Stockholms universitet

A Statistical Analysis of Students' Time-To-Degree at the Department of Mathematics

Huixin Zhong

Kandidatuppsats i matematisk statistik
Bachelor Thesis in Mathematical Statistics

Kandidatuppsats 2016:30
Matematisk statistik
December 2016

www.math.su.se

Matematisk statistik
Matematiska institutionen
Stockholms universitet
106 91 Stockholm

Mathematical Statistics
Stockholm University
Bachelor Thesis **2016:30**
http://www.math.su.se

A Statistical Analysis of Students' Time-To-Degree at the Department of Mathematics

Huixin Zhong*

December 2016

Abstract

This study aims to identify factors that are associated with students' time-to-degree, such as gender, age and grade in Mathematics I, as well as to build a model for predicting students' degree completion within a time period. The sample subjects are first-time, full time undergraduate programme students who entered the Department of Mathematics at Stockholm University between autumn 2007 and autumn 2013. Binary logistic regression analysis is implemented to study whether students obtained a bachelor degree within three years or not. We utilize also ordinal logistic regression analysis to study students' time-to-degree more specifically, which means that we are not only interested in students' degree completion within three years, but also students' degree completion within seven terms as well as within eight terms. A binary logistic regression model and a proportional odds model have been developed. The results have shown that gender, programme, grade in Mathematics I and that whether students finished Mathematics I within the same term are highly associated with students' time-to-degree in both models. That whether students finished Mathematics I within the same term has the strongest effect in terms of an odds ratio. The discussion focuses on explaining the effects of the influential factors and giving suggestions for future studies.

*Postal address: Mathematical Statistics, Stockholm University, SE-106 91, Sweden. E-mail: huixinsandy.zhong@gmail.com. Supervisor: Jan-Olov Persson.

Acknowledgement

This paper constitutes a bachelor's thesis of 15 ECTS in Mathematical Statistics at the Department of Mathematics at Stockholm University. I sincerely thank my supervisor Jan-Olov Persson for his constant support and valuable guidance in carrying out this project work. I am grateful to Martin Sköld for providing the data for this study. I would also like to thank my sister for her understanding and endless love. Last but not least, I wish to thank Tran Hao Nguyen and Mattias Andersson for taking time to read and give critique.

Contents

1 Introduction

The Bachelor's Programme in Mathematics and the Bachelor's Programme in Mathematics and Economics are two Bachelor's degree programmes (comprise three years of full-time studies) with the most applications and admissions in a mathematical subject at Stockholm University. The number of applications increases almost every year. According to the Swedish Higher Education Authority (UKÄ)[1], universities have to report the number of registered full-time students and the number of completed credits (converted to students' annual performance equivalent) each year in order to get the funding cap from the government. Besides, universities should bear the responsibility for improving and assuring the quality of their work by doing a self-evaluation, which will be needed to be submitted to allow the UKÄ to perform evaluations of programmes of higher education in Sweden.

Thus, the administration of the Department of Mathematics has an obligation to evaluate their work after each year's graduation. Assessing students' performance becomes one of the most important evaluation processes as graduated students can be indicators of institutional quality and institutes are partially responsible for the lower rate of degree graduates. Therefore, it would be essential for the department to know how many of the new admitted students will graduate on time each year or how many terms students need to obtain their degree. Answering these questions will not only help the administration to see how successful they were in accepting new students but also help them to see how successful they were in spending budgets. Having students who do not complete their programme of study on time certainly puts a burden on the department's next year's budget. Accordingly, the administration expects the students to complete their degree on time and they may want to predict the probability of students' time-to-degree. Based on the reasons above, this paper is conducted to build a tool to address such kinds of situation.

1.1 Aim

The aim of this paper are to identify factors that are associated with students' time-to-degree, for instance age, gender and study programme, as well as to develop a model that can be used to predict the probability of students' degree completion. Early prediction of students' time to graduation near the beginning of the degree programme will not only help the administration of the Department of Mathematics to make a better plan and manage their budget for the following year but also help the department to promote more students to complete their degree on time and provide immediate support for those whose graduation might be delayed.

1.2 Disposition of the paper

In section 2, we describe how we process the data and present all potential variables. In section 3, the reader is introduced to the statistical concepts and methods needed for the whole paper. Section 4 presents the statistical analysis and section 5 interprets results of fitted models. Conclusion will be made in section 6. Discussion of the models and suggestions for future studies can be found in the last section.

2 Data Description

The sample of the study was generated from seven years' data files which were obtained from Ladok - A national system for study administration within higher education in Sweden. Subjects were the first-time undergraduate students who entered in the Department of Mathematics at Stockholm University between autumn 2007 and autumn 2013 and they registered full-time for the first term. The total number of subjects is 665. Notice that first-time students who registered full-time for an independent course (fristående kurs) during this time period were not included because we do not know if these students planned to obtain a degree within a period of three years from the day when they registered for an independent course. Based on previous experience, we know that many students who register for independent courses at the department do not continue their studies after a few terms. It is hard to know their persistence to degree completion. If they were included in the study but it turns out that they only wanted to take a few courses from the beginning for some reasons or that they initially had not planned to obtain a degree, then this would have led to misinterpretation.

One of our goals is to identify factors that may have impacts on raising or declining the probability of degree completion directly or indirectly. We have three data files which enable us to build up a desirable data set and to extract as many potential variables as possible. The extracted variables are presented in table 1 below. In Appendix A.1, the detailed information of each data file is presented.

Table 1: Potential Variables

Variables	Description
Age	A student's age when he/she registered for a programme
Gender	Gender of a student
Maths1sameterm	If a student passed Mathematics I within the same term
Grade	Grade from A to E that a student received in Mathematics I
Programme	Study programmes
Regyear	The calender year when a student registered for a programme
Within3Years	If a student obtained a degree within three years or not
TimetoDegree	The number of terms a student needed to obtain a degree

However, when we looked through the subjects, we found that as many as 389 subjects missed a grade in Mathematics I (the first mandatory mathematical course) and did not obtain a degree. It is very possible that they have already dropped out of the programme at the very beginning of the term when studied Mathematics I. This will result in a large number of missing data if we consider the grade in Mathematics I as an influential factor, which we actually will do in this study. Since we do not know if they have planned to obtain a degree and we want to avoid the missing data problem, we decided to exclude those subjects from the data set. The study data set ends up with 276 subjects. We should note that the exclusion leads to a limitation of our study, that is, we study students' time-to-degree given that students have passed Mathematics I. Before performing our analysis, let us consider some descriptive statistics. The statistical software SAS was used to process the data.

2.1 Response variables

In this paper, time-to-degree is defined by the number of academic terms enrolled between the time of entering the university and of the degree completion. Students who received a degree certificate are referred to as students who graduated or obtained their degree. Those who meet the general requirements for obtaining the degree but have not applied for the certificate are not counted as students who completed their degree. A binary variable *Within3Years* was created to describe whether a student obtained a degree within three years or not. It will be used as our main response variable. Among 276 students in the data set 33 (12%) obtained a degree within three years. We have another variable which describes specifically the number of academic terms students needed to complete a degree is called *TimetoDegree*. This variable can be treated as an ordinal response and we will discuss it further in the Analysis section.

2.2 Explanatory variables

As mentioned previously, the data was obtained manually. We select the variables *Age*, *Gender*, *Programme*, *Maths1sameterm*, *Grade* and *Regyear* as our explanatory variables. We should mention clearly that there are other variables that may have impacts on the degree completion length, but we stick to what we are given. Now we study some basic features of the given explanatory variables and understand their relationship with the main response variable.

2.2.1 Age

For the calculation of the variable *Age*, registration year of the study programme and student's birthday were used, for example a student who was born in 1992 and registered for autumn 2012 is $2012 - 1992 = 20$ years old. Table 2 shows descriptive statistics for *Age*. We see that students who obtained a degree within three years are slightly younger than students who did not obtain a degree within three years.

Table 2: Descriptive statistics for *Age*

Within3Years	Mean	Median	Std Dev	Minimum	Maximum
276	21.92	21	3.99	18	46
No	21.97	21	4.14	18	46
Yes	21.55	21	2.67	19	28

2.2.2 Gender

The data set contains 174 females and 102 males. Table 3 shows that females are more likely to graduate sooner than males.

Table 3: Table of Within3Years by Gender

Within3Years	Male	Female
No	159 (91.4%)	84 (82.3%)
Yes	15 (8.6%)	18 (17.7%)
Total	174 (100%)	102 (100%)

2.2.3 Programme

Study programmes included in the study are the Bachelor's Programme in Mathematics and Economics (denote by M+E) and the Bachelor's Programme in Mathematics (denote by M). The Department of Mathematics offers two more study programmes in Mathematics, which are the Bachelor's Programme in Mathematics and Philosophy and the Bachelor's Programme in Biomathematics and Computational Biology, but they were not included

in the study because very few students applied for these programmes and we have substantial interest in those popular programmes with the most applications. Table 4 shows that students who study *M+E* seem to graduate earlier.

Table 4: Table of Within3Years by Programme

Within3Years	M + E	M
No	104 (83.9%)	139 (91.4%)
Yes	20 (16.1%)	13 (8.6%)
Total	124 (100%)	152 (100%)

2.2.4 Grade in Mathematics I

Mathematics I is the first compulsory course in mathematics within both programmes. Our study is conditioned on students who have received a passing grade in the course. Passing grades are designated on a five-point scale : A, B, C, D or E with A as the highest grade and E as the lowest. So the variable *Grade* has ordered categories. From Table 5, we see that students who received a higher grade in Mathematics I seem to finish early. *Grade in Mathematics I* will be written as *Grade* for simplification in the following context.

Table 5: Table of Within3Years by Grade

Within3Years	A	B	C	D	E
No	44 (74.6%)	38 (84.4%)	48 (94.1%)	71 (94.7%)	42 (91.3%)
Yes	15 (25.4%)	7 (15.6%)	3 (5.9%)	4 (5.3%)	4 (8.7%)
Total	59 (100%)	45 (100%)	51 (100%)	75 (100%)	46 (100%)

2.2.5 Maths1sameterm

The variable *Maths1sameterm* describes whether a student passed Mathematics I within the same term or not. One term is set to from January to August and from September to January in order to allow students have another chance to retake the exam without taking one more term. Reported in Table 6, 15.9% of students who finished the Mathematics I within the same term graduated on time while only 2.5% of students who failed to pass the course within the same term graduated on time. Students who finished the first mathematical course on time seem to be much more likely to obtain a degree within three years.

Table 6: Table of Within3Years by Maths1sameterm

Within3Years	No	Yes
No	79 (97.5%)	164 (84.1%)
Yes	2 (2.5%)	31 (15.9%)
Total	81 (100%)	195 (100%)

2.2.6 Regyear

The variable *Regyear* describes the registration year which extends from 2007 to 2013. It seems that the number of students who graduated on time follows some trend, but meanwhile we cannot discern any pattern in Table 7.

Table 7: Table of Within3Year by Regyear

Within3Years	2007	2008	2009	2010	2011	2012	2013
No	21(87.5%)	22(100%)	36(87.8%)	26(74.3%)	45(97.8%)	41(74.6%)	52(98.1%)
Yes	3(12.5%)	0(0%)	5(12.2%)	9(25.7%)	1(2.2%)	14(25.5%)	1(1.9%)
Total	24(100%)	22(100%)	41(100%)	35(100%)	46(100%)	55(100%)	53(100%)

3 Theory

In this section, we will present the theory needed throughout the paper. A large majority of theories are cited in Agresti(1) *Categorical Data Analysis* [2], Agresti(2) *Analysis of Ordinal Categorical Data* [3] or Hosmer et al. *Applied Logistic Regression* [4]. The theory about odds ratio and binary logistic regression is retrieved from Chapter 2 and 5 of Agresti(1). The theory about ordinal logistic regression and proportional odds assumption is taken from Chapter 3 of Agresti(2). Two model selection methods: *stepwise selection* and *purposeful selection* are taken from Chapter 4 of Hosmer et al. And most of the model fit and diagnostic tests can be referred to Chapter 5 and 6 of Agresti(1).

3.1 Odds and Odds ratio

Odds express the probability of an event occurring relative to the probability of an event not occurring. Mathematically, the odds are defined as

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

where π is the probability of success. The odds always have values greater than zero since π must fall in the $(0, 1)$ range. If the odds are greater than one, then a success event occurs more likely than a failure event. On the contrary, if the odds are less than one, then a failure event is more likely to

happen than a success one.

An odds ratio is used to compare the odds for two groups. It is the ratio of the odds of an event occurring in one group to the odds of an event occurring in another group. Mathematically, the odds ratio is defined as

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

where Ω_1 and Ω_2 are the odds for group 1 and group 2 respectively and π_1 and π_2 refer to the probability of success in group 1 and group 2 respectively.

Just as the odds, the odds ratio is always positive. If the odds ratio is greater than one, then subjects in group 1 are more likely to have success than subjects in group 2. If the odds ratio is smaller than one, the interpretation holds in a opposite way. If the odds ratio equals one, then there is no difference between the two groups and subjects of odds for both groups will be equally likely to occur.

3.2 Logistic regression

Logistic regression modelling has a wide variety of applications in many areas, including clinical studies, social science research, engineering and marketing. Binary logistic regression was developed to describe the relation between a binary response variable and a set of continuous and/or categorical explanatory variables. So when a response variable is dichotomous, which means that it can only take two values (an success event and a failure event), the binary logistic regression model is usually appropriate and useful.

Suppose a binary response variable $\mathbf{Y_i}$ has success event coded as 1 for each observation $i = 1, ..., n$ and p explanatory variables $\mathbf{x_i} = (x_{i1}, ..., x_{ip})$. Denote the probability of success by $\pi_i = Pr(\mathbf{Y_i} = 1)$, then the binary logistic regression model is written as

$$logit(\pi_i) = log(\frac{\pi_i}{1-\pi_i}) = \alpha + \sum_{j=1}^{p} \beta_j x_{ij} = \alpha + \beta_1 x_{i1} + ... + \beta_p x_{ip}, \qquad (1)$$

which is also called *logit model* in a linear relationship.

If we rewrite Equation (1) by exponentiating both its sides, we can see that the odds are an exponential function of $\mathbf{x_i}$. The expression is given as follows:

$$\frac{\pi_i}{1-\pi_i} = exp\left(\alpha + \sum_{j=1}^{p} \beta_j x_{ij}\right). \qquad (2)$$

This expression enables us to interpret the β_j in a basic way. That is, for every one-unit increase in the predictor variable x_j, the odds are expected

to increase multiplicatively by e^{β_j}, given the other predictor variables in the model are held constant.

In fact, we can express the logit in terms of the probability of success π_i by solving Equation (2). We obtain

$$Pr(Y = 1|X = x) = \pi_i = \frac{exp\left(\alpha + \sum_{j=1}^{p} \beta_j x_{ij}\right)}{1 + exp\left(\alpha + \sum_{j=1}^{p} \beta_j x_{ij}\right)}. \tag{3}$$

3.3 Ordinal Logistic regression

Although binary logistic regression is most common, logistic regression is extensible to more than two response levels. Based on differences in which and how the response levels are compared, several types of cumulative logits can be used to describe the relationship between an ordinal response variable and one or more explanatory variables. The most popular one is probably the cumulative logits which will be described below in detail. Other types of cumulative logits such as adjacent-categories logits and continuation-ratio logits is referred to Chapter 3 of Agresti(2) for those of you who are interested.

For c outcome categories response variable Y with probabilities $\pi_1, \cdots, \pi_c$, the cumulative logits are defined as

$$\text{logit}[Pr(Y \leq j)] = \log \frac{Pr(Y \leq j)}{1 - Pr(Y \leq j)} = \log \frac{\pi_1 + \cdots + \pi_j}{\pi_{j+1} + \cdots + \pi_c} \tag{4}$$

for $j = 1, \cdots, c-1$. Then each cumulative logit can be defined as

$$\begin{aligned}
\text{logit}[Pr(Y \leq 1)] &= \log \frac{\pi_1}{\pi_2 + \pi_3 \cdots + \pi_c} && (5)\\
\text{logit}[Pr(Y \leq 2)] &= \log \frac{\pi_1 + \pi_2}{\pi_3 + \pi_4 \cdots + \pi_c} && (6)\\
&\vdots && (7)\\
\text{logit}[Pr(Y \leq c-1)] &= \log \frac{\pi_1 + \pi_2 + \cdots + \pi_{c-1}}{\pi_c}. && (8)
\end{aligned}$$

Now we present a model for the cumulative logits which incorporates explanatory variables. The model is usually called the proportional odds model and it is the most frequently used ordinal logistic regression model in practice. It compares the probability of an equal or smaller response with the probability of a larger response.

For subject i, let y_i denote the outcome category for the response variable, and let $\mathbf{x_i} = (x_{i1}, \cdots, x_{ip})$ denote p explanatory variables. The model simultaneously uses all $c-1$ cumulative logits and it is formalized as

$$\text{logit}[Pr(Y_i \leq j|\mathbf{x_i})] = \alpha_j + \boldsymbol{\beta'}\mathbf{x_i} = \alpha_j + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}. \quad j = 1, ..., c-1 \tag{9}$$

In model (9), the logit for each cumulative probability j has its own intercept, α_j, but common slopes, β:s. The $\{\alpha_j\}$ are increasing in j and the logit is an increasing function of this probability. In the following context, we call the logit for cumulative probability j with the corresponding intercept α_j a *cumulative logit function*. Note that the slope for each explanatory variable stays the same across different cumulative logit functions. For simplicity of notation, we write $Pr(Y \leq j)$ instead of $Pr(Y_i \leq j|\mathbf{x_i})$ in the following context.

3.4 Variable Selection

There are many variable selection methods for finding a "best" model. In this paper, we will apply two methods which are *Stepwise Procedures* and *Purposeful Selection*. Stepwise procedures are probably the most commonly used and simplest of all variable selection procedures. They are more automated and statistical driven. Purposeful selection is introduced by Hosmer, Lemeshow and Sturdivant (Hosmer et al. 2013). They point out that mechanical selection procedures, such as stepwise procedures have its limitations to scrutinize the results carefully. They prefer purposeful selection since it gives the analyst control over every step of the selection process and can identify confounders correctly. For large sample size different methods perform roughly the same but purposeful selection is in favour when the sample size is smaller (Bursac et al. 2008)[5].

3.4.1 Stepwise Procedures

Stepwise procedures usually refer to *forward selection*, *backward elimination* and *stepwise selection*.

Forward selection starts with the intercept in the model. For all explanatory variables not in the model, their p-value will be checked to see if they can be added to the model. The one with the lowest p-value that is lower than the required cut-off such as 0.05, is chosen. This process continues until no new variables can be added.

Backward elimination reverses the forward selection. It starts with all the explanatory variables. The variable with the highest p-value greater than the required cut-off will be removed. This process will stop until all p-values are lower than the given cut-off. Note that using this method a problem of *quasi-complete separation*, that is, no overlap of sample points is likely to occur when there is a relatively small sample size[6]. More information about this problem can be found in Agresti(2) (p.64, 2010).

Stepwise selection is a combination of forward selection and backward elimination and there are several ways to do. One way is to start with a model containing only the intercept. The variable with the lowest p-value that is less than the required cut-off will be added to the model. And then

it will be checked if it should be removed based on the given cut-off. These two steps will be repeated until no variables can be added or removed.

3.4.2 Purposeful Selection

Purposeful selection is presented rigorously in Chapter 4 of Hosmer et al.(2013). It is a procedure that contains 7 steps and is summarized below.

Step 1: Purposeful selection starts with a careful univariale analysis of each explanatory variable. We select the variables that have a p-value for the likelihood ratio statistic less than 0.25 as candidates for developing a first multivariable model later. A higher significance level rather than the standard level (such as 0.05) is used here. This is because we do not want to exclude variables which are not significant alone, but might be when other variables are in the model.

Step 2: Use all selected variables from step 1 to fit a multivariable logistic regression model and check the significance of each variable using the Wald statistic. Variables that have a p-value larger than the level 0.05 are excluded and a new/smaller model fits. The new, smaller model is compared to the previous, old, larger model by performing a likelihood ratio test.

Step 3: Compare the values of the parameter estimates in the smaller model with the ones in the larger model. We are particularly interested in any variable whose coefficient has changed remarkably in magnitude. An indicator $\Delta\hat{\beta}_i = |(\hat{\theta}_i - \hat{\beta}_i)/\hat{\beta}_i|$ is used to assess whether a parameter estimate has changed "too much", where $\hat{\beta}_i$ denotes the parameter estimate for variable i in the larger model and $\hat{\theta}_i$ denotes the parameter estimate for variable i in the smaller model. A value of $\Delta\hat{\beta}_i > 0.2$ implies that the excluded variable(s) should be added back into the model because they are needed to adjust the effect of other variables.

Step 4: Add each variable not selected in Step 1 one at a time to the model obtained from Step 3 and check its significance by the Wald statistic p-value.

Step 5: In this step we examine more carefully the variables in the model obtained from Step 4. The level of the categorical variables should be reasonable and the continuous variables should have a linear relationship with the logit. We refer the model at the conclusion of this step as the *main effects model*.

Step 6: Now consider all possible interactions among the variables in the model. We add all possible interactions one at a time to the main effects model and check the significance at the standard level. All significant inter-

actions are then added to the model at the same time and we investigate if some of them can be removed by following Step 2. No main effects are removed at this step and we refer the model at the end of this step as *preliminary final model*

Step 7: Assess the adequacy and check the goodness of fit of the preliminary final model (see section 3.5 below).

3.5 Model fit and diagnostics

3.5.1 Proportional odds assumption

When fitting a proportional odds model, we need to check the proportional odds property of the model. Some software, such as SAS reports a score test of the proportional odds assumption, for example, a test for whether the slopes of the explanatory variable are equal across the cumulative logit functions. The proportional odds model which has one parameter for each explanatory variable

$$\text{logit}[Pr(Y \leq j)] = \alpha_j + \boldsymbol{\beta'}\boldsymbol{x}, \qquad j = 1, ..., c-1$$

is compared to a more complex model which has different parameter for each explanatory variable,

$$\text{logit}[Pr(Y \leq j)] = \alpha_j + \boldsymbol{\beta}_j'\boldsymbol{x}, \qquad j = 1, ..., c-1.$$

The null hypothesis is that the proportional odds model is true, and the alternative is that the more complex model holds, in other words, that different slops are needed. The test statistic (not showing here, see (Agresti(2),p.70) for details) can be shown to be approximately chi-squared distributed with degree of freedom $p(c-2)$, where p is the number of the explanatory variables in the proportional odds model. The test statistic that is not significant at a standard level (such as 0.05) indicates that the proportional odds assumption of the model holds, otherwise we need to justify the proportional odds assumption (See (Agresti(2), p.70) for more information).

3.5.2 Likelihood-ratio test

A likelihood-ratio test is used to compare the goodness of fit of two nested models, a smaller model M_0 and a larger model M_1 with one or more parameters compared to M_0. A hypothesis test between M_0 and M_1 can be formalized as follows:

$$H_0 : M_0 \text{ holds}$$

$$H_1 : M_1 \text{ holds but not } M_0.$$

Let L_0 be the maximized likelihood under the null hypothesis and L_1 be the maximized likelihood under the alternative hypothesis. Let l_0 and l_1 be the corresponding maximized log-likelihoods. The form of the test statistic is namely the ratio of two likelihood functions and given as follows:

$$-2 \cdot log\left(\frac{L_0}{L_1}\right) = -2(l_0 - l_1)$$

Asymptotically, the likelihood ratio test statistic is distributed as a *chi-squared* random variable under the null hypothesis, with a degree of freedom equal to the difference in the number of parameters between two models.

$$-2(l_0 - l_1) \overset{H_0}{\approx} \chi^2_{df}$$

3.5.3 The Hosmer-Lemeshow test

When the data is ungrouped or when there is at least one continuous variable, traditional goodness-of-fit test such as the Deviance or the Pearson χ^2 are not valid since they do not have limiting χ^2 distribution (Hosmer et al, p.155-157). Hosmer et al. introduce a more suitable test for these situations, usually called the *Hosmer-Lemeshow test*. It is a χ^2 test formed by partitioning the data according to the estimated probabilities and then dividing them into g approximately equal sized groups. One usually forms 10 groups and from these one can create a Pearson statistic $\hat{C}$ for comparing the observed and estimated values. Asymptotically, $\hat{C}$ is chi-squared distributed with $(g-2)$ degree of freedom under the null hypothesis that the model is true. If the $\hat{C}$ is significant at a given level then the model in question does not fit the data well. More details regarding this test can be found in Chapter 5.2.2 of Hosmer et al (2013).

3.5.4 AIC

Akaike information criterion(AIC) is probably the best known criteria that can help select a good model in terms of estimating quantities of interest. It is used to compare different models within the same data set. The AIC is defined as

$$AIC = -2(\log(L) - k)$$

where L is the maximized likelihood function and k is the number of parameters in the model. Given a collection of candidate models for the data, the preferred model is the one with the minimum AIC value.

3.5.5 ROC, AUC and Concordance Index

A common way of judging predictive power of a binary logistic model is to create a *Receiver Operating Characteristic* (ROC) curve. Let

$$\hat{y_i} = \begin{cases} 1 & \text{if } \hat{\pi_i} > \pi_0 \\ 0 & \text{otherwise} \end{cases}$$

where $\hat{y_i}$ is the predicted values of y_i for some cutoff π_0. The following definitions are given in Agresti(1) (p.228, 2002):

$$sensitivity = P(\hat{y_i} = 1|y = 1) \text{ and } specificity = P(\hat{y_i} = 0|y = 0).$$

A ROC curve with a concave shape is then obtained by plotting *sensitivity* against *1-specificity* for all the possible cutoffs. The area under the ROC curve (AUC) is a measure of predictive power. A high area under the curve indicates a good prediction ability. Hosmer et al.(p.177, 2013) provide rough guidelines on how to evaluate the AUC:

$$\text{if} = \begin{cases} 0.5 < AUC < 0.7 & Poor \\ 0.7 < AUC < 0.8 & Acceptable \\ 0.8 < AUC < 0.9 & Excellent \\ 0.9 < AUC & Outstanding \end{cases}$$

If one wants to validate the predictive power of an ordinal logistic regression, the above-mentioned ROC analysis is not appropriate as the method has not been extended to the ordinal logistic regression model (Hosmer et al, p.289). Nevertheless, one can check for an index of predictive power, called the *concordance index* (Agresti(2), p.65). Here all possible pairs of subjects having different response values are considered. The concordance index is the probability of concordance, between predicted probability and response, that is, that an observation with a large y-value also has a higher predicted probability. A value of the concordance index of 0.5 suggests that the predictions of a model are no better than random guessing and a value of 1 indicates perfect predictions. The higher the value of the concordance index, the better the predictive power.

3.5.6 Cross Validation

When assessing predictive power of a model comes into question, the data used to evaluate how well one can predict the response variable based on the explanatory variables is usually the data used to fit a model. If fitting the model and assessing its statistical performance are on the same data, the assessment of a model can be optimistically biased. One way of dealing with this situation is cross validation, discussed by Feng Zhang in Chapter 3 of his doctoral dissertation [7]. Cross validation is an algorithm to estimate

predictive errors and it is often used as a model validation technique. There are several types of cross validation and the one we will use in this paper is called "leave-one-out". It is done by omitting one observation at a time and then predicting on the observation and measuring the predictive error after fitting a model without that observation. An unbiased assessment can therefore be achieved.

As described in the previous section, the area under the ROC is a measure of predictive power. We can create a ROC curve using cross validation in SAS. This is done by fitting the model to the complete data set and using the cross validated predicted probabilities to provide a ROC curve[8]. The idea behind this is "*The cross validated predicted probability for an observation simulates the process of fitting the model ignoring the observation and then using the model fit to the remaining observations to compute the predicted probability for the ignored observation.*"[8]. Note that the AUC is lower when cross validation is used.

4 Analysis

In this section, the model selection process is described in details. We will select and evaluate fit of the final model. Binary logistic regression will be used to analyse the data since the main response variable *Within3Years* is dichotomous. The other response variable *TimetoDegree* which has the ordinal property will be studied after fitting a binary logistic regression model. Six explanatory variables which are *Age*, *Gender*, *Programme*, *Maths1sameterm* and *Grade* and *Regyear* were presented in section 2.2. *Gender*, *Programme* and *Maths1sameterm* are categorical variables with two levels. *Age* and *Regyear* will be treated as continuous variables in the first place since the observations are taking values between a certain set of real numbers and we do not have proper reasons to treat them as categorical. The variable *Grade* is yet to be discussed whether to be treated as categorical or continuous since it has ordered categories, from the lowest grade E to highest grade A. A simple model treats this variable as continuous and grade is assumed to have a linear effect for a set of monotone scores such as $1, 2, 3, 4, 5$. A complex, larger model treats this variable as categorical with five levels. Either model may be adequate but we will choose the one that we deem most useful to us.

All the analysis will be performed in the statistical software SAS (University Edition).

4.1 Degree Completion Within Three Years

The event "A student obtained a bachelor degree within three years given that the student has finished the Mathematics I" is modelled. *Male* of *Gender*, *M* of *Programme* and *No* of *Maths1sameterm* are specified as the

reference levels. We want to find a model that is able to predict the data well, while still being parsimonious and easy to interpret.

4.1.1 Fitting the binary logistic regression model

We utilize first purposeful selection and then some stepwise procedures. We treat grade as continuous here and discuss its feasibility at a later step of the selection procedure.

Step 1: We start with a careful univariable analysis of each explanatory variable. The results of this analysis are shown in Table 8.

Table 8: Results of Fitting Univariable Logistic Regression Models

Parameter	Coeff.	SE	P-value
Age	-0.0301	0.0524	0.5499
Gender	0.4102	0.1874	0.0284
Programme	0.3804	0.1896	0.0541
Maths1sameterm	1.0051	0.3711	0.0004
Grade	0.4688	0.1441	0.007
Regyear	-0.0849	0.0974	0.9307

Table 9: Results of Fitting the Multivariable Model with All Variables Significant at the 0.25 Level in the Univariable Analysis

Parameter	Coeff.	SE	P-value
Gender	0.9658	0.4045	0.0170
Programme	1.1903	0.4278	0.0054
Maths1sameterm	1.6856	0.7714	0.0289
Grade	0.5396	0.1683	0.0013

Step 2: We now include all variables that are significant at the 0.25 level from step 1 to fit our first multivariable model. We examine the Wald statistic for each variable. Variables that have p-value over the standard level 0.05 are excluded from the model. Table 9 shows that all variables are significant in the fitted model. There is no model comparison between a smaller one and a larger one here. So we jump over step 3 and move on to step 4.

Step 4: Now it is time to revisit the insignificant variables in Step 1. We add *Age* ($p = 0.6386$) and *Regyear* ($p = 0.9464$) to the multivariable model one at a time but they do not show any significance at level 0.05. Since *Regyear* only takes 7 values, we tried to include it as a categorical variable in the model. However, it did not become significant and quasi-complete separation of data points was detected. Thus, the model at this step is remained the same as the first fitted multivariable model.

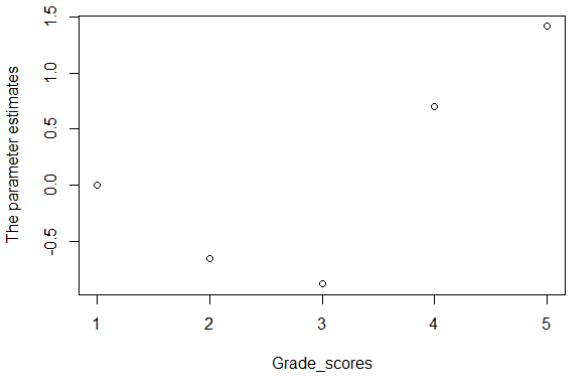
Step 5: In this step we examine the variables from step 4 closely. We have three categorical variables which are *Gender*, *Programme* and *maths*1*sameterm*. From Table 9, we see that parameter estimates of these variables are positive, which is reasonable regarding the level of the categories. We must also check whether the variable *Grade* has a linear relationship with the logit since we treated it as continuous. We do this by comparing the current model having grade treated in a continuous manner to a more complex model having grade treated in a categorical manner. We present the results of fitting the multivariable model when treating grade as categorical in Table 10.

Table 10: Results of Fitting the Multivariable Model having treated *Grade* as categorical

Parameter	Coeff.	SE	P-value
Gender	1.1174	0.4181	0.0075
Programme	1.2937	0.4484	0.0039
Maths1sameterm	1.7800	0.7834	0.0231
Grade2	-0.6501	0.7671	0.3967
Grade3	-0.8807	0.8374	0.2929
Grade4	0.6971	0.7309	0.3402
Grade5	1.4138	0.6901	0.0405

From Table 10, we can see that grade has 4 parameters and two of them have a negative sign. The parameters do not seem to follow a linear trend when plotted against the grade scores (see Figure 1). However, the parameter estimates contain some uncertainties as the standard errors are quite large. We may be mistaken if we think that the current simple model does not hold because we do not see a linear trend from the plot. Thus, we examine more closely these two models.

Figure 1: A plot of the parameter estimates for *Grade* against the grade scores



First, we check the AIC for comparing two models. We find that AIC values for the two models are approximately equal. The simple model has AIC value equal to 178.370 and the complex model has AIC value equal to 178.057. Still, we cannot tell which model fits the data substantially better. So we perform a likelihood-ratio test to compare the fit of these two models. Denote the simple model by M_0 and the complex model by M_1. The maximized log likelihood for the two models are 168.370 and 162.057 respectively. As discussed in Section 3.5.2, the likelihood-ratio statistic for comparing models M_0 and M_1 is

$$G^2(M_0|M_1) = -2(l_0 - l_1) = 168.370 - 162.057 = 6.313 < \chi^2_{0.05}(3) = 7.8147$$

The likelihood-ratio test statistic 6.313 does not exceed the chi-square value with a degree of freedom 3 (P-value $= 0.0973$). This supports that the simple model having *Grade* treated as continuous is adequate. Besides, tests of the *Grade* effect are more powerful when it has a single parameter rather than several parameters. Reported in SAS, the p-value of the Wald χ^2-test is 0.0013 for continuous grade while the p-value of the Wald χ^2-test is 0.0038 for categorical grade. For all the reasons mentioned above, we decide to keep treating *grade* as continuous in further analysis. We refer the model at this step as our main effects model.

Step 6: This step in the analysis is to select interactions. With only four main effects, we consider all 6 two-way interactions between the selected variables. We add all of them to the model one by one and check the significance at level 0.05. We find that none of these interactions became significant. The resultant model is then referred as our preliminary final model, denote $\mathbf{M_b}$. Parameter estimates of $\mathbf{M_b}$ can be found in Table 9.

Step 7: In this step we assess the fit of the preliminary final model. Before testing how well $\mathbf{M_b}$ fits the data, we consider other models using

some stepwise procedures.

Three stepwise procedures presented in section 3.4.1 will be implemented. We use all six explanatory variables as well as two-way interaction terms between all of them. The variables are treated in the same way as with the purposeful selection and the cut-off of 0.05 is chosen to allow a variable enter or stay in the model. We found that all three procedures gave the same final model as the purposeful selection when taking grade as continuous. However, a quasi-complete separation of data points was detected when running the method of backward elimination and validity of this model fit is therefore questionable.

Summarizing the analysis above, we choose $\mathbf{M_b}$ as our final model despite the fact that the same model chosen by the backward elimination is questionable. We evaluate the goodness of fit of the model. Because the model contains a continuous variable, normal global fit statistics such as the Deviance or Pearson χ^2-test are not appropriate. The Hosmer-Lemeshow test as introduced in section 3.5.3 was proposed to deal with such case. For $\mathbf{M_b}$, the Hosmer-Lemeshow statistic with $g = 9$ groups equals 9.0247, with df $= 7$ and p-value of 0.2509. This indicates that the difference between the observed counts and fitted values is quite small. In other words, the model fits the data quite well.

4.2 Time to Degree

In the previous section, we focused on identifying the factors that are associated with students' degree completion within three years. A binary logistic regression model was fitted. Using this model we could easily estimate the probability of degree completion within three years. However, we cannot estimate the probability of degree completion more than that time period. Among students who could not obtain a degree within three years, quite many students successfully graduated by the seventh term or the eighth term. It would also be interesting to find out how likely these students graduated within those time periods. So in these section, we perform a more detailed analysis and try to develop a model that can be used for estimating the probability of students' degree completion within three years as well as for estimating the probability of graduation within seven terms and within eight terms.

To obtain a model mentioned above, another response variable *TimetoDegree* with four categories can be used here. It is extent of students' time-to-degree: less than or equal to six terms, seven terms, eight terms and more than eight terms. Thus, the response variable can be treated as ordinal here. When a response is ordinal, one can consider a proportional odds model, described in section 3.3.

Before performing the analysis, we should note that the data of 2013 was excluded. The study sample only consists of 223 subjects. Table 11 shows the frequency of each response level. Descriptive tables for the explanatory variables can be found in Appendix A.2.

Table 11: Summary of the ordinal response

TimetoDegree	Frequency	Percent	Cumulative Frequency	Cumulative Percent
$\leq$ 6 terms	32	14.35	32	14.35
7 terms	9	4.04	41	18.39
8 terms	14	6.28	55	24.66
$\geq$ 8 terms	168	75.35	223	100.00

4.2.1 Fitting the proportional odds model

In our study sample, a proportional odds model, which has a different intercept for each cumulative logit function but with the same slopes is:

$$\text{logit}[Pr(TimetoDegree \leq j)] = \alpha_j + \boldsymbol{\beta'}\mathbf{x}. \quad j = 1, ..., 3$$

Since our response has four levels, we will have three cumulative logit functions. We choose to compare less time-to-degree levels to more time-to-degree levels, for instance, the first cumulative logit function compares time-to-degree less than or equal to 6 terms to time-to-degree more than 6 terms (combined equal to 7 terms, equal to 8 terms and more than 8 terms categories). The cumulative logit functions are formalized as follows:

$$\text{logit}[Pr(TimetoDegree \leq 6\ terms)] = \alpha_1 + \boldsymbol{\beta'}\mathbf{x}$$

$$\text{logit}[Pr(TimetoDegree \leq 7\ terms)] = \alpha_2 + \boldsymbol{\beta'}\mathbf{x}$$

$$\text{logit}[Pr(TimetoDegree \leq 8\ terms)] = \alpha_3 + \boldsymbol{\beta'}\mathbf{x}.$$

Our goal here is to obtain a model that could be used for estimating the cumulative probabilities of the ordinal response, as well as, to some extent, for quantifying the effect of individual factors. Hosmer et al. point out that the method of purposeful selection can also be applied to develop an ordinal logistic regression model (p.305, 2013). Therefore, we use this method to find a proper proportional odds model as well. Treatment of the explanatory variables are the same as in the binary logistic regression model. The steps are presented as follows:

Step 1: We fit a univariable proportional odds model for each variable. The results of this fit are shown in Table 12. Three intercepts for each univariable model are not showing here.

Table 12: Results of Fitting Univariable Proportional Odds Model

Parameter	Coeff.		P-value
Age	-0.0667	0.0459	0.1144
Gender	0.6753	0.3109	0.0300
Programme	0.3314	0.3081	0.2820
Maths1sameterm	1.2881	0.4403	0.0010
Grade	0.3571	0.1140	0.0015
Regyear	0.2082	0.0988	0.0321

Step 2: We now fit our first multivariable proportional odds model (denoted by M_1) using all the significant variables from step 1. We find that the Wald test for the coefficient for *Regyear* is not significant with $p = 0.1016$. We exclude this variable and denote the simple model by M_0. A test similar to the Wald test, the likelihood ratio test (P-value = 0.099) also suggests that the simple model excluding *Regyear*is adequate. So we continue the analysis with model M_0. The results of these two fitted model can be found in Appendix A.3 (Table 24 and Table 25).

Step 3: In this step we compare the parameter estimates of M_0 and M_1. In results not shown, we find that the largest percent change is 9.27% for the coefficient of *Grade*. This does not exceed our criterion of 20%. So we continue our analysis using M_0.

Step 4: On univariable analysis the variables *Age* and *Programme* were not significant. When each of these variables is added, one at a time, to M_0, the coefficient of *Programme* became significant. The results of this fitted model is shown in Table 13. Denote this model by M_p. With a p-value of 0.1205, *Age* was not shown to be significant.

Table 13: Results of the fitted model at the end of Step 4

Parameter	Coeff.	SE	P-value
Gender	0.7639	0.3263	0.0192
Maths1sameterm	1.0269	0.4669	0.0279
Grade	0.4025	0.1324	0.0024
Programme	0.7240	0.3440	0.0353

Step 5: Now it is time to examine more closely the variables in M_p. The positive sign of parameter estimates shows that categories for the categorical variables are appropriate. For the continuous variable *Age* we must check its linear relation with the logit. As we did in the previous section, we develop a more complex model having treated grade as categorical (denoted by M_c) and compare its fit to M_p. We find that the AIC value for M_p (341.846) is clearly smaller than the AIC value for M_c (343.617). The likelihood-ratio statistic for comparing M_c and M_p equals 4.229, with df = 3, suggesting also that the simple model M_p is adequate. Thus, based on the AIC and

the likelihood ratio test, we choose to continue treating grade as continuous in further analysis. M_p is referred as our main effect model.

Step 6: Once we have obtained the main effect model, we check for all possible interactions among the variables in M_p. Since we only have four main effects, we add each of the 6 two-way interaction terms one at a time to the model. We find that no interaction became significant at the 5% level. Hence, M_p become our preliminary final model.

Step 7: In this step the fit of our preliminary final model needs to be evaluated. Before assessing the goodness of fit, we should check for the proportional odds assumption since the proportional odds model only applies to data that meet the assumption. We look at a score test which is reported by SAS (see section 3.5.1). A chi-square test statistic equal to 7.1244 with df = 8 (P-value = 0.5233) indicates that the proportional odds assumption holds.

As for the binary logistic regression model, global goodness-of-fit tests such as the Deviance and the Pearson χ^2-test are not valid for proportional odds model when the model contains at least one continuous variable. Hosmer et al (2013) introduces several alternative tests in Chapter 8 for such cases, for example a Lipsitz test proposed by Lipsitz et al (1996), a Pulkstenis and Robinson test suggested by Pulkstenis and Robinson (2004) and an extension of the Hosmer-Lemeshow test for the multinomial logistic regression model developed by Fagerland and Hosmer (2012b). However, they have not been widely used and do not seem to be available in the current statistical software (Agresti(2), 2010). Due to time limitation and considering the difficulty of implementation of these methods in practice, we do not check the global goodness-of-fit of M_p.

5 Results

5.1 Interpretation of the binary logistic regression model

Using binary logistic regression we identified the factors that are related to student's degree completion within three years and we developed a model that could be used to predict degree completion within three years. Purposeful selection and stepwise procedures gave us the same final model $\mathbf{M_b}$. We checked the predictive power of $\mathbf{M_b}$ by assessing the area under a ROC curve (AUC) without and with cross validation. We found that the AUC for $\mathbf{M_b}$ is 0.7888 without cross validation. With cross validation, the AUC estimate drops slightly to 0.7345. Two values of AUC indicate that our model has an *Acceptable* predictive capacity (see section 3.5.5). The small drop suggests that $\mathbf{M_b}$ predicts the data well when cross validation is used. The corresponding ROC curves can be found in Appendix 3.

5.1.1 Example

Analysis of maximum likelihood estimates of $\mathbf{M_b}$ is presented in the following table.

Table 14: Analysis of Maximum Likelihood Estimates of $\mathbf{M_b}$

Parameter		Coeff.	SE	P-value
Intercept		-6.2498	1.0071	<.0001
Gender	Male	0	-	-
	Female	0.9658	0.4045	0.0170
Programme	M	0	-	-
	M+E	1.1903	0.4278	0.0054
Maths1sameterm	No	0	-	-
	Yes	1.6856	0.7714	0.0289
Grade		0.5396	0.1683	0.0013

We use an example to illustrate our final model $\mathbf{M_b}$. To estimate a student's probability of completing a degree within three years using *Gender*, *Programme*, *Maths1sameterm* and *Grade*, Equation (1) (see section 3.2) would be applied as follows:

$$\text{logit}[\Pr(\text{Obtained a degree} \leq 3 \text{ years})] = \alpha + \sum_{j=1}^{4} \beta_j x_{ij}$$

$$= \alpha + \beta_{Gender} * Gender_{F/M} + \beta_{Progr} * Progr_{M+E/M} + \beta_{Maths1} * Maths1_{Y/N} + \beta_{Grade} * Grade$$

A number of variable combinations or student groups can be made among the four significant variables. One of them can be that a female student studies the Bachelor's programme of Mathematics, has finished Mathematics I within the same term and obtained a B. The equation would be:

$$\text{logit}[\Pr(\text{Obtained a degree} \leq 3 \text{ years})] =$$

$$= -6.2498 + 0.9658 + 0 + 1.6856 + 0.5396 * 4$$
$$= -1.44$$

Using the Equation (2) presented in section 3.2 we can calculate the estimated probability of degree completion within three years:

$$\Pr(\text{Obtained a degree} \leq 3\text{years}) = \frac{e^{-1.44}}{1 + e^{-1.44}} = 19.2\%$$

This yields a three-year degree completion probability of 19.2% for this student and students in the same group have the same probability to graduate on time. The corresponding 95% confidence interval for the estimated $\Pr(\text{Obtained a degree} \leq 3 \text{ years})$ is $(0.102, 0.332)$. The full calculation is omitted here but can be found in Appendix A.4.

5.1.2 Interpretation of the Odds Ratio

Table 15 shows the odds ratio estimates for $\mathbf{M_b}$.

Table 15: Odds Ratio Estimates of $\mathbf{M_b}$

Parameter		Point Estimate	95% Confidence	Wald Limits
Gender	Female vs Male	2.627	1.189	5.804
Programme	M+E vs M	3.288	1.422	7.605
Maths1sameterm	Yes vs No	5.396	1.190	24.472
Grade		1.715	1.233	2.385

Gender: The odds ratio estimate of *Gender* shows that the odds of obtaining a degree within three years is 162.7% (CI, (1.189, 5.804)) higher for a female student than for a male student if all of the other variables in the model are held constant. This means that females are more likely to graduate within three years than males given that all of them have passed the Mathematics. In the example shown above, the estimated probability of degree completion within three years was 19.2%. If it is a male student in the same situation, the estimated probability will drop to 8.3%. The odds ration between these two events is $\frac{0.192/(1-0.192)}{0.083/(1-0.083)} = 2.63$, which corresponds to the odds ratio stated in Table 14.

Programme: There is a quite large difference between students who study the programme in Mathematics and Economics and those who study the programme in Mathematics. The odds that a student obtained a degree within three years given that the student studied the programme in Mathematics and Economics is 3.3 times (CI, (1.422, 7.605)) as high as the odds that a student obtained a degree given that the student studied the programme in Mathematics. If we control the other variables, the estimated probability of degree completion within three years will increase to 43.8% if the student studied the Bachelor's programme in Mathematics and Economics.

Maths1sameterm: The factor that describes whether a student passed the first mathematical course, Mathematics I within the same term has a strongest effect on students' timely degree completion compared to effect of the other factors. The odds that a student obtained a degree within three years when the student finished Mathematics I within the same term is 5.4 times as high as when a student did not finish Mathematics I within the same term. In other words, students who passed Mathematics I on time are more likely to complete a degree on time compared to those who did not. Using the example in section 5.1.1 again, and controlling for the other variables, a student who did not finished Mathematics I within the same term has only a probability of 4.2% to obtain a degree within three years. However, the level of precision of the odds ratio is quite low as the confidence interval is

really wide (1.190, 24.472). When the uncertainty in the odds ratio estimate is large, we cannot be reasonably sure that the true effect actually lies.

Grade in Mathematics I: The variable grade in Mathematics I was treated as continuous and there was no direct statistical evidence of nonlinearity in the logit. For each one level increases in the grade in Mathematics I, the odds that a student obtained a degree within three years multiply by 1.7 (CI, (1.233, 2.385)), that is, there is a 70% increase in the odds of obtaining a degree. Students who received a better grade in Mathematics I are more likely to graduate sooner.

5.2 Interpretation of the proportional odds model

We applied the method of purposeful selection and obtained a model containing the following variables: *Gender*, *Programme*, *Maths1sameterm* and *Grade*. We summarize how well the response can be predicted by checking the *concordance index*, presented in section 3.5.5. The estimated concordance index for $\mathbf{M_p}$ reported by SAS is 0.699. Based on the given guideline that a value of 0.5 indicates random predictions and that a value of 1.0 indicates perfect predictions, we would say that the predictive power of $\mathbf{M_p}$ is fairly poor.

Table 16 provides the maximum likelihood estimates of model $\mathbf{M_p}$.

Table 16: Analysis of Maximum Likelihood Estimates of $\mathbf{M_p}$

Parameter		Coeff.	SE	P-value
Intercept	$\leq$ 6 terms	-4.6306	0.6802	<.0001
Intercept	7 terms	-4.2974	0.6690	<.0001
Intercept	8 terms	-3.8797	0.6557	<.0001
Gender	Male	0	-	-
	Female	0.7639	0.3263	0.0192
Programme	M	0	-	-
	M+E	0.7240	0.3440	0.0353
Maths1sameterm	No	0	-	-
	Yes	1.0269	0.4669	0.0279
Grade		0.4025	0.1324	0.0024

5.2.1 Example

Now let us also use an example to illustrate the proportional odds model $\mathbf{M_p}$. By analogy with the binary logistic regression model, we can exponential the cumulative logits to produce the cumulative odds for the proportional odds model, and then we can solve the cumulative logit probabilities. Let θ_j

denote the cumulative probabilities $Pr(TimetoDegree \leq j)$, then

$$\text{logit}(\theta_j) = \alpha_j + \boldsymbol{\beta'}\mathbf{x} \rightarrow \frac{\theta_j}{1-\theta_j} = e^{\alpha_j + \boldsymbol{\beta'}\mathbf{x}} \rightarrow \theta_j = \frac{e^{\alpha_j + \boldsymbol{\beta'}\mathbf{x}}}{1 + e^{\alpha_j + \boldsymbol{\beta'}\mathbf{x}}}$$

For the cell probabilities themselves,

$$Pr(TimetoDegree = j) = \frac{e^{\alpha_j + \boldsymbol{\beta'}\mathbf{x}}}{1 + e^{\alpha_j + \boldsymbol{\beta'}\mathbf{x}}} - \frac{e^{\alpha_{j-1} + \boldsymbol{\beta'}\mathbf{x}}}{1 + e^{\alpha_{j-1} + \boldsymbol{\beta'}\mathbf{x}}}$$

We use the same example from the previous section 5.1.1. A female student studies the Bachelor's Programme of Mathematics, has finished Mathematics I within the same term and obtained a B grade. Then we can obtain the estimated cumulative probabilities:

$$Pr(TimetoDegree \leq 6\ terms) = 22.6\%$$

$$Pr(TimetoDegree \leq 7\ terms) = 29.0\%$$

$$Pr(TimetoDegree \leq 8\ terms) = 38.3\%$$

We see that in this model the estimated probability of obtaining a bachelor degree within three years is 22.6%, which is roughly similar to the value of the estimated probability using the binary logistic regression model (19.2%). The probability of completing a degree within seven terms and within eight terms are 29.0% and 38.3%, respectively.

From these, we can even obtain the estimated cell probabilities.

$$Pr(TimetoDegree \leq 6\ terms) = 22.6\%$$

$$Pr(TimetoDegree = 7\ terms) = 28.97\% - 22.62\% = 6.4\%$$

$$Pr(TimetoDegree = 8\ terms) = 38.25\% - 28.97\% = 9.3\%$$

$$Pr(TimetoDegree \geq 8\ terms) = 1 - 38.25\% = 61.8\%$$

5.2.2 Interpretation of the Odds ratio

Table 17 displays the estimated odds ratio of model $\mathbf{M_p}$.

Table 17: Odds Ratio Estimate of Model $\mathbf{M_p}$

Parameter		Point Estimate	95% Confidence	Wald Limits
Gender	Female vs Male	2.147	1.132	4.069
Programme	M+E vs M	2.063	1.051	4.048
Maths1sameterm	Yes vs No	2.792	1.118	6.974
Grade		1.496	1.154	1.939

The interpretation of these proportional odds ratios is pretty much the same as the interpretation of odds ratios from a binary logistic regression. For gender, we would say that the odds of obtaining a degree within six terms versus more than six terms (the combined equal to 7 terms, equal to 8 terms and more than 8 terms categories) are 115% higher for a female student compared to a male student given that all of the other variables in the model are held constant. Likewise, the odds of obtaining a degree within seven terms (the combined less than or equal to 6 terms and equal to 7 terms categories) versus more than seven terms (the combined equal to 8 terms and more than 8 terms categories) are 115% higher. The odds of obtaining a degree within eight terms versus more than eight terms are 115% higher. That is, female students are more likely to spend less time to obtain a degree than male students. Similarly, students who study the Bachelor's Programme in Mathematics and Economics are more likely to complete a degree on time compared to students who study the Bachelor's Programme in Mathematics. Students who finished Mathematics I within the same term are more likely to graduate earlier compared to those who failed to pass the course within the same term. Grade was treated as continuous and there was no direct statistical evidence of nonlinearity in the logit. The interpretation is that for one level increase in grade in Mathematics I, the odds of obtaining a degree within six terms versus more than six terms are 1.5 times higher, given the other variables in the model are held constant. Because of the proportional odds assumption, the same increase is found between within seven terms and more than seven terms, as well as between within eight terms and more than eight terms.

6 Conclusion

Before making conclusions, we must remember the limitations of this study. It is based on the information that is available to obtain from Ladok. Furthermore, we have only analysed the subjects who have passed Mathematics I. Subjects who have met the general requirement for obtaining a degree but have not applied for it did not appear in the study.

Our goals were to identify factors that were associated with students' time-to-degree, as well as to build a model that could be utilized to predict the probability of students' degree completion at the Department of Mathematics at Stockholm university. A binary logistic regression model with $Within3years$ as response variable was obtained through both purposeful selection and stepwise procedures. While this model is parsimonious, it still has an adequate fit and an acceptable predictive ability. It could be used to predict the probability of an individual student's degree completion within three years. Variables that are highly associative with the response are gender, programme, that whether a student finished the Mathematics I within

the same term and grade in Mathematics I. Among all the statistically significant variables, that whether a student finished Mathematics I within the same term has the strongest effect while its estimate contains the largest uncertainty in terms of odds ratio effect.

After obtaining the binary logistic regression model, we built a proportional odds model with a four-level response variable, which was not intended to be compared with the binary logistic regression model in the first place. We wanted to provide an alternative method that enabled us to study students' time-to-degree more specifically. Since we could make full use of the data, the proportional odds model contains actually more information. Using also the purposeful selection, we obtained a model that could be able to use for predicting degree completion within three years, within seven terms and within eight terms, as well as to some extent, for quantifying the effect of individual factors. The model consists of the same variables as in the binary logistic regression model. It is simple but the predictive power is relatively weak and we could not assess the fit of the model due to the difficulty of implementation of the methods in reality.

7 Discussion

In both studied models, the variable that describes whether a student has finished the Mathematics I within the same term has the strongest effect in terms of the odds ratio among all the influential variables. Those who passed Mathematics I on time have a higher probability to graduate on time. I personally do not find it surprising because Mathematics I offers fundamental knowledge in mathematics and it is a prerequisite for higher level courses in mathematics. Not having passed the Mathematics I on time certainly makes students spent more time to pass higher level courses. Programme directors and the administration of the department may want to emphasise the importance of passing Mathematics I on time on the introduction day of Mathematics I.

Programme has the second strongest effect. Students who study the Bachelor's Programme in Mathematics and Economics (M+E) are more likely to complete their degree on time compared to those who study the Bachelor's Programme in Mathematics (M). Personally, I did not expect that there was such a large difference between the two programmes regarding its association with degree completion length, since students who applied for M had a higher level of mathematics knowledge. The prerequisite for entering M+E is Mathematics D while for entering the M+E is only Mathematics C. One of the reasons that could be possible is that students who study M can only specialize in a mathematics subject while students who study M+E can not only specialize in a mathematics subject but also an economics subject. Mathematics subjects are often considered as difficult subjects. If a student

who studies M+E and has chosen a mathematics subject as he/her major, he/she still has another option to choose if he/she finds it hard to complete the programme and wants to change his/her specialization. Another reason could be that students are motivated to finish the programme earlier due to more job opportunities in the mathematical, statistical and economical labour market

We suspected that older students were less able to graduate on time, but it was shown that age was not a significant factor at 0.05 level in both models. It may be interesting to investigate why we did not see any significant impact. Would it become significant if our study sample had been larger?

Remember that grade in Mathematics I is an ordered variable. We examined whether the model having treated grade as continuous was substantially better than the model having treated grade as categorical. We found that the model having treated grade as continuous was adequate, and considering also other reasons we decided to treat grade as continuous in both final models. However, we could not really say that grade had a positive linear relationship with the responses. We could have tried to add a quadratic term of grade and checked if such model would perform better after we had decided to treat grade as continuous. Though we have already presented our results, it is still interesting to find out if adding a quadratic term really improves the performance of the final models. We add a quadratic term of grade to the final binary logistic regression model. We find that the model containing a quadratic term of grade performs actually better. It has a lower value of AIC (175.695), a higher value of AUC (0.8055 without cross validation and 0.7626 with cross validation) and a better goodness of fit (p-value of the HL test = 0.5460). However, due to time limitations, we do not analyse this model in more detail. Based solely on the above-mentioned advantages, the model containing a quadratic term of grade can be used as a new working model. Nevertheless, more analysis and further discussion on this model are needed and readers are encouraged to study the model further.

For future studies on the same subject, I would suggest using a larger sample size in order to quantify the effects precisely. The 95% confidence intervals for the odds ratios in both models are very wide. This is because we have a small sample size. Furthermore, other factors such as high school performance, admittance group, marital status can be taken into consideration if the information is available. Including such kinds of variables may describe students' time-to-degree in a better way. And last but not least, I would recommend that one can try to assess the global fit of a proportional odds model by hand. A few tests were listed at the end of section 4.2.1 and readers are encouraged to read Chapter 8 of Hosmer et al.(2013) for detailed information.

Appendix

A.1 Description of the data files from Ladok

Data file	Variables in the data files
All registered full-time students in programmes from autumn 2007 to autumn 2013 *N = 665*	Birthday Gender Study programme code Registration year of the study programme
All students who obtained a bachelor degree from 2008.05 to 2016.09 *N = 161*	Birthday Gender Study programme code Degree code (type of degree) Specialization ECTS-Credits for the programme The date when the degree certificate was issued The date when the study programme was finished
All students who passed Mathematics I from autumn 2007 to spring 2016 *N = 1298*	Birthday Gender Study programme code Registration term ECTS-Credits for the course The date when students passed the course

A.2 Frequency tables for the second data set

Table 18: Summary of *Age*

N	Mean	Median	Std Dev	Minimum	Maximum
223	22.08	28.5	4.22	18	46

Table 19: Table of TimetoDegree by Gender

	Gender		
TimetoDegree	Male	Female	Total
1	15	17	32
2	5	4	9
3	7	7	14
4	110	58	108
Total	137	86	223

Table 20: Table of TimetoDegree by Programme

	Programme		
TimetoDegree	M+E	M	Total
1	19	13	32
2	5	4	9
3	5	9	14
4	77	91	108
Total	117	106	223

Table 21: Table of TimetoDegree by Maths1sameterm

	Maths1sameterm		
TimetoDegree	No	Yes	Total
1	2	30	32
2	1	8	9
3	4	10	14
4	56	112	168
Total	63	160	223

Table 22: Table of TimetoDegree by Grade

	Grade					
TimetoDegree	A	B	C	D	E	Total
1	15	6	3	4	4	32
2	4	0	1	4	0	9
3	3	1	6	2	2	14
4	29	30	32	46	31	168
Total	51	37	42	56	37	223

Table 23: Table of TimetoDegree by Regyear

	Regyear						
TimetoDegree	2007	2008	2009	2010	2011	2012	Total
1	3	0	5	9	1	14	32
2	1	1	2	1	1	3	9
3	0	1	2	3	5	3	14
4	20	20	32	22	39	35	168
Total	24	22	41	35	46	55	223

A.3 Tables and figures for the fitted models

Table 24: Results of the first fitted multivariable model using all the significant variable from step 2, -2log(L) = 329.573

Parameter	Coeff.	SE	*P*-value
Gender	0.7849	0.3273	0.0165
Maths1sameterm	1.0038	0.4654	0.0310
Grade	0.2899	0.1259	0.0214
regyear	0.1680	0.1026	0.1016

Table 25: Results of the fitted multivariable model with exclusion of *Regyear*, -2log(L) = 332.295

Parameter	Coeff.	SE	*P*-value
Gender	0.8164	0.3248	0.0119
Maths1sameterm	1.0057	0.4628	0.0297
Grade	0.3168	0.1241	0.0107

Figure 2: ROC Curve for $\mathbf{M_b}$

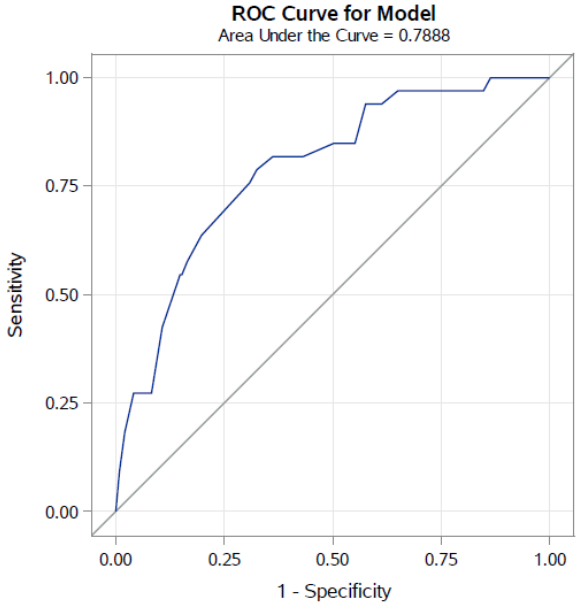


Figure 3: ROC Curve for $\mathbf{M_b}$ with cross validation

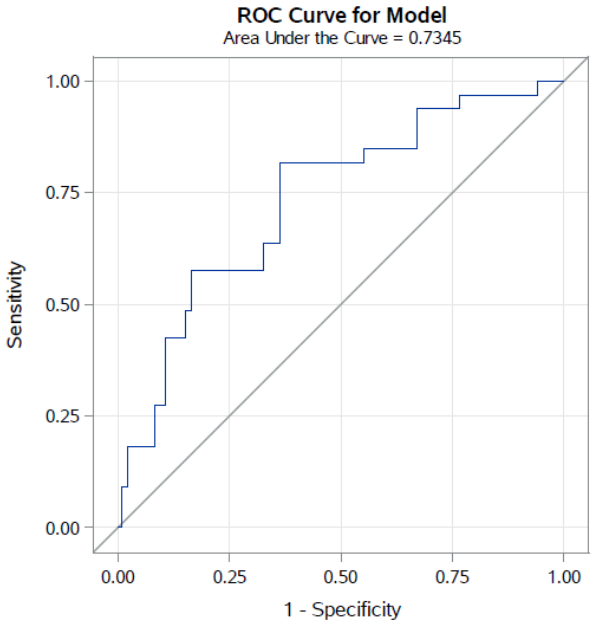
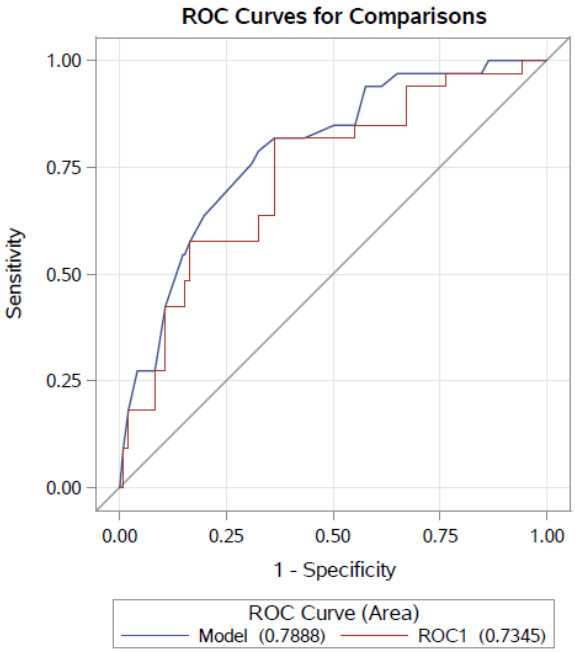


Figure 4: ROC Curve for $\mathbf{M_b}$ with and without cross validation



A.4 Full calculation of the 95% confidence interval for logit[Pr(Obtained a degree ≤ 3 years)]

The formulae for calculating a $100(1-\alpha)\%$confidence interval for a true logit is given by (Agresti(1), p194)

$$\text{CI} = \text{logit}[\hat{\pi}(\mathbf{x})] \pm z_{\alpha/2}\sqrt{\mathbf{x}\hat{cov}(\hat{\boldsymbol{\beta}})\mathbf{x}'}$$

where $\mathbf{x} = (1, x_1, \cdots x_{p-1})$ = dummy variable 1 for intercept and values of the $p-1$ explanatory variables, $\hat{cov}(\hat{\boldsymbol{\beta}})$ is the estimated covariance matrix of the parameter and $z_{\alpha/2}$ is the percentile point of a standard normal distribution.

That logit[Pr(Obtained a degree ≤ 3 years)] = -1.44 was calculated and $z_{\alpha/2} = 1.96$, $\mathbf{x} = (1, 1, 0, 1, 4)$. The covariance matrix was obtained by SAS (see Figure 5).

Figure 5: Estimated Covariance Matrix

NAME	Intercept	gender1	progr1	maths1sameterm1	ngrade
within3years	-6.24984	0.96578	1.19032	1.68565	0.53956
Intercept	1.01432	-0.12241	-0.18970	-0.48135	-0.09967
gender1	-0.12241	0.16363	-0.00558	0.01590	0.00810
progr1	-0.18970	-0.00558	0.18303	0.00551	0.02330
maths1sameterm1	-0.48135	0.01590	0.00551	0.59503	-0.02196
ngrade	-0.09967	0.00810	0.02330	-0.02196	0.02832

Plugging in the values in $\sqrt{\mathbf{x}\hat{cov}(\hat{\boldsymbol{\beta}})\mathbf{x}'}$ we obtain 0.3770. Plugging the values in the formulae, the corresponding confidence interval for logit[Pr(Obtained a degree ≤ 3 years)] is

$$CI = (a, b) = -1.44 \pm 1.96 \cdot 0.377 = (-0.701, 0.101).$$

Transforming back to the original probability scale, we finally obtain a 95% confidence interval

$$(0.1017, 0.3316)$$

for logit[Pr(Obtained a degree ≤ 3 years)].

References

[1] Universitets Kanslers Ämbetet *Så finansieras högskolan*, Retrieved from http://www.uka.se/fakta-om-hogskolan/universiteten-och-hogskolorna.html, 2016-10-17.

[2] Alan Agresti (2002) *Categorical Data Analysis*, 2nd Edition, New Jersey: Johan Wiley & Sons.

[3] Alan Agresti (2010) *Analysis of Ordinal Categorical Data*, 2nd Edition, New Jersey: Johan Wiley & Sons.

[4] David W. Hosmer, Jr., Stanley Lemeshow, Rodney X Sturdivant (2013) *Applied Logistic Regression*, 3nd Edition, New Jersey: Johan Wiley & Sons. Retrieved from http://onlinelibrary.wiley.com.ezp.sub.su.se/book/10.1002/9781118548387.

[5] 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

[6] SAS homepage https://support.sas.com/documentation/cdl/en/statug/63347/HTML/default/viewer.htm#statug_logistic_sect036.htm, 2016-12-06

[7] (2011) Feng Zhang, *Cross-validation and Regression Analysis in High-dimensional Sparse Linear Models*, Stanford University.

[8] SAS Institute Inc. (2015) *"Usage Note 39724: ROC analysis using validation data and cross validation"*.Retrieved from http://support.sas.com/kb/39/724.html, 2016-11-24