Report 1 - Regression analysis SF2930

Group 21
Magnus Axén maaxen@kth.se 980915-7119
Henrik Fagerlund hfag@kth.se 991012-2655

November 8, 2022

1 Scenario

We have selected to work with Scenario I: Body fat assessment. Here we are working with a dataset of 14 variables, e.g. density, age and other anatomical variables, with 248 data points, i.e. 248 men. Of these coefficients we wish to predict the variable, density.

2 Introduction

The project goal is, given a dataset of certain values for bodyfat of either men or women, analyse and develop a model to estimate BFM, *body fat mass.* In this given scenario we are in the situation where $p < n$, i.e. we have more observations than variables. In this project it is said that BMI is a poor predictor of actual fatness, so BFM should instead be modelled. In this project we selected the male dataset for analysing and to develop a well fitted model.

3 Project goals

The specific project goals are listed as follows,

- Thorough residual analysis for model adequacy checking, including various types of residual scaling and plotting.
- Diagnostics and handling of outliers, leverage and influential observations using e.g. Cook's distance and CovRatio.
- Possible transformations of the variables to correct model inadequacies.
- Multicollinearity diagnostics and treatments.
- Different types of variable selection using model evaluation criteria such as e.g. MSE, AIC, BIC, Malow's C_p and adjusted R2. Evaluate MSE and adjusted R2using cross validation.
- Computer-intensive procedures for the model assessment, such as bootstrap residuals or bootstrap based confidence intervals for regression coefficients using the percentile method .

4 Analysis and theory

4.1 Residual analysis

4.1.1 Normality assupmtions

When discussing the residual we are more formally talking about,

$$Y = X\beta + \epsilon,$$

where epsilon is the error and beta is the estimated values with the real data X.

In our model assessment the following assumptions are made and examined to be satisfied about the errors.

- There is a linear relationship between the regressors and the response.
- The errors ϵ_i have mean 0, variance σ^2, and are uncorrelated.
- The error $\epsilon \in N(0, \sigma^2 I)$

4.1.2 Studentized residuals

When looking at the residuals we can rescale them, this is to get a more clear grasp as to how they behave. With studentized residuals we approximate σ with MS_{res} and defines it as follows,

$$r_i = \frac{e_i}{\sqrt{MS_{res}(1 - h_{ii)}}}$$

4.1.3 Fitted values vs studentized

A plot between the fitted values and the studentized residuals will be used to deduce various conclusions about the residuals.

4.1.4 Press

A press statistic is a measure of how well the model can predict new data. Generally we wish the press to be small since R^2 is aspired to be as close as possible to 1, seeing as it tells us what percentage of variation can be explained via the model. The following relationship is present,

$$R^2 = 1 - \frac{PRESS}{SS_T}$$

where $PRESS$ is defined as,

$$PRESS = \sum (y_i - \hat{y_i}) = \sum \left(\frac{e_i}{1 - h_{ii}} \right)^2 .$$

Seeing as this is the case, we would ideally want $PRESS$ to be close to 0.

4.2 Outlier handling

In investigating influential, i.e. points that excerpt a lot of pull on the model, one should be careful about classifying points as outliers without reason. In this section we investigate and look at the data to try to justify and research if certain points are outliers or just high leverage points. Listed below are some methods we have found to investigate the topic.

4.2.1 Cooks distance

Cooks distance is a measure as to how much pull one single variable have on the model, its given by the following formula,

$$D_i = \frac{(\hat{y_i} - y_i)^T (\hat{y_i} - y_i)}{(k+1) MS_{res}}.$$

4.2.2 DFBETAS and DFFITS

Two other measueres, DFBETAS and DFFITS,

$$DFBETAS_{jj} = \frac{r_{ji}}{\sqrt{\mathbf{r}'_{\mathbf{r}} \mathbf{r}_j}} \frac{t_i}{\sqrt{1 - h_i}}, \quad DFFITS_i = \frac{\hat{y_i} - \hat{y_i}}{\sqrt{S_0^2 h_i}}.$$

DFBETAS measures tell us about the effect β_j have if we delete the ith observation. Whilst DFFITS measures how much the ith observation effects the $\hat{y_i}$. We will use the suggested cutoffs of $|DFFITS_i| > 2\sqrt{(k+1)/n}$, $|DFBETAS_{ij}| > 2/\sqrt{n}$ for investigating.

4.2.3 CovRatio

The earlier metrics, DFBETAS and DFFITS tell us about the effect observations have on the various coefficients β_j and their respected fitted values.

However if we want insight into the overall precision of the another metric need be used. we can thus use CovRatio metric, which describes the role of the

ith observation by its precision of estimation by,

$$Covratio_i = \left| \frac{(X_i^{'}X_i)^{-1}S_i^2}{(X'X)^{-1}MS_{res}} \right|$$

In short if the $Covratio > 1$ it improves the precision and if the $Covratio < 1$ it worsens the precision. Moreover high leverage points will in turn have high $Covratio$ values, this will matter in the anlysis later.

4.3 Transformations

Transformations of variables can be needed in instances where the underlying assumptions of the model can not be satisfied or are contradicted. This can be indicated by σ^2 not being constant. We can solve it by transforming the variable y before applying the model, then methods such as box-cox can be used.

In our residual section we will investigate the normailty assumption and verify the constnat variance of the residuals. After having verified this we can conclude that no transformation is indeed needed.

4.4 Multicollinearity

If a near-linear relationship between regressors can be found in a dataset, it it said to exhibit multicollinearity. This near-linear relationship exist if there is non-zero $t_1, t_2, ..., t_n$ such that $t_0X_0 + t_1X_1 + t_2X_2 + ... + t_nX_n \approx 0$, i.e such that $X\mathbf{t} \approx 0$.

A dataset that exhibit multicollinearity will cause the equation $X^TX\hat{\beta} = X^T\mathbf{y}$ to have a unique solution for $\hat{\beta}$ but to be very sensitive to changes in X. This can result in very large confidence intervals for β_j.

4.4.1 VIF

VIF (Variance Inflation Factors) is a method for detecting mulitcollinearity by examine a VIF-value for all regressors. These factors/values shows to which extent each regressor is linearly dependent of the other regressors and is calculated using equation 1. The rule to follow is that the dataset exhibit multicollinearity if $VIF_j \geq 5$ for some regressor j.

$$VIF_j := \frac{\text{Var}\left(\hat{\beta}_j\right)}{\sigma^2} = \left(X^TX\right)^{-1}(j,j) = \left(1 - R_{(j)}^2\right)^{-1} \tag{1}$$

where

$$R_{(j)}^2 = 1 - SS_{Res}^{(j)} / SS_T^{(j)}$$

4.4.2 Eigenvalue analysis of $X^T X$

A consequence for a dataset where multicollinearity can be found is that $X^T X$ is ill-conditioned and have small determinant. Since the determinant is equal to the product of the eigenvalues, at least one of the eigenvalues will be small if the determinant is small.

The rule to follow is that if the condition indices $\kappa_j := \lambda_{max}/\lambda_j \geq 100$ for some j, the dataset suffer from multicollinearity.

The condition number is defined as $\kappa := \lambda_{max}/\lambda_{min}$. The dataset is said to have moderate to strong multicollinearity if $100 \leq \kappa < 1000$, and severe mulitcollinearity if $\kappa \geq 1000$.

4.4.3 Ridge regression

When a model suffer from some multicollinear independent variables, a good way of estimating the coefficients for a better fitted regression model is by using ridge regression method. A "good" estimator is generally **unbiased** and have **small variance**.

Least squares regression tries to minimize the sum of squared residuals, RSS,

$$RSS = \sum (y_i - \hat{y}_i)^2$$

Ridge regression on the other hand, seeks to minimize the following expression,

$$RSS + \lambda \sum \beta_j^2$$

where the second term is referred to as *shrinkage penalty* and one seeks to select the λ that minimizes mean squared error (MSE).

Ridge regression method will result in a final model can be fitted using the optimal λ.

4.5 Variable selection

Often we are interested in knowing which models actually improve the model accuracy, this is where variable selection comes into play. Moreover sometimes its more practical to keep a model simple and easy whilst in other cases we want to have a more complex model with more variables.

4.5.1 Forward/backward elimination

Forward- and backward stepwise selection are two stepwise regression methods used to build a model containing only the most useful variables.

Forward stepwise selection method starts with an empty model and then add one variable at a time. The variable added is the one that gives the single best improvement to the model. Once the model no longer improves by adding variables, some pre-specified stopping rule is achieved, or all of the variables have been added, the process stops.

Backward stepwise selection is basically the opposite of forward. It starts with a maximum number of variables and eliminating the least significant variable one by one. Similar to the forward method, once the model no longer improves by eliminating variables, some pre-specified stopping rule is achieved, or all of the variables have been removed, the process stops.

The terms that are added/removed are being chosen by a F-statistic, by finding the variable with the highest correlation to y. Here adjusting various thresholds will give us different models.

Often forward selection is used to find an easier model with fewer coefficients, and backward is used to find a more complex model describing all its intricacies.

4.5.2 BIC,C(p), Adjusted R^2

BIC, Mallwos C_p and Adjusted R^2 are various metrics of looking at which model might be best suited.

Adjusted R^2 is computed as,

$$R^2_{adj} = 1 - \frac{\left(1 - R^2\right)(N-1)}{N-p-1} \tag{2}$$

where

$$R^2 = 1 - \frac{SS_{\mathrm{Res}}}{SS_{\mathrm{R}}}$$

N - Total sample size
p - Number of independent variables

here, we seek the model that maximizes the adjusted R^2.

The other values, BIC, and Mallows C_p can be computed as follows,

$$C_p = \frac{1}{n}(RSS + 2d\hat{\sigma}), \quad BIC = \frac{1}{n}\left(RSS + \log(n)d\hat{\sigma}^2\right),$$

(here d is the number of predictors).

These are values that ideally is sought to be minimized, which can be deduced simply by looking at the equations.

4.6 Bootstrapping

Bootstrapping is a computer intensive procedure in which we can reliably estimate the error in the predicted estimated coefficients with confidence intervals. The bootstrapping selects a smaller sample of the original data, which will omit some data and can contain duplicates of some data. This is done to empirically determine the coefficients and create confidence intervals for said coefficients.

In the bootstrapping example we use the function Boot in R to construct 1000 resamplings with a 95 percentage confidence interval.

5 Results

5.1 Residual analysis

5.1.1 Normality assumption

In order to examine whether or not the normality assumptions are satisfied for the residuals e_i we plot a given QQ-plot and examine it visually. This is done for the full linear model, i.e. taking all coefficients into consideration. As can be seen, it does follow the line relatively well, only comment is that the ends are a bit off, indicating heavier tails for the Gaussian distribution. But seeing as it it barely deviates from the diagonal we conclude that the errors are indeed Gaussian.

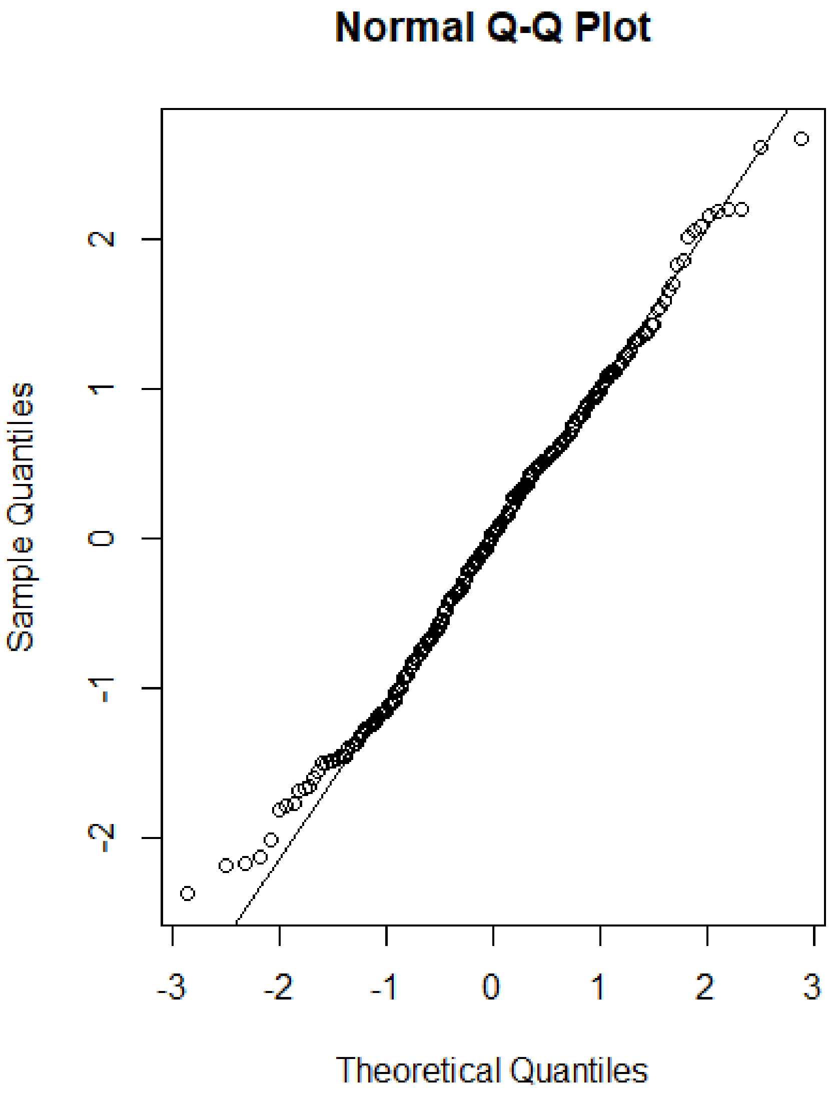


Figure 1: QQ-Plot residuals

5.1.2 Press

From the full fitted linear model, we got a $PRESS = 0.02588875$, which might not be that telling (even though its small), not knowing the magnitude of the coefficients and variables. But instead $R^2_{prediction} = 0.7052601$, there is some ambiguity as to what constitute a "good" R^2. From the large data set we could perhaps say its not random that its roughly 70 percentage, but also that the full LM doesn't tell the full story.

5.1.3 Fitted values vs studentized

As we can by inspection from the plot, the residuals seem to be gathered around 0, varying with some variance. Moreover they seem also to be distributed evenly above and below the horizontal line. Which would suggest that the residuals are well behaved.

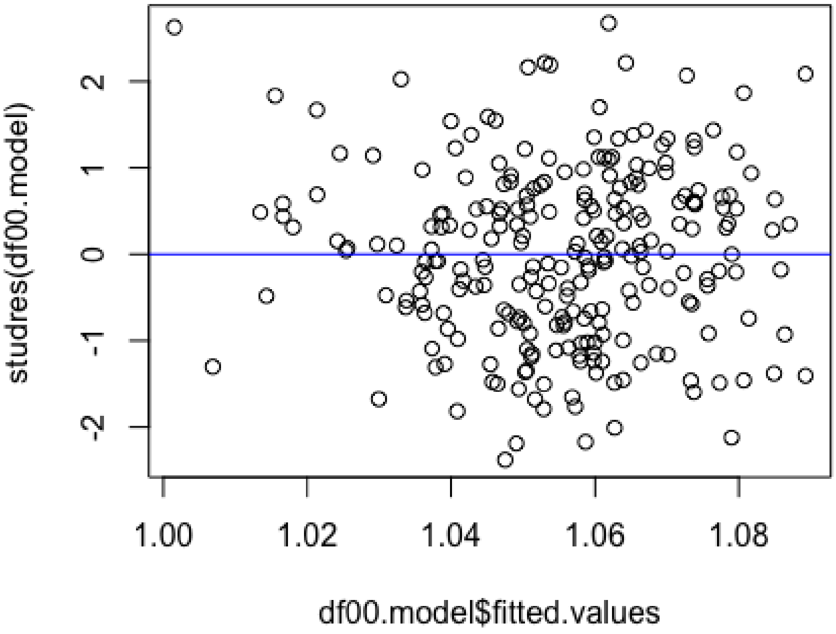


Figure 2: Student residual vs fitted values

5.2 Outlier handling

5.2.1 Cooks distance

After performing the above mentioned test and with the given threshold we attain a plot as follows. It is thus seen that point 39 and 83 are above the threshold. Moreover we can also note that none of the other points are even close in reaching the threshold.

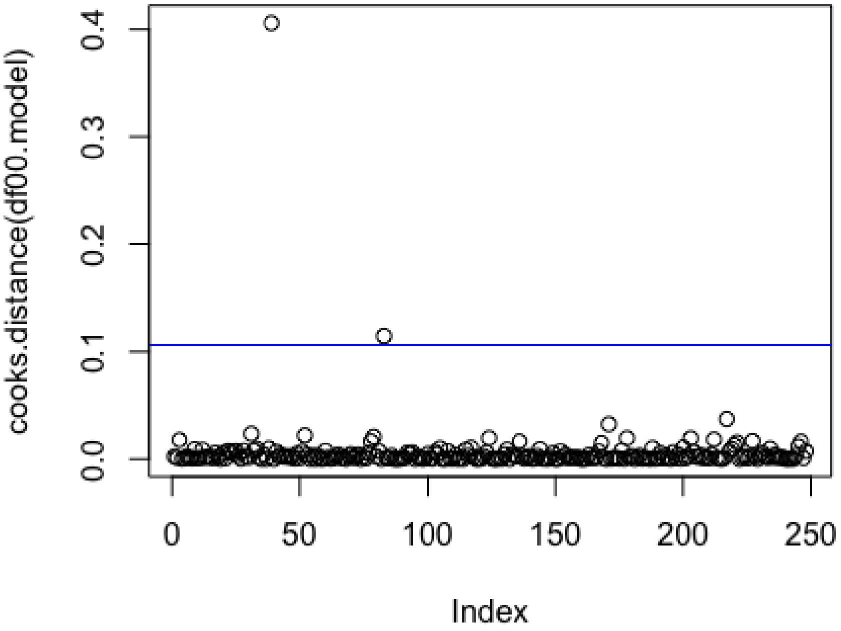


Figure 3: Cooks distance

5.2.2 DFBETAS and DFFITS

Listed below we have the DFFITS plot which is indicative of the similar high influence points as before. Comment thresholds basically filling no function.

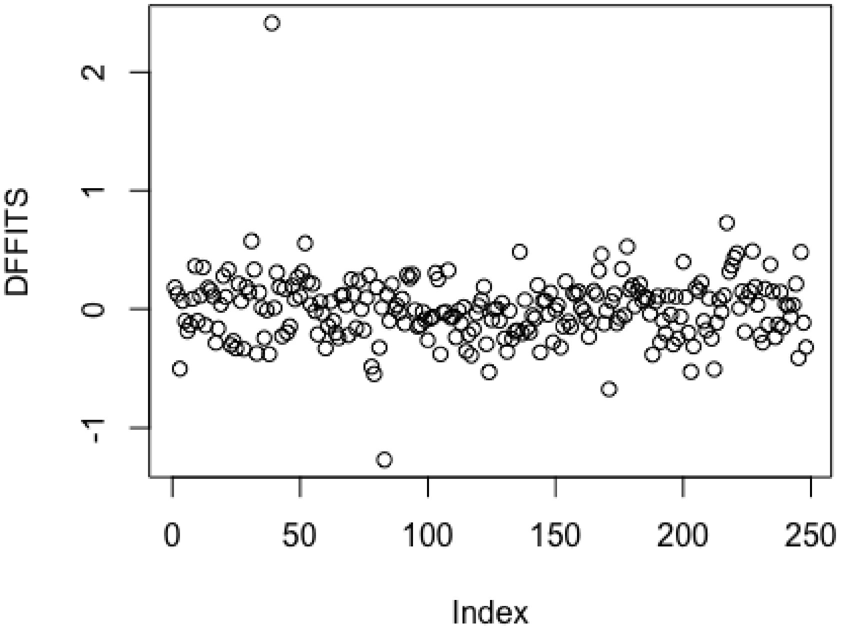


Figure 4: DFFITS

5.2.3 CovRatio

Below listed is the CovRatio plot of the full LM. From this we can tell the two former points which were mentioned have indeed high covratio.

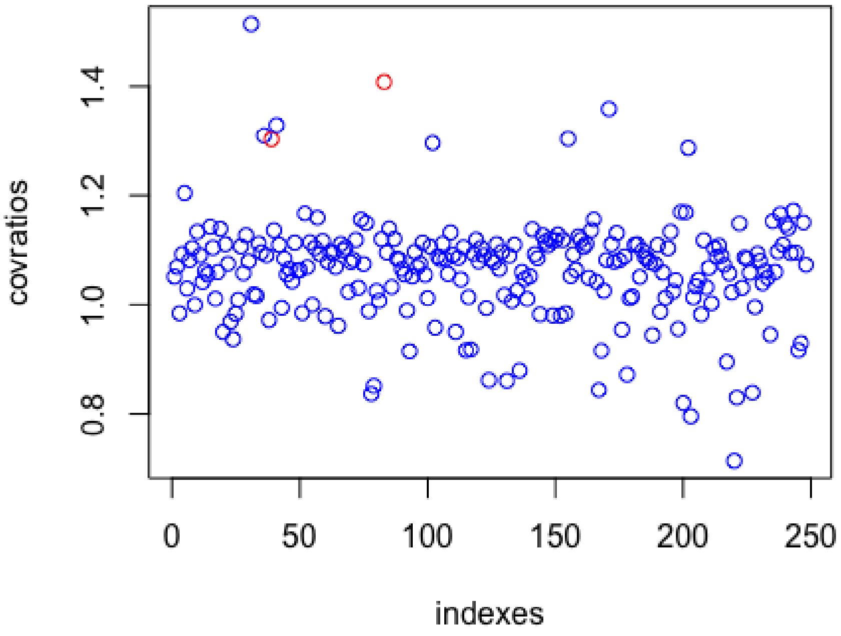


Figure 5: Covariance ratio.

5.2.4 Final words about outliers

The above analysis gave us reason to inspect the points closer. Upon doing so we discovered that the points 39 and 83 looks as if they could be potential outliers. The point 83, looks as if though the ankle value have had accidentally a 3 (30) inputted instead of a 2 (20), causing this large shift. And the point 39 is someone who is severely overweight. However since we don't know how the data was attained or have any more insight we can not confidently remove any of the points.

5.3 Multicollinearity

The levelplot of the $X^T X$ matrix is shown in figure 6 and indicate that some near-linear relationship exists (the dark-green area, e.g. abdomen to chest).

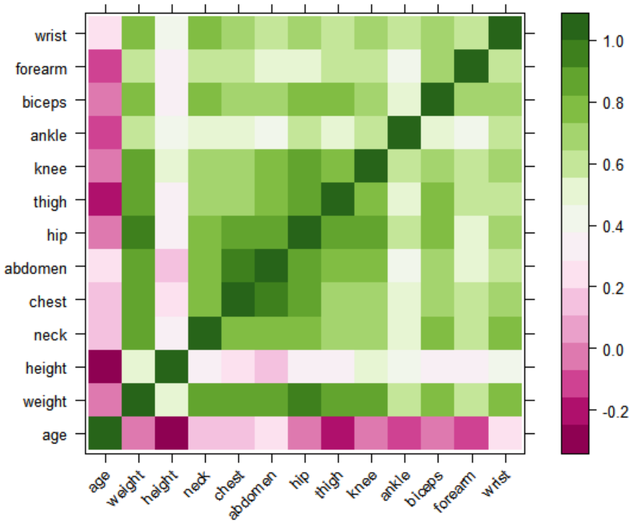


Figure 6: Levelplot of $X^T X$.

5.3.1 VIF

The following plot present the VIF-values for all regressors done on the original fitted model. The plot is showing multiple regressors to have VIF-values higher than 5 (indicated by the blue line), resulting in that multicollinearity can be assumed to be found in the dataset.

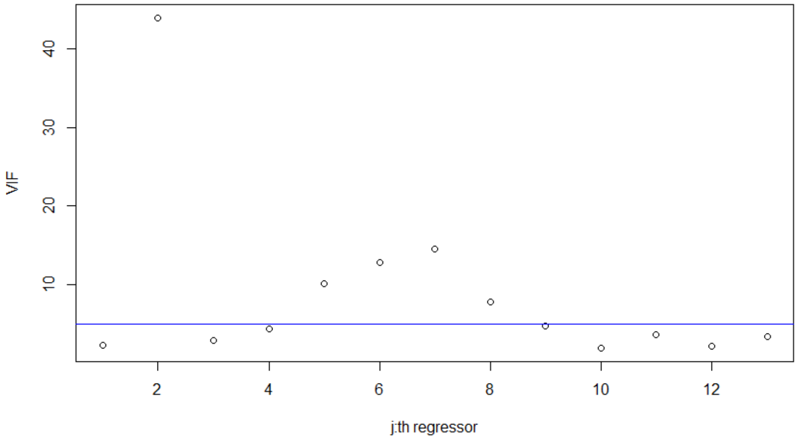


Figure 7: VIF-values for the regressors.

5.3.2 Eigenvalue analysis of X^TX

The condition indices for all regressors is calculated and presented in figure 8.

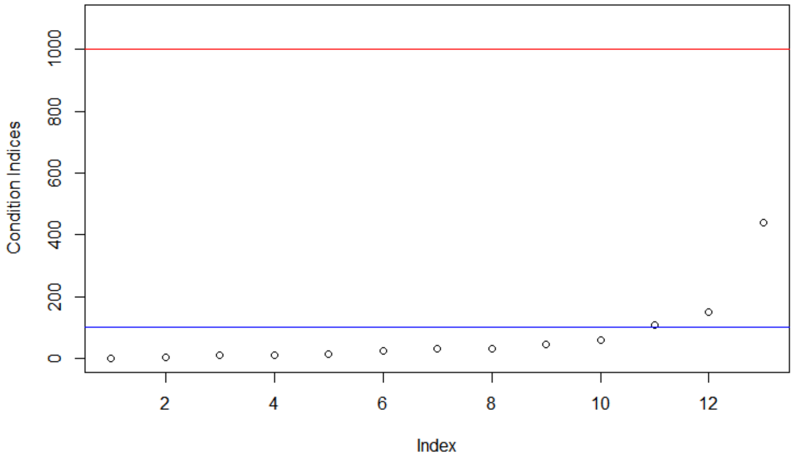


Figure 8: Condition indices for the regressors.

From this plot, it is shown that the condition indices takes values above 100 for three regressors and indicate that mulicollinearity exist in the dataset. This agrees with the conclusion based the VIF-method.

The condition number, which is the maximum condition index seen in figure 8, is equal to 441. This is lower than 1000, thus the dataset can be said to exhibit moderate to strong mulitcollinearity, and severe mulicollinearity is not quite the case.

5.3.3 Ridge regression

The specific λ that minimizes MSE is identified using k-fold cross validation. The number of folds performed is the default $k = 10$ folds, this to make sure that it is sufficient amount of folds for the dataset, a few less folds would also work. Test MSE (Mean Squared Error) is plotted for different values of λ and presented in figure 9.

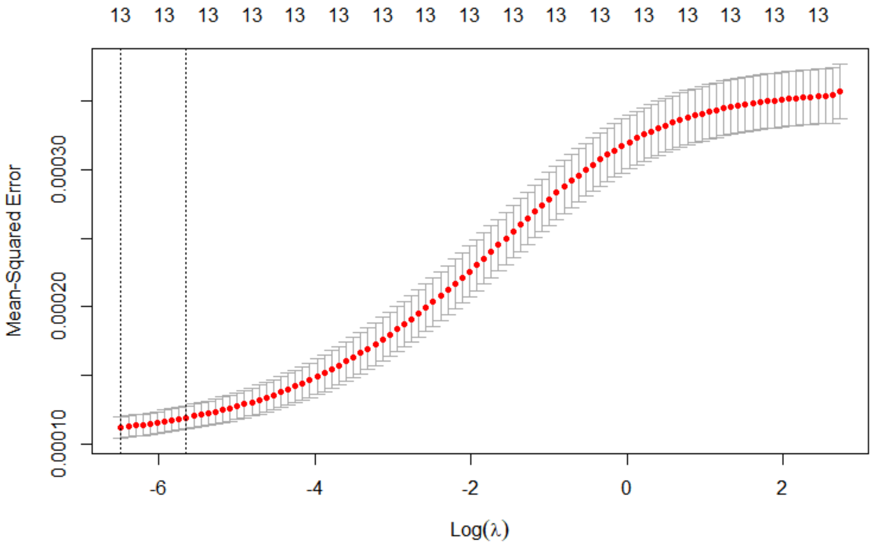


Figure 9: Test MSE for different λ.

The optimal value for λ turned out to be $\lambda_{best} = 0.00153$ and is used to produce the final model. The coefficient estimates obtained for the final model is shown in the following table,

Regressor	Coefficient estimate
(Intercept)	1.109593e+00
age	-2.351671e-04
weight	-2.372756e-06
height	6.941755e-04
neck	8.574702e-04
chest	-2.788383e-04
abdomen	-1.162754e-03
hip	2.222674e-05
thigh	-4.321560e-04
knee	-1.474968e-04
ankle	3.636851e-05
biceps	-1.548729e-04
forearm	-7.410275e-04
wrist	3.606373e-03

A ridge trace plot, figure 10, present how the coefficient estimates change as a result of increasing λ.

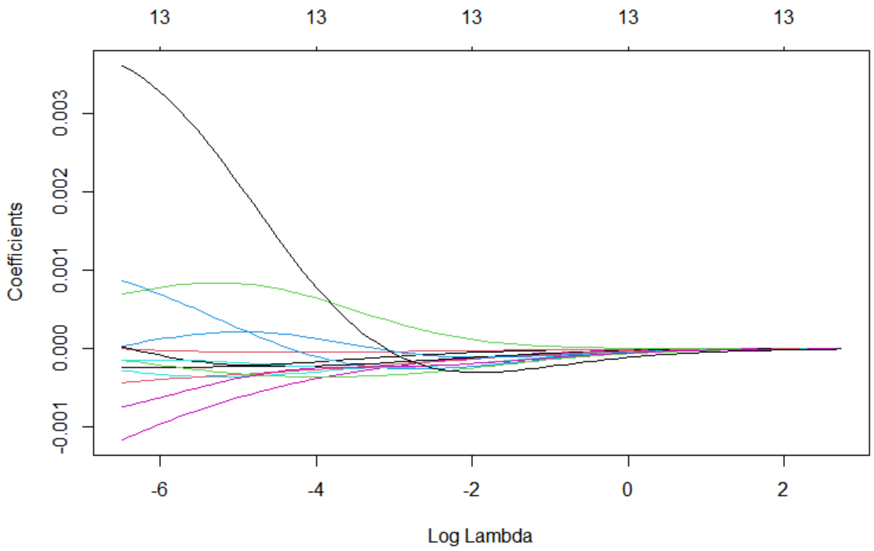


Figure 10: Coefficient estimates for different λ.

The final model produces is used make predictions and then calculate the adjusted R^2_{adj}-value. The value obtained was $R^2_{adj} = 0.698$.

5.4 Variable selection

5.4.1 Forward/backward elimination

The coefficient estimates obtained by **forward** stepwise selection is presented in the following table. The variables ignored is: height, chest, knee, and ankle.

Regressor	**Coefficient estimate**
(Intercept)	1.147
age	-1.29e-04
weight	2.09e-04
height	-
neck	11.50e-04
chest	-
abdomen	-21.63e-04
hip	5.35e-04
thigh	-6.52e-04
knee	-
ankle	-
biceps	-3.95e-04
forearm	-10.36e-04
wrist	33.41e-04

The coefficient estimates obtained by **backward** stepwise selection is presented in the following table. The variables ignored is: height, chest, ankle, and biceps. This shows that the two methods, forward and backward elimination, result in two slightly different models.

Regressor	**Coefficient estimate**
(Intercept)	1.146
age	-1.35e-04
weight	1.89e-04
height	-
neck	10.95e-04
chest	-
abdomen	-21.50e-04
hip	5.61e-04
thigh	-7.32e-04
knee	-1.47e-04
ankle	-
biceps	-
forearm	-11.78e-04
wrist	33.21e-04

Here we can tell the models have similar coefficients and amounts of coefficients, even if backward has one more. This could be done via just manually checking the, p-values, or adjusting the thresholds.

5.4.2 BIC,C(p), Adjusted R^2

As for the amount of coefficients we got ,9 for the adjusted R^2 was 9, 4 for BIC and 8 for the C_p as can be seen in the plot.

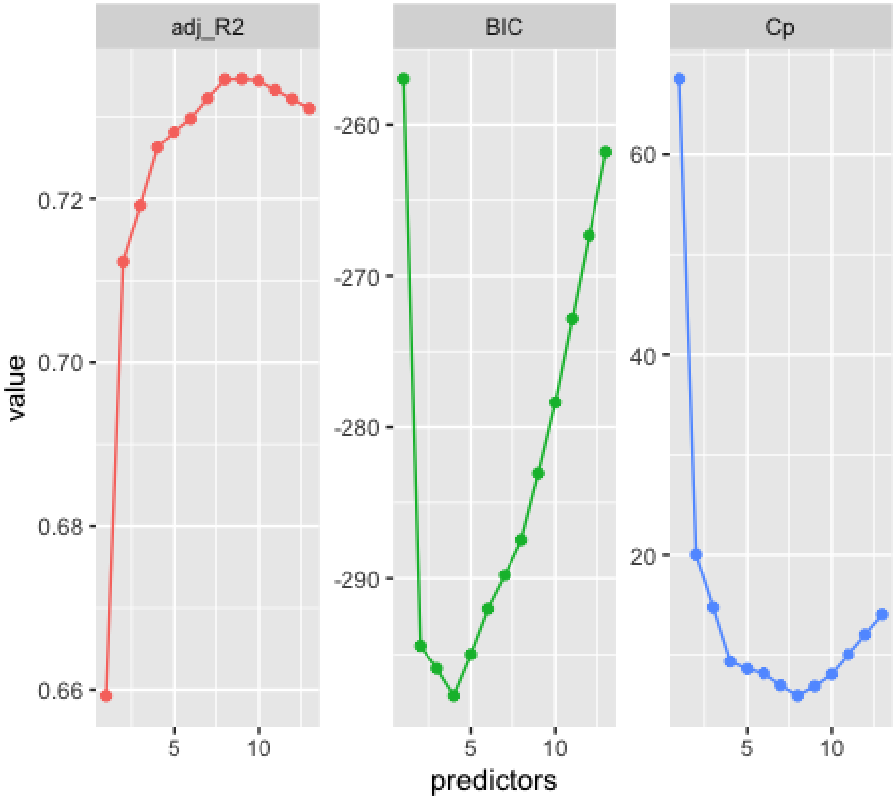


Figure 11: Adjusted R-squared,BIC,CP

When selecting trade-offs we are obviously dealing with various trade-offs between what we want our model to be "best" in. If we are primarily just interested in just minimizing MSE for example then we would select the coefficients in the adjusted R^2 as they are equivalent in that regard. Whilst the BIC value indeed takes into consideration the number of coefficients use, so it could be preferred for a simpler model. Lastly, Mallows works similarly but punishes the model less for the added coefficients.

5.5 Bootstrap

Firstly, since we are not experiencing Heteroskedasticity we can apply Bootstrapping. Moreover for the bootstrap we decided to use the model with 8 coefficients that was suggested by backward selection, seeing as it seems like a reasonable amount of coefficients by the previous section.

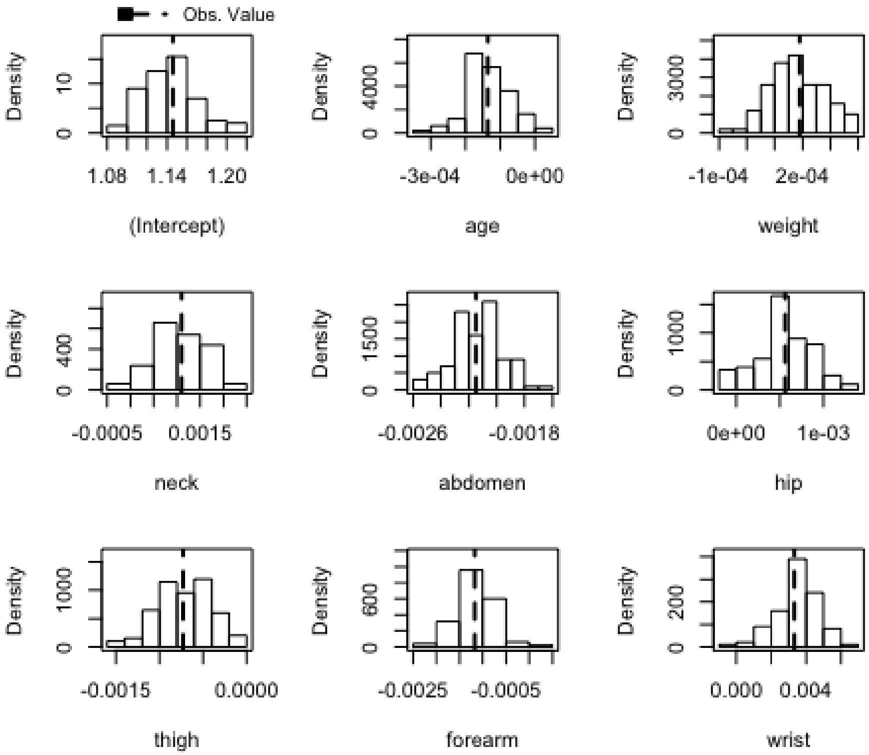


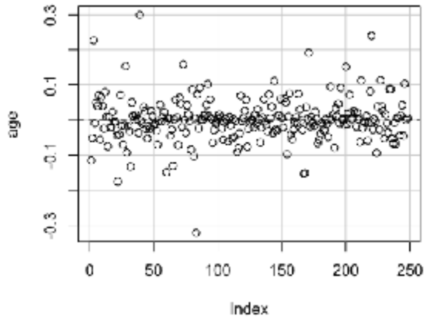
Figure 12: Bootstrapping, n=100

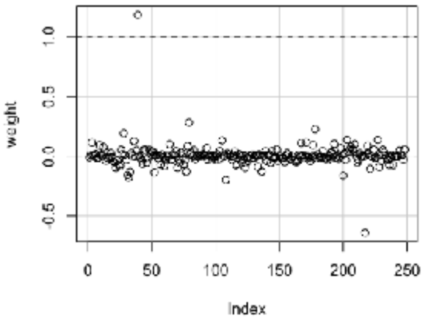
In the various confidence intervals we can tell its quite Gaussian, and indeed extending the sampling from 100 to 1000 would improve it.

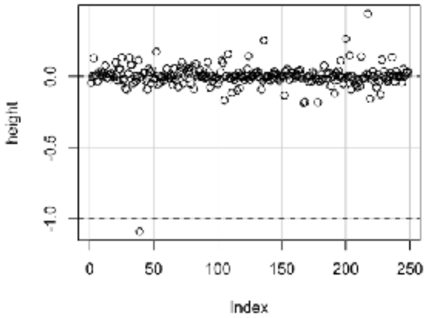
6 Conclusion

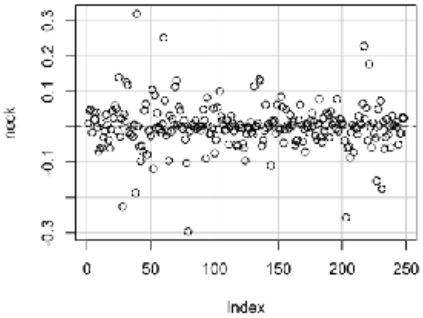
A full LM model is not necessarily the best here, instead we a model with fewer coefficients would be ideal since multicollinearity can be found. Moreover there is great skepticism for certain values that could be potential outliers, however seeing as we don't know a great deal as to how data was attained it is difficult to remove these.

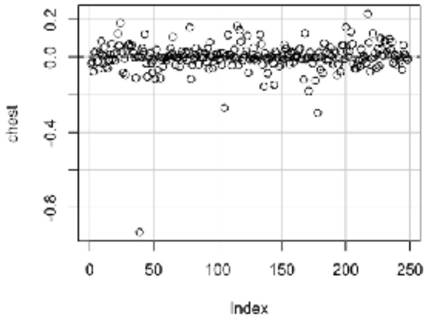
7 Appendix

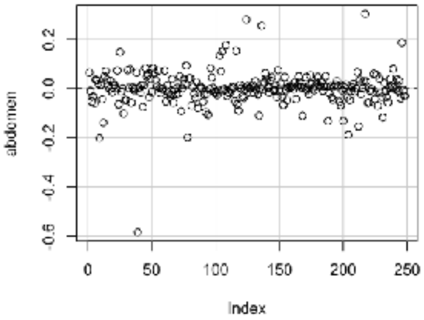

age
-0.3
-0.1
0.1
0.3
0
50
100
150
200
250
Index

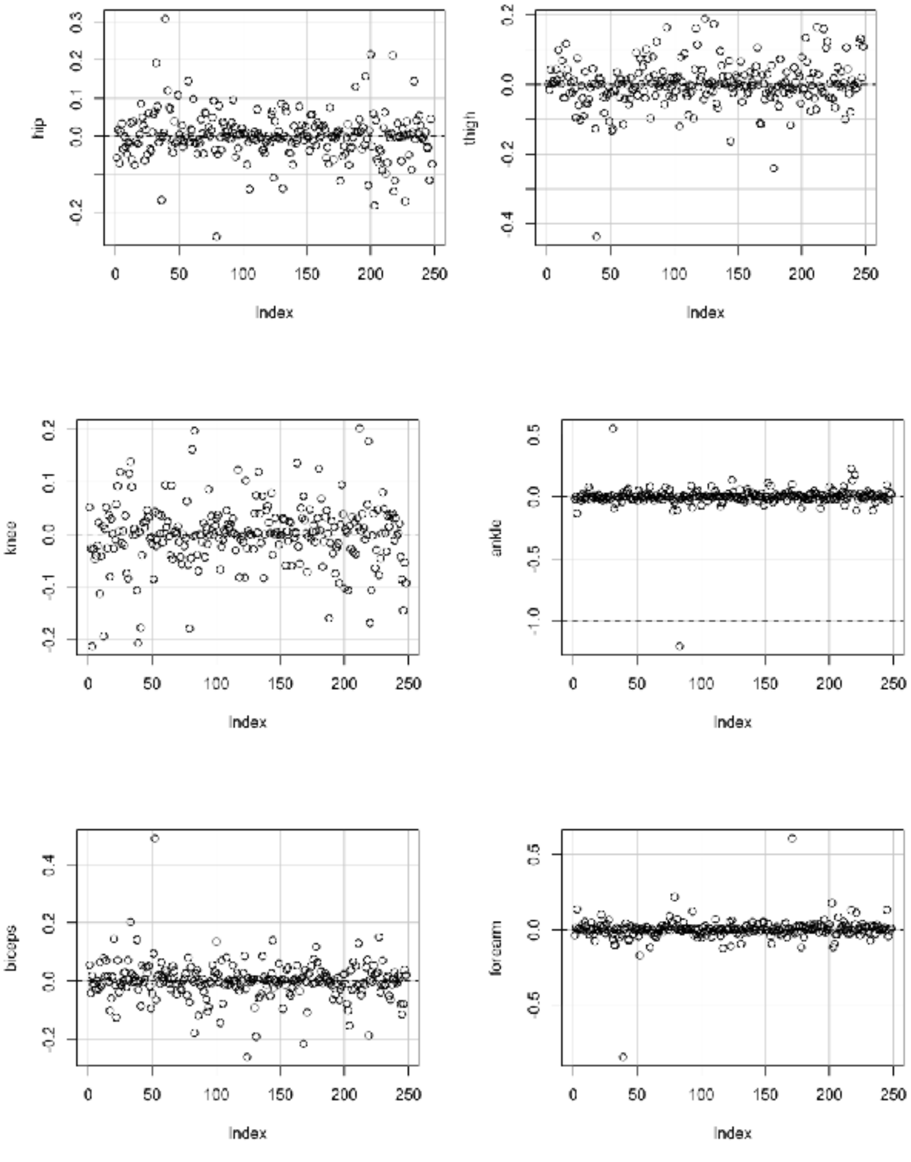

weight
-0.5
0.0
0.5
1.0
0
50
100
150
200
250
Index

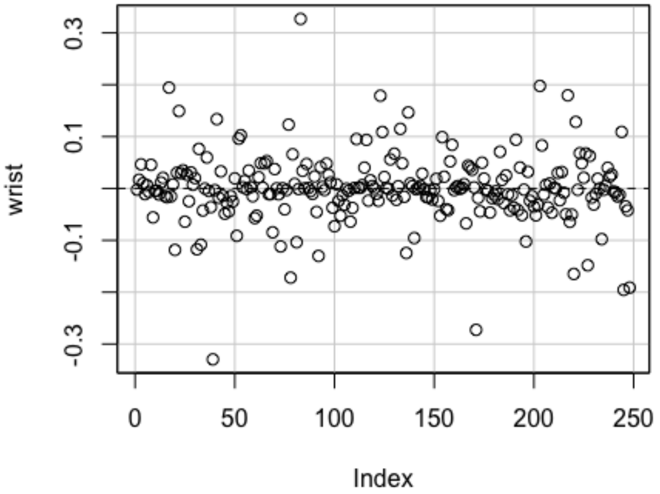

height
-1.0
-0.5
0.0
0
50
100
150
200
250
Index


neck
-0.3
-0.1
0.1
0.2
0.3
0
50
100
150
200
250
Index


chest
-0.8
-0.4
0.0
0.2
0
50
100
150
200
250
Index


abdomen
-0.6
-0.4
-0.2
0.0
0.2
0
50
100
150
200
250
Index


hip
0.3
0.2
0.1
0.0
-0.2
thigh
0.2
0.0
-0.2
-0.4
knee
0.2
0.1
0.0
-0.1
-0.2
ankle
0.5
0.0
-0.5
-1.0
biceps
0.4
0.2
0.0
-0.2
forearm
0.5
0.0
-0.5
0
50
100
150
200
250
Index


0.3
0.1
-0.1
-0.3
wrist
0
50
100
150
200
250
Index