KU LEUVEN

LINEAR MODELS

Regression Analysis

Report

Group 5

Ana Sofia Mendes - r0925549

Ishika Jain - r0915387

Shreekar Araveti - r0919044

Sounak Ghosh - r0914328

Stefan Panggabean - r0865831

December 2022

Introduction - The assignment aims to find the best fitting model for the response variable HAZ, which measures the stuntedness of growth in a sample of children from Ghana. The given data set has other variables, and we aim to regress these variables to find the best fitting model for HAZ .

Exploratory data analysis - Exploratory data analysis is performed to make better sense of the data set given. It is seen that there are a few categorical variables, since they have a finite number of numeric values - dummy variables are generated. To better infer the influence of these categorical variables (*akan*, *wealth* and *edu*) on the regressand, they have been modeled as factors. The given data set is then split into two as training and validation data sets. EDA is performed on the training data set. The *gg-plot* shows the scatter plots of the regressors and their correlation. From the realized correlation values it is seen that there is not much correlation between the numerical variables, implying little to no multicollinearity. It is however still possible that there exists some amount of correlation with the categorical variables, and so there may be interaction between other variables and the categorical ones.

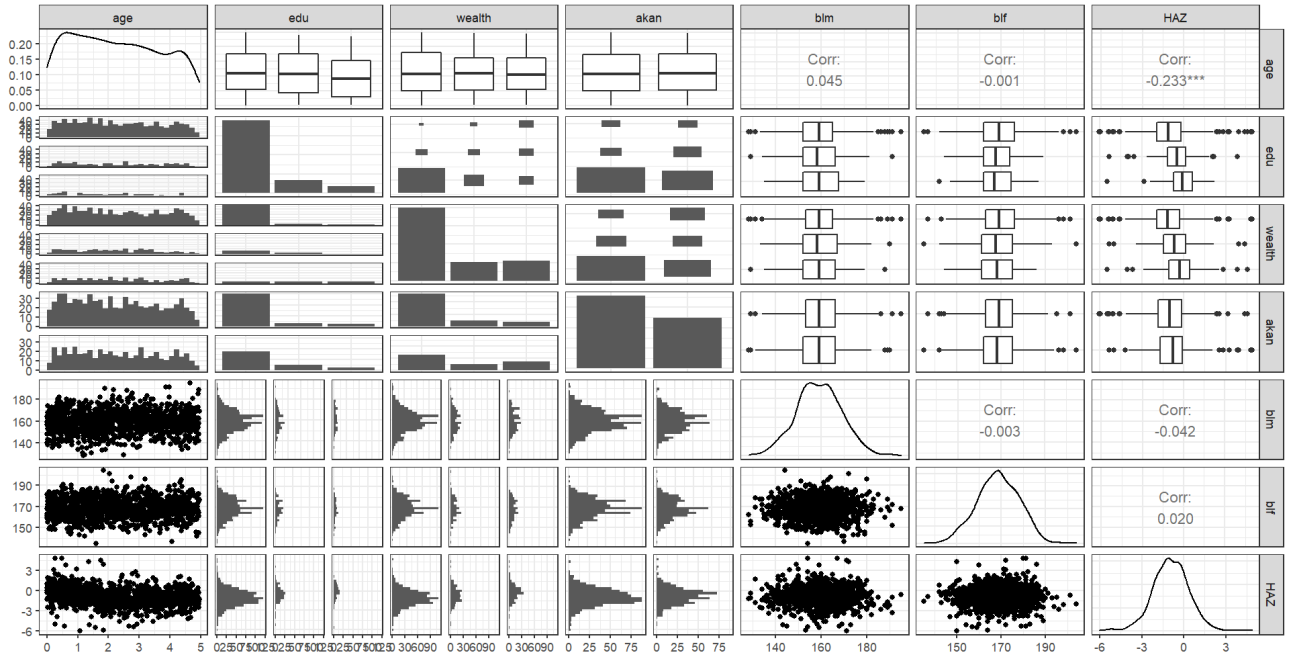


Figure 1 - GGpairs plot of the given data set

The count of the categorical variables in their categories (groups) have been plotted in Figure 3 (edu, wealth and akan). From histograms of the other variables (Figure 2 - age, blf and blm), it is seen that the variables *blf* and *blm* have a rather normal distribution, however the same can't be said about *age*. This implies that the number of children in the study across all ages are somewhat comparable.

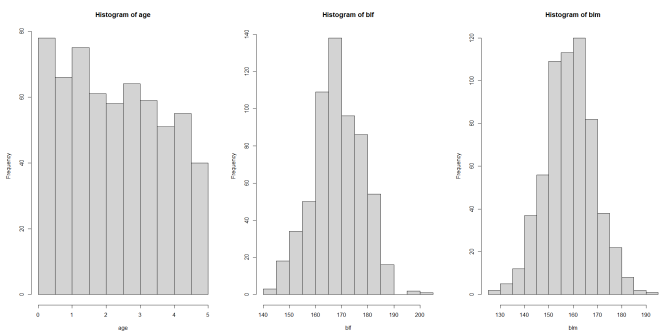


Figure 2

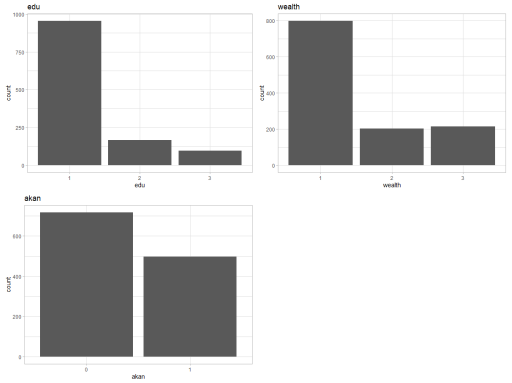


Figure 3

Boxplots of the categorical variables (Figure 4) in their groups indicate: *HAZ* value for *akan* of group 1 is on average higher than that of group 0, meaning that children of Akan ethnicity have lesser deformation. The *edu* indicated that *HAZ* value in *edu* group 3 is higher than that of 2 which in turn is higher than 1 - this implies that mothers with higher education will have children with growth that is normal or less stunted.

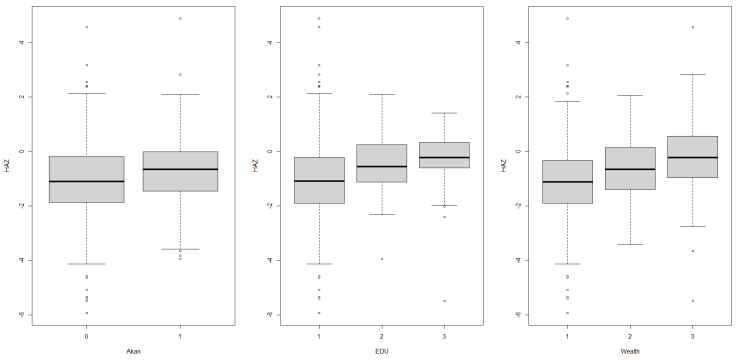


Figure 4 - Box plot as factors (akan, edu, wealth)

Wealth behaves similar to that of *edu* - children from rich households tend to have lesser deformities than those from average and poor, with children from poor households having the most stunted growth of the three. It is to be noted that points outside 75% of the first and second IQR are taken as outliers by R, boxplot implies potential outliers.

Correlation & multicollinearity - The Variance Inflation Factor (VIF) of the numeric predictors with the results in the following table. The largest VIF (1.009) is significantly smaller than 10 and the mean of the VIF values (1.006) is fractionally larger than 1. Therefore, the VIF data indicates a lack of strong multicollinearity.

The eigenvalues and condition numbers are also in support of the previous indications, since there does not exist an eigenvalue close to zero and no condition number is above 30. Another conclusion then derived is that Principal Component Regression (PCR) and ridge regression is deemed not necessary for this case.

Table 1

Variables	age	blm	blf
Eigenvalue	1.099	0.998	0.902
Condition Number	1	1.049	1.103
VIF	1.009	1.008	1.001

The following model (here on referred to as full model) has been regressed:

$$HAZ = \beta_0 + \beta_1 * age + \beta_2 * edu + \beta_3 * wealth + \beta_4 * akan + \beta_5 * blm + \beta_6 * blf$$

The *summary* shows that *age, wealth, blm* and *blf* are the only significant regressors, an undesirable adjusted R^2 value of 0.1131 is realized, implying that only 11.31% of the variance in HAZ is explained by the regressors.

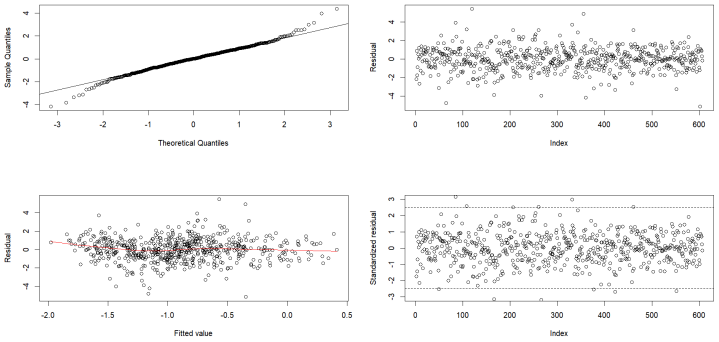


Figure 5

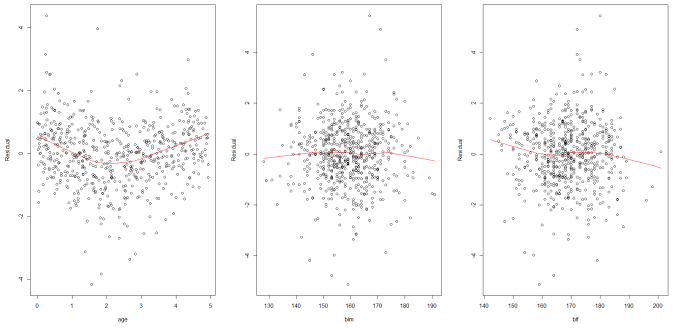


Figure 6 - Standardized Residuals (age, blm, blf)

The figure 5 (top left) shows the *qq-plot* with tails, which implies that there is data at the extremes (the data is not normally distributed). The *residuals vs fitted values* (bottom left) imply heteroscedasticity and the *standardized residuals* (bottom right) may imply outliers.

The partial residual plots (Figure 6) indicate the presence of a few nonlinear terms. The concave up nature of *residual vs age* (left) indicates presence of a quadratic regressor (age^2), however not much can be made of the remaining terms.

As the full model has realized abysmal adjusted R^2 value we will have to incorporate procedures and terms into the model seeking its betterment.

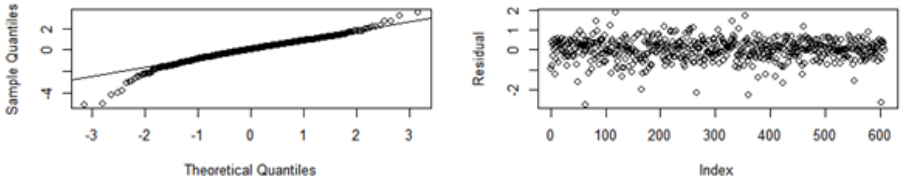
Transformations

Box-Cox Transformation - This transformation is a statistical technique that, when applied on a model, transforms it in such a way that it closely resembles a normal distribution. We use Box-Cox transformation in hope of normalizing the error distribution, stabilizing the error variance, and straightening the relation between HAZ and the regressors. The fitted model is:

((HAZ+7)^lambda - 1)/lambda ~ age + edu + wealth + akan + blm + blf

where lambda = 0.5 (value of lambda at which log-likelihood attains maximum).

Although the p-value associated with the F-statistic suggests that the regressors are jointly significant we can see that *age* and *wealth* are the only significant regressors at 5% level of significance, while all the other predictors (*edu, akan, blm, blf)* are insignificant. Moreover, from adjusted R^2 = 0.113 we can conclude that only 11.3% of the variance is explained by the regressors, which is again, very low.



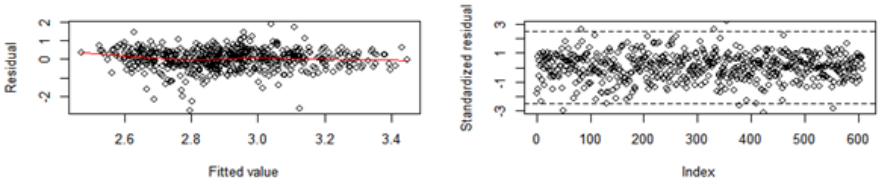


Figure 7

From the realized graphics (Figure 7), it is observed that there is very little improvement in the *qq-plot*, or the *residual vs fitted values*, and the regressors. R^2 value is very similar to the R^2 realized for the full model, which implies that the Box-Cox Transformation did not achieve the intended effect.

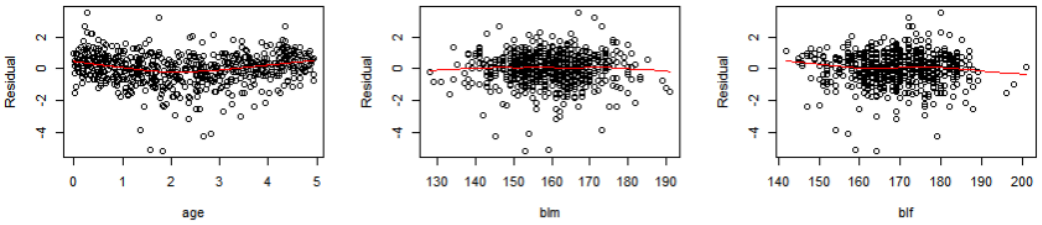


Figure 8

<u>Logarithmic Transformation of Response HAZ</u> -

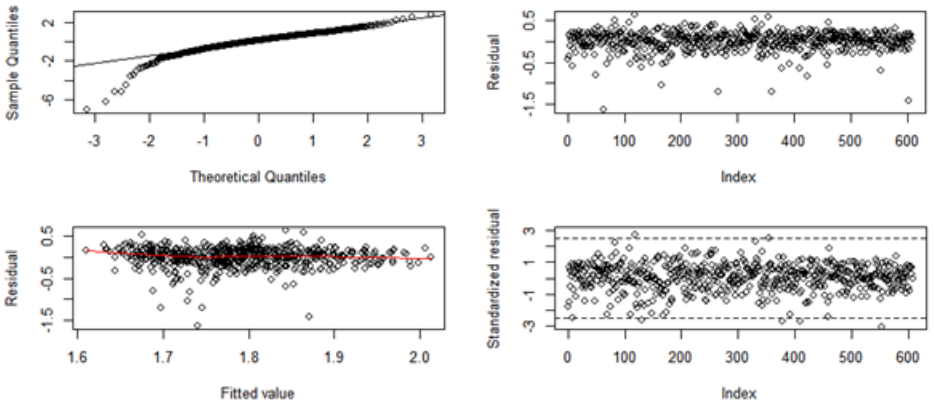


Figure 9

We fitted the model by transforming the response variable HAZ into log(HAZ + 7), and found that the normality assumption worsened since one of the tails seems heavier than the full model. There is no significant improvement in the plots of *standardized residuals vs the regressors* or the R^2 value.

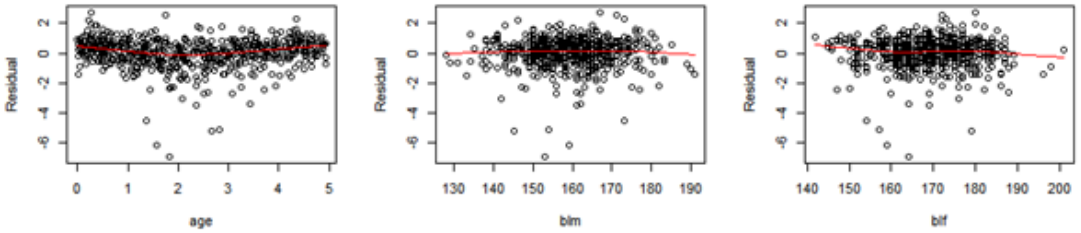


Figure 10

Weighted Least Squares Regression -

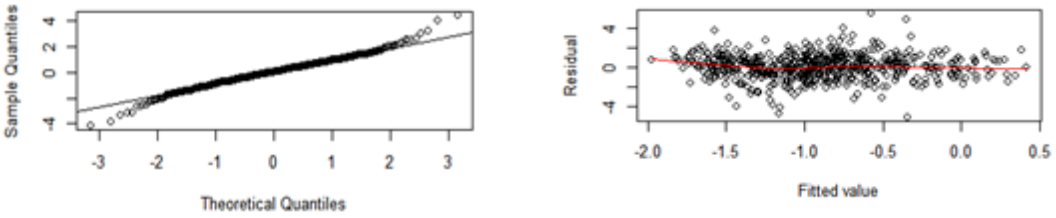


Figure 11

Transforming HAZ would change the relation between HAZ and the regressors, so we now try weighted least squares in order to reduce heteroscedasticity. We see that the coefficients are similar to the full model, thus extra reweighting is not required, there is no reduction in the standard errors, and nor did the values of R^2 improve.

The plots of weighted *residuals vs the regressors* are also almost the same as before. We find no significant improvement in the assumptions, thus we discard the Weighted Least Squares.

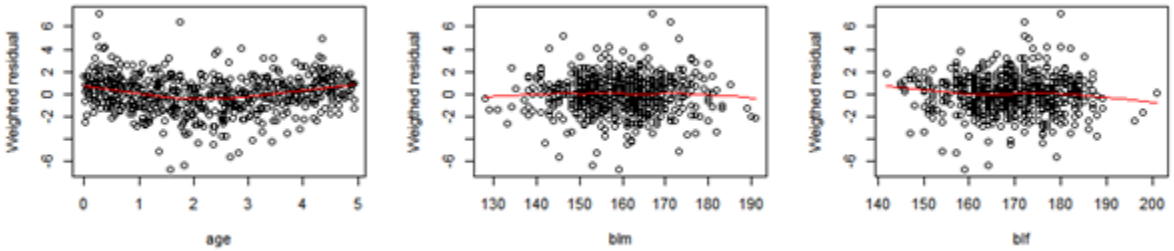


Figure 12

The reason behind the heteroscedasticity could be the omission of an important variable or interaction terms among regressors, as neither transforming HAZ, nor adding weights improved our model assumptions.

Model with higher order terms - Since we found curvature in the plots of *residuals vs fitted values*, in this part, we transform the linear model by inserting higher order terms, particularly squared terms.

On adding age^2 to the previous model we see that the p-value of the F-statistic indicates that the regressors are jointly significant. While from the p-value of the individual regressors we see that *age*, *wealth* and age^2 are the only significant predictors the value of adjusted R^2 also increases from 0.1131 to 0.207 indicating that some of the variance is explained by the age^2 term.

```
> #age^2
> fit_agesq <- lm(HAZ ~ age + edu + wealth + akan + blm + blf+ I(age^2), data = data.trainin
g)
> summary(fit_agesq)

Call:
lm(formula = HAZ ~ age + edu + wealth + akan + blm + blf + I(age^2),
    data = data.training)

Residuals:
    Min      1Q  Median      3Q     Max
-4.8158 -0.7011  0.0017  0.7037  5.1477

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -2.2097520  1.1625206  -1.901  0.05781 .
age         -1.2196988  0.1346293  -9.060  < 2e-16 ***
edu2         0.2072304  0.1558158   1.330  0.18404
edu3         0.0082451  0.2144639   0.038  0.96935
wealth2      0.4751600  0.1452303   3.272  0.00113 **
wealth3      0.8110537  0.1632998   4.967 8.90e-07 ***
akan         0.1140161  0.1029612   1.107  0.26858
blm          0.0001176  0.0048729   0.024  0.98075
blf          0.0131372  0.0051056   2.573  0.01032 *
I(age^2)     0.2144353  0.0272610   7.866 1.72e-14 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.192 on 597 degrees of freedom
Multiple R-squared:  0.207,     Adjusted R-squared:  0.1951
F-statistic: 17.32 on 9 and 597 DF,  p-value: < 2.2e-16

> # age^2 is significant, R^2 increased
```

(Figure 13) However, on adding blf^2 or blm^2, to the previous model, we see that although the p-value of F-statistic suggests that the regressors are jointly significant, only *age* and *wealth* are of any significance. Also, the R^2 hardly has any change from the one in the full model.

Therefore, we might want to consider the model: *HAZ ~ age + edu + wealth + akan + blm + blf + age*^{2}, since it has some improvements compared to the full model.

Model with Interaction Terms - In this section we tried to understand if inserting interaction effects can make an improvement in our model. Even though there is no significant correlation between numerical variables, it is still possible to have interaction between categorical and numeric variables, or between categorical variables. We fit models of all possible combinations of interaction terms . We notice that the interaction term between *age* and *wealth3* has a weak significance (*p*-value= 0.0150), and it increases the value of adjusted R^2 from 0.1131 to 0.1195 as obtained from the full model.

Also, the interaction between *edu2* and *blm* is weakly significant (*p*-value = 0.0314), and it increases the value of adjusted R^2 to 0.1201. Apart from these two interaction terms all the other ones are not significant.

```
> #age*wealth- weak significance *
> fitI3 <- lm(HAZ ~ age + edu + wealth + akan + blm + blf +age*wealth, data = data.training)
> summary(fitI3)

Call:
lm(formula = HAZ ~ age + edu + wealth + akan + blm + blf + age *
    wealth, data = data.training)

Residuals:
    Min      1Q  Median      3Q     Max
-4.9950 -0.7384  0.0148  0.7977  5.3561

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -2.5564709  1.2170597  -2.101   0.0361 *
age         -0.2351908  0.0436086  -5.393 9.99e-08 **
edu2         0.2445579  0.1630434   1.500   0.1342
edu3         0.0837499  0.2241767   0.374   0.7088
wealth2      0.4419152  0.2882078   1.533   0.1257
wealth3      0.2020381  0.2735510   0.739   0.4605
akan         0.1256965  0.1078991   1.165   0.2445
blm         -0.0000915  0.0051163  -0.018   0.9857
blf          0.0112879  0.0053375   2.115   0.0349 *
age:wealth2 -0.0173773  0.1067704  -0.163   0.8708
age:wealth3  0.2367507  0.0970044   2.441   0.0150 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.247 on 596 degrees of freedom
Multiple R-squared:  0.1341,    Adjusted R-squared:  0.1195
F-statistic: 9.228 on 10 and 596 DF,  p-value: 2.806e-14
```

```
> #edu*blm- weak significant
> fitI8 <- lm(HAZ ~ age + edu + wealth + akan + blm + blf +blm*edu, data = data.training)
> summary(fitI8)

Call:
lm(formula = HAZ ~ age + edu + wealth + akan + blm + blf + blm *
    edu, data = data.training)

Residuals:
    Min      1Q  Median      3Q     Max
-5.0680 -0.7182  0.0186  0.7677  5.4946

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.386304   1.280051  -1.083   0.2792
age         -0.194894   0.036128  -5.395 9.92e-08 ***
edu2        -4.627072   2.269767  -2.039   0.0419 *
edu3        -5.286106   3.199770  -1.652   0.0991 .
wealth2      0.391721   0.151940   2.578   0.0102 *
wealth3      0.729600   0.170438   4.281 2.17e-05 ***
akan         0.137342   0.107820   1.274   0.2032
blm         -0.008005   0.005764  -1.389   0.1654
blf          0.011235   0.005331   2.107   0.0355 *
edu2:blm     0.030825   0.014287   2.157   0.0314 *
edu3:blm     0.033455   0.019844   1.686   0.0923 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.247 on 596 degrees of freedom
Multiple R-squared:  0.1347,    Adjusted R-squared:  0.1201
F-statistic: 9.274 on 10 and 596 DF,  p-value: 2.335e-14
```

Figure 14

Since both the interaction terms are significant along with the *age*^{2} term, we can suggest the model:

$$HAZ \sim age + edu + wealth + akan + blm + blf + age^2 + age*wealth + edu*blm,$$

from here referred to as the best fitting model.

```
> fitI12 <- lm(HAZ ~ age + edu + wealth + akan + blm + blf +blm*edu + age*wealth + I(age^
2), data = data.training)
> summary(fitI12)

Call:
lm(formula = HAZ ~ age + edu + wealth + akan + blm + blf + blm *
    edu + age * wealth + I(age^2), data = data.training)

Residuals:
    Min      1Q  Median      3Q     Max
-4.5983 -0.7017  0.0138  0.7061  5.3104

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.343622   1.216173  -1.105  0.26970
age         -1.269273   0.136422  -9.304  < 2e-16 ***
edu2        -4.358222   2.155368  -2.022  0.04362 *
edu3        -5.342879   3.036141  -1.760  0.07896 .
wealth2      0.503701   0.273668   1.841  0.06619 .
wealth3      0.260912   0.259689   1.005  0.31545
akan         0.125623   0.102420   1.227  0.22048
blm         -0.005595   0.005501  -1.017  0.30948
blf          0.013958   0.005066   2.755  0.00605 **
I(age^2)     0.216183   0.027033   7.997 6.71e-15 ***
edu2:blm     0.028708   0.013571   2.115  0.03482 *
edu3:blm     0.033175   0.018827   1.762  0.07858 .
age:wealth2 -0.011015   0.101395  -0.109  0.91353
age:wealth3  0.251556   0.092098   2.731  0.00649 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.181 on 593 degrees of freedom
Multiple R-squared:  0.2268,    Adjusted R-squared:  0.2099
F-statistic: 13.38 on 13 and 593 DF,  p-value: < 2.2e-16
```

(FIgure 15) On fitting this model, the *summary* shows the added interaction terms *age:wealth*, *edu:blm* and the *age*^{2} term are significant. The *p*-value associated with the F-statistic also suggests that the regressors are jointly significant. Also, the value of the adjusted R^2 has increased to 0.2099, which means that the regressors explain almost 21% of the total variance.

So, we might consider this model, as it shows considerable improvements compared to the full model.

Outlier detection (on the best fitting model)

Investigating outliers is done with multiple diagnostic tools. With standardized and studentized residuals aimed at investigating vertical outliers, several points had been found to exceed a threshold of 2.5.

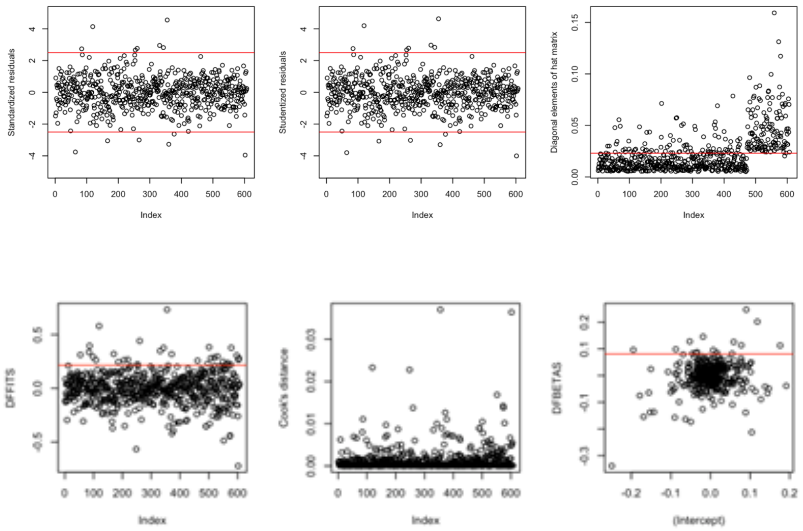
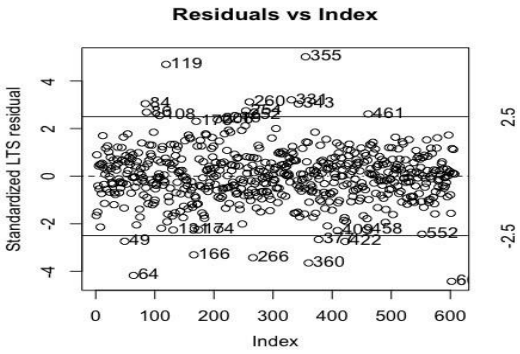


Figure 16

Next was the usage of diagonal elements of the hat matrix with a threshold of 2p/n≈0.023, it can be observed that there was over a third of the training data (aprox. 200 points) above said threshold indicating high leverage, although further research would be needed to ascertain the type of leverage points.

To evaluate the influence a single observation has on the fitted values, the DFFITS diagnostic is used with a threshold of 2*sqrt(p/n)≈0.214. A large DFFITS value can indicate a large effect of an observation on its fitted value or a large standardized residual value which is seen on multiple occasions as noted in the graphs. Based on the data provided, said outliers relative to the DFFITS diagnostic results are observations: 112, 170, 174, 200, 223, 240, 354, 387, 420, 444, 504, 515, 647, 680, 711, 757, 927, 1007, 1013, 1022, 1044, 1067, 1102, 1121, 1136, 1152, 1153, 1196, 1208, 1212.

The next evaluation used is the Cook's distance metric which measures the influence of a single observation on all fitted values. With prior information on declaring an influential data point as having Cook's distance larger than 1, the training data did not suggest any influential points.



The last evaluation measure is the DFBETAS diagnostic which measures the influence of a single observation on the regression coefficients. Setting a threshold of 2/sqrt(n)≈0.081, several of the observations noted previously are deemed influential as well.

(Figure 17) Another avenue explored for outlier analysis is with Reweighted Least Trimmed Squares (RLTS). Like Least Trimmed Squares, RLTS uses a breakdown value parameter to rank and

include residuals based on any estimate of regression coefficients. A weight function is used in order to exclude standardized residuals over 2.5, then the residuals are recalculated with WLS regression. For a breakdown value of 0.1, there were many observations considered as vertical outliers (119, 84, 86, 355, 108, 49, 64, 166, 260, 266, 360, 377, 422, 602, 461, 343, 331) with an adjusted R^2 of 0.24.
In the process, the data indicates that it is an awkward match for procedures involving robust distance since computation of the MCD covariance matrix often leads to errors caused by its near singular property. This has then led to difficulty in distinguishing between bad and good leverage points.

Considering the large difference in indices between outlier assessment tools, added to the lack of clear linear relationship in exploratory data analysis, outlier analysis has instead indicated the possibility of an underfitting scenario for linear models in general. In addition, results have thrown certainty of recurring outliers for further case-by-case analysis into question and thus the training data is left fully intact.

Variable Selection / Model Reduction

A form of stepwise regression is performed to add significant variables or remove non-significant regressors. We will use backward elimination, forward selection, and stepwise regression, based on the AIC score (i.e. the lowest value). We perform these methods on our best fitting model:

Best fitting model: *HAZ ~ age + edu + wealth + akan + blm + blf + I(age^2) + age*wealth + edu*blm*

With the backward elimination, only *akan* is removed, while with the other two methods the variables *blm, edu,* and *edu:blm* are also removed. When performing the model on the training set, we obtain an R^2=21.38% (a little lower than the one obtained for the non-reduced model). The ANOVA test shows a *p*-value of 0.13, this will imply that the reduced model can't be strictly preferred over the best model (the best model can't be rejected), meaning the prediction capacity of the two models is not significantly different. It is suggested to work with the best (non-reduced) model in this case.

Model Comparison and Validation + Discussion

table of Selected Models

MODEL 1	*HAZ ~ age + edu + wealth + akan + blm + blf* *(full model)*
MODEL 2	*HAZ ~ age + edu + wealth + akan + blm + blf + I(age^2) + age*wealth + edu*blm*
MODEL 3	*((HAZ+7)^lambda - 1)/lambda ~ age + edu + wealth + akan + blm + blf* *(Box-Cox with lambda=0.5)*

Table 2

To perform model validation, we evaluate the R^2 and adjusted R^2 along with the MSEP, MSE, AIC and PRESS scores for all models. Models with a lower PRESS value tend to have better predictive performance (i.e. less chance of overfitting). The results obtained are summarized in the table below:

Model	Set	R^2	Adj. R^2	MSE	MSEP	RMSE	AIC	PRESS
1	Training	0.1248	0.1131	1.5664	-	1.2516	2005.94	963.71
	Test	0.1305	0.1189	1.6266	2.0038	1.2754	2032.15	1004.32
2	Training	0.2268	0.2099	1.3956	-	1.1813	1940.74	862.65
	Test	0.1808	0.1628	1.5455	2.1011	1.2432	2005.97	960.99
3	Training	0.1159	0.1041	0.2826	-	0.5316	966.39	173.68
	Test	0.1212	0.1094	0.2878	0.3536	0.5365	979.16	177.61

Table 3

From the table, it can be seen that model 2 achieves the best values of R^2 and adjusted R^2 compared to the other models, both on the training and validation sets. It can also be stated that model 3 is the optimal model with respect to the PRESS, MSEP and AIC scores of the validation data set since they are all the lowest values. Moreover, the results of the training and validation sets are similar. This consistency is important as it ensures that there is no overfitting of data. However, compared to the full model, the prediction capacities get a little worse (and approximately half of the R^2 values obtained for model 2).

Conclusion - We select model 2 as our best model since it has the highest adjusted R^2 on the test set (16%). The values of PRESS and MSEP are important to check for overfitting. However, as can be seen, we never achieve a high R^2 with the models obtained. In fact, the R^2 indicates the existence of **underfitting**, meaning that a linear model might not be the most appropriate to regress on this dataset.

```
> summary(fit2.val)

Call:
lm(formula = HAZ ~ age + edu + wealth + akan + blm + blf + I(age^2) +
    age * wealth + blm * edu, data = data.test)

Residuals:
    Min      1Q  Median      3Q     Max
-5.2135 -0.7306  0.0266  0.7454  5.2731

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.068827   1.300518   1.591  0.11219
age         -0.948851   0.143372  -6.618 8.14e-11 ***
edu2        -0.416552   2.274310  -0.183  0.85474
edu3        -5.367328   3.137780  -1.711  0.08769 .
wealth2     -0.335190   0.256060  -1.309  0.19103
wealth3      0.077286   0.281866   0.274  0.78403
akan1       -0.053216   0.107927  -0.493  0.62214
blm         -0.008893   0.005736  -1.550  0.12156
blf         -0.003488   0.005307  -0.657  0.51125
I(age^2)     0.134492   0.028163   4.776 2.26e-06 ***
age:wealth2  0.330069   0.101306   3.258  0.00119 **
age:wealth3  0.213964   0.099446   2.152  0.03183 *
edu2:blm     0.004227   0.014197   0.298  0.76602
edu3:blm     0.037614   0.019727   1.907  0.05704 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.243 on 594 degrees of freedom
Multiple R-squared:  0.1808,    Adjusted R-squared:  0.1628
F-statistic: 10.08 on 13 and 594 DF,  p-value: < 2.2e-16
```

(Figure 18) **Interpretation** - The predictor variable HAZ, is a measure of the deviation from standard child growth. In the model, we observe that all significant regressors are related to the variable *age*, which makes sense when taking the response variable into consideration.

The interaction *age:wealth* also indicates that average and richer people tend to have higher HAZ (lesser deformity observed in children) than poor people for a given age as the interaction term is positive and significant. It can also be seen as poor people do not have access to proper nutrition due to a lack of money and resources, and ostensibly have a lower height-to-age score as compared to the average and rich people.

Whereas for different levels of age, the total effect of age for poor people depends on the sign of $-0.94age+0.1344age^2$. Similarly, for *average:wealth*, it depends on the sign of $-0.94age+0.1344age^2+0.33$. For rich people, it depends on the sign of $-0.94age+0.1344age^2+0.213$.

Also, *edu3:blm* and *edu3* have borderline values, so we can assume that mothers with higher education might also have an influence on the value of HAZ, making it higher and closer to 0.

It is important to note that some variables lost their significance when applying the model to the validation set.