Stockholms universitet

Prediction of graduation success and time to exam

A statistical analysis of the natural science bachelor's programs at Stockholm University

Hiam Shaba

Kandidatuppsats i matematisk statistik
Bachelor Thesis in Mathematical Statistics

Kandidatuppsats 2019:11
Matematisk statistik
Juni 2019

www.math.su.se

Matematisk statistik
Matematiska institutionen
Stockholms universitet
106 91 Stockholm

Stockholm University

Mathematical Statistics
Stockholm University
Bachelor Thesis **2019:11**
http://www.math.su.se

Prediction of graduation success and time to exam
A statistical analysis of the natural science bachelor's programs at Stockholm University

Hiam Shaba*

June 2019

Abstract

The aim of this thesis is to investigate which factors that have an impact on the probability to graduate and the time it takes to graduate at natural science bachelor's programs, given that the programs begin with the same basic course in mathematics, Mathematics I, at Stockholm University. The factors used are the time to finish their first course in mathematics and the grade in the course, the gender, the age of the student when starting the education and which bachelor's program the student is enrolled in. Students not following a program are also included in the model predicting the time to exam. The sample data consist of undergraduate students that were registered in the course Mathematics I from 2007 until January 2019. A model for the prediction of graduation success is computed with a binomial logistic regression model, where we investigate whether students obtain a bachelor degree within six years, using a data sample from year 2007 until 2012. The results of the analysis is that the time to finish Mathematics I, the bachelor's program, the grade in Mathematics I and the student's age when starting the education, are associated with the probability to graduate. Though, when visualizing the performance of the model the AUC-value indicate bad predictability. The 'time to degree'-model implements a gamma generalized linear model with a data sample of students that have completed their thesis given that they have completed Mathematics I, from year 2007 until 2017. In this model we include students taking stand-alone courses. The analysis of the time it takes to obtain a degree resulted in two gamma generalized linear models with an identity and logarithmic link function, and the time to finish Mathematics I as a significant explanatory variable.

*Postal address: Mathematical Statistics, Stockholm University, SE-106 91, Sweden. E-mail: hish0194@student.su.se. Supervisor: Tom Britton, Martin Sköld.

Acknowledgements

This bachelor's thesis in mathematical statistics covers 13.5 university points at the Department of Mathematics, Stockholm university.

I would like to acknowledge my supervisors Martin Sköld and Tom Britton and thank them for their commitment and valuable work in guiding and supporting me during the thesis. I would again like to thank Martin Sköld for providing me the data material used in the study.

Contents

1 Introduction

The natural science bachelor's programs at Stockholm University consist of students enrolled in programs but also those who frame their degree with stand-alone courses. Given the possibility to adapt the studies according to the student's needs, it results in students with a variety of aims. This thesis aims to firstly analyze if the students will finish their thesis or not, given that they enrolled the basic course in mathematics year 2012 and before, with a binomial logistic regression model. Secondly to predict the time it takes to finish the thesis with a gamma generalized linear model, given that they enrolled the basic course in mathematics year 2007 until 2017. The basic course in mathematics is Mathematics I, which constitutes of 30 university credits, given the first term or first year in the used bachelor's programs. The course is split into two parts which are analysis and linear algebra. The explanatory variables in both models are the time to finish Mathematics I, the bachelor's program the student is enrolled in with four or five levels depending on the model (including stand-alone courses in the 'time to degree'-model), the grade in Mathematics I with five levels, the student's age in the beginning of the education and the gender.

Zhong's thesis (2016) has a similar subject but she does her analysis only on students in the bachelor's program in mathematics and mathematics and economics, while this thesis analyze all natural science bachelor's programs at Stockholm University whose first course is Mathematics I. The author's models are also formed differently. She builds two generalized linear models that are binomial logistic regression models. The first model predicts the probability to obtain a thesis within three years while this thesis predicts the probability to finish the thesis within five to six years to include the delayed students, and the students that studies in half-time. Zhong's second model predict the time to finish the degree with a categorical response variable with four factor levels while this thesis uses a continuous response variable measured in years, for the gamma distributed generalized linear model.

The interest in this topic comes from institutions wanting to know the factors influencing students' course of studies and how large the impact is. The data material is given by the Department of Mathematics at Stockholm University from LADOK, a student administration system where student's merits and registrations are documented.

2 Description of the data material

The data material used to obtain the models in this thesis is a sample of 1381 undergraduate students from year 2007 until 2019 in January, that

started their program with Mathematics I. This is not the entire sample of 1629 observations which includes all students that are enrolled in programs, even those that are not bachelor's programs or stand-alone courses, and also includes students majoring in economics. In the sample of 1381 students we exclude all students that are not enrolled in bachelor's programs or stand-alone-courses, and also the students majoring in economics since it is not a subject of natural science. We are with this sample computing two models with different approaches and motives, so the data is tidied in different ways. In the binary model we investigate the probability that someone finishes the thesis given that he or she has finished Mathematics I. We filter the data such that we only use observations of bachelor's students that were registered in Mathematics I year 2012 and before. This is built on the assumption that the students that still have not finished their theses after five to six years on the bachelor's programs, most likely never will, including the students that studies in half-time. Note, that since we are investigating the probability to finish the bachelor's thesis, we assume that the students enrolled in bachelor's programs have intentions to finish their theses unlike the students taking stand-alone courses. Since we cannot be certain about their intentions, the later ones are not included in the analysis. This leaves us with 381 bachelor's students that have finished Mathematics I, to fit the model.

In the analysis of the time to finish the thesis we exclude all students that have not finished their thesis, since we are investigating the time to finish the thesis given that the student has finished Mathematics I. This leaves us with 361 students enrolled in bachelor's programs and stand-alone courses, that have finished their thesis. The time to finish the thesis is measured by taking the time difference of when the student is enrolled in the basic course until it has received a grade in the bachelor's thesis. It is important to stress that since we are only using observations with finished theses, some students of year 2015 and 2016 may have not gotten the chance to finish their theses which gives us skewed data. The students of those years might seem "lazy" compared to older classes. The variables will be presented in Table 1 below.

Table 1: List of variables

Variable	Type	Description
Time.Finish.MM2001	Continuous	Time to finish Mathematics I and obtain a grade in years
Program	Nominal	The bachelor's program the student is enrolled in
Grade.MM2001	Ordinal	The student's grade in Mathematics I, grade A to E
Age	Continuous	Age of the student when starting the education
Gender	Nominal	The student's gender
Finish.Thesis	Nominal	If the student has finished his/her thesis or not within five to six years
Time.Finish.Thesis	Continuous	Time to finish the thesis counting from registration date of Mathematics I until an approved thesis, in years

2.1 Explanatory variables

Both models that we will fit will contain the given explanatory variables Time.Finish.MM2001, Program, Grade.MM2001, Age and Gender from Table 1. Each independent variable will be shortly presented in the following subsections and the figures presented will contain both those that have finished the thesis, those who have not and the students taking stand-alone courses.

2.1.1 Bachelor's Program

The natural science bachelor's programs, at Stockholm University, are manually put into five categories: computer science, physics, mathematics, stand-alone courses and the rest which includes the bachelor's programs in astronomy, meteorology, biomathematics and biomathematics and computational biology. Table 2 presents the size proportions in each category and we can see that the students taking stand-alone courses consist of almost half of the data sample. They are followed by the student majoring in mathematics with 29% etc. Note that this data sample consists of all students that have enrolled and passed Mathematics I from year 2007 until 2019 with and without a finished bachelor's thesis.

Table 2: Proportion of students in each program

Program	Frequency	Proportion
Stand-alone	686	49.7 %
Mathematics	404	29.3 %
Physics	128	9.3 %
Computer Science	39	2.8 %
The Rest	124	9.0 %

2.1.2 Time to finish Mathematics I

The time to finish the first term basic course Mathematics I, of 30 university credits (hp), is a continuous variable, obtained by taking the difference in years between the date of registration and the date when the grade of the course is registered. Table 3 presents the median of time it takes to finish the basic course in each program. We see that physics and mathematics students take shorter time than the rest, and that students in computer science take the longest time. This is due to the computer science program have the basic course on half-time which is a year. Though, it is important to stress that we have only 39 computer science students which is 2.8% of the data. Among those we have 24 students whom have taken more than 1 year to finish this course and 18 of those students have not finished their thesis while 6 students have.

Table 3: Median time to finish Mathematics I in each program

Program	Time (years)
Physics	0.36
Mathematics	0.37
Stand-alone	0.41
Computer Science	1.78
The Rest	0.38

2.1.3 Grade in Mathematics I

The grades in Mathematics I is in five levels from A to E and Grade.MM2001 is a categorical variable of the ordinal type. Below in Figure 1 we find the box plots where the time to finish the thesis is plotted against the grades in Mathematics I. We can see that all plots are skewed and that there exist outliers in every grade of the course except for the grade C. Though, for the grade C, and the grade D, the data is more spread indicating that the time to finish the thesis vary more for the students with the given grades in Mathematics I. Note that a few students have finished their thesis in a very short amount of time were e.g. one student finished the thesis after 0.5 years and received the grade A. This student is a physics major and must have taken courses before being enrolled in the course Mathematics I.

The same goes to other students that have finished their thesis in less than three years. It is important to stress that you can earn a bachelor's thesis in e.g. mathematics or mathematical statistics by just taking the obligatory courses, and sufficient amount of courses in the subject, if you have other courses to include in your degree. It can also be the case that some are not finished with their bachelor's programs but have been able to write their thesis despite of that.

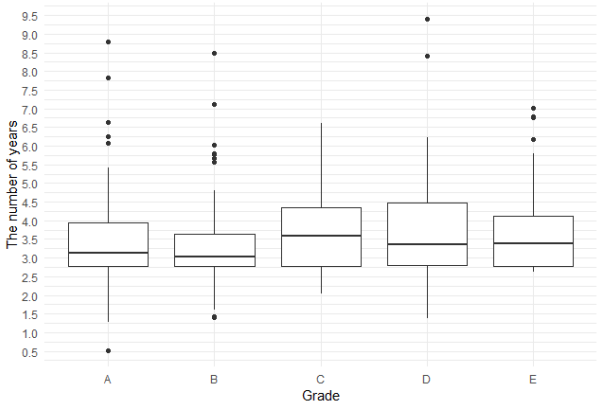


Figure 1: Time to finish the thesis vs. grade in MM2001

2.1.4 Age

Age is a continuous variable which gives us the age of the student when he or she was first registered in the course Mathematics I. The variable is obtained by taking the difference in years between the date of birth and the registration date of Mathematics I. By observing Figure 2 below we see that the majority of students are 19-20 years old when starting their education. This variable will be quite important to investigate when evaluating if the age of the student when they begin the education is important or not.

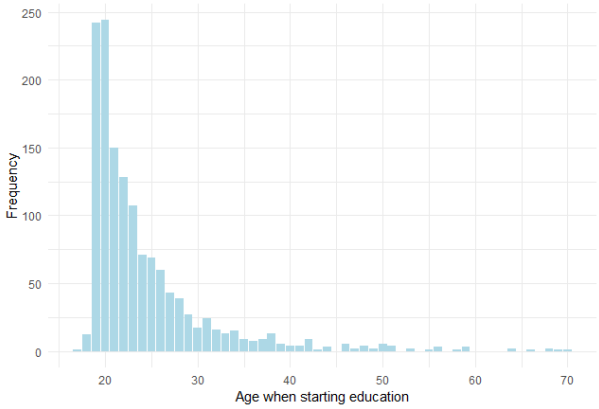


Figure 2: The distribution of age when starting the education

2.1.5 Gender

Gender is a nominal variable taking two values. The distribution of gender is presented in Table 4 below where it is shown that the students of the data sample of 1381 observations only consist of one third females. Looking at the box plots of Figure 3 we see that they are very much alike with medians close to each other. Also, note that there are more outliers of the male students which is self-evident since there are twice as many males as females.

Table 4: The distribution of gender

Gender	Proportion
Male	66.7 %
Female	33.3 %

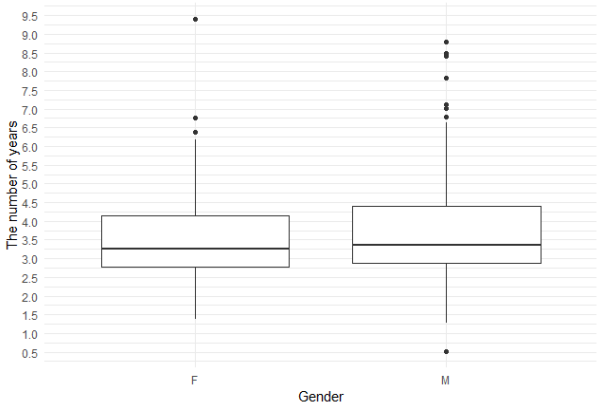


Figure 3: Time to finish the thesis vs. gender

3 Theory

3.1 Models

3.1.1 Generalized linear models

In the case of having non-normal response distributions *generalized linear models*, (GLM), are to be used as an extension of ordinary regression models (Agresti, 2012, p. 114). The GLM consist of three components which is the random response component Y, a systematic component related to the explanatory variables and a link function that links together the random and systematic component. The random component is the response variable with independent observations with the probability density function

$$f(y_i;\theta_i) = a(\theta_i)b(y_i)\exp[y_iQ(\theta_i)] \tag{1}$$

where θ_i varies depending on the outcome of explanatory variables for $i = 1, ..., N$ and $Q(\theta_i)$ is the natural parameter.

The systematic components illustrates a linear relationship between a vector $(\eta_1, ..., \eta_N)$ and explanatory variables (Agresti, 2012, p. 114). If x_{ij} is the value of the predictor j for observation i we obtain

$$\eta_i = \beta_0 + \sum_j \beta_j x_{ij} \tag{2}$$

which is called the linear predictor for $j = 1, ..., p$ and $i = 1, ..., N$.

The link function links $\mu_i = E(Y_i)$ to η_i by $\eta_i = g(\mu_i)$ when $i = 1, ..., N$, where g is the monotonic and differentiable link function (Agresti, 2012, p. 114). The function g has the formula

$$g(\mu_i) = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} \tag{3}$$

which links $E(Y_i)$ to the explanatory variables.

The first assumption made for the distribution is that y_i, given x_i, are conditionally independent, and the second assumption is that the distribution of $y_i|x_i$ is from a simple exponential family (Fahrmeir & Tutz, 2001, p. 19). The expected value will be $E[y_i|x_i] = \mu_i$ and the distribution may depend on a dispersion parameter ϕ.

3.1.2 Odds ratio

In a logistic regression when using categorical data we look at the odds and odds ratio (Agresti, 2012, p. 44). The odds is the proportion of a successful event and a failed event. Given a probability π of success the odds is

$$\Omega = \frac{\pi}{1-\pi} \tag{4}$$

where $\pi \in (0, 1)$. Having odds greater than one means that a success is more likely than a failure, and if they are less than one a failure is more likely to happen. The odds ratio is when you measure the difference in proportion to succeed for two groups like the following:

$$\theta = \frac{\Omega_1}{\Omega_2} = \frac{\pi_1/(1-\pi_1)}{\pi_2/(1-\pi_2)} \tag{5}$$

where Ω_1 is the odds ratio of group 1 and Ω_2 is the odds ratio of group 2. If the odds ratio is greater than one it means that group 1 is more likely to have success than group 2, and the opposite holds if the odds ratio is smaller than one. In the case of odds ratio being one, both groups will equally likely have success.

3.1.3 Logistic regression for the probability to finish the thesis

With a binary response variable Y and a vector of explanatory variables X, we use a logistic regression model to fit the data (Agresti, 2012, p. 119, 163-164). The assumption is that there exist a non-linear relationship between x and $\pi(x)$ which is defined as:

$$\pi(x) = \frac{\exp(\alpha + \beta x)}{1 + \exp(\alpha + \beta x)} \tag{6}$$

where the sign of β illustrates an increase or decrease of $\pi(x)$ as x increases. Hence, as x increases and $\beta > 0$, $\pi(x)$ increases, and the opposite holds for when $\beta < 0$. The probability that $Y = 1$ given that we have observed $X = x$ is denoted as $\pi(x) = P(Y = 1|X = x)$. Most often the non-linear relationship between $\pi(x)$ and x is monotonic meaning that a change in $\pi(x)$ is monotonic and continuous depending on x. Having a β close to zero or zero means that Y is independent of X. Using multiple predictors in the logistic regression we get the GLM and taking the logit of it we obtain the linear relationship

$$\text{logit}[\pi(\mathbf{x}_i)] = \log \frac{\pi(\mathbf{x}_i)}{1-\pi(\mathbf{x}_i)} = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} \tag{7}$$

which is the log odds for the explanatory variable j and observation i. The models have a logit link function and a binomial random component. The

monotonic and continuous relationship between $\pi(\mathbf{x}_i)$ and $\mathbf{x}_i$ gives us therefore a linear relationship between the log odds and the data. We know that the logit can be a real number since $\pi(\mathbf{x}_i) \in (0, 1)$.

The odds are an exponential function of $\mathbf{x}_i$ which we can see by exponentiating the logit function of equation (7). Hence, e^{β_j} is the odds of the explanatory variable $\mathbf{x}_j$, meaning that the odds increase multiplicatively by e^{β_j} for every increase in the unit of the explanatory variable $\mathbf{x}_j$. In this case, we will use the binomial GLM with the logarithmic link function $g(\mu_i) = \log(\mu_i)$.

3.1.4 Multiple regression for the time to finish the thesis

A multiple regression model with a response variable y_i and explanatory variables x_{ij} for observation i and variable j has the following expression

$$y_i = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} + \epsilon_i \tag{8}$$

where ϵ_i is the error term for observation i (Sundberg, 2016, p. 64). In this case the response variable is $y_i = \text{Time.Finish.Thesis}_i$.

3.1.5 Assumptions

When constructing a regression model there are some assumptions to make (Gelman & Hill, 2007, p. 45). Following are the assumptions, listed in a decreasing order of importance:

1) Validity: appropriate data is used to solve the problem.

2) Additivity and linearity: Additivity is one of the most important assumptions for a regression model where the effects of the explanatory variables on the expected value of the dependent variable are additive. Hence, the explanatory variables' influence on the response variables is independent of other influences, like the intercept. If this is violated transformations can be used. There should be a linear relationship between the response variable and the predictors.

3) Independence of errors: The error terms should be independent.

4) Homoscedasticity: The error terms should be constant meaning $Var(\epsilon_i) = \sigma^2$.

5) Normality: All error terms should be normal distributed i.e. $\epsilon_i \in N(0, \sigma^2)$.

3.1.6 Gamma distribution

In the case of having continuous and positive variables the gamma distribution is appropriate when computing a regression analysis (Fahrmeir & Tutz, 2001, p. 23). The density will be

$$f(y|\mu,\nu) = \frac{1}{\Gamma(\nu)}(\frac{\nu}{\mu})^{\nu} y^{\nu-1} \exp{(-\frac{\nu}{\mu}y)}, y \geq 0 \tag{9}$$

where $\mu > 0$ is its mean and the shape parameter is $\nu > 0$ which gives us the form of the density. The dispersion parameter is $\phi = 1/\nu$. Depending on the shape parameter we will obtain the following different cases:

$$f(y) = \begin{cases} \text{decreases monotonically,} & \text{if } 0 < \nu < 1 \\ \text{exponential distribution,} & \text{if } \nu = 1 \\ \text{0 at y=0 or mode at } y = \mu - \mu/\nu \text{ and positively skewed,} & \text{if } \nu > 1 \end{cases} \tag{10}$$

In case of the gamma distribution the link function can be the inverse link function i.e. $g(\mu_i) = 1/\mu_i$, the identity meaning $g(\mu_i) = \mu_i$ and the log-link meaning $g(\mu_i) = \log(\mu_i)$.

3.2 Selection of models

The method of choosing a model from data that reflects the true outcome, and minimize the number of variables, is to create a more numerically stable model (Hosmer & Lemeshow, 2013, p. 90). The standard error gets larger the more variables included in the model which makes the model more dependent on the observed data. These methods differs depending on if it is a logistic regression model or another regression model. Hence, in the following sections we will present different approaches depending on the type of model. We will implement *purposeful selection* and *stepwise procedures*.

3.2.1 Purposeful Selection

Purposeful selection is used in logistic regression and consists of seven steps normally used when choosing a model (Hosmer & Lemeshow, 2013, p. 90). According to Bursac et al. (2008) purposeful selection works well for smaller sample sizes i.e. 240-600 observations, like in this study. This method is preferred since it allows the statistical analysts to control every step. In this way the analyst can find the variables that influence the dependent and

explanatory variables and cause false relation, i.e. confounders.

Step 1
The method starts with an analysis of the explanatory variables. In the case of categorical variables a contingency table is to be made for each variable, with frequency counts where the outcome $y = 0, 1$ is analyzed against the k levels of the explanatory variable. Pearson's chi-square test of independence (H_0: independence) will be used in the test since it is asymptotically equivalent to the likelihood ratio chi-square test, were both have $k - 1$ degrees of freedom. For the continuous variables the likelihood ratio test will be used to test each variable. In both cases we will use the significance level of 25% since variables can be insignificant in the initial stage but later on show significance to the model together with other variables. The equation of the Pearson chi-square statistics is (Agresti, 2012, p. 75)

$$\chi^2 = \sum_i \sum_j \frac{(n_{ij} - \hat{\mu}_{ij})^2}{\hat{\mu}_{ij}} \tag{11}$$

where $\hat{\mu}_{ij} = n\hat{\pi}_{i+}\hat{\pi}_{+j} = n_{i+}n_{+j}/n$, $\hat{\pi}_{i+} = n_{i+}/n$ and $\hat{\pi}_{+j} = n_{+j}/n$ for row i and column j. The equation of the likelihood ratio statistics is

$$G^2 = -2\log\Lambda = 2\sum_i \sum_j n_{ij}\log(n_{ij}/\hat{\mu}_{ij}) \tag{12}$$

Step 2
Now a model with the significant variables from step 1 is fitted (Hosmer & Lemeshow, 2013, p. 91). We investigate what variables that are not significant on a 5% significance level according to the Wald statistic and exclude these. Thereafter we use the partial likelihood ratio test to test this model against the one we obtained from step 1. The partial likelihood ratio test is the following (Agresti, 2012, p. 11)

$$-2(L_0 - L_1) \tag{13}$$

where L_0 and L_1 are the maximized log-likelihood functions.

Step 3
Now we will compare the models' values of the estimated coefficients from step 2, i.e. the smaller model with significant variables with the larger model (Hosmer & Lemeshow, 2013, p. 92). We will compute the following ratio (Hosmer & Lemeshow, 2013, p. 67)

$$\Delta\hat{\beta}\% = 100\frac{\hat{\theta}_i - \hat{\beta}_i}{\hat{\beta}_i} \tag{14}$$

where $\hat{\theta}_i$ is the coefficient estimate i of the smaller model and $\hat{\beta}_i$ is the coefficient estimate i of the larger model. They state that if the $\Delta\hat{\beta}$ is larger than 20% then there is a risk that the excluded variables are important to the other variables as a group, and should be added back.

Step 4
In this step each variable that was insignificant in step 1 will be added back one at a time, and the significance will be checked with the Wald statistic p-value. In the case of categorical variables with more than two levels the likelihood ratio test will be used. This is done due to investigate which variables that contribute more in presence of other variables than individually.

Step 5
This step consists of examining the continuous variables, and check if the logit increases or decreases linearly against the variable. This can be computed by splitting up the data in four categories based on the data's quartile (Hosmer & Lemeshow, 2012, p. 95). Thereafter we compute a categorical variable with the four levels, the lowest one is used as reference level. Then, a new model is created where the continuous variable is replaced by the categorical version. We plot the estimated factor levels of the coefficient against the midpoints of the three upper quartiles. Additionally, we plot a coefficient that euqals zero against the first quartile. By observing the plot and comparing the models we decide what variable is more appropriate.

Step 6
We will now check if there are two-way interactions among the variables in the model we obtained from step 5 (Hosmer & Lemeshow, 2012, p. 92). The interaction terms should be added one at a time, and their significance tested with a likelihood ratio test at a 5% significance level. Thereafter we redo step 2 to get a simpler model.

Step 7
The final step is to check if the model fits the data well and the models prediction ability. This will be done by Hosmer-Lemeshow test, AUC and ROC-curves and also k-fold cross-validation, which will be discussed more thoroughly in the next sections.

3.2.2 Goodness of fit

The *goodness of fit* illustrates how well a certain model can predict the data (Agresti, 2012, p. 173). A logistic regression model can be tested with Pearson's chi-square test and the likelihood ratio G^2-test. When the data is ungrouped the statistics will not converge leaving us with the *Hosmer-*

Lemeshow test. In this test, using the ungrouped original data and according to the estimated probabilities of success, we partition observed and fitted values. They should have equal sizes, and most often they are partitioned in ten groups g that are ordered in a rising order. Then the partitions are compared with the observed and fitted counts. Having a binary outcome, y_{ij} in group i for observation j, and the fitted probabilities for the model for ungrouped data, $\hat{\pi}_{ij}$, we get the statistic K

$$K = \sum_{i=1}^{g} \frac{(\sum_j y_{ij} - \sum_j \hat{\pi}_{ij})^2}{(\sum_j \hat{\pi}_{ij})[1 - (\sum_j \hat{\pi}_{ij})/n_i]} \tag{15}$$

where $j = 1, ..., n_i$ and $i = 1, ..., g$. The test is approximately chi-squared distributed with degrees of freedom $g - 2$, and tests the null hypothesis if the model fits the data well.

3.2.3 ROC-curve

To measure a model's prediction ability we use the *Receiver Operating Characteristic Curve* (Agresti, 2012, p. 224; Hosmer and Lemeshow, 2013, p. 173). This is a plot of true signal, sensitivity, against the probability of false signal which is computed by (1 − specificity) for range of possible cutoffs. Since classification tables are less informative than ROC-curves they will not be applied, this is due to the ROC-curve that summarizes predictive powers for all possible cutoffs. The sensitivity, also called *the true positive rate*, is defined as the probability of a correct classification when $y = 1$, so the predicted value is successful given that the true value is successful. *The false positive rate* is another definition of 1-specificity and is the probability of an incorrect classification when $y = 0$, so the predicted value is failed given that the true value is failed. Plotting these against each other it will result in a convex curve where the area between this curve and a straight line going through $(0, 0)$ and $(1, 1)$ yields the AUC, *the Area Under Curve*, evaluating the model's ability of prediction. This is illustrated in the figure below taken from Agresti (2013, p. 225).

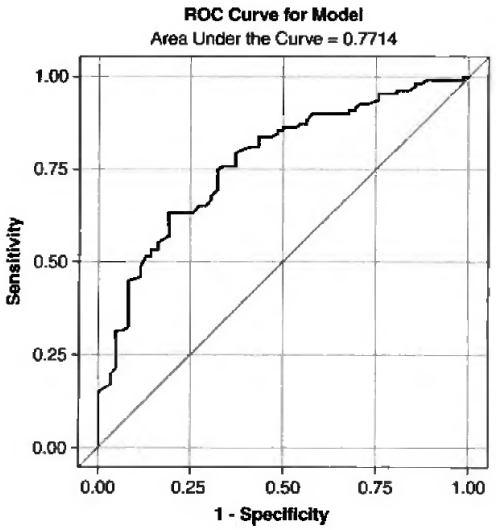


Figure 4: Example of a ROC-curve with AUC

According to Hosmer and Lemeshow (2013, p. 177) there are some general guidelines that interprets the given AUC-value which is presented in the following table:

AUC:	**Discrimination:**
0.5	No
(0.5, 0.7)	Poor
[0.7, 0.8)	Acceptable
[0.8, 0.9)	Excellent
≥ 0.9	Outstanding

3.2.4 K-fold Cross-validation

Cross-validation is very often used to estimate the prediction error (Hastie et al., 2009, p. 241). When splitting the data set into K data sets, we use $K-1$ part of it fitting the model and the remaining to test the model's fit to the data. Here the model's prediction error is calculated in the prediction of the kth part. Usually k is five or ten and it's then called *k-fold cross-validation*. The procedure is repeated K times so each fold can be used as the test set. For this type of cross-validation the method estimates the expected error. With 5-fold the method has lower variance but depending on the training set's size, it can be biased. By looking at the performance of the method in relation to the size they come to the conclusion that the training set of size 40 is not good for 5-fold cross-validation but with 160 observations the bias would lower. With higher data sets the conclusion would be the same.

3.2.5 Variance Inflation Factor (VIF)

The *variance inflation factor* is used in a correlation analysis to examine collinearity between the explanatory variables in both linear, gamma GLM and logistic regression models (Sundberg, 2016, p. 73). It shows how much larger variance each variable has in a model with other variables, than if it would be the only explanatory variable in the model, or if it would be orthogonal to the rest. The variance of a regression coefficient is expressed as

$$\mathrm{Var}(\hat{\beta}_j) = \frac{\sigma^2}{s_{jj}^2}\mathrm{VIF} \tag{16}$$

where the VIF-factor is expressed as

$$\mathrm{VIF} = \frac{1}{1 - R_j^2} \tag{17}$$

where R_j^2 is the coefficient of determination for variable x_j in relation to the other variables. This coefficient is defined as the portion of the variance that is explained by the model (Sundberg, 2016, p. 69). The following is the equation of the coefficient of determination

$$R_j^2 = 1 - \frac{SSE}{SST} = 1 - \frac{\sum_i (y_{ij} - y_{ij}^*)^2}{\sum_i (y_{ij} - \overline{y_j})^2} \tag{18}$$

where SSE is the sum of squared errors, SST is the total sum of squares, $y_{ij} - y_{ij}^*$ is the residuals and $\overline{y_j}$ is the mean.

3.2.6 Akaike Information Criterion (AIC)

The *Akaike information criterion* is another way of evaluating different models (Agresti, 2012, p. 212). The criterion uses the maximized log-likelihood function and the model's degrees of freedom. So with this criterion you get how close the fitted values are to the true mean values. It is given by

$$\mathrm{AIC} = -2(\log(\hat{L}) - k) \tag{19}$$

where $\hat{L}$ is the maximized likelihood-function and k is the number of parameters of the model.

3.2.7 Forward selection

The following three model selection procedures are the ones that can be both used for linear models, gamma GLM and logistic regression models. Forward selection starts its algorithm with the intercept only and the model is expanded with one variable every time (Sundberg, 2016, p. 71). The explanatory variable chosen into the model is the most significant one or

the one that yields the smallest AIC-value for the model. This study will compute the model selection based on the AIC and not the p-value. So the variables are chosen depending on the model with the most minimized AIC value.

We are using the R-function step() which does this procedure based on the models' AIC. This function is used for the stepwise and forward selection, and backward elimination for both models in this thesis.

3.2.8 Backward elimination

This method assumes that we have a model with all its explanatory variables (Sundberg, 2016, p. 71). In every step the variable that gives the model the largest AIC-value is removed, and the model is compared against other reduced models. If multiple variables give us relatively high AIC-values the method chooses the variables that give us the model with the lowest AIC. The procedure is repeated until we obtain the model with the lowest possible AIC.

3.2.9 Stepwise selection

In the stepwise selection we start with an empty model and its intercept (Sundberg, 2016, p. 71). In every step the algorithm apply an explanatory variable that gives the model a lower AIC value, as in forward selection, but it also searches for the model with the lowest AIC-value considering the variables that are already applied. Hence, if the model gets a higher AIC-value with the newly applied variable but only in relation to one already applied variable, the old variable will be eliminated.

4 Analysis

4.1 Logistic regression for the probability to finish the thesis

The model built to evaluate if a natural science bachelor's student at Stockholm University graduates or not, given that they have finished the course Mathematics I, is a generalized linear model in the binomial family. Hence, a logistic regression model will be built with the explanatory variables Time.Finish.MM2001, Program, Grade.MM2001, Gender and Age, with the response variable Finish.Thesis that gives us 1 for a student that has finished its thesis within six years, or 0 for the one that has not. The variables Program, Grade.MM2001 and Gender are factors which means that one level of each variable is chosen as reference level.

4.1.1 Purposeful selection

To evaluate which model has the best predictability, and fits the data best, we will implement purposeful selection.

Step 1: The first step consists of an univariate analysis of the explanatory variables. We will first test the continuous variables with the null hypothesis $H_0 : \beta_i = 0$, of the i:th variable, with the log likelihood test. Both Time.Finish.MM2001 and Age were significant on a 25% level. To test the categorical variables with chi-square tests we compute contingency tables, presented in Appendix A Table 1-3, that result in the variable Gender being insignificant on a 25% level. The table below presents the univariate analysis of the variables:

Table 5: The p-values obtained from the tests

Parameter	P-value
Time.Finish.MM2001	0.0133
Age	0.0902
Program	0.00764
Grade.MM2001	0.00114
Gender	0.616

We proceed to step 2 with a multiple logistic regression model excluding the gender variable.

Step 2: Now we use the model with the explanatory variables that were significant on a 25% level from step 1, the result is presented in Table 6. Looking at the Wald statistics we see from the p-values that all variables are insignificant except for a factor level of Grade.MM2001 and two factor levels of Program, therefore we cannot exclude the variables. We compute a likelihood ratio test to see if the data needs the other variables by testing a model, without these variables, against a model including them. Hence, we fit a new model with only Program and Grade.MM2001, since all the other variables were insignificant in our multiple logistic regression model. We test this model with the likelihood ratio test against the model we obtained from step 1. The result indicates that we cannot reject the null hypothesis on a 5% level since the test's p-value was 9.8%. This means that a reduced model may be appropriate for the data so we continue on with the reduced model that only includes the variables Program and Grade.MM2001, and the result is presented in Table 7 in Step 3.

Table 6: The multiple logistic regression model excluding Gender

Parameter	Levels	Coefficient	SE	P-value
Intercept		0.257	0.604	0.671
Time.Finish.MM2001		-0.139	0.0957	0.148
Program	Physics	0.613	0.268	0.0223*
	Computer science	0.435	0.583	0.455
	The Rest	0.851	0.272	0.00173 **
	Mathematics	1.0	-	-
Grade.MM2001	A	0.931	0.321	0.00369 **
	B	0.109	0.327	0.738
	C	0.146	0.317	0.646
	D	1.0	-	-
	E	-0.558	0.399	0.162
Age		-0.0363	0.0244	0.138

Step 3: In this step we will compare the variables we have in both models and evaluate if the coefficients differ more than 20%. The two variables that are in the smaller model, which is presented in Table 7, are Program with 4 levels, were the bachelor's program in mathematics is used as reference level, and Grade.MM2001 with five levels, were the grade D is used as reference level. The old model, presented in Table 6, is the model we compare the reduced model against. The calculations of the proportional differences in the coefficients between the models are the following:

$$\Delta\hat{\beta}_{Physics}\% = \frac{0.643 - 0.613}{0.613} * 100 = 4.89\%$$

$$\Delta\hat{\beta}_{Comp.Science}\% = \frac{0.0829 - 0.435}{0.435} * 100 = -80.94\%$$

$$\Delta\hat{\beta}_{Rest}\% = \frac{0.854 - 0.851}{0.851} * 100 = 0.35\%$$

$$\Delta\hat{\beta}_{A}\% = \frac{1.010 - 0.931}{0.931} * 100 = 8.49\%$$

$$\Delta\hat{\beta}_{B}\% = \frac{0.198 - 0.109}{0.109} * 100 = 81.65\%$$

$$\Delta\hat{\beta}_{C}\% = \frac{0.206 - 0.146}{0.146} * 100 = 41.10\%$$

$$\Delta\hat{\beta}_{E}\% = \frac{-0.568 - (-0.558)}{(-0.558)} * 100 = 1.80\%$$

We see a large proportional difference in the bachelor's program in computer science where the proportion is approximately 81%, though negative. So the difference is larger than 20% but not for the bachelor's programs in

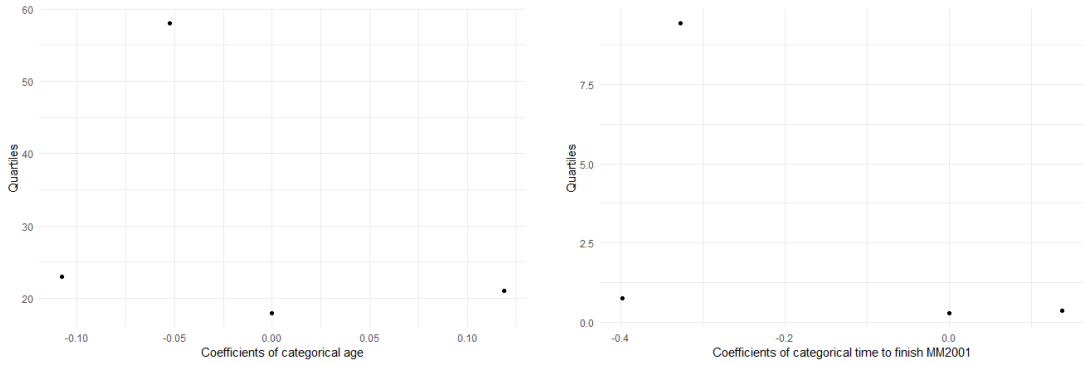
physics and the rest. The grades in the course Mathematics I show that two grades, B and C, have differences larger than 20% but not in the other two cases, for the grades E and A. This creates a difficulty in drawing a conclusion about the models. Though, since the mentioned factors indicate that there is a risk that the excluded variables are important to the data, we choose to add them back. Not the same conclusion is drawn in step 2 with the LR-test but since the proportional differences are so large for the coefficients we choose to continue with the model that only excludes the gender.

Table 7: The multiple logistic regression with Program and Grade.MM2001

Parameter	Levels	Coefficient	SE	P-value
Intercept		-0.707	0.247	0.00424 **
Program	Physics	0.643	0.267	0.0158*
	Computer science	0.0829	0.529	0.876
	The Rest	0.854	0.270	0.00155 **
	Mathematics	1.0	-	-
Grade.MM2001	A	1.010	0.313	0.00125 **
	B	0.198	0.321	0.537
	C	0.206	0.313	0.510
	D	1.0	-	-
	E	-0.568	0.400	0.151

Step 4: In the univariate analysis in step 1 the variable Gender was not significant. By adding it back to the model it will still be insignificant on a 5% level. We continue to the next step with the model excluding Gender.

Step 5: In this step we will examine the continuous variables by controlling that the logit increases or decreases linearly against the variable. We start off by computing the quartiles and by those split up the data in categories, using cutpoints based on the quartiles. This gives us a four factor level variable where the first quartile will be the reference group. Thereafter, we plot the three categorical factor levels of Age against the upper three quartiles, and a coefficient that equals zero against the first quartile, and get Figure 5a. By the figure it is obvious that they are not linear but scattered, so we compute a likelihood-ratio test of the model with categorical age against the one with continuous age. The test shows insignificance which indicates that a reduced model is more appropriate for the data. Hence, we continue with using the continuous age in our model. The same approach was applied to test the second continuous variable in the model, Time.Finish.MM2001. With a scattered plot, Figure 5b, and an insignificant p-value we draw the same conclusion as for the age, that we keep the continuous variable and proceed with the model from step 4.



(a) (b)

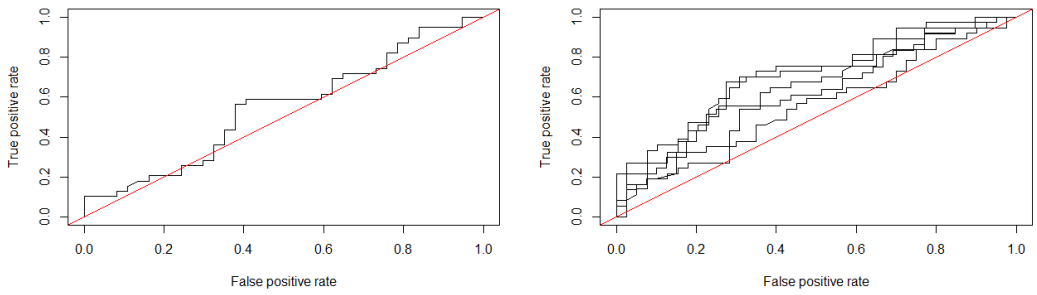
Figure 5: Categorical age coefficients vs. quartiles to the left (a) and categorical time to finish MM2001 coefficients vs. quartiles to the right (b)

Step 6: In this step we will check for all possible two-way interactions among our main effects. The four main effects result in six two-way interaction terms. By adding them one at a time we get no significant likelihood ratio test at a 5% significance level so we continue with the model that has only excluded the gender-variable.

Step 7: In this step we evaluate the model's prediction ability. First we will do this by the AUC-value and the ROC-curve. We will sample out 80% of the data for the training sample and the rest for testing, which is computed once. The AUC with this approach is 0.548 meaning that the model has a poor ability to predict the data and is almost making random guesses. By computing a 5-fold cross validation we obtain the AUC-value 0.645 for our last model which is larger than what the sampled data obtained, but still a poor prediction. The 5-fold cross validation is computed by splitting up the data into five data sets, use four sets to fit the data and the remaining to test the model's fit. Hence, the procedure is computed five times where each fold is used as both train and test set. Due to the small data sample we could not have a validation set. Since the 5-fold cross-validation has lower variance and bias, this method is more reliable than the sampling method described before.

The ROC-curve for the sampled data, Figure 6a, is presented below along with the ROC-curves for the cross validated data in Figure 6b. Figure 6b is obtained by printing all five folds from the cross-validation and the AUC is also obtained from the cross validation. The R-package and functions used for this procedure is the package caret with the functions trainControl and

train. The Hosmer-Lemeshow test shows that the model does not fit the data well since we got a small p-value, i.e. a significant result. This means that the difference between the model and the observed data is significant so the results can be misleading. Though, we will continue with the stepwise procedures and thereafter conclude which model to choose.

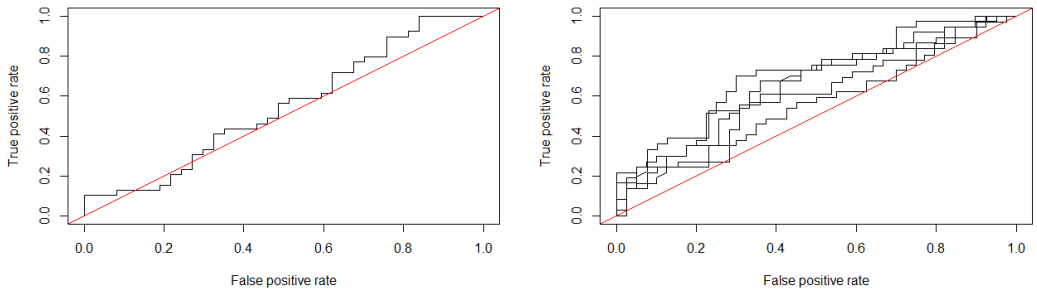


(a) (b)

Figure 6: ROC-curves from sampled data with AUC 0.548 to the left (a) and 5-fold cv data with AUC 0.645 to the right (b)

4.1.2 Stepwise regression

Now we will use stepwise selection, backward elimination and forward selection to fit a model. By using the AIC as a measure for choosing the right model and choosing the model with the lowest AIC we come to the conclusion that both stepwise selection and backward elimination give us the same model as purposeful selection, while forward selection give us the original full model. The AUC with sampled data for the full model will be 0.545 which is slightly lower than the AUC for the reduced model and still a bad prediction. By using the 5-fold cross-validation the AUC will again be higher than the AUC obtained from sampled data, with an AUC of 0.637, which is lower than the AUC of the reduced model. This is again a poor value but better than the AUC we obtained by sampling the data. Since 5-fold cross validation has lower variance and bias in this case, its result is more reliable than the sampling method. The ROC-curves for the full model are presented in Figure 7a and Figure 7b below. Figure 7b is obtained in a similar manner as Figure 6b.



(a) (b)

Figure 7: ROC-curves from sampled data with AUC 0.545 to the left (a) and 5-fold cv data with AUC 0.637 to the right (b)

The Hosmer-Lemeshow test gives us a p-value higher than 5% i.e. an insignificant result for the full model, meaning that we have no significant difference between the model and the observed data. Hence, the model fits the data well. Conversely, the AUC-values have been slightly higher for the reduced model than the full model. It is also important to stress that since three out of four used model selection methods resulted in the reduced model, and that we aim to obtain the most predictive model based on the AUC, we can conclude that our reduced model is more appropriate for the data material despite the outcome of the Hosmer-Lemeshow test.

4.2 Multiple linear and gamma regression for the time to finish the thesis

The multiple regression model that evaluated the students' time to finish the thesis given that they have finished the course Mathematics I is either a linear or a gamma distributed model. The model will be built with the explanatory variables Time.Finish.MM2001, Program including stand-alone courses, Grade.MM2001, Gender and Age, with the response variable Time.Finish.Thesis which consists of continuous, positive decimal numbers measured in years.

4.2.1 Correlation analysis

In the first stage when evaluating data material with a continuous dependent variable and multiple explanatory variables, we have to investigate if they are collinear or multicollinear, meaning that they are linearly predicted by other variables. Firstly we will compute Pearson's correlation coefficient of the continuous variables of both the dependent and independent variables. To analyze the categorical variables' collinearity we create contingency tables of

two variables at a time and compute chi-square tests to see if they are independent or have some kind of correlation. The continuous variables show no severe correlation but Time.Finish.Thesis and Time.Finish.MM2001 has a weak correlation of 0.31. The other variables have even weaker correlations, where Time.Finish.Thesis and Age have a negative correlation of -0.026, and Time.Finish.MM2001 and Age have a correlation of 0.0078. The reason for the higher but weak correlation between the time to finish the thesis and Mathematics I could be that both variables' starting point is when the student enrolled the course Mathematics I.

Before computing the chi-square tests we create contingency tables for two categorical variables at a time which are presented in Appendix A Table 4-6. The chi-square test between Program and Grade.MM2001 is insignificant meaning that we cannot reject the null hypothesis that they are independent. The test between Gender and Program is significant so we reject that they are independent and therefore correlation can exist between these variables. The chi-square test between Gender and Grade.MM2001 show insignificance meaning that we cannot reject that they are independent. To check multicollinearity we use the variance inflation factor which shows generalized vif-values not larger than 1.3 for the linear model, the values are presented in Appendix B. The correlation between a continuous and a categorical variable is more complicated since the correlation measures if there is a linear relationship between two variables, which cannot be the case.

4.2.2 Stepwise regression

The data is fitted in a linear model that even with transformations cannot show evenly spread observations around zero of the fitted values against the residuals of the model, which is used to show linearity and non-constant error variances. The explanatory variable is not linear to the independent variables either. We start off by implementing backward elimination, stepwise and forward selection with AIC to reduce the variables, and again test the linearity-assumptions. Forward selection gave us the full model. Both backward elimination and stepwise selection gave us a model with only Time.Finish.MM2001 as an explanatory variable, since that was the model with the lowest AIC. Even then the fitted values against the residuals did not imply linearity nor constant variance of error terms, so the conclusion can be drawn that the error terms are not normally distributed. We try to transform the dependent variable in the reduced models by taking the logarithm or the square root of the dependent variable, but it does not confirm the assumption of the error terms normality nor constant variance. Since the dependent variable is positive, skewed and continuous we can compute a generalized linear model with the gamma distribution as stated in section

3.1.6. The skewness is presented in Figure 8 below which shows that the time to finish the thesis is right skewed i.e. positively skewed. This means that the mean is greater than the median which is a correct illustration since the mean is 3.5 years while the median is 3.1 years. Therefore we continue with gamma distributed generalized linear models. Also, note the discussion we had in section 2.1.3 about outliers. It is important to stress that some students that have finished their thesis in less than three years, have been able to due to the opportunity to include courses in other subjects in their degrees. We do not have information about this manner, but since the opportunity is given we can conclude this.

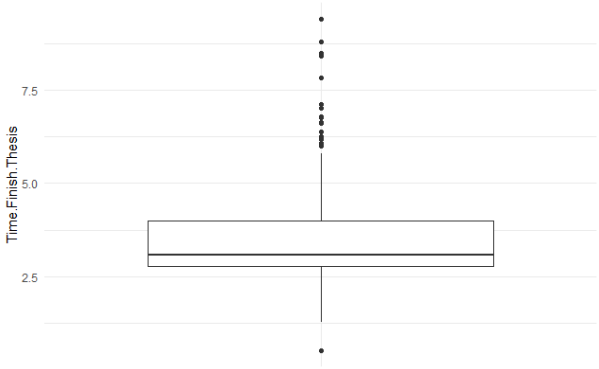


Figure 8: Boxplot of time to finish the thesis

Computing the gamma distributed model with the identity, the logarithmic and the inverse link-functions, it gives us models presented in Appendix C Table 8-10. From the tables we see that the AIC-values are close to each other but the model with the identity link function has the lowest AIC-value. The plots from Appendix D illustrates the fitted values against the residuals of each of these models in Figure 9a, 10a and 11a. The result is skewed but since the assumed distribution is the gamma distribution we cannot draw the conclusion that the residuals should be normal. Therefore we fit a model where the response variables are simulated from the gamma distribution, where the shape and scale parameters are determined from the models we want to compare the simulated residuals against. The result, presented in Appendix D, of the gamma GLM with inverse link function show no similarities since the simulated residuals in Figure 11b are more spread to the right while the model's residuals are in a cluster to the left. The other two models' simulated residuals, Figure 9b and 10b, show similar patterns as the models' residuals so we draw the conclusion that we can proceed in investigating the two gamma distributed models with the identity and logarithm link functions.

We will now use stepwise and forward selection and backward elimination to examine which explanatory variables contribute to the models depending on the AIC-value. According to backward elimination and stepwise selection the GLM with the lowest AIC value for both the logarithm and identity link function is the one consisting of only Time.Finish.MM2001 as a dependent variable. Obviously, the models have different coefficients and AIC-values due to different link-functions and these are presented in Appendix C, Table 11-12. Forward selection choose the full model for both link functions leading us to investigate four models. Firstly, we will start by plotting the residuals of the two reduced models and the simulated residuals for each model. Looking at Appendix D, Figure 12a-b and 13a-b, we see that the real and simulated residuals are spread or clustered in a similar way for both models. Hence, we continue with all four models to look closely on their AIC-values. As stated, in Appendix C, Table 8-9 and Table 11-12, the reduced model with the identity link function has the lowest AIC of 1095.1 so that could be our last model. However, due to the minor differences in the AIC value both link functions are appropriate to the data. Since both stepwise selection and backward elimination resulted in a model with only Time.Finish.MM2001 as explanatory variable we can conclude that the reduced models fit the data better than the full model.

5 Results

5.1 Interpretation of the logistic regression model for the probability to finish the thesis

With a binomial logistic regression analysis we obtained the model presented in Table 6 to predict graduation success. This model was given by both purposeful selection and the stepwise procedures except for the forward selection. Hence, the variables that are significant to predict graduation success are the time to finish the course Mathematics I, the bachelor's program the student is enrolled in, the grade in Mathematics I and the student's age when first enrolled in Mathematics I. The predictive power of this model, with and without cross validation was poor according to section 3.2.3. With just sampled data we obtained an AUC of 0.548 which is very poor and indicate that the model is almost as bad as random guesses. The 5-fold cross validation gave the AUC-value 0.645 which is higher than the AUC of sampled data but still poor. Though, since 5-fold cross validation give us a model with smaller variance and bias, the later AUC is more reliable. For the AUC to be on an acceptable level it should have been at least 0.7. The Hosmer-Lemeshow test show significance, meaning that this model does not fit the data sample which can be misleading. Though, as mentioned, since three out of four methods chose this model, and it has the highest predictive power among the models, we choose it. The model's odds ratios and their corresponding

confidence intervals are presented in Table 8 below:

Table 8: Frequency and odds ratios of the last model

Parameter	Levels	Frequency	Odds ratio	95 % Wald	Confidence Interval
Time.Finish.MM2001		381	0.871	0.722	1.050
Program	Physics	97	1.846	1.0912	3.123
	Computer science	17	1.545	0.493	4.846
	The Rest	96	2.342	1.375	3.988
	Mathematics	171	1.0	-	-
Grade.MM2001	A	90	2.538	1.353	4.758
	B	71	1.116	0.588	2.116
	C	76	1.157	0.621	2.155
	D	100	1.0	-	-
	E	44	0.572	0.262	1.250
Age		381	0.964	0.919	1.0117

5.1.1 Time to finish Mathematics I

The odds ratio for the continuous time to finish the course Mathematics I is 0.871 which means that for each increase in year the odds of finishing the thesis is 0.871 with the confidence interval $(0.722, 1.050)$. Hence, there is a 12.9% decrease in the odds of finishing the thesis for each year. We can interpret this as that students who take longer time to finish Mathematics I are less likely to finish their thesis. Note that the Wald confidence interval is obtained by taking the exponential of the interval $\hat{\beta} \pm z_{\alpha/2}(SE)$.

5.1.2 Bachelor's program

The odds ratio for each level of the categorical variable Program is in relation to the reference level Mathematics. This means that the students majoring in physics have a 84.6% increase in the odds of graduating in relation to the mathematics students. The students majoring in computer science have a 54.5% increase in the odds of finishing the thesis compared to the mathematics students. The students majoring in all the other program, the Rest, have a 134.2% increase in the odds of finishing the thesis in relation to the mathematics majors. This can also be seen in the coefficients presented in Table 6 were all factor levels have positive coefficients, so all levels in relation to mathematics majors have an increase in odds of graduating. The significant levels are Physics and The Rest. Since the reference level is mathematics, which is the largest bachelor's program group, the groups of physics majors and the rest are significant since they are as well large groups. The computer science students are few, only 17 students, which in comparison to the

mathematics group is insignificant.

5.1.3 Grade in Mathematics I

The odds ratio for the levels of the grade in the course Mathematics I are in relation to the grade D in the course. This means that the odds of graduating in relation to the students getting a D in Mathematics I is an 153.8% increase in the odds for the students getting an A, an 11.6% increase in the odds for students getting a B, an 15.7% increase in the odds for students getting a C and a 42.8% decrease in the odds for students getting an E. This can also be observed in the coefficients presented in Table 6 where only the factor level of the grade E is negative. The only significant level is the grade A and among those 24% are mathematics majors, 35% are physics majors, 18% are computers science majors and the rest are 13%, measured in relation to their corresponding bachelor's program. So each percentage is computed by taking the number of students with the significant grade in each bachelor's program and divide it with the total number of students in that program. Hence, the largest proportion of group with the highest grade in Mathematics I is the physics-group which is a significant factor level. The second largest proportion in the programs with the highest grade in Mathematics I are the mathematicians but they are the reference level. It is important to stress that these are the largest groups among the bachelor's programs and with the highest percentage of students that have received an A in the course Mathematics I.

By observing Table 13 in Appendix C, we can see that the bachelor's programs with the largest proportion of graduates are in mathematics, physics and the rest-group. So it is self-confident that the physics and the rest factor levels will be significant in the bachelor's program-variable. The physics majors also have the largest proportion of A-students in the course Mathematics I, as mentioned earlier. The conclusion can be drawn that having earned an A in the basic course will result in a higher probability to graduate, leading us to the larger groups of majors. Though, the same conclusion cannot be drawn about the rest-group, since the proportion of students that have earned the grade A in the course Mathematics I is not as large as for the physics majors. The rest group consist though of different kinds of students with different aims, which can create extreme results.

5.1.4 Age when starting education

The students age when starting the bachelor's program, or being enrolled in the course Mathematics I, is a continuous variable. The odds ratio shows

that for each increase in age the odds of finishing the thesis is a decrease in 3.6%. Hence, students who are older when starting their programs are less likely to finish their thesis in time. Note that this variable's coefficient in Table 6 can be interpreted in a similar way.

5.2 Interpretation of the gamma GLM for the time to finish the thesis

The gamma generalized linear regression analysis resulted in two models with the identity and logarithm link function. These models were obtained from both stepwise regression and backward elimination. The models are presented in Appendix C Table 11-12, and are presented below respectively:

$$\text{Identity: } \mu_i = 3.206 + 0.396\text{Time.Finish.MM2001}_i$$

$$\text{Logarithm: } \log(\mu_i) = 1.177 + 0.0956\text{Time.Finish.MM2001}_i$$

In both cases the significant variable to predict the time to graduation is the time to finish Mathematics I, with a positive coefficient and a low standard error. This means, for the first model, that as the time to finish Mathematics I increases with one year, so does the mean of the time to finish the thesis with $3.206 + 0.396 = 3.602$ years. Hence, when the time to finish Mathematics I increases with a year, the time to finish the thesis increases with 3.602 years. For the model with a logarithmic link function the change is multiplicative, meaning that for each increase in years the change is $\exp(1.177 + 0.0956) = 3.57$. So, as the time to finish Mathematics I increases with a year, the time to finish the thesis increases with 3.57 years. The AIC-values are 1095.1 and 1095.6 respectively and due to the small difference we conclude that both models are appropriate to use for the given data material.

6 Discussion

6.1 Discussion of the logistic regression model for the probability to finish the thesis

The chosen logistic regression model has the explanatory variables Time.Finish.MM2001, Program, Grade.MM2001 and Age. We start off by confirming that the Hosmer-Lemeshow test indicate that the model does not fit the data well, so the result can be misleading. Though, we want the most predictive model and both the AUC-value of the sampling method and 5-fold cross validation suggests the chosen model. The majority of the used model selection methods also come to the same conclusion, i.e. both stepwise selection, backward elimination and purposeful selection. Therefore, we choose this model and continue with discussing it.

From Table 8 we can see that the variable with the strongest effect on the probability to finish the thesis, given that the student has finished Mathematics I, is the grade A in Mathematics I compared to the grade D. Hence, a student which has obtained the grade A in the basic course has a higher probability to finish the thesis in time compared to the ones that obtained a D. This is quite logic since a high grade in the basic course means that the student has a solid fundamental knowledge in the subject, and will not need to extend higher level courses in more than six years, as we defined the variable Finish.Thesis.

The variable with the second strongest effect on the probability to finish the thesis is the Rest in Program. This means that students in bachelor's programs in astronomy, meteorology, biomathematics and biomathematics and computational biology are more likely to finish their degree in time as defined. The logistic model investigates the probability to finish the thesis or not given that they have finished Mathematics I, as mentioned earlier. So, the students taking stand-alone courses are not included in the analysis since we cannot predict their motives. Removing the students taking stand-alone courses we will obtain that the students in the Rest-category represent 25% of the data material. This means that compared to the mathematics students, who make up 45% of data material, they are more likely to finish their thesis on time. Since the rest-category consists of such a small group it can have extreme values. It consists of different bachelor's programs that works differently, which might be one of the reasons of the extreme result.

The variable of the time to finish Mathematics I and the age when starting the education are continuous variables and are both relevant to the model, according to the methods used. The 95% confidence intervals for these variables are not as wide as for the categorical variables. However, we can still draw the conclusion that larger sample size for the analysis is needed. In our attempt to get smaller intervals for the coefficients, we changed reference level from the computer science students to the mathematics students in the variable for bachelor's programs, since the mathematics majors are a majority. This did not result in much smaller intervals, since the average width of the intervals with computer science as reference level is 3.24 while the average width of the intervals with mathematics as reference level is 3.0. Though, we continued with mathematics as reference level since it is the largest bachelor's program and the study is from a mathematicians point of view. In the case of grade in Mathematics I the largest group was the students that obtained the grade D, which is the reference level of that factor. The model's predictability is certainly questionable since the AUC-value with sampled data is 0.548 and 0.645 with 5-fold cross-validated data, meaning that the model has poor predictability. With sampled data you could say that the model's predictability is like random guesses and with 5-fold cross

validation it is a poor prediction. Though, since the 5-fold cross-validation method reduced the variance and bias, we choose to rely on its result more.

6.2 Discussion of the gamma GLM for the time to finish the thesis

The gamma generalized linear models with the identity and logarithm link functions, and time to finish the course Mathematics I as explanatory variable, are our last obtained models. The reason that we choose to assume a gamma distribution for the response variable is due to it being positive, skewed and continuous according to Figure 8. This is logic since the time to finish the thesis consist of a majority finishing their thesis in approximately three years, with Mathematics I as starting point. The fitted linear model got residuals that did not show linearity nor constant variance, despite transformations of the response variable. Hence, with this knowledge we concluded that a gamma generalized linear model would be more appropriate. The link function depends on the data, therefore we tested the possible link functions when fitting the models. In Appendix D we find our simulated residuals and obtained residuals from the models, that helped us conclude if the models have residuals as the gamma distribution. This method of confirming the models' distribution can certainly be questioned, but since we compare the models against simulated models, using their corresponding scale and shape parameters, it is a quite reasonable approach since the simulated part is the response variable. We believe that regardless of sample size, we would still have not obtained normally distributed error terms since the majority finish their thesis in approximately 3 years. Though, transformations might had given us a linear result so it would have been interesting to investigate this manner further with a larger sample size.

6.3 Suggestions for improvement

For further studies it would have been valuable to use a larger sample size, which might give a better result. This might had help reducing our wide 95% confidence intervals of the odds ratios, and compute a linear model even though it would have needed transformations. It would also have been valuable in both analyses to include more relevant explanatory variables. The factors that can affect if a student finishes the degree and the time it takes are if they already have a degree when starting the education, their high school grades in mathematics courses, how much the students work parallel with the university studies and if they have children when obtaining their degree. It would also have been interesting to include data from other universities, and make an analysis that works for the different institutions.

A Contingency Tables

Table 1: Contingency table of Finish.Thesis and Program

Finish.Thesis\Program	Computer science	Physics	Mathematics	The Rest	Total
No	10	41	104	42	197
Yes	7	56	67	54	184
Total	17	97	171	96	381

Table 2: Contingency table of Finish.Thesis and Grade.MM2001

Finish.Thesis\Grade.MM2001	A	B	C	D	E	Total
No	31	38	40	57	31	197
Yes	59	33	36	43	13	184
Total	90	71	76	100	44	381

Table 3: Contingency table of Finish.Thesis and Gender

Finish.Thesis\Gender	F	M	Total
No	69	128	197
Yes	69	115	184
Total	138	243	381

Table 4: Contingency table of Grade.MM2001 and Program

Grade.MM2001\Program	Computer science	Physics	Mathematics	The Rest	Stand-alone	Total
A	1	32	34	10	27	104
B	1	14	25	11	19	70
C	2	10	28	19	17	76
D	3	10	29	21	20	83
E	0	5	9	7	7	28
Total	7	71	125	68	90	361

Table 5: Contingency table of Gender and Program

Gender\Program	Computer science	Physics	Mathematics	The Rest	Stand-alone	Total
Female	1	13	54	28	32	128
Male	6	58	71	40	58	233
Total	7	71	125	68	90	361

Table 6: Contingency table of Gender and Grade.MM2001

Gender\Grade.MM2001	A	B	C	D	E	Total
Female	35	21	30	31	11	128
Male	69	49	46	52	17	233
Total	104	70	76	83	28	361

B The Variance Inflation Factor

Table 7: VIF for the linear full model

Parameter	General VIF	Degrees of Freedom
Time.Finish.MM2001	1.16	1
Program	1.22	4
Grade.MM2001	1.17	4
Age	1.05	1
Gender	1.06	1

C The Gamma Generalized Linear Models

Table 8: The full gamma GLM with identity link function and AIC 1104.6

Parameter	Levels	Coefficient	SE	P-value
Intercept	3.292	0.383	2.65e-16***	
Time.Finish.MM2001		0.414	0.0900	5.84e-06***
Program	Stand-alone	-0.212	0.166	0.200
	Physics	0.0769	0.189	0.684
	Computer science	-0.213	0.540	0.694
	The Rest	-0.101	0.185	0.584
	Mathematics	1.0	-	-
Grade.MM2001	A	-0.00710	0.189	0.970
	B	-0.218	0.196	0.268
	C	0.0439	0.199	0.825
	D	1.0	-	-
	E	0.385	0.292	0.188
Age		-0.00530	0.0152	0.727
Gender	Female	1.0	-	-
	Male	0.131	0.134	0.330

Table 9: The full gamma GLM with logarithm link function and AIC 1106.1

Parameter	Levels	Coefficient	SE	P-value
Intercept		1.212	0.113	<2e-16***
Time.Finish.MM2001		0.0985	0.0195	7.09e-07***
Program	Stand-alone	-0.0471	0.0491	0.338
	Physics	0.0255	0.0537	0.636
	Computer science	-0.0491	0.141	0.728
	The Rest	-0.0294	0.0534	0.582
	Mathematics	1.0	-	-
Grade.MM2001	A	-0.00655	0.0547	0.905
	B	-0.0656	0.0584	0.262
	C	0.0116	0.0563	0.837
	D	1.0	-	-
	E	0.105	0.077	0.175
Age		-0.00192	0.00453	0.671
Gender	Female	1.0	-	-
	Male	0.0334	0.0396	0.400

Table 10: The full gamma GLM with inverse link function and AIC 1107.6

Parameter	Levels	Coefficient	SE	P-value
Intercept		0.294	0.0331	<2e-16***
Time.Finish.MM2001		-0.0235	0.00421	4.84e-08***
Program	Stand-alone	0.00837	0.0141	0.552
	Physics	-0.00843	0.0151	0.577
	Computer science	0.0104	0.0347	0.766
	The Rest	0.00852	0.0152	0.576
	Mathematics	1.0	-	-
Grade.MM2001	A	0.00249	0.0156	0.873
	B	0.0190	0.0171	0.268
	C	-0.00350	0.0156	0.823
	D	1.0	-	-
	E	-0.0282	0.0202	0.164
Age		0.000611	0.00133	0.647
Gender	Female	1.0	-	-
	Male	-0.00768	0.0113	0.499

Table 11: The reduced gamma GLM with identity link function and AIC 1095.1

Parameter	Coefficient	SE	P-value
Intercept	3.206	0.0831	<2e-16***
Time.Finish.MM2001	0.396	0.0847	4.16e-06***

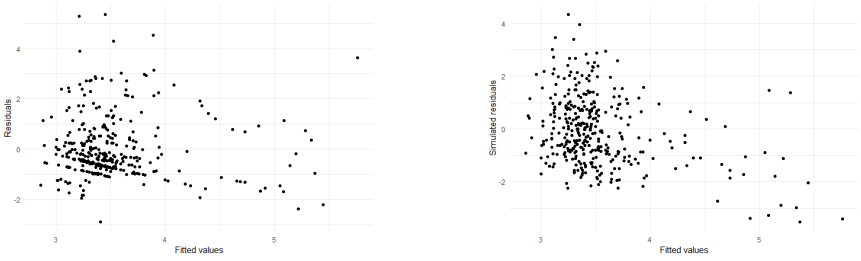
Table 12: The reduced gamma GLM with logarithm link function and AIC 1095.6

Parameter	Coefficient	SE	P-value
Intercept	1.177	0.0232	<2e-16***
Time.Finish.MM2001	0.0956	0.0181	2.27e-07***

Table 13: Frequency and proportion of students' graduation success

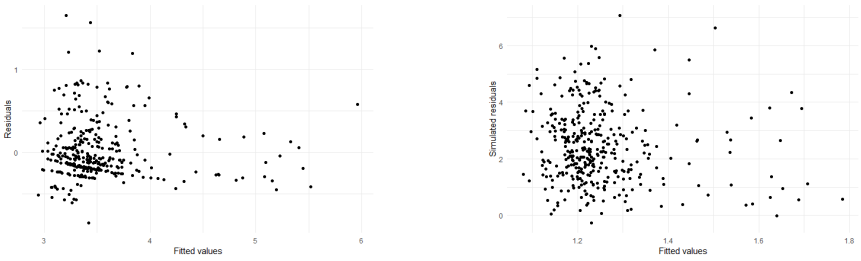
Bachelor's program	Finished thesis		Not finished thesis	
	Frequency	Proportion	Frequency	Proportion
Mathematics	67	36.4 %	104	52.8 %
Physics	56	30.4 %	41	20.8 %
Computer science	7	3.8 %	10	5.1 %
The rest	54	29.3 %	42	21.3 %

D Residual plots for data and simulated response



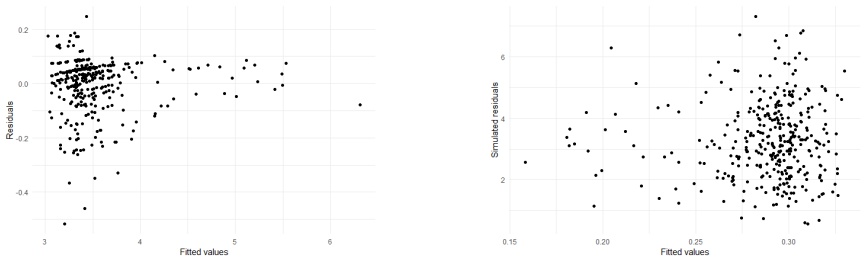
(a) (b)

Figure 9: Residuals of the full gamma GLM with identity link function to the left (a) and the corresponding simulated residuals to the right (b)



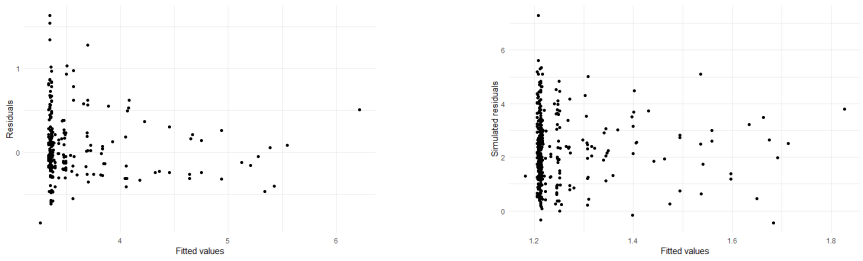
(a) (b)

Figure 10: Residuals of the full gamma GLM with logarithm link function to the left (a) and the corresponding simulated residuals to the right (b)



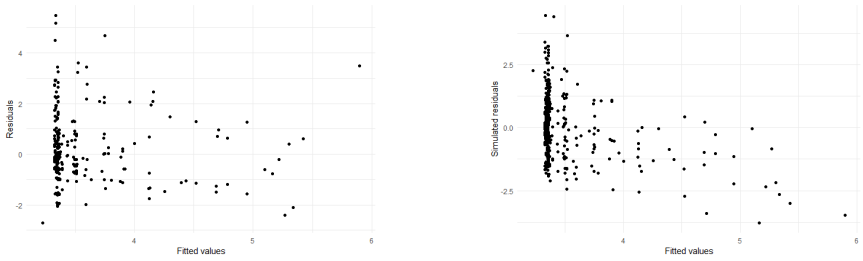
(a) (b)

Figure 11: Residuals of the full gamma GLM with inverse link function to the left (a) and the corresponding simulated residuals to the right (b)



(a) (b)

Figure 12: Residuals of the reduced gamma GLM with logarithm link function to the left (a) and the corresponding simulated residuals to the right (b)



(a) (b)

Figure 13: Residuals of the reduced gamma GLM with identity link function to the left (a) and the corresponding simulated residuals to the right (b)

References

Agresti, A. (2012). *Categorical data analysis*. 3. ed. Hoboken, N.J.: Wiley

Bursac, Z., Gauss, C.H., Williams, D.K. (2008). *Purposeful selection of variables in logistic regression.* Source Code for Biology and Medicine, 2008, 3:17

Fahrmeir, L. & Tutz, G. (2001). *Multivariate statistical modelling based on generalized linear models.* 2. ed. New York: Springer

Gelman, A. & Hill, J. (2007). *Data analysis using regression and multilevel/hierarchical models.* Cambridge: Cambridge University Press

Hastie, T., Tibshirani, R. & Friedman, J. (2009).*The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* Second. New York, NY: Springer New York

Hosmer, D., Lemeshow, S. & Sturdivant, R. (2013). *Applied logistic regression.* 3rd edition Hoboken, N.J.: Wiley

Sundberg, R. Kompendium Oktober 2016. *Lineära Statistiska Modeller.* Stockholm University

Zong, H. (2016). *A Statistical Analysis of Students' Time-To-Degree at the Department of Mathematics.* Stockholm University