Report I - SF2930 Regression Analysis

William Gan, 980619-9510 Knut Dahlström, 981226-1551

February 12, 2023

1 Project 1

We decided to work with **Scenario I** and used the data set for women's BFM in order to improve our large-sample regression model to fit the relevant data. By using large-sample regression we assumed that $p < n$, where n is the number of observations and p our predictors.
The template for the project report along with instructions are given below:

Introduction and project goals

The goal of this project is to develop a regression model in order to predict the response variable which in this case is our Body fat mass (BFM) or also called DEXFAT-index, denoted with y. Obesity is a major issue regarding public health and has been increasing lately, therefore by analyzing the factors that affects obesity could be of big interest to prevent it.[1] Our regressor variables represents age and different physiological measurements of the women's body, see Table 1.

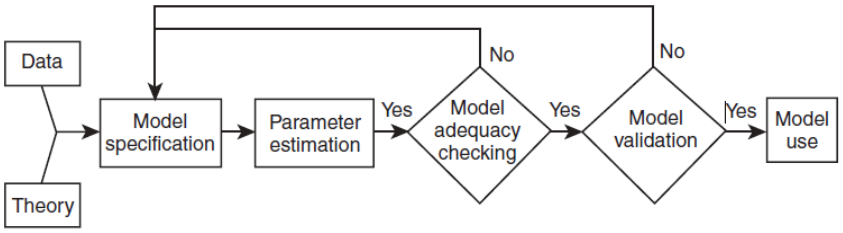


Figure 1: Regression building model

The project process follows the model presented in Figure 1, as such the data was first analysed with the assumptions of linearity with well fitted variables and without any extreme data points (outliers). This model was discarded through further steps in this model and the underlying assumptions were therefore restated multiple times.[2]

Analyses and Model Development

The model development mainly consisted of residual analysis, outlier diagnostics, transformations and variable selection. Additionally, in order to assess the model bootstrapping was used combined with multicollinear diagnostics. For instance, variable selection was used with the variables in the table below.

Variable	Name	Measurement
y	DEXfat/BFM	Density
x_1	Age	Age
x_2	Waistcirc	Waist Circumference
x_3	Hipcirc	Hip Circumference
x_4	Elbowbreadth	Breadth of Elbow
x_5	Kneebreadth	Breadth of Knee
x_6	anthro3a	Transformed with log
x_7	anthro3b	Transformed with log
x_8	anthro3c	Transformed with log
x_9	anthro4	Transformed with log

Table 1: Relevant variables used to improve our model.

The general model that was constructed followed the matrix equation:

$$y = \mathbf{X}\beta + \epsilon, \tag{1}$$

where $\mathbf{X}$ is the matrix containing our data with the independent variables, β the vector consisting of the coefficients of the model $\beta_1, \beta_2, ..., \beta_9$ and ϵ the vector of the random errors $\epsilon_1, \epsilon_2, ..., \epsilon_9$. The coefficients for the regression model β has the least square estimator $\hat{\beta}$ which is:

$$\hat{\beta} = (\mathbf{X'X})^{-1}\mathbf{X'}y \tag{2}$$

Due to the usage of the estimator $\hat{\beta}$ in the resulting model, it is necessary to add the errors ϵ in order for the model to be valid since these are "estimators" and compensate for the differences, which the exact solution would yield. The matrix with our independent regressors looked as following:

$$\mathbf{X} = \begin{bmatrix} 1 & \mathrm{x}_{11} & \mathrm{x}_{12} & ... & \mathrm{x}_{19} \\ 1 & \mathrm{x}_{21} & \mathrm{x}_{22} & ... & \mathrm{x}_{29} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & \mathrm{x}_{n1} & \mathrm{x}_{n2} & ... & \mathrm{x}_{n9} \end{bmatrix} \tag{3}$$

and as seen in equation 2, $\mathbf{X'}$ is the transpose of $\mathbf{X}$.

1.1 Model Development Steps

We proceeded with our regression model in the following steps:

1. Residual analysis
2. Outlier diagnostics
3. Transformations of regressors and response variable
4. Multicolinearity diagnostics
5. Model assessment

Some of the steps were not used but were necessary to mention in order to understand how the resulting model was concluded. A more general illustration of the model development process can be seen in Figure 1

1.2 Model Assumptions

In order to fit a linear regression model to the data some assumption had to be made. The **first** assumption that was assumed was that the relation between the response variable y and our regressors is linear. The second assumption was that the random errors ϵ have expectation zero $E(\epsilon) = 0$ and constant variance $\sigma^2 = constant$. The third assumption was that the random errors ϵ are normally distributed, indicating that $\epsilon \in N(0, \sigma^2)$.

2 Residual Analysis

Residual analyses are preformed to find the discrepancies between the fit of the model and the data. Thus the definition is the predicted value of y subtracted from the actual value of y as in equation 4.

$$r_i = y_i - \hat{y}_i \tag{4}$$

By this measure an observer can find outliers and get a sense if the fit truly fits the assumptions of normality. The residuals further have some useful properties, such as a zero mean and a well defined estimated variance, see equation (5). In Figure 2, the regular residual plot can be seen for the different data points. Potential outliers can also be seen which is discussed later in the report.

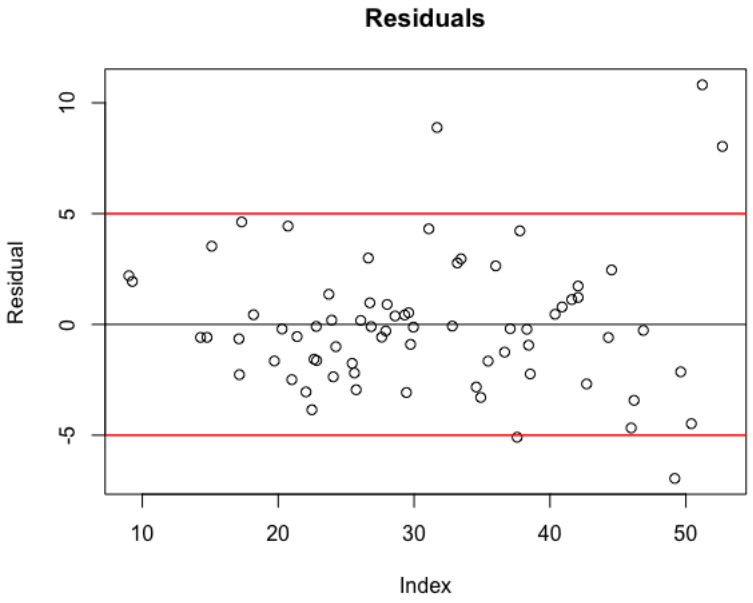


Figure 2: The residuals plotted against the data points.

The standard variance of the residuals are denoted M_{res} and are calculated as in (5) . As the numerator in equation (5) implies, the residuals have $n - p$ degrees of freedom. Residuals for a model with many degrees of freedom are more reliable to use for adequacy checking for a model than if there are few degrees of freedom.

$$MS_{Res} = \frac{\sum_{i=1}^{n}(e_i - \vec{e})}{n - p} \tag{5}$$

2.1 Hat matrix

The hat matrix is useful for identifying hidden interpolaton, and its diagonal is also used to scale the residuals and is created as follows.[2]

$$\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}\boldsymbol{X})^{-1}\boldsymbol{X} \tag{6}$$

In the residual analysis of the womens BFM data set, there is a close to zero mean for most of the response variable spectrum and only a few potential outliers. However, from this information alone there is little indication of the outliers ultimate influence on the regression.[2]. By looking at Figure 2 the points that might be outliers lie above the first red line are 73, 87 and 94.

2.2 Standardized Residuals

To further investigate if there were any outliers, standardized residual analysis was used. The standardized residual for one point is defined as the deviation between the observed point and the expected value or how significant it is compared to the χ^2 value. Mathematically it is defined as the residual divided by it's estimated standard deviation MS_{Res}, see equation 7.[2]

$$rs_i = \frac{r_i}{\sqrt{MS_{Res}}}, \quad i = 1, 2, ..., n. \tag{7}$$

As seen in Figure 3, the same points appear to deviate a lot as was seen in Figure 2. The points with indices 73, 87 and 94 approximately have 3 as deviation and could be very unusual for the dataset. Additionally, there is another potential outlier below the second red line, which makes the total of four potential outliers.

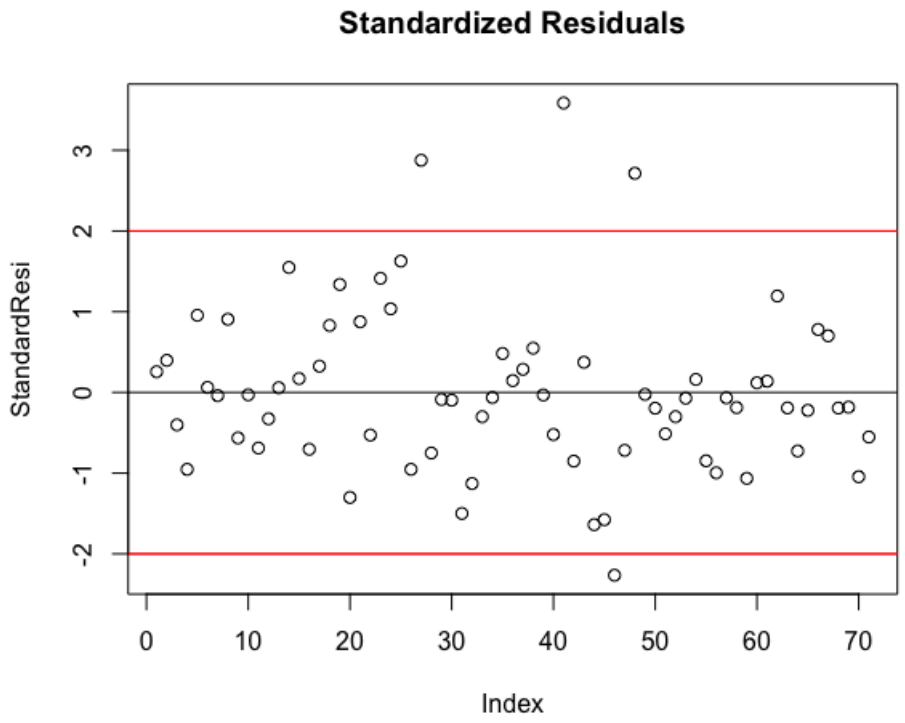


Figure 3: The standardized residuals plotted against the data points.

The conditions for determining whether a data point is an outlier or not, depends if the standardized residual is large or not. For instance, standardized residuals that deviate with ± 3 might indicate that the related data point is quite unusual.[2]

2.3 Studentized Residuals

The scaling of the standardized residuals has the downside of only using the mean estimate MS_{Res} as seen in equation (7). Thus residuals of data on the extremes of the response variable spectrum are in small datasets severly overestimated. To compensate the regular studentized residual model adjusts the error by individual scaling. This is achieved by scaling the numerator, MS_{Res}, by the distance of the independent variables from their mean. This scale factor corresponds to the diagonal elements of $\boldsymbol{H}$ from equation (6).[2]

$$r_i = \frac{e_i}{\sqrt{MS_{Res}(1 - [\frac{1}{n} + \frac{(x_i - \vec{x})^2}{S_{XX}}])}} = \frac{e_i}{\sqrt{MS_{Res}(1 - h_{ii})}} \tag{8}$$

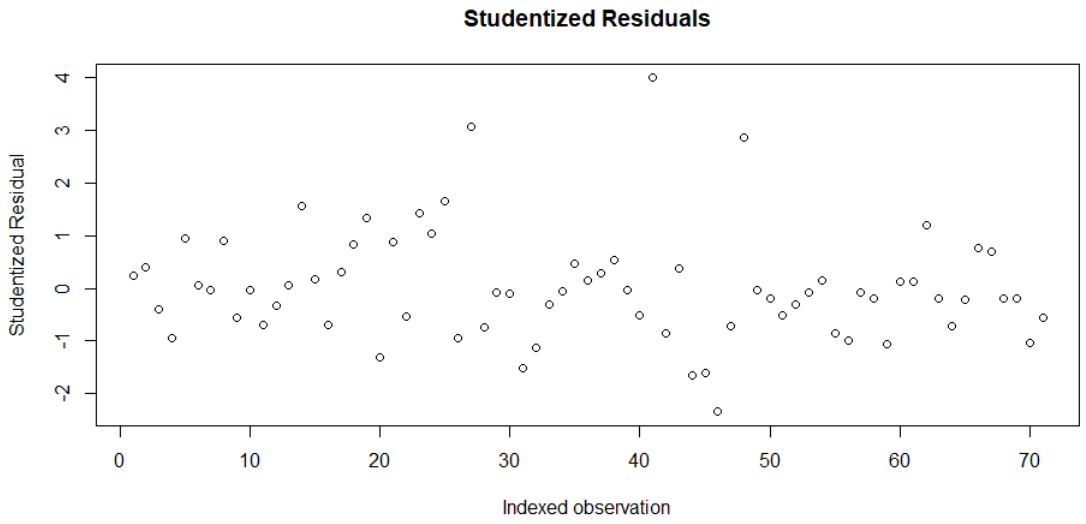


Figure 4: The studentized residuals plotted against their index in the data set

In figure 4 the studentized residuals are not ordered by their expected response value, instead they are ordered by their index. We can therefore recognise four potential outliers which were same as mentioned before, see Table 2.

2.4 Predicted Residual Error Sum of Squares

The predicted residual error sum of squares (**PRESS**) metric denoted $e_{(i)}$ in equation (9), is a measure of the discrepancy of the data and the regression given that the current data point was excluded from the initial regression. These too are scaled residuals as the discrepancy can be described as a function of the corresponding diagonal $\boldsymbol{H}$ matrix element as presented in equation (6) and the standard residual. [2]

$$e_{(i)} = y_i - \hat{y}_{(i)} = \frac{e_i}{1 - h_{ii}} \tag{9}$$

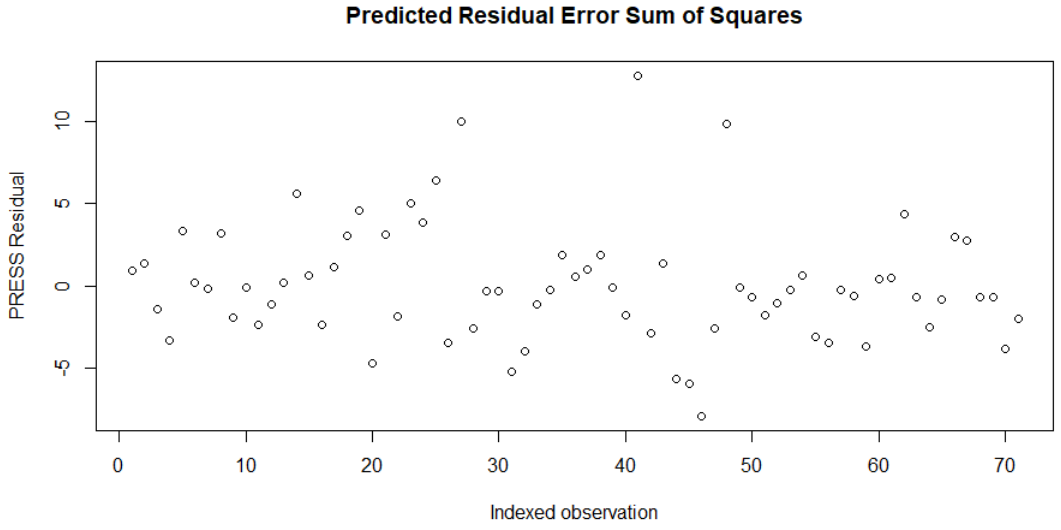


Figure 5: The PRESS metric plotted against the data points.

In figure 5 we find a roughly zero mean, and residuals that have a roughly normal spread.

2.5 Externally Studentized Residuals

The externally studentized residuals or commonly known as r-studentized residuals are a combination of PRESS and the studentized residuals, denoted t_i in equation (10). Scaled residuals using MS_{Res} as seen in equation (7) are called internally scaled, since MS_{Res} is calculated using the entire data set. S_i as calculated in equation (11) is similar to PRESS in the aspect that it removes the current data point before estimating the variance, and scaling using it are thus called externally scaled. R-studentized residuals use both using the individual variance S_i and the $\boldsymbol{H}$ diagonal elements calculated as in equation (6) to scale the residual. This thus accounts for distance from mean in both dependent and independent variables, thus reveling outliers when compared to ordinary studentized residuals.[2]

$$t_i = \frac{e_i}{\sqrt{S_i^2(1-h_{ii})}} \tag{10}$$

$$S_i^2 = \frac{(n-p)MS_{Res} - \frac{e_i^2}{1-h_{ii}}}{n-p-1} \tag{11}$$

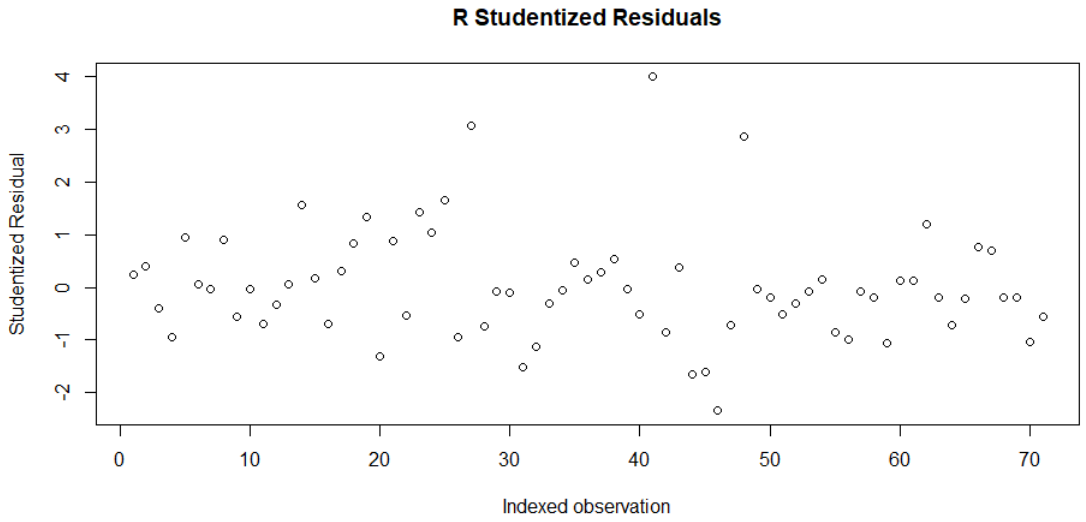


Figure 6: The R-studentized residuals plotted against the data points.

When compared to the studentized residuals there is little to no discernible difference in the plots.

2.6 Outliers in regression

One of the main reasons to use and compare scaled residual analyses is to identify true outliers. Outliers are datapoints which differ greatly from the rest of the dataset. As these extreme values have great influence on the regression, one should investigate their validity and in certain cases they should be discarded entierly.[2]

Data nr from WHO	73	87	94

Table 2: Table of outliers from the residual analyses

In the residual analyses we found 3 data points which were potential outliers in every case. Since we don't know under what circumstances these were recorded we cannot discard these without question.

3 Regression diagnostics and handling of outliers

3.1 Influence Points

Points that affect the regression model due to extreme y or x values can be referred as "*influence points*". These points affect the regression model by shifting the values of the coefficients β of the model. This is because the point deviates from the expected value. Measures to evaluate whether the influence points are valid or not can be done be either removing them and analyzing the model. However, these points could still be relevant for the model even if they are extreme from the expected subset. To determine if an observation is an influence point or not, Cook's Distance or analyzing the covariance ratio could be applied.[2]

3.2 Leverage Points

In regression analysis a "*leverage point*" can be considered as a point or observation that deviates a lot from the subset of the other data points that fit the regression model. Even if the point deviates a lot it is still in the regression model, this is called a "*leverage point*". The significance of these points may not affect the regression model that much, i.e our coefficients β, but could easily affect the statistic metrics when summarizing the data set. A common general rule is if one of the diagonal elements in the "hat matrix", described in equation 6, is larger than $\frac{2p}{n}$ then it could be considered as a leverage point. Shortly, points with extreme values of x are said to have high leverage. meaning that these points have a higher probability to affect the regression line. These points could also be influence points.[2]

3.3 Cook's distance

As mentioned in section "*3.1 Influence Points*", points that are extreme in x-space and y-space are considered to be enough influencial to affect the stability of the regression model. Cook's Distance was used to determine whether a point had these extreme value in x-space and y-space.

Cook's Distance is defined as the square distance of the least-square approximation when using all points n with $\hat{\beta}$ compared to the least-square estimation when the influence point/points i are not taken into account $\hat{\beta}_i$. The general formula for Cook's Distance can be seen in equation 12.[2]

$$D_i = (\mathbf{M}, c) = \frac{(\hat{\beta}_i - \hat{\beta})\mathbf{M}(\beta_i - \hat{\beta})}{c}, \quad i = 1, 2, ..., n. \tag{12}$$

The $\mathbf{M}$ parameter is set to be $\mathbf{M} = \mathbf{X'X}$ and $c = pMS_{res}$. Thus, leading to that large D_i values have a significant impact on the least-square estiamtion $\hat{\beta}$. Another way to rewrite equation 12 can be seen below:

$$D_i = \frac{r_i^2}{p}\frac{Var(\hat{y}_i)}{Var(e_i)} = \frac{r_i^2}{p}\frac{h_{ii}}{1 - h_{ii}}, \quad i = 1, 2, ..., n. \tag{13}$$

The fraction $\frac{r_i^2}{p}$ is the square of the studentized residual for the ith point and multiplied with $\frac{h_{ii}}{1-h_{ii}}$ gives us D_i. The resulting fraction can be summarized as the distance from the vector x_i to the center of the remaining data set that does not contain the influence point i. The quantity D_i therefore measures the influence that an observation has by taking the residual magnitude and the position in x-space into account.

Applying Cook's Distance to our dataset we got the following result:

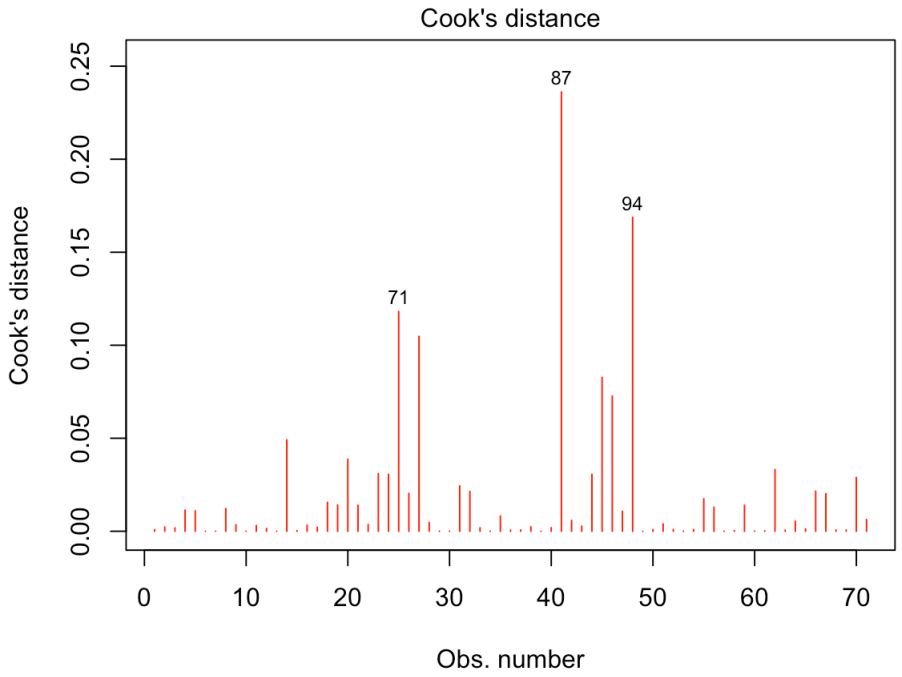


Figure 7: Plot of Cook's Distance for BFM Women.

As seen in Figure 7, the same points appear to have the largest Cook's Distance in the figure. The points 94 and 87 are the obvious ones but point 71 seem to have a great magnitude, similar to point 73, as well. However since point 71 did not show any obvious deviations in the residual analysis it is disgarded as an outlier. The points 73, 87 and 94 were removed from the dataset.

4 Multicoliniarity Diagnostics and Treatments

When producing the $\hat{\beta}$ matrix one uses the form of $(\boldsymbol{X'X})^{-1}$. In the case that $\boldsymbol{X}$ has some linear dependency, $\boldsymbol{X'X}$ becomes singular. Singular matrices can not be inverted, thus one can not regress variables that are linearly dependent. Furthermroe, close to linearly dependent variable matrices result in $\boldsymbol{X'X}$ with close to singular qualities. Multicolinearity is thus a measure of how close to exact linear dependency a variable is.[2]

4.1 Analysis of Variance Inflation Factors

One way of determining if $\boldsymbol{X'X}$ is close to linearly dependent is to check its diagonal elements. These are called variance inflation factors (VIF) and any factors over 5 or 10 indicate significant multicollinearity.

$$VIF_i = (\boldsymbol{X'X})_{ii} \tag{14}$$

Variable	Name	VIF_i
x_1	Age	1.207378
x_2	Waistcirc	5.770604
x_3	Hipcirc	5.121272
x_4	Elbowbreadth	1.416972
x_5	Kneebreadth	2.853228
x_6	anthro3a	39.471164
x_7	anthro3b	49.100539
x_8	anthro3c	9.037833
x_9	anthro4	111.261438

Table 3: Variance Inflation Factors

As seen in Table 3, six factors exceed 5 which is cause for concern and further, 3 significantly exceed 10. It is therefore important that some measure against multicollinearity is taken.

4.2 Eigensystem analysis

Another way to examine multicollinearity is to analyse the condition values (κ_i) as seen in equation (16). The λ_i refer to the eigenvalues of the matrix $\boldsymbol{X'X}$ as calculated in equation (15), where λ_{max} is the largest eigenvalue.

$$\lambda\vec{\boldsymbol{v}} = \boldsymbol{X'X}\vec{\boldsymbol{v}} \tag{15}$$

$$\boldsymbol{\kappa}_i = \frac{\lambda_{max}}{\lambda_i} \tag{16}$$

Variable	Name	λ_i	κ_i
x_1	Age	154876.2215	1
x_2	Waistcirc	11199.61793	138.2870581
x_3	Hipcirc	2340.894063	661.6114071
x_4	Elbowbreadth	51.22901939	30232.12689
x_5	Kneebreadth	30.97508008	50000.26509
x_6	anthro3a	13.81574029	112101.2832
x_7	anthro3b	2.563147626	60424.22994
x_8	anthro3c	0.5877981905	263485.3663
x_9	anthro4	0.1337720159	1157762.485

Table 4: Eigensystem Analysis: condition factors

The naive approach to the fit resulted in variables with high condition numbers as seen in table 4. To indicate troubling multicolinearity κ_{max} only has to have the magnitude of 1000. [2] As κ_{max} is an order of magnitude larger than that it is a good indication that the regression has severe problems and does require adjusting.

To combat multicolinearity one usually omits variables and remakes the regression, another opportunity would be to create a ridge regression. Variable omission is treated later in this report, and ridge regression was deemed outside the scope.

5 Transformations

We assumed that there is a linear relationship between the response variable y and our regressors, which needs to be further validated. Sometimes the linear relationship might not apply and thus making a regression model more complicated to construct. Since it is easier to work with a linear regression model sometimes a suitable "*transformation*" is enough to create the linear relationship between the response variable and the regressors.[2]

Transformations can be done on several regressors and the response variable. For instance, a logarithmic transformation of a variable is simply taking the logarithm of the value for that specific variable. This Other mathematic manipulations accounted as transformations could be taking the square and square root of an variable.

In order to decide if a transformation was necessary, the studentized residuals where plotted against the fitted values $\hat{y}_i$, see Figure 8.

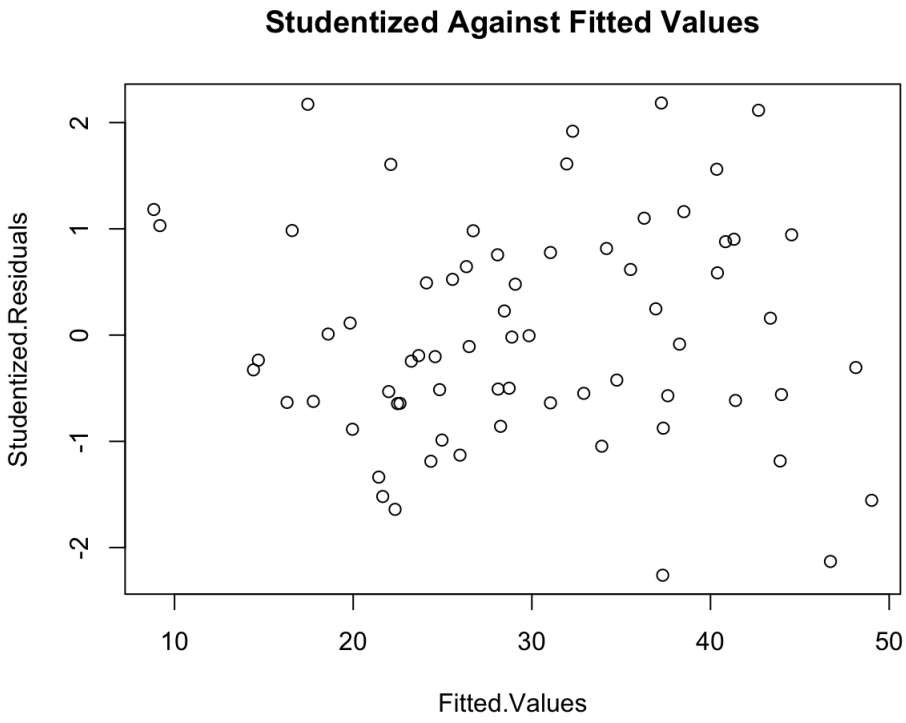


Figure 8: Studentized residuals plotted against the fitted values.

The different points in Figure 8 look random and does not form a specific pattern. This usually indicates that a transformation is not necessary. Note that the dataset retrieved from WHO about women BFM had some regressors that were already transformed.

5.1 Box Cox Method

The Box-Cox transformation is used on a non-normal dependent variable to transform it into a normal shape. We earlier highlighted the importance of the assumption of normality. Maximum likelihood is used to estimate the parameters of the regression model and λ. For instance, a power tranformation is used on the reponse variable, meaning that y is raised to the power of λ. By determining λ one can decide if a power transformation of kind y^{λ} is necessary.[2]

One way to estimate λ is to construct a confidence interval. By applying maximum likelihood to the regression model, equation 17 needs to be maximized.

$$L(\lambda) = -\frac{1}{2} n ln(SS_{res}(\lambda)) \tag{17}$$

The approximate 100(1-α) percent confidence interval for λ include the values of λ that satisfy equation 18.

$$L(\hat{\lambda}) - L(\lambda) \leq \frac{\chi^2_{\alpha,1}}{2n} \tag{18}$$

Then by plotting $L(\lambda)$ against λ and plotting a horizontal line at height (see equation 19) the confidence interval is constructed.

$$L(\hat{\lambda}) - \frac{\chi^2_{\alpha,1}}{2} \tag{19}$$

This horizontal line then cuts $L(\lambda)$ on two points on the λ axis which indicates the span of the confidence interval. By setting $\alpha = 0.95$ the following plot was attained:

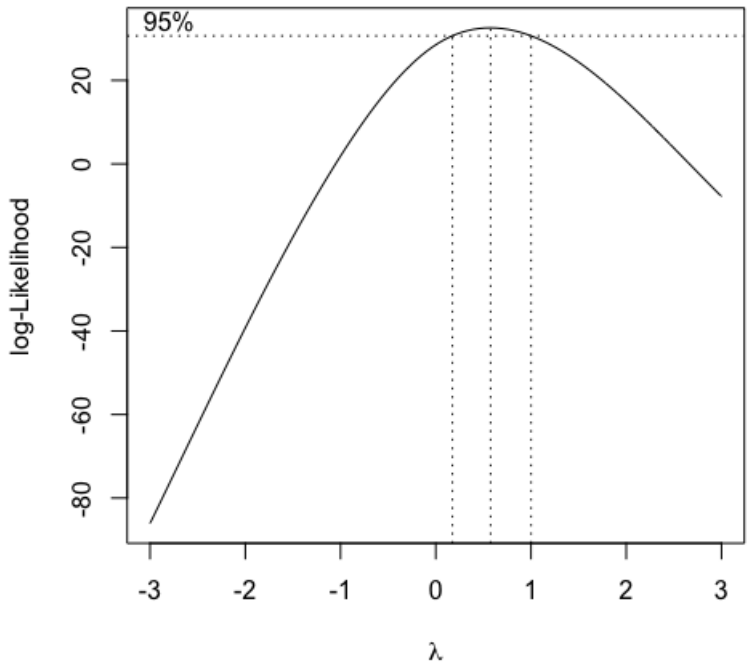


Figure 9: The figure displays the confidence interval for λ.

As seen in Figure 9 the CI for λ is just over zero and one. Conclusively, we decided to go with $\lambda = 1$ and did not transform any variables. This was simulated with a 95 % CI.

6 Normal Distribution Analysis

To determine whether the data sample is normally distributed or not, the sample quantiles was plotted against the theoretical quantiles for a normal distribution. As for the model with the outliers 73, 87 and 94, the plot can be seen in Figure 10.

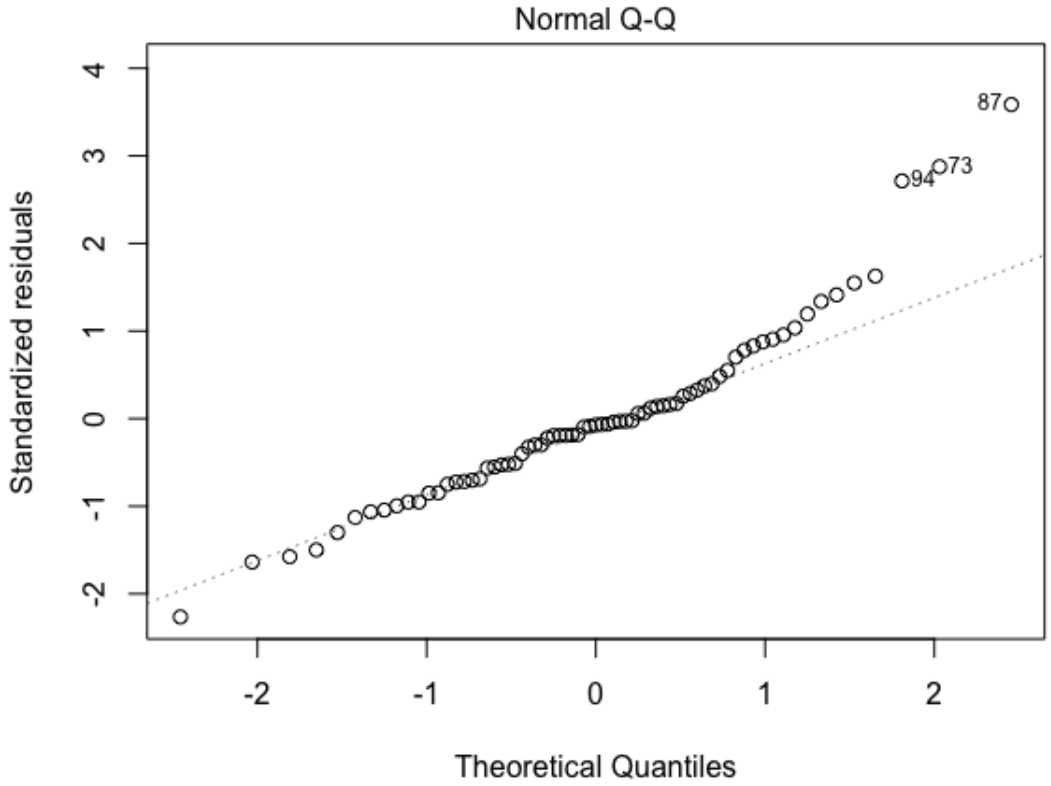


Figure 10: Normal qqplot for our data points.

As seen in Figure 10 some points do not lie on the theoretical line for a normal distribution, in particular the three outliers deviated a lot from the theoretical line. Due to the their extreme values both in the residual analysis and in the normal plot, the outliers 73,

87 and 94 were removed and plotted again, see Figure 11.

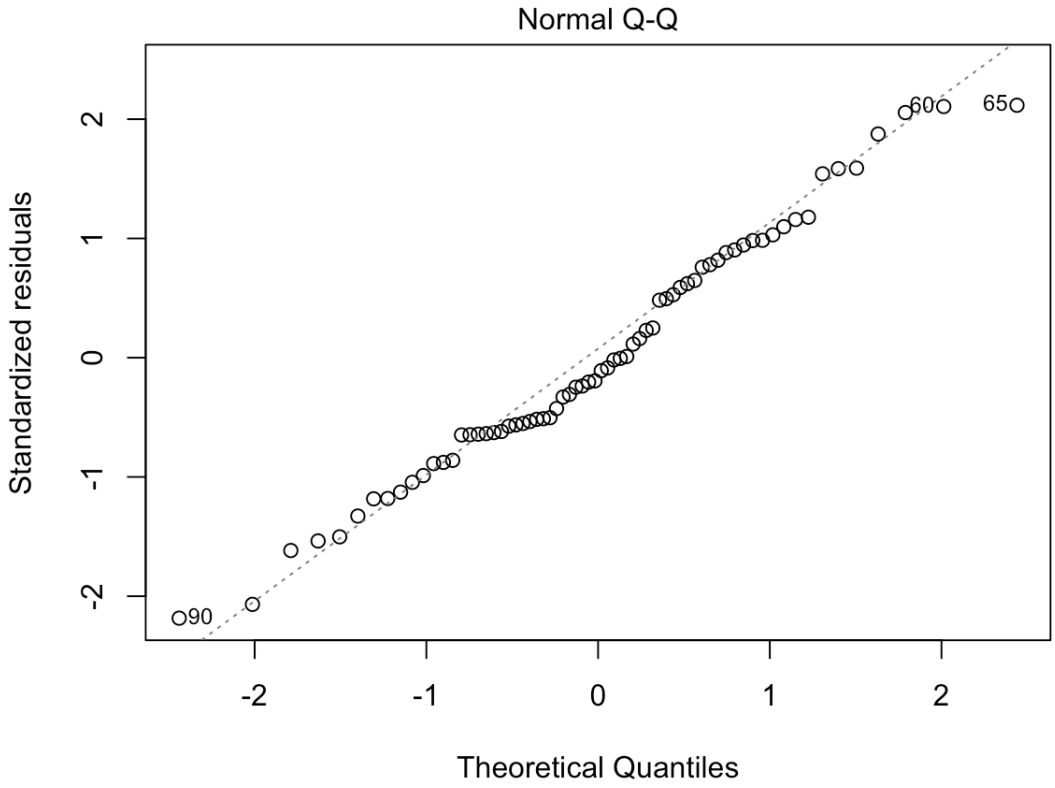


Figure 11: Normal qqplot without outliers as our data points.

The removal of the outliers increased the normality of the model, thus, confirming that the previous points that were assumed to be outliers were true. This conclusion fits the assumption that the errors ϵ are normally distributed, as mentioned in section "*1.2 Model Assumptions*".

7 Variable Selection

The approach of analyzing the residuals, any multicollinearity between the variables and a transformation evaluation is a classic approach in terms of model building. What needs to be taken into account as well is the variables or regressors selected for the model. Multicollinear diagnostics opens the room for determining if some regressors should be left out from the model in order to increase accuracy of the prediction, however this could also have the opposite effect if the removed regressors might be relevant for the model. The new subset of regressors is then applied in a new model and compared to the old one to

detect any unnecessary variables, which was done in this case.

7.1 Criterias for Keeping a Variable

In order to decide which variables to keep in the model, several methods where used to determine if the specific variable fulfilled the criteria to be kept. In particular, the criterias that were evaluated are "*Adjusted* R^2", "*Mallow's* C_P", "*Bayesian Information Criterion*" and "*Akaike Information Criterion*".

7.1.1 Adjusted R^2

The adjusted R^2 denoted $R^2_{Adj.p}$ is originally derived from the coefficient of determination. The latter case describes the proportion of the variance in the response variable that is predictable from the regressors. Why the adjusted R^2 is used instead of R^2 is because it tells if an added predictor variable actually improves the model by looking if the added variable improves the model more than expected by chance. The adjusted R^2 is generally defined as seen in equation 20.[2]

$$R^2_{Adj.p} = 1 - (\frac{n-1}{n-p})(1 - R^2_p) \tag{20}$$

7.1.2 Mallow's C_p

Mallow's C_P criterion addresses the issue of overfitting. In it's core the metric of goodness-of-fit is tested and a smaller C_p-value usually indicates that a model fits more accurately. The C_p-value is defined as:

$$C_p = \frac{SS_{Res}(p)}{\hat{\sigma}^2} \tag{21}$$

As seen in equation 21, the C_p-value depends on the mean square prediction error which aims to be minimized.[2]

7.1.3 Bayesian Information Criterion

Another criterion that tests goodness-of-fit or the improvement to a model when adding predictor variables is the Bayesian Information Criterion or BIC. This is done by introducing an extra penalty term depending on the number of parameters in the model.[2] Usually a smaller BIC-value is preferred and it is defined as:

$$BIC = -2ln(L) + pln(n) \tag{22}$$

7.1.4 Akaike Information Criterion

The Akaike Information Criterion is derived from maximizing the expected entropy of the model. The entropy of a model can be seen as a measure of the expected information. Similar to the BIC-value an extra penalty term is added therefore AIC can be seen as a penalized log-likelihood measure and is defined as:

$$AIC = -2ln(L) + 2p \tag{23}$$

The log-likelihood function is L and p is the number of parameters in the model.[2]

7.2 Possible Regressions

In this step all the possible regression models were evaluated. The different models were created by starting with one predictor/regressor and iteratively adding another regressor so it becomes two and so on. Then each model is evaluated by looking at the stastical metrics to determine which variables are relevant or not. For instance, the adjusted R^2 was calculated. The number of models that are evaluated depends on the number of variables by 2^k, where k is the number of variables.[2]

Var.	1	2	3	4	5	6	7	8	9
Interc.	1	1	1	1	1	1	1	1	1
x_1	0	0	0	0	0	0	1	1	1
x_2	1	0	1	1	1	1	1	1	1
x_3	0	1	1	1	1	1	1	1	1
x_4	0	0	0	0	0	0	0	0	1
x_5	0	0	0	1	1	1	1	1	1
x_6	0	0	0	0	0	0	0	1	1
x_7	0	1	1	1	1	1	1	1	1
x_8	0	0	0	0	0	1	1	1	1
x_9	0	0	0	0	1	1	1	1	1
rsq	0.830	0.919	0.945	0.947	0.949	0.950	0.951	0.951	0.951
adjr2	0.827	0.916	0.942	0.944	0.944	0.945	0.945	0.944	0.943
C_p	137.3	34.2	5.04	4.81	4.90	5.71	6.30	8.09	10.0
BIC	-112.0	-158.0	-180.4	-178.6	-176.5	-173.6	-170.9	-167.0	162.9
rss	1104.6	528.0	356.8	344.6	334.1	327.6	319.9	318.7	318.2

Table 5: Each column describes how many regressors and which ones were used with its corresponding statistical metrics. The table is constructed by using the dataset with the outliers removed.

By investigating Table 5, one can decide which model to use by looking at the stastical metrics for each model. In this case by choosing the regressors "*waistcirc, hipcirc*" and "*anthro3b*" gives the best summary statistics for the model. However in order to decide the total number of regressors to select for an improved model further investigation was done.

7.3 Cross Validation

Another way of determining the effectiveness of the model in terms of variable selection and estimation of predictiveness is "Cross Validation". There are different types of cross validation such as k-fold cross validation which was used in this model. The first step of k-fold cross validation is to divide the dataset into k equally large subsets, where one of the subset will be the validation set for testing the model and the rest training sets. The validation set is used as a reference after the training sets are "trained" or fitted on to the model. Everytime the training set is compared to the validation set a so called validation error or mean square error can be determined, since this is done k-times the mean of these validation errors are calculated and used as the estimate error. This can then be used to conclude which regressors to use.[2]

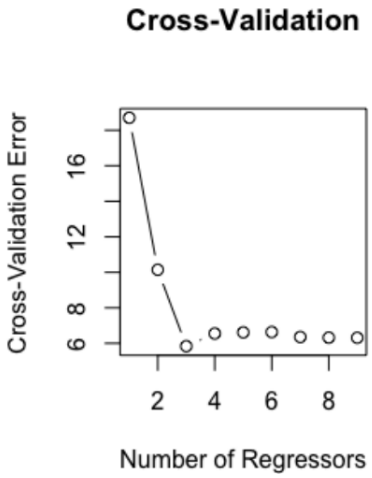


Figure 12: K-fold cross-validation.

The most common values for k are $k = 5, 7$ and 10, where $k = 10$ is chosen in this case. The aim is to find a balance between the computation time when k is large and the accuracy of the method when k does not have a too low value. As expected from analyzing the variables from section "*7.2 Possible Regressions*", the regressors x_2, x_3 and x_7 have the smallest errors.[2]

8 Model Assessment

8.1 Bootstrapping

Bootstrapping residuals are computation intensive methods to validate regression. Similar to the various studentizing methods, this method removes samples of the original data to compute new βs. Studentizing the data, only removes them one-by-one, and really only uses the hat matrix to preform these calculations. Thus the differing sizes and combinations of the samples removed by bootstrapping increases the workload immensely.

8.1.1 Bootstrapping regression

By the means of bootstrapping the linear model proposed trough cross validation one can find if it is a good model. I.e. the comparison between the model based on the regressors x_2, x_3 and x_7 and its bootstrapped counterpart, reveals if the proposed betas are accurate or not. The bootstrap compared to the regression as seen in table 6, was re-sampled randomly 1000 times. The proposed coefficients are identical as far as 7 decimal places, which can be said as that there is very little *bias* in the bootstrap. This is a good indication that the new model is accurate.

Name	Intercept	waistcirc	hipcirc	anthro3b
Variable	x_0	x_2	x_3	x_7
Bootstrap	-58.7400426870	0.2439055010	0.3559851371	7.0431938712
Regression	-58.7400426870	0.2439055010	0.3559851371	7.0431938712

Table 6: Bootstrapping Regression

8.1.2 Bootstrapping Residuals

The residual analysis of the bootstrapped regression produced similar results as one would expect from the observations from the regression analysis. The residual results were so similar that they produced no discernible difference in any of the points, to such a degree that one would have to increase the number of significant figures to more than 7 decimal places. This should be enough to give a strong indication that the new linear model is somewhat accurate.

8.1.3 The Percentile Method

The final assessment through bootstrapping used in this report was through checking the bootstrap confidence intervals. Asessmentwise intervals indicating large uncertainty are indications that the regression would be of poor quality, and a small interval a indication of a good regression. Furthermore a confidence interval that includes 0, would indicate that there is uncertainty to if the model adequately can describe the dependent variable.

Name	Intercept	waistcirc	hipcirc	anthro3b
Variable	x_0	x_2	x_3	x_7
2.5 %	-67.045402624105	0.13676488362310	0.23714234509193	5.5798326175170
97.5 %	-51.914460961969	0.33450946653259	0.47312531695896	9.6603081108623

Table 7: Bootstrapping Confidence Interval 95%

As seen in table 7 the 95% confidence interval indicates that the model has significance, as the interval excludes 0. The confidence interval with confidence of 95% for the coefficients have some variation, although its quite small.So therefore we can conclude that the model proposed in cross validation is valid.

8.1.4 Scale location

Scale location is checked for heteroscedasticity. A heteroescedastic model would have residuals dependent on the fitted values. Since the data set is quite small and centered, one would not expect the plot to be exactly flat. The slight variation amongst the fitted value therefore is expected. Further the endpoints are of the same approximate value, which further strengthens the assertion that the data has normally distributed residuals.

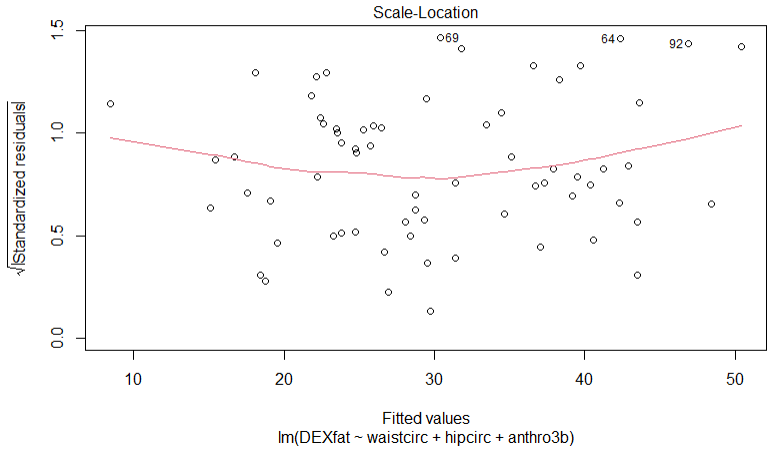


Figure 13: The scaled residuals plotted against the data points.

The figure above is the scaled on the data without the outliers and only with the predictors x_2, x_3 and x_7.

8.2 Other Methods

It should be noted that there are some more methods to validate data. The best method would be to collect more data, so since that was out of the question the validation methods were quite few. Other than bootstrapping, one can also manually divide the data, so that there is some test set, that one then tests the predictive power of the regression on. This requires somewhat large sets of data to be done efficiently, since removing chunks from small datasets could reduce the predictive power of the regression. So for this set with 71 points of data was deemed to small for such procedures.

9 Results

The result of the final model, after removing outliers and some regressors, was on the following form:

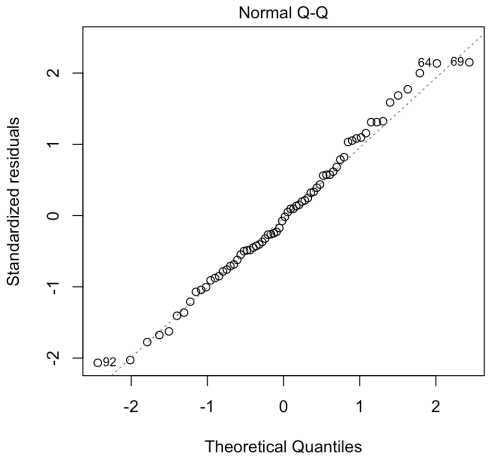
$$y = X_2\hat{\beta}_2 + X_3\hat{\beta}_3 + X_7\hat{\beta}_7 + \hat{\beta}_0 + \epsilon. \tag{24}$$

Thus, implying that the regressors that were conclusively taken into account were x_2, x_3 and x_7 with it's corresponding coefficient $\hat{\beta}_2$, $\hat{\beta}_3$, $\hat{\beta}_7$ and the intercept $\hat{\beta}_0$. The values of the coefficients can be seen in Figure 14, which were $\hat{\beta}_2 = 0.24$, $\hat{\beta}_3 = 0.36$, $\hat{\beta}_7 = 7.04$ and $\hat{\beta}_0 = -58.74$.

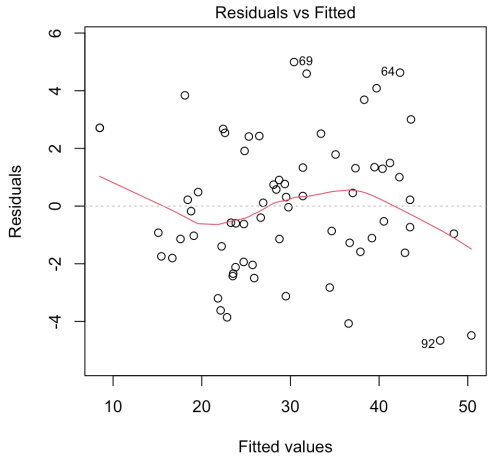
```
Coefficients:
                    Estimate       Std. Error   t value   Pr(>|t|)
(Intercept) -58.740042687035   3.404133510737 -17.25550 < 2.22e-16 ***
waistcirc     0.243905501004   0.044022095305   5.54052 6.0801e-07 ***
hipcirc       0.355985137105   0.053590158916   6.64273 7.8468e-09 ***
anthro3b      7.043193871240   0.835024873886   8.43471 5.5331e-12 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.3612660862 on 64 degrees of freedom
Multiple R-squared:  0.94505068783,     Adjusted R-squared:  0.94247493882
F-statistic: 366.90325218 on 3 and 64 DF,  p-value: < 2.22044605e-16
```

Figure 14: Summary of the regression model.

Furthermore, the quantiles were plotted again seen in Figure 15a and the residuals in Figure 15b to also display the constant variance of the error ϵ.



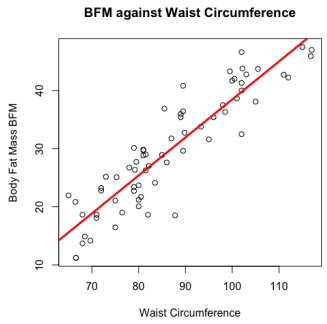
(a) The final normality plot.



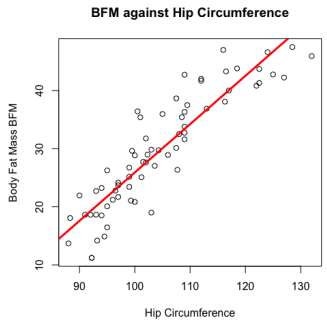
(b) Figure showing constant variance.

Figure 15: The following figure shows the resulting regression evaluations applied on the resulting data.

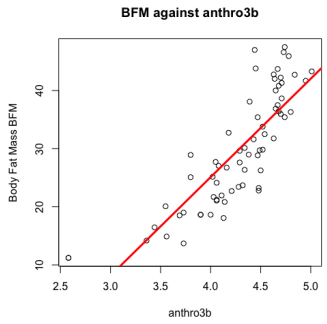
In the figures below the final results are presented. The linear regression model is plotted together with the data points for the three different predictor variables.



(a) The BFM plotted against the first regressor x_2.



(b) The BFM plotted against the second regressor x_3.



(c) The BFM plotted against the third regressor x_7.

Figure 16: The following figure shows the resulting regression model applied on the resulting data.

10 Conclusion

The final model that was concluded can be seen in section "*9 Results*", which was a linear regression model with intercept, three regressor variables taken into account out of nine and three outliers removed from the dataset. The final model seemed to fit the data pretty well compared to the earlier models. The outliers were detected by doing a thorough residual analysis of different residuals and strongly argued for that the points removed were outliers. The variable selection phase involved multicollinear analysis, cross-validation and all possible regressions. These concluded that x_2, x_3 and x_7 were strongly affecting the model, however, adding the other predictors that were removed could also improve the model if combined with other procedures for analyzing the regression model. In conclusion, the final model is significantly more accurate than the original this can be seen by comparing Figure 15 to Figures 10, 11 and 8. In these figures the normality for the errors improves as the different measures are taken, as for the residuals the variance seems more constant in the final figure.

10.1 Future Work

Some investigation areas for this regression model and project:

- Remove points that are on the edge of being outliers such as point 93.
- Iteratively add regressors to see if the model improves.

These are just suggestions but one could of course consider taking other measures to investigate improvements.

References

[1] World Health Organization. Obesity: preventing and managing the global epidemic (Report). p. 1–2. Retrieved February 12, 2023. World Health Organization, 2000.

[2] Montgomery, D.C., Peck, E.A., Vining, G.G. Introduction to Linear Regression Analysis. 5th edition. Wiley, 2012

[3] Belsey, Kuh, Welsch. Regression Diagnostics: Identifying Influential Data and Sources of Collinearity, 1980