Kindergarten, Parents' Education and Reading Literacy Achievement

a multiple regression model

Kaja Horvat

Bachelor thesis

Supervisor: Anna Lindgren

Faculty of Science
Centre for Mathematical Sciences
Mathematical Statistics

Sammanfattning

Skolelevers läsförmåga är ett komplext fenomen som har många bakomliggande faktorer, men kan möjligtvis förutsägas av en mindre modell. Påverkas läsförmågan vid slutet av årskurs fyra av dagisnärvarande och föräldrars utbildningsnivå? Kan läsförmåga förutsägas av kombinationen av dagisnärvarande och föräldrars utbildningsnivå?

Resultat från IEA PIRLS (Progress in International Reading Literacy Survey) användes vid besvarande av forskningsfrågeställningarna. Barns läsförmågepoäng, bedömt med rimligvärdesmetodologi, förutspåddes med multipla regressionsmodeller baserade på dagisnärvarande och föräldrars utbildningsnivå. De bästa modellerna valdes baserat på deras respektive residualstandardfel, R^2-koefficient, justerad R^2-koefficient och AIC- samt BIC-värden, och fastställs med 10-delad korsvalidering.

Den viktigaste faktorn för ett barns läsförmåga är moderns utbildningsnivå. Den näst viktigaste faktorn är faderns utbildningsnivå. Faktorn som har minst betydelse är barnets dagisnärvarande, med en mycket liten påverkan på läspoängen, och är således försumbar vid förutsägelse av resultat i praktiken.

Nyckelord: statistisk modell, multipel regressionsmodell, läsförmåga, dagis, föräldrautbildning.

Abstract

Student achievement in reading is a complex phenomenon that depends on many factors, but perhaps it can be predicted with a smaller model. Does attending kindergarten as well as parental education affect reading achievement at the end of the fourth grade? Can reading achievement be predicted by the combination of attending kindergarten and parental education?

Results from the IEA PIRLS (Progress in International Reading Literacy Survey) were used for answering the research questions. Children's reading achievement scores, represented with the Plausible Values methodology, were predicted with multiple regression models using variables about their kindergarten attendance and parents' education. The best models were chosen based on their respective residual standard error, R^2 coefficient, adjusted R^2 coefficient, and AIC and BIC values, and assessed using 10-fold cross validation.

The most important factor in a child's reading achievement is their mother's education. The second most important factor is their father's education, and the least important factor is kindergarten attendance, which has a very small influence on the reading score and is not important for predicting results in practice.

Keywords: statistical model, multiple regression model, reading literacy achievement, kindergarten, parental education.

Table of contents

List of Tables

List of Figures

1. Introduction

Student achievement in reading depends on many factors. Researchers are trying to estimate the importance of specific conditions in students' lives that may contribute to their success in gaining knowledge. Among the most important factors for reading are socio-economic (SES) factors, as proven by many researchers (White 1982, Coleman 1966, Sirin 2005). Socio-economic factors may include: parental education, number of books at home (as a predictor of wealth and lifestyle), urban/rural environment, attending kindergarten, music schools, sports, parental attitude towards reading, teachers' education etc. It is not easy to determine whether these factors are related among themselves or they may be distinguished.

1.1 Background

In Slovenia, more than three quarters of the children aged 1-5 attend kindergartens (Republic of Slovenia, Statistical Office 2014). The system has been well-established for more than 50 years and enables parents (especially mothers) to fully engage in their professional life as children can attend kindergartens for up to 9 hours per day, 5 days per week. The payment depends on a family income and lets the least wealthy families take advantage of kindergartens. Kindergarten teachers are expected to have a tertiary level diploma (Act on kindergartens 2005), thus we expect them to be adequately qualified to educate children, also because preschool curriculum exists and kindergartens are required to follow it (Act on Kindergartens 2005).

"Typically, the correlation between SES and student achievement is about .30 at the individual student level" (Sirin 2005 and White 1982 in Gustafsson et al. 2013, 183). However, socio-economic factors are a complex and multidimensional concept. In most countries, parents' formal education level has been identified as a key component of cultural capital, which is a term that is used to label the most important dimensions of socio-economic factors (Gustafsson et al. 2013, 183). The relationship between parents' education and reading skills and academic achievements of the child is in general attributed to "parents' beliefs, values, expectations, attitudes and behaviors: well educated parents appear to have high expectations of their children, while at the same time adapting their expectations to the performance of their children. In contrast, parents with little education tend to have lower, or sometimes unrealistically high, expectations of their children" (Gustafsson et al. 2013, 186).

1.2 Aim

In this thesis we want to explore whether children who attended kindergartens before school have better reading achievements at the end of the fourth grade of elementary school, whether the achievement depends on years spent in kindergarten, and if the achievement depends on parental education. We are interested if children from specific SES groups (as defined by parental education) get more benefits in reading than others.

1.3 Research questions

1. Does attending kindergarten as well as parental education affect reading achievement at the end of the fourth grade?
2. Can reading achievement be predicted by the combination of attending kindergarten and parental education?

2. Theory

Regression analysis is the analysis of relationships among variables, which is expressed in the form of an equation

$$y = b_0 + b_1 x + b_2 x + \cdots b_p x_p$$

where $x_1, x_2, \ldots, x_p$ are independent variables, y is the dependent or response variable, and $b_1, b_2, \ldots, b_n$ are regression coefficients which are determined from the data. When an equation contains more than one independent variable, it is called a multiple regression equation (Chatterjee and Price 1977, 1). "The task of regression analysis is to learn as much as possible about the environment represented by the data" (Chatterjee and Price 1977, 2).

2.1 Generalized Linear Models

Generalized linear models have three components: random, which identifies the response variable Y and assumes its probability distribution; systematic, which specifies the explanatory variables; and the link, which describes the functional relationship between the systematic component and the expected value of the random component (Agresti 1996, 72). "The GLM relates a function of that mean to the explanatory variables through a prediction equation having linear form" (Agresti 1996, 72).

The link function is a function

$$g(\mu) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$$

It specifies how $\mu = E(Y)$ relates to the explanatory variables. The simplest possible link function has the form $g(\mu) = \mu$. It directly models the mean and is called the identity link. It specifies a linear model for the mean response

$$\mu = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p ,$$

which is an ordinary regression model for continuous responses (Agresti 1996, 73).

2.3 Multiple Regression Models

Data in a multiple regression model "consists of *n* observations on a dependent or response variable *y* and *p* independent (explanatory) variables $x_1, x_2, \ldots, x_p$" (Chatterjee and Price 1977, 51). The relationship between the independent and dependent variables is formulated as a linear model

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_p x_{pi} + u_i,$$

where β_j are constants - model partial regression coefficients and u_i are random disturbances, where $i = 1, \dots, n$ are indices of individual observations and $j = 1, \dots, p$ are indices of explanatory variables.

We assume that for any set of fixed values of $x_1, x_2, \dots, x_p$ that are within the range of the data, the linear model provides an acceptable approximation of the true relationship between dependent and independent variables. u_i measures discrepancy in the approximation for the *i*th observation and contains no systematic information for determination of y that is not already included in the x's. We assume that u's are random, independently distributed, and have a zero mean and constant variance σ^2. The regression coefficients β_j are the increment in y that corresponds to a unit increase in x_j when all other variables are kept constant. The coefficients are estimated by the method of least squares, which minimizes the sum of squared residuals (Chatterjee and Price 1977, 51-52).

With estimated regression coefficients $\hat{b}_i$ we define a predicted value

$$\hat{y}_i = \hat{b}_0 + \hat{b_1} x_{1i} + \hat{b}_2 x_{2i} + \cdots \hat{b}_p x_{pi}$$

and observe residuals

$$e_i = y_i - \hat{y}_i,$$

which are then used for evaluating model specifications (Chatterjee and Price 1977, 53). The *i*th standard residual is defined as

$$e_{is} = \frac{e_i}{s},$$

where s is the standard deviation of all the residuals. These residuals should have a zero mean and a unit standard deviation and should be distributed approximately as independent, normal deviates. They are not strictly independently distributed, but when we have a large number of observations, the lack of independence may be ignored. Studying plots of the residuals is one of the main tools in regression analysis, and examining plots of residuals can reveal various different model violations (Chatterjee and Price 1977, 9). "In general, when the model is correct, the standardized residuals tend to fall between 2 and -2 and are randomly distributed about zero. The residual plots should show no distinct pattern of variation" (Chatterjee and Price 1977, 9-10). "After fitting a linear model one should examine the residuals for any evidence of heteroscedasticity" (Chatterjee and Price 1977, 50), which is revealed if the residuals tend to decrease or increase with the values of x_j. If heteroscedasticity is present, we

should take it into account when fitting the model, otherwise the resulting least square estimates will not have the maximum precision and therefore smallest variances. We can remove heteroscedasticity by working with transformed variables (Chatterjee and Price 1977, 50).

2.4 Multicollinearity

It might be impossible to change one variable while holding all the others constant, which means that there exists a linear relationship among the explanatory variables - they are not orthogonal (Chatterjee and Price 1977, 143). Independent variables will usually not be orthogonal, and that we expect nonorthogonality with observational data. Because of nonorthogonality, least squares results for each independent variable are dependent on which other variables we have in the model (Rawlings et al. 2001, 210).

Nonorthogonality means that two or more variables are highly correlated. "Multicollinearity is associated with unstable estimated regression coefficients. This situation results from the presence of strong linear relationships among the explanatory variables. It is not a problem of misspecification" (Chatterjee and Price 1977, 155). Indicators of multicollinearity are large changes in estimated coefficients when we add or delete a variable, large changes in coefficients when we alter or drop a data point, algebraic signs of the estimated coefficients that are different than expected, and coefficients of variables that we expect to be important have large standard errors (Chatterjee and Price 1977, 155-156).

3. Description of the Method and Data

This chapter describes data and methods used in the analysis.

3.1 Data

Data is taken from IEA PIRLS 2011 (PIRLS stands for Progress in International Reading Literacy Survey) International Database (PIRLS 2011 International Database). We have analyzed the Slovenian data, from which we have taken 10 variables which were included in our test models. There are 4512 students in the database. After removing the students missing one or more of the answers to kindergarten attendance or education of mother or father, there are 4088 students left in the sample. The sample of students is representative for the 4th grade students in Slovenia (Foy and Joncas 2012).

3.1.2 Variables

In the following section the variables used in the thesis will be described.

3.1.2.1 Reading achievement (5 variables)

The reading achievement is derived according to the Plausible Values methodology. "Plausible values are multiple imputations of the unobservable latent achievement for each student. /…/ One way to describe plausible values is to say that plausible values represent the range of abilities that a student might reasonably have, given the student's item responses" (Wu 2005, 114-115). With the Plausible Values methodology, a probability distribution for a student's θ, the student ability parameter, is estimated, instead of directly estimating θ, which means that instead of a point estimate of θ, "a range of possible values for a student's θ with an associated probability for each of these values; is estimated. Plausible values are random draws from this (estimated) distribution for a student's θ" (Wu 2005, 116).

Each student has 5 reading achievements (5 plausible values - 5 PVs). "Typically, five plausible values are generated for each student, although there does not seem to be strong support in the literature for five" (Wu 2005, 116). They are not intended to estimate individual student scores but are "imputed scores for like students—students with similar response patterns and background characteristics in the sampled population— that may be used to estimate population characteristics correctly" (Martin and Mullis 2012, 6). They are used in PIRLS to ensure the accuracy of estimates for a population as a whole and an "advantage of

this method is that the variation between the five plausible values generated for each student reflects the uncertainty associated with proficiency estimates for individual students" (Martin and Mullis 2012, 8). Plausible values can be used in two ways: we can only use the first vector of plausible values to estimate the result, or we can use all the five vectors and estimate the result as the average of what we got for the five plausible values (Martin and Mullis 2012, 8).

The scale centerpoint for PV is 500 and is set to correspond to the mean of the overall achievement distribution. 100 points are set to correspond to the standard deviation (Mullis et al. 2012, 36). Mean achievement for Slovenia is 530 points (Mullis et al. 2012, 38). To understand what the points from the PVs mean, Cliffordson and Gustafsson have shown that 40 points is the difference a year makes - pupils, who are a year older and have been in school for one more year, get a reading achievement score that is 40 points higher. They also calculated that two thirds of this difference of 40 points is due to school, and the remaining third is due to the children being chronologically older (Cliffordson and Gustafsson 2008).

3.1.2.2 Kindergarten attendance and parental education (5 variables)

Background data in PIRLS survey is provided with 4 different questionnaires: Student Questionnaire, Home Questionnaire, Teacher Questionnaire and School Questionnaire (PIRLS 2011 Contextual Questionnaires). Data about attending kindergartens and parents' education are based on Home Questionnaires (these are questionnaires for students' caregivers, one student takes home one Home Questionnaire).

Question 17 from the Home Questionnaire was: "What is the highest level of education completed by the child's father (or stepfather or male guardian) and mother (or stepmother or female guardian)?" (PIRLS 2011 Home Questionnaire). In the database there are two variables on the education of the child's parents (one for each parent) and the categories in the database for both variables are:

1 = "NO SCH"
2 = "<ISCED 1 OR 2>"
3 = "<ISC 2>"
4 = "<ISC 3>"
5 = "<ISC 4>" (not applicable for Slovenia)

6 = "<ISC 5B>"

7 = "<ISC 5A,1ST DEG>"

8 = "BEYOND <ISC 5A,1ST DEG>"

9 = "NOT APPLICABLE"

99 = "OMITTED OR INVALID" (PIRLS 2011 International Database)

In Slovenia, ISCED 1 or 2 corresponds to elementary school (1 is finished 6th grade and 2 is finished 9th grade, i.e. completed elementary education), ISCED 3 corresponds to high school or gymnasium, ISCED 4 is not applicable for Slovenia, ISCED 5A corresponds to a university Bachelor's degree, 5B to tertiary education that is not university, and beyond 5A corresponds to a Master's degree or a PhD (Classification of Categories of the Slovenian Education System to ISCED 1997 Categories 2012).

The variable containing information about the highest education of either parent was calculated from the variables about mother and father's educations and it contains the following categories:

1 = "UNIVERSITY OR HIGHER"

2 = "POST-SECONDARY BUT NOT UNIVERSITY"

3 = "UPPER SECONDARY"

4 = "LOWER SECONDARY"

5 = "SOME PRIMARY,LOWER SECONDARY OR NO SCHOOL"

6 = "NOT APPLICABLE" (PIRLS 2011 International Database)

Home questionnaire also contains a question on whether a child attended kindergarten. Question 4A was written as: "Did your child attend kindergarten" (PIRLS 2011 Home Questionnaire)? In the database the answer contains the following categories:

1 - yes

2 - no

9 - omitted or invalid (PIRLS 2011 International Database)

There was an additional question about years spent in kindergarten for children who attended kindergartens. Question 4B was written as: "How long did a child attend kindergarten before

school" (PIRLS 2011 Home Questionnaire)? The answer in the database contains the following categories:

1 = "3 YEARS OR MORE"
2 = "BETWEEN 2 AND 3 YEARS"
3 = "2 YEARS"
4 = "BETWEEN 1 AND 2 YEARS"
5 = "1 YEAR OR LESS"
6 = "LOGICALLY NOT APPLICABLE"
9 = "OMITTED OR INVALID" (PIRLS 2011 International Database).

3.1.3 Descriptive Statistics

Some summary statistics for the reading achievements' PVs are seen in Table 1. There is also a summary of the average of the PVs. We can see that the means are very close together for all PVs, however, their average is different than the Slovenian average, which is 530. This is due to the fact that some students were removed from the analysis because of missing data. Minimum and maximum values between the PVs have bigger differences in between different PVs.

Table 1: Summary Statistics for PVs

Statistic	n	Mean	Standard deviation	Min	Max
PV1	4088	532.101	68.886	299.147	765.886
PV2	4088	531.299	68.558	255.114	741.018
PV3	4088	531.447	69.065	275.556	829.159
PV3	4088	532.538	68.717	241.183	766.970
PV5	4088	531.950	68.212	210.827	734.936
AvgPV	4088	531.867	65.431	255.624	737.715

Qualitative factors in explanatory variables have been transformed into quantitative factors. This could be done because there exists a quantitative ordering of classes (i.e. we can order years in kindergarten and parents' education quantitatively).

There are 3756 children who have attended kindergarten in the sample, and only 332 who have not. We can see this in Figure 1 and Figure 2 below. When we look at the chart for the variable years in kindergarten, we can again see the 332 children who did not attend kindergarten. For the children who did attend, we can see how many attended for how much time: 209 attended kindergarten for 1 year or less, 109 attended for between 1 and 2 years, 328 attended for 2 years, 659 attended for between 2 and 3 years, and 2451, the majority, attended for 3 years or more.

Figure 2: Kindergarten attendance

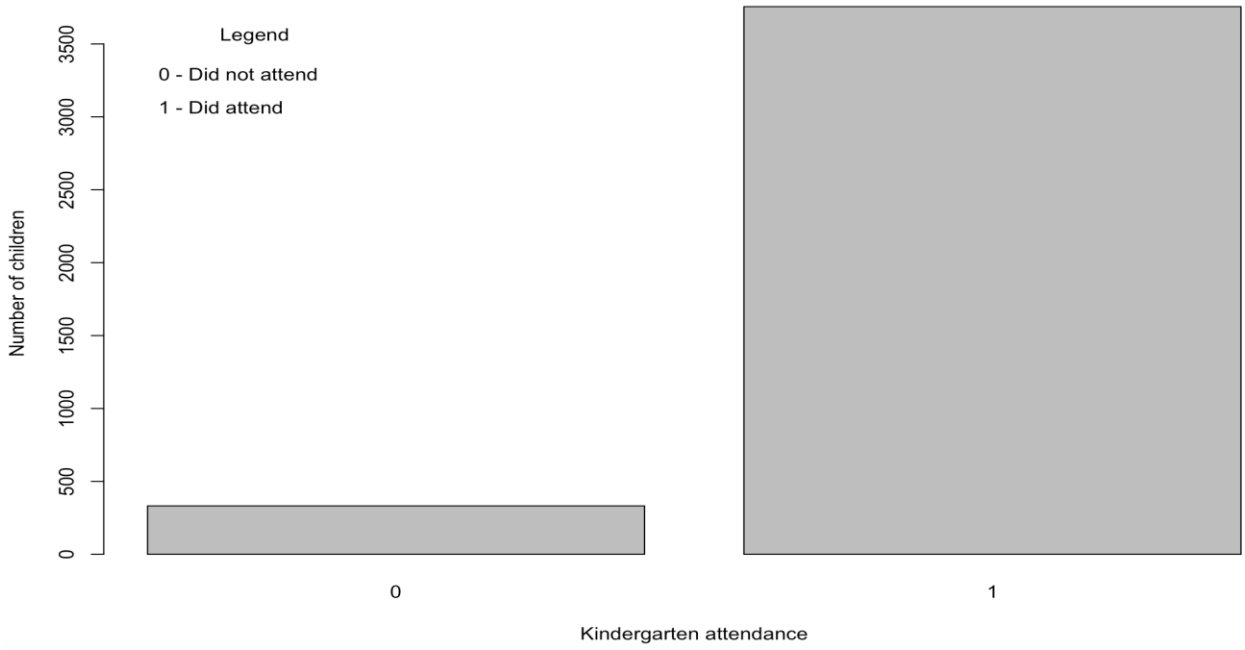
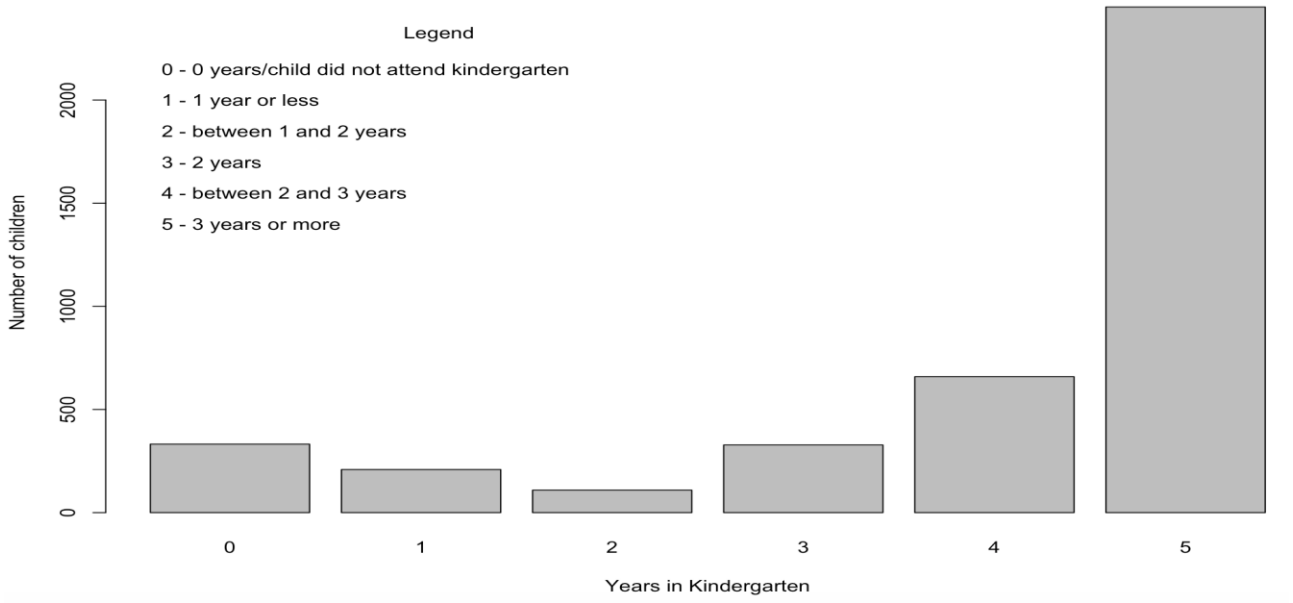


Figure 1: Years in kindergarten



When we look at parents' education in Figure 2 and Figure 3, we can see that most parents have upper secondary education. There are 66 fathers and 22 mothers with some primary, lower secondary or no school, 309 fathers and 298 mothers with lower secondary school, 2594 fathers and 2146 mothers with upper secondary school, 590 fathers and 823 mothers with post-secondary school but not university, and 529 fathers and 799 mothers with university or higher.

Figure 4: Fathers' educations

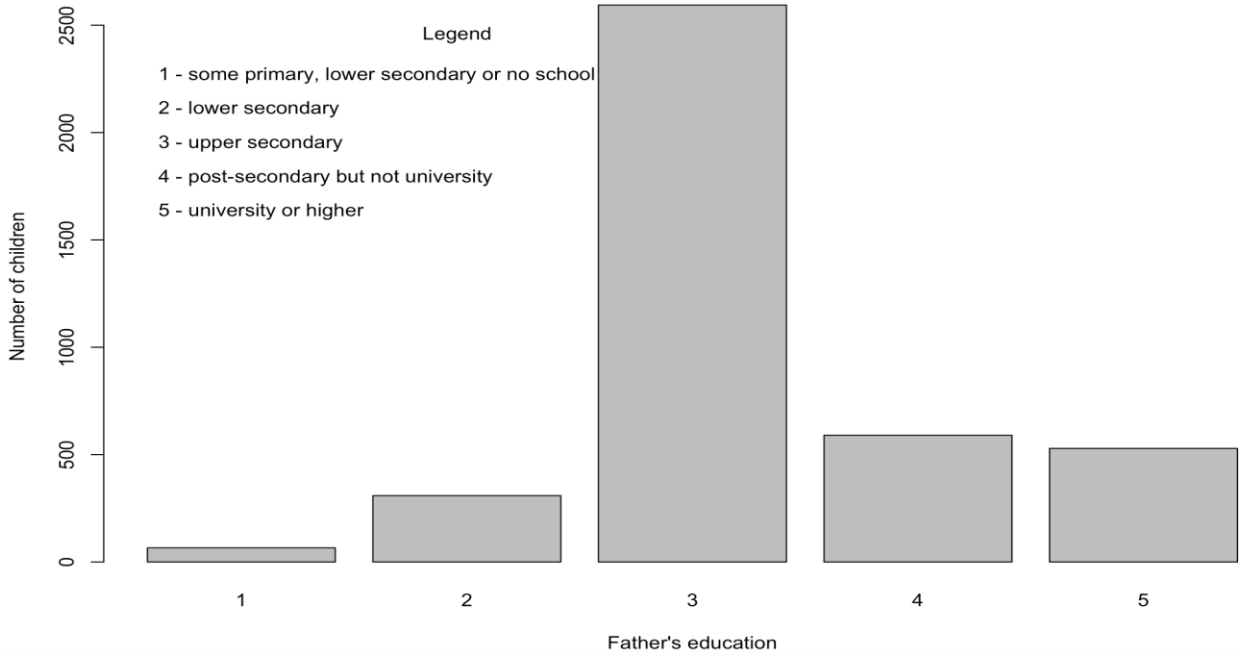
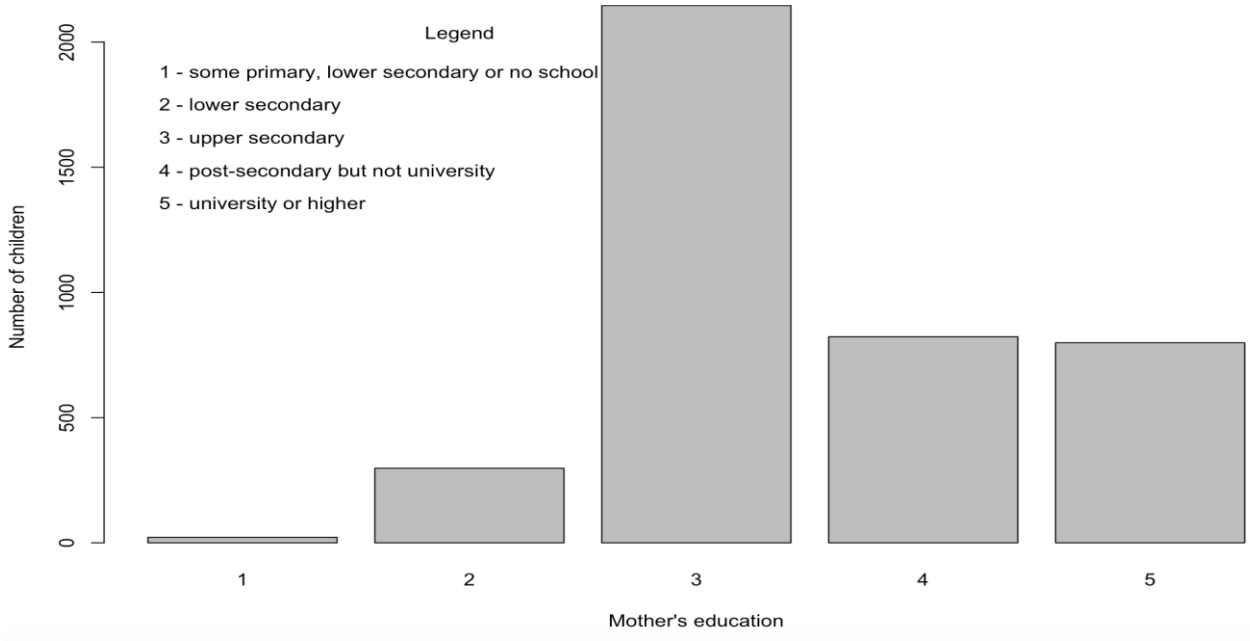
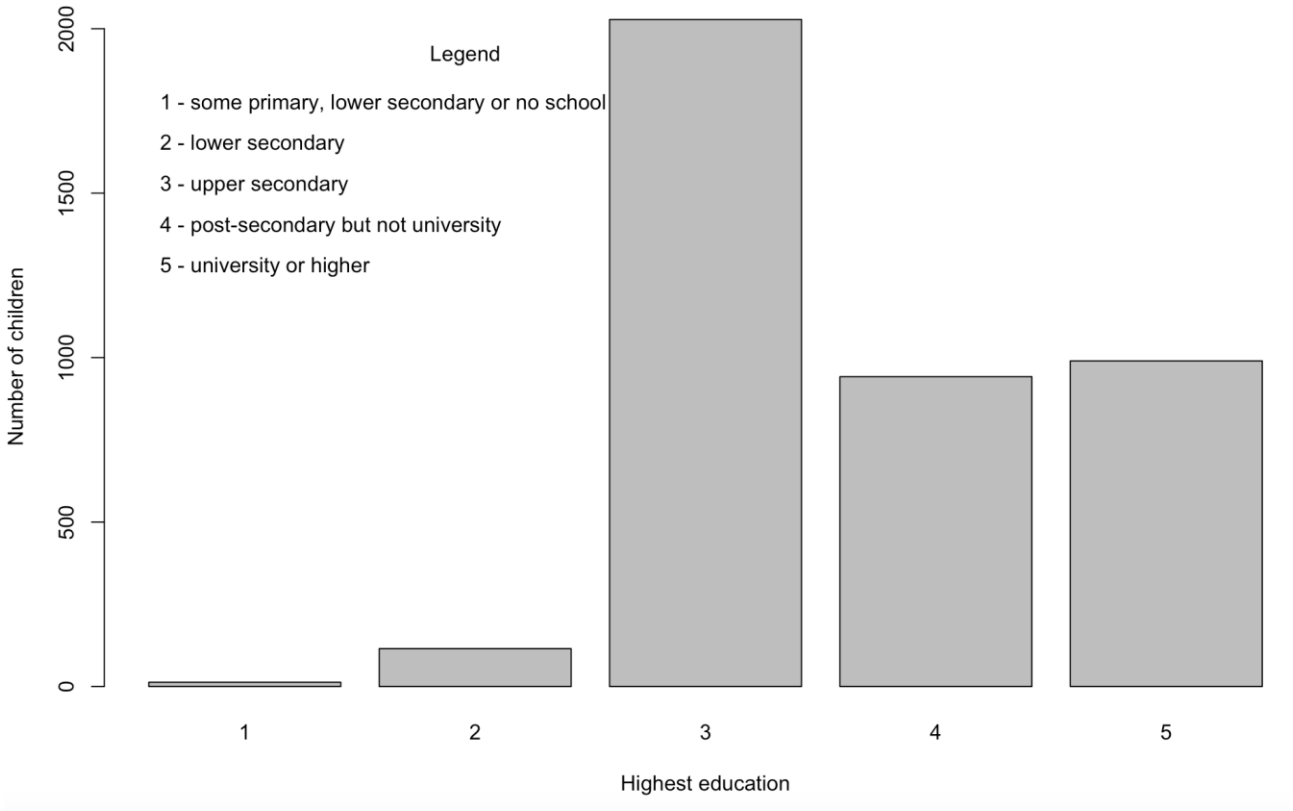


Figure 3: Mothers' educations



Looking at the variable about the highest education of either parent in Figure 5, we learn that 13 children's most educated parent has some primary, lower secondary or no school, 115 children's parents have at most lower secondary school, 2028 have at most upper secondary school, 942 have at most post-secondary school but not university, and 990 have both parents with at least university or higher. Most children's parents' highest education is again upper secondary school.

Figure 5: Highest education of either parent



Tables 2-7 are contingency tables for education and kindergarten. There number of children in each of the group is not similar to the number of children in other groups because neither parents' education nor kindergarten attendance groups have a similar number of children in them.

Table 2: Kindergarten attendance and father's education

Father's Education Kindergarten	some primary, lower secondary or no school	lower secondary	upper secondary	post-secondary but not university	university or higher	Total
Did not attend	10	45	215	32	30	332
Did attend	56	264	2379	558	499	3756
Total	66	309	2594	590	529	4088

Table 3: Kindergarten attendance in years and father's education

Father's Education Kindergarten	some primary, lower secondary or no school	lower secondary	upper secondary	post-secondary but not university	university or higher	Total
0 years/child did not attend kindergarten	10	45	215	32	30	332
less than 1 year	3	32	140	14	20	209
between 1 and 2 years	4	11	79	9	6	109
2 years	5	24	211	49	38	327
between 2 and 3 years	11	36	432	107	73	586
3 years or more	32	161	1517	379	362	2451
Total	65	309	2594	590	529	4088

Table 4: Kindergarten attendance and mother's education

Mother's Education Kindergarten	some primary, lower secondary or no school	lower secondary	upper secondary	post-secondary but not university	university or higher	Total
Did not attend	4	60	189	45	34	332
Did attend	18	238	1957	778	765	3756
Total	22	298	2146	832	799	4088

Table 5: Kindergarten attendance in years and mother's education

Mother's Education Kindergarten	some primary, lower secondary or no school	lower secondary	upper secondary	post-secondary but not university	university or higher	Total
0 years/child did not attend kindergarten	4	60	189	45	34	332
less than 1 year	1	27	128	31	22	209
between 1 and 2 years	3	11	70	13	12	109
2 years	3	22	176	65	62	328
between 2 and 3 years	4	37	353	132	133	659
3 years or more	7	141	1230	537	536	2451
Total	49	141	2146	823	799	4088

Table 6: Kindergarten attendance and highest education

Highest Education Kindergarten	some primary, lower secondary or no school	lower secondary	upper secondary	post-secondary but not university	university or higher	Total
Did not attend	3	29	199	54	47	332
Did attend	10	86	1829	888	943	3756
Total	13	115	2028	942	990	4088

Table 7: Kindergarten attendance in years and highest education

Highest Education Kindergarten	some primary, lower secondary or no school	lower secondary	upper secondary	post-secondary but not university	university or higher	Total
0 years/child did not attend kindergarten	3	29	199	54	47	332
less than 1 year	0	15	129	37	28	209
between 1 and 2 years	3	3	73	15	15	109
2 years	1	6	165	82	74	328
between 2 and 3 years	1	15	318	166	159	659
3 years or more	5	47	1144	588	667	2451
Total	13	142	2028	912	990	4088

As is evident in the correlation table, Table 8, kindergarten and years in kindergarten are, as expected, highly correlated and therefore not orthogonal. They are measuring the same phenomenon but with a different scale, which is why we should only include one of them in our model. The same is true for mother/father's education - one of them is already included in the highest education, which is why we should choose to include either both parents' education or only the highest education in our model. If we would include those highly correlated variables together in our model, there would be a lot of multicollinearity and our regression coefficients could be unstable.

The correlations between the variables for kindergarten and variables for parents' education are low, which means that we can use them together in our model without the risk for unstable coefficients that would come with multicollinearity.

The correlations between the variables for kindergarten and variables for parents' education are low, which means that we can use them together in our model without the risk for unstable coefficients that would come with multicollinearity.

Table 8: Correlation matrix

	Kindergarten attendance	Years in kindergarten	Father's education	Mother's education	Highest education
Kindergarten attendance	1	0.743	0.0752	0.123	0.115
Years in kindergarten	0.743	1	0.117	0.163	0.160
Father's education	0.0752	0.11672615	1	0.499	0.706
Mother's education	0.123	0.163	0.499	1	0.877
Highest education	0.115	0.160	0.706	0.877	1

3.2 Model

Regression equations are used for different purposes and depending on the objective, we have to choose how much emphasis is placed on eliminating variables from the model. The objective is to build a realistic model, and there is desire to identify important variables (Rawlings et al. 2001, 206-208). There exists no best set of variables to be included in a linear model, because a regression equation can be used for different purposes, and the purpose for which it will be used should be kept in mind when choosing the variables to include in the model (Chatterjee and Price 1977, 193). When choosing the model, we will take that into account. We want a model that will be able to predict a child's score based on their kindergarten attendance and parents' education. However, we want our model to have an appropriate number of paramenters. It is also important to control covariates that can influence the relationship, because otherwise the observed effect may simply reflect effects of those covariates on X and Y (Agresti 1996, 53). This is why we will never put kindergarten attendance and kindergarten attendance in years, or father/mother's education and highest education into the same model.

There are different criteria for choice of subset size (Rawlings et al. 2001, 220), i.e. for how many independent variables to use in the model. We will test different subset sizes and different variables to see which model fits best. We have two different goals: model selection, where we choose the best model among different models, and model assessment, where we estimate the prediction error of our final model. If we have enough data, it is best to divide the

dataset into three parts: training (to fit the models), validation (to estimate the error), and test set (to assess the general error in our final model) (Hastie et al. 2009, 222).

3.2.1 Model selection

To build a model, we will try models with different variables as well as a different number of them. Coefficient of determination R^2, as well as AIC and BIC values will help us determine which model to choose.

"Coefficient of determination R^2 is the proportion of the total (corrected) sum of squares of the dependent variable "explained" by the independent variables in the model" (Rawlings et al. 1998, 220) and it is calculated by the following formula:

$$R^2 = \frac{SS(Regr)}{SS(Total)},$$

where

$$SS(Regr) = \sum_{i=1}^{n} (\hat{y}_i - \mu)^2$$

and

$$SS(Total) = \sum_{i=1}^{n} (y_i - \mu)^2.$$

We want to find the model that accounts for as much variation in Y as is practical. The model that explains the most of the variation is the model that contains all the independent variables, and it gives the maximum R^2. The less independent variables the model has, the lower R^2 is. When using the R^2 criterion, we have to judge if the increase in R^2 from additional variables justifies the increased complexity of the model, and we usually choose the biggest model for which the increase in R^2 from the previous model is big - after that size the increase in R^2 with expanding the model should be small (Rawlings et al. 2001).

We can also use the adjusted R^2, which rescales the previous R^2 by degrees of freedom (it involves a ratio of mean squares instead of a sum of squares):

$$R^2_{\text{adj}} = 1 - \frac{(1-R^2)(\text{n}-1)}{(n-p-1)},$$

where p is the number of variables in the model. When R^2 is adjusted, it removes the impact of degrees of freedom and therefore gives a quantity that is more comparable over models involving different numbers of parameters (Rawlings et al. 2001, 222-223).

The Akaike information criterion (AIC) is calculated as

$$AIC = n \ln\left(\frac{SS(Res)}{n}\right) + 2\,(p+1)\,,$$

where

$$SS(Res) = SS(Total) - SS\,(Regr).$$

The first term decreases with p, but the second one increases with p and is a penalty for using a model with more variables. The best model is the one with the lowest AIC. AIC is widely used even though it tends to select models with larger subset sizes than the true model. Because of that, alternative criteria have been developed. One of them is the Schwarz Bayesian criterion, or the Bayesian information criterion (Rawlings et al. 2001), which is given by

$$BIC = n \ln\left(\frac{SS(Res)}{n}\right) + (p+1)\ln(n).$$

It uses the multiplier $ln(n)$ instead of 2 as AIC for the number of parameters k in the model. Therefore it penalizes models with a larger number of parameters more. Again, we want the model with the minimum BIC value (Rawlings et al. 2001, 225). "To use AIC for model selection, we simply choose the model giving smallest AIC over the set of models considered" (Hastie et al. 2009, 230). The Bayesian information criterion is related to AIC, with the factor 2 in AIC replaced by log N in BIC. Compared to AIC, BIC penalizes complex models more heavily (Hastie et al. 2009, 233).

3.2.2 Model assessment

When we have built a model, we should validate its effectiveness for the purpose for which it was intended. This requires assessing the effectiveness of the fitted equation against an independent set of data. We expect that the fitted equation will fit the data from which it was computed better than it will fit any other independent set of data - it will likely fit the sample data even better than the true model would (if it were known). Because it is often impractical to obtain an adequate independent data set for validating a model, we can, if the existing data is sufficiently large, split it and use it for both estimation and validation (Rawlings et al. 2001, 228-230).

The simplest and most widely used method for prediction of error estimation is cross-validation. "K-fold cross-validation uses part of the available data to fit the model, and a different part to test it. We split data into *K* roughly equal-sized parts" (Hastie et al. 2009, 241). "For the *k*th part /…/ we fit the model to the other *K* - 1 parts of the data, and calculate

the prediction error of the fitted model when predicting the *k*th part of the data. We do this for $k = 1, 2, \ldots, K$ and combine the K estimates of prediction error" (Hastie et al. 2009, 242). Typically, we choose K to be 5 or 10. Cross-validation effectively estimates the average error. To do cross-validation correctly, we must retrain the model completely for each fold of the process (Hastie et al. 2009, 242-249). In this thesis we will use 10-fold cross validation.

3.3 Implementation

We use R Studio for modeling the data, and we use different additional packages:

1. The package plyr (Wickham 2016),
2. The package MASS (Ripley et al. 2016),
3. The package cvTools (Alfons 2015).

4. Results

In the following chapter we will test different multiple regression models. We will make different combinations of variables to include in the model, but, as mentioned before, we will not use both variables for kindergarten attendance or mother/father's education and highest education of parents at the same time. We will test different models for every vector of PVs, including the average PVs. We will then describe the best models further, validate them, and in the end we will compare the results.

4.1 Model Selection

We combined the variables into 14 regression models seen below. The abbreviations for the variables as following:

KG = Kindergarten attendance

YKG = Years in kindergarten

FEd = Father's education

MEd = Mother's education

HighestEd = Highest education of both parents

The models are:

1. Y = KG + MEd + FEd
2. Y = KG + HighestEd
3. Y = KG + MEd
4. Y = KG + FEd
5. Y = YKG + MEd + FEd
6. Y = YKG + HighestEd
7. Y = YKG + MEd
8. Y = YKG + FEd
9. Y = KG
10. Y = YKG
11. Y = FEd
12. Y = MEd
13. Y = HighestEd
14. Y = MEd + FEd

4.1.1 Models for PV1

We tested the 14 models with the response variable Y set to the variable PV1. From Table 9 we can read that the model with the least residual standard error is model 5. This is also the model with the highest coefficient of determination R^2 (both adjusted and non-adjusted). The model with the lowest AIC value is, again, model 5. However, model 14 has the lowest BIC value but is closely followed by model 5. This is probably due to the fact that BIC is a criterion that penalizes big models. This means that model 5 is chosen as the best model.

Table 9: Model choice for PV1

		Residual standard error	Degrees of freedom	R^2	Adjusted R^2	AIC	BIC
1	**PV1 = KG + MEd + FEd**	63.91	4084	0.1398	0.1392	33995.94	34021.20
2	**PV1 = KG + HighestEd**	64.73	4085	0.1176	0.1171	34098.29	34117.23
3	**PV1 = KG + MEd**	64.43	4085	0.1256	0.1252	34060.93	34079.87
4	**PV1 = KG + FEd**	66.1	4085	0.07959	0.07914	34270.49	34289.43
5	**PV1 = YKG + MEd + FEd**	**63.88**	4084	**0.1408**	**0.1402**	**33991.16**	34016.43
6	**PV1 = YKG + HighestEd**	64.67	4085	0.1190	0.1186	34091.58	34110.53
7	**PV1 = YKG + MEd**	64.38	4085	0.1269	0.1265	34054.77	34073.72
8	**PV1 YG + FEd**	66.00	4085	0.08246	0.08201	34257.69	34276.64
9	**PV1 = KG**	68.80	4086	0.002784	0.00254	34596.13	34608.76
10	**PV1 = YKG**	68.59	4086	0.008953	0.008711	34570.76	34583.39
11	**PV1 = FEd**	66.13	4086	0.07858	0.07835	34272.96	34285.60
12	**PV1 = MEd**	64.43	4086	0.1255	0.1253	34059.34	34071.97
13	**PV1 = HighestEd**	64.73	4086	0.1174	0.1172	34097.14	34109.77
14	**PV1 = MEd + FEd**	63.91	4085	0.1397	0.1293	33994.20	**34013.15**

4.1.2 Models for PV2

Again, we tested the 14 models. Now we set the responsible variable Y to be PV2. As seen in Table 10, the model with the smallest residual standard error is model 7, followed by model 5. The model with the highest R^2 and adjusted R^2 coefficients is model 5. Model 5 also has the lowest AIC score, and it's followed by model 14. The first two places are switched for BIC score: model 14 has a lower score than model 5, but the differences are very small. Therefore, we choose model 5 as the best model again.

Table 10: Model choice for PV2

		Residual standard error	Degrees of freedom	R^2	Adjusted R^2	AIC	BIC
1	**PV2 = KG + MEd + FEd**	63.75	4084	0.1361	0.1354	33974.64	33999.90
2	**PV2 = KG + HighestEd**	64.47	4085	0.1162	0.1158	34065.33	34084.28
3	**PV2 = KG + MEd**	64.29	4085	0.1209	0.1205	34043.58	34062.53
4	**PV2 = KG + FEd**	65.80	4085	0.07929	0.07883	34232.82	34251.76
5	**PV2 = YKG + MEd + FEd**	63.70	4084	**0.1374**	**0.1368**	**33968.29**	33993.55
6	**PV2 = YKG + HighestEd**	64.40	4085	0.1181	0.1177	34056.76	34075.71
7	**PV2 = YKG + MEd**	**63.23**	4085	0.1227	0.1222	34035.56	34054.51
8	**PV2 = YG + FEd**	65.67	4085	0.08285	0.0824	34216.94	34235.89
9	**PV2 = KG**	68.49	4086	0.002324	0.00208	34558.99	34571.63
10	**PV2 = YKG**	68.24	4086	0.009584	0.009342	34529.14	34541.77
11	**PV2 = FEd**	65.82	4086	0.07855	0.07832	34234.1	34246.73
12	**PV2 = MEd**	64.29	4086	0.1209	0.1207	34041.73	34054.36
13	**PV2 = HighestEd**	64.46	4086	0.1162	0.1159	34063.72	34076.35
14	**PV2 = MEd + FEd**	63.74	4085	0.1360	0.1356	33972.7	**33991.65**

4.1.3 Models for PV3

We test the models for PV3. Looking at table 11 we can see that model 5 is the best model for this response variable by four out of five criteria: it has the lowest residual standard error, the highest R^2 and adjusted R^2 coefficients, and the lowest AIC. However, model 14 has the lowest BIC, slightly lower than model 5, which has the second lowest one. The proposed model by each criterion is the same as the proposed model by that same criterion for PV1. We choose model 5 as the best model for PV3.

Table 11: Model choice for PV3

		Residual standard error	Degrees of freedom	R^2	Adjusted R^2	AIC	BIC
1	**PV3 = KG + MEd + FEd**	64.00	4084	0.1419	0.1413	34006.98	34032.24
2	**PV3 = KG + HighestEd**	64.72	4085	0.1223	0.1218	34097.7	34116.64
3	**PV3 = KG + MEd**	64.62	4085	0.1250	0.1246	34084.84	34103.79
4	**PV3 = KG + FEd**	66.07	4085	0.0852	0.08475	34266.77	34285.72
5	**PV3 = YKG + MEd + FEd**	**63.96**	4084	**0.1430**	**0.1424**	**34002.00**	34027.26
6	**PV3 = YKG + HighestEd**	64.67	4085	0.1237	0.1232	34091.14	34110.09
7	**PV3 = YKG + MEd**	64.57	4085	0.1264	0.1260	34078.35	34097.3
8	**PV3 = YG + FEd**	65.98	4085	0.08788	0.08743	34254.77	34273.72
9	**PV3 = KG**	68.95	4086	0.003629	0.003385	34613.93	34626.56
10	**PV3 = YKG**	68.74	4086	0.009586	0.009343	34589.42	34602.05
11	**PV3 = FEd**	66.12	4086	0.08371	0.08348	34271.42	34284.05
12	**PV3 = MEd**	64.62	4086	0.1247	0.1245	34084.2	34096.83
13	**PV3 = HighestEd**	64.73	4086	0.1218	0.1216	34097.62	34110.26
14	**PV3 = MEd + FEd**	64.00	4085	0.1417	0.1413	34006.04	**34024.99**

4.1.4 Models for PV4

When choosing the model for PV4 (see Table 12), we get the result that model 5 is the best model by four criteria: it has the lowest residual standard error, the highest R^2 and adjusted R^2 coefficients, and the lowest AIC value. Model 1, model 14 and model 7 follow as the next best choices by 3 criteria: both R^2 coefficients, as well as by AIC. Model 14 has the lowest BIC value. It is worth mentioning that the numbers are very close. Because it is best by 4 out of 5 criteria, model 5 is chosen as the best model.

Table 12: Model choice for PV4

		Residual standard error	Degrees of freedom	R^2	Adjusted R^2	AIC	BIC
1	**PV4 = KG + MEd + FEd**	63.89	4084	0.1361	0.1355	33993.36	34018.62
2	**PV4 = KG + HighestEd**	64.67	4085	0.1148	0.1143	34091.04	34109.99
3	**PV4 = KG + MEd**	64.36	4085	0.1231	0.1227	34052.31	34071.26
4	**PV4 = KG + FEd**	66.07	4085	0.07593	0.07547	34266.65	34285.6
5	**PV4 = YKG + MEd + FEd**	**63.86**	4084	**0.1370**	**0.1364**	**33989.12**	34014.38
6	**PV4 = YKG + HighestEd**	64.62	4085	0.1161	0.1157	34084.92	34103.87
7	**PV4 = YKG + MEd**	64.32	4085	0.1243	0.1239	34046.8	34065.75
8	**PV4 = YG + FEd**	65.97	4085	0.07868	0.07823	34254.44	34273.39
9	**PV4 = KG**	68.64	4086	0.002478	0.002234	34577.31	34589.94
10	**PV4 = YKG**	68.44	4086	0.008402	0.008160	34552.96	34565.59
11	**PV4 = FEd**	66.10	4086	0.07507	0.07484	34268.43	34281.07
12	**PV4 = MEd**	64.36	4086	0.1231	0.1229	34050.53	34063.16
13	**PV4 = HighestEd**	64.67	4086	0.1147	0.1144	34089.60	34102.24
14	**PV4 = MEd + FEd**	63.89	4085	0.1361	0.1357	33991.48	**34010.42**

4.1.5 Models for PV5

The proposed models by each criterion for PV5 are the same as for PV1, PV3, and PV4. We can see this in Table 13. It means that model 5 has the lowest residual standard error, the highest R^2 and adjusted R^2 coefficients, the lowest AIC and the second lowest BIC, as model 14's BIC lower. However, the numbers are very close for both AIC and BIC, as well as for the R^2 coefficients (both adjusted and non-adjusted), but since model 5 is a little better by all criteria except one, there is no doubt. We choose model 5 as the best model for PV5.

Table 13: Model choice for PV5

		Residual standard error	Degrees of freedom	R^2	Adjusted R^2	AIC	BIC
1	**PV5 = KG + MEd + FEd**	63.38	4084	0.1374	0.1368	33926.94	33952.21
2	**PV5 = KG + HighestEd**	64.18	4085	0.1152	0.1147	34028.91	34047.85
3	**PV5 = KG + MEd**	63.86	4085	0.1240	0.1235	33988.08	34007.03
4	**PV5 = KG + FEd**	65.55	4085	0.07706	0.07661	34201.32	34220.27
5	**PV5 = YKG + MEd + FEd**	**63.35**	4084	**0.1380**	**0.1374**	**33923.97**	33949.23
6	**PV5 = YKG + HighestEd**	64.14	4085	0.1162	0.1158	34024.22	34043.17
7	**PV5 = YKG + MEd**	63.83	4085	0.1248	0.1244	33983.99	34002.93
8	**PV5 = YG + FEd**	65.47	4085	0.07935	0.0789	34191.16	34210.11
9	**PV5 = KG**	68.14	4086	0.002288	0.002044	34517.78	34530.41
10	**PV5 = YKG**	67.96	4086	0.007549	0.007307	34496.17	34508.8
11	**PV5 = FEd**	65.57	4086	0.07632	0.07610	34202.58	34215.21
12	**PV5 = MEd**	63.85	4086	0.1239	0.1237	33986.19	33998.82
13	**PV5 = HighestEd**	64.17	4086	0.1151	0.1149	34027.28	34039.91
14	**PV5 = MEd + FEd**	63.37	4085	0.1374	0.1370	33924.98	**33943.93**

4.1.6 Models for Average PV

In the end we also tested the models for the average of the PVs (see table 14). The results are very similar to the results for the other PVs. Model 5 has the lowest residual standard error, the highest R^2 coefficients, the lowest AIC and the second lowest BIC (model 14 has lower BIC). The differences are very small, so model 5 is chosen as the best model for average PVs.

Table 14: Model choice for average PV

		Residual standard error	Degrees of freedom	R^2	Adjusted R^2	AIC	BIC
1	**AvgPV = KG + MEd + FEd**	60.26	4084	0.1523	0.1517	33515.32	33540.58
2	**AvgPV = KG + HighestEd**	61.08	4085	0.1291	0.1287	33623.57	33642.52
3	**AvgPV = KG + MEd**	60.82	4085	0.1363	0.1359	33589.68	33608.63
4	**AvgPV = KG + FEd**	62.52	4085	0.08747	0.08703	33814.61	33833.55
5	**AvgPV = YKG + MEd + FEd**	**60.23**	4084	**0.1534**	**0.1528**	**33510.08**	33535.34
6	**AvgPV = YKG + HighestEd**	61.02	4085	0.1307	0.1303	33616.26	33635.21
7	**AvgPV = YKG + MEd**	60.77	4085	0.1378	0.1373	33582.9	33601.85
8	**AvgPV = YG + FEd**	62.41	4085	0.09060	0.09016	33800.57	33819.52
9	**AvgPV = KG**	65.34	4086	0.002955	0.002711	34174.72	34187.35
10	**AvgPV = YKG**	65.12	4086	0.009698	0.009455	34146.98	34159.61
11	**AvgPV = FEd**	62.55	4086	0.08643	0.08620	33817.29	33829.92
12	**AvgPV = MEd**	60.82	4086	0.1362	0.1360	33588.08	33600.71
13	**AvgPV = HighestEd**	61.07	4086	0.1290	0.1287	33622.4	33635.03
14	**AvgPV = MEd + FEd**	60.26	4085	0.1523	0.1518	33513.57	**33532.51**

4.2 Model Assessment

Model 5 was chosen as the best model for all the PVs, including the average. We now have to assess the model, which we will, as mentioned, do using 10-fold cross validation. Since the folds are split differently in every trial, we did 10 trials and then averaged them, to be sure there are no discrepancies. The average prediction errors for model 5 are very similar to residual standard errors, which means that the models have a good predictive performance. The results of the 10-fold cross validation are seen in Table 15.

Table 15: Results of 10-fold cross validation for model 5

	Prediction errors for model 5										
Y	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**Avg**
PV1	63.906	63.926	63.900	63.883	63.915	63.913	63.943	63.913	63.918	63.924	63.914
PV2	63.930	63.915	63.919	63.923	63.906	63.930	63.892	63.894	63.924	63.927	63.916
PV3	63.896	63.927	63.923	63.904	63.878	63.903	63.894	63.909	63.902	63.902	63.904
PV4	63.922	63.891	63.892	63.899	63.921	63.897	63.896	63.903	63.909	63.895	63.903
PV5	63.896	63.887	63.898	63.898	63.936	63.903	63.921	63.898	63.905	63.901	63.904
Avg PV	63.929	63.903	63.904	63.905	63.896	63.905	63.909	63.894	63.921	63.883	63.905

4.3 Final Models

In the following section we will describe and interpret the final models. For all PVs, Model 5 was chosen as the best, which makes sense since this is the model that has the most information about both kindergarten attendance and parents' educations included in the variables. Model 5 has the form:

$$Y = YKG + MEd + FEd$$

We will now look at the models, interpret and compare them.

4.3.1 Coefficients

In Table 16 - Table 21 we can see the estimates of coefficients, their significance levels and confidence intervals. They are similar for all our models, even though when looking at minimum and maximum values of PVs in each of the groups in Table 1 one could think that

the data in each vector of PVs is quite different from the data in other vectors. However, the means are very similar, and therefore it makes sense that our coefficients are also similar, since the link function we use in our models is $g(\mu) = \mu$.

Table 16: Coefficients for model with Y=PV1

	Constant	YKG	MEd	FEd
Estimate	414.854	1.426	21.377	11.095
significance level	<0.001	0.01	<0.001	<0.001
95% confidence interval	[405.179, 424.529]	[0.180, 2.671]	[18.860, 23.894]	[8.419, 13.771]

Table 17: Coefficients for model with Y=PV2

	Constant	YKG	MEd	FEd
Estimate	415.249	1.604	20.573	11.372
significance level	<0.001	0.05	<0.001	<0.001
95% confidence interval	[405.601, 424.897]	[0.362, 2.846]	[18.063, 23.083]	[8.703, 14.040]

Table 18: Coefficients for model with Y=PV3

	Constant	YKG	MEd	FEd
Estimate	412.084	1.563	20.832	12.150
significance level	<0.001	0.05	<0.001	<0.001
95% confidence interval	[402.395, 421.772]	[0.317, 2.810]	[18.312, 23.352]	[9.471, 14.830]

Table 19: Coefficients for model with Y=PV4

	Constant	YKG	MEd	FEd
Estimate	417.59	1.33	21.32	10.58
significance level	<0.001	0.05	<0.001	<0.001
95% confidence interval	[407.9211, 427.267]	[0.0805, 2.570]	[18.806, 23.838]	[7.900, 13.251]

Table 20: Coefficients for model with Y=PV5

	Constant	YKG	MEd	FEd
Estimate	417.858	1.093	21.229	10.697
significance level	<0.001	0.05	<0.001	<0.001
95% confidence interval	[408.262, 427.453]	[-0.142, 2.328]	[18.732, 23.725]	[8.043, 13.351]

Table 21: Coefficients for model with Y=AvgPV

	Constant	YKG	MEd	FEd
Estimate	415.528	1.402	21.067	11.178
significance level	<0.001	0.05	<0.001	<0.001
95% confidence interval	[406.405, 424.650]	[0.228, 2.576]	[18.693, 23.440]	[8.655, 13.701]

4.3.2 Plots of residuals

We plot the residuals from our models to confirm that they do not follow some sort of a pattern. As mentioned before, studying residuals is a very important tool in regression analysis, because it can reveal model violation. With the Figure 6 – Figure 35 we can confirm that our models are correct, since they show no distinct pattern of variation (this can be seen in plots of residuals), the standardized residuals tend to be between -2 and 2 and are normally distributed (this can be seen in the normal Q-Q plot with standardized residuals), and there is no heteroscedasticity, since the residuals do not tend to increase or decrease with the values of the *x*'s (this can be seen in the plots where residuals are plotted against the *x*'s).

Figure 6: Residuals for PV1

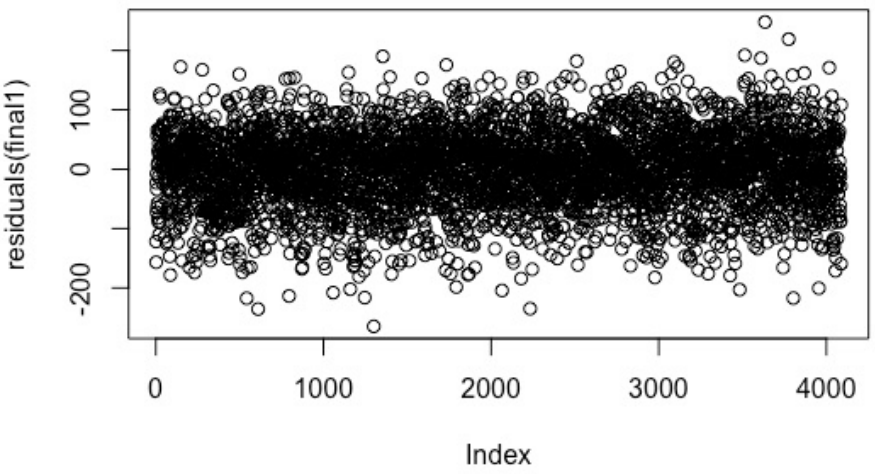


Figure 7: QQ plot with standardized residuals for PV1

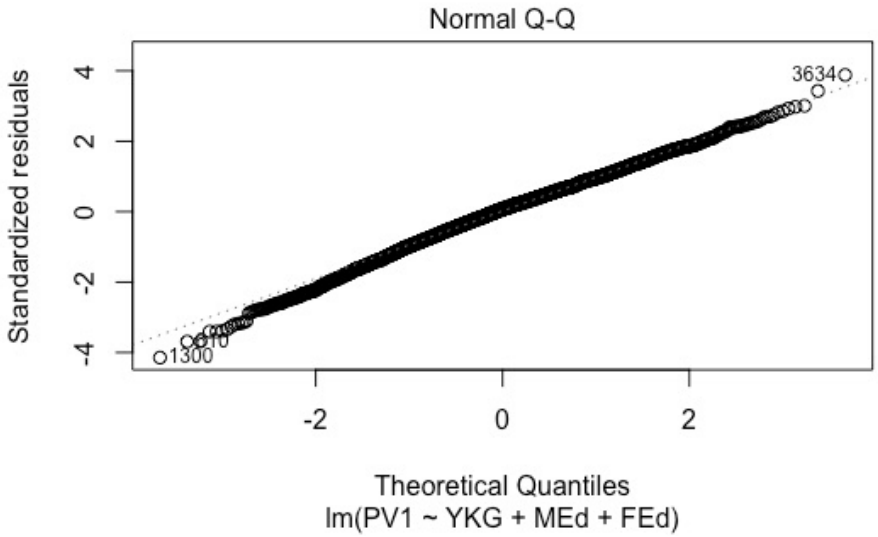


Figure 8: Residuals for PV1 against YKG

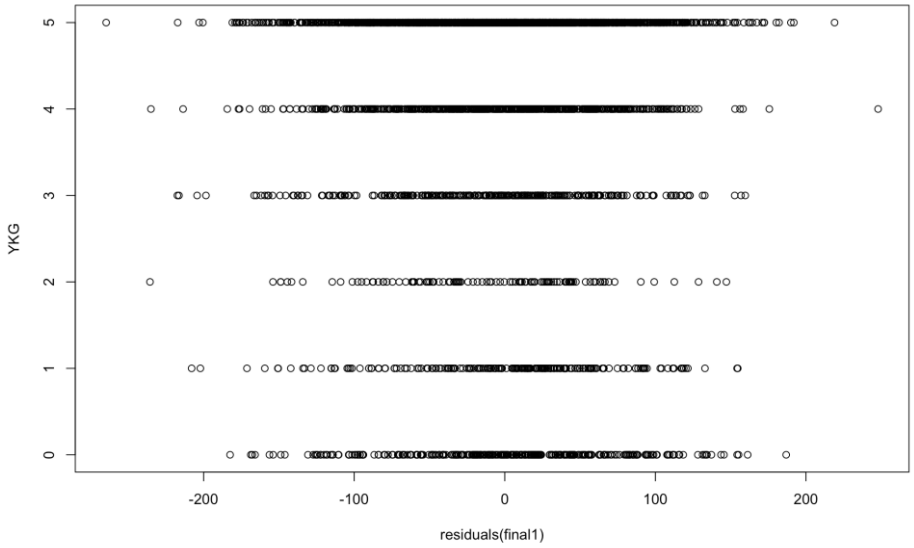


Figure 9: Residuals for PV1 against MEd

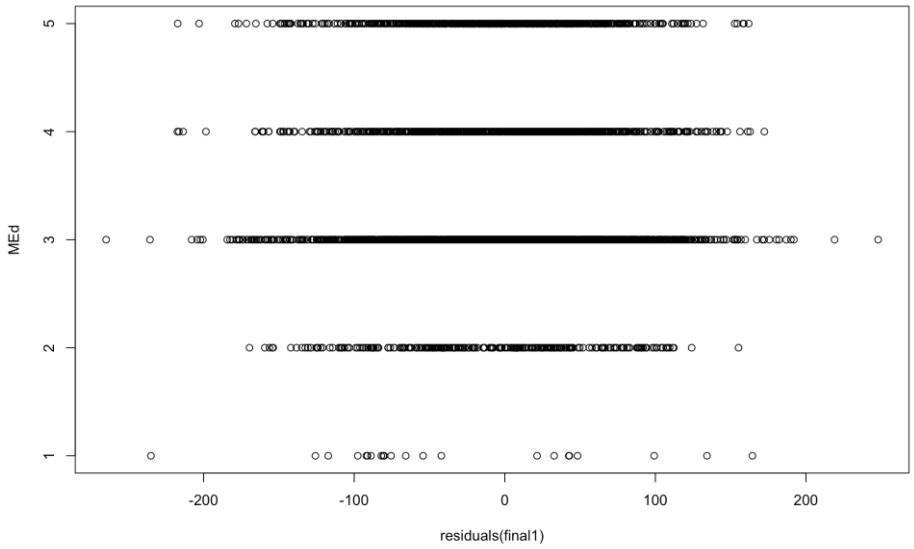


Figure 10: Residuals for PV1 against FEd

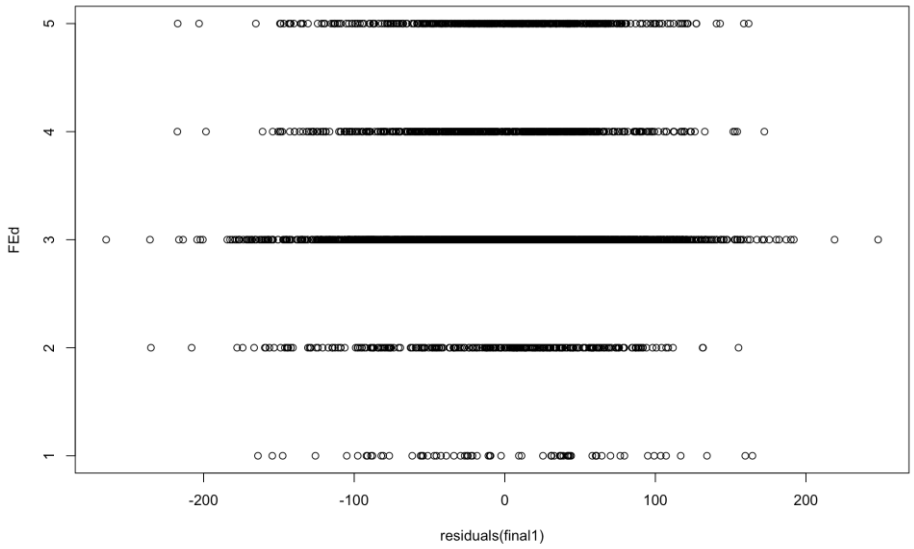


Figure 11: Residuals for PV2

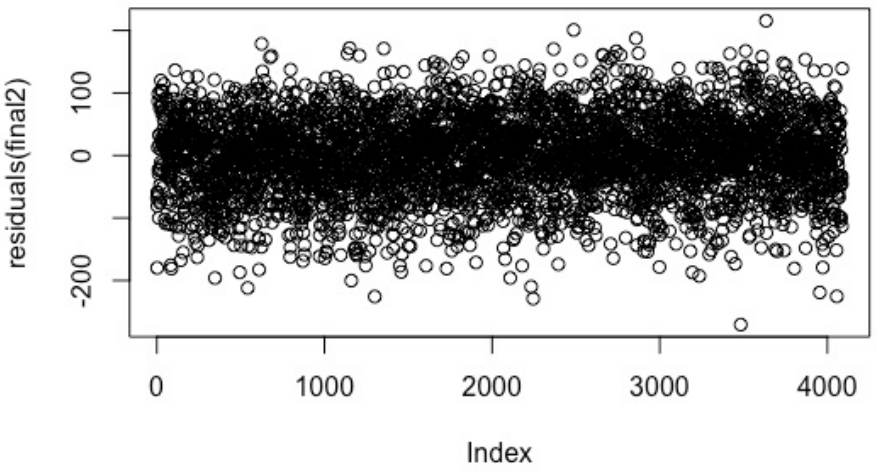


Figure 12: QQ plot with standardized residuals for PV2

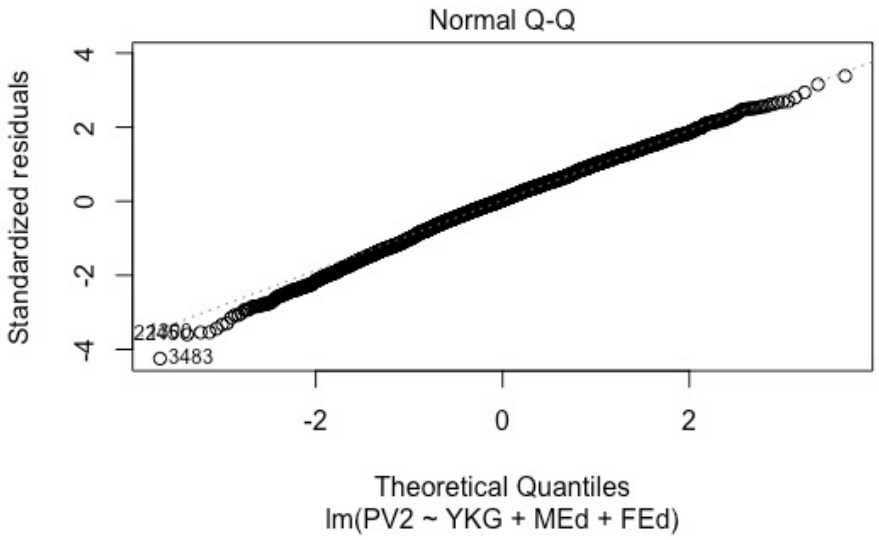


Figure 13: Residuals for PV2 against YKG

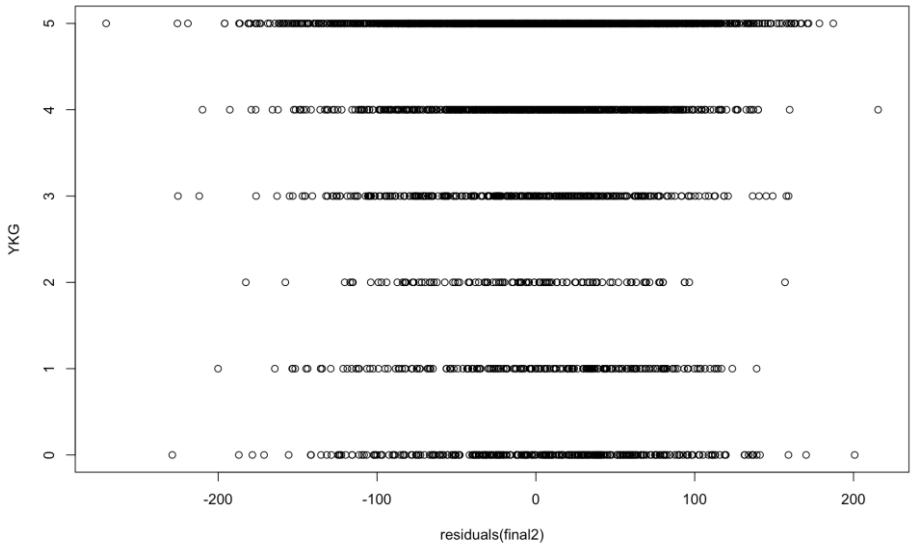


Figure 14: Residuals for PV2 against MEd

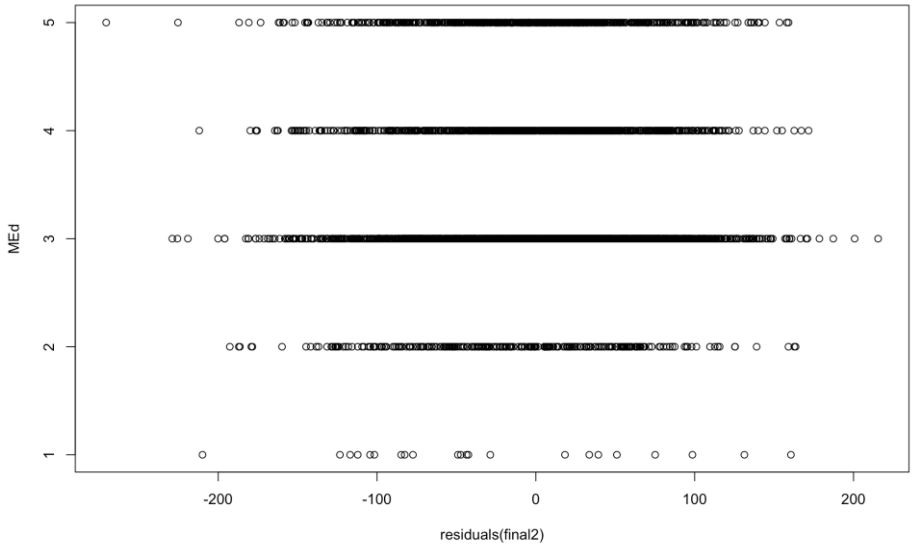


Figure 15: Residuals for PV2 against FEd

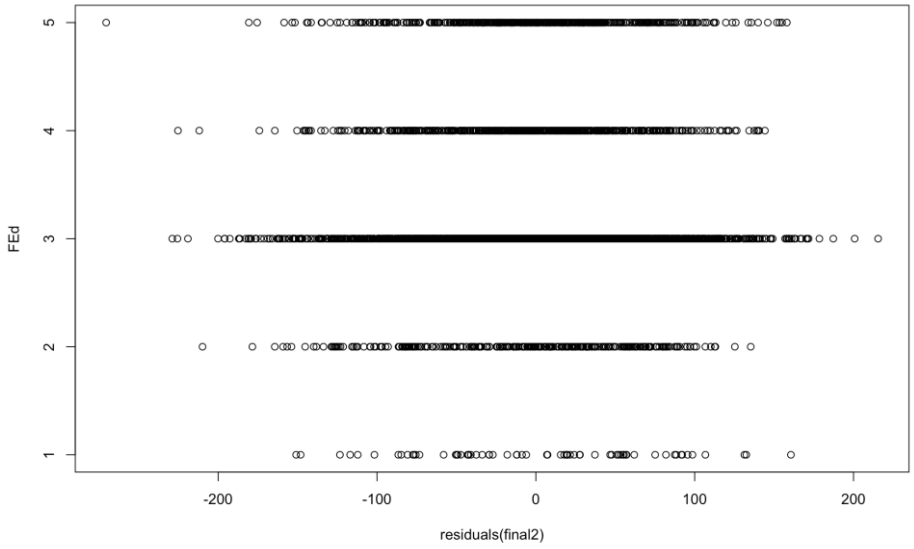


Figure 16: Residuals for PV3

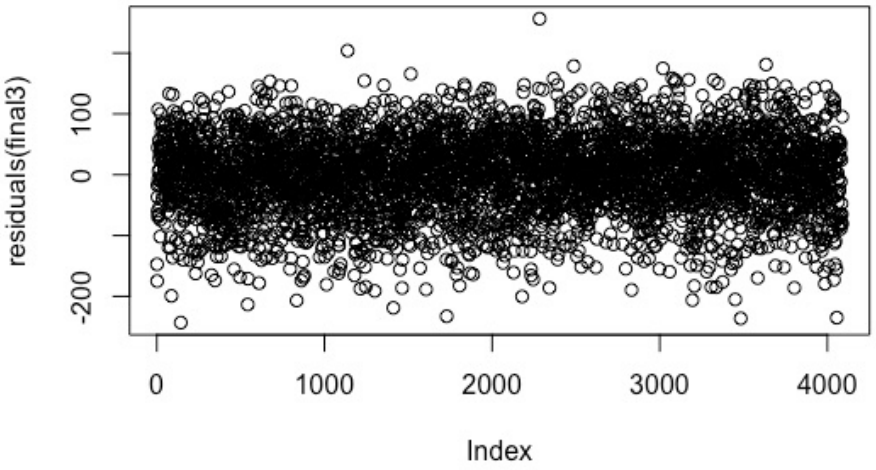


Figure 17: QQ plot with standardized residuals for PV3

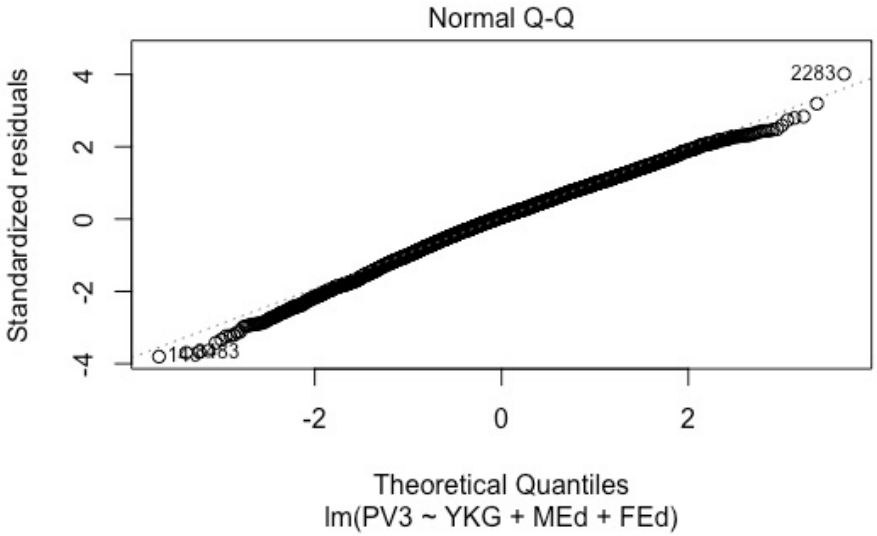


Figure 18: Residuals for PV3 against YKG

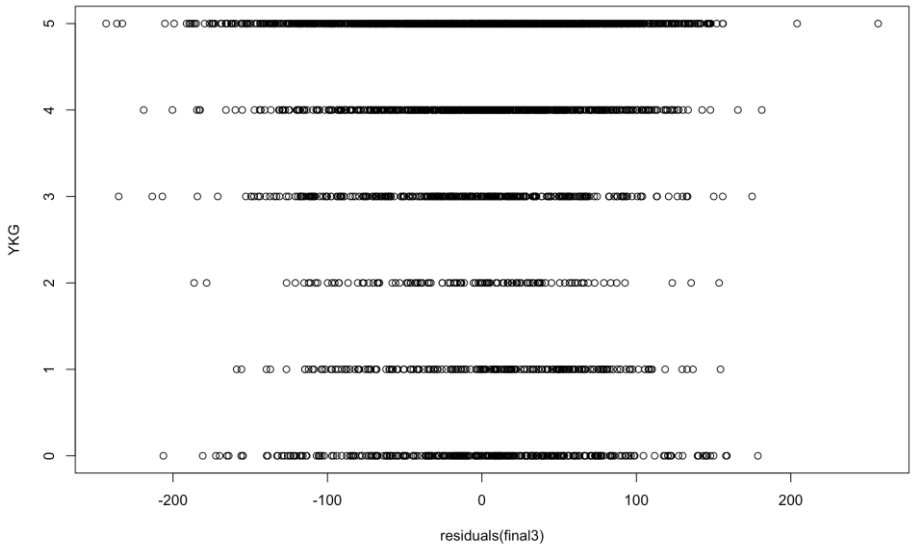


Figure 19: Residuals for PV3 against MEd

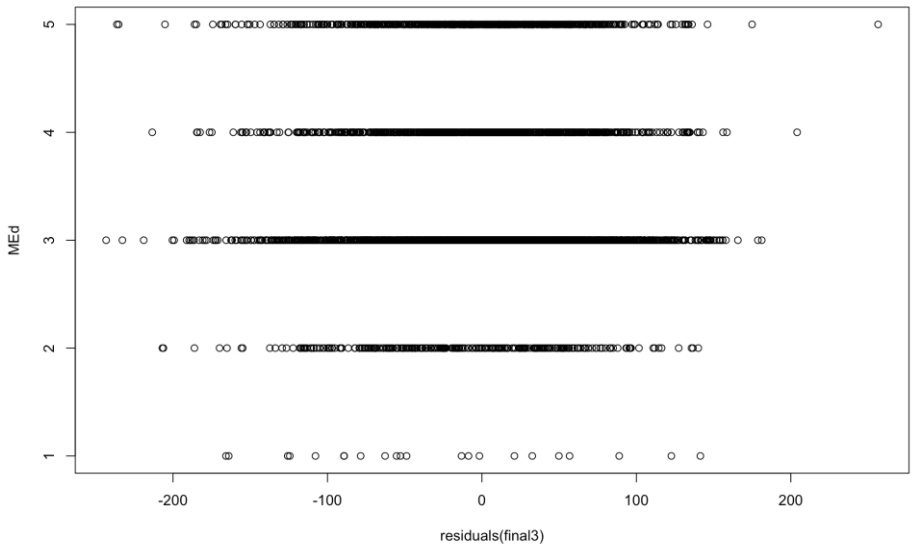


Figure 20: Residuals for PV3 against FEd

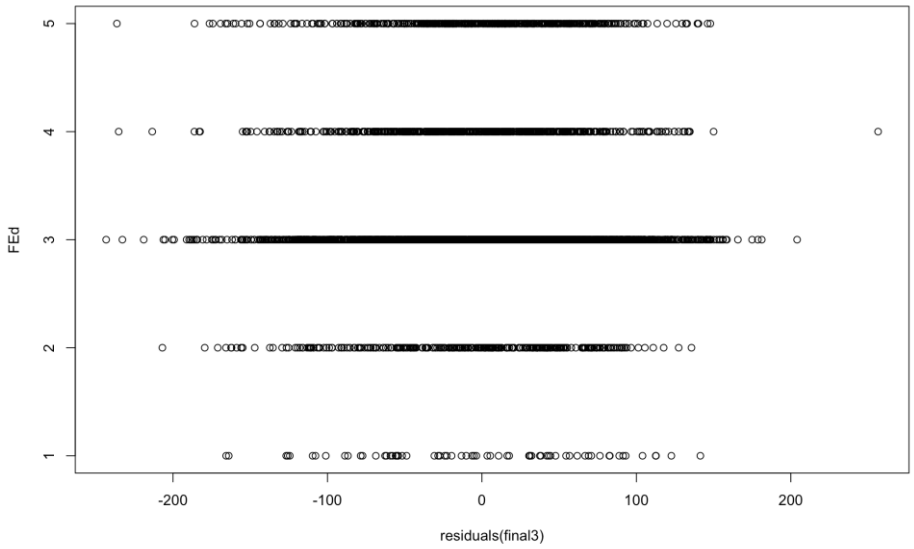


Figure 21: Residuals for PV4

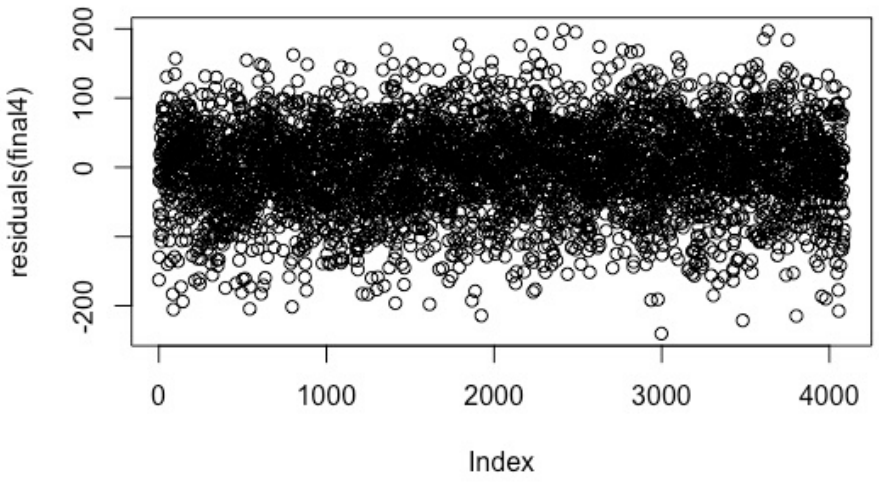


Figure 22: QQ plot with standardized residuals for PV4

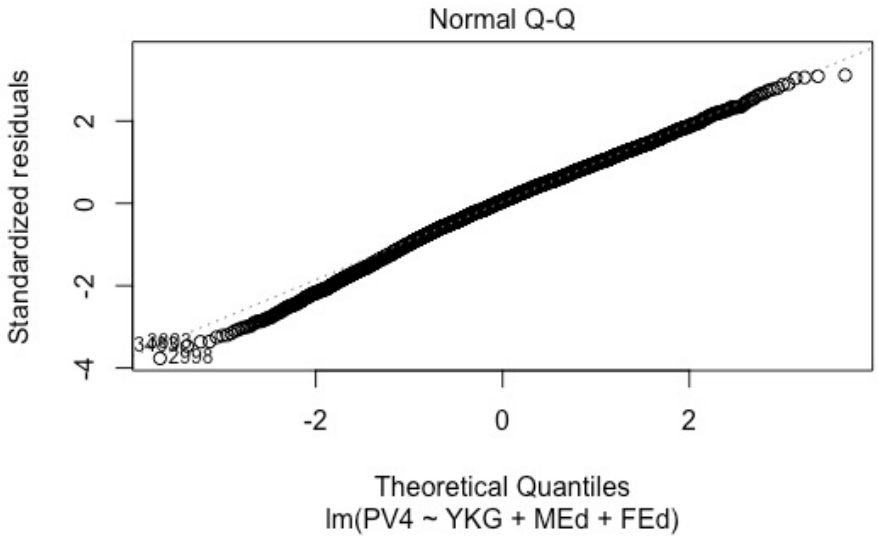


Figure 23: Residuals for PV4 against YKG

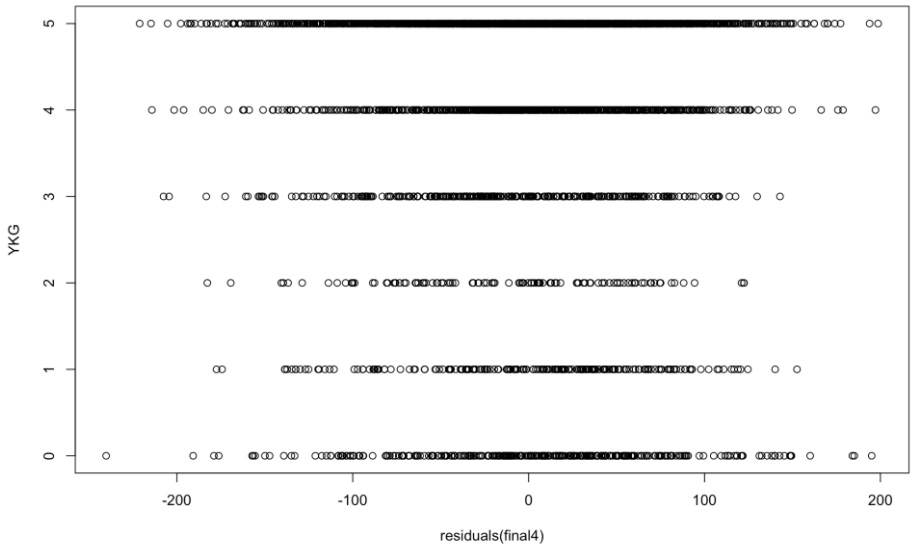


Figure 24: Residuals for PV4 against MEd

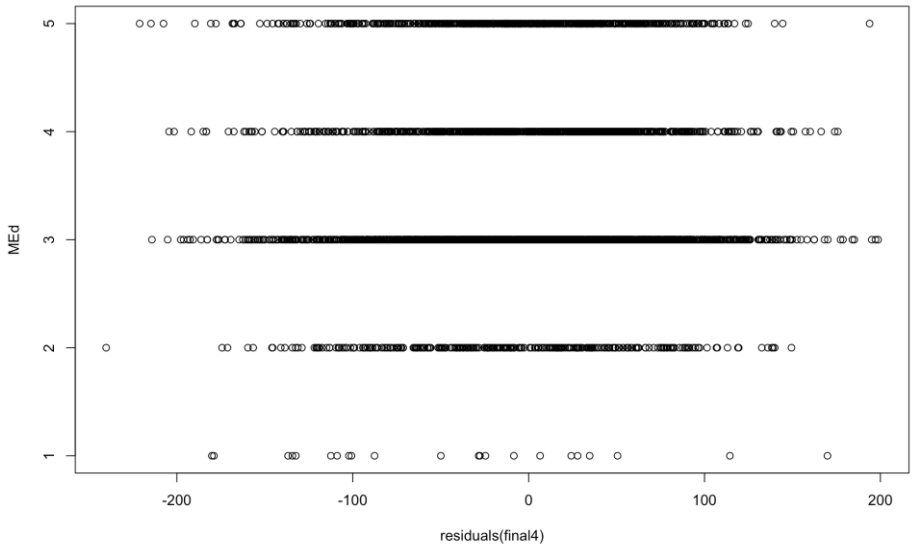


Figure 25: Residuals for PV4 against FEd

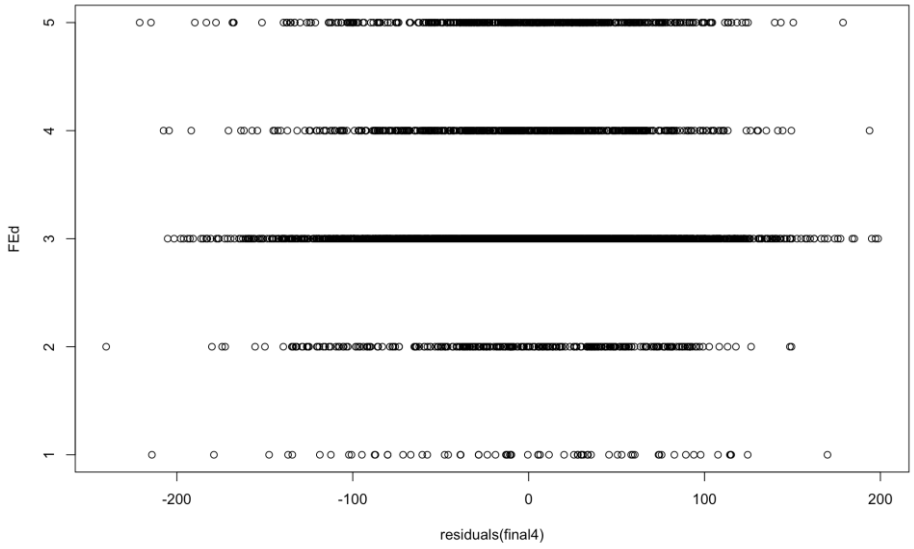


Figure 26: Residuals for PV 5

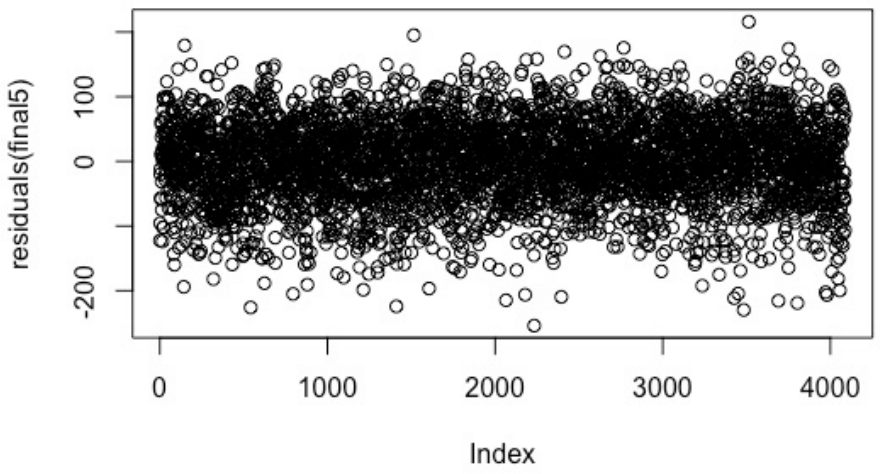


Figure 27: QQ plot with standardized residuals for PV5

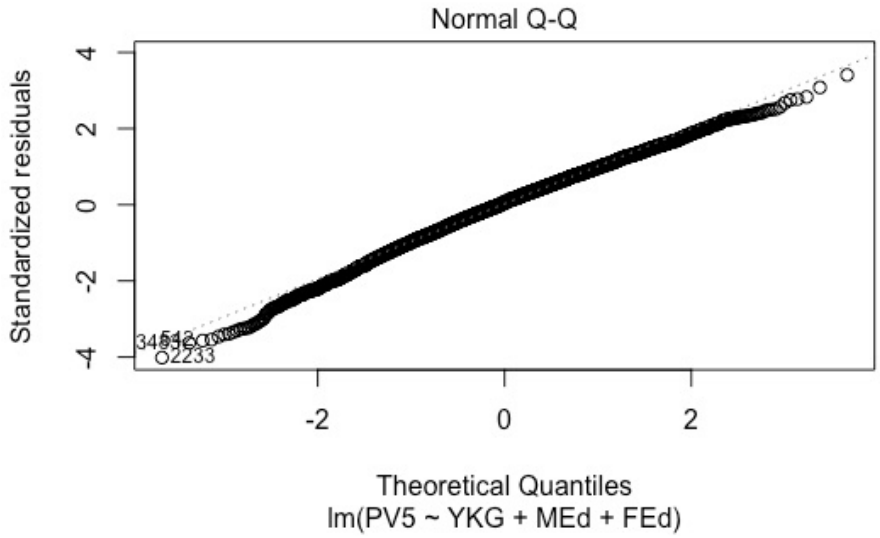


Figure 28: Residuals for PV5 against YKG

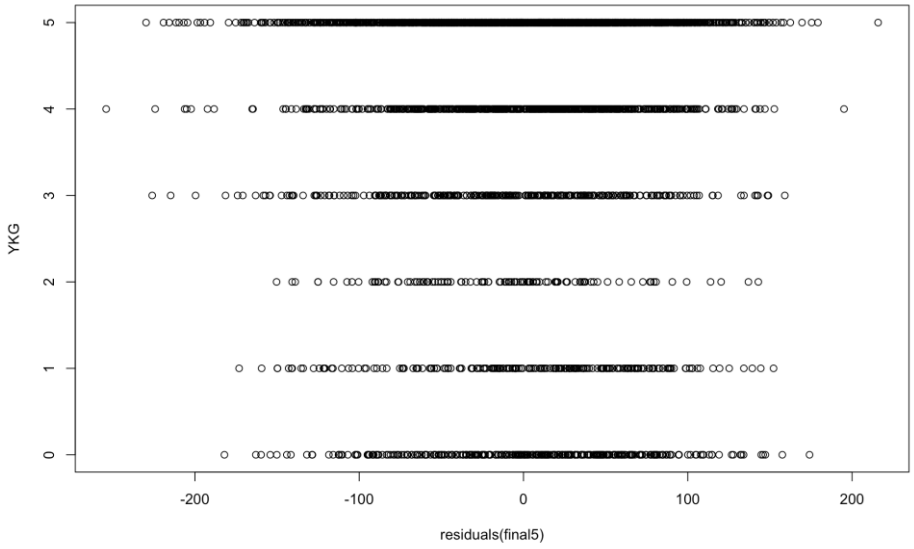


Figure 29: Residuals for PV5 against FEd

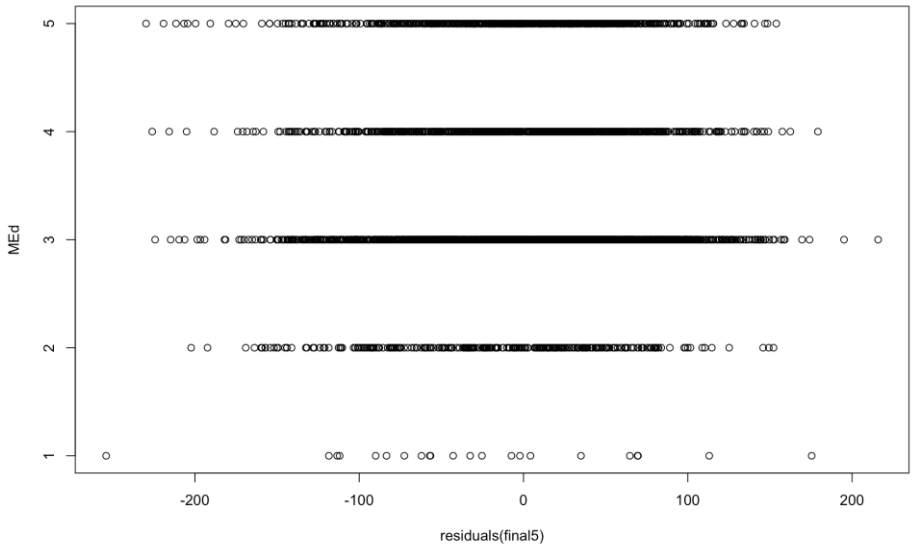


Figure 30: Residuals for PV5 against FEd

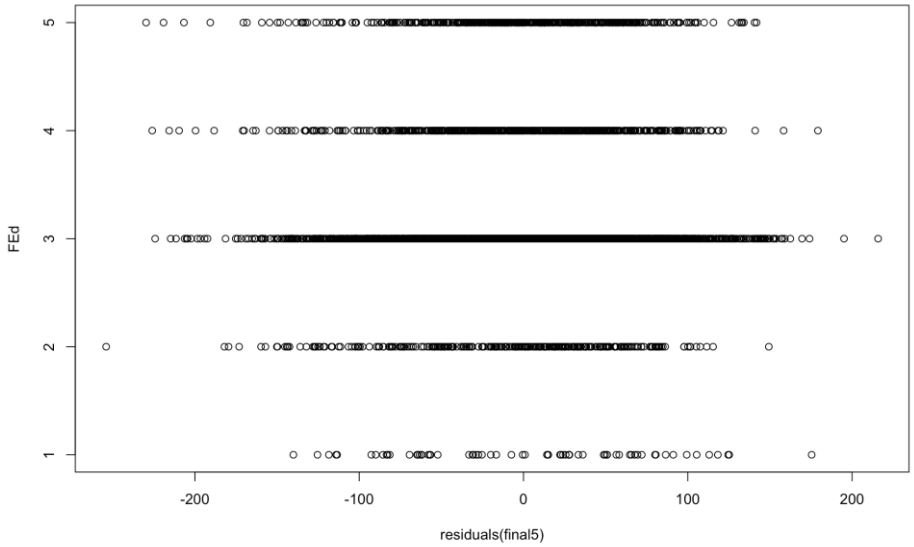


Figure 31: Residuals for AvgPV

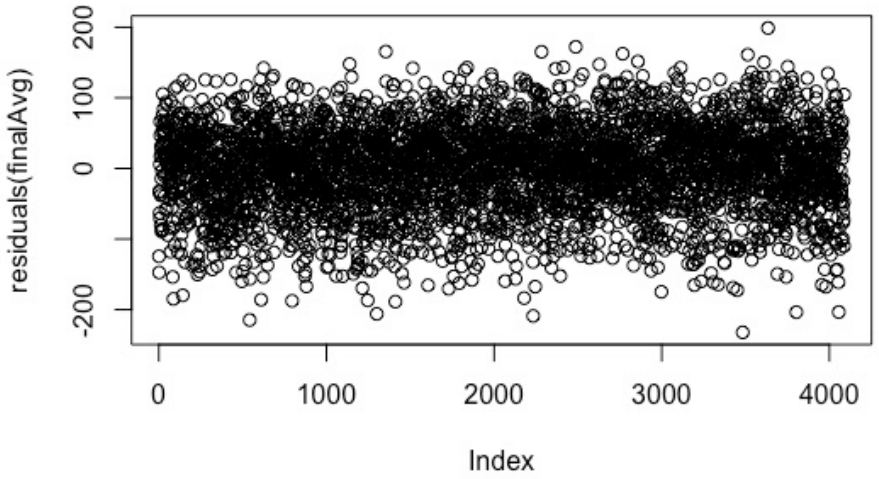


Figure 32: QQ plot with standardized residuals for AvgPV

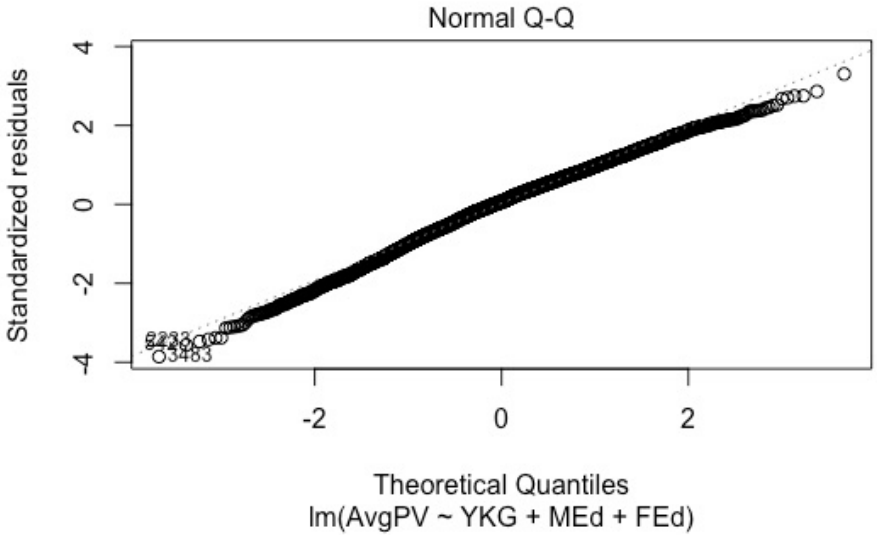


Figure 33: Residuals for AvgPV against YKG

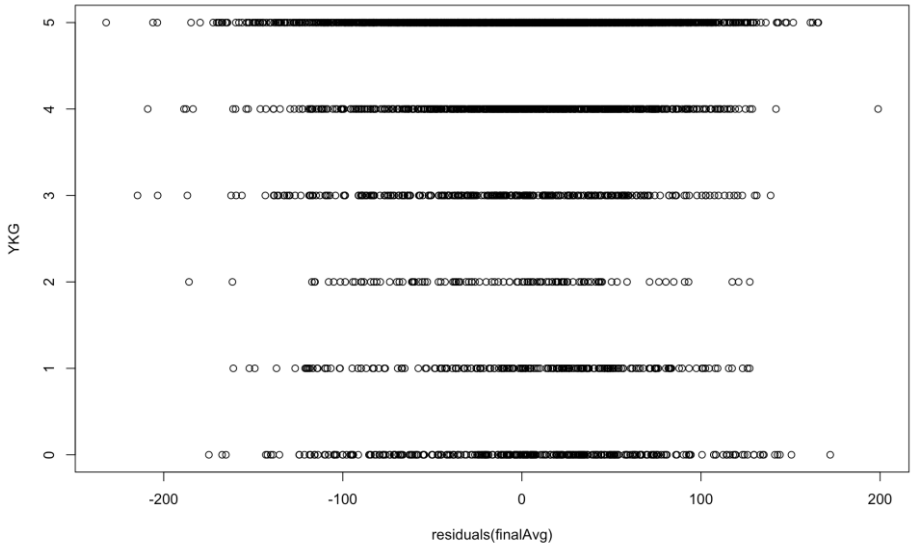


Figure 34: Residuals for AvgPV against MEd

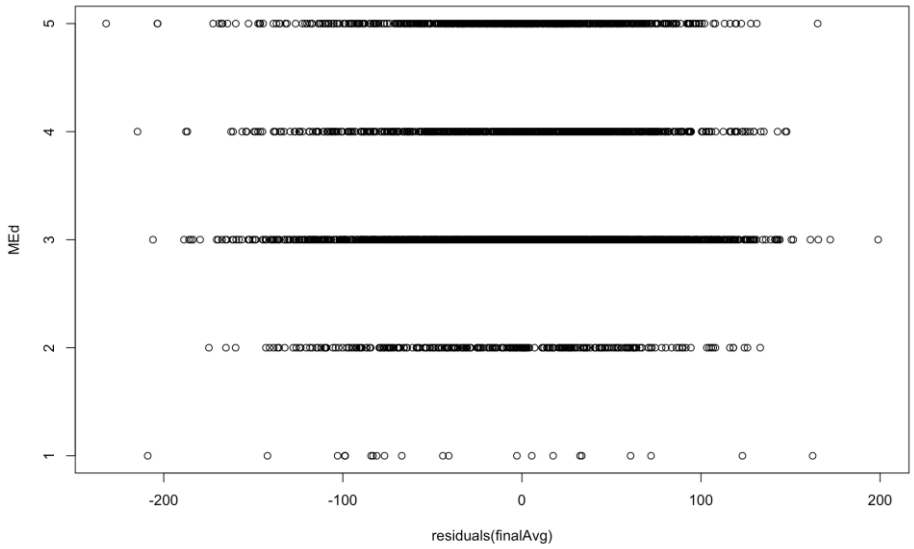
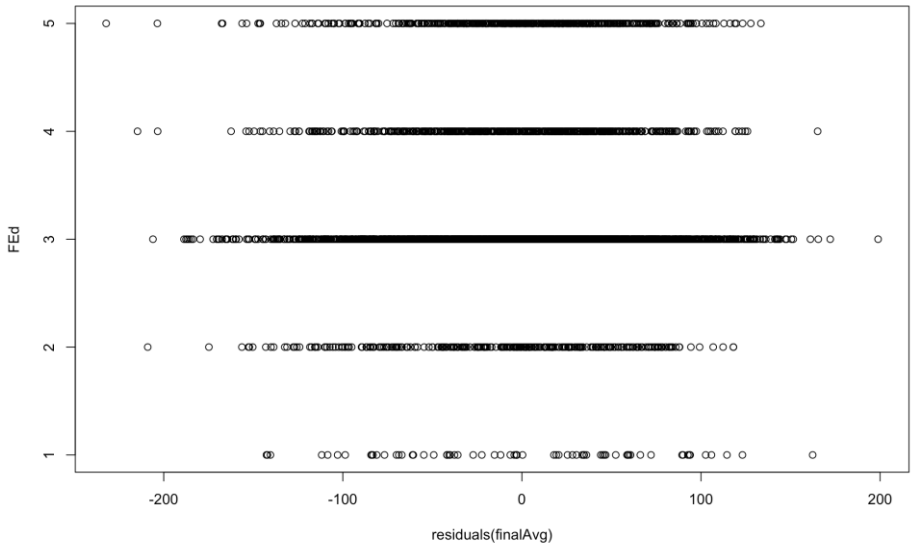


Figure 35: Residuals for AvgPV against FEd



4.3.3 Interpretation of the models

In the following section, we will explain and interpret the chosen models. Since they are very similar (they only have slightly different coefficients), the explanation holds for all of them.

The models that we chose as best are:

Final model for PV1: Y = 414.8538 + 1.4256 YKG + 21.3767 MEd + 11.0949 FEd

Final model for PV2: Y = 415.2489 + 1.6040 YKG + 20.5727 MEd + 11.3718 FEd

Final model for PV3: Y = 412.0840 + 1.5630 YKG + 20.8320 MEd + 12.1500 FEd

Final model for PV4: Y = 417.5940 + 1.3250 YKG + 21.3220 MEd + 10.5760 FEd

Final model for PV5: Y = 417.8575 + 1.0933 YKG + 21.2291 MEd + 10.6969 FEd

Final model for AvgPV: Y = 415.5275 + 1.4023 YKG + 21.0665 MEd + 11.1780 FEd

From the coefficients we can see that the most important factor in a child's reading achievement is their mother's education, followed by their father's education. The coefficient for kindergarten is small compared to the other two, even when we look at it in terms of the possible values of YKG: they are integers from 0 to 5, while possible values for FEd and MEd are integers from 1 to 5.

Let's look, for example, at the rounded numbers from the final model for AvgPV (the numbers for the other models are very similar, we just choose one model to illustrate the results, which are extremely similar for all models): A child who has a mother with the lowest education only gets 21 points from that part of the model, while the child who has a mother with the highest level of education gets 105. Each additional level of education that a child's mother reaches means that the child will get 21 additional points. This means that the predicted difference between a child with the least educated mother and a chlid with the most educated mother is up to 84 points, and for father that difference is up to 44 points (11 points per level of education), while the difference between a child who did not attend kindergarten and a child who did attended it is only up to 7 points.

5. Discussion

The goal of this thesis was to answer the questions if attending kindergarten and parental education affect reading achievement at the end of the fourth grade and test if reading achievement can be predicted by the combination of attending kindergarten and parental education. To be able to answer those questions, we need to give meaning to the results we got by using multiple regression models.

"PIRLS reports achievement at four points along the scale as international benchmarks: Advanced International Benchmark (625), High International Benchmark (550), Intermediate International Benchmark (475), and Low International Benchmark (400)" (Mullis et al. 2012, 62). This means that the average Slovenian child, who reaches 530 points, reaches an Intermediate International Benchmark and is 20 points or half a year away from reaching the High International Benchmark (students who are a year older and have spent a year longer in school get a reading achievement score that is 40 points higher (Cliffordson and Gustafsson 2008)). Since we have excluded some children from the study because of missing data, the average achievement of a child from our sample was approximately 532 points.

We tried to predict the achievement by the combination of attending kindergarten and parental education. As the best model for predicting this we chose a model of the form

$$Y = \beta_0 + \beta_1 YKG + \beta_2 MEd + \beta_3 FEd,$$

where YKG is the variable representing how many years a child spent in kindergarten, MEd is the education of the mother and FEd is the education of the father. All the coefficients were very statistically significant, so from a mathematical point of view, the answer to the research question about predicting reading achievement by the combination of attending kindergarten and parental education is: yes, it is possible to predict reading achievement by the combination of these two variables.

However, if we look at the predictions our models give us - let's again look at model for AvgPV (the results for all models are very similar, we just choose one to illustrate the situation) - we find out that the minimum AvgPV the model can predict is 447.772 points, and the highest is 583.7615, while the minimum AvgPV in the data sample is 255.624 and the highest is 737.715. This means that it is impossible to predict results that are below the Low International Benchmark or that reach the Advanced International Benchmark.

This could be due to the fact that there are, as shown by other researches (White 1982, Coleman 1966, Sirin 2005), many other socio-economic factors that might influence children's reading abilities, and to get a better prediction of the reading achievements, it would be necessary to include more of them in the model. We can still answer our first question: Does attending kindergarten as well as parental education affect reading achievement at the end of the fourth grade? Yes, it does. However, parental education, especially the mother's, is much more important than kindergarten, which has a very small influence when included in a model together with parents' education and even though it is statistically significant, it is practically not important for predicting results in the real world.

A small amount of variables in a model trying to explain a complex phenomenon such as reading achievement is something that can be criticized. The model lacks many variables for socio-economic factors. This is an obvious implication for further research, which should focus on identifying more socio-economic factors that influence children's reading abilities and including them in a model.

6. Conclusion

In this thesis, we tested 14 different multiple regression models to see if attending kindergarten in relation to parental education affects reading achievement at the end of the fourth grade and to test if reading achievement can be predicted by the combination of attending kindergarten and parental education.

The models we tested were made for the purpose of estimating a child's reading ability measured by Plausible Values methodology, which assigns 5 plausible values to each child. We also calculated and included their average. Each model had up to three explanatory variables, which were chosen among the following variables: kindergarten attendance (KG), kindergarten attendance in years (YKG), mother's education (MEd), father's education (FEd), and the highest education of either parent (HighestEd).

When choosing a model, we checked every model's residual standard error, R^2 coefficient, adjusted R^2 coefficient, and AIC and BIC values. The following model was chosen as the best for all PVs:

$$Y = \beta_0 + \beta_1 YKG + \beta_2 MEd + \beta_3 FEd.$$

After choosing the model, we assessed it with 10-fold cross validation. Then we described the coefficients for all 6 versions of the model (one for each PV, including the average PV), and plotted the residuals. From the coefficients we concluded that the most important factor in a child's reading achievement is their mother's education. The second most important factor is their father's education, and the least important one is kindergarten attendance. The answer to our first research question, if kindergarten attendance and parental education affect reading achievement at the end of the fourth grade, is therefore positive. However, parental education, especially mother's education, is much more important than kindergarten, which, when included in the model together with parents' education, has a very small influence on the reading score, and even though it is statistically significant it is not important for predicting results in practice.

To give meaning to the results, we also explained what the points from the PVs mean. 40 points represent a difference that a year of schooling makes. We also mentioned the 4 PIRLS benchmarks and realized that our model cannot predict values which fall below the lowest of them or reach the highest one. Still, from a mathematical point of view, the answer to the

second research question, if reading achievement can be predicted by the combination of attending kindergarten and parental education is yes, because all the coefficients of explanatory variables were statistically significant. However, there are many more socio-economic factors that might influence a child's reading abilities, and to get a better prediction, more of them should be taken into account. Reading achievement is a complex phenomenon that is hard to describe with a small amount of variables.

7. Literature

1. *Act on kindergartens*. 2005. Accessed 10. June 2016. https://www.uradni-list.si/1/content?id=58651.
2. Agresti, Alan. 1996. *An Introduction to Categorical Data Analysis*. New York, Chichester, Brisbane, Toronto, Singapore: John Wiley & Sons.
3. Alfons, Andreas. 2012. *Package 'cvTools'*. Accessed 10. August 2016. https://cran.r-project.org/web/packages/cvTools/cvTools.pdf.
4. Chatterjee, Samprit, and Bertram Price. 1977. *Regression Analysis by Example*. New York, London, Sidney, Toronto: John Wiley & Sons.
5. Cliffordson, Christina and Jan-Eric Gustafsson. 2008. Effects of Age and Schooling on Intellectual Performance: Estimates Obtained from Analysis of Continuous Variation in Age and Length of Schooling. *Intelligence* 36(2): 143-152.
6. Coleman, James S. et al. 1966. *Equality of educational opportunity*. Washington, DC: Government Printing Office.
7. Foy, Pierre and Marc Joncas. 2012. Sample Design in TIMSS and PIRLS. In: Martin, M.O. & Mullis, I.V.S. (Eds.): *Methods and procedures in TIMSS and PIRLS 2011*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College. Acessed 10. June 2016.: http://timss.bc.edu/methods/t-achievement-scales.html.
8. Gustafsson, Jan-Eric, Yang Hansen, Kajsa, and Monica Rosén. 2013. Effects of Home Background on Student Achievement in Reading, Mathematics, and Science at the Fourth Grade. In TIMSS and PIRLS 2011: Relationships Among Reading, Mathematics, and Science Achievement at the Fourth Grade - Implications for Early Learning, edited by Michael O. Martin and Ina V. S. Mullis, 181-287. Chestnut Hill: TIMSS & PIRLS International Study Center, Boston College. Acessed 10. June 2016. http://timssandpirls.bc.edu/timsspirls2011/downloads/TP11_Relationship_Report.pdf.
9. Hastie, Trevor, Tibshirani, Robert and Jerome Friedman. 2009. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. New York: Springer.
10. *Home Questionnaire PIRLS 2011*. 2011. Accessed on 25. May 2016. http://timss.bc.edu/pirls2011/downloads/P11_HQ.pdf.
11. Martin, Michael O. and Mullis, Ina V.S. 2012. *Methods and procedures in TIMSS and PIRLS 2011*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College. Accessed 10. June 2016. http://timss.bc.edu/methods/t-achievement-scales.html.

12. Mullis, Ina V.S., Martin, Michael O., Foy, Pierre and Kathleen T. Drucker. 2012. *PIRLS 2011 International Results in Reading*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.
13. *PIRLS 2011 International Database.* 2011. Accessed 25. May 2016. http://timssandpirls.bc.edu/pirls2011/international-database.html.
14. *Pirls 2011 Contextual questionnaires*. 2011. Accessed on May 25 2016. http://timss.bc.edu/pirls2011/international-contextual-q.html.
15. Rawlings, John O., Pantula, Sastry G., and David A. Dickey. 2001. *Applied Regression Analysis: A Research Tool*. New York, Berlin, Heidelberg: Springer.
16. *Razvrstitev kategorij slovenskega sistema izobraževanja v kategorije po ISCED 1997 (Classification of Categories of the Slovenian Education System to ISCED 1997 Categories).* 2012. Accessed 10. June 2016. https://www.stat.si/Klasius/Docs/PretKLASIUS-SRV_%20ISCED97.pdf.
17. Republic of Slovenia, Statistial Office. 2014. *Kindergartens, Slovenia, school year 2013/14 – final data*. Accessed 10. August 2016. http://www.stat.si/StatWeb/en/mainnavigation/data/show-first-release-old?IdNovice=6123.
18. Ripley, Brian, Venables, Bill, Bates, Douglas M., Hornik, Kurt, Beghardt, Albrecht and David Firth. 2016. *Package 'Mass'*. Accessed on June 20 2016. https://cran.r-project.org/web/packages/MASS/MASS.pdf.
19. Sirin, Selcuk R. 2005. Socioeconomic status and academic achievement. A Meta-analytic review of research 1990-2000. *Review of educational research* 75(3): 417-45.
20. White, Karl R. 1982. The relation between socio-economic status and academic achievement. *Psychological Bulletin* 91(3): 461-481.
21. Wu, Margaret. 2005. The Role of Plausible Values in Large-Scale Surveys. *Studies in Educational Evaluation* 31(2-3): 114-128.
22. Wickham, Hadley. 2016. *Package 'plyr'*. Accessed on June 15 2016. https://cran.r-project.org/web/packages/plyr/plyr.pdf.