Report I - SF2930 Regression Analysis

David Haapanen, 010813-3794 Gustav Karlander, 020317-3299

February 19, 2025

Project 1

Scenario I: – Body fat assessments was chosen for this project. The model presented in the report model were made using the given the data set labeled 'bodyfatmen.csv' containing 252 observations of 14 variables regarding physical measurements of of men.

Introduction and project goals

Obesity increases the risk for chronic diseases such as diabetes, cardiovascular diseases and cancer. This has previously only been a concern for high-income countries but obesity is now rising in low- and middle-income countries. Thus from a medical stand point there is an interest of having reliable methods for identify people with excess fat.

Body mass index (BMI) has before been widely used for this manner but has proven to be an ineffective predictor for assessing this, instead it would be more beneficial to look at a persons body fat mass (BFM). There exist complex and expensive methods that can asses peoples BFM including X-ray densitometry, hydrodensitometry etc. However it would be more practical to have a cheaper and accessible option for this.

Thus the goal for this project is to develop a multiple linear regression model (MLRM) and asses its performance. The model will use anthropometric measurements (waist circumference, skin-fold thicknes etc.) as regressors for predicting the BFM as the single response variable. The steps concerning the development of the final model will be discussed in this report which will consider:

- Model assumptions and validation
- Identifying leverage and influential points with purpose of handling outliers
- Possible variable transformation if needed
- Assessment of Multicollinearity

- Variable selection
- Performance assessment

Analyses and model development

Instructions: For the specified scenario, focus on the corresponding data set (at least one of the two data sets for **Scenario I:**) and report (display and comment on) all the steps of the regression analysis following thoroughly the instructions provided in the project description.

Data set

The given data set 'bodyfatmen.csv' contained, without exclusion of any data points, 252 observations of 14 different anthropometric measurements of men. These were specifed as:

- y (g/cm^3): Body density measured by hydrodensitometry.
- x_1 (years): Age of the individual.
- x_2 (lbs): Weight of the individual.
- x_3 (inches): Height of the individual.
- x_4 (cm): Neck circumference.
- x_5 (cm): Chest circumference.
- x_6 (cm): Abdomen circumference.
- x_7 (cm): Hip circumference.
- x_8 (cm): Thigh circumference.
- x_9 (cm): Knee circumference.
- x_{10} (cm): Ankle circumference.
- x_{11} (cm): Biceps circumference.
- x_{12} (cm): Forearm circumference.
- x_{13} (cm): Wrist circumference.

General Multiple Linear Regression Model

A general MLRM with y_j as the response variable and regressor variable x_{ij} corresponding to the regression coefficients β_i for $i = 1, ..., k$ at observation j can be described as by

$$y_i = \beta_0 + \beta_1 x_{1j} + ... + \beta_k x_{kj} + \epsilon_j$$

, where ϵ_j is the error term. Given that one has n observations the model can be expressed with matrix notation as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

Here $\mathbf{y} \in \mathbb{R}^{n\times 1}$ is the response vector, $\mathbf{X} \in \mathbb{R}^{n\times(k+1)}$ is the model matrix where the first column consists of ones for the intercept term. Moreso $\boldsymbol{\beta} \in \mathbb{R}^{(k+1)\times 1}$ is the vector of regression coefficients and $\boldsymbol{\epsilon} \in \mathbb{R}^{(k+1)\times 1}$ is the vector of error terms for each observation.

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \quad \mathbf{X} = \begin{bmatrix} 1 & x_{11} & x_{21} & \dots & x_{k1} \\ 1 & x_{12} & x_{22} & \dots & x_{k2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{1n} & x_{2n} & \dots & x_{kn} \end{bmatrix},$$

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix}, \quad \boldsymbol{\epsilon} = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}.$$

When applying the least squares method on this system of equations to obtain the least-squares estimators $\hat{\boldsymbol{\beta}} = \left[\hat{\beta}_0, \hat{\beta}_1, \cdots, \hat{\beta}_k\right]^T$. there are som assumptions being made, namely

- $\mathbb{E}[\epsilon] = \mathbf{0}$
- $\mathrm{Cov}(\epsilon) = \sigma^2\mathbf{I}$

Thus it is very important to determine of the assumptions are not violated when constructing a MLRM with least-squares method.

Initial MLRM

Using every observation and regressor variable provided from the data set a MLRM was made using the least-squares method to determine the least squares estimators $\hat{\boldsymbol{\beta}}$. This was considered as the full model, since the model takes into account for all of the 13 regressor variables. Thus for a given observation from the data set the response variable y was the BFM value whilst $x_1, x_2, ..., x_{13}$ was the regressor variables. To summarize the full model

was made using $n = 252$ observations with $k = 13$ regressor variables in order to estimate $p = k + 1 = 14$ coeffient estimates, including the intercept.

The full model showed to have $R^2 = 0.731$ and $R^2_{adj} = 0.7238$, whilst the residual standard error was 0.01 on 238 degrees of freedom. Furthermore a global F-test was done on the full model with significance level $\alpha = 0.05$. The p-value was evaluated as $2.2 \cdot 10^{-16}$, whilst the F-statistic for the full model was 51.6. Hence $H_0 : \beta_1 = \beta_2 = ... = \beta_k = 0$ was ruled out against $H_1 : \beta_1 \neq 0$ for at least one j, indicating that at there is a linear relationship between at least one regressor variable and the response variable.

Evaluation of Assumptions

The assumptions made was that the error term for every observation is normally distributed with mean zero, hence that $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$ for $i = 1, ..., n$ where n is the amount of observations used. In order too assess that these assumptions were true residual analysis were conducted consisting of

- Evaluation of normal distribution using Q-Q plot
- Evaluation if error terms have constant variance with residual vs. fitted values plots.

A Q-Q plot of the full model can be seen below in fig. 1, which shows some slight tailing but not large enough to indicate that the assumption that the used data follows a normal distribution is violated. Since fig. 1 shows that the data points follows a straight line it was concluded that the assumption that the data follows a normal distribution was not violated.

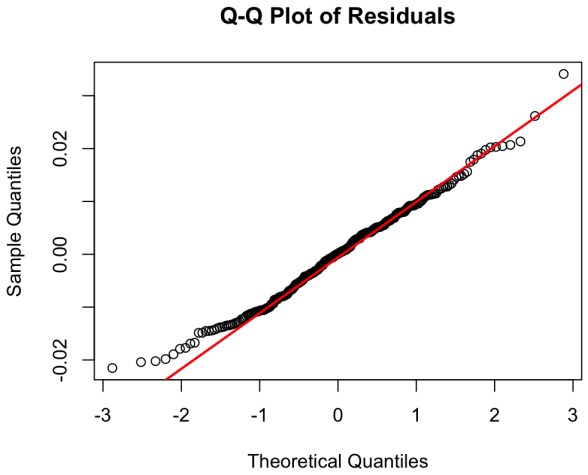

Figure 1: Q-Q plot of the full model

Furthermore plots regarding residuals against the fitted values were made. The plots were residuals vs. fitted values seen in fig. 2, as well as studentized residuals vs. fitted values and standardized residuals vs. fitted values which can be seen in fig. 3 and fig. 4, respectively.

All plots show a horizontal band centered around zero. Thus there is no clear indication that the model assumption regarding linearity between response and regressor have been violated, nor that the assumption of constant variance for the error terms was violated. Furthermore the mean of residuals was determined as -6.95266210^{-19} which is very close to 0. It was thus concluded that the assumption of the errors terms having the mean zero had not been violated.

Moreso in fig. 3 and fig. 4 very few residuals fall under or over the ± 2 lines which is promising, since for a normally distributed sample about 95% of the studentized and standardized residuals should fall within this range. Thus giving more indication of the assumption of the error terms being normally distributed.

With all of these observed results it was concluded that the assumptions made for the development of the full model were not violated.

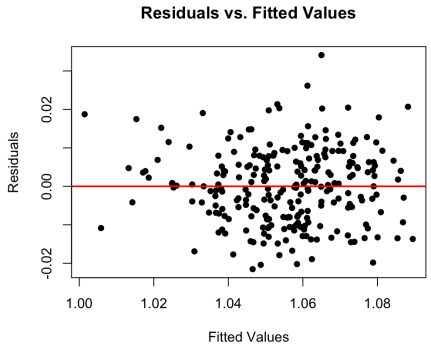

Figure 2: Residuals vs. fitted values from full model

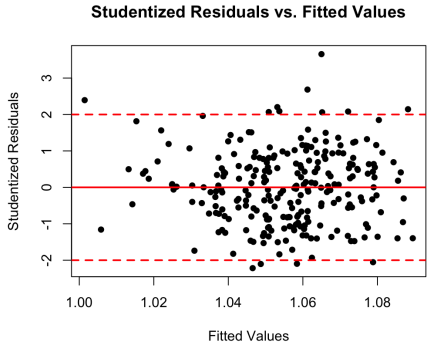

Figure 3: Studentized residuals vs. fitted values from full model

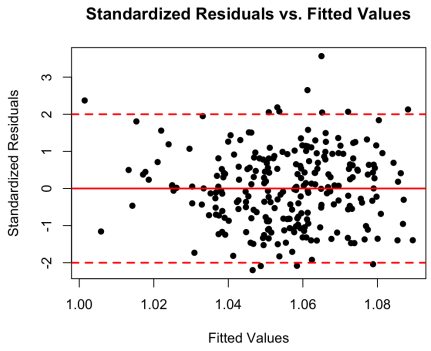

Figure 4: Standardized residuals vs. fitted values from full model

Leverage and Influential points

Leverage points refers to points that are remote in comparison to other observation in $\mathbf{X}$. These points does not necessarily have to affect a model, but can have an impact dependent on how big their corresponding residuals are. For large residuals then a leverage point might be considered an influencial point, which affect the model. It is expected that if these types of points are removed the the model will significantly change.

Thus detecting and analyzing these points is important to assess whether they should be excluded when developing a model depending on the impact that they can have. The things that were considered in this project in order to detect these types of points were

- The diagonol elements of the hat matrix; h_{ii} is the diagonal element at row i and column i of the hat matrix $\mathbf{H}$ defined as

  $$\mathbf{H} = \mathbf{X}(\mathbf{X}^\top \mathbf{X})^{-1}\mathbf{X}^\top$$

  Moreso h_{ii} has the properties of measuring the distance of observation $\mathbf{x}_i$ from the centroid (mean of all predictor values) in x-space. Thus a larger value for h_{ii} indicates that $\mathbf{x}_i$ has more leverage, by convention if $h_{ii} > \frac{2p}{n}$ then $\mathbf{x}_i$ may be a leverage point.

- Cook's distance; D_i quantifies how much the coefficients of $\hat{\boldsymbol{\beta}}$ would change if observation i was removed.

  $$D_i = \frac{r_i^2}{p}\frac{h_{ii}}{1-h_{ii}}$$

  where r_i is the internally studentized residual at i. D_i measures the squared distance between $\hat{\boldsymbol{\beta}}_{(i)}$, the coefficient estimates if observation i is removed from the data set, and $\hat{\boldsymbol{\beta}}$. A larger value of D_i thus give strogner indication that point i may be a influenctial point. By praxis a cutoff rule is that point i is a influencial point if

  $$D_i > F_{\alpha,p,n-p}, \quad \text{for } \alpha \geq 0.5$$

- DFFITS; Evaluates how the removal of observation i will impact the fitted value $\hat{y}_i$

  $$\text{DFFITS}_i = \frac{\hat{y}_i - \hat{y}_{(i)}}{S_{(i)}\sqrt{h_{ii}}} = t_i\frac{\sqrt{h_{ii}}}{\sqrt{1-h_{ii}}} \tag{1}$$

  with $\hat{y}_{(i)}$ being the fitted value excluding observation i and t_i being the externally studentized (R-student) residual. A larger value indicates that observation i has more of an influence on predictions and as praxis the cutoff for considering a observation being influencial is $2\sqrt{\frac{p}{n}}$.

Firstly value for h_{ii} was determined for every observation $\mathbf{x}_i$ with $i = 1, ..., 252$, in order to evaluate the points which had a value $h_{ii} > \frac{2p}{n} = \{p = 14, n = 252\} = \frac{28}{252}$. Each h_{ii} was then plotted for every observation in a leverage plot, where every point with a h_{ii} value then the set threshold was detected. This can be seen below in fig. 5

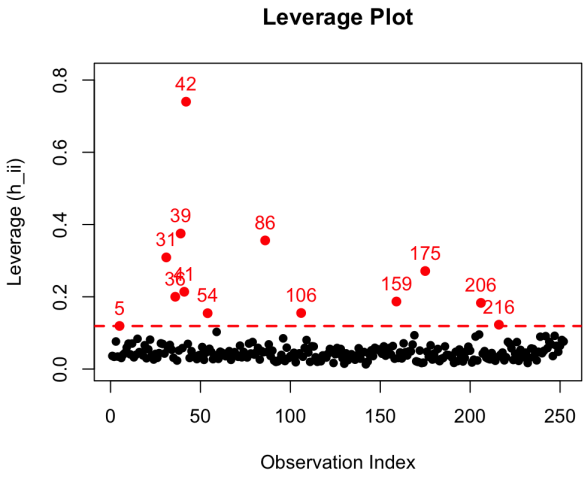

Figure 5: Leverage plot of full model

Then the corresponding studentized residuals from the detected points above the set threshold from fig. 5 was determined, these can seen below in table 1, ordered with regards to magnitude of the residual. This was inorder to get an estimate of potential influential points.

Index	**Studentized Residual**
39	2.39396175
86	-1.33621792
175	-1.18712209
216	-1.15910847
54	1.22855149
31	0.94821285
41	0.44508968
36	-0.46409453
206	-0.44031603
159	-0.19511573
5	-0.13136177
42	-0.05938964
106	-0.03952432

Table 1: Studentized Residuals corresponding to $\mathbf{x}_i$ that had h_{ii} value above set threshold

Next Cook's distance D_i was determined for every observation i and the cutoff threshold of $F_{\alpha,p,n-p}$ using the signficance level $\alpha = 0.5$ aswell as $p = 14$ and $n = 252$. A plot showing each value of D_i for corresponding i was made, including a line representing the cutoff aswell as the indices of the observations showing higher values of D_i. This can be seen in fig. 6

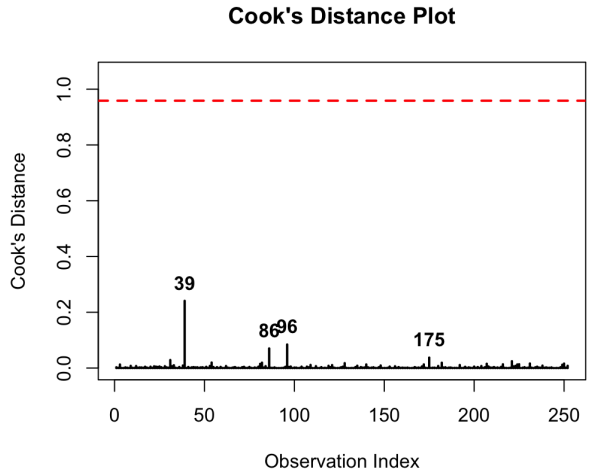

Figure 6: Plot Cook's distance for each observations

It can be seen in fig. 6 that no observation has a value of D_i that exceeds the cutoff, but some points do have a higher value than others. The exact value of Cook's distance for these observations are listed below in table 2

Index	**Cook's Distance**
39	0.24095272
96	0.08434597
86	0.07028532
175	0.03742113

Table 2: Cook's Distance for different data points

Thirdly the value of DFFITS$_i$ was evaluated for every observation i, these were then plotted with the corresponding index of each data point with a line representing the cutoff $\pm 2\sqrt{\frac{14}{252}}$ inserted. This plot can be seen below in fig. 7

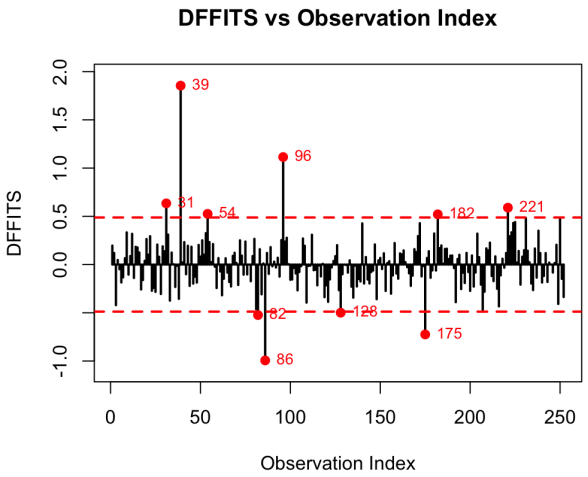

Figure 7: Plot of DFFITS from full model

From fig. 7 we can see that several observations exceed the cutoff, with observation 39 having the largest DFFTIS value.

To summarize table 1 shows that observations 39,86 and 175, ordered in magnitude, are the three points that had the largest magnitude for their corresponding studentized residuals while having h_{ii} values greater than the cutoff threshold. Moreso fig. 6 shows that no observation had a value for Cook's distance exceeding the threshold. But table 2 shows that the points 39, 96, 86 and 175 , ordered with regards to magnitude, had more significant values of Cook's distance in comparision to all other points. Lastly fig. 7 show that several points exceeds the threshold cutoff for their DFFITS value, among these were the points of 39, 86, 96 and 175, which in magnitude exceeds the cutoff most compared to all other points.

From this it was concluded that point 39 can be considered a influencial point and that points 86, 96 and 175 most likely are also quite influencial on the model. Upon more inspection of these points it was noted that point 39 corresponded to quite a high weight of 363.15 lbs and that points 86 corresponded to an age of 67 years. Points 96 and 175 didn't seem to have any large abnormalities. Thus one could argue that point 39 and 86 might be outliers, however it was decided that these wouldn't be removed from the dataset. This was with the intent that the final model should be applicable for men with a higher weight aswell as older age.

Transformation of variables

Since it was concluded that assumptions were not violated, it was concluded that there was no need to transform any variables.

Multicollinearity diagnostics

For this part of the report $\mathbf{X}$ underwent unit length scaling, such that

$$x_{ij}^s = \frac{x_{ij} - \bar{x}_j}{\sqrt{\sum_{i=1}^n (x_{ij} - \bar{x}_j)^2}}, \quad y_i^s = \frac{y_i - \bar{y}}{\sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}}$$

This centers and removes β_0 from the model while each regressors is transformed to have mean zero and variance of one. The transformed matrix can then be expressed as

$$X_s^T X_s = \begin{bmatrix} 1 & r_{12} & \dots & r_{1k} \\ r_{12} & 1 & \dots & r_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ r_{1k} & r_{2k} & \dots & 1 \end{bmatrix}$$

where $r_{ij} = r_{ji}$ are the off-diagonal elements which states the correlation between regressor i and j. For the rest of this section $\mathbf{X}$ refers to $\mathbf{X}_s$

Exact multicollinearity refers to that all columns of X are linearly dependent, but this is rearly the case. However regressors can be nearly linearly dependent hence

$$\sum_{j=1}^{k} t_j X_j \approx 0, \quad \text{for some } t_1, t_2, \dots, t_k$$

This refers to multicollinearity which has several implications.

- $X^\top X$ is ill-conditioned
- $X^\top X \approx 0$
- $(X^\top X)^{-1}$ will have large entries thus making the variance from least squared estimates large since

$$\text{Cov}(\hat{\beta}) = \sigma^2 (X^\top X)^{-1}$$

The effects of multicollinearity is that least squared methods overestimates the $\hat{\beta}$ in comparision to the true β, such that one regressor estimate is far off from the correct value. Thus it is important to consider multicollinearity and try to reduce it in order to have a good model.

The metrics that were considered in order to assess multicollinearity of the full model was

- Assessing off-diagonal elements of the matrix $X^\top X$. If elements denoted as r_{ij} has a magnitude close to one, $|r_{ij}| \approx 1$, then it means that $\mathbf{X}_i$ and $\mathbf{X}_j$ are nearly linearly dependent.

- Variance Inflation Factor;

$$\text{VIF}_j = C_{jj}$$

  where C_{jj} is the diagonal element in $\mathbf{C} = (X^\top X)^{-1}$ which tells how much the variance of $\hat{\beta}_j$ is increased to due multicollinearity. As a rule it is often said that a VIF value greater than 5 or 10 indicates multicollinearity which causes poor estimation of the corresponding coefficients. For this project we considered the rule of a VIF value greater than 10.

- Eigenvalue analysis of $X^\top X$; By determining the eigenvalues of $X^\top X$ and looking at their magnitude one can estimate the level of multicollinearity. Eigenvalues that are smaller indicate greater multicollinearity within $X^\top X$. By getting the largest eigenvalue λ_{max} and the smallest eigenvalue λ_{min} one can assess the so called condition number defined as

$$\kappa = \frac{\lambda_{max}}{\lambda_{min}}$$

  Where the level of multicollinearity can be assessed by the conventions of:
  - $\kappa < 100$, no serious multicollinearity
  - $100 \leq \kappa < 1000$, moderate to serious multicollinearity
  - $1000 \leq \kappa$, severe multicollinearity

  Furthermore by assessing each eigenvalue λ_j for $j = 1, ..., k$ one can compute condition indices

$$\kappa_j = \frac{\lambda_{max}}{\lambda_j}$$

  The number of $1000 \leq \kappa_j$ indicates the amount of nearly linear dependencies between regressor.

Firstly the off-diagonal elements of the correlation matrix $X^\top X$ were quickly assessed. It could be seen that the magnitude of some elements were closer to one than others for example the correlation between x_7 ("hip") and x_2 ("weight") was ≈ 0.94 while the correlation between x_5 ("chest") and x_6 ("abdomen") was ≈ 0.92. The correlation matrix with some of the off-element with higher magnitude highlighted can be seen in fig. 8

	density	age	weight	height	neck	chest	abdomen	hip	thigh	knee	ankle	biceps	forearm	wrist
density	1.00	-0.28	-0.59	0.10	-0.47	-0.68	-0.80	-0.61	-0.55	-0.50	-0.26	-0.49	-0.35	-0.33
age	-0.28	1.00	-0.01	-0.17	0.11	0.18	0.23	-0.05	-0.20	0.02	-0.11	-0.04	-0.09	0.21
weight	-0.59	-0.01	1.00	0.31	0.83	0.89	0.89	0.94	0.87	0.85	0.61	0.80	0.63	0.73
height	0.10	-0.17	0.31	1.00	0.25	0.13	0.09	0.17	0.15	0.29	0.26	0.21	0.23	0.32
neck	-0.47	0.11	0.83	0.25	1.00	0.78	0.75	0.73	0.70	0.67	0.48	0.73	0.62	0.74
chest	-0.68	0.18	0.89	0.13	0.78	1.00	0.92	0.83	0.73	0.72	0.48	0.73	0.58	0.66
abdomen	-0.80	0.23	0.89	0.09	0.75	0.92	1.00	0.87	0.77	0.74	0.45	0.68	0.50	0.62
hip	-0.61	-0.05	0.94	0.17	0.73	0.83	0.87	1.00	0.90	0.82	0.56	0.74	0.55	0.63
thigh	-0.55	-0.20	0.87	0.15	0.70	0.73	0.77	0.90	1.00	0.80	0.54	0.76	0.57	0.56
knee	-0.50	0.02	0.85	0.29	0.67	0.72	0.74	0.82	0.80	1.00	0.61	0.68	0.56	0.66
ankle	-0.26	-0.11	0.61	0.26	0.48	0.48	0.45	0.56	0.54	0.61	1.00	0.48	0.42	0.57
biceps	-0.49	-0.04	0.80	0.21	0.73	0.73	0.68	0.74	0.76	0.68	0.48	1.00	0.68	0.63
forearm	-0.35	-0.09	0.63	0.23	0.62	0.58	0.50	0.55	0.57	0.56	0.42	0.68	1.00	0.59
wrist	-0.33	0.21	0.73	0.32	0.74	0.66	0.62	0.63	0.56	0.66	0.57	0.63	0.59	1.00

Figure 8: Correlation matrix of full model

This indicates that some multicollinearity is prevelent in the model.

Secondly each value of VIF was determined for each regressor and plotted against the set cutoff threshold of 10. This can be seen in fig. 9.

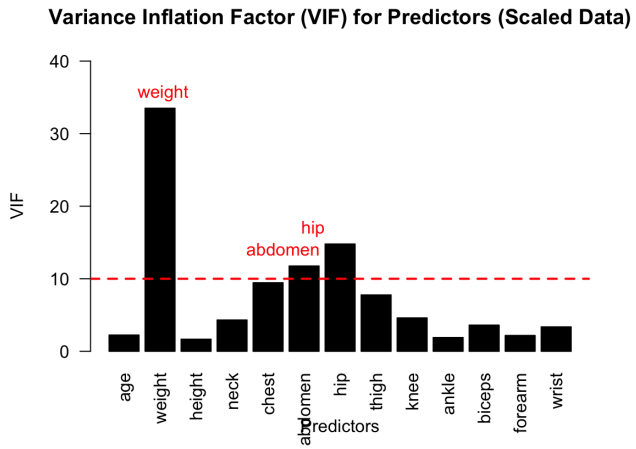

Figure 9: Plot of VIF for each regressor of full model

Moreso the exact VIF values for each regressor can be seen in table 3

Variable	Name	VIF Value
x_2	Weight	33.509
x_7	Hip	14.797
x_6	Abdomen	11.767
x_5	Chest	9.461
x_8	Thigh	7.778
x_9	Knee	4.612
x_4	Neck	4.324
x_{13}	Wrist	3.378
x_{11}	Biceps	3.620
x_1	Age	2.250
x_{12}	Forearm	2.192
x_{10}	Ankle	1.908
x_3	Height	1.675

Table 3: Variance Inflation Factor (VIF) sorted by magnitude

Thus from fig. 9 and table 3 it could be concluded that the regressors of x_2, x_7 and x_6 were affected of multicollinearity. Thus these regressors might be worth removing from the model since they are near-linear dependent of other regressors.

Lastly the eigenvalues of $X^\top X$ were determined, which made it possible to compute the condition number κ and the condition indices for each regressor. The condition number was determined to be $\kappa \approx 340.61$ and the condition indices for every regressor was plotted against the threshold of 100, which can be seen in fig. 10

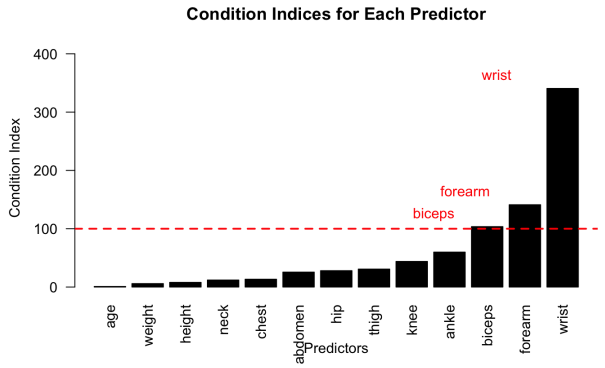

Figure 10: Plot of condition indices for every regressor

From fig. 10 one can see that three regressors are over the threshold of 100 one being close

to 400. This indicates that three regressor have moderate to significant multicollinearity, which is somewhat consistent with what can be seen in fig. 9. This gives stronger indication that the regressors of x_2, x_7 and x_6 might be near-linear dependent of other regressors.

Treatment of Multicollinearity

Since the option of collecting additional data was not available it was chosen to eliminate some variables. It can be seen in fig. 9 that x_2, x_7 and x_6 were the regressors that exceeded the threshold set to regards to VIF values and table 3 shows that the regressors of x_2 and x_7 exceeded the set threshold the most. Thus it was decided that the regressor variables of x_2 and x_7 should be removed from the full model in order to decrease multicollinearity. This led to a reduced model were the regressor variables corresponding to "weight" and "hip" were removed. Let this model be denoted as the reduced model, it was determined that it had the values of $R^2 = 0.7229$ and $R^2_{adj} = 0.7102$ which doesn't deviate far from corresponding values when considering the full model. Hence the model was made simpler while the values of R^2 and R^2_{adj} was more or less the same, which can be seen as an improvement.

Moreso the VIF value for every regressor was once again evaluated to quickly check for multicollinearity, with the cutoff of 10. This can be seen in fig. 11, which now show that no regressor has a VIF that exceeds the threshold thus indicating that the multicollinearity from the full model had been reduced.

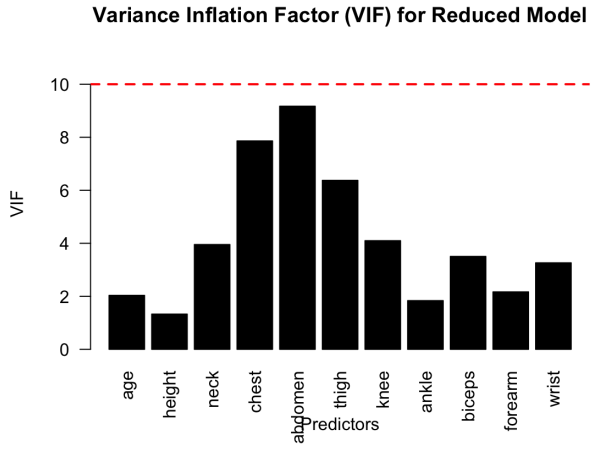

Figure 11: VIF values for each regressor in the reduced model

Variable Selection

Variable selection beyond the reduced model, hence the possibility of removing further regressors from the reduced model, was considered. Reducing the amount of regressors in a model has some benefit since it can enhance precision of estimates of the coefficients corresponding to the kept regressors. However this comes with the downside that bias can become prevelent, but if the if the deleted regressors has a neglible effect then the decrease in the variance of the estimated parameters outweight the introduced bias, this is often indicated by a reduced mean square error (MSE). At the same time keeping regressors that increase the variance of the estimated parameters and predicted values are is not benificial.

Thus subset regression models were developed using R's built in backward stepwise elimination and forward stepwise selection, which eliminates the least significant predictors iteratively based on AIC (Akaike Information Criterion)

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

, where L is the likelyhoof function. It evaluates the fit of the model while penalizing complexity. Models with a lower AIC value are preferred.

Firstly the reduced model, which coefficient estimates can be seen summarized below in table 4, was used to get a subset regression model using backwards stepwise elimination.

Variable	**Name**	**Coefficient Estimate**
Intercept	-	1.087e+00
x_1	Age	-2.048e-04
x_3	Height	3.388e-04
x_4	Neck	1.346e-03
x_5	Chest	3.386e-04
x_6	Abdomen	-1.928e-03
x_8	Thigh	-1.306e-04
x_9	Knee	6.109e-04
x_{10}	Ankle	-3.599e-04
x_{11}	Biceps	-4.158e-04
x_{12}	Forearm	-1.162e-03
x_{13}	Wrist	4.754e-03

Table 4: Coefficient estimates reduced model

The estimated coefficients of the subset model using stepwise backward elimination is

summarized below in table 5, which shows that the regressor variables of $x_2, x_7, x_8, x_9, x_{10}$ and x_{11} were removed from the reduced model.

Variable	**Name**	**Coefficient Estimate**
Intercept	-	1.090e+00
x_1	Age	-1.736e-04
x_3	Height	3.889e-04
x_4	Neck	1.200e-03
x_5	Chest	3.068e-04
x_6	Abdomen	-1.926e-03
x_{12}	Forearm	-1.290e-03
x_{13}	Wrist	4.634e-03

Table 5: Coefficient estimates from backward stepwise elmination

Starting from an empty model with no regressor a subset model was developed using forward stepwise selection. The coefficient estimates for this model can be seen in table 6, which shows that like the reduced model excludes the regressors x_2 and x_6.

Variable	**Name**	**Coefficient Estimate**
Intercept	-	1.087e+00
x_1	Age	-2.048e-04
x_3	Height	3.388e-04
x_4	Neck	1.346e-03
x_5	Chest	3.386e-04
x_6	Abdomen	-1.928e-03
x_8	Thigh	-1.306e-04
x_9	Knee	6.109e-04
x_{10}	Ankle	-3.599e-04
x_{11}	Biceps	-4.158e-04
x_{12}	Forearm	-1.162e-03
x_{13}	Wrist	4.754e-03

Table 6: Coefficient estimates from forward stepwise selection

Moreso both the value of R^2_{adj} and C_p was plotted for each iteration were the algorithm removed a regressor, these plots can be seen in fig. 8 and fig. 1 respectively.

Looking table 4, table 5 and table 6 we notice that the regressors that are included in the subset model developed by foward stepwise selections are the same ones that are included in the so called reduced model. Both of these models thus use 11 regressor variables,

but the subset model that was developed from the backward stepwise elimination only uses 7 regressors. In order to asses which of these models that would be the best we performed cross validation

Cross Validation

Cross validation was done for all three models by utilizing 10-fold cross validation. This yielded in the results seen in table 7

Model	MSE	R
Forward Stepwise Selection	0.0001094527	0.7141686
Backward Stepwise Elimination	0.0001115152	0.7144787
Reduced Model	0.0001164284	0.7011015

Table 7: Cross-Validation Results

Since it can be seen in table 7 that the models had similar values for MSE and R^2 we decided that the final model would be the reduced model, since we from fig. 11 saw that multicollinearity for this model was reduced.

Results; Model assessment using bootstrapping

The final model was assessed using boostrap based confidence intervals for the regressor coefficients. The 95% confidence intervals for each regressor coefficient can be seen in fig. 12

```
             Lower 95% CI  Upper 95% CI
(Intercept)  1.030848e+00  1.129494e+00
age         -3.291418e-04 -6.351881e-05
height       3.675050e-05  1.241249e-03
neck         2.810335e-04  2.462985e-03
chest       -8.206735e-05  8.295165e-04
abdomen     -2.309071e-03 -1.582313e-03
thigh       -6.840649e-04  4.773580e-04
knee        -5.308285e-04  1.537469e-03
ankle       -1.445511e-03  1.008565e-03
biceps      -1.287580e-03  3.778124e-04
forearm     -2.516613e-03 -1.900931e-05
wrist        2.149339e-03  7.247764e-03
```

Figure 12: Bootstrapped confidence intervals

From fig. 12 one can see that the intervals for the coefficients estimates for "chest", "thigh", "knee" and "ankle" include 0. Which may indicate that these predictors may not be significant since they don't have a purely positive or negative effect of the response variable.

Conclusion and Final model

During the project a MLRM model was developed with the purpose of estimating a mans BFM value based on several anthropometric measurements. Firstly a full model was developed from all regressor variables $x_1, x_2, ..., x_{13}$. From residual analysis it was concluded that model assumptions regarding the error terms were not violated. Thus it was concluded that no variable transformation was needed

Next leverage and influencial points were studied were some points could be considered influencial. Upon looking closer at these points it was concluded that even though they may be outliers they were kept for the model development with the intention that the model should be able to predict the BFM value for older and heavier men aswell.

Thirdly multicollinearity was assessed, which led to the conclusion that three regressor variables had slight to moderate multicollinearity. To reduce this the regressors of x_2 and x_7 was removed which led to the so called reduced model. The regressors of this reduced model then all had VIF values lower than the set threshold of 10, indicating that the multicollinearity was successfully reduced. This is illustrated in fig. 11

Further variable selection was then evaluated using backward stepwise elimination from the reduced model, based iteratively on AIC values. This led to a subset model with 7 regressors. Another subset model was developed from beginning from a completely empty model with no regressor variables applied to foward stepwise selection process based on AIC values. This subset model led to a model with 11 regressors, which were the same that were left from the reduced model. From cross validation of all three models it was concluded that the MSE and R^2 values between the models were very similiar, see table 7. Because of this and that the VIF values for the regressors in the reduced model had been seen not exceed the set threshold, the final model was determined to be the reduced model.

Lastly model assessment of the final model was done with bootstrapped confidence intervals, this can be seen in fig. 12. It can be seen that three of the coefficients have a confidence interval that includes 0, indicating that the corresponding regressors don't have a purely positive or negative influence on the response variable y (BFM). The values for the estimated regressor coefficient of the final can be seen in table 8. The regressors that aren't seen in the table have the coefficient 0, hence is not included in the final model.

Variable	**Name**	**Coefficient Estimate**
Intercept	-	1.087e+00
x_1	Age	-2.048e-04
x_3	Height	3.388e-04
x_4	Neck	1.346e-03
x_5	Chest	3.386e-04
x_6	Abdomen	-1.928e-03
x_8	Thigh	-1.306e-04
x_9	Knee	6.109e-04
x_{10}	Ankle	-3.599e-04
x_{11}	Biceps	-4.158e-04
x_{12}	Forearm	-1.162e-03
x_{13}	Wrist	4.754e-03

Table 8: Coefficient estimates of the final model