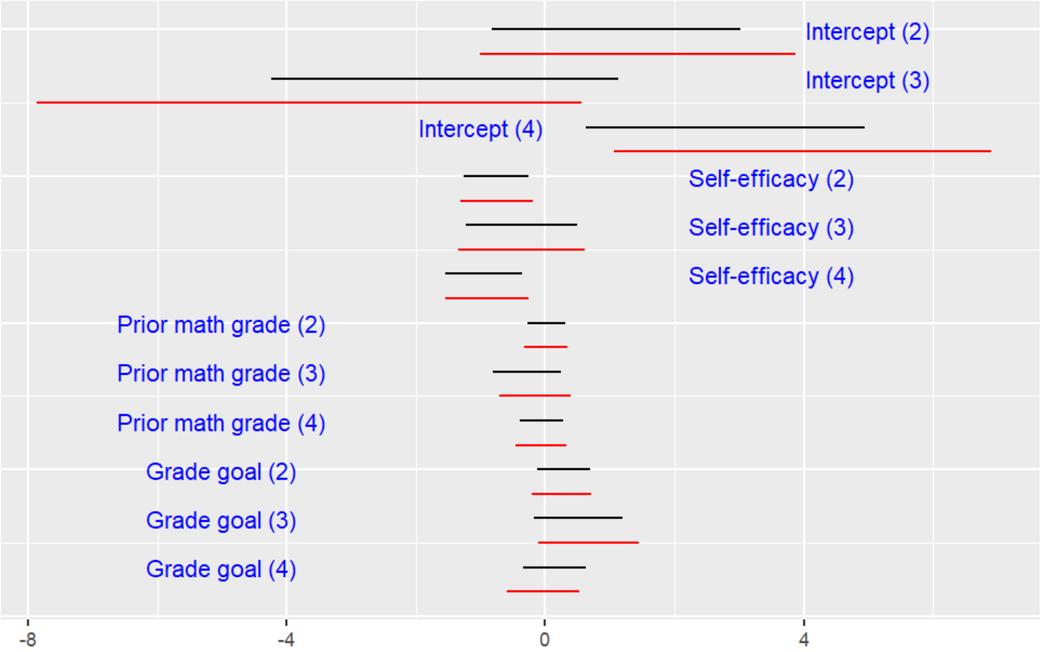

Figure 6.1: 95% confidence (black, frequentist) and 95 % credibile intervals (red, Bayesian) for the dependent variable why_quiz

Therefore the results of the frequentist approach and Bayesian approach seem quite similar, especially since the statistically significant coefficients of the frequentist approach correspond to the coefficients of the Bayesian approach that have a credible interval which does not contain zero (and are hence indicated with an asteriks in the `UPG` package, just like statistical significance is always indicated).

However, the interpretation of the confidence interval and credible interval is considerably different. In the frequentist approach it is assumed that the parameter value of the coefficients is fixed and can be estimated by an experiment. In this approach, a confidence interval is based on repeated sampling theory, which means that if this experiment is repeated many times, then 95 percent of the confidence intervals of exactly the same experiment will capture the 'true' value of the coefficient [van de Schoot et al., 2021]. In the Bayesian approach it is assumed that the coefficient has a certain distribution, which is a random variable and hence not fixed. The credible interval denotes the interval within which the value of the coefficient falls with 95% probability and is computed by looking at the posterior distribution. The Bayesian interpretation of the credible interval is more intuitive than the frequentist interpretation of the confidence interval. Furthermore the posterior distribution can provide much more information about the coefficients, compared to the point estimates of the frequentist approach.

6.1.3 Posterior distribution

Hence an advantage of the Bayesian approach is that the posterior distributions of all simulated Markov Chains can also be extracted from the model. The model contains a $(k+1) \cdot j \cdot \#draws$-matrix with all estimated parameters β_{kj}, where k denotes the number of independent variables and j the number of categories of the dependent variable. Thus for the variable 'why_quiz' a $4 \cdot 4 \cdot 10000$-matrix is extracted. This can be rearranged such that the `bayesplot` package can be used for visualizations of the Markov Chains. The code for rearranging and visualization can be found in appendix H.

This can give for example histograms, such as figure 6.2. This provides the posterior distribution of the estimates. It can be seen that for chain 1 (representing category 2) and chain 3 (representing category 4) the effect of self-efficacy on the probability of the outcome seems to have a negative effect for choosing these two categories. Moreover it can be seen that for the second chain (representing category 3), self-efficacy is much more spread and closer to zero. Lastly for the fourth chain (representing the baseline category, category 1), the beta is zero, which is indeed assumed in the model. The histograms for the other independent variables are included in appendix H.

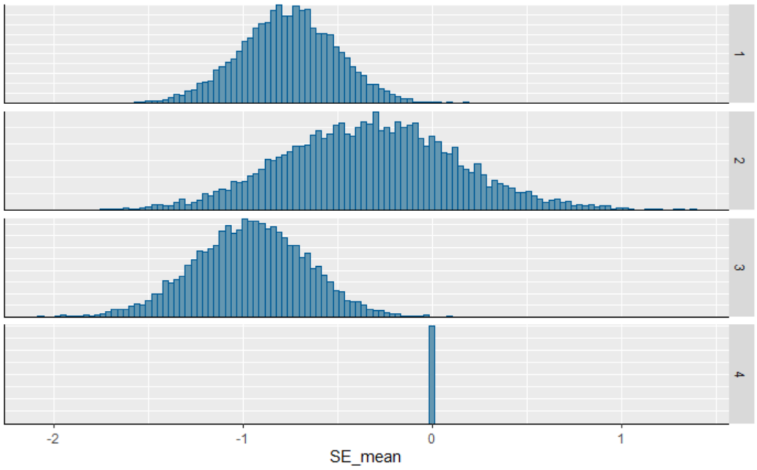

Figure 6.2: Histograms of posterior distribution Self-efficacy, histograms are often used to visualize the posterior distribution

Moreover scatterplots can be created, which can also be found in appendix H. As an example we include the following scatterplots:

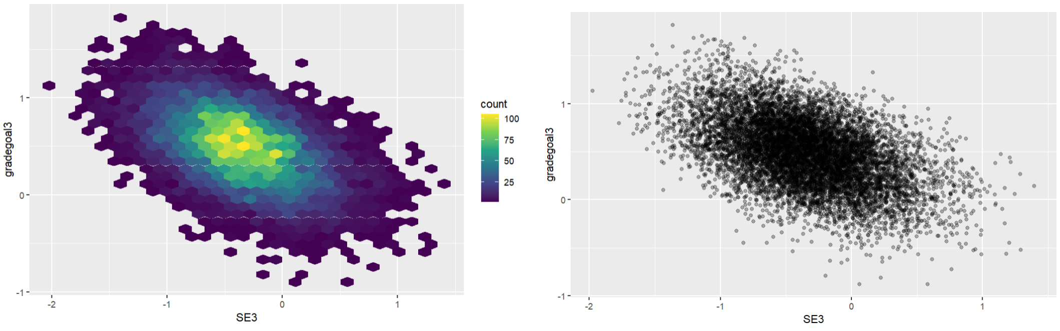

Figure 6.3: Self-efficacy(3) and grade goal(3)

In these scatterplots it is shown that when doing a binomial logistic regression comparing category 3 to category 1 (baseline category), it seems that there is a negative relationship between the coefficients of self-efficacy and grade goal. Meaning that if the effect of self-efficacy is lower, the effect of grade goal is higher.

All in all, the posterior distribution that is computed by the Bayesian approach, can provide more insight in the actual distribution of the data and relations between variables. Additionally, the model fit of the Bayesian approach is considerably better than the frequentist models, as indicated by the AIC and BIC values. Lastly the confidence and credible intervals are very similar, but the interpretation of these is considerably different, with the Bayesian interpretation being more intuitive.

6.2 Comparison packages

Three different R were used in the current study: `mlogit`, `VGAM` and `UPG`. Of course the frequentist packages differ considerably in their results compared to the Bayesian package, since both adopt a different approach. However the differences between the frequentist and Bayesian approach have already been discussed in the previous section, hence these will not be detailed anymore. Moreover it should be noted that the results of both frequentist packages are very similar, although both report different model statistics. In table 6.1 the main differences in the packages are summarized.

	Frequentist or Bayesian approach	**Based on GLM or RUM**	**Reordering data required**	**Detection Hauck-Donner effect**	**Inclusion different variables**	**Estimation time**
`VGAM`	Frequentist	GLM	No	✓	✗	< 1 seconds
`mlogit`	Frequentist	RUM	Yes	✗	✓	< 1 seconds
`UPG`	Bayesian	RUM	Yes	Not relevant*	✗	≈ 35 seconds**

Table 6.1: Comparison of the three packages

* The detection of the Hauck-Donner effect is not relevant for the `UPG` package since the Hauck-Donner effect occurs when computing the Wald statistic, which is only used in the frequentist approach to statistics. The Hauck-Donner effect can indicate separability in the data. In the paper introducing the `UPG` package it is mentioned that the applied method solves the issue of perfect separation [Zens et al., 2021].

** The estimation time of the `UPG` package is based on a run with 2000 burn-in and 10000 draws.

6.2.1 Different variables

The inclusion of different variables is explained in this section. The random utility model is very well suited for including other types of variables, which is thus included in the package based on the random utility model: `mlogit`. This might be very interesting in some specific research questions. As mentioned before, the utility function is a sum of a observed (V_{ij}) and unobserved (ε_{ij}) part. In the previous sections the observed part of the utility function, V_{ij}, was not yet specified. This is a linear combination of independent variables. Three types of variables can be included in the random utility model, in addition to the intercept [Croissant, 2010], namely:

1. Alternative specific variables with a generic coefficient. These are variables that are specific to an alternative, for example when you are comparing travel modes, the price per alternative travel mode differs and is hence alternative specific.

2. Alternative specific variables with alternative specific coefficients. These are variables that are also specific to an alternative, but also have an different impact. For example when comparing travel modes, this can be the travel time. The effect that travel time has on the utility function can be different across alternatives: ten minutes in the car can have a different utility effect than ten minutes in a train.

3. Individual specific variables with alternative specific coefficients. These are variables that are specific for an individual, hence when comparing travel modes this could be the age of the traveler.

In the generalized linear model, and hence in packages `VGAM` and `UPG`, the inclusion of these other types of variables is not introduced and it seems like only individual specific variables are possible to include. In the current study, this was also done in the package `mlogit`, but the possibility of including more variables can be very useful in some study designs.

All in all, the main differences between the packages are: need for reordering the data, detection of the Hauck-Donner effect and the inclusion of different types of variables. Since all packages have their shortcomings and advantages, combining them when conducting a multinomial logistic regression could be desirable.

Chapter 7

Discussion and conclusion

The aim of the current study was to compare the two main approaches of statistics, namely Bayesian and frequentist statistics, when performing a multinomial logistic regression. In addition, the aim was to compare three different R packages for performing such a multinomial logistic regression. Lastly an aim was to shed some more light on the 'black box' that statistical programs sometimes seem to be: requiring input and producing standard output. In order to do so, firstly the theory behind multinomial logistic regression was outlined based on two different models: the generalized linear model and the random utility model. Both models generated an equivalent probability function, which is shown in section 2.3. Subsequently this probability function is used in both the frequentist and the Bayesian approach, for which the methods of estimating coefficients were outlined in section 3. Following this the PRIME study was introduced, along with data preparation and assumption checks. In the fifth chapter, the results of the `mlogit`, `VGAM` and `UPG` are described, along with an interpretation of the coefficients. Lastly in the sixth chapter the Bayesian and frequentist approach to multinomial logistic regression are compared. In addition also a comparison of the use of the different packages is given, which might serve as a useful guide for researchers who want to choose a package for conducting their analyses.

In comparing the frequentist and Bayesian approach, the first thing that is noticed is that the point estimates and intervals provided by both methods are actually quite comparable. However, the interpretation of the Bayesian credible interval is more intuitive than the frequentist confidence interval. Furthermore the $p-$value is often taken as some measure for 'certainty', while it cannot be interpreted as such. As McElreath (2020) notes: "the $p-$value plays into the human need for certainty, but it may be time for both scientists and the public to embrace the comfort of being unsure". In my opinion the Bayesian statistics reflects this unsureness more, since the estimates of coefficients are also not seen as fixed values but rather as random variables. However, on the other hand, Bayesian statistics can be seen as more subjective, since the prior can be chosen by the researcher, and can influence the results.

When comparing the three packages, it is clear that both frequentist packages (`mlogit` and `VGAM`) give almost the same results. However, the theoretical background of both differs: `mlogit` is based on the random utility model, while `VGAM` is based on the generalized linear model. This provides a subtle difference between both packages and therefore it is suggested to use the random utility model in problems that include students, since this theory is mostly based on modelling choices of customers (or in this case: choices of students). An advantage of RUM is that alternative specific variables can be included, which can provide better models. However, to achieve this specific questions should be included into questionnaires to assess alternative specific variables. The generalized linear model is more common for making predictions and assessing relationships between variables. The Bayesian package (`UPG`) does not provide an elaborate summary of the results, but when playing around in R with the computed posterior distribution, interesting results can be visualized, such as scatterplots and histograms of the posterior distributions (see appendix H).

7.1 Limitations and future research

A limitation is that there was no correction for imbalanced data in the current study. Some categories of dependent variables were less frequently chosen by participants, leading to imbalanced data. For example, the analyzed dependent variable ‘why_quiz’ has one category that less than 8% of the participants chose. Both in the frequentist and Bayesian approach there are methods to correct for this imbalance. In the frequentist approach it is often suggested to incorporate a penalized likelihood, while the Bayesian equivalent consists of choosing a regularizing prior to correct for the implausibility of some parameter values [McElreath, 2020]. The current study aimed to thoroughly comprehend the theory behind multinomial logistic regression, and as such, incorporating penalized likelihood and a regularizing prior were not included in the scope. Future research can expand on this theoretical framework.

Another suggestion for future research is to run the analysis for all dependent variables and propose practical implications based on these results. The current study is mostly focused on the statistical background of multinomial logistic regression and therefore less on the practical implications for PRIME. In addition, the analysis of only one dependent variable is included. However, with the R code that is also included in appendix A, the analyses for all dependent variables can be run and interpreted.

Additionally a suggestion for future research is to not compute averages for the variables ‘grade goal’ and ‘self-efficacy’. By averaging these variables, information in variance gets lost. Therefore it might be better to not include the average, but instead for example use multilevel modelling [McElreath, 2020].

Lastly, for future research adapting the Bayesian approach it is suggested to do proper sensitivity analyses for different priors. In doing so, the ‘subjectivity’ of Bayesian analysis might be considerably less. Moreover it might be good to include informative priors. In the current study not much information was available to motivate the choice of particular priors, hence the default priors were chosen. However, when a new study would be conducted, the current posterior distributions of the coefficients could be used as prior distributions.

Appendix A

Appendix: Main program to run analysis one dependent variable

```r
library(car)
library(MASS)
library(knitr)
library(readxl)
library(stringr)
library(survival)
library(plotly)
library(EnvStats)
library(dplyr)
library(arsenal)
library(psych)
library(summarytools)
library(stargazer)
library(mlogit)
library(ggplot2)
library(scatterplot3d)
library(GGally)
library(Formula)
library(UPG)
library(VGAM)
library(bamlss)
library(bayesplot)
library(rstanarm)
library(base)
library(ltm)
library(rgl)
library(magick)
library(hexbin)

#####data loading and preparation
my_data_all = read_excel("C:/Users/sterr/OneDrive/Documents/Studie/Technische wiskunde/BEP/R/StudyHabits_286v3.xls")

#remove columns with information about study strategies
my_dataall = my_data_all %>%
  select(16, 51:66)

#remove rows with missing data (total is 286, of which 7 incomplete questionnaires)
my_data = data.frame(my_dataall[complete.cases(my_dataall[ , 16]),])

#format all dependent variables as categorical variables
dv = c("study_next","study_more_for","material_more_than_once","convinced_do_next","why_quiz","pattern_of_study","time_of_day","time_of_day_belief")
for (variable in dv) {mydatavar = paste("my_data$",variable,sep="")
  mydatavar = factor(mydatavar)}

```

```r
#Create variable for grade goal as the third independent variable
my_data$grade_goal = (as.numeric(my_data$grade_aim)+as.numeric(my_data$grade_expected)+as.numeric(my_data$grade_lowest_satisfied))/3

#format all independent variables as numeric (categories undesirable for multinomial logistic regression)
ivnumbers = c(1,2,18)
for (iv in ivnumbers) {
  my_data[, iv] <- as.numeric(my_data[, iv])
}

#####Define dependent variables and independent variables
dependent_var = c("Study next", "Study more for","Material more than once","Why quiz","Convinced do next","Pattern of study","Time of day","Time of day belief")
independent_var = c("Prior math grade","Grade goal","Self-efficacy score")

#check class of other variables -> change class from character to numeric
sapply(my_data, class)
my_data$age= as.numeric(my_data$age)

#####Descriptive statistics

ggplot(my_data, aes(grade_goal)) +
  geom_histogram(color = "#56B4E9", fill = "#0099F8", alpha = 0.5) +
  geom_vline(aes(xintercept = mean(grade_goal)), color = "#000000", size = 0.5) + labs(x = "Grade goal (Scale: 1-10)")

summary(my_data$grade_goal)
sd(my_data$grade_goal)

ggplot(my_data, aes(SE_mean)) +
  geom_histogram(color = "#56B4E9", fill = "#0099F8", alpha = 0.5) +
  geom_vline(aes(xintercept = mean(SE_mean)), color = "#000000", size = 0.5) + labs(x = "Self-efficacy (Scale: 1-5)")

summary(my_data$SE_mean)
sd(my_data$SE_mean)

ggplot(my_data, aes(prior_math_grade)) +
  geom_histogram(color = "#56B4E9", fill = "#0099F8", alpha = 0.5) +
  geom_vline(aes(xintercept = mean(prior_math_grade)), color = "#000000", size = 0.5) + labs(x = "Prior math grade (Scale: 1-10)")

summary(my_data$prior_math_grade)
sd(my_data$prior_math_grade)

########ANALYSIS ONE VARIABLE#######
###Choose variable (study_next, study_more_for, material_more_than_once, why_quiz, convinced_do_next, pattern_of_study, time_of_day, time_of_day_belief):
dependentvar = "why_quiz"
###The code is written such that the baseline category is always category 1, in all three packages.


####mlogit package
mlogitdata = my_data %>%
  select(1,2,3:10,18)
mlogitdata$choiceid = 1:nrow(my_data) # First preparing the data for mlogit

mldata = dfidx(mlogitdata, shape = "wide", choice = dependentvar, idx = list(c("choiceid")),pkg="mlogit")
mlogitformula = as.formula(paste(dependentvar, "~ 0 | SE_mean + prior_math_grade + grade_goal | 0", sep = ""))
ml.iv = (mlogit(mlogitformula, mldata,reflevel="1"))
summary(ml.iv)

####vgam package
GLMformula = as.formula(paste(dependentvar, "~ SE_mean + prior_math_grade + grade_goal", sep = ""))
```

```r
glm.iv =vglm(GLMformula, data=my_data, family = multinomial(refLevel =1))
summary(glm.iv)


####UPG package
ivnumber = which(colnames(my_data)==dependentvar)+1 #picking the right column as dependentvar
print(ivnumber)

bayesdata = my_data %>%
  select(1,2,3:10,18)
bayesdata= data.matrix(bayesdata) #first preparing the data for UPG
intercept = rep(1,nrow(bayesdata))
bayesdata=cbind(intercept,bayesdata)

#Default values are: A0 = 4, B0 = 4, burnin = 1000, draws = 1000. Baseline category is by default chosen as category that occurs most often. Pick category 1 with "baseline".

y = bayesdata[,ivnumber] #Pick number of the right dependent variable for in the analysis (represented as y)
X = data.matrix(bayesdata[,c(1,2,3,12)])
bayes.iv = UPG(y=y,X=X,model='mnl',baseline="1",draws=10000,burnin=2000,A0=4,B0=4)
summary(bayes.iv)
plot(bayes.iv)


###Determine model fit with loglikelihood, AIC and BIC.
AIC=AIC(ml.iv,k=3)
BIC = AIC(ml.iv, k=log(nrow(my_data)))
loglikelihood = logLik(ml.iv)[1]
print(paste0("For the dependent variable and the package mlogit ", dependentvar, " the AIC is ", AIC, " and BIC is ", BIC, " and the loglikelihood is ", loglikelihood))


AIC2=AIC(glm.iv,k=3)
BIC2 = AIC(glm.iv, k=log(nrow(my_data)))
loglikelihood2 = logLik(glm.iv)[1]
print(paste0("For the dependent variable and the package VGAM ", dependentvar, " the AIC is ", AIC2, " and BIC is ", BIC2, " and the loglikelihood is ", loglikelihood2))

AIC3=AIC(bayes.iv,k=3)
BIC3 = AIC(bayes.iv, k=log(nrow(my_data)))
loglikelihood3 = logLik(bayes.iv)[1]
print(paste0("For the dependent variable and the package UPG ", dependentvar, " the AIC is ", AIC3, " and BIC is ", BIC3, " and the loglikelihood is ", loglikelihood3))


###Plot confidence and credible intervals
confidence= confint(ml.iv)                                #confidence intervals
credint = coef(bayes.iv, q = c(0.025, 0.975))             #credible intervals


## Voor 4 categorie n uitkomstvariable
##Rood is confidence, zwart is credible
sp = ggplot()
sp + geom_segment(aes(x = credint[[1]][,1][1], y = 15, xend = credint[[3]][,1][1], yend = 15)) + annotate(geom="text", x=-5, y=15, label="Intercept (2)",color="blue")+ geom_segment(aes(x = credint[[1]][,1][2], y = 14, xend = credint[[3]][,1][2], yend = 14)) + annotate(geom="text", x=-5, y=14, label="Intercept (3)",color="blue") + geom_segment(aes(x = credint[[1]][,1][3], y = 13, xend = credint[[3]][,1][3], yend = 13)) + annotate(geom="text", x=-5, y=13, label="Intercept (4)",color="blue")+ geom_segment(aes(x = credint[[1]][,1][4], y = 12, xend = credint[[3]][,1][4], yend = 12)) + annotate(geom="text", x=3.5, y=12, label="Self-efficacy (2)",color="blue")+ geom_segment(aes(x = credint[[1]][,2][1], y = 11, xend = credint[[3]][,2][1], yend = 11)) +  annotate(geom="text", x=3.5, y=11, label="Self-efficacy (3)",color="blue") + geom_segment(aes(x = credint[[1]][,2][2], y = 10, xend = credint[[3]][,2][1], yend = 10)) + annotate(geom="text", x=3.5, y=10, label="Self-efficacy (4)", color="blue")+ geom_segment(aes(x = credint[[1]][,2][3], y = 9, xend = credint[[3]][,2][3], yend = 9)) + annotate(geom="text", x=-5, y=9, label="Prior math grade (2)",color="blue") + geom_segment(aes(x = credint[[1]][,2][4], y = 8, xend = credint[[3]][,2][4], yend = 8)) + annotate(geom="text", x=-5, y=8, label="Prior math grade (3)",color="blue")+ geom_segment(
    aes(x = credint[[1]][,3][1], y = 7, xend = credint[[3]][,3][1], yend = 7)) + annotate(geom="
    text", x=-5, y=7, label="Prior math grade (4)",color="blue") + geom_segment(aes(x = credint
    [[1]][,3][2], y = 6, xend = credint[[3]][,3][2], yend = 6)) + annotate(geom="text", x=-5, y
    =6, label="Grade goal (2)",color="blue") + geom_segment(aes(x = credint[[1]][,3][3], y = 5,
    xend = credint[[3]][,3][3], yend = 5)) + annotate(geom="text", x=-5, y=5, label="Grade goal
    (3)",color="blue") + geom_segment(aes(x =credint[[1]][,3][4], y = 4, xend = credint
    [[3]][,3][4], yend = 4)) + annotate(geom="text", x=-5, y=4, label="Grade goal (4)",color="
    blue")  + geom_segment(aes(x = confidence[1,1], y = 14.5, xend = confidence[1,2], yend =
    14.5), colour="red") +  geom_segment(aes(x = confidence[7,1], y = 13.5, xend = confidence
    [7,2], yend = 13.5), colour="red") +  geom_segment(aes(x = confidence[4,1], y = 12.5, xend =
     confidence[4,2], yend = 12.5), colour="red") +  geom_segment(aes(x = confidence[10,1], y =
    11.5, xend = confidence[10,2], yend = 11.5), colour="red") +  geom_segment(aes(x =
    confidence[2,1], y = 10.5, xend = confidence[2,2], yend = 10.5), colour="red")  +  geom_
    segment(aes(x = confidence[8,1], y = 9.5, xend = confidence[8,2], yend = 9.5), colour="red")
     +  geom_segment(aes(x = confidence[5,1], y = 8.5, xend = confidence[5,2], yend = 8.5),
    colour="red") +  geom_segment(aes(x = confidence[11,1], y = 7.5, xend = confidence[11,2],
    yend = 7.5), colour="red") +  geom_segment(aes(x = confidence[3,1], y = 6.5, xend =
    confidence[3,2], yend = 6.5), colour="red") +  geom_segment(aes(x = confidence[9,1], y =
    5.5, xend = confidence[9,2], yend = 5.5), colour="red") + geom_segment(aes(x = confidence
    [6,1], y = 4.5, xend = confidence[6,2], yend = 4.5), colour="red") + geom_segment(aes(x =
    confidence[12,1], y = 3.5, xend = confidence[12,2], yend = 3.5), colour="red") + xlab(paste0
    ("Confidence intervals (red) and credible intervals (black) for ", dependentvar))
```

Appendix B

Appendix: Cronbach's alpha

```r
#Select relevant data gradegoal
gradedata = my_data %>%
  select(12:14)
#Select relevant data for self-efficacy
SEcdata = my_data_all %>%
  select(16, 45:66)
#Remove rows with missing data (total is 286, of which 7 incomplete questionnaires)
SEc_data = data.frame(SEcdata[complete.cases(SEcdata[ , 22]),])
SE_data = SEc_data %>%
  select(2:6)

#Cronbach alpha: define own function to compute Cronbach alpha
cronbachalpha <- function(dataframe){
  k = length(dataframe)
  totalsum = rowSums(dataframe)
  sd_total_squared = sd(totalsum)^2
  if (k == 5){                          #for Self-efficacy, 5 items
    sd_sum_squared = sd(dataframe[,1])^2 + sd(dataframe[,2])^2 + sd(dataframe[,3])^2 + sd(dataframe[,4])^2  + sd(dataframe[,5])^2
  }
  else{                                   #for Grade goals, 3 items
    sd_sum_squared = sd(dataframe[,1])^2 + sd(dataframe[,2])^2 + sd(dataframe[,3])^2
  }
  alpha = k/(k-1) * (1 - sd_sum_squared/sd_total_squared)
  print(alpha)
}

#Cronbach alpha: gradedata
cronbachalpha(gradedata) #with own function
cronbach.alpha(gradedata) #with R-function

#Cronbach alpha: self-efficacy
cronbachalpha(SE_data) #with own function
cronbach.alpha(SE_data) #with R-function
```

Appendix C

Appendix: Assumption check - code to generate 3D plots & scatterplots

```
###3D plots for each dependent variable to assess separability
###Choose variable (study_next, study_more_for, material_more_than_once, why_quiz, convinced_do_next, pattern_of_study, time_of_day, time_of_day_belief):
dependentvar = "study_next"

#define colors
color3 = c("coral","cornsilk", "midnightblue")
colors4 = c("coral","cornsilk", "limegreen","midnightblue")
colors5 = c("coral","cornsilk", "limegreen","midnightblue","black")

if (dependentvar == "material_more_than_once" || dependentvar == "pattern_of_study" || dependentvar == "convinced_do_next") { print(paste0("This 3D plot is for the dependent variable ",dependentvar))  #For dependent variables with 3 categories
numberdv = which(colnames(my_data) == dependentvar)
dv_df = as.numeric(unlist(my_data[numberdv]))
my_data$color = colors[as.numeric( as.factor(dv_df)) ]
plot3d(
  x=my_data$prior_math_grade,
  y=my_data$SE_mean,
  z=my_data$grade_goal,
  col = my_data$color,
  type = ’s’,
  radius = .1,
  xlab="Prior math grade", ylab="Self-efficacy", zlab="Grade goal")
rglwidget()
} else if(dependentvar == "study_more_for" || dependentvar == "why_quiz" || dependentvar == "time_of_day" || dependentvar == "time_of_day_belief"){
 print(paste0("This 3D plot is for the dependent variable ",dependentvar))  #For dependent variables with 4 categories
 numberdv = which(colnames(my_data) == dependentvar)
 dv_df = as.numeric(unlist(my_data[numberdv]))
  my_data$color = colors4[ as.numeric(as.factor(dv_df) ) ]
plot3d(
  x=my_data$prior_math_grade,
  y=my_data$SE_mean,
  z=my_data$grade_goal,
  col = my_data$color,
  type = ’s’,
  radius = .1,
  xlab="Prior math grade", ylab="Self-efficacy", zlab="Grade goal")
rglwidget()
} else {            #For dependentvar study_next (only one with 5 categories)
  print(paste0("This 3D plot is for the dependent variable ",dependentvar))
numberdv = which(colnames(my_data) == dependentvar)
dv_df = as.numeric(unlist(my_data[numberdv]))
  my_data$color = colors5[as.numeric( as.factor(dv_df) )]
plot3d(
  x=my_data$prior_math_grade,
  y=my_data$SE_mean,
  z=my_data$grade_goal,
  col = my_data$color,
  type = ’s’,
  radius = .1,
  xlab="Prior math grade", ylab="Self-efficacy", zlab="Grade goal")

rglwidget()

}

###Assumption multicollinearity: Pearson’s correlation coefficient and scatterplots. Check all pairs of independent variables (hence self-efficacy, grade goal and prior math grade)

corr = round(cor(my_data[, c(1,2,18)], method = "pearson"),5)     #Create dataframe IV’s
cor_SE_pmg = paste("Pearson’s correlation coefficient: ", corr[1,2])   #Self-efficacy and prior math grade
cor_SE_gg = paste("Pearson’s correlation coefficient: ", corr[1,3]) #Self-efficacy and grade goal
cor_pmg_gg = paste("Pearson’s correlation coefficient: ", corr[2,3])  #Grade goal and prior math grade



ggplot(my_data, aes(x=SE_mean,y=prior_math_grade)) + geom_point(position = "jitter") + annotate("text", x=4, y=5, label= cor_SE_pmg) + xlab("Self-efficacy mean score (Scale: 1-5)") + ylab("Prior math grade (Scale: 1-10)")
ggplot(my_data,aes(x=SE_mean,y=grade_goal)) + geom_point(position=position_jitter(h=0.1, w=0.1)) + annotate("text", x=4, y=5, label= cor_SE_gg) + xlab("Self-efficacy mean score (Scale: 1-5)") + ylab("Grade goal (Scale: 1-10)")
ggplot(my_data,aes(x=grade_goal,y=prior_math_grade)) + geom_point(position = "jitter")+ annotate("text", x=8, y=5, label= cor_pmg_gg) + xlab("Grade goal (Scale: 1-10)")  + ylab("Prior math grade (Scale: 1-10)")
```

Appendix D

Appendix: Output of mlogit

```

Call:
mlogit(formula = why_quiz ~ 0 | SE_mean + prior_math_grade +
    grade_goal | 0, data = mldata, reflevel = "1", method = "nr")

Frequencies of alternatives:choice
       1        2        3        4
0.275986 0.433692 0.075269 0.215054

nr method
5 iterations, 0h:0m:0s
g'(-H)^-1g = 1.63E-07
gradient close to zero

Coefficients :
                    Estimate Std. Error z-value Pr(>|z|)
(Intercept):2       1.429528   1.245902  1.1474 0.251223
(Intercept):3      -3.655668   2.148428 -1.7016 0.088839 .
(Intercept):4       3.985226   1.489282  2.6759 0.007452 **
SE_mean:2          -0.747113   0.286382 -2.6088 0.009086 **
SE_mean:3          -0.366232   0.502425 -0.7289 0.466045
SE_mean:4          -0.899107   0.328348 -2.7383 0.006176 **
prior_math_grade:2  0.011188   0.168281  0.0665 0.946993
prior_math_grade:3 -0.158517   0.283995 -0.5582 0.576729
prior_math_grade:4 -0.067749   0.199946 -0.3388 0.734734
grade_goal:2        0.245657   0.233603  1.0516 0.292982
grade_goal:3        0.663917   0.395737  1.6777 0.093411 .
grade_goal:4       -0.036622   0.287937 -0.1272 0.898790
---
Signif. codes:  0    ***    0.001    **    0.01    *    0.05    .    0.1         1

Log-Likelihood: -333.65
McFadden R^2:  0.037763
Likelihood ratio test : chisq = 26.188 (p.value = 0.001903)
```

Appendix E

Appendix: Output of VGAM

```

Call:
vglm(formula = GLMformula, family = multinomial(refLevel = 1),
    data = my_data)

Coefficients:
                   Estimate Std. Error z value Pr(>|z|)
(Intercept):1       1.42953    1.24590   1.147  0.25122
(Intercept):2      -3.65567    2.14836      NA       NA
(Intercept):3       3.98523    1.48928   2.676  0.00745 **
SE_mean:1          -0.74711    0.28638  -2.609  0.00909 **
SE_mean:2          -0.36623    0.50241  -0.729  0.46603
SE_mean:3          -0.89911    0.32835  -2.738  0.00618 **
prior_math_grade:1  0.01119    0.16828   0.066  0.94699
prior_math_grade:2 -0.15852    0.28399  -0.558  0.57672
prior_math_grade:3 -0.06775    0.19995  -0.339  0.73473
grade_goal:1        0.24566    0.23360   1.052  0.29298
grade_goal:2        0.66392    0.39573   1.678  0.09340 .
grade_goal:3       -0.03662    0.28794  -0.127  0.89879
---
Signif. codes:  0   ***   0.001   **   0.01   *   0.05   .   0.1       1

Names of linear predictors: log(mu[,2]/mu[,1]), log(mu[,3]/mu[,1]), log(mu[,4]/mu[,1])

Residual deviance: 667.3093 on 825 degrees of freedom

Log-likelihood: -333.6547 on 825 degrees of freedom

Number of Fisher scoring iterations: 5

Warning: Hauck-Donner effect detected in the following estimate(s):
’(Intercept):2’

Reference group is level  1  of the response
```

Appendix F

Appendix: Output of UPG

```

Checking data & inputs ...
Initializing Gibbs Sampler ...
Simulating from posterior distribution ...
  |========================================================================================|
    100%
Sampling succesful!
Saving output ...
Finished! Posterior simulation took 35.95 seconds.

--- Bayesian Multinomial Logit Results ---

N = 279
Analysis based on 10000 posterior draws after
an initial burn-in period of 2000 iterations.
MCMC sampling took a total of 35.95 seconds.

Category ’1’ is the baseline category.


|                  |  Mean|   SD|  Q2.5| Q97.5| 95% CI excl. 0 |
|:-----------------|-----:|----:|-----:|-----:|:--------------:|
|Category ’2’      |      |     |      |      |                |
|intercept         |  1.11| 0.98| -0.77|  3.02|                |
|prior_math_grade  |  0.02| 0.15| -0.27|  0.32|                |
|SE_mean           | -0.75| 0.25| -1.25| -0.26|       *        |
|grade_goal        |  0.28| 0.21| -0.14|  0.71|                |
|                  |      |     |      |      |                |
|Category ’3’      |      |     |      |      |                |
|intercept         | -1.41| 1.38| -4.20|  1.24|                |
|prior_math_grade  | -0.28| 0.26| -0.80|  0.23|                |
|SE_mean           | -0.34| 0.45| -1.21|  0.57|                |
|grade_goal        |  0.49| 0.36| -0.21|  1.19|                |
|                  |      |     |      |      |                |
|Category ’4’      |      |     |      |      |                |
|intercept         |  2.77| 1.08|  0.67|  4.89|       *        |
|prior_math_grade  | -0.04| 0.18| -0.39|  0.31|                |
|SE_mean           | -0.95| 0.30| -1.54| -0.39|       *        |
|grade_goal        |  0.13| 0.25| -0.37|  0.62|                |
```

Appendix G

Appendix: Different runs of UPG

These runs are performed with 2000 burn-in and 10000 draws and $A0 = 4$, $B0 = 4$ as the parameters of the prior distribution. Only the point estimates of four coefficients are included to provide a clear overview. It can be seen that the values provided by all five runs are approximately the same, hence the Markov chains in the different runs all seem to converge to the desired posterior distributions.

	Run 1	**Run 2**	**Run 3**	**Run 4**	**Run 5**
AIC	472.459	470.267	469.545	471.424	470.211
BIC	504.033	501.841	501.119	503.120	502.983
Time posterior simulation	34.72	34.39	33.89	33.94	34.92
Intercept (2) β_{02}	1.111	1.106	1.123	1.121	1.109
Self-efficacy (2) β_{12}	0.027	0.020	0.022	0.024	0.025
Prior math grade (2) β_{22}	-0.751	-0.742	-0.750	-0.748	-0.750
Grade goal (2) β_{32}	0.273	0.276	0.276	0.285	0.277

Appendix H

Appendix: Visual display of posterior distribution

```
####Constructing visuals posterior distribution UPG - for dependent variable with 4 categories

######Rearranging the matrix
#The element with the posterior distribution returned from UPG (bayes.iv$posterior$beta) is a 3D-matrix
#Columns of matrix: Intercept, prior_math_grade, SE_mean, grade_goal
#Layers of matrix: First layer = Category 2, Second layer = Category 3, Third layer = Category 4, Fourth layer = Reference category 1
posterior = as.array(bayes.iv$posterior$beta)
column.names = c("intercept","prior_math_grade","SE_mean","grade_goal")
matrix.names=c("2","3","4","1")
#rename and rearrange the structure of the matrix such that the histograms can be constructed with function mcmc_hist_by_chain
dimnames(posterior)[[2]]=column.names
dimnames(posterior)[[3]]=matrix.names
newdata = aperm(posterior, c(1,3,2))

##Histograms, binwidth can be specified
mcmc_hist_by_chain(newdata, pars = c("intercept"), binwidth=0.1)
mcmc_hist_by_chain(newdata, pars = c("prior_math_grade"),binwidth=0.02)
mcmc_hist_by_chain(newdata, pars = c("SE_mean"),binwidth=0.03)
mcmc_hist_by_chain(newdata,pars = c("grade_goal"), binwidth=0.02)


##Scatter plots for all coefficients, first extract right columns from the 3D matrix
intercept2 = bayes.iv$posterior$beta[,,1][,1]
intercept3 = bayes.iv$posterior$beta[,,2][,1]
intercept4 = bayes.iv$posterior$beta[,,3][,1]
intercept1 = bayes.iv$posterior$beta[,,4][,1]
priormath2 = bayes.iv$posterior$beta[,,1][,2]
priormath3 = bayes.iv$posterior$beta[,,2][,2]
priormath4 = bayes.iv$posterior$beta[,,3][,2]
priormath1 = bayes.iv$posterior$beta[,,4][,2]
SE2 = bayes.iv$posterior$beta[,,1][,3]
SE3 = bayes.iv$posterior$beta[,,2][,3]
SE4 = bayes.iv$posterior$beta[,,3][,3]
SE1 = bayes.iv$posterior$beta[,,4][,3]
gradegoal2 = bayes.iv$posterior$beta[,,1][,4]
gradegoal3 = bayes.iv$posterior$beta[,,2][,4]
gradegoal4 = bayes.iv$posterior$beta[,,3][,4]
gradegoal1 = bayes.iv$posterior$beta[,,4][,4]

#create dataframe for scatterplots (not for reference category): total 18 scatterplots
dataframe = data.frame(intercept2,intercept3,intercept4,intercept1,priormath2,priormath3,priormath4,priormath1,SE2,SE3,SE4,SE1,gradegoal2,gradegoal3,gradegoal4,gradegoal1)
#two possibilities for the scatterplots: 'heat map' and regular scatter plot
```

```
ggplot(dataframe,aes(x=intercept2,y=priormath2)) +  geom_hex() +scale_fill_viridis_c()
ggplot(dataframe,aes(x=intercept2,y=priormath2)) + geom_point(alpha = 0.1)+  geom_rug(alpha = 0.01)

ggplot(dataframe,aes(x=intercept2,y=SE2)) +  geom_hex() +scale_fill_viridis_c()
ggplot(dataframe,aes(x=intercept2,y=SE2)) + geom_point(alpha = 0.3)

ggplot(dataframe,aes(x=intercept2,y=gradegoal2)) +  geom_hex() +scale_fill_viridis_c()
ggplot(dataframe,aes(x=intercept2,y=gradegoal2)) + geom_point(alpha = 0.3)

ggplot(dataframe,aes(x=SE2,y=gradegoal2)) +  geom_hex() +scale_fill_viridis_c()
ggplot(dataframe,aes(x=SE2,y=gradegoal2)) + geom_point(alpha = 0.3)

ggplot(dataframe,aes(x=SE2,y=priormath2)) +  geom_hex() +scale_fill_viridis_c()
ggplot(dataframe,aes(x=SE2,y=priormath2)) + geom_point(alpha = 0.3)

ggplot(dataframe,aes(x=gradegoal2,y=priormath2)) +  geom_hex() +scale_fill_viridis_c()
ggplot(dataframe,aes(x=gradegoal2,y=priormath2)) + geom_point(alpha = 0.3)

ggplot(dataframe,aes(x=intercept3,y=priormath3)) +  geom_hex() +scale_fill_viridis_c()
ggplot(dataframe,aes(x=intercept3,y=priormath3)) + geom_point(alpha = 0.3)

ggplot(dataframe,aes(x=intercept3,y=SE3)) +  geom_hex() +scale_fill_viridis_c()
ggplot(dataframe,aes(x=intercept3,y=SE3)) + geom_point(alpha = 0.3)

ggplot(dataframe,aes(x=intercept3,y=gradegoal3)) +  geom_hex() +scale_fill_viridis_c()
ggplot(dataframe,aes(x=intercept3,y=gradegoal3)) + geom_point(alpha = 0.3)

ggplot(dataframe,aes(x=SE3,y=gradegoal3)) +  geom_hex() +scale_fill_viridis_c()
ggplot(dataframe,aes(x=SE3,y=gradegoal3)) + geom_point(alpha = 0.3)

ggplot(dataframe,aes(x=SE3,y=priormath3)) +  geom_hex() +scale_fill_viridis_c()
ggplot(dataframe,aes(x=SE3,y=priormath3)) + geom_point(alpha = 0.3)

ggplot(dataframe,aes(x=gradegoal3,y=priormath3)) +  geom_hex() +scale_fill_viridis_c()
ggplot(dataframe,aes(x=gradegoal3,y=priormath3)) + geom_point(alpha = 0.3)

ggplot(dataframe,aes(x=intercept4,y=priormath4)) +  geom_hex() +scale_fill_viridis_c()
ggplot(dataframe,aes(x=intercept4,y=priormath4)) + geom_point(alpha = 0.3)

ggplot(dataframe,aes(x=intercept4,y=SE4)) +  geom_hex() +scale_fill_viridis_c()
ggplot(dataframe,aes(x=intercept4,y=SE4)) + geom_point(alpha = 0.3)

ggplot(dataframe,aes(x=intercept4,y=gradegoal4)) +  geom_hex() +scale_fill_viridis_c()
ggplot(dataframe,aes(x=intercept4,y=gradegoal4)) + geom_point(alpha = 0.3)

ggplot(dataframe,aes(x=SE4,y=gradegoal4)) +  geom_hex() +scale_fill_viridis_c()
ggplot(dataframe,aes(x=SE4,y=gradegoal4)) + geom_point(alpha = 0.3)

ggplot(dataframe,aes(x=SE4,y=priormath4)) +  geom_hex() +scale_fill_viridis_c()
ggplot(dataframe,aes(x=SE4,y=priormath4)) + geom_point(alpha = 0.3)

ggplot(dataframe,aes(x=gradegoal4,y=priormath4)) +  geom_hex() +scale_fill_viridis_c()
ggplot(dataframe,aes(x=gradegoal4,y=priormath4)) + geom_point(alpha = 0.3)
```

This code produces histograms of the posterior distribution.

Moreover it produces scatter plots for the pairs of coefficients. The baseline category is not included, since the values for the baseline coefficients are all 0, hence it does not produce informative scatterplots, as seen in figure H.

The scatter plots of the other baseline coefficient pairs are displayed below. Only pairs of the same categories are plotted, hence for example in figure H.6 the relation between the effect of intercept and prior math grade for category 2 are plotted.

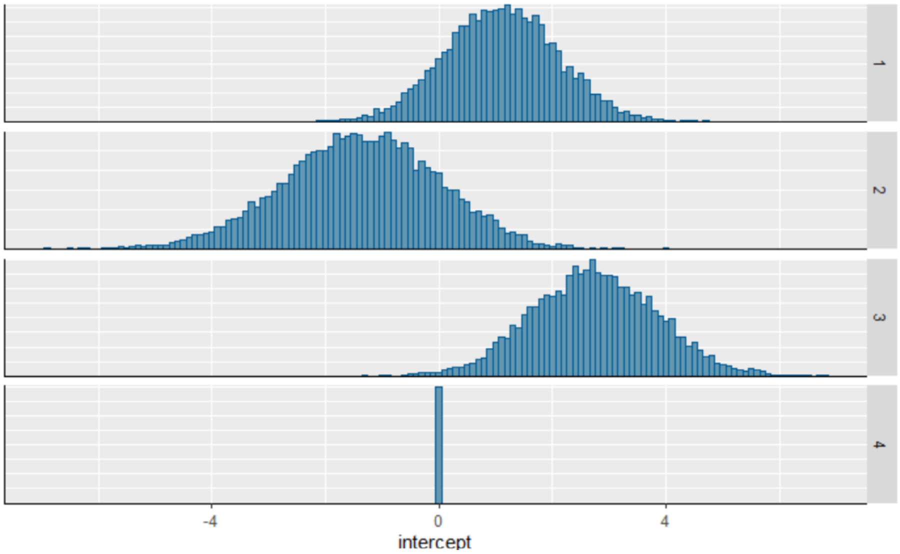

Figure H.1: Histograms of posterior distribution Intercept

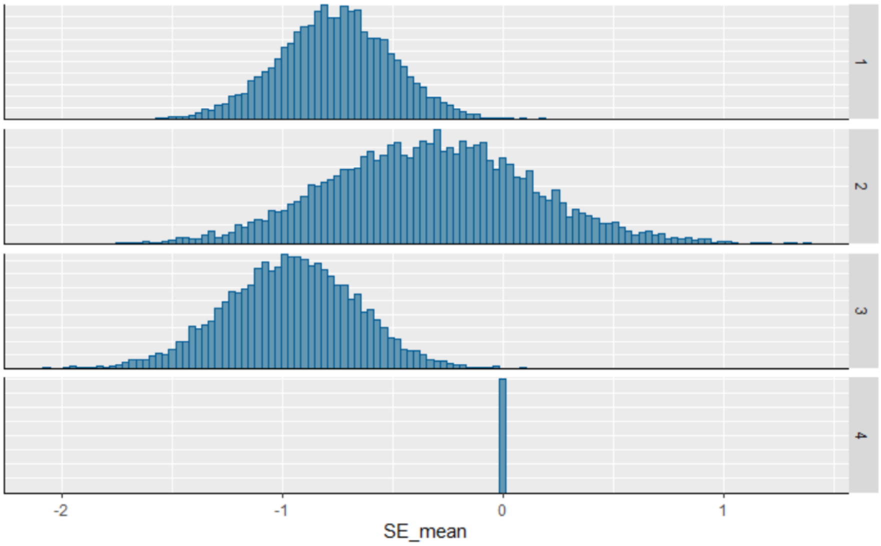

Figure H.2: Histograms of posterior distribution Self-efficacy

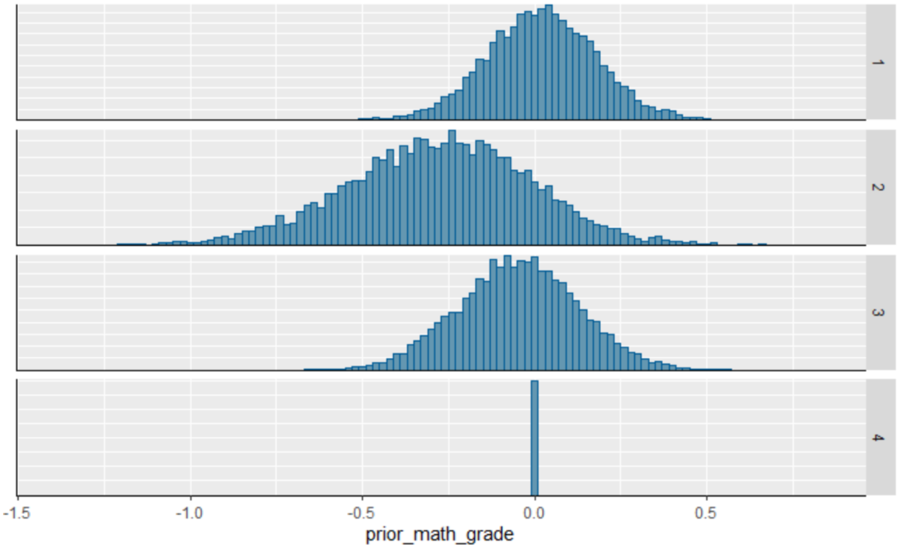

Figure H.3: Histograms of posterior distribution Prior math grade

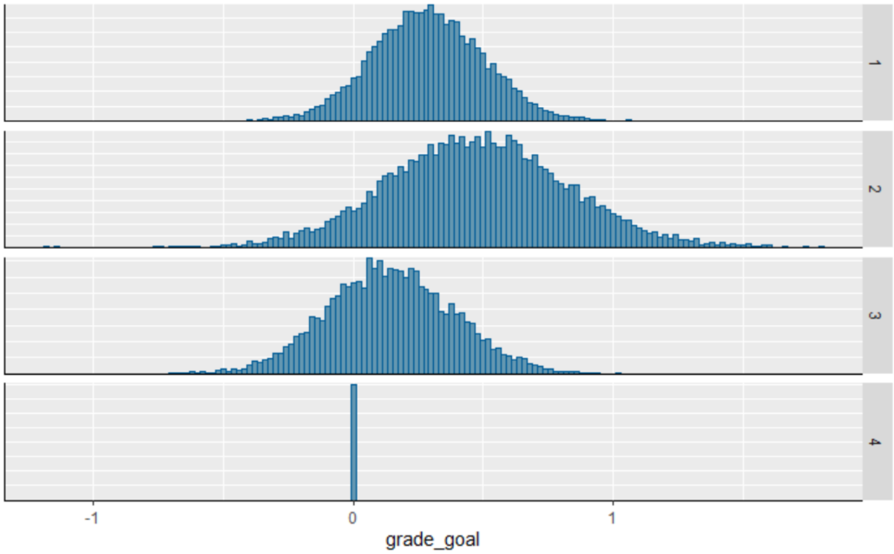

Figure H.4: Histograms of posterior distribution Grade goal

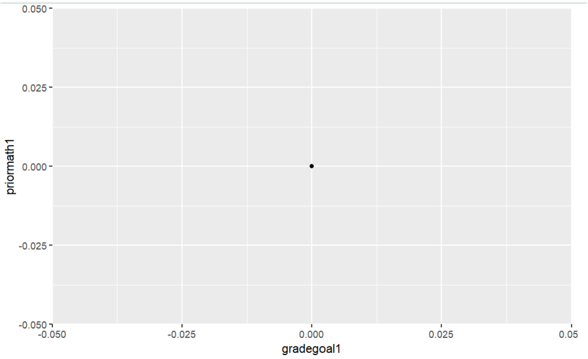

Figure H.5: Scatterplot of baseline category, uninformative since all coefficients are set to 0

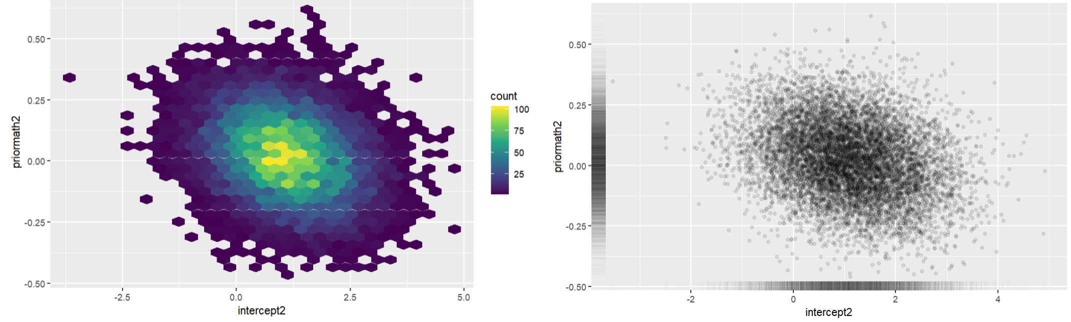

Figure H.6: Intercept(2) and prior math grade(2)

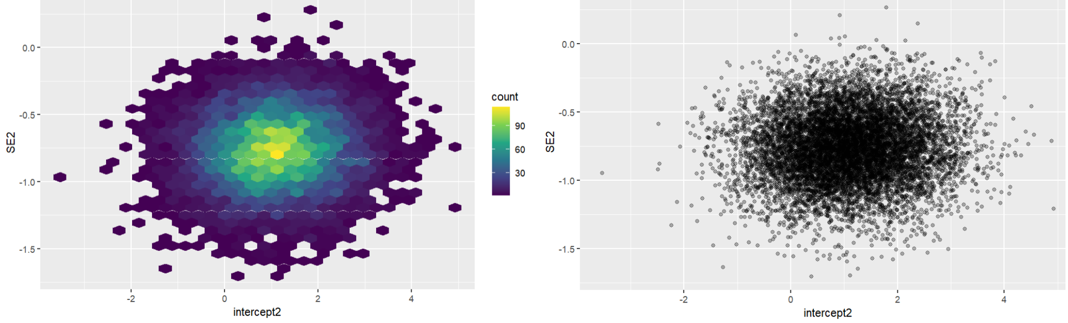

Figure H.7: Intercept(2) and self-efficacy(2)

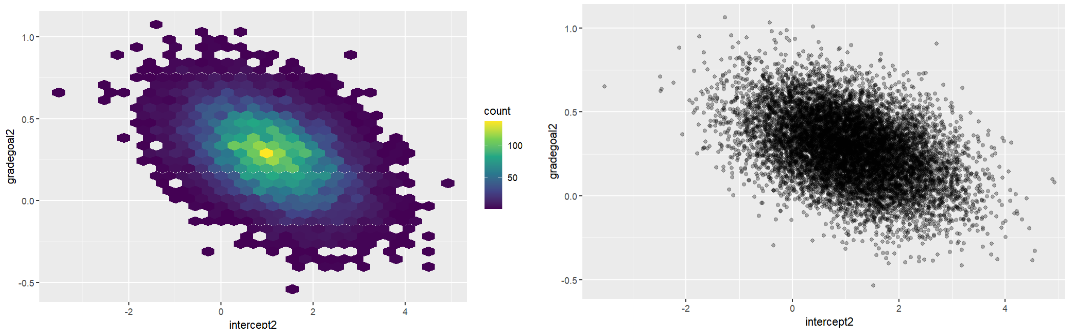

Figure H.8: Intercept(2) and grade goal(2)

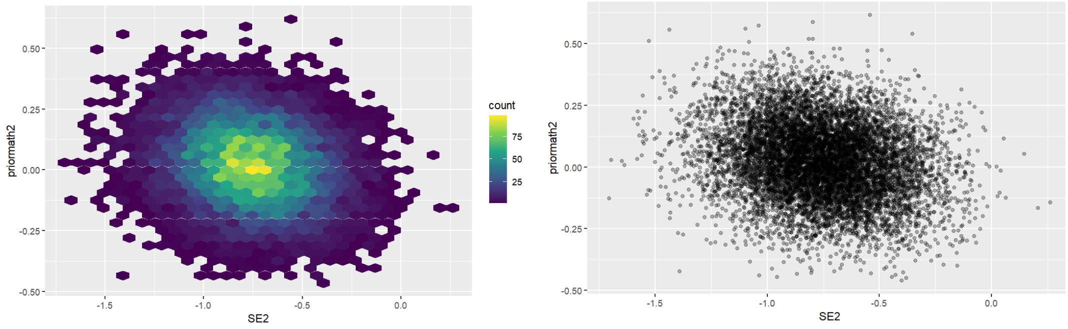

Figure H.9: Self-efficacy(2) and prior math grade(2)

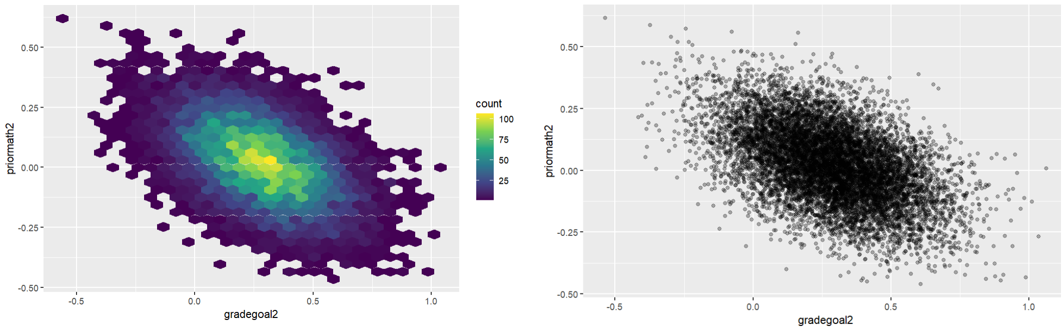

Figure H.10: Grade goal(2) and prior math grade(2)

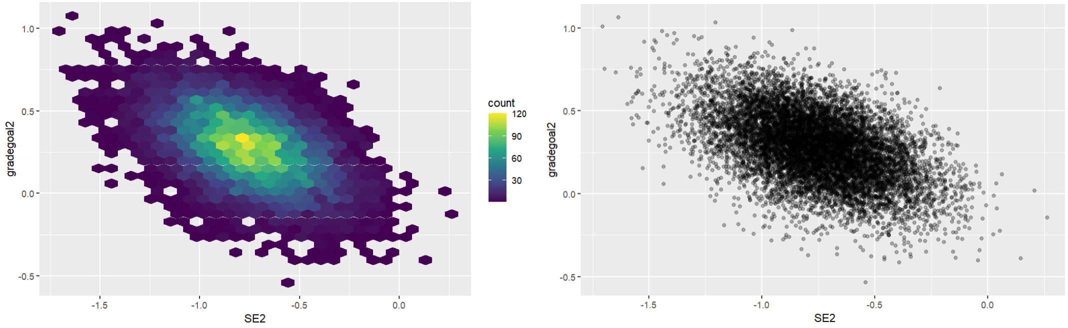

Figure H.11: Self-efficacy(2) and grade goal(2)

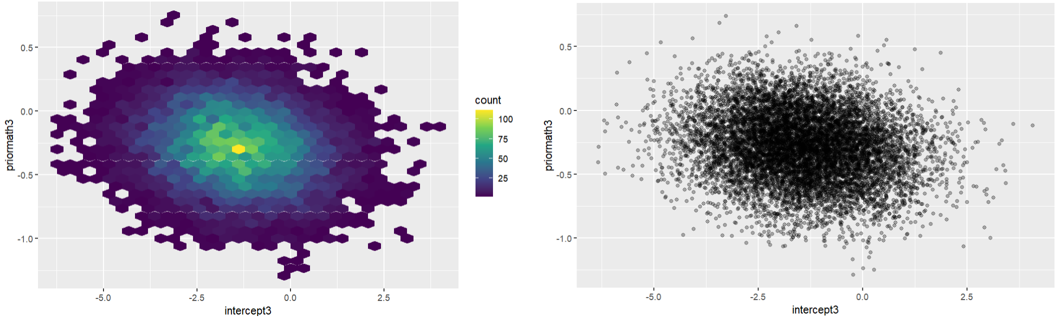

Figure H.12: Intercept(3) and prior math grade(3)

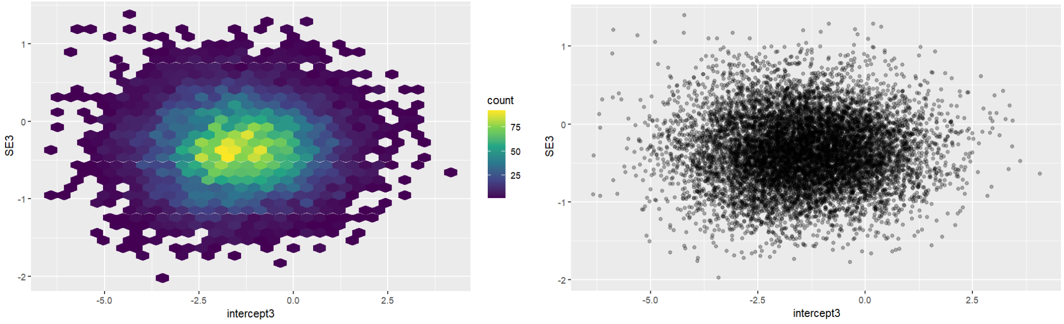

Figure H.13: Intercept(3) and self-efficacy(3)

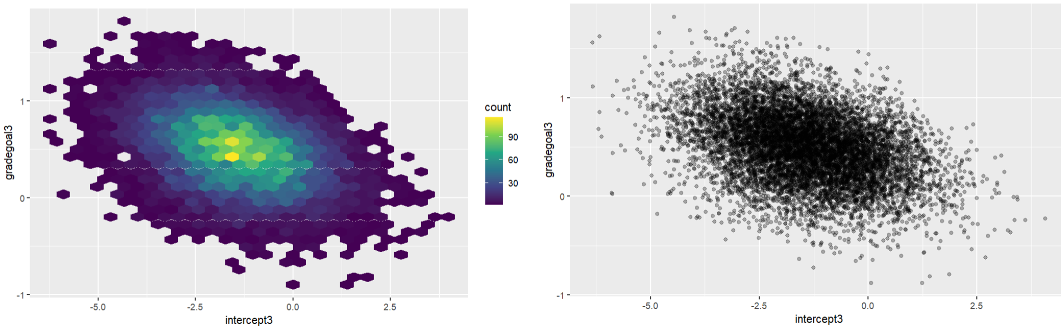

Figure H.14: Intercept(3) and grade goal(3)

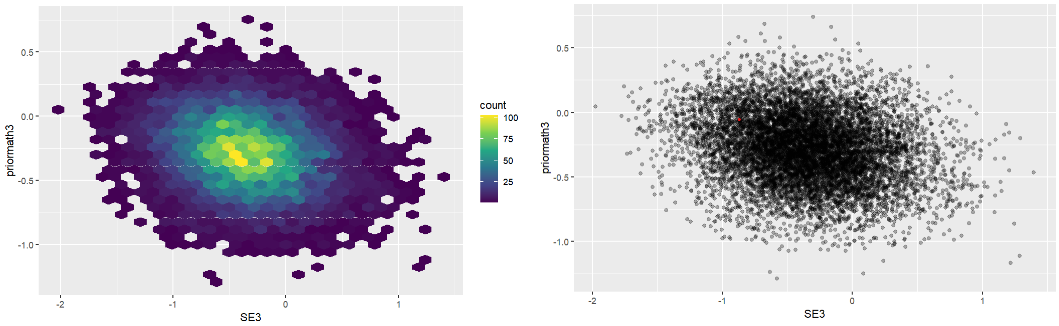

Figure H.15: Self-efficacy(3) and prior math grade(3)

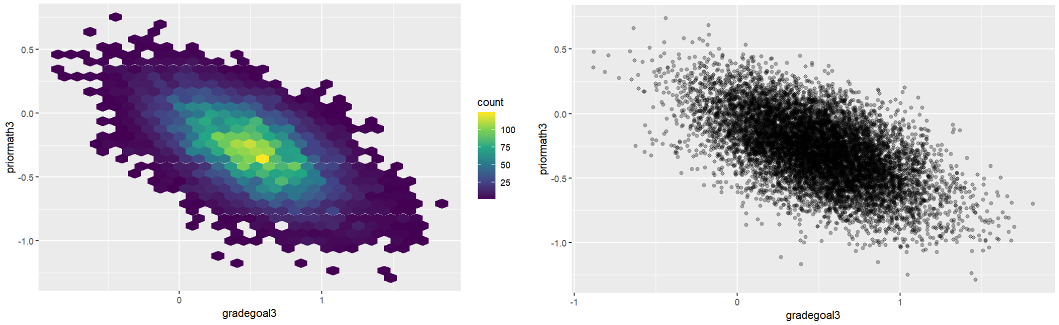

Figure H.16: Grade goal(3) and prior math grade(3)

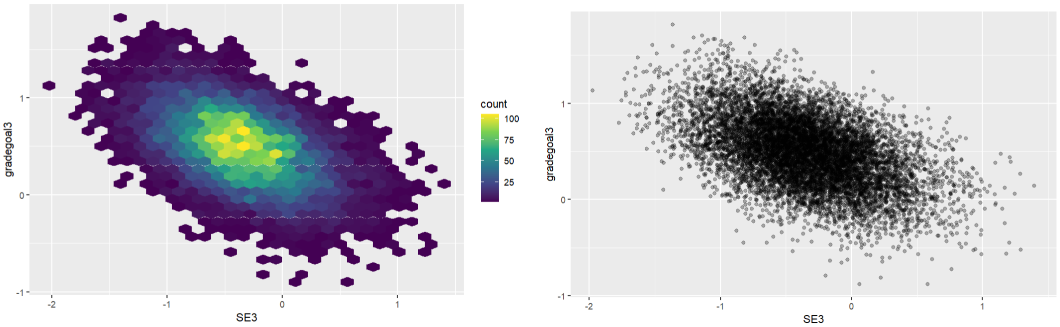

Figure H.17: Self-efficacy(3) and grade goal(3)

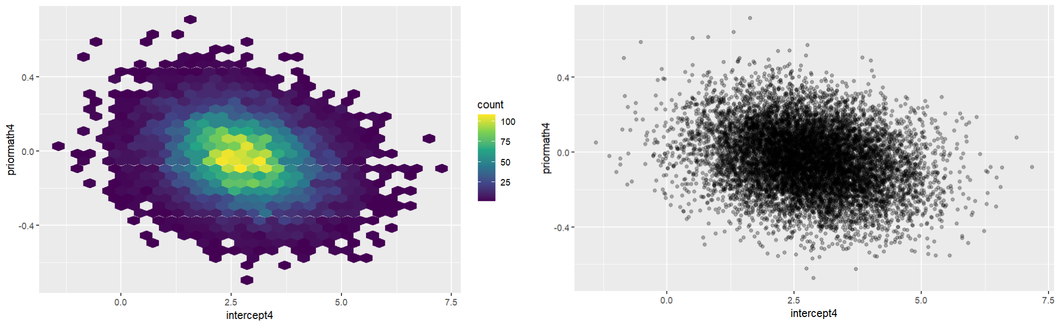

Figure H.18: Intercept(4) and prior math grade(4)

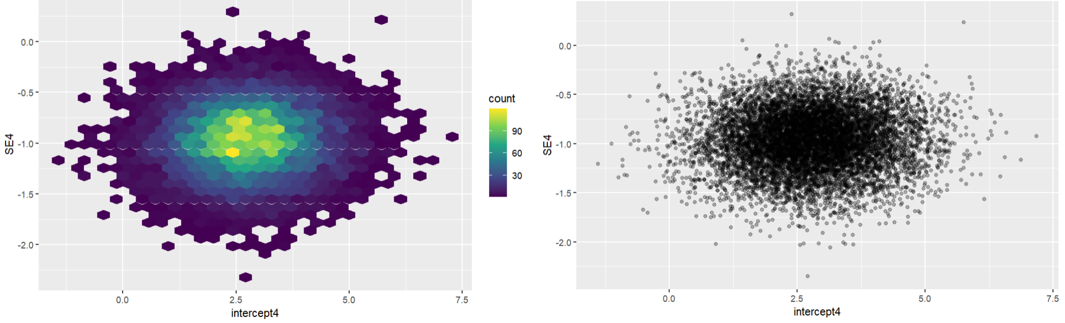

Figure H.19: Intercept(4) and self-efficacy(4)

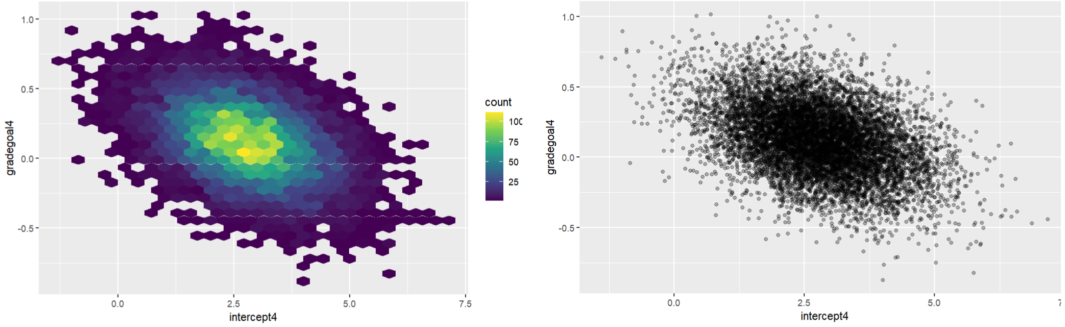

Figure H.20: Intercept(4) and grade goal(4)

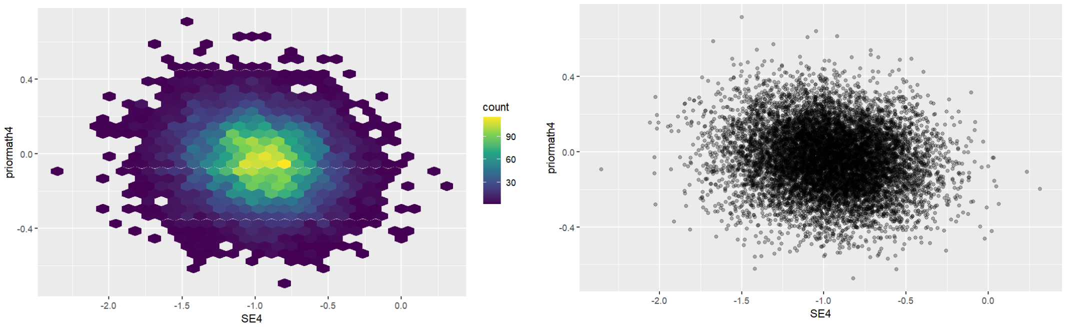

Figure H.21: Self-efficacy(4) and prior math grade(4)

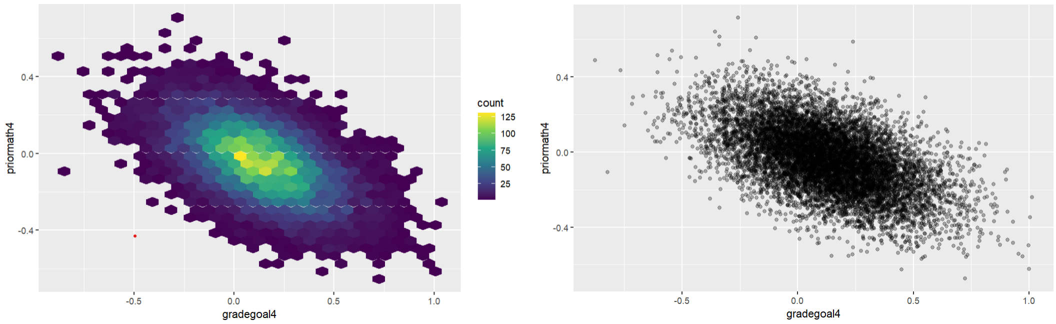

Figure H.22: Grade goal(4) and prior math grade(4)

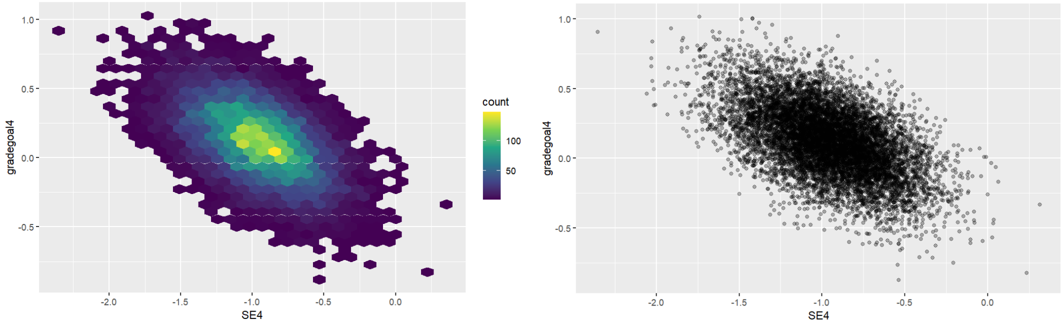

Figure H.23: Self-efficacy(4) and grade goal(4)

Bibliography

[Bandura, 1977] Bandura, A. (1977). Self-efficacy: toward a unifying theory of behavioral change. *Psychological review*, 84(2):191.

[Bland and Altman, 1997] Bland, J. M. and Altman, D. G. (1997). Statistics notes: Cronbach's alpha. *Bmj*, 314(7080):572.

[Bolstad and Curran, 2016] Bolstad, W. M. and Curran, J. M. (2016). *Introduction to Bayesian statistics*. John Wiley & Sons.

[Colley, 2012] Colley, S. J. (2012). *Vector Calculus*. Pearson, 4 edition.

[Croissant, 2010] Croissant, Y. (2010). Estimation of multinomial logit models in R : The `mlogit` package.

[Cronbach, 1951] Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, 16(3):297–334.

[Epperson, 2013] Epperson, J. (2013). *An Introduction to Numerical Methods and Analysis*. Wiley.

[Field et al., 2012] Field, A., Miles, J., and Field, Z. (2012). *Discovering statistics using R*. Sage publications.

[Firth, 1991] Firth, D. (1991). Generalized linear models. In D.V. Hinkley, N. R. and Snell, E., editors, *Statistical theory and modelling*, chapter 3, pages 55–82. Chapman and Hall.

[Fisher, 1925] Fisher, R. (1925). *Statistical Methods for Research Workers*. Oliver & Boyd, 1 edition.

[Fox, 2016] Fox, J. (2016). *Applied Regression Analysis and Generalized Linear Models*. SAGE Publications, London.

[Grigg et al., 2018] Grigg, S., Perera, H. N., McIlveen, P., and Svetleff, Z. (2018). Relations among math self-efficacy, interest, intentions, and achievement: A social cognitive perspective. *Contemporary Educational Psychology*, 53:73–86.

[Grimmett and Welsh, 2014] Grimmett, G. and Welsh, D. (2014). *Probability: an introduction*. Oxford University Press.

[Kornell and Bjork, 2007] Kornell, N. and Bjork, R. A. (2007). The promise and perils of self-regulated study. *Psychonomic bulletin & review*, 14(2):219–224.

[Lee, 2012] Lee, P. (2012). *Bayesian statistics: an introduction*, volume 4. John Wiley & Sons.

[Lehmann, 2011] Lehmann, E. L. (2011). *Fisher, Neyman, and the Creation of Classical Statistics*. Springer New York.

[Locke and Bryan, 1968] Locke, E. A. and Bryan, J. F. (1968). Grade goals as determinants of academic achievement. *The Journal of General Psychology*, 79(2):217–228.

[McElreath, 2020] McElreath, R. (2020). *Statistical Rethinking: A Bayesian Course with examples in R and STAN*.

[McFadden, 1974] McFadden, D. (1974). Conditional logit analysis of qualitative choice behavior. In Zarembka, P., editor, *Frontiers in Econometrics*, chapter 4, pages 105–142. Academic press.

[McFadden, 1977] McFadden, D. (1977). Quantitative methods for analyzing travel behaviour of individuals: some recent developments. *Cowles foundation discussion paper*, 474:1–47.

[Nelder and Wedderburn, 1972] Nelder, J. A. and Wedderburn, R. W. (1972). Generalized linear models. *Journal of the Royal Statistical Society: Series A (General)*, 135(3):370–384.

[Pernet, 2016] Pernet, C. (2016). Null hypothesis significance testing: A guide to commonly misunderstood concepts and recommendations for good practice. *F1000 Research 2016*, 4(621).

[PRIME TU Delft, 2022] PRIME TU Delft (2022). Research. `https://www.tudelft.nl/ewi/over-de-faculteit/afdelingen/applied-mathematics/studeren/prime/research`. Last checked on 08-12-2022.

[Schober et al., 2018] Schober, P., Boer, C., and Schwarte, L. A. (2018). Correlation coefficients: appropriate use and interpretation. *Anesthesia & Analgesia*, 126(5):1763–1768.

[Tabachnick et al., 2007] Tabachnick, B. G., Fidell, L. S., and Ullman, J. B. (2007). *Using multivariate statistics*, volume 5. pearson Boston, MA.

[Train, 2009] Train, K. E. (2009). *Discrete Choice Methods with Simulation.* Cambridge University Press, 2nd edition.

[van de Schoot et al., 2021] van de Schoot, R., Depaoli, S., King, R., Kramer, B., Märtens, K., Tadesse, M. G., Vannucci, M., Gelman, A., Veen, D., Willemsen, J., et al. (2021). Bayesian statistics and modelling. *Nature Reviews Methods Primers*, 1(1):1–26.

[van der Vaart et al., 2017] van der Vaart, A., Jonker, M., and Bijma, F. (2017). *An introduction to mathematical statistics.* Amsterdam University Press.

[Wasserstein and Lazar, 2016] Wasserstein, R. L. and Lazar, N. A. (2016). The ASA Statement on p-Values: Context, Process, and Purpose. *The American Statistician*, 70:129–133.

[Yee, 2008] Yee, T. W. (2008). The `VGAM` package. *R News*, 8(2):28–39.

[Yee, 2015] Yee, T. W. (2015). *Vector generalized linear and additive models: with an implementation in R*, volume 10. Springer.

[Yee, 2021] Yee, T. W. (2021). On the Hauck–Donner Effect in Wald tests: Detection, tipping points, and parameter space characterization. *Journal of the American Statistical Association*, pages 1–12.

[Zens et al., 2021] Zens, G., Frühwirth-Schnatter, S., and Wagner, H. (2021). Efficient bayesian modeling of binary and categorical data in r: The `UPG` package. *arXiv preprint arXiv:2101.02506.*