Modelling Factors Affecting Academic Performance in Swedish Schools with Multiple Linear Regression

JOHANNA BREIVOLD
Abstract

This bachelor thesis examines factors affecting the academic performance in Swedish schools. Specifically, the average qualification point among ninth grade students in schools in Stockholm municipality during the academic year 2021-2022 are studied. Multiple linear regression is used to identify individual, social, and school specific factors which have a significant impact on the average qualification point in schools. The purpose is to identify factors affecting the academic performance, and by that contribute to the knowledge base constituting the foundation for the work to improve the academic performance and provide equal opportunities for all students. The Swedish grading system, previous research on factors affecting students’ performance, and the Swedish school in a societal perspective are also discussed. The findings indicate that the background of the students, the parents’ level of education, and the number of students per teacher are good predictors for academic performance.

Keywords

Academic performance, average qualification point, foreign background, educational level, teacher’s degree, students per teacher, staffing ratio, organizer, multiple linear regression
Sammanfattning


Nyckelord

Studieresultat, genomsnittligt meritvärde, utländsk bakgrund, utbildningsnivå, lärarexamen, lärartäthet, huvudman, multipel linjär regression
## Contents

1 Introduction ........................................ 1
   1.1 Background ...................................... 1
   1.2 Purpose .......................................... 1
   1.3 Research questions ............................... 1
   1.4 Limitations ....................................... 2

2 Theory ............................................. 3
   2.1 The Swedish grading system .................... 3
   2.2 Previous research on factors affecting students’ performance .... 3
   2.3 Societal perspective ............................. 5

3 Methodology and results .......................... 6
   3.1 Data collection .................................. 6
   3.2 Initial model ................................... 8
   3.3 Model selection ................................ 11

4 Discussion ........................................ 15
   4.1 Model accuracy ................................ 15
   4.2 Explanatory variables ........................... 15
   4.3 Future work .................................... 16

5 Conclusion ....................................... 17

References ......................................... 18
1 Introduction

1.1 Background

We know that all children are born under unique circumstances and with different opportunities to succeed in life, regardless of what success means to the individual [1]. These differences are reflected in various areas of life, from the ability to attach to other people, to the risk of falling into criminality. The different opportunities we are born with also affect our chances to perform well in school and, later in life, at work.

In this bachelor thesis, we will study how the academic performance in schools are affected by individual factors, such as the students’ ethnic background and parents’ level of education, and by school-specific factors, such as proportion of teachers with teachers’ degree and staffing ratio. This will be done by performing multiple linear regression. The goal is to identify factors that affect the academic performance, positively or negatively, to gain knowledge about where and what kind of additional resources are needed.

1.2 Purpose

The objective of the project is to identify factors affecting the academic performance in Swedish schools. Specifically, we will look at data showing the average qualification point of students in grade nine of Swedish compulsory schools, based on different factors, such as parents’ background, parents’ level of education, and number of students per teacher. The purpose is to identify possible ways to improve the overall academic performance in Swedish schools, which is desirable for the students, the schools as well as the society.

1.3 Research questions

The research questions are:

- Which factors affect the academic performance?
- How do these factors affect academic performance?
1.4 Limitations

The research is based on a dataset containing information about compulsory schools only in Stockholm municipality [11, 12]. Since the dataset used covers a relatively large amount of schools, and the schools are relatively widespread, this is not considered a major issue. However, when doing generalizations to the country as a whole or to other countries with similar school systems, this should be kept in mind.
2 Theory

2.1 The Swedish grading system

In Sweden, students are graded from the sixth year of the compulsory school (grundskola) [10]. The grade scale contains six grades, A-F, which are translated to values according to Table 2.1.

Table 2.1: The Swedish grade scale

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>20</td>
<td>17.5</td>
<td>15</td>
<td>12.5</td>
<td>10</td>
<td>0</td>
</tr>
</tbody>
</table>

The average qualification point (meritvärde) is then calculated as the sum of the grade values for the 16 best grades. If a student has received a grade in modern languages as an elective, they can count this as a 17th grade to increase their average qualification point [9]. By modern languages we mean a third language, in addition to Swedish and English [8]. The most common languages to study are French, German, and Spanish, but several other languages can be counted as modern languages as well. The maximum average qualification point is 340 points.

2.2 Previous research on factors affecting students’ performance

Over the years, a significant amount of research has been conducted to identify factors affecting students’ performance in school. The Swedish National Agency for Education has published a report to provide an overview of the research that has been conducted in the field [14]. The report summarizes a broad overview of the factors that research and evaluation have shown to be significant. The factors can be divided into individual factors, social factors, school factors, and teacher and teaching factors.

2.2.1 Individual factors

Research shows that individual factors, such as gender, social background, and ethnicity, have a significant impact on students’ academic performance.
The family’s socio-economic status, the parents’ involvement in their children’s schooling, and their expectations on their children also play an essential role. An educational goal of the Swedish school that has been identified as important is to reduce the impact that these factors have on students’ academic performance, in order for schools to provide equal opportunities for all students to achieve their goals.

2.2.2 Social factors

The Swedish National Agency for Education has also identified a shift towards more segregated schools, as a consequence of a more segregated society. By segregated schools they mean that students from similar backgrounds are attending the same schools. This has lead to more significant differences in academic performance between schools and between groups of students based on social background. The research indicates that the effect of socio-economic background is stronger at school level than at individual level. It also indicates that the more homogeneous the student composition is, the stronger the effect becomes.

2.2.3 School factors

The authors also mention that the Swedish school system has become more decentralized in various ways. One aspect of decentralization is that the responsibility for resource allocation to the schools has been transferred from the government to the municipalities. The aim was to achieve a more efficient distribution of resources and to direct resources to where they are needed the most. There are namely significant differences between and also within municipalities regarding need for school resources. The school costs vary greatly, as do the proportion of qualified teachers and teacher density. However, research does not provide any clear answers to what extent the decentralization has contributed to handling these variations. Overall, research shows that different resources are important for students’ results, but there is reason to distinguish between general effects, i.e., effects for all students, and effects for certain groups of students.
Regarding teacher density and class size, both Swedish and international research shows that the general effects are relatively weak. However, these resources are much more important for younger students, students with weaker academic backgrounds, and students with weak support from home. According to the Swedish National Agency for Education, it is therefore of great importance to relate the municipalities’ principles for resource allocation to the result development in the schools. The authors mean that by allocating resources within municipalities based on their varying conditions, the disparities in results between schools could be decreased.

2.2.4 Teacher and teaching factors

The research also indicates that the teacher plays a major role when it comes to how well they succeed in achieving good results for their students. Subject didactic competence, i.e., the ability to teach in a varied way in a certain subject, is more important than just subject knowledge. Hence, it is important that teachers not only possess knowledge about the subjects that they teach, but also about how to convey the knowledge in various ways.

2.3 Societal perspective

In 2021, the total costs for Swedish municipalities’ were 740 billion SEK [16]. Of this, the educational sector accounted for 320 billion SEK, where the compulsory school, including preschool class (förskoleklass), accounted for 140 billion SEK [15]. This represents 19.5% of the municipalities’ total costs for the year.

According to Sveriges kommuner och regioner (SKR), there is no evidence that more money to schools automatically leads to better academic performance among the students [17]. Instead they argue that it is how the money is being spent that determines the academic performance of the students, and that the schools, the municipalities, and the state together must contribute to create a joint understanding of how the funds should be allocated to optimize the students’ academic performance.
3 Methodology and results

3.1 Data collection

The data used in this research was collected from the Swedish National Agency for Education which has a statistical database containing information about all Swedish schools [13]. The relevant data was analyzed in R using the packages car, GGally, MASS, Metrics, olsrr, and stats [2–7].

3.1.1 The dataset

The dataset contained 143 observations, i.e., information about 143 compulsory schools in Stockholm municipality. Out of these, 66 observations contained missing data, leaving 77 observations for the multiple linear regression. Missing data were observed for the variables Meritvarde, AntalElever, SvBakgrund and EftergymnUtb for schools with less than ten students, where the values were not published to ensure the students’ integrity [13]. The properties of the dataset, after removing observations containing missing data, are presented in Table 3.1. Potential consequences of missing data and removal of these observations are discussed in Section 4.1.

3.1.2 The variables

Below is a presentation of the variables included in the dataset.

Average qualification point (genomsnittligt meritvärde)

The response variable is the students’ average qualification point for the academic year 2021/2022. The variable can take values between 0 and 340.

Number of students (antal elever)

The total number of students at a specific school.

Swedish background (svensk bakgrund)

The variable is continuous and shows the proportion of students who are born in Sweden and have at least one parent who is born in Sweden. Students who do not have a personal identity number, i.e., has unknown background, are not included.
<table>
<thead>
<tr>
<th>Variable</th>
<th>Variable name</th>
<th>Type</th>
<th>Observed values</th>
</tr>
</thead>
<tbody>
<tr>
<td>Average qualification point</td>
<td>Meritvarde</td>
<td>Response variable</td>
<td>126.5 - 316.3</td>
</tr>
<tr>
<td>Number of students</td>
<td>AntalElever</td>
<td>Continuous</td>
<td>10 - 210</td>
</tr>
<tr>
<td>Swedish background (%)</td>
<td>SvBakgrund</td>
<td>Continuous</td>
<td>11.2% - 100%</td>
</tr>
<tr>
<td>Post-secondary education (%)</td>
<td>EftergymnUtbc</td>
<td>Continuous</td>
<td>21.2% - 100%</td>
</tr>
<tr>
<td>Teacher’s degree (%)</td>
<td>Lararexamen</td>
<td>Continuous</td>
<td>45.7% - 100%</td>
</tr>
<tr>
<td>Students per teacher</td>
<td>EleverPerLarare</td>
<td>Continuous</td>
<td>1.7 - 23.5</td>
</tr>
<tr>
<td>Organizer</td>
<td>Huvudman</td>
<td>Discrete</td>
<td>0 = Private 1 = Public</td>
</tr>
</tbody>
</table>

Post-secondary education (*eftergymnasial utbildning*)
This variable is also continuous and shows the proportion of students who have at least one parent that has completed post-secondary education. Students whose parents’ highest level of education is unknown, e.g., if the student’s has unknown background, are not included.

Teacher’s degree (*lärarexamen*)
Similar to the previous variables, this variable is continuous. It shows the proportion of teachers who are qualified, i.e., has a teacher’s degree.

Students per teacher (*elever per lärare*)
The variable is calculated as the total number of students divided by the number of teachers with full time position. Hence, a low value corresponds to a high staffing ratio, whereas a high value corresponds to a low staffing ratio.

Organizer (*huvudman*)
This binary variable shows if the school is private (*fristående*) or public (*kommunal*). It is represented by a dummy variable, where 0 indicates that the
school is private and 1 indicates that the school is public.

3.2 Initial model

Initially, a full model containing all explanatory variables were created. The coefficient estimates and diagnostics for the initial model are presented in Table 3.2 and Table 3.3. The p-value for the entire model was zero up to machine precision, thus the hypothesis that all explanatory variables were equal to zero was rejected. However, some of the variables, namely, AntalElever, Lararexamen, and Huvudman, had individual p-values greater than the significance level. This finding will be analyzed further in Section 3.3.

Table 3.2: Summary, initial model

<table>
<thead>
<tr>
<th>Variable</th>
<th>Estimate</th>
<th>Std. error</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>135.63</td>
<td>20.71</td>
<td>8.24 · 10⁻⁹</td>
</tr>
<tr>
<td>AntalElever</td>
<td>1.64 · 10⁻³</td>
<td>0.05</td>
<td>0.97</td>
</tr>
<tr>
<td>SvBakgrund</td>
<td>-0.59</td>
<td>0.16</td>
<td>4.49 · 10⁻⁴</td>
</tr>
<tr>
<td>EftergymnUtth</td>
<td>2.00</td>
<td>0.22</td>
<td>9.16 · 10⁻⁴</td>
</tr>
<tr>
<td>Lararexamen</td>
<td>-0.25</td>
<td>0.27</td>
<td>0.34</td>
</tr>
<tr>
<td>EleverPerLarare</td>
<td>2.49</td>
<td>0.70</td>
<td>7.27 · 10⁻⁴</td>
</tr>
<tr>
<td>Huvudman</td>
<td>-1.27</td>
<td>5.51</td>
<td>0.82</td>
</tr>
</tbody>
</table>

Table 3.3 also shows that the $R^2$ value of the initial model is 0.728, indicating that the model explains 72.8% of the variation in the data. The adjusted $R^2$ value is 70.5%.

Table 3.3: Goodness of fit, initial model

<table>
<thead>
<tr>
<th>$R^2$</th>
<th>Adjusted $R^2$</th>
<th>Degrees of freedom</th>
<th>Residual std. error</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.728</td>
<td>0.705</td>
<td>70</td>
<td>16.45</td>
</tr>
</tbody>
</table>

3.2.1 Model validation

Linearity assumption

To verify the assumption of a linear relationship between the response variable and the explanatory variables, a scatter plot with the studentized residuals plotted...
against the fitted values were created. The plot is shown in Figure 3.1. It is clear from the figure that the residuals are relatively equally dispersed around the horizontal axis and that no distinct pattern exists, indicating that the studentized residuals approximately have a mean of zero.

![Residuals vs. Fitted Values](image)

**Figure 3.1: Initial model, studentized residuals vs. fitted values**

**Normality assumption**

To check the normality assumption, the normal Q-Q plot, i.e., the standardized residuals plotted against the quantiles, were plotted. The plot is shown in Figure 3.2. Since the residuals nearly forms a straight line with 45-degree angle, the residuals closely follow a normal distribution. There is indeed a small tendency of a heavy-tail, but since it is just a small tendency this is not considered an issue for the model building and evaluation.

**Multicollinearity**

If there is multicollinearity between two or more variables, the variables may become insignificant when they are both present in the model. Hence, the goal is to create a model without multicollinearity. To get an overview of possible multicollinearity that may exist among the variables, a scatter plot matrix with all variables were created. The scatter plot matrix is shown in Figure 3.3.

**Outliers**

To detect potential outliers, a bar plot of Cook’s distance for each observation, i.e.,
the influence of the ith observation if it is removed from the sample, was plotted.
The plot is shown in Figure 3.4. Since all distances are smaller than 1, none of the observations were considered highly influential.

**Variance inflation factors (VIF)**

Since the correlation matrix only gives a visual overview over correlation between the variables, the potential presence of multicollinearity was further examined by calculating the variance inflation factors (VIF). The values for the different variables are presented in Table 3.4. Since all values are smaller than 10, there is no indication of a severe multicollinearity in the model. Hence, no explanatory variables were removed.

<table>
<thead>
<tr>
<th><strong>Explanatory variable</strong></th>
<th><strong>VIF value</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>AntalElever</td>
<td>1.58</td>
</tr>
<tr>
<td>SvBakgruend</td>
<td>3.33</td>
</tr>
<tr>
<td>EftergymnUtb</td>
<td>3.36</td>
</tr>
<tr>
<td>Lararexamen</td>
<td>1.79</td>
</tr>
<tr>
<td>EleverPerLarare</td>
<td>1.35</td>
</tr>
<tr>
<td>Huvudman</td>
<td>1.76</td>
</tr>
</tbody>
</table>
3.3 Model selection

The initial model were then reduced, to get a subset of promising models for further evaluation. As a first step, forward selection and backward elimination were performed. A model based on the individual p-values of the variables were also created.

3.3.1 Forward selection and backward elimination

The forward selection was performed based on the individual p-values of the variables. The p-value threshold was set to default, i.e., 0.3. According to the forward selection, four variables should be selected, namely, *EftergymnUtby*, *SvBakgrund*, *EleverPerLarare*, and *Lararexamen*. The details of the forward selection are presented in Figure 3.5. The figure shows the $R^2$ value, adjusted $R^2$ value, Mallows’s $C_p$, Akaike information criterion (AIC), and root mean square error (RMSE) for the model at each step of the selection.

The backward elimination gave the results that the two variables *AntalElever* and
Figure 3.4: Cook's distance

Figure 3.5: Forward selection, summary

*Huvudman* should be removed. This means that forward selection and backward elimination arrive at the same model. The coefficient estimates and diagnostics for the model are presented in Table 3.5 and Table 3.6.

### 3.3.2 Model based on individual p-values

In Section 3.2, it was found that the variables *AntalElever, Lararexamen*, and *Huvudman* had individual p-values greater than the significance level. Hence, a multiple linear regression model excluding these variables were created. The coefficient estimates and diagnostics for the model are presented in Table 3.7 and Table 3.8.
Table 3.5: Summary, model based on forward selection and backward elimination

<table>
<thead>
<tr>
<th>Variable</th>
<th>Estimate</th>
<th>Std. error</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>137.58</td>
<td>17.83</td>
<td>5.18 · 10^{-11}</td>
</tr>
<tr>
<td>SvBakgrund</td>
<td>-0.58</td>
<td>0.15</td>
<td>2.54 · 10^{-4}</td>
</tr>
<tr>
<td>EftergymnUtbo</td>
<td>1.99</td>
<td>0.21</td>
<td>1.72 · 10^{-14}</td>
</tr>
<tr>
<td>Lararexamen</td>
<td>-0.29</td>
<td>0.21</td>
<td>0.17</td>
</tr>
<tr>
<td>EleverPerLarare</td>
<td>2.51</td>
<td>0.66</td>
<td>3.24 · 10^{-4}</td>
</tr>
</tbody>
</table>

Table 3.6: Goodness of fit, model based on forward selection and backward elimination

<table>
<thead>
<tr>
<th>R²</th>
<th>Adjusted R²</th>
<th>Degrees of freedom</th>
<th>Residual std. error</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.728</td>
<td>0.713</td>
<td>72</td>
<td>16.22</td>
</tr>
</tbody>
</table>

### 3.3.3 Final model

To compare the models, their respective adjusted $R^2$ value and root mean square error (RMSE) were calculated. The models and their corresponding values are presented in Table 3.9. It is clear from the table that Model 1 has the highest adjusted $R^2$ value and the lowest RMSE. Hence, Model 1 was considered the most suitable model for describing the average qualification point in Swedish schools.
Table 3.7: Summary, model based on individual p-values

<table>
<thead>
<tr>
<th>Variable</th>
<th>Estimate</th>
<th>Std. error</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>117.36</td>
<td>10.26</td>
<td>&lt; 2 · 10⁻¹⁶</td>
</tr>
<tr>
<td>SvBakgrund</td>
<td>-0.56</td>
<td>0.15</td>
<td>3.78 · 10⁻⁴</td>
</tr>
<tr>
<td>EftergymnUtb</td>
<td>1.97</td>
<td>0.21</td>
<td>2.63 · 10⁻¹⁴</td>
</tr>
<tr>
<td>EleverPerLarare</td>
<td>2.22</td>
<td>0.63</td>
<td>8.00 · 10⁻⁴</td>
</tr>
</tbody>
</table>

Table 3.8: Goodness of fit, model based on individual p-values

<table>
<thead>
<tr>
<th>R²</th>
<th>Adjusted R²</th>
<th>Degrees of freedom</th>
<th>Residual std. error</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.721</td>
<td>0.709</td>
<td>73</td>
<td>16.32</td>
</tr>
</tbody>
</table>

Table 3.9: Model summary

<table>
<thead>
<tr>
<th>Model</th>
<th>Variables</th>
<th>Adjusted R²</th>
<th>RMSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model 1</td>
<td>SvBakgrund EftergymnUtb Lararexamen EleverPerLarare</td>
<td>0.713</td>
<td>15.69</td>
</tr>
<tr>
<td>Model 2</td>
<td>SvBakgrund EftergymnUtb EleverPerLarare</td>
<td>0.709</td>
<td>15.89</td>
</tr>
</tbody>
</table>
4 Discussion

4.1 Model accuracy

The data consisted of observations of schools in Stockholm municipality during the academic year 2021-2022. The reason behind the choice of data was the time aspects of consolidating data for several municipalities and academic years. Since the number of schools were relatively large and the schools were relatively wide spread, this was not considered a major issue. However, one needs to be careful when doing generalization to the whole country, as mentioned in Section 1.4.

The choice of geographical area and time period resulted in 143 observations, among which 77 where left after removing 66 of the observations due to missing data. Hence, the final model was based on a dataset with relatively few observations. The small amount of observations and the fact that the model only took one academic year into consideration could have a negative impact on the model accuracy.

4.2 Explanatory variables

The final model contained the explanatory variables SvBakgrund, EftergymnUtb, Lararexamen, and ElevenPerLarare. Below is an interpretation of the different factors’ role in the model.

Swedish background (SvBakgrund)

The coefficient estimate is -0.58, which means that the average qualification point decreases when the proportion of students who were born in Sweden and have at least one parent who are born in Sweden increases. The p-value is $2.54 \cdot 10^{-4}$, meaning that the effect is statistically significant. This indicates that the students’ background is an important predictor of the average qualification point.

Post-secondary education (EftergymnUtb)

The coefficient estimate of this variable is 1.99, which indicates that the average qualification point increases when the proportion of students who have at least one parent that has completed post-secondary education increases. This effect is highly statistically significant, with a p-value of $1.72 \cdot 10^{-14}$. Hence, the proportion
of students who has at least one parent that has completed post-secondary education is also an important predictor of the average qualification point.

**Teacher’s degree (Lararexamen)**
The coefficient estimate is -0.29, suggesting that the average qualification point decreases when the proportion of teachers who are qualified, i.e., have a teacher’s degree, increases. However, due to the large p-value of 0.17, the effect is not statistically significant. Hence, we cannot conclude that the proportion of teachers who are qualified has a significant effect on the average qualification point.

**Students per teacher (EleverPerLarare)**
The coefficient estimate of this variable is 2.51, suggesting that the average qualification point in a school increases by 2.51 units for each additional student per teacher. The p-value of this variable is 3.24 · 10⁻⁴, indicating that the effect is statistically significant and that the staffing ratio indeed is an important predictor. However, it is reasonable to believe that a low staffing ratio is tolerable in schools where the students are already performing well, not that a lower staffing ratio would improve the students’ results.

### 4.3 Future work

This thesis provides insights about factors affecting academic performance in Swedish schools. However, there are areas that can be further explored to create a more extensive knowledge base. Two major areas for improvement are presented below.

Firstly, this research is based on a relatively small and homogeneous dataset, more specifically data for compulsory schools in Stockholm municipality. To be able to reliably generalize the results, data for all Swedish compulsory schools could be used. This data can be retrieved from the data source of this research [17].

Secondly, other factors that may affect the academic performance could also be included in the dataset. Even if this thesis arrived at a certain set of factors affecting academic performance, this does not mean that other factors does not affect the academic performance. By including other factors in the dataset, other relationships could be found and further explored upon.
5 Conclusion

The aim of this thesis was to determine which factors affect the academic performance in Swedish schools and how these factors affect the academic performance. The answers to the targeted research questions are presented below.

Which factors affect the academic performance?
The results indicate that three of the variables described in Section 3.1.2 have a statistically significant impact on the average qualification point of schools. The variables are listed below.

- Swedish background of the students
- Post-secondary education of the students’ parents
- Number of students per teacher

How do these factors affect academic performance?
The impact of each statistically significant factor is presented below.

- Schools with a higher proportion of students with Swedish background tend to have a lower average qualification point.
- Schools with a higher proportion of students who has at least one parent that has completed post-secondary education tend to have a higher average qualification point.
- Schools with a higher students per teacher quotient tend to have a higher average qualification point.
References


vagledningsvarpavanligafragor / samycketkostarskolan . 2785 . html.