VRIJE UNIVERSITEIT

ECONOMETRICS & OPERATIONS RESEARCH

INTREGRATIVE PRACTICAL

Education

Author
STIJN SNEL
BRAM GRIFFIOEN
FABIËNNE TUIJP
SAFIYAH NIAMAT

Coordinator
Mr. S. TELG
Mrs. A. ESTEVEZ FERNANDEZ

January 31, 2022

1 Introduction

For this project, we have studied the performance of students in the courses Mathematics and Dutch in University, in relation to the performance of secondary education in high school and in relation to the success of students in their first year of university studies. Next to this, we have conducted an operational study towards the economical question of how to locate students that follow the course Mathematics D in smaller schools. This assignment will be of particular interest to high schools and universities that are affected by the under performance of students in the courses Dutch and Math provided in University. As well as an operational solution for cost-efficiency for the course mathematics D provided in smaller schools. With help of the data obtained in the questions we eventually form a useful advice for the Ministry of Education whether to change the hours devoted to the high school subjects.

In order to find an answer to our first main question, we looked into two sub-questions regarding the given data set containing information about the grades, subjects, studies and ECT scores. To know if the the grades in high school actually say something about the performance of the students in college, we studied if the grades of the subjects Mathematics and Dutch in various university programs can be explained by the grades obtained in high school. We subdivided the data into Gymnasium and Atheneum students in high school, meaning that we divided the data into students who follow Latin and Greek and students who do not. In order to solve this questions we formed three sub-questions:

1. Do the different type of High school Profiles in Gymnasium and Atheneum explain the grade for Mathematics in University?
2. Do the different type of High school Profiles in Gymnasium and Atheneum explain the grade for Dutch in University?
3. Do the different type of University Studies explain the grades in University?

We made use of our knowledge of the Ordinary Least Squares to implement it into R. With use of several OLS regression results we were able to solve the sub-questions and therefore the question if the high school grades can explain the grades in various university programs.

For the the second sub-question of the first part we studied if the high school subjects can explain the success of students in the first year of university. In other words whether a students gets at least 42 ECT's in the first year of studies in relation to having followed the different subjects. Again, we made use of the OLS with regards to the required data. This provided us enough results to find the desired answer.

For the second main question about the four schools, we also looked at two sub-questions:

1. Where to give the classes to the students?
2. How to share the teaching costs of using a classroom and paying the teacher among the four schools?

Firstly, we answered the question about where to give the classes to the students. Secondly, we focused on the question about how to share the costs of offering these classes among the four schools. In order to answer these questions, we made use of the provided data. These data tables contained, among other things, information about the number of students, the travelling distances and times, the available rooms in the schools including the costs of these rooms.

The second sub-question was about how to share the teaching costs of using a classroom and paying the teacher among the four schools. Of course there were multiple solutions to this problem. In the sub-section about this question we will give our distribution and the reason why we chose for this way.

This report is structured in the following way: First we will give a brief explanation about our data set, to discuss the main features such as missing data and particular interesting parts. Second, we will explain which models and techniques we have used to analyze the data, methods and assumptions that are important such as the OLS assumptions. Third, we will go over the results for both of the main questions and discuss various methods we have tried to find in our opinion the best models. And finally, our recommendation and final conclusion.

2 Data

This report will be providing answers to two main questions, divided into several sub-questions. For both questions we will use different kinds of data.

Question 1:
The data used for the first main question was provided to us by the Ministry of Education, via Canvas and consisted of a data frame with a sample of 1,844 observations. This data frame included the grades of students in the different subjects in high school, the grades in the first year of university studies, and the number of credits obtained in the first year of university studies.

First, what made this the data set interesting at first glance are the different types of high school profiles. As not every student follows the same type of courses, there is a differentiation in groups that do and do not follow ancient languages including Latin and Greek. Also the types of Math differs for every student. The different types of mathematics in our data set are A, B, C and D. However, D is considered a subset of B because it can not be followed independently. We decided to split up our data into several subcategories. Our two subcategories are Gymnasium, students that follow Latin and Greek, and Atheneum. Within these subcategories we combined all the students according to the different types of math they follow. So, in total we will be working with 8 different subcategories in this report.

Table 1: Summary Statistics: Education Data

Statistic	N	Mean	St. Dev.	Min	Pctl(25)	Pctl(75)	Max
MathUni	1,854	7.0	1.3	2.7	6.1	7.8	10.0
DutchUni	1,854	7.2	1.2	3.3	6.4	8.0	10.0
Mathematics.A	552	7.4	1.5	5.0	6.1	8.6	10.0
Mathematics.C	374	7.4	1.5	5.0	6.0	8.7	10.0
Mathematics.B	928	7.4	1.5	5.0	6.2	8.7	10.0
Mathematics.D	404	7.5	1.4	5.0	6.3	8.7	10.0
Dutch	1,854	7.5	1.5	5.0	6.3	8.8	10.0
English	1,854	7.5	1.5	5.0	6.2	8.8	10.0
Latin	868	7.5	1.5	5.0	6.3	8.7	10.0
Greek	868	7.5	1.5	5.0	6.3	8.7	10.0
ECTs	1,854	38.7	12.4	6	30	48	60

Second, as shown in the summary statistics, table 1, we observed that there are a total of 11 different numerical variables that, in our opinion, is of particular interest to this project. We observed that both Dutch and English for both University and high school courses, are indeed followed by every student. Further we observed that the standard deviation, mean and max have the same value in every course, concluding that there is almost no distinction between these course grades on the first sight. Also to answer the last sub question whether the University grades can be explained by the student success in their first year, we saw that the mean is below the minimum passing requirements of 42 ECTs.

Table 2: Correlation Matrix: Mathematics Courses

	mathUni	dutchUni	MathA	MathB	MathC	MathD
mathUni	1	0.050	0.582	0.537	0.551	0.358
dutchUni	0.050	1	0.047	−0.038	0.124	−0.044
MathA	0.582	0.047	1			
MathB	0.537	−0.038		1		
MathC	0.551	0.124			1	−0.015
MathD	0.358	−0.044			−0.015	1

Table 3: Correlation Matrix: Alpha Courses

	mathUni	dutchUni	Dutch	English	Latin	Greek
mathUni	1	0.050	0.070	0.071	0.070	0.022
dutchUni	0.050	1	0.292	0.217	0.239	0.180
Dutch	0.070	0.292	1	−0.013	−0.001	−0.031
English	0.071	0.217	−0.013	1	0.008	0.017
Latin	0.070	0.239	−0.001	0.008	1	0.056
Greek	0.022	0.180	−0.031	0.017	0.056	1

And third, as shown in the correlation matrices in table 2 and 3 we saw some very interesting relationship results between our variables. Both between the alpha courses: English, Dutch, Latin and Greek with the Dutch grade in University and the relationship between the mathematics courses in high school with the mathematics in University we saw a very high positive correlation. And we could show some first indications that there could be an overall positive relation between the performance of students in university and the performance on some of the high school courses.

Question 2:
For the second main question we were given a data set on canvas. In this data set in excel, information about the four schools was given. First, there was a table with travel distances in kilometres between the four schools, followed by a table with the travel time in minutes. Then, the information was split per school. For each school, we were given the costs of the teacher per hour, the number of students for mathematics D in year 1 and year 2, separately. Next to this, there was an overview of the available rooms in the schools, including their capacities and costs per hour. In total there were five rooms per school. We also knew that the costs per minute of travelling were 40 cents per student.

To solve the first sub-question, we needed to minimise the total costs. This means that we needed information about travel costs, teaching costs and room costs. To calculate these costs we needed the table about travel time in minutes, the 40 cents per student, the teaching costs given in the table per school, the capacity of the room and the accompanying costs of this room. As said above, we were also given a table with the travel distance between the schools in kilometres. However, we did not use this table. We only used the table with the travel time in minutes. It is true that the

higher the distance between the schools, the higher the travel time between the schools. Hence, we did not use the table in kilometres. This is also due to the fact that we were given the travel costs per student per minute instead of the travel costs per kilometre.

For the second sub-question, we needed the total costs computed in the first sub-question. To give an answer to this question we did not need other tables from our given data set. So only table we did not need to use to answer this main question was the table with distances in kilometres between the schools.

3 Models

3.1 Models used in Question 1

During this project we have made use of the OLS model. This model was introduced to us in last period's course, Econometrics 1. We see this tool as a very useful method to find a relationship between our dependent variables, grades in university and our explanatory variables such as the high school courses, studies and ECT's. For our model to be consistent we have devoted a list of assumptions the models have. We will discuss these assumptions below. An example of the models that we used in this project can be seen below.

Interpretation of our Model:
We interpreted our models in the following way: first we looked at the adjusted R squared value, because with both the gymnasium model and the Atheneum model being a multiple linear regression model showed that was the best fit for our model. Second we looked at the F-statistics and P-value. Before we made our models, we decided to choose an alpha value of 0.05 because we thought it was a good balance between a Type 1 error and a Type 2 error. Not small enough to prevent that there was more of a chance that we would NOT have rejected the null, when in fact we should have and large enough to be sure that we were correctly accepting/ rejecting the null hypothesis. We looked for a large F-statistic and a smaller p-value than 0.05, for the model to be significantly interesting. Third we looked at the t-statistics of our residuals. This was to make sure that assumption 4 as mentioned below, is normally distributed. Fourth, we looked at the coefficients to see which one has a large enough influence on the fitted values, and lastly we checked if the standard errors were not too large.

Our models:
As shown below, we mostly used these 3 models for our two questions of part one. The first and second model applies a multiple linear regression model to explain the Variable Y_i, this value can be Mathematics or Dutch in University. When applying the model we did not simultaneously use beta 3 till beta 6, however beta 4 and beta 6 can be simultaneously taken in one model. The third one is a simple linear regression model.

Atheneum OLS with Mathemtics Univeristy grade as dependent variable:

$$\begin{aligned} \mathrm{Y}_i = \beta_0 &+ \beta_1 \mathrm{Dutch}_i \\ &+ \beta_2 \mathrm{English}_i \\ &+ \beta_3 \mathrm{MathematicsA}_i \\ &+ \beta_4 \mathrm{MathematicsB}_i \\ &+ \beta_5 \mathrm{MathematicsC}_i \\ &+ \beta_6 \mathrm{MathematicsD}_i + u \end{aligned}$$

Gymnasium OLS with Dutch Univeristy grade as dependent variable:

$$\begin{aligned} \mathrm{Y}_i = \beta_0 &+ \beta_1 \mathrm{Dutch}_i \\ &+ \beta_2 \mathrm{English}_i \\ &+ \beta_3 \mathrm{MathematicsA}_i \\ &+ \beta_4 \mathrm{MathematicsB}_i \\ &+ \beta_5 \mathrm{MathematicsC}_i \\ &+ \beta_6 \mathrm{MathematicsD}_i \\ &+ \beta_7 \mathrm{Latin}_i \\ &+ \beta_8 \mathrm{Greek}_i + u \end{aligned}$$

ECT's OLS with Mathemtics Univeristy grade as dependent variable:

$$\mathrm{Y}_i = \beta_0 + \beta_1 \mathrm{ECT's}_i + u$$

Verifying the OLS assumptions: [1]
As mentioned above, our main models for this assignment were based on a linear regression model. For our OLS model to be consistent and to find the best linear unbiased estimator, our model has to pass five underlying assumptions. These assumptions can be defined as: the model is linear in its parameters, there is a random sampling of observations, the mean of the error terms should be zero, there is no multi collinearity and there is homoscedasticity.

Assumption 1: The model is linear
First, to prove that our model is linear in its parameters we plotted all the individual parameters against the explanatory variables such as Mathematical University to show the relation between the various variables. In addition, we checked the correlation matrix between each variable and the dependent variable, which measures how closely the variables are linearly related to each other, see appendix figures 2 and 3. The second way we will show linearity is according to the correlation between variable used in the model, see the correlation matrices in tables 3 and 2. If there is a high positive correlation we can assume that there indeed is linearity.,

Assumption 2: There is a random sampling of observations
Second, as our data set was given to us, we assumed that the observations were sampled randomly. And therefore assumed this assumption is passed.

Assumption 3: The error terms mean should be zero
Third, the expected value of the mean of the error terms of OLS regression should be zero given the values of independent variables. In other words, the distribution of error terms has mean zero and does not depend on the independent variable. We did verify this by taking the mean of our OLS models' residuals.

Assumption 4: The error terms follow a normal distribution:
Fourth, this can be verified in multiple ways. However, we chose to verify it with three measurements. We first looked at the histogram of the error term and checked if this indeed showed a bell

curve as the distribution, then we looked at the t-statistic.

Assumption 5: homoscedasticity
This assumption requires that the variance of the error term is the same for every value. We will be reviewing this by applying homoscedasticity on the regression residuals, if the variance is constant we will conclude that the error term is homoscedastic.

3.2 Models used in Question 2

To solve question 2, we formulated the question as an optimization problem. We have learned how to do this during the course Operations Research 1. We formulated this model for one year, so that it can be used for year 1 and year 2 separately. We wanted to know where to give the classes, which means we needed to decide on a school and in this school we needed to pick rooms for year 1 and year 2. We wanted to choose the school and the rooms such that the total costs were as low as possible. The model that we used to solve this question is as follows:

Number the schools from 1 to 4 and the rooms in each school from 1 to 5. Then introduce binary variables X_{ij} with $i = 1, ..., 5$ and $j = 1, ..., 4$ for classroom i from school j. These variables can take value 0 or 1. If room i from school j is used, $X_{ij} = 1$, otherwise $X_{ij} = 0$.

Introduce variables y_j with $j = 1, ..., 4$ for the number of students in school j and the variables d_{kj} with $k = 1, ..., 4$ and $j = 1, ..., 4$ for the distance in minutes between school k and school j. In this model, y_T represents the total number of students in one year.

Let v be the travelling costs per minute per student and let R_{ij} be the room capacity of room i in school j. In our case, the travelling costs per student per minute were given and were 40 cents.

Now, let C_{ij} be the room costs of room i in school j, Tr_j the travel costs when the classes are given in school j and Te_j the teaching costs when the classes are given in school j.
Then the model is:

$$
\begin{aligned}
\min \quad & \sum_{j=1}^{4}\sum_{i=1}^{5}(C_{ij} * X_{ij}) + (Tr_j + Te_j) * X_{ij} \\
\text{s.t.} \quad & \sum_{j=1}^{4}\sum_{i=1}^{5} R_{ij} * X_{ij} \geq y_T \\
& \sum_{j=1}^{4} y_j = y_T \\
& \sum_{k=1,k\neq j}^{4} (d_{kj} * v) * y_k = Tr_j \qquad j = 1, \ldots, 4 \\
& X_{ij} \in \{0, 1\} \qquad i = 1, \ldots, 5 \quad j = 1, \ldots, 4 \\
& \sum_{j=1}^{4}\sum_{i=1}^{5} X_{ij} = 1
\end{aligned}
$$

4 Results

4.1 Question 1

4.1.1 Question 1.1

We wanted to know if the grades of the subjects of Mathematics and Dutch in various university programs can be explained by the grades obtained in high school on the subjects of Mathematics, Dutch, English, Latin and Greek. In other words, we wanted to find out if the mentioned grades in the various university programs depended on the grades in high school. In order to find this out, we divided the data into two parts. The division consists of an Atheneum and a Gymnasium direction. This enabled us to find an answers to following questions:

1. Do the different type of High school Profiles in Gymnasium and Atheneum explain the grade for Mathematics in University?
2. Do the different type of High school Profiles in Gymnasium and Atheneum explain the grade for Dutch in University?
3. Do the different type of University Studies explain the grades in University?

Because we divided this question into three 3 sub-questions we, therefore, had a hypothesis for each sub-question. For the first sub-question, as stated above, we thought that a significant part of the university mathematics grade can indeed be explained by the obtained high school math grades. However, we did think there is a difference between the impact of the different types of math but that there is no significant difference between the Atheneum and Gymnasium math grades.
For the second sub-question, we did not think that the alpha subjects affect the Dutch university grades significantly. Maybe at most, only the Dutch course shows a bit of impact. However, we did think that having followed Latin and Greek is an advantage in relation to the Dutch University grade. And lastly, we thought that the type of studies at the university could play a part in the impact on the University grades.

In order to solve the sub-questions we have used our knowledge of the Ordinary Least Squares to calculate different OLS regression results. This way we gained data such as the correlation coefficients, R-squared, Adjusted R-squared, coefficient's p-values and F-statistics.

We have combined, in our opinion, the most useful data from table 14 and 15 from the appendix into table 4. The reason why we combined this particular data is, because they have the highest R-squared values out of all the results. R-squared measures the strength of the relationship between your model and the dependent variable on a convenient 0-100 percent scale. Since we used multiple variables, we actually preferred looking at the adjusted R-squared, since it is a modified version of R-squared that accounts for predictors that are not significant in a regression model. Therefore with the model we used, we thought that the results in table 4 are the most explainable results, because they have the highest adjusted R-squared values.

For the first sub-question we looked at the different mathematics data of both Atheneum and Gymnasium with regards to Mathematics in University. In table 14 and 15 we looked at coefficients

with a p-value smaller than an Alpha of 0.05. Looking at the math university results of the Gymnasium part of the table we see that all the mathematics coefficients (at points 1, 2, 3 and 4) have a p-value smaller than 0.05. In addition, the subjects Dutch (1) and English (3) also have a p-value lower than 0.05. Just as at Gymnasium, the math coefficients at Atheneum have a p-value lower than 0.05, but this time only Dutch (3) has a p-value lower than 0.05. Since the eligible coefficients of Dutch and English are already very small values, we can say that their impact is virtually negligible, despite the small p-value. This leaves only the math courses, therefore we can say that the math grades at the various university programs can be explained by all the math grades obtained in high school. However, there is also a difference in the amount of impact of the math courses. To indicate this distinction we looked at the adjusted R-squared values (adj. R-squared values). In table 4 we have assembled the four coefficients with the highest adj. R-squared values. As you can see in both situations, math B explains a big part of the math university grade and math A and the combination of math B/D also shows some obvious impact. This proves our hypothesis to be true.

In comparison to the Math OLS results, we see that for the Dutch results only the alpha subjects have a p-value smaller than 0.05. For Gymnasium: Dutch, English, Greek and Latin all explain the Dutch grade with different margins. For Atheneum we also saw that Dutch and English have impact. However for both Gymnasium and Atheneum there are several study programs with very low adjusted R-squared. In table 4 we added the only study program that shows coefficients with an acceptable p-value and a relevant adjusted R-squared value, namely 0.534. Therefore, we can claim with a certain degree of certainty that the Alpha subjects do explain various (mostly small) margins of the Dutch University grades. The dutch course mainly has the biggest impact of the four alpha courses, but not nearly as much as several math courses have on the Math university grades.

It is clear that our prediction was not totally correct. At first, we thought that the alpha subjects would have any significant impact which appeared not to be true. However, our impression that the Dutch high school grades would explain the most, is proven right. Lastly, it is interesting to see that indeed the Gymnasium grades explain more of the University grades than the Atheneum grades. Yet, this does not tell anything about Gymnasium students getting higher grades, as can be seen in figure 1.

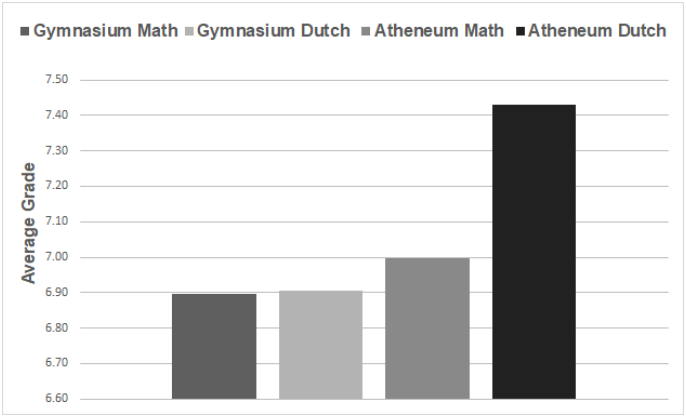

Figure 1: Average University grades per Profile

In table 16 in the appendix you can see our OLS regressions results of the various study programs with regard to both the university Math and Dutch grades. It is clear to see that none of the study programs have any affect on the two grades. For both Math and Dutch the adjusted R-squared is so small that we can state that the studies do not explain any margin of the two grades. These results reject our hypothesis.

To conclude, the grades of Mathematics in various university programs can be explained by the four math subjects in high school, namely math A, B, C and B/D. Out of these four, Math B explains the most of the Math University grade. The subjects Dutch, English, Latin and Greek do not explain any significant margin of the university Mathematics grades. The Dutch university grades on the other hand, can be explained by the language subjects. The mathematics subjects though, have no influence at all. Dutch has the greatest impact on the university grades out of the four subjects.

Table 4: OLS Regression Results Gymnasium and Atheneum

	Dependent variable:				
	MathUni Gymn.	MathUni Gymn.	DutchUni Gymn.	MathUni Ath.	MathUni Ath.
	(1)	(2)	(3)	(4)	(5)
Dutch	0.041*	0.091**	0.268***	0.077*	0.038
	(0.021)	(0.045)	(0.021)	(0.046)	(0.047)
English	0.020	0.073*	0.161***	0.071	0.044
	(0.021)	(0.043)	(0.022)	(0.047)	(0.048)
'Mathematics B'	0.580***		0.116***	0.558***	0.440***
	(0.021)		(0.021)	(0.048)	(0.045)
'Mathematics A'		0.554***			
		(0.044)			
Latin	−0.005	0.033	0.122***		
	(0.021)	(0.044)	(0.021)		
Greek	0.051**	0.064	0.129***		
	(0.021)	(0.045)	(0.022)		
'Mathematics D'					0.307***
					(0.048)
Constant	1.923***	0.834	0.845**	2.047***	0.352
	(0.362)	(0.721)	(0.364)	(0.592)	(0.723)
Observations	262	259	262	262	202
R^2	0.761	0.400	0.543	0.361	0.421
Adjusted R^2	0.756	0.388	0.534	0.354	0.410

Note: *p<0.1; **p<0.05; ***p<0.01

4.1.2 Question 1.2

For the second question of part one, we looked into the relationship between the success of students in their first year of university studies and their results in high school. Rather, whether a student gets at least 42ECT's in the first year of studies in relation to having followed the different subjects.

OLS model:
Since we already know from question one, see the part above, that several high school subjects are indeed affecting the grades in University, we now only make use of the variables ECT's, Math and Dutch University grades. We used these variables to create a simple linear regression model, see table 5 below. We did not include the several high school studies as mentioned above and shown in the appendix in table 16. Because these studies are insignificant to our model.

What we did observe from our model is that both the coefficients for Math and Dutch have an acceptable p-value, since they are smaller than 0.05. The coefficients of both the ECT'S for Math and Dutch University have a relevant adjusted R-squared value (0.530 and 0.392 respectively). However, we also observed that the coefficients seem in some way very small, meaning that the impact of the university grades, and therefore the high school grades are virtually negligible.

Since the impact of both the university and high school grades on the obtained ECT's is virtually negligible, we can state with a certain degree of certainty that the success of students in their first year of university is unrelated to the grades obtained in high school. Moreover, the subjects Dutch and Math in university also tell very little to none about obtaining 42 ECT's in the first year of university.

Table 5: OLS regresson: ECT's

	Dependent variable:	
	Math Uni	Dutch Uni
	(1)	(2)
ECTs	0.074***	0.061***
	(0.002)	(0.002)
Constant	4.072***	4.828***
	(0.066)	(0.071)
Observations	1,854	1,854
R^2	0.530	0.393
Adjusted R^2	0.530	0.393
F Statistic (df = 1; 1852)	2,090.028***	1,201.504***
Note:	*p<0.1; **p<0.05; ***p<0.01	

4.2 Question 2

4.2.1 Question 2.1

For the following questions, four schools are considering offering the subject of Mathematics D to their students. However, this subject is taken by a small proportion of students, which makes it too costly to offer this subject in small schools. That is why the four schools have decided to work together to jointly provide Mathematics D to their students. Given our data, we were asked to solve two sub-questions. We needed to find the best school to give the classes of Mathematics D, and then decide on the best rooms for year 1 and year 2 to give the classes. As already said in the section Models, we wanted to minimize the total costs. Hence to solve this question, we formulated the problem as an optimization problem, in which we minimized the objective function using various constraints. We already described this model before in the section Models.

To start, we divided the total number of students into two groups: students from year 1 and students from year 2. For these two groups, we calculated all the costs separately, since the students from year 1 and 2 are getting separately classes for mathematics D. There were three types of costs that we considered to minimize the total costs: travel costs, classroom costs, and teaching costs. We started off with the travel costs. We calculated the travel costs for year 1 and 2 from each school by multiplying the number of students of that year by the travel time in minutes, which we then multiplied by the 40 cents, the travel cost per minute per student. We obtained the result of how much it would cost per school to travel to each school. So, for school A for example, we got the travel costs of school B travelling to A, school C travelling to A and school D travelling to A. This is what we did for all the schools. The results can be seen in table 6.

Table 6: Travelling costs per school

Travelling costs per school	Year 1	Year 2	Total
From school A to School B	€ 126.00	€ 92.40	€ 218.40
From school A to school C	€ 60.00	€ 44.00	€ 104.00
From school A to school D	€ 42.00	€ 30.80	€ 72.80
From school B to school A	€ 92.40	€ 117.60	€ 210.00
From school B to school C	€ 114.40	€ 145.60	€ 260.00
From school B to school D	€ 114.40	€ 145.60	€ 260.00
From school C to school A	€ 72.00	€ 48.00	€ 120.00
From school C to school B	€ 187.20	€ 124.80	€ 312.00
From school C to school D	€ 28.80	€ 19.20	€ 48.00
From school D to school A	€ 36.40	€ 44.80	€ 81.20
From school D to school B	€ 135.20	€ 166.40	€ 301.60
From school D to school C	€ 20.80	€ 25.60	€ 46.40

We wanted to compare the total costs including the costs of travelling back with the total costs excluding the costs of travelling back. When we included these costs, the travel costs had a bigger influence on the total costs than the teaching costs and the classroom costs.

So next, we calculated the total costs per school when all the schools travelled to one school, including the travelling costs when the students travel back to their own school. These costs should then be minimized. These results can be seen in table 7. You can see that travelling to school D has the lowest travelling costs for the students from both years.

Table 7: Total travelling costs when travelling to one school

Travelling costs per school	Year 1	Year 2 back	Year 2	Year 2 back
Schools going to school A	€ 200.80	€ 401.60	€ 210.40	€ 420.80
Schools going to school B	€ 448.40	€ 896.80	€ 383.60	€ 767.20
Schools going to school C	€ 195.20	€ 390.40	€ 215.20	€ 430.40
Schools going to school D	**€ 185.20**	**€ 370.40**	**€ 195.60**	**€ 391.20**

For the classroom costs, we needed to keep in mind that there is a capacity per room. In year 1 there were in total 57 students and in year 2 there were in total 53 students. We kept year 1 and year 2 separately and we wanted a whole year to fit into one classroom, since both years would not fit together in every classroom. We decided to find the best classrooms for each school by looking at the capacity and the costs for that room. We thought that one hour of mathematics D per week would not be enough, so we decided to give 2 hours per week, which meant that we had to double the costs of the rooms. We found the best classrooms at every school for every year and these are shown in table 8, including the costs of these rooms per 2 hours.

Table 8: Classroom costs per 2 hours

Room costs per school	Year 1 room	Year 1 costs	Year 2 room	Year 2 costs
School A	2	€ 240.00	3	€ 180.00
School B	2	€ 280.00	2	€ 280.00
School C	3	**€ 170.00**	3	**€ 170.00**
School D	4	€ 190.00	4	€ 190.00

The students from year 1 and year 2 will not have the mathematical D lesson at the same time, which meant that we could use the same class rooms for year 1 and year 2. So for some schools both year 1 and 2 use the same classroom. We came to the conclusion that school C has the lowest classroom costs.

The last costs that we considered are the teaching costs, which are constant for every school. The lowest teaching costs are from School A, which can be seen in table 9. We thought it would be best that the mathematical D lessons will be given on just one day, so the one lesson per week will have a duration of 2 hours. This means that the travel costs will stay the same, however the classroom costs and the teaching costs will be multiplied by 2.

Table 9: Teaching costs

Teaching costs	Total costs per hour	Total costs per 2 hours
School A	**€ 95.00**	**€ 190.00**
School B	€ 106.00	€ 212.00
School C	€ 110.00	€ 220.00
School D	€ 100.00	€ 200.00

We obtained the minimized costs by adding all the costs together for each school. These results can be seen in table 10 and in table 11. Here, we looked at two different scenario's. First, we looked at the total costs when the students travel back to their school. Second, we looked at the total costs when the students do not travel back to their school.

In table 10 we can see what the total costs are if the students travel back, so the total travel costs are bigger. In table 11 you can see the total costs when we looked at the case where the students travel once to the school where the class is given. We can see from these tables that the travel costs have a big influence on the total costs when you include travelling back.

Table 10: Total costs per school when travelling back

Total costs	Year 1	Year 2	Total
School A	€ 831.60	€ 790.80	€ 1622.40
School B	€ 1388.80	€ 1259.20	€ 2648.00
School C	€ 780.40	€ 820.40	€ 1600.80
School D	**€ 760.40**	**€ 781.20**	**€ 1541.60**

Table 11: Total costs per school when not travelling back

Total costs	Year 1	Year 2	Total
School A	€ 630.80	**€ 580.40**	€ 1211.20
School B	€ 940.40	€ 875.60	€ 1816.00
School C	€ 585.20	€ 605.20	€ 1190.40
School D	**€ 575.20**	€ 585.60	**€ 1160.80**

Looking at table 11, we can see that for year 1 school D is the best choice and that for year 2 school A is the best choice. In table 8, we already noted what rooms should be used in which school. So for school D this is room 4 and for school A this is room 3. However, when we look at table 10 we can see that for both years school D is the best choice. This is because the travel costs of school A are much larger than those of school D. The optimal solution would then be found when the students do not travel back. To answer this, we look at the numbers that are bold in table 11. The total costs for this case are €575.20 + € 580.40 which is equal to €1155.60. We can see that this amount is smaller than €1160.80, which would be the total costs when the students travel back.
As we already said, we did not use the table including the distances in kilometres between the schools. We did calculate the distances in kilometres between the schools and came to the conclusion that when all schools travel to school D, the travel distance was the smallest: 10.5 km. School D was followed by school A with 11.4 km.

4.2.2 Question 2.2

The second sub-question focused on how to share the costs among the four schools. There were two types of costs that needed to be shared. The travel costs are not taken into account when sharing the total costs, hence we only focused on the costs of using a classroom and the costs of the teacher. As seen above in sub-question 2.1, we chose to teach the classes for year 1 and 2 in school D, using room 4 for both years, if we take into account that the students should travel back. Later in this report we focus on the case when students do not travel back. This results in the total costs (teaching costs and room costs) of € 780.00, consisting of € 390.00 for year 1 and € 390.00 for year 2.

As already mentioned above, there were multiple ways to share the costs among the four schools. For instance, dividing the costs by 4, which results in every school paying the same amount of money. However, we decided to share the costs based on the number of students per school. Since we wanted to minimize total effort and total costs, we focused on the number of students. Opportunity costs explain this: since we wanted minimal effort and costs, we needed to make sure that there will be a minimum amount of students moving from one school to the other. We first focused on the case where the students travel back to their own school. School C has 30 students and school D has 29 students, which are the highest numbers. Because of this, it seemed better to move to school C or D, since you move less students than when you move the students of school C and D to another school. For question 2.1 above, we already computed that school D will be the cheapest, and by splitting the costs among the four schools based on the number of students we minimized the total effort put into this.

Also for this question, we split the number of students in two groups: the number of students with Mathematics D from year 1 and the number of students with Mathematics D from year 2. For each of these groups we calculated the amount of costs per student. This resulted in two numbers: €6.84 per student for all students of year 1 and € 7.36 per student for all students of year 2.

The total costs per school for year 1 can be computed by multiplying € 6.84 with the number of students in year 1 in that school. For year 2, the same holds. Multiply € 7.36 with the number of students in year 2 of that school. Adding these two numbers gave the total costs per school. The results can be found in table 12 below.

Table 12: Teaching costs and room costs per school when travelling back

Costs per school	Year 1	Year 2	Total
School A	€ 102.63	€ 80.94	€ 183.57
School B	€ 75.26	€ 103.02	€ 178.28
School C	€ 123.16	€ 88.30	€ 211.46
School D	€ 88.95	€ 117.74	€ 206.68
Total costs			€ 780.00

Table 13: Teaching costs and room costs per school when not travelling back

Costs per school	Year 1	Year 2	Total
School A	€ 102.63	€ 76.79	€ 179.42
School B	€ 75.26	€ 97.74	€ 173.00
School C	€ 123.16	€ 83.77	€ 206.93
School D	€ 88.95	€ 111.70	€ 200.65
Total costs			€ 760.00

When the students do not travel back to their own school, there was a different result as already mentioned in question 2.1. Then, we will teach the classes for year 1 in school D in room 4 and the classes for year 2 in school A in room 3. This results in total costs (excluding travelling costs) of € 760.00, consisting of € 390.00 for year 1 and € 370.00 for year 2. We use the same method for splitting costs, which results in the costs per school given in table 13.

5 Conclusion

As explained in the introduction, our project consists of two parts both divided into two sub-questions. For the first part, we had to find out if the grades of the subjects of Mathematics and Dutch in various university programs can be explained by the grades obtained in high school on the subjects Math, Dutch, English, Latin and Greek. We divided this question into three sub-questions for convenience.

With the use of several OLS regressions in R we came to the conclusion that only Mathematics subjects can explain a margin of the Math grade in university, where Math B has the most impact of the four. In contrast, the Dutch grade in college is explained precisely not by mathematics, but by the four language subjects. Out of the four Dutch explains the biggest margin of Dutch in university, but does not have as much influence as the several math subjects in high school have on the Math university grade. We also found that the different study programs have no affect on the university Dutch and Math grades. Whilst we thought that for example Econometric students would have better Math grades than say Law students. On the other hand our hypothesis for the first sub-question did appear to be true. Nevertheless, we did not predict the second sub-question correctly. We did not think that the language subjects besides Dutch would explain an significant margin of the Dutch grade. Interestingly enough all the alpha subjects did have impact. Moreover, following the subjects Latin and Greek means that more of the Dutch university grade could be explained by the four language subjects in comparison to only having followed Dutch and English.

At the second sub-questions of part one we had to find out if the success of students in the first year of university could be explained by the different high school subjects. Since we already found that the Math and Dutch grade in the various university programs could be explained in some way by the high school subjects, we could use the ECT's, Math and Dutch university grades for our OLS regression. This regression showed us that the adjusted R-squared- and p-value were acceptable to claim anything with a certain degree of certainty. However, the coefficient margins were virtually negligible, meaning that the impact of the high school and Math and Dutch grade in university is also negligible. The results were just as we predicted, since we thought that the content of the high school subjects were very different in comparison to the content of each study program.

For the second main question, we were asked to find the best school to give the classes of mathematics D in and then find the most suitable room. We were given a data set containing information about the four schools included. This information focused on travelling distances, travelling time, number of students per school split into two years and available classrooms including the costs of these rooms. Next, we needed to answer the question about how to divide the total costs among the four schools.

We worked in excel to answer both of these questions. Using different formulas, we calculated the travel costs when travelling to each school. We also calculated the costs of the classrooms and the teaching costs per school. We eventually wanted to find the school, classrooms and teacher with the lowest costs. To come to an answer to this question, we calculated the total costs per school, with year 1 and year 2 calculated separately. To calculate these total costs, we also looked at two different cases: one where the students travel back and one where the student do not travel back. For these two cases we made two different tables. We compared these costs and found that when we looked at the case where students travel back, school D was the cheapest school, using room 4 for both years. However, this is not the optimal solution, due to the fact that the travel cost will have a big influence on the total costs. The optimal solution was found when the students do not travel back. In this case, students from year 1 should go to school D, using room 4. Students from year 2 should go to school A, using room 3.

For the second part of this question, we divided the total costs (classroom costs and teaching costs) among the four schools based on the number of students. We wanted the schools with the most students to pay the most. This results in school C and D paying the most.
Although we did not use the information about the travel distance in this report, it could be interesting if we would have used this. We did not have further information, so we could not use this table. If we did have information about the travelling costs based on kilometres, we could maybe come to different results.

In conclusion, if the main goal of the Ministry of Education (ME) is to positively influence the Math and Dutch grade in various university programs by changing the hours devoted to the high school subjects, we would advice the Ministry of Education to increase the hours devoted to especially Mathematics B since this shows to have the biggest impact on the Math grade in university. We would also advice to increase Dutch out of the four language subjects, since this subject explains the biggest margin of Dutch in university. However, if the ministry has to choose between Dutch and Math B, we strongly advice to increase the hours for Math B, because it's impact on the Math grade in university is significantly bigger.

However, if the ME wants to increase the rate of success of the students in university by changing the devoted hours of subjects in high school, we advice not to change any subject. Because as shown in the second question in the first part, the impact of the high school grades with regards to the ECT's is virtually negligible.

6 Bibliography

References

[1] Jan R. Magnus. *LaTeX: Introduction to the Theory of Econom trics*. VU University Press, Amsterdam, 2017.

7 Appendix/Appendix

In this appendix we present our tables used in the project.

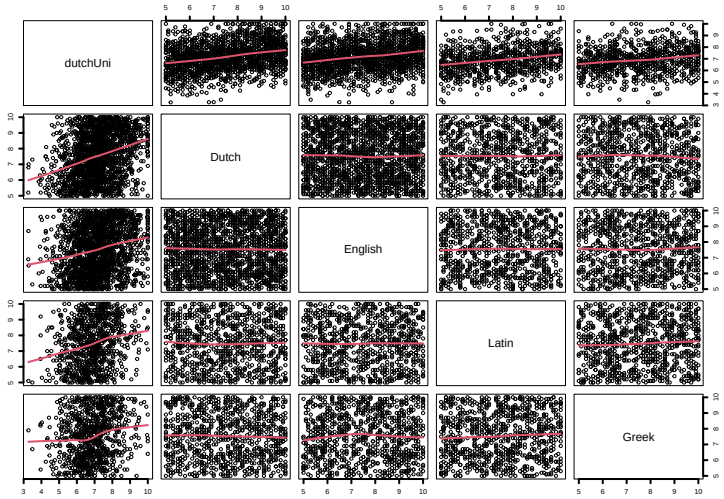

Figure 2: Scatter Plot: Alpha courses vs Dutch University

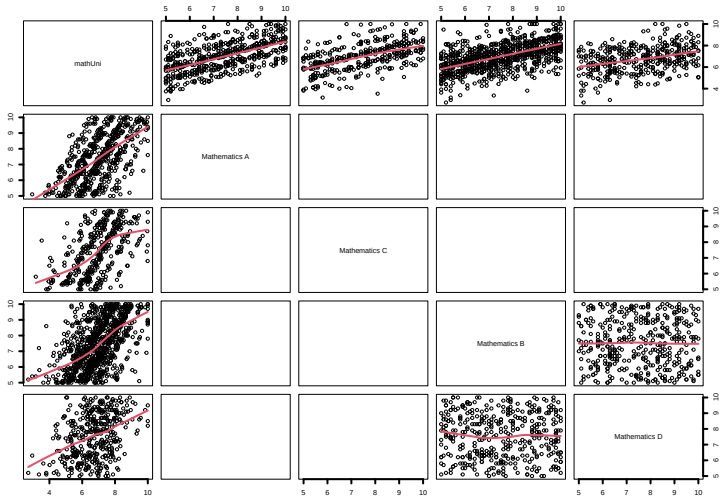

Figure 3: Scatter Plot: Math courses vs Math University

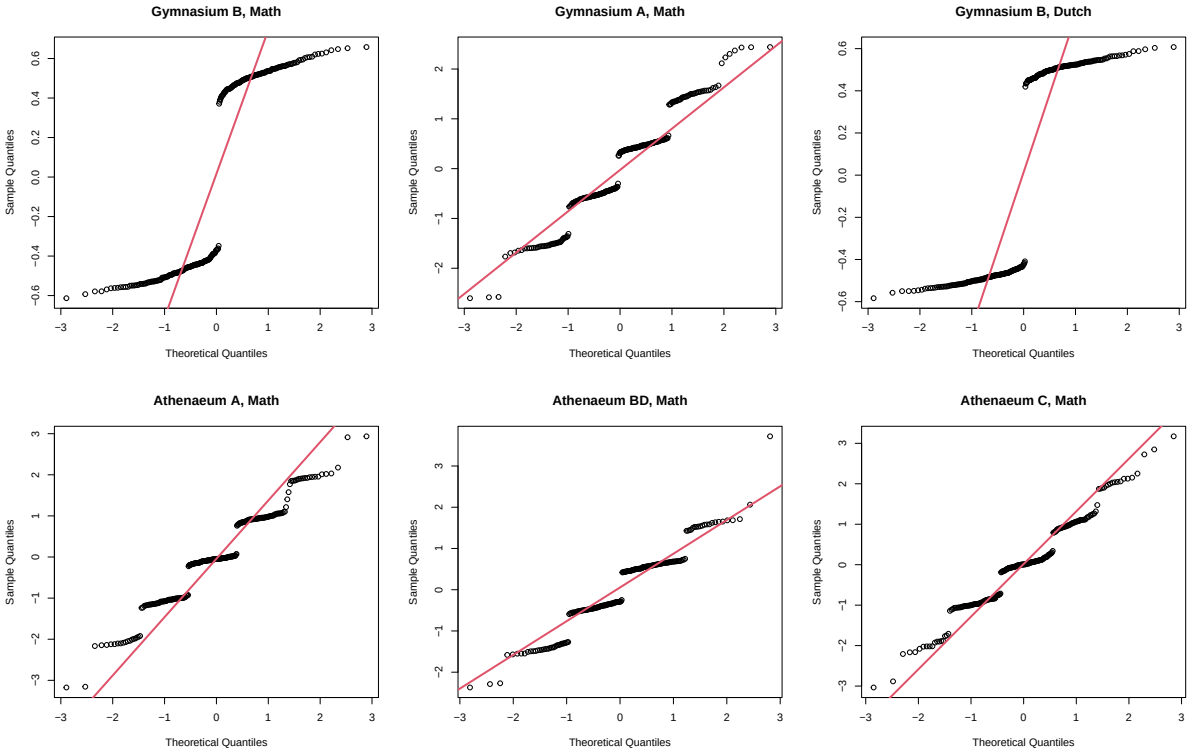

Figure 4: QQ plot with the OLS Residuals for table 4

Table 14: OLS Regression Results Gymnasium

	Dependent variable:							
	MathUni	MathUni	MathUni	MathUni	DutchUni	DutchUni	DutchUni	DutchUni
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Dutch	0.091**	0.041*	0.064	0.068	0.233***	0.268***	0.201***	0.271***
	(0.045)	(0.021)	(0.058)	(0.050)	(0.048)	(0.021)	(0.057)	(0.054)
English	0.073*	0.020	0.128**	0.080*	0.221***	0.161***	0.223***	0.177***
	(0.043)	(0.021)	(0.061)	(0.048)	(0.046)	(0.022)	(0.060)	(0.052)
'Mathematics A'	0.554***				0.034			
	(0.044)				(0.047)			
'Mathematics B'		0.580***		0.216***		0.116***		0.084
		(0.021)		(0.050)		(0.021)		(0.053)
'Mathematics C'			0.466***				−0.057	
			(0.057)				(0.057)	
'Mathematics D'				0.307***				−0.112**
				(0.053)				(0.056)
Latin	0.033	−0.005	−0.005	−0.005	0.171***	0.122***	0.163***	0.232***
	(0.044)	(0.021)	(0.061)	(0.052)	(0.047)	(0.021)	(0.060)	(0.056)
Greek	0.064	0.051**	−0.021	−0.056	0.140***	0.129***	0.137**	0.119**
	(0.045)	(0.021)	(0.060)	(0.047)	(0.048)	(0.022)	(0.059)	(0.051)
Constant	0.834	1.923***	1.984*	2.216**	1.027	0.845**	1.850*	1.113
	(0.721)	(0.362)	(1.006)	(0.882)	(0.767)	(0.364)	(0.997)	(0.945)
Observations	259	262	145	202	259	262	145	202
R^2	0.400	0.761	0.333	0.233	0.226	0.543	0.226	0.271
Adjusted R^2	0.388	0.756	0.309	0.210	0.210	0.534	0.198	0.249

Note: *p<0.1; **p<0.05; ***p<0.01

Table 15: OLS Regression Results Atheneum

	Dependent variable:							
	MathUni	MathUni	MathUni	MathUni	DutchUni	DutchUni	DutchUni	DutchUni
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Dutch	0.052	0.077*	0.096**	0.038	0.387***	0.199***	0.209***	0.158***
	(0.047)	(0.046)	(0.048)	(0.047)	(0.046)	(0.050)	(0.052)	(0.056)
English	0.020	0.071	0.086*	0.044	0.153***	0.160***	0.131**	0.238***
	(0.044)	(0.047)	(0.050)	(0.048)	(0.043)	(0.050)	(0.054)	(0.057)
'Mathematics A'	0.503***				0.030			
	(0.045)				(0.044)			
'Mathematics B'		0.558***		0.440***		0.073		0.031
		(0.048)		(0.045)		(0.051)		(0.054)
'Mathematics C'			0.490***				0.021	
			(0.049)				(0.053)	
'Mathematics D'				0.307***				0.053
				(0.048)				(0.057)
Constant	2.785***	2.047***	2.054***	0.352	3.338***	3.944***	4.851***	3.732***
	(0.573)	(0.592)	(0.650)	(0.723)	(0.563)	(0.637)	(0.703)	(0.869)
Observations	293	262	229	202	293	262	229	202
R^2	0.310	0.361	0.321	0.421	0.228	0.103	0.089	0.111
Adjusted R^2	0.302	0.354	0.312	0.410	0.220	0.092	0.077	0.093

Note: *p<0.1; **p<0.05; ***p<0.01

Table 16: OLS Regression Results Studies

	Dependent variable:	
	DutchUni	MathUni
	(1)	(2)
Antropology	0.190	0.148
	(0.155)	(0.163)
Computer Science	−0.144	0.012
	(0.147)	(0.155)
Chemistry	−0.202	−0.030
	(0.151)	(0.160)
Farmacology	0.042	−0.012
	(0.183)	(0.193)
Physics	0.191	0.141
	(0.154)	(0.163)
Sociology	0.044	−0.141
	(0.143)	(0.151)
Dentistry	0.011	−0.100
	(0.132)	(0.139)
Law	0.085	−0.310
	(0.182)	(0.192)
Economics	−0.177	0.029
	(0.156)	(0.164)
Econometrics	0.199	−0.177
	(0.141)	(0.149)
Mathematics	−0.190	−0.059
	(0.152)	(0.161)
Dutch	0.214	−0.048
	(0.182)	(0.192)
Business Administ.	−0.057	−0.054
	(0.146)	(0.154)
Geology	0.084	−0.258
	(0.153)	(0.161)
Medicine	7.173***	7.009***
	(0.105)	(0.111)
Observations	1,854	1,854
R^2	0.013	0.008
Adjusted R^2	0.006	0.001
F Statistic (df = 14; 1839)	1.737**	1.068
Note:	*p<0.1; **p<0.05; ***p<0.01	

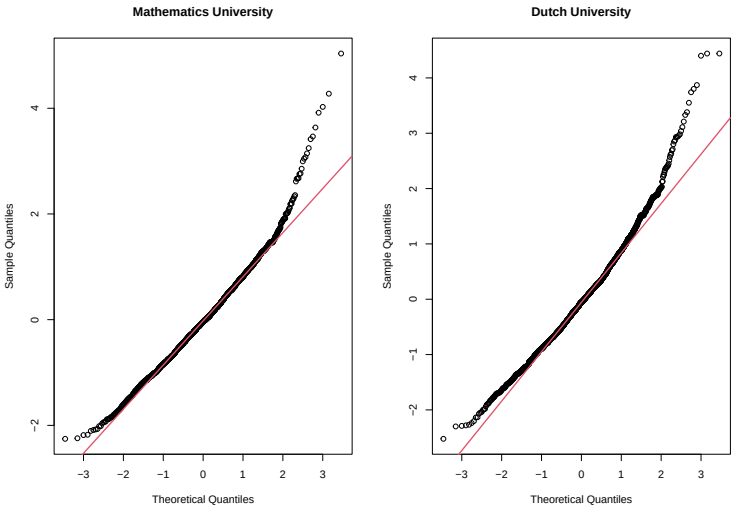

Figure 5: QQ plot: OLS Residuals of table 5 (ECTS)

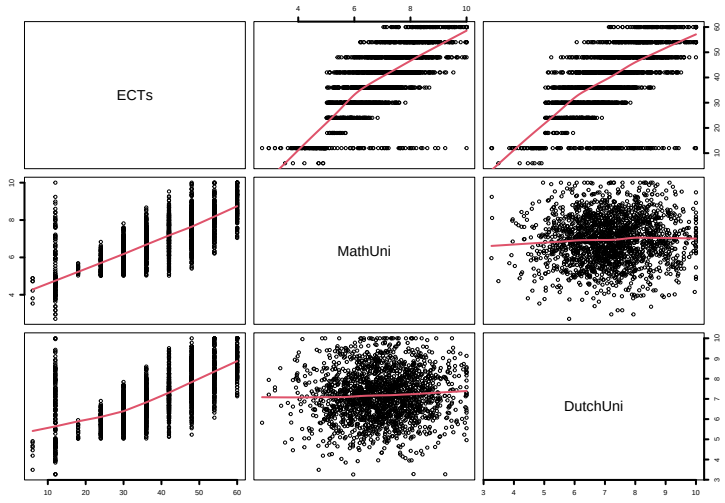

Figure 6: Scatter Plots of ECTs vs University courses