Project 1

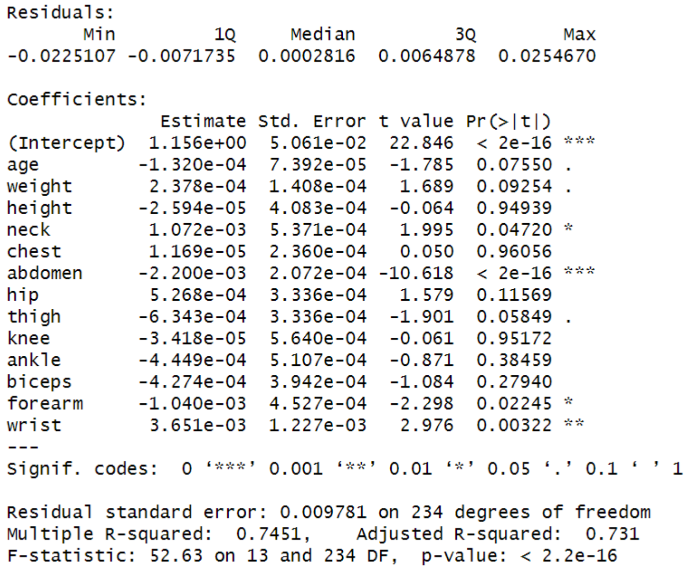
Klara Zimmerman and Ville Wassberg

February 2024

Full model

We begin by fitting all the data from ”bodyfatmen.csv” to a multivariate linear model with the 13 variables x_1 =”age”, ”x_2 =weight”, x_3 =”height”, x_4 =”neck”, x_5 =”chest”, x_6 =”abdomen”, x_7 =”hip”, x_8 =”thigh”, x_9 =”knee”, x_{10} =”ankle”, x_{11} = ”biceps”, x_{12} =”forearm” and x_{13} =”wrist”. The response variable is ”density”. Including the intercept, we thus get 14 regressors $\beta_0, \beta_1 ..., \beta_{13}$. Their approximated values are presented in figure 1.

Our model is thus $Y = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + ... + \hat{\beta}_{13} x_{13} + \epsilon = \hat{\beta} X + \epsilon$, where ϵ is the error term. The residual standard error for this model is 0.009781, and the adjusted R^2 value is 0.731.



```
Residuals:
       Min         1Q     Median         3Q        Max 
-0.0225107 -0.0071735  0.0002816  0.0064878  0.0254670 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  1.156e+00  5.061e-02  22.846  < 2e-16 ***
age         -1.320e-04  7.392e-05  -1.785  0.07550 .  
weight       2.378e-04  1.408e-04   1.689  0.09254 .  
height      -2.594e-05  4.083e-04  -0.064  0.94939    
neck         1.072e-03  5.371e-04   1.995  0.04720 *  
chest        1.169e-05  2.360e-04   0.050  0.96056    
abdomen     -2.200e-03  2.072e-04 -10.618  < 2e-16 ***
hip          5.268e-04  3.336e-04   1.579  0.11569    
thigh       -6.343e-04  3.336e-04  -1.901  0.05849 .  
knee        -3.418e-05  5.640e-04  -0.061  0.95172    
ankle       -4.449e-04  5.107e-04  -0.871  0.38459    
biceps      -4.274e-04  3.942e-04  -1.084  0.27940    
forearm     -1.040e-03  4.527e-04  -2.298  0.02245 *  
wrist        3.651e-03  1.227e-03   2.976  0.00322 ** 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.009781 on 234 degrees of freedom
Multiple R-squared:  0.7451,	Adjusted R-squared:  0.731 
F-statistic: 52.63 on 13 and 234 DF,  p-value: < 2.2e-16
```

Figure 1: Summary of the full model.

Residual Analysis

We now want to perform a residual analysis of this model, in order to verify that the relationship between the regressors x_i and the response variable Y is linear, and that the error terms ϵ_i are multivariate normal distributed.

Normal probability plots

We first check whether the residuals of the observations are normally distributed. This is checked by plotting the scaled residuals versus a theoretical standard normal distribution, in a Normal QQ-plot. We expect a linear relationship in this plot. As seen in the top right plot of figure 2, the residuals do seem to be normal distributed.

Residuals vs. fitted values

We then plot the residuals vs the fitted values of the regressors. Here, we expect a horizontal band. As seen in the the top left plot in figure 2, this is indeed what we find. This suggests that we do not need to transform the response variable.

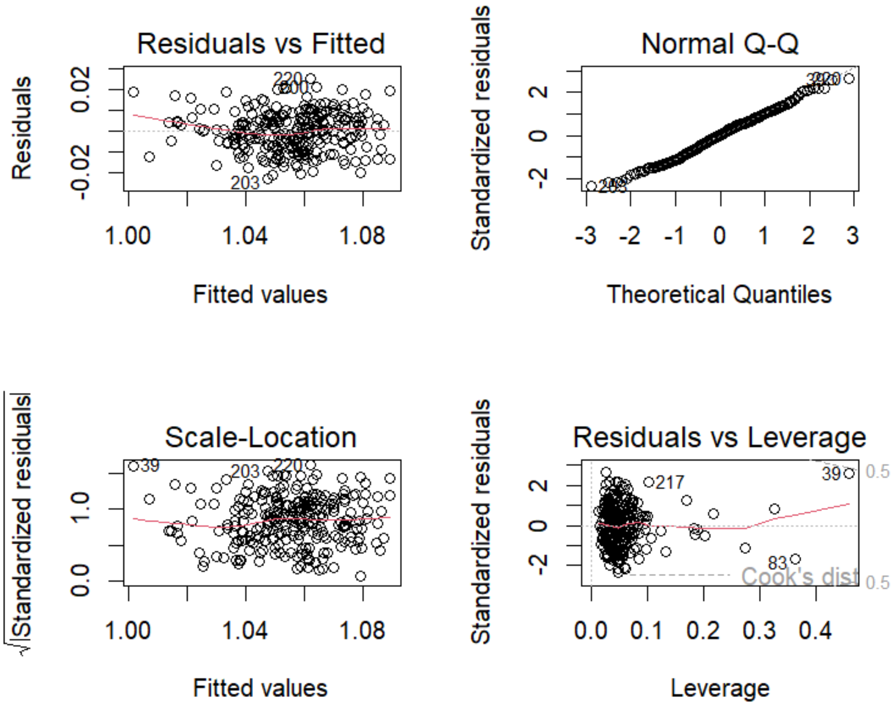


Figure 2: Plot of full model, including residual plots and normal Q-Q plot

Residuals vs regressor variables

Next, we look at all the observations and for each predictor plot them against the residuals for the full model. If the plots show horizontal bands, it suggests constant variance. If the data is scattered in a funnel shape, it suggests that the variance is not constant, and if the data is plotted as a curved band, it suggests the regressor is not well specified. In the two later cases, we may need to transform the data for the regressor in question. These plots are presented in figures 3 and 4.

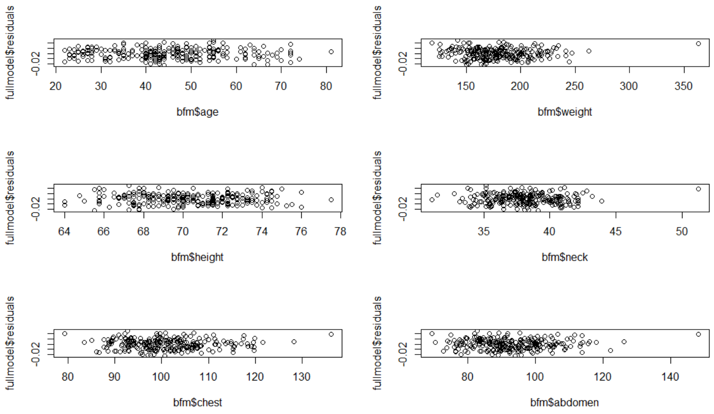


Figure 3: Residuals vs regressor values.

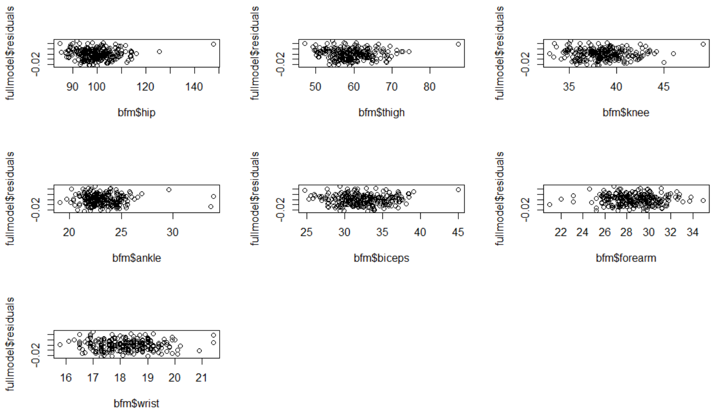


Figure 4: Residuals vs regressor values continued.

As seen in figures 3 and 4, funnel shapes appear for the regressors ”height”, ”neck”, ”ankle”, ”forearm” and ”wrist”. The ”chest” and ”hip” regressors show slightly curved shapes.

Added Variable plots

In order to further asses the unique contribution of each predictor, we plot each regressor against the response variable, while keeping the remaining regressors constant. If a linear relationship can be detected in these plots, it supports the claim that the regressor in question belongs in the model. Again, if a non-linear relationship is detected, it suggests that a variable transformation could be useful. These plots are presented in figures 5 and 6.

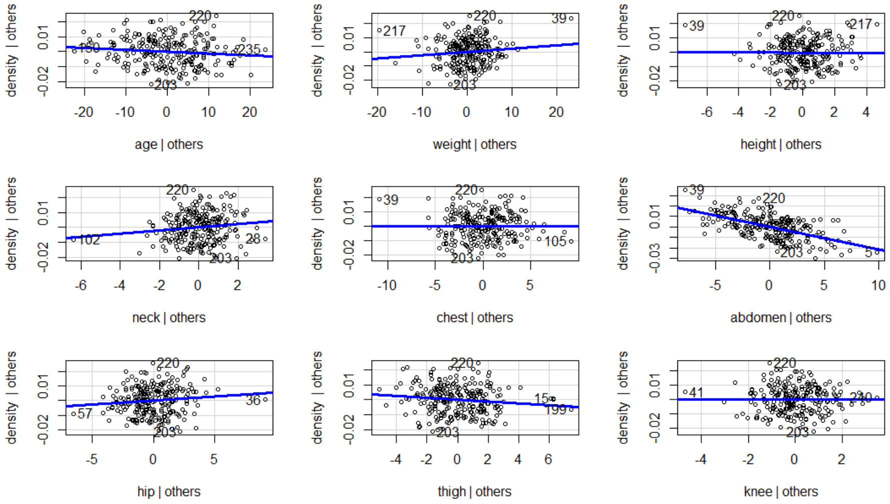


Figure 5: Added variable plots for the full model.

As seen in figures 5 and 6, the regressors that may need to be transformed are ”chest” and ”knee”, as these exhibit slightly non-linear relationships.

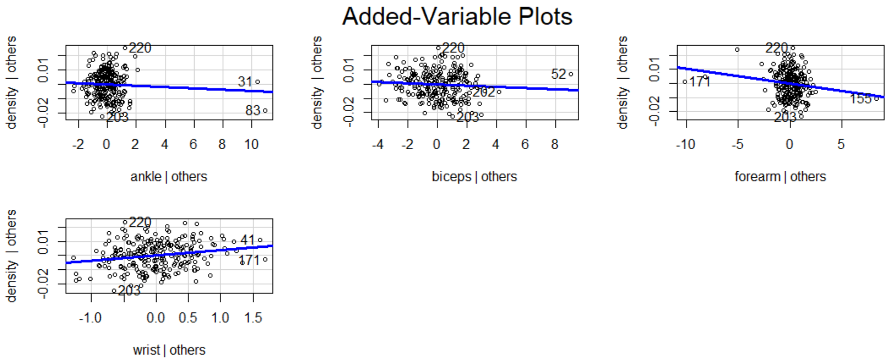


Figure 6: Added variable plots for the full model continued.

Studentized Residual and R-student residuals

Next, we plot the studentized residuals and R-student residuals. These plots can help detect potential outliers, as well as check whether our model seems to accurately capture patterns in the data.

In figure 12, we see some potential outliers. These are to be determined in the influence analysis. Other than that, we see that the observations seem to be within a linear band, spread randomly across the axes.

Results of residual analysis

After this residual analysis, we claim that the residuals seem to be normal distributed, and the overall model seems to capture a linear relationship in the data. However, the model could be improved by possibly transforming the variables ”chest”, ”knee”, ”wrist” and ”neck”. Further, we want to perform diagnostics to determine whether there are influential data points that should be considered outliers, and thus be excluded from the data set.

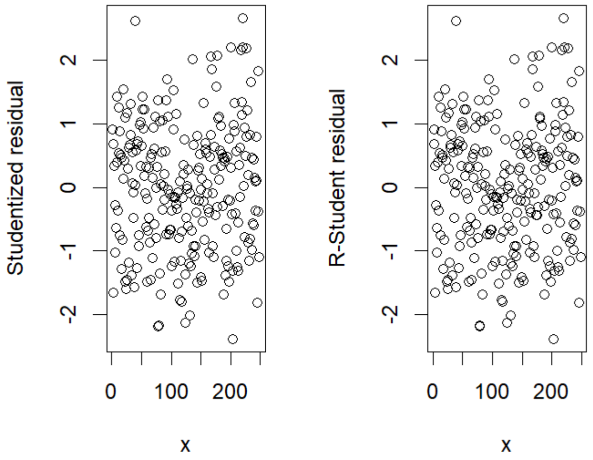


Figure 7: Studentised and R-student plots

Diagnostics and Handling of Outliers

In this section, we want to study the data set to detect potential outliers. Outliers are observations that have a big influence on the model, and that differ significantly from the other observations. If such data points are found, we study them further to see if eliminating the data point in question is reasonable. To do this, will look at residuals, Cook's distance and QQ-plots to help us find the potential outliers.

- **Residuals**

  As can be seen in the "Residuals vs Fitted" plot (figure 8), if we choose a threshold of $|error| < 2$, a total of 14 point are outside this interval. The observations with the largest residuals are 220 and 203. Looking at the "Scale-Location" plot (figure 8), we see that taking the square root of the standardized residuals shows that observation 39 also has a deviating value.

- **Cook's distance**

  We use Cook's distance test, with a standard threshold of $4/(n - k - 1)$, where n is the number of observations and $k-1$ is the number of regressors. This test yields 12 different data points that are at a greater distance from the rest of the data set. In particular, observations 39 and 83 appear to be at a relatively large distance from the rest of the data set.

- Looking at the leverage of the observations plotted against the standardized residuals, we see that obervation 39 seems to be a strong candidate for an outlier in the data set.
- **QQ-plot**

  Next, we look at which residuals appear to deviate the most from a normal distribution in the QQ-plot. According to the QQ Residuals in figure 8, we see that points 203 and 220 deviate the most.

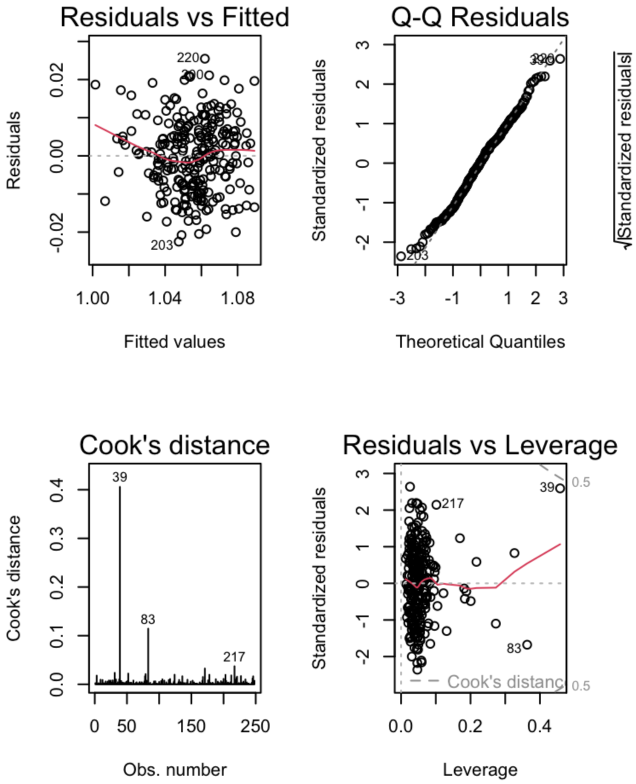


Figure 8: Plots used to identifying potential outliers.

Data points 203 and 220 are both influential and have a deviating residual. However, such points are not necessarily subject to be removed from the data set. Upon looking closer at these two observations, we see that they do not exhibit any apparent differences compared to the values of the other observations. We thus find no clear reason to exclude observations 203 and 220 from our data set.

We also take a closer look at observation 39, since this data point has a Cook's distance of 0.4, the largest in the set, which is almost 4 times larger than the data point with the second largest Cook's distance. This reveals that observation 39 exhibits a significantly larger "weight"-value than any other observation in the cohort. We therefore choose to remove this point from our data set, since it deviates from the rest of the set, with a clear explanation.

Transforming and Cleaning the Data

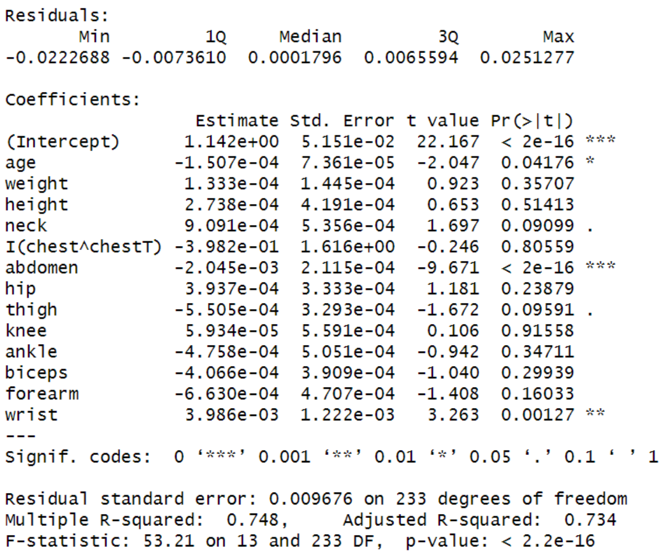
In the residual analysis, we saw that potentially transforming the variables "chest", "knee", "wrist" and "neck", we could obtain a model that captures the structure of the data better.

We use the boxTidwell function to find the best transformation for each of these variables. However, the only transformation that noticably improved the linear relationship between regressor and response variable was transforming the variable "chest". The best transformation according to the boxTidwell strategy the values of this regressor to the power of -0.89.

As decided in the diagnostics tests for outliers, we remove observation 39 from the data set.

The Clean Model

After transforming and cleaning up the data set, we fit a new model, summarized in figure 9. The adjusted R^2-value has now increases slightly, from 0.731 to 0.734, and the residual standard error is now 0.009676. We re-run the residual analysis, and plot the results in figure 10.



```
Residuals:
       Min         1Q     Median         3Q        Max 
-0.0222688 -0.0073610  0.0001796  0.0065594  0.0251277 

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)    
(Intercept)      1.142e+00  5.151e-02  22.167  < 2e-16 ***
age             -1.507e-04  7.361e-05  -2.047  0.04176 *  
weight           1.333e-04  1.445e-04   0.923  0.35707    
height           2.738e-04  4.191e-04   0.653  0.51413    
neck             9.091e-04  5.356e-04   1.697  0.09099 .  
I(chest^chestT) -3.982e-01  1.616e+00  -0.246  0.80559    
abdomen         -2.045e-03  2.115e-04  -9.671  < 2e-16 ***
hip              3.937e-04  3.333e-04   1.181  0.23879    
thigh           -5.505e-04  3.293e-04  -1.672  0.09591 .  
knee             5.934e-05  5.591e-04   0.106  0.91558    
ankle           -4.758e-04  5.051e-04  -0.942  0.34711    
biceps          -4.066e-04  3.909e-04  -1.040  0.29939    
forearm         -6.630e-04  4.707e-04  -1.408  0.16033    
wrist            3.986e-03  1.222e-03   3.263  0.00127 ** 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.009676 on 233 degrees of freedom
Multiple R-squared:  0.748,     Adjusted R-squared:  0.734 
F-statistic: 53.21 on 13 and 233 DF,  p-value: < 2.2e-16
```

Figure 9: Summary Model trained with cleaned and transformed data.

As we see in figure 10, the Cook's distances of the remaining data points are much closer together, and the leverage seems more equally spread.

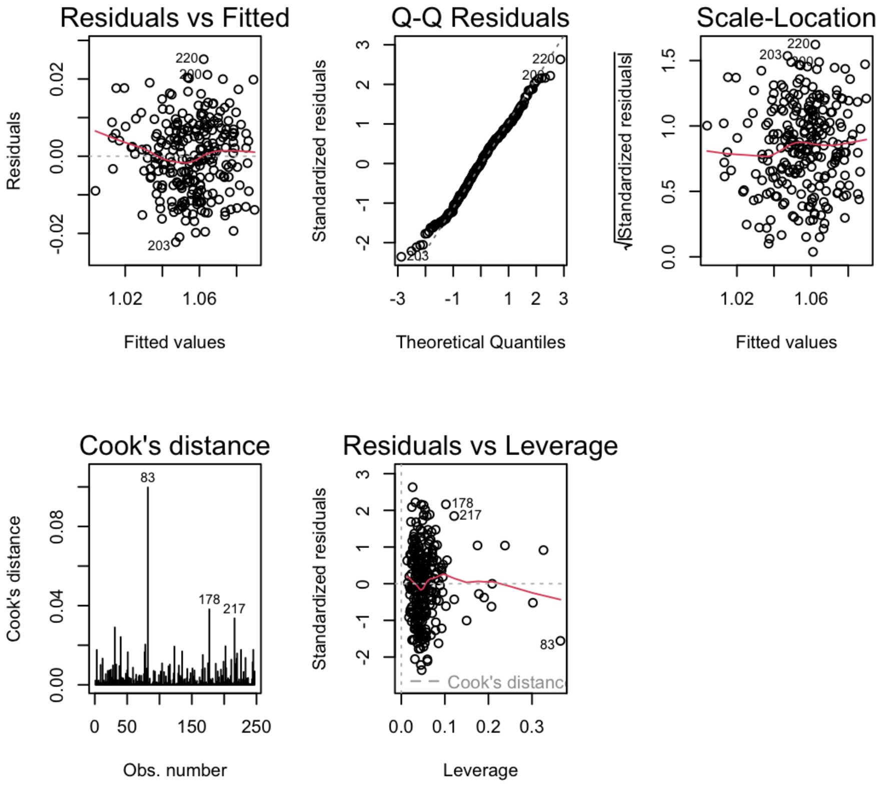


Figure 10: Residual analysis after transforming regressor ”chest” and removing observation 39.

Multicolinearity

We check for multicolinearity using built in functions (with the transformed and cleaned model). In figure 11, we see that there is a very strong correlation between many of the regressors. For example, ”weight” and ”hip” have a correlation of 0.932, and ”abdomen” and ”chest” have a correlation of -0.905. Most regressors seem to correlate strongly with almost all other regressors. An effect of this could be a greater uncertainty in estimating the regressor coefficients, making the model unstable. It also makes it difficult to understand which regressors are truly of importance to the model. This suggests that our model will benefit strongly from a reduction of parameters. This will be discussed in the coming sections.

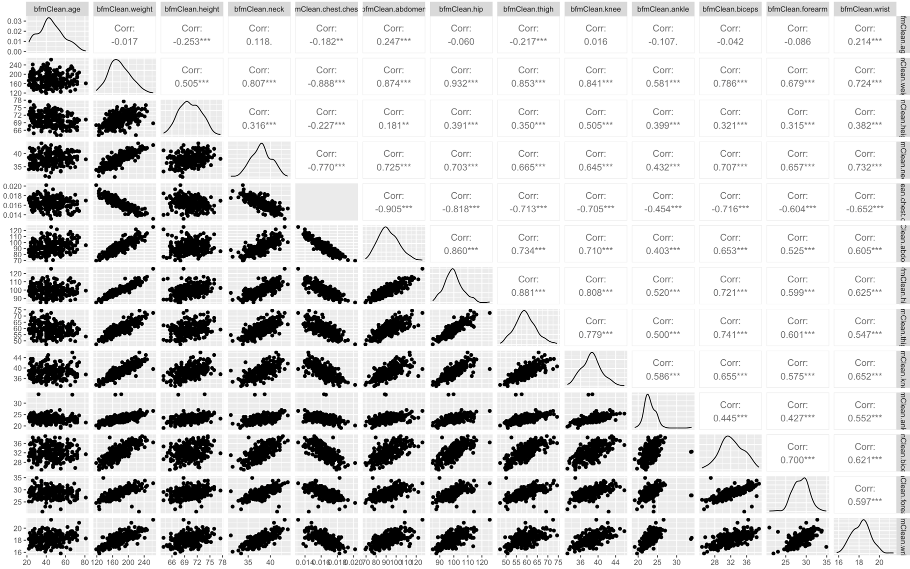


Figure 11: Covariance between all sets of regressors

Variable Selection

Splitting the data

We now split the data into a training set a test set. This is done in order to evaluate how well our model predicts new data points. We fit a model using the training set, and test it on the test set. Using all regressors, with the cleaned and transformed data set, we get a mean squared error of 0.0001052547. We now want to see if variable selection can decrease this value, as well as simplify our model.

Stepwise selection

We then use forward, backward and simultaneous stepping techniques to reduce the number of regressors in our model. Through these methods we hope to decrease the MSE and improve the multicolinearity of our model. With all three methods we get the same results:

- To maximize R^2, ($maxR^2 = 0.7525$), we use 9 regressors: age, height, chest, abdomen, hip, ankle, biceps, forearm and wrist
  - MSE = 0.0001081481
- To minimize bic ($min(bic) = 6.028116$), we use 3 regressors: height, abdomen and wrist
  - MSE = 0.0001034654

- To minimize C_p ($minbic = -149.1595$), we use 6 regressors: age, height, chest, abdomen, biceps and wrist.
  - MSE = 0.0001057558

As seen in figure 11, the variables "abdomen" and "chest" have a very large colinearity (-0.905). The only model found using stepwise selection where both of these variables do not appear is the model with only 3 regressors. This model is also the most simple model, since it has the fewest number of parameters. It is also the model with the lowest MSE out of the four presented. This suggests that this model, which is $Y_1 = X\hat{\beta} = 1.05503 + 0.00284x_3 - 0.01808x_6 + 0.00368x_{13}$, is our best canditate at this point.

To further asses the predictive power of a model such as Y_1, that is trained on a data set containing observations of the three variables $x_3, x_6 and x_{13}$, we perform a cross validation. This can be done using all observations, not only those in the training set. This enlarges our data set. Choosing the number of folds in the cross-validation is a trade-off between bias and variance. We thus chose $k = 10$ folds, since this is a standard value that tends to create neither an overfitted nor underfitted model for a data set of this size.

We now get a slightly different model, with MSE $= 9.650921e - 05$. The model we get is the following.

$$Y_2 = X\hat{\beta} = 1.05558 + 0.00228x_3 - 0.01802x_6 + 0.00352x_{13}$$

where $x_3 =$ "height", $x_6 =$ "abdomen" and $x_{13} =$ "wrist".

The multicolinearity of these three regressors is presented in figure 12. None of these three regressors have a colinearity of above 0.7, which is a clear improvement from the full model. However, "abdomen" and "wrist" do seem to have similar structures.

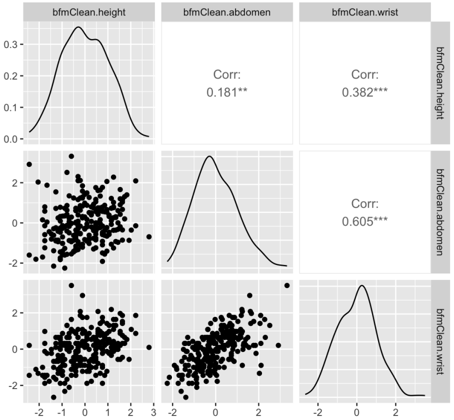


Figure 12: Multicolinearity of the three regressors "height", "abdomen" and "wrist": chosen using stepwise selection to minimize *bic*.

Ridge, Lasso and PCR

Other techniques for model selection are Ridge and lasso regression, and PCR. We use built in functions to find such models, and get the following results.

- Ridge Regression
  - number of regressors: all
  - MSE = 0.0001099346
- Lasso Regression
  - number of regressors: 3: height, abdomen, wrist (+ intercept)
  - MSE = 0.0001092377
  - Model:
    $Y = 1.0550645873 + 0.0018292446x_3 - 0.0145269403x_6 + 0.0005462876x_{13}$
- PCR
  - number of regressors: 13 (no intercept)
  - MSE = 0.0001051365

Out of these three methods, PCR is the model which minimizes the mean square error the best. This model does not include an intercept, but does include all of the regressors. Hence, this model is very different from the one found using Stepwise techniques to minimize the mean squared error. In the next section, we will look at the confidence intervals for some of the models we have found.

Model Assessment

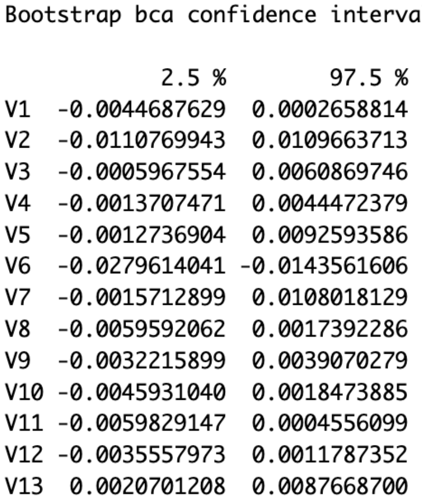
We want to look at computer-intensive procedures to estimate the standard errors for the models we have found. To do this, we will use a bootstrapping method. By doing so, we can approximate the confidence intervals.

The bootstrap sample is a randomly selected subset, chosen with replacement, of our observations. By fitting a model using different such subsets, we produce bootstrap estimates for each regression coefficient.

According to the summary of the model with 3 predictors, the Residual standard error is 0.009508. For each predictor, the standard error is 0.0009549 (height), 0.0010593 (abdomen), 0.0010924 (wrist).

We now use bootstrapping to find the confidence intervals for the PCR model:

PCR confidence intervals



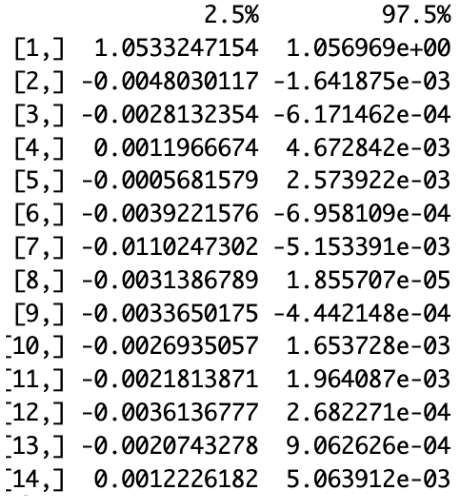
```
Bootstrap bca confidence intervals

             2.5 %        97.5 %
V1  -0.0044687629  0.0002658814
V2  -0.0110769943  0.0109663713
V3  -0.0005967554  0.0060869746
V4  -0.0013707471  0.0044472379
V5  -0.0012736904  0.0092593586
V6  -0.0279614041 -0.0143561606
V7  -0.0015712899  0.0108018129
V8  -0.0059592062  0.0017392286
V9  -0.0032215899  0.0039070279
V10 -0.0045931040  0.0018473885
V11 -0.0059829147  0.0004556099
V12 -0.0035557973  0.0011787352
V13  0.0020701208  0.0087668700
```

Figure 13: The confidence intervals from the PCR model, found using bootstrapping.

If 0 is included in the interval, and the interval is small, this suggests that the regressor in question might not contribute very much to the model. This is not the case for approximations $\hat{\beta}_6$ (abdomen) and $\hat{\beta}_{13}$ (wrist). Since these parameters are included in all models we have found above, we have reason to believe they contribute to the model. This is also corroborated by the very first model we created (figure 1) which also suggested that x_6 and x_{13} where the most influential parameters.

We use bootstrapping to find the approximated confidence intervals for the model found using ridge regression as well. This is presented in figure 14.

Ridge confidence intervals

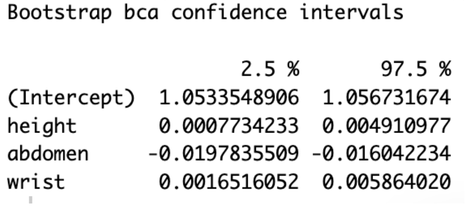


```
               2.5%        97.5%
 [1,]  1.0533247154  1.056969e+00
 [2,] -0.0048030117 -1.641875e-03
 [3,] -0.0028132354 -6.171462e-04
 [4,]  0.0011966674  4.672842e-03
 [5,] -0.0005681579  2.573922e-03
 [6,] -0.0039221576 -6.958109e-04
 [7,] -0.0110247302 -5.153391e-03
 [8,] -0.0031386789  1.855707e-05
 [9,] -0.0033650175 -4.442148e-04
[10,] -0.0026935057  1.653728e-03
[11,] -0.0021813871  1.964087e-03
[12,] -0.0036136777  2.682271e-04
[13,] -0.0020743278  9.062626e-04
[14,]  0.0012226182  5.063912e-03
```

Figure 14: The confidence intervals from the ridge regression model, found using bootstrapping.

Finally, we look at the confidence intervals for the model with only three parameters, which minimized the *bic*-value. We see in figure 15 that the intercept and all the predictors are within narrow confidence intervals, and that these are not centered around 0. This again further confirms the accuracy of this parameter choice.

Confidence intervals - model that minimizes bic

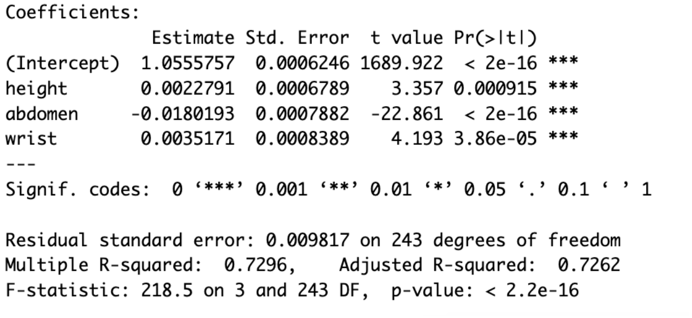


```
Bootstrap bca confidence intervals

                    2.5 %       97.5 %
(Intercept)  1.0533548906  1.056731674
height       0.0007734233  0.004910977
abdomen     -0.0197835509 -0.016042234
wrist        0.0016516052  0.005864020
```

Figure 15: The confidence intervals from the model found when minimizing bic.

Conclusion

We now conclude that a reasonable model to estimate body fat amongst men from the dataset provided, is: $Y_2 = X\hat{\beta} = 1.05558 + 0.00228x_3 - 0.01802x_6 + 0.00352x_{13}$. This is a simple model of only three predictors, namely abdomen, height and wrist, indicating measurements made on these is enough for predicting the body fat density of men. Several methods have been used such as a



```
Coefficients:
              Estimate Std. Error  t value Pr(>|t|)    
(Intercept)  1.0555757  0.0006246 1689.922  < 2e-16 ***
height       0.0022791  0.0006789    3.357 0.000915 ***
abdomen     -0.0180193  0.0007882  -22.861  < 2e-16 ***
wrist        0.0035171  0.0008389    4.193 3.86e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.009817 on 243 degrees of freedom
Multiple R-squared:  0.7296,    Adjusted R-squared:  0.7262 
F-statistic: 218.5 on 3 and 243 DF,  p-value: < 2.2e-16
```

Figure 16: Summary of final model.

thorough residual analysis, handling and diagnostics of outliers, transformation of data; multicollinearity has been diagnosed and treated using methods such as Lasso, Ridge regression and principal component regression. Variable selection using forward and backward methods as well as evaluating the models by MSE, BIC, C_p and adjusted R^2 using cross-validation, with a number of k-folds as 10, which was chosen with a balance of variance-bias-tradeoff in mind. Moreover, computer-intensive procedures such as Bootstrap based confidence intervals for regression coefficients were used. The resulting model predicts density with a mean square error (MSE) of $9.650921e-05$ and adjusted R^2 value of 0.7262. Finally, the final model to be used is displayed in 16.