Kungliga Tekniska Högskolan

SF2930 Regression Analysis

Report I

Isac Karlsson
Ludvig Wärnberg Gerdin

Examiner
Tatjana Pavlenko

March 6, 2020

Contents

1 Introduction and Project Goals

1.1 Introduction

Our choice of scenario is Scenario I: Body fat assessment, which involves Large-Sample regression ($p < n$). According to the World Health organization (WHO) obesity, the state where excess body fat is causing extensive health effects, is a large risk factor for some chronic diseases. Some examples are cancer and diabetes. Since the number of cases of obesity is increasing one may want to identify these people quickly and reliably.

1.2 Data Description

Since BMI has shown to be a bad predictor of actual fatness, this project focuses on body fat mass (BFM). Accurate methods for calculating BFM have been developed, but because of the high costs and efforts associated with these methods researchers turn to techniques such as regression models for computing BFM.

The given dataset `bodyfatmen.csv` describes data of body density (calculated using underwater weighing), age and other anthropometric variables about 252 men.

1.3 Project Goals

The main goal of the project is to create and validate our own regression model in order to predict BFM. This includes the following:

1. Residual analysis for model adequacy checking
2. Handling of outliers, influential observations and leverage
3. Transformations of variables in order to correct model inadequacies
4. Multicollinearity treatments and diagnostics
5. Different types of variable selection and evaluation of these using cross validation
6. Computer-intensive procedures for model assessment (e.g. bootstrap residuals)

2 Analyses and Model Development

The information presented in the proceding sections are primarily taken from {Introduction to Linear Regression Analysis} [6]. If not, the information is cited.

2.1 Residual analysis

Some major assumptions we use in our analysis are:

1. The errors ϵ_i for observation i are independently and identically normally distributed.
2. Mean of $\epsilon = 0$
3. Variance of $\epsilon = \sigma^2$, where σ is a constant.
4. There is approximately a linear relationship between the regressors and the response (y).

When analysing violations of the assumptions given above, the primary tool is using the model residuals. We define the residual for observation i as

$$e_i = y_i - \hat{y_i},\ i = 1, ..., n$$

One may view a residual as the difference between the data and the fit although it is also a way to analyze the variability in the response variable that cannot be explained by the regression model. Plotting residuals is a effective method to examine how the regression model fits the data and make sure the assumptions listed are not violated.

2.1.1 R-Student

One type of residual is the externally studentized residual, which is given by

$$t_i = \frac{e_i}{\sqrt{S_i^2(1 - h_{ii})}},\ i = 1, ..., n$$

The externally studentized residual is also called the R-student residual. Here an estimate of σ^2 is used instead of MS_{Res} in order to create an *externally* studentized residual.

Now we introduce some basic residual plots, which are commonly generated using computers. These should be analyzed routinely when solving any kind of regression modelling problem. Note that the externally studentized residuals are often the ones plotted since they have constant variance.

2.1.2 Normality of residuals

This is a tool for analysing if two datasets (of quantiles) come from the same probability distribution. By plotting the quantiles against each other we will hopefully see somewhat of a straight line. This corresponds to them originating from the same distribution.

Here some small departures from the normality assumption does not have a large impact. Meanwhile large nonnormality could have more of an impact on the regression modelling process. Mainly, the problems relate to inference of the model building - for example, prediction intervals depend on the normality assumption. One may check the normality assumption simply by constructing a normal probability plot of the residuals.

2.1.3 Fitted Values Against Residuals

Simply a plot of the, often externally studentized, residuals versus the fitted values. This is useful because it allows an easy way to detect model inadequacies. If the plot shows the residuals contained in a horizontal band, then the model does not contain any obvious defects. If this is not the case one may conclude that there are likely model imperfections.

2.1.4 Added Variable Analysis

Particularly useful when analysing if the relationship between the regressor variables and the response has been defined accurately. Another way to use these plots are when evaluating the marginal usefulness of some variable that is not presently a part of the model. Here y (the response variable) and x_j (regressor) is regressed against the regressors (currently present in the model) and the residuals that follow for each regression. When plotting these residuals against each other one may analyse the marginal relationship for the regressor x_j that has caught our attention.

2.2 Diagnostics and handling of Outliers

2.2.1 Treatment of outliers

An observation that is noticeably different from the rest of the data is considered an outlier. A way to spot y space outliers is simply by analyzing the residuals. The ones that are noticeably larger (when considering the absolute value of these residuals) than the other residuals is an indication of potential outliers. The magnitude of the impact caused by these outliers depends on their location in x space. An example of identifying potential outliers is by using scaled residuals (e.g. R-student).

Note that outliers that are considered bad values, e.g. values from mis-measuresments, should preferably be discarded. Meanwhile there should always be non-statistical confirmation that the outlier really is a bad value before discarding it. One could argue that outliers are the most important part of the data since it often control many properties when modelling.

One way to analyse the effect of each outliers is by simply not including the data point and refitting. In general we prefer it when the model is not too sensitive to a small number of observations.

The hat matrix can be very useful when detecting potential outliers, since it determines the variances and covariances of $\hat{y_j}$ and $\mathbf{e}$. Each element h_{ij} corresponds to the amount of leverage exercised by the ith observation y_i on the jth, fitted value, $\hat{y_j}$.

It appears that large hat diagonals may correspond to an influential outlier since they are remote in x space when compared to the rest of the data. Knowing this analysts also want to observe the studentized residuals of each observation. Large hat diagonals along with large residuals are likely an influential observation.

2.2.2 Cook's Distance

One way to both of these at the same time is by using the squared distance between the least-squares estimate (based on all n points) and also the estimate obtained when deleting the ith point. This is called Cooks distance and can be interpreted as the euclidean distance that the vector containing fitted values is moved when deleting the ith observation.

The Cook's distance is arguably one of the more important metrics for our prediction purpose, since is highlight's the observation's effect on the predicted y-values. [3]

2.2.3 DFFITS & DFBETAS

Two other measures of the effects when deletion an observation is $DFBETAS$ and $DFFITS$. $DFBETAS$ tells us about the effects on the regression coefficient $\hat{\beta_j}$ when deleting the ith observation. It is defined as follows and is given in units of standard deviation.

$DFFITS$ analyses the effects on the fitted value when deleting the ith observation. Here $DFFITS$ tells us the number of standard deviations that the fitted value is changed by when deleting observation i. Since the $DFFITS$ values consider the effect on the fitted value, this metric is arguably one of the more important ones for our purpose.

$DFBETA$ is presumably more interesting from an explanatory point-of-view [3], which is not the primary purpose of this report. We therefore analyse the Cook's distance and the $DFFITS$ values more thoroughly that the $DFBETA$ values.

2.3 Transformations of variables

Whenever an assumption mentioned in 2.1 violated it is usually a good idea to consider data transformation. In some cases expressing the regressor and or the response variables using another measurement results in violations no longer being present, e.g. inequality of variance.

If we wish to transform y, in order to correct for example nonconstant variance, we can use the power transformation y^λ where λ is what we want to determine. We can do this by using the Box-Cox method which also allows us to estimate the parameters of the regression model simultaneously, using maximum likelihood.

Further, we might want to transform the regressors x_j, for example if the regressor expresses a non-linear relationship with the response variable. Partial regression plots can be used to determine if such a transformation is needed. If the regressors x_j enters the model linearly, then the partial regression plot will show a straight line. When x_1 is considered a candidate variable for the model, if the partial regression plot shows a horizontal band, that tells us that no additional information for predicting y is described by x_1. When the partial regression plot shows a curvilinear band, then one may use a transformation (e.g. replacing x_1 with $1/x_1$).

2.4 Diagnostics and handling of Multicollinearity

As a result of multicollinearity, the model fit with the least-squares method may be very deficient. This may cause the usefulness of the regression model to decrease significantly.

2.4.1 Pair-wise Correlation Matrix

One simple way to detect multicollinearity is by inspecting the off-diagonal element r_{ij} in the $\mathbf{X'X}$ matrix. $\mathbf{X}$ is an $n-\text{by}-p$ matrix, where n is the number of observations and p is the number of predictors. A near linear dependency between x_i and x_j will result in $|r_{ij}|$ to be near unity. Note that this is useful for detecting linear dependence between pairs of regressors and that this can not be used as a tools for detecting anything more complex than that. Therefore, this method of detecting multicollinearity will only be considered as a complementary method to more appropriate methods described here.

2.4.2 Variance Inflation Factors and Eigensystem analysis

The diagonal elements of the matrix $C = (\mathbf{X'X})^{-1}$ can also be used for detecting multicollinearity. Note that the jth element of C can be written as

$$C_{jj} = (1 - R_j^2)^{-1}$$

where R_j^2 is obtained when x_j is regressed on the other $p-1$ regressors.

When x_j is almost orthogonal to the other regressors, R_j^2 is small and C_{jj} is close to unity. Meanwhile if x_j is nearly linear dependent on a subset of the other regressors, R_j^2 is close to unity and C_{jj} is large.

One may also analyze the characteristic roots/eigenvalues of $\mathbf{XX}$ to measure the extent of multicollinearity. When one or more of the eigenvalues are small, then there exists one or more near-linear dependencies.

As an ending note, we should mention the inhererent multicollinearity in this dataset. Most candidate predictors are measures of body size, which naturally causes the predictors to be linearly related in to each other. That said, it is appropriate to investigate methods to alleviate the effect of multicollinearity since the stability of the model is heavily influenced by multicollinearity.

2.5 Computer-Intensive Procedures and Variable Selection

2.5.1 Bootstrap

Bootstrapping is a computer-intensive technique that allow us to compute, for example, reliable estimates of the standard errors of regression estimates when there is no standard procedure available or cases where the results are only approximate techniques (e.g. based on large-sample theory).

Say, for instance, that we are interested in the standard error for a particular predictor coefficient $\hat{\beta}_j$. Then, we are required to select a random sample of size n with replacement from this original sample, this is called the bootstrap sample. We proceed to fit the model to this sample by using the procedure as for the original sample. This gives us the first bootstrap estimate $\hat{\beta}_1^*$. We repeat this process many times and in each repetition, a new bootstrap sample is selected, the model is fit, and an estimate $\hat{\beta}_i^*$ is determined. From these bootstrapped estimates, an approximate standard error is computed.

2.5.2 Variable Selection

If multicollinearity is present, variable selection methods are very useful. Note that variable selection does not result in complete elimination of multicollinearity, in some cases two or more regressors are highly related even though some subset of them indeed should be a part of the model, instead it helps us justify the presence of multicollinearity in the final model. One should also note that experience and subjective considerations should always be considered as a part of the variable selection problem.

2.5.3 All Possible Regression and Other Methods

Simply requires to fit all the regression equations starting with one candidate regressor, then two candidate regressors and so on. These are later analyzed regarding some criterion and the best one is selected.

Since evaluating all possible regressions can sometimes be computationally infeasible, there are other methods for evaluating only a smaller number of subset regression models by adding/removing regressors one at a time. These methods are generally called stepwise procedures, and examples are forward selection and backward elimination.

Note that we have not included any of the stepwise regression methods mentioned above. Primarily because of the list of problems connected with these methods [4], which are for example that they yield R-squared values that are highly biased and cause severe problems in the presence of collinearity. The use of all possible regression have been recommended in favour of the stepwise techniques, if computationally feasible. [6] Since we are dealing with a relatively small dataset, and less than 20 candidate predictors, we choose to use all possible regression in our variable selection.

2.5.4 Cross-validation

The all possible regression procedure were nested into a cross-validation procedure. The most primitive form of cross-validation is when the data is split into two parts, and the model is trained one subset of the data (usually referred to as the training set) and validated on the other part of the data (usually referred to as the validation set).

The most extreme form of cross-validation is the leave-one-out cross-validation, where the model is fit on $n-1$ observations and tested on one the last observation. Then the "validation"-observation is swapped for one of the $n-1$ observations and the procedure is looped until every observation has acted as the validation observation.

In this project we first make a simple split of the data as described in the first paragraph, and end up with a training and validation sample. We combine this split with the use of K -fold cross-validation. In K -fold cross-validation, the sample is split into K parts, and the model is fitted on $K-1$ parts of the data and validated on the last part. The validation part is swapped for one of the $K-1$ parts, and the procedure is re-run until all "folds" has acted the validation fold. The cross-validation is conducted on the training sample and the final model error is validation sample. That way we derive a model from one part of the data, and get an performance estimate of the chosen model on unseen data.

We set the $K = 10$ in this analysis. This has been recommended as a compromise for keeping low both the bias of overestimating the generalization error of the model and the variance of the model. [5]

Table 1: Sample characteristics.

	Overall
n	248
density (median [IQR])	1.05 [1.04, 1.07]
age (median [IQR])	43.00 [35.00, 54.00]
weight (median [IQR])	176.50 [159.25, 196.81]
height (median [IQR])	70.00 [68.25, 72.25]
neck (median [IQR])	38.00 [36.40, 39.42]
chest (median [IQR])	99.65 [94.55, 105.30]
abdomen (median [IQR])	90.95 [85.05, 99.33]
hip (median [IQR])	99.30 [95.57, 103.28]
thigh (median [IQR])	59.00 [56.08, 62.35]
knee (median [IQR])	38.50 [36.90, 39.90]
ankle (median [IQR])	22.80 [22.00, 24.00]
biceps (median [IQR])	32.05 [30.28, 34.40]
forearm (median [IQR])	28.75 [27.30, 30.00]
wrist (median [IQR])	18.30 [17.60, 18.80]

3 Results

3.1 Sample characteristics

Table 1 reports the sample characteristics. These are left for the reader, in particular to compare with the outliers presented in section 3.6.

3.2 Residual analysis

3.2.1 Normality of residuals

Figure 1 illustrates a quantile-quantile plot of the externally studentized residuals. The observer may say that the points exhibit a pattern that indicates that the residuals are distributed with heavier tails than that of a normal distribution. [6]. Still, the deviations from the diagonal line is relatively small, and hence we conclude that the residuals are normally distributed.

3.2.2 Fitted Against Residuals

Figure 2 illustrates the fitted values $\hat{y}_j$ against the R-student residuals. No apparent pattern is formed by the points, i.e. the points seem to be randomly scattered along the dotted horizontal line. Hence we conclude that the residuals have constant variance, and thus assume that the errors do as well.

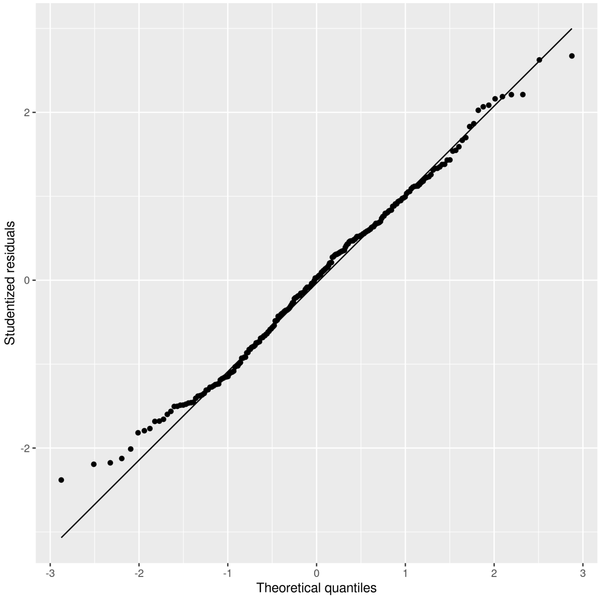


Figure 1: Normality plot of residuals.

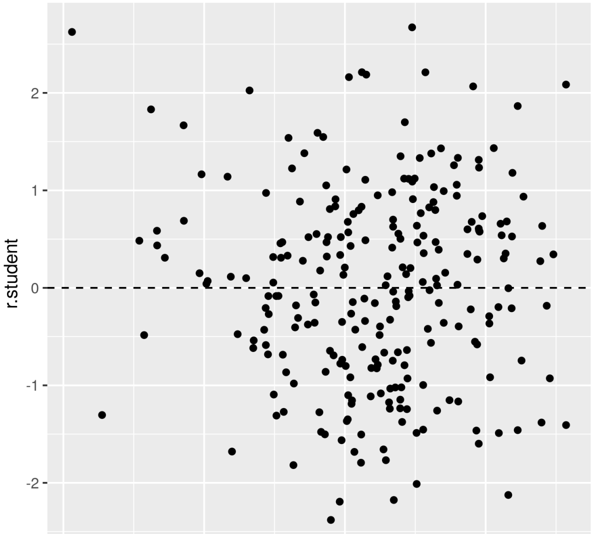


Figure 2: Fitted values against R-student residuals.

3.2.3 Added Variable Analysis

Partial regression plots are found in figure 3, 4, 5, and 6. All figures exhibit potential points that are unusually large in the x-space and hence their influence on the model fit should be examined further. This will be considered in section 2.2. All regressors seem to enter the model linearly. The `height` regressor exhibit a slight double-bow pattern, however the pattern is not obvious. With these points in mind, we choose not to transform any of the predictors.

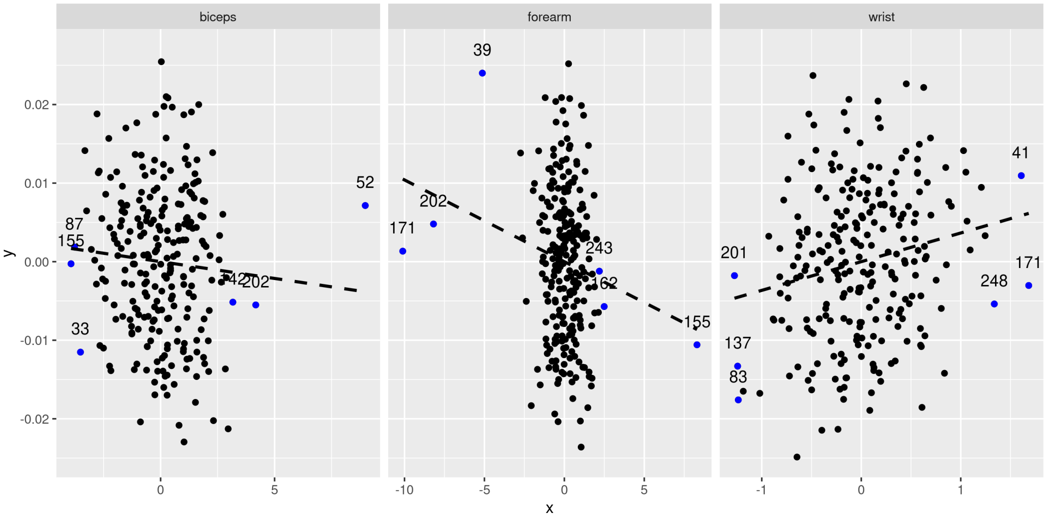


Figure 3: Partial regression plots of regressors `biceps`, `forearm`, and `wrist`.

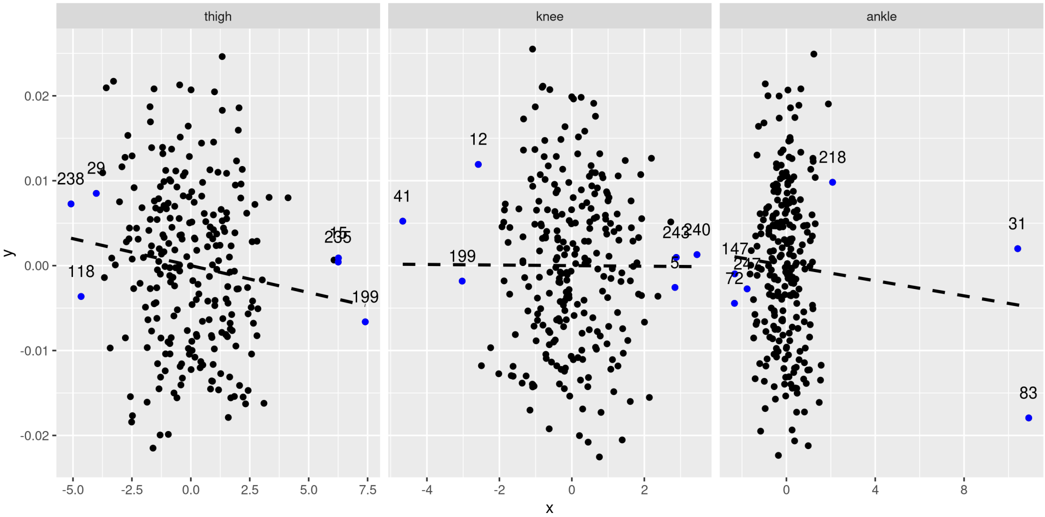


Figure 4: Partial regression plots of regressors `thigh`, `knee`, and `ankle`.

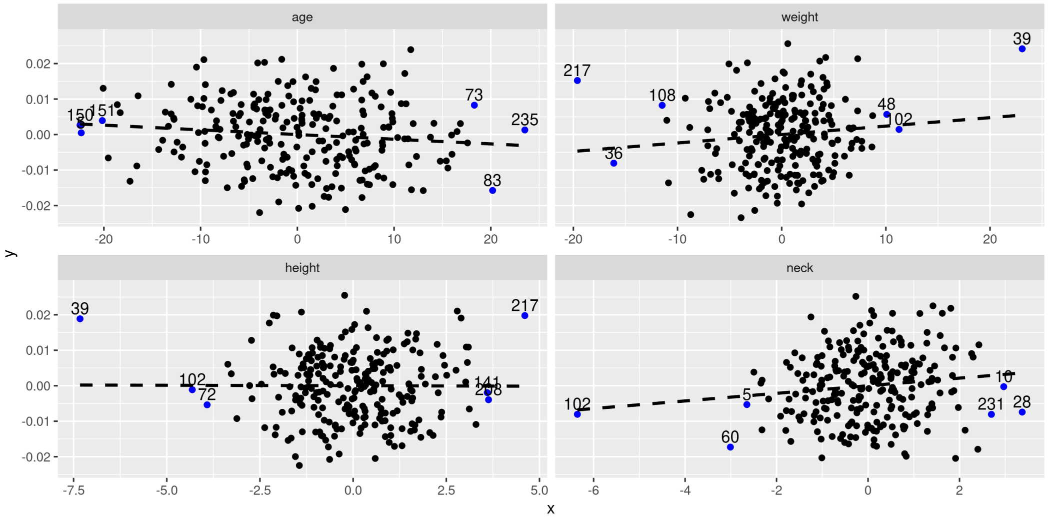


Figure 5: Partial regression plots of regressors `age`, `weight`, `height`, and `neck`.

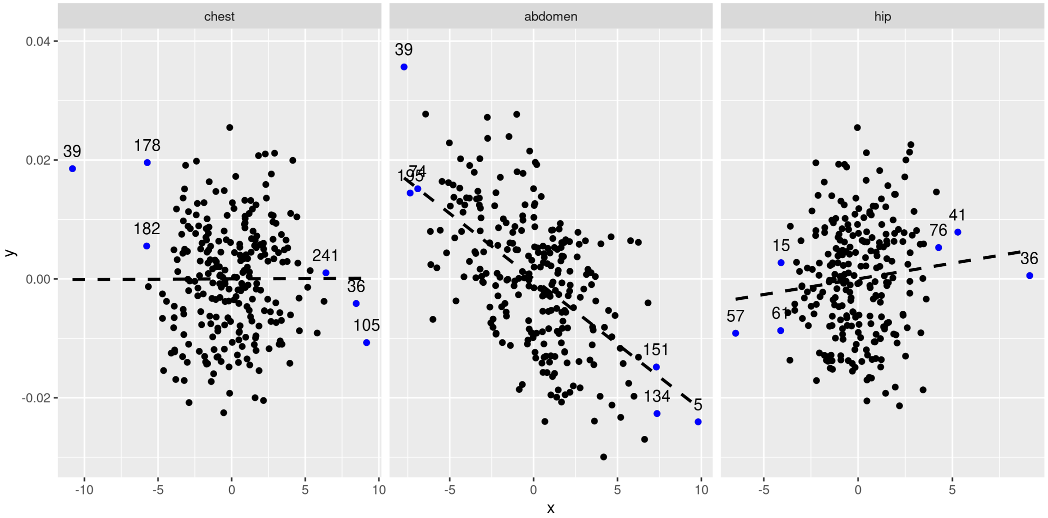


Figure 6: Partial regression plots of regressors `chest`, `abdomen`, and `hip`.

Table 2: ANOVA table for full model.

	Sum sq	Mean sq	F value	Pr(>F)
age	0.007	0.007	78.377	0
weight	0.034	0.034	351.406	0
height	0.007	0.007	72.021	0
neck	0.002	0.002	16.446	0
chest	0.001	0.001	14.799	0
abdomen	0.012	0.012	127.789	0
hip	0.000	0.000	1.437	0.232
thigh	0.001	0.001	7.291	0.007
knee	0.000	0.000	0.001	0.979
ankle	0.000	0.000	0.011	0.916
biceps	0.000	0.000	2.483	0.116
forearm	0.000	0.000	3.241	0.073
wrist	0.001	0.001	8.858	0.003
Residuals	0.022	0.000	Not applicable	Not applicable

3.3 Significance tests

Table 2 presents the Analysis of Variance table (ANOVA) for the full model. In the preceding sections we concluded that the R-student residuals seem to be randomly scattered and that the R-student residuals approximately follows a normal distribution. Therefore, we assume that the significance tests presented here are valid.

The results from the ANOVA analysis will not be covered in detail in the upcoming sections. Since our primary purpose is prediction, not explanation, the results presented here are left for the reader. Instead, we place greater emphasis on handling multicollinearity (see section 3.5) and conducting cross-validation for model development (see section 3.7), since these methods affect the stability of our predictions and generalizability of our model.

3.4 Transformations of variables

In section 2.1 we noted that there was no indication that a transformation was needed on the response variable. Here, we will see that the transformation of the response variable skews the results negatively. Figure 7 displays the values of λ to be used in a potential Box-Cox transformation of the dependent variable. The λ that maximized the log-likelihood is 0.9 (0.7-1.1 approximate 95% CI). Using $\lambda = 0.9$ gives us the quantile-quantile plot displayed on the right hand side in figure 7. We notice that this affects the distribution of residuals by making it more light-tailed.

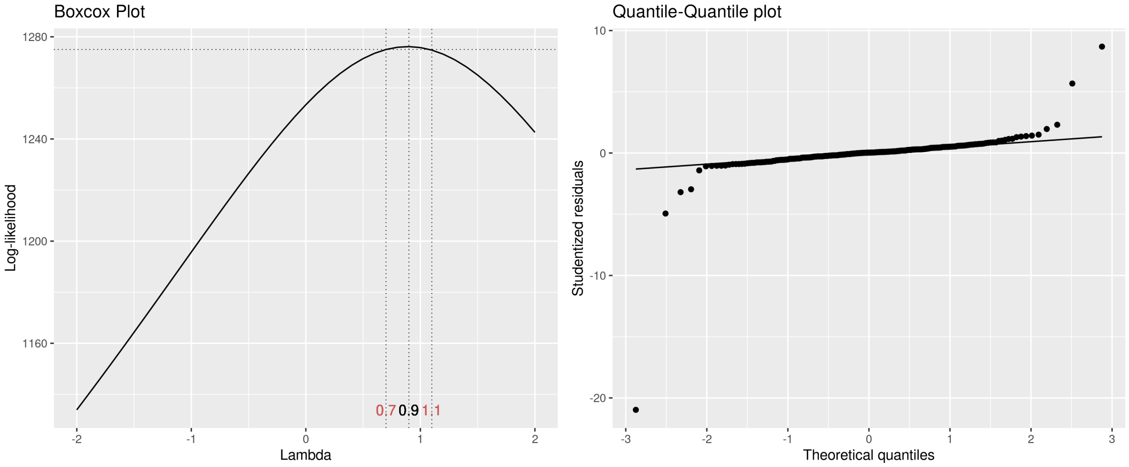


Figure 7: Values for lambda against the log-likelihood for Box-Cox transformations.

3.5 Diagnostics and Handling of Multicollinearity

Table 3 presents the VIF for each respective regressor and eigen values of the $\mathbf{XX}'$ matrix. The eigen values for the `biceps`, `forearm`, and `wrist` regressors are relatively close to zero, and the VIF of the `weight`, `chest`, `abdomen`, and `hip` regressors are larger than 10. Hence, there appears to be multicollinearity in the data.

A correlation matrix for the full model is found in section 5. The strong collinearity between the `weight` regressor and other predictors is apparent in the correlation matrix in figure 13. The `weight` regressor shows a strong correlation with all but the `age` and the `height` regressors.

In order to handle the multicollinearity in the data, we replace the variables that appear to be involved in the multicollinearity with a summary variable. [6] The summary variable is referred to as `combo` and was defined as

$$\frac{\texttt{hip} \times \texttt{thigh} \times \texttt{abdomen}}{\texttt{weight}}$$

The rationale for this particular combination of predictors was that it minimizes the MSE, as well as makes sure that the VIF are below 10 and that the eigen values of the $\mathbf{XX}'$ are kept relatively large. The resulting VIF are presented in figure 8.

The residual analysis were re-run in order to make sure that the assumptions for normality still hold. The plots are presented in 6. Also the ANOVA table including the `combo` variable is shown. We note that the effort to reduce multicollinearity did not affect the other diagnostics in a noticeable way. Therefore, we keep the summary variable and move to handling of outliers.

Table 3: Multicolinearity measures.

	age	weight	height	neck	chest	abdomen	hip	thigh	knee	ankle	biceps	forearm	wrist
Eigen	8.16	1.44	0.86	0.68	0.55	0.32	0.27	0.25	0.19	0.13	0.07	0.05	0.02
VIF	2.26	43.94	2.87	4.39	10.17	12.88	14.55	7.82	4.74	1.95	3.68	2.17	3.35

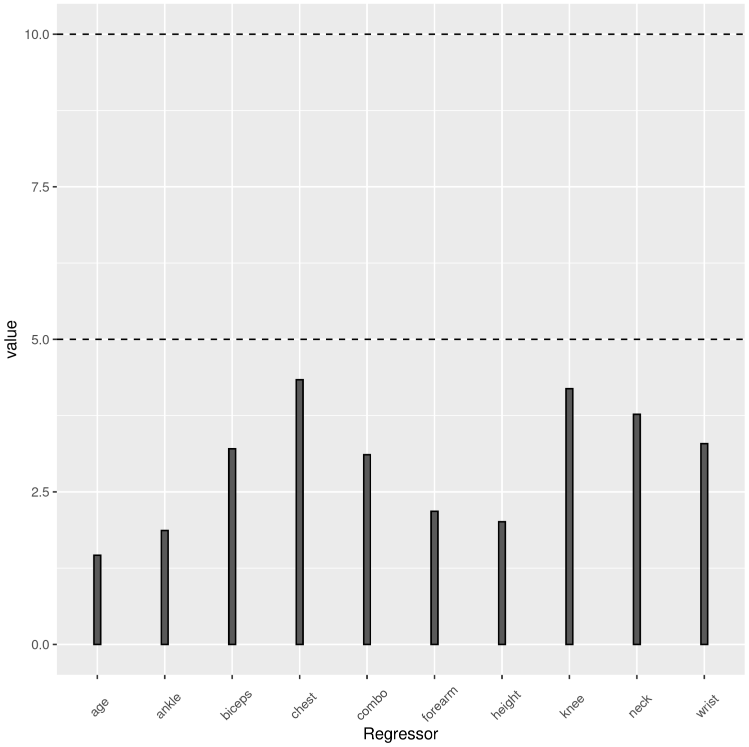


Figure 8: Variance Inflation Factors (VIF) when using the summary variable `combo`.

3.6 Diagnostics and Handling of Outliers

Figure 9 illustrates Cook's distance for all points, where the three observations with the largest Cook's distance are labelled. Considering the cut-off $D_i = 1$ as proposed in [6], where D_i is the Cook's distance for observation i, we note that none of the observations would be considered influential. Still, observation 39, 83, and 41 are large relative to the other points in terms of their Cook's distance. Noting the relative differences, rather than relying to a specific cut-off-value, has been mentioned as a diagnostic for further inspection of influential points. [2] These observations are therefore considered as influence points that may affect our model fit in a considerable way.

Figure 10 reports the $DFFITS$ values. The recommended cutoff-value mentioned in [6], i.e. $\pm 2\sqrt{\frac{p}{n}}$ where $p = 13$ is the number of potential regressors and $n = 248$ is the sample size, is plotted as a dotted line, and the points that lie below or above this cut-off value are

Table 4: Observations considered as outliers from the Cook's distance and *DFFITS* analysis.

Observation	age	height	neck	chest	knee	ankle	biceps	forearm	wrist	combo
39	46	72.25	51.2	136.2	49.1	29.6	45.0	29.0	21.4	5258.523
41	45	68.75	43.2	128.3	39.6	26.6	36.4	32.7	21.4	4373.653
83	67	67.50	36.5	98.9	37.8	33.7	32.4	27.7	18.2	2826.431
5	24	71.25	34.4	97.3	42.2	24.0	32.2	27.7	17.7	3495.294
78	67	67.75	38.4	97.7	38.2	23.7	29.4	27.2	19.0	3113.035
168	35	65.50	34.0	90.8	34.8	22.0	24.8	25.9	16.9	2660.040
212	51	64.00	41.2	119.8	36.9	23.6	34.7	29.1	18.4	3930.616

labelled. We observe that several points are considered influential points when using the this cut-off value.

Figure 16, 18, 17, and 19 in section 7 presents $DFBETA$ values for groups of regressors. Observation 39 is present in a number of these figures, as well as observation number 83 and 217. Using the aforementioned cut-off value of $\frac{2}{\sqrt{n}}$, we note however that none of these points would be considered influential points.

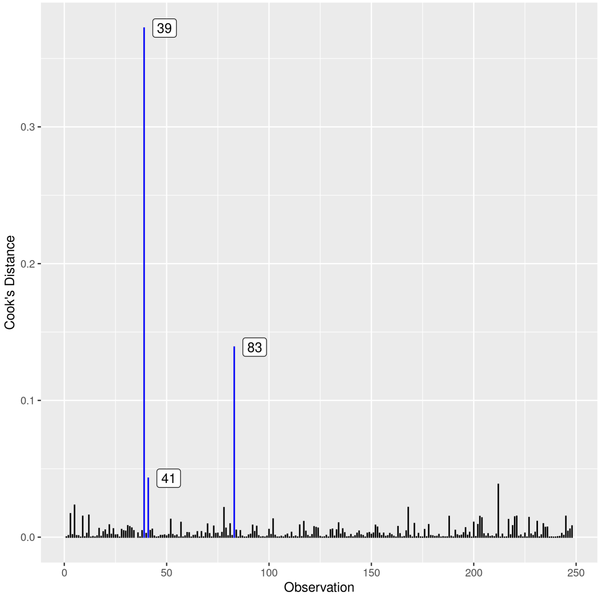


Figure 9: Cook's distance for all observations.

We present the observations noted in the Cook's distance and DFFITS plots in Table 4. The points labelled in the $DFBETA$ plots are not considered by the reason noted previously in section 2.2.3. The points that was identified as potential outliers in the added-variable plots can be compared to the points that are considered as influential in the Cook's distance plots and the DFFITS plot. For example, we see that observation 39 would be noted as an outlier in a number of added-variable plots, and is also in included as one of the more influential observations considering its DFFITS and Cook's distance values.

When handling the outliers we consider two perspectives: Cause of outlier tendencies and effect on fit of the model. Looking at the observations, and comparing it to the sample characteristics in Table 1 we note that some observations are indeed outliers in the x-space however plausible measurements, for example observation 39. In other words, they are likely not results of mis-measurements, and hence should not be removed for that reason. The second perspective, the outliers effect on the model, is discussed in section 3.7.

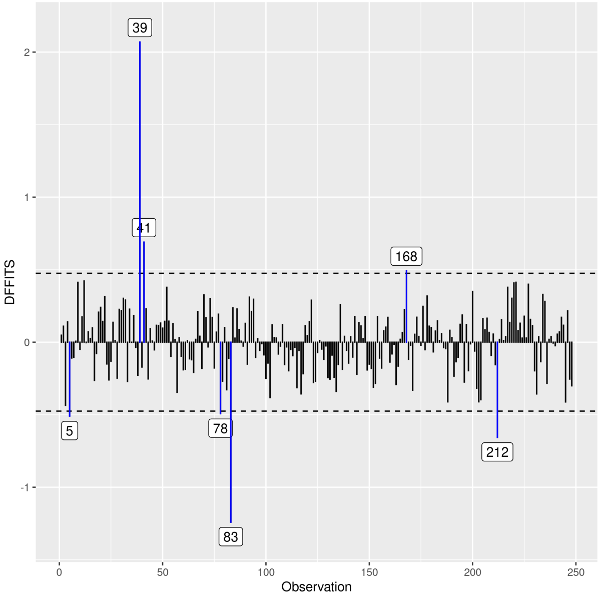


Figure 10: $DFFITS$ for all observations.

3.7 Variable selection

The measurements for BIC, the C(p) criterion, and adjusted R^2 of the best subset models are presented in figure 12. The cross-validated MSE for the full model, the model with a summary variable, and the model the summary variable without the influential observations are presented in Table 5.

Several methodological considerations were made in this step. Firstly, regarding influential and outlier observations. By removing influential observations we reduce the mean squared error by a considerable amount. However, we have no quantitative nor qualitative reason for removing them. Therefore, we will leave the outliers in the dataset.

Secondly, regarding our method of handling multicollinearity. Since our primary purpose was prediction, one could argue that we should proceeded with the model that minimizes the MSE on the test sample, that is the full model without the summary variable. In fact, the predictions made by the model may still be perfectly accurate even if the model is misspecified, as long as predictions are mode on observations within or close to the x-space on which the model was fitted. [6] We would argue, however, that by handling multicollinerity we ensure stable least-squares estimators for the model, and hence predictions that

Table 5: Cross-validated MSE for models.

Type	MSE
Full model	0.0001116
With summary variable	0.0001378
Summary variable without inf. obs.	0.0001146

more valid outside the training x-space. In doing so, we sacrifice a gain in MSE. There are also other methods of handling multicollinearity that were not considered here, for example Principal Component Regression (PCR) or ridge regression, that could have been better options for our purpose.

Thirdly, the choice to bootstrap confidence intervals around the model coefficients. Another method would be bootstrap prediction intervals [1]. This would arguably be more useful for our prediction purpose. However, the CI boostrap around the regression coefficients give us a confidence estimate around the stability of the coefficients of our model and is therefore useful for prediction.

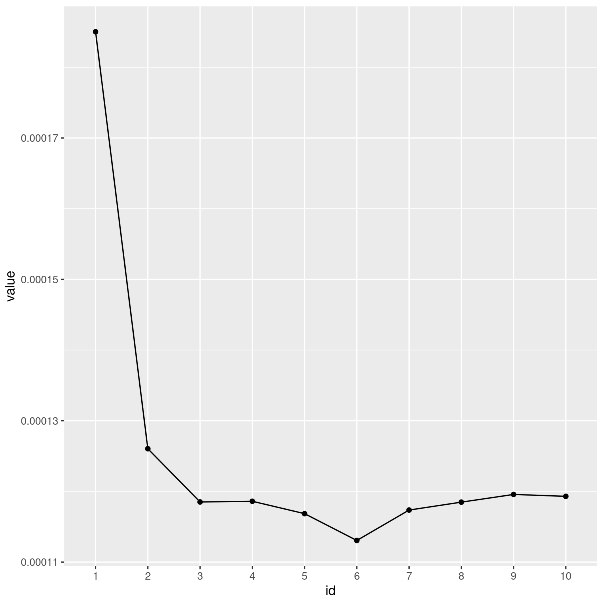


Figure 11: Cross-validated mean squared error for the best subset model for each number of regressors.

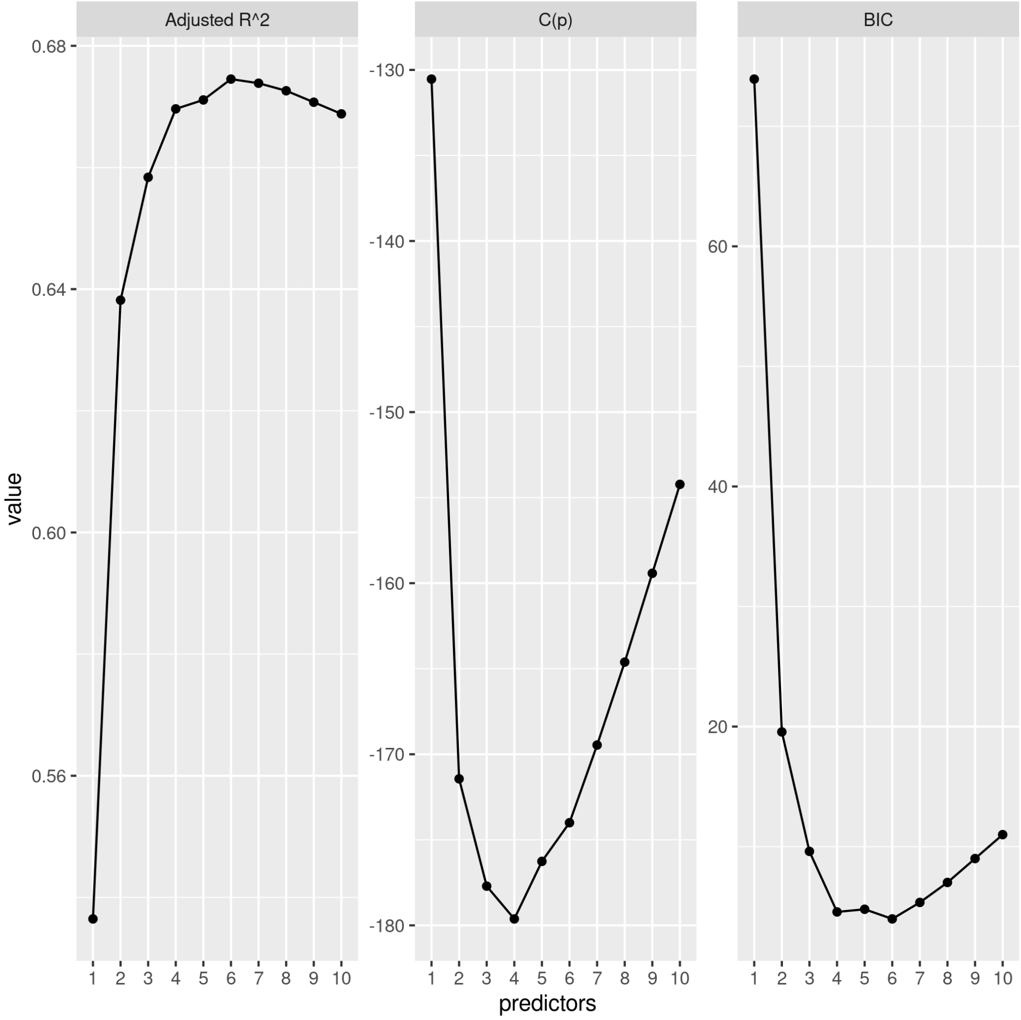


Figure 12: Number of regressors against multiple performance measures for the best subset models.

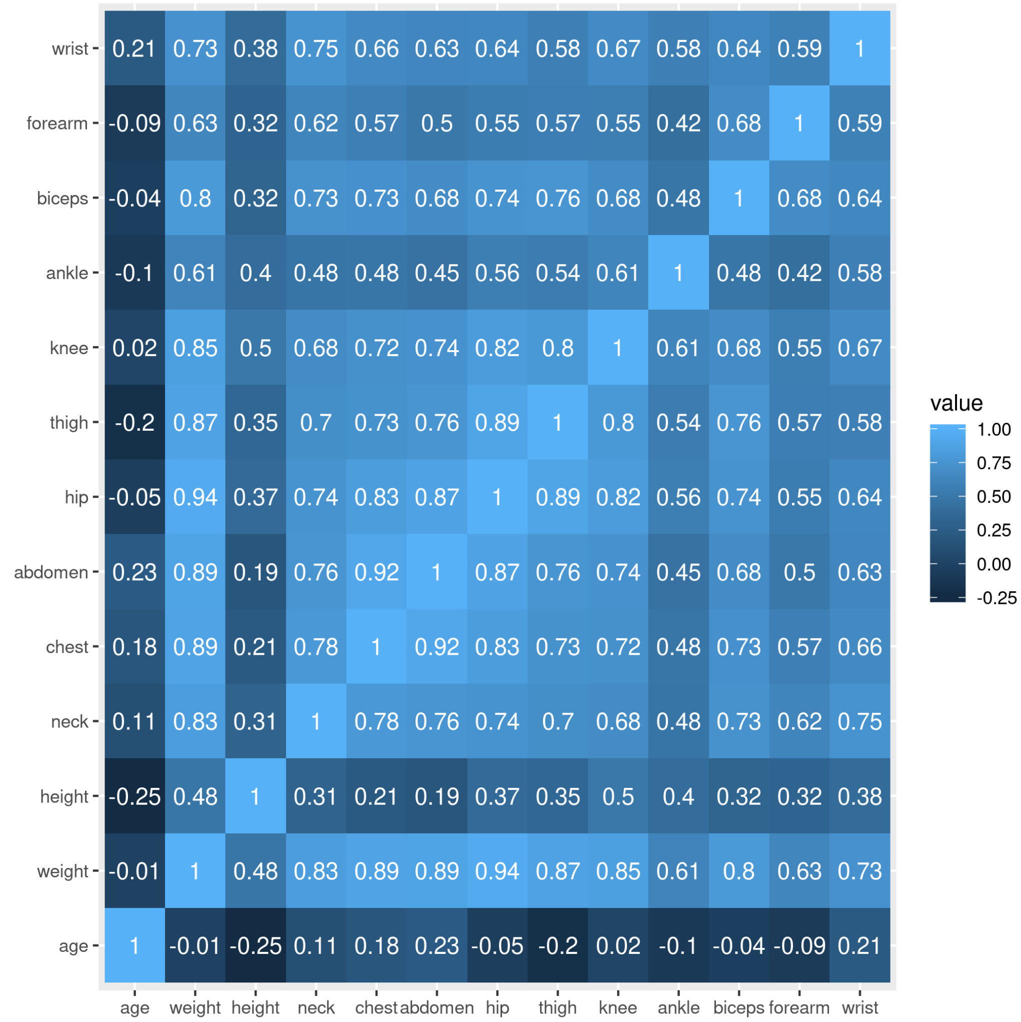
4 Conclusion

The most well performing model, determined by its cross-validated mean squared error, its predictors and the corresponding coefficients along with 95% confidence intervals are presented in Table 6.

Table 6: Coefficients (95% CI) of final model.

Predictor	Coefficient (95 %)
(Intercept)	1.19078 (1.14198 to 1.21116)
age	-3e-04 (-0.00044 to -0.00022)
chest	-0.00102 (-0.00131 to -0.00073)
wrist	0.00329 (0.00173 to 0.00583)
combo	-3e-05 (-3e-05 to -2e-05)

5 Appendix A



	age	weight	height	neck	chest	abdomen	hip	thigh	knee	ankle	biceps	forearm	wrist
wrist	0.21	0.73	0.38	0.75	0.66	0.63	0.64	0.58	0.67	0.58	0.64	0.59	1
forearm	-0.09	0.63	0.32	0.62	0.57	0.5	0.55	0.57	0.55	0.42	0.68	1	0.59
biceps	-0.04	0.8	0.32	0.73	0.73	0.68	0.74	0.76	0.68	0.48	1	0.68	0.64
ankle	-0.1	0.61	0.4	0.48	0.48	0.45	0.56	0.54	0.61	1	0.48	0.42	0.58
knee	0.02	0.85	0.5	0.68	0.72	0.74	0.82	0.8	1	0.61	0.68	0.55	0.67
thigh	-0.2	0.87	0.35	0.7	0.73	0.76	0.89	1	0.8	0.54	0.76	0.57	0.58
hip	-0.05	0.94	0.37	0.74	0.83	0.87	1	0.89	0.82	0.56	0.74	0.55	0.64
abdomen	0.23	0.89	0.19	0.76	0.92	1	0.87	0.76	0.74	0.45	0.68	0.5	0.63
chest	0.18	0.89	0.21	0.78	1	0.92	0.83	0.73	0.72	0.48	0.73	0.57	0.66
neck	0.11	0.83	0.31	1	0.78	0.76	0.74	0.7	0.68	0.48	0.73	0.62	0.75
height	-0.25	0.48	1	0.31	0.21	0.19	0.37	0.35	0.5	0.4	0.32	0.32	0.38
weight	-0.01	1	0.48	0.83	0.89	0.89	0.94	0.87	0.85	0.61	0.8	0.63	0.73
age	1	-0.01	-0.25	0.11	0.18	0.23	-0.05	-0.2	0.02	-0.1	-0.04	-0.09	0.21

Figure 13: Correlation matrix of the full model

6 Appendix B

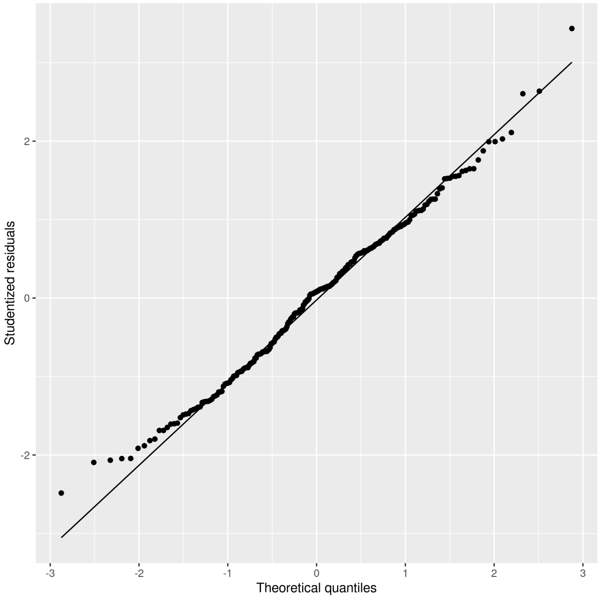


Figure 14: Normality plot of residuals when using the `combo` variable.

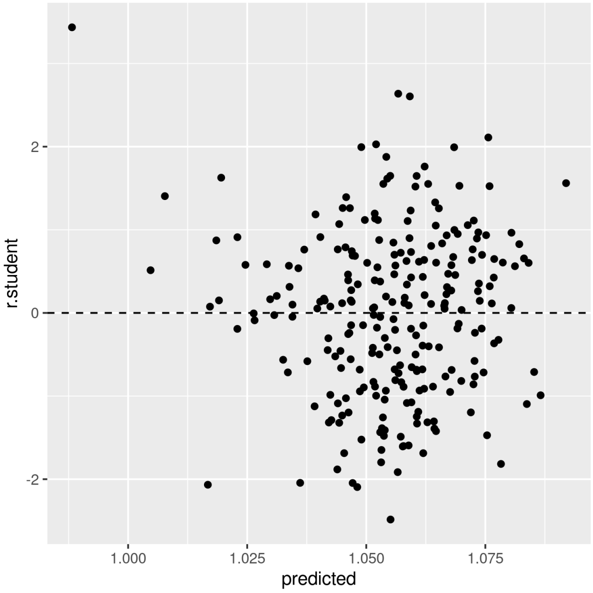


Figure 15: R-student residuals agaist fitted values when using the `combo` variable.

Table 7: ANOVA table for model with `combo` variable.

	Sum sq	Mean sq	F value	Pr(>F)
age	0.007	0.007	60.715	0
height	0.000	0.000	2.768	0.097
neck	0.020	0.020	160.383	0
chest	0.020	0.020	164.179	0
knee	0.001	0.001	10.078	0.002
ankle	0.000	0.000	1.553	0.214
biceps	0.000	0.000	0.962	0.328
forearm	0.000	0.000	0.057	0.811
wrist	0.002	0.002	14.472	0
combo	0.007	0.007	59.082	0
Residuals	0.029	0.000	Not applicable	Not applicable

7 Appendix C

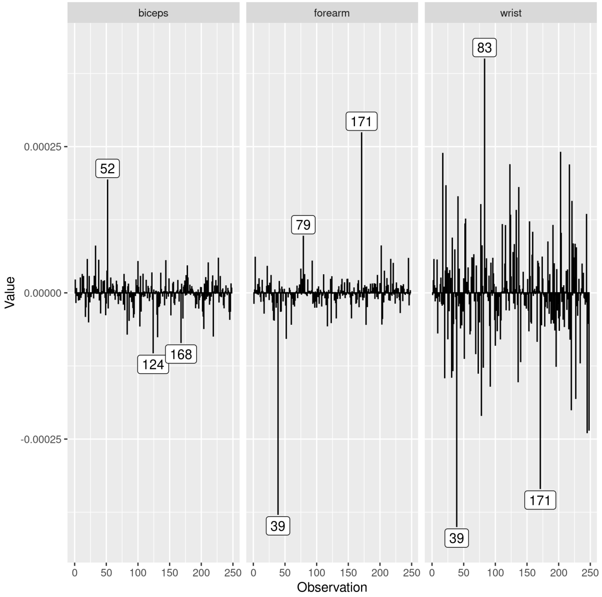


Figure 16: *DFBETA* for regressors `biceps`, `forearm`, and `wrist`.

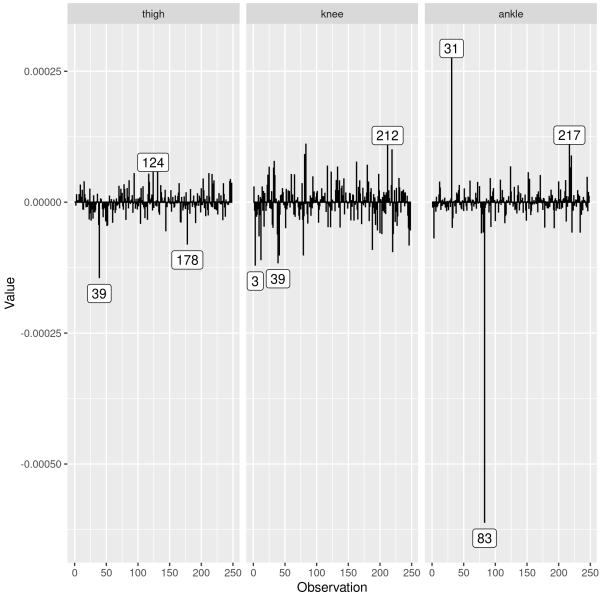


Figure 17: $DFBETA$ for regressors `thigh`, `knee`, and `ankle`.

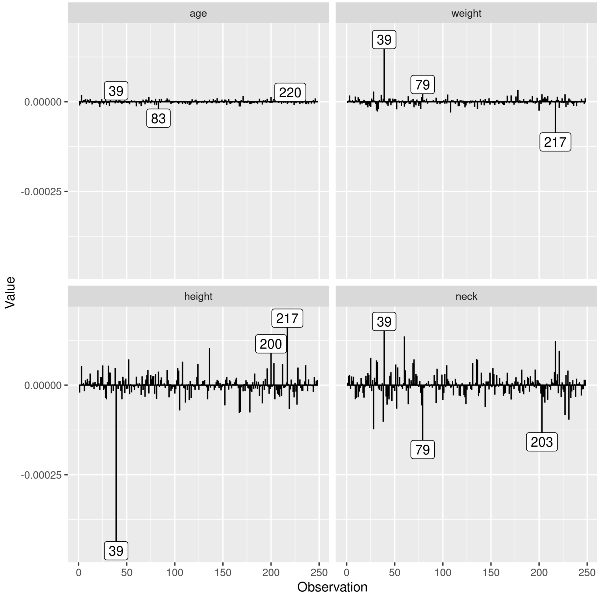


Figure 18: $DFBETA$ for regressors `age`, `weight`, `height` and `neck`.

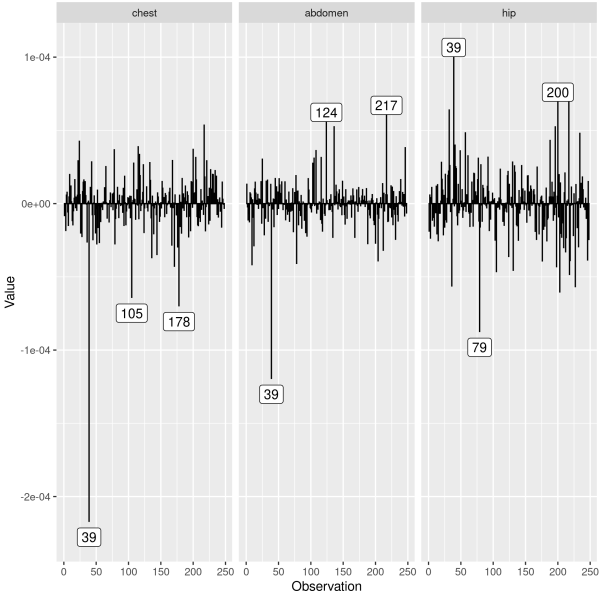


Figure 19: $DFBETA$ for regressors `chest`, `abdomen`, and `hip`.

References

[1] A. C. Davison and D. V. Hinkley. *Bootstrap Methods and their Application.* Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 1997.

[2] John Fox. *Regression Diagnostics: An Introduction.* Sage Publications, 1991.

[3] gung Reinstate Monica (https://stats.stackexchange.com/users/7290/gung-reinstate monica). How to read cook's distance plots? Cross Validated. URL:https://stats.stackexchange.com/q/22286 (version: 2012-02-05).

[4] gung Reinstate Monica (https://stats.stackexchange.com/users/7290/gung-reinstate monica). Algorithms for automatic model selection. Cross Validated. URL:https://stats.stackexchange.com/q/20856 (version: 2012-01-11).

[5] T. Hastie, R. Tibshirani, and J.H. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* Springer series in statistics. Springer, 2009.

[6] Douglas C Montgomery, Elizabeth A Peck, and G Geoffrey Vining. *Introduction to Linear Regression Analysis.* Wiley-Interscience, 5 edition, 2012.