Kungliga Tekniska Högskolan

SF2930 Regression Analysis

Report I

Isac Karlsson
Ludvig Wärnberg Gerdin

Examiner
Tatjana Pavlenko

February 27, 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). There exists very accurate methods for calculating BFM but because of high costs and efforts cheaper methods such as regression models are widely used.

The given dataset (BFM MEN) 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

2.1 Residual analysis

Some major assumptions we use in our analysis are:

1. The errors ϵ_i for observation i are iid. 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, or error, 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

It is possible to use an externally studentized residual given by [1]

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

which is often called R-student. 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 impact because, for example, prediction intervals depend on the normality assumption. One may check the normality assumption simple by constructing a normal probability plot of the residuals.

The normality of residuals therefore ensures that the confidence intervals presented in section 3 are valid.

2.1.3 Fitted 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.

2.1.5 Other useful plots

One may want to analyze the possibility of multicollinearity being present in the data. Knowing that this can disturb the least-squares fit in ways that results in the regression model ending up being nearly useless. One way to do this is by create a scatter-plot of two regressors against each other (i.e. analyzing the relationship between regressor variables. If two regressors are correlated one may not need to include them both in the model. If they are highly correlated the mentioned possibility of multicollinearity is larger.

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 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. Each element h_{ij} corresponds to the amount of leverage exercised by the ith observation y_i on the jth, fitted value, $\hat{y_j}$.

The hat matrix is can be very useful when detecting potential outliers, since it determines the variances and covariances of $\hat{y}$ and e.

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.

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 $ 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.

2.3 Transformations of variables

2.4 Diagnostics and handling of Multicolinearity

Table 1: 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 Results

3.1 Significance tests

Table 1 presents the ANOVA table for the full model.

3.2 Residual analysis

3.2.1 Normality of residuals

Figure 1 illustrates QQ plot of the model residuals. The observer may say that the points exhibit a pattern that indicates that the residuals come from a distribution with heavier tails than that of a normal distribution. [1]. Still, the deviations from the diagonal line is relatively small, and hence we conclude that the first Gauss-Markov condition is fulfilled. That is, the model errors seem to be 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 horizontal

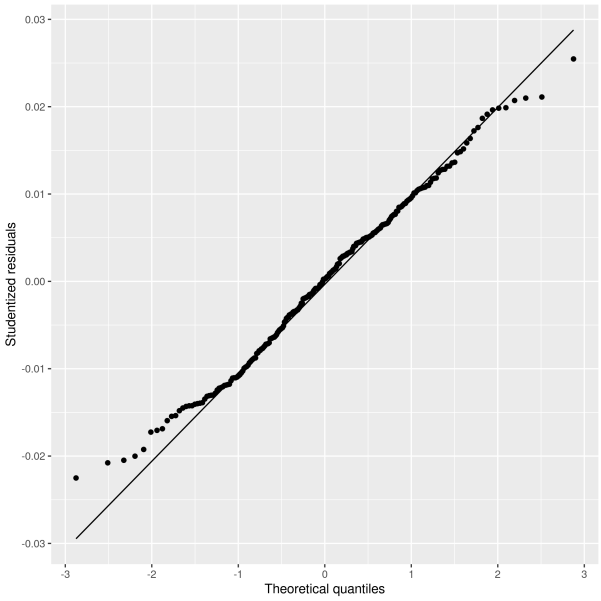

Figure 1: Normality plot of residuals.

line. Hence we conclude that the second Gauss-Markov condition is fulfilled, that is the errors have a constant variance.

3.2.3 Added Variable Analysis

Partial regression plots are found in figure 3, 4, 5, and 6. All figures exhibits potential outliers (which will be further considered in section 2.2). More specifically, in figure 3 we note a few potential outliers on the right hand side of the plot for the `biceps` regressor, and on the right and left hand side for the `forearm` regressor. Moreover, in figure 4, we notice outliers on the right hand side of the `ankle` plot, and a group of potential outliers on the `thigh` plot. Finally, we notice a few potential outliers in figure 5 and 6.

Figure 4, 5, and 6 conveys important information about the information that `knee`, `height`, and `chest` adds to the model. These regressors seem to follow a horizontal band along a fitted line from the origin, which may suggest that none of the regressors adds additional information to the predictions.

3.3 Transformations of variables

Figure 7 displays the values of λ to be used in a potential Box-Cox transformation of the dependent variable `density`. The λ that maximized the log-likelihood is 0.9 (0.7-1.1 95% CI).

Using $\lambda = 0.9$ gives us the normal probability 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. That is, the the tails of the distribution are too light for the distribution to be considered normal.

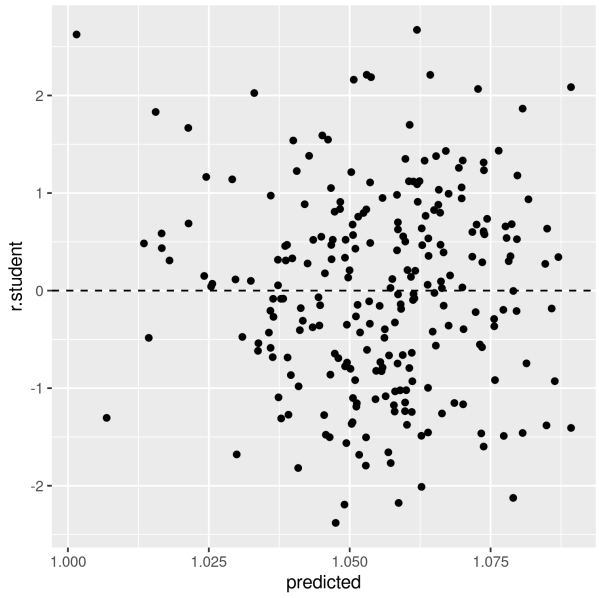

Figure 2: Fitted values against R-student residuals.

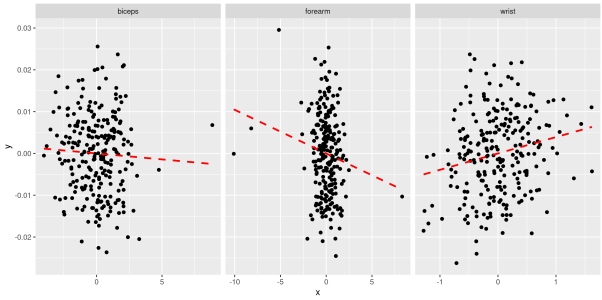

Figure 3: Partial regression plots of regressors `biceps`, `forearm`, and `wrist`.

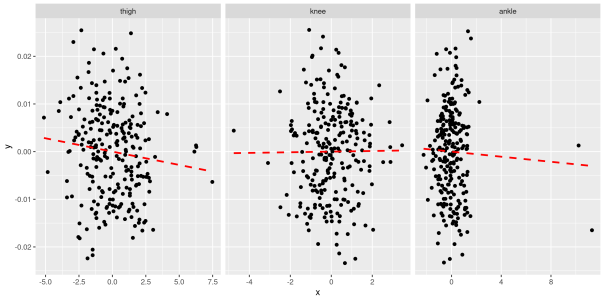

Figure 4: Partial regression plots of regressors `thigh`, `knee`, and `ankle`.

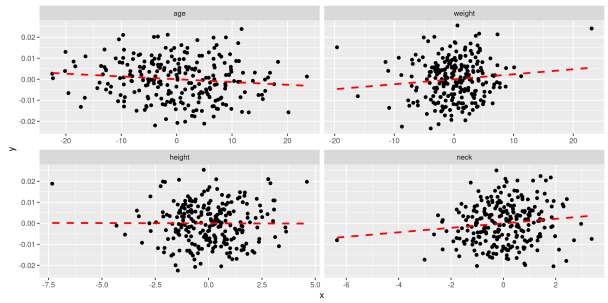

Figure 5: Partial regression plots of regressors `age`, `weight`, `height`, and `neck`.

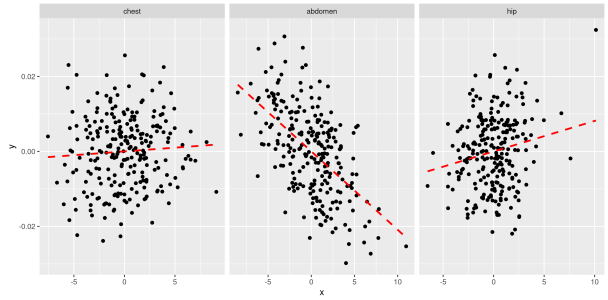

Figure 6: Partial regression plots of regressors `chest`, `abdomen`, and `hip`.

In section 2.1 we noted that there was no indication that a transformation was needed. Here, we see that the transformation of the response variable only makes matters worse.

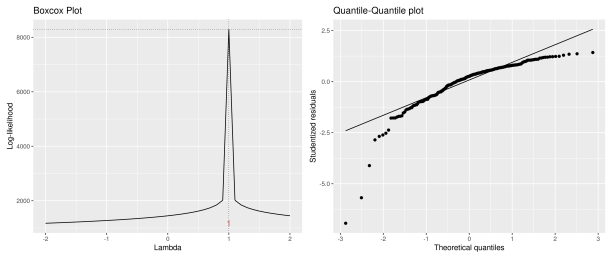

Figure 7: Values for lambda against the log-likelihood of `density` for Box-Cox transformations.

3.4 Diagnostics and handling of Outliers

Figure 8 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 [1], 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 and 83 are largely different relative to the other points in terms of their Cook's distance.

Figure 9 reports the $DFFITS$ values. We label observations as in figure 8. We observe

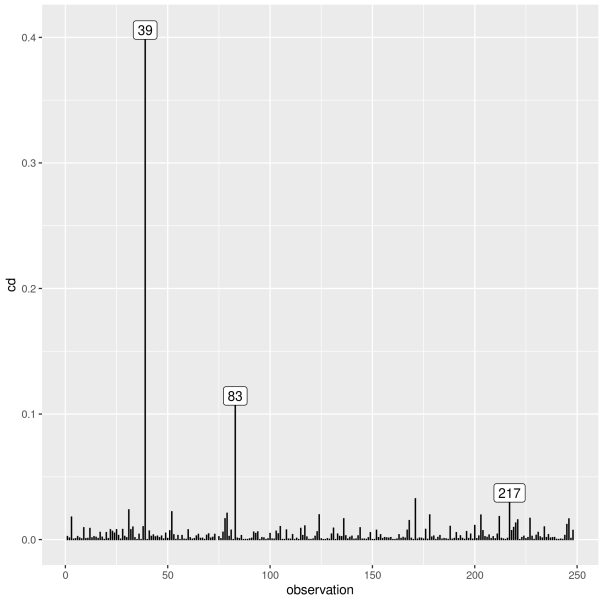

Figure 8: Cook's distance for all observations.

that the three largest absolute $DFFITS$ correspond to the same observations as in the Cook's distance plot. The recommended cutoff-value referred to in [1], i.e. $2\sqrt{\frac{p}{n}}$ 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 is labelled. We observe that several points are considered influential points when using that cut-off value.

Figure 10, 11, 12, and 13 presents $DFBETA$ values for groups of regressors. Observation 39 is present in a number of these figures. Using the aforementioned cut-off value of $\frac{2}{\sqrt{n}}$, we we note that none of these points would be considered influential points.

3.5 Diagnostics and Handling of Multicolinearity

Table 2 presents the VIF and eigen value for each respective regressor. 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 (NOTERA DETTA I METOD). Hence, there appears to be an indication of multicolinearity amongst the candidate regressors.

A coorelation matrix for the full model is found in section 5. The strong multicolinearity associated with the `weight` regressor is apparent in the correlation matrix in figure 16. The `weight` regressor shows a strong correlation with all but the `age` and the `height` regressors.

In a later stage we use all possible regression in order to determine the candidate regression models. In that stage we further examine the results of removing the regressors with an indication of high multicolinearity.

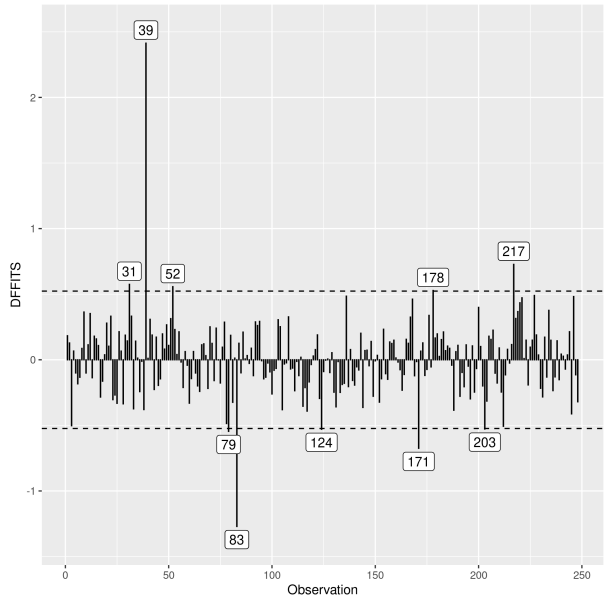

Figure 9: $DFFITS$ for all observations.

Table 2: 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

3.6 Variable selection

The full results of the best subsets regression is presented in section 6. The model corresponding to each model index in Table 3 in 6 is presented in Table 4. The best subset plots are presented in figure 15.

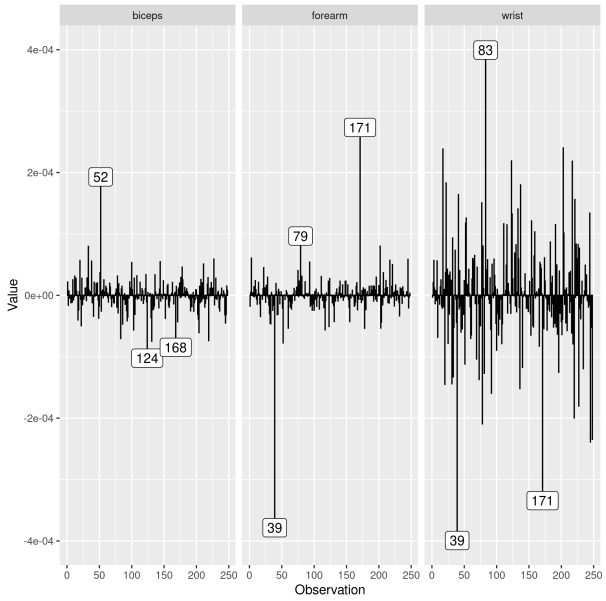

Figure 10: $DFBETA$ for regressors `biceps`, `forearm`, and `wrist`.

4 Conclusion

References

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

5 Appendix A

6 Appendix B

Table 3: Coefficients (95% CI) of final model.

Predictor	Coefficient (95 %)
(Intercept)	1.1458 (1.087 to 1.1935)
age	-1e-04 (-3e-04 to 0)
weight	2e-04 (0 to 4e-04)
neck	0.0011 (2e-04 to 0.0022)
abdomen	-0.0022 (-0.0025 to -0.0018)
hip	6e-04 (-1e-04 to 0.0012)
thigh	-7e-04 (-0.0012 to -2e-04)
forearm	-0.0012 (-0.0021 to -2e-04)
wrist	0.0033 (0.001 to 0.0053)

Table 4: Performance measures for candidate models, where p refers to the number of regressors

Model index	p	Adjusted R-squared	AIC	BIC	C(p)
1	1	0.6592859	-1528.774	-2233.467	67.563246
2	2	0.7122838	-1569.713	-2273.834	20.030208
3	3	0.7191761	-1574.741	-2278.747	14.709324
4	4	0.7262143	-1580.054	-2283.820	9.307883
5	5	0.7281405	-1580.828	-2284.453	8.557443
6	6	0.7297570	-1581.334	-2284.787	8.098682
7	7	0.7322138	-1582.630	-2285.817	6.902379
8	8	0.7345127	-1583.804	-2286.676	5.864506
9	9	0.7345759	-1582.902	-2285.580	6.821711
10	10	0.7343691	-1581.753	-2284.240	8.017302
11	11	0.7332561	-1579.765	-2282.131	10.006266
12	12	0.7321254	-1577.769	-2280.015	12.002451
13	13	0.7309834	-1575.772	-2277.897	14.000000

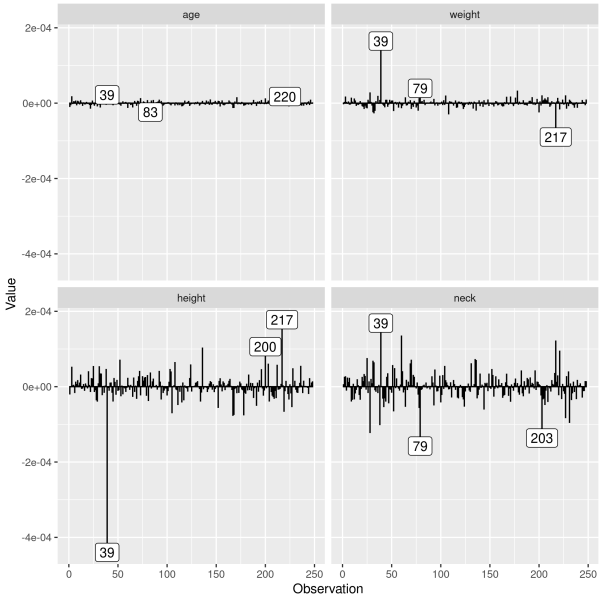

Figure 11: $DFBETA$ for regressors `age`, `weight`, `height` and `neck`.

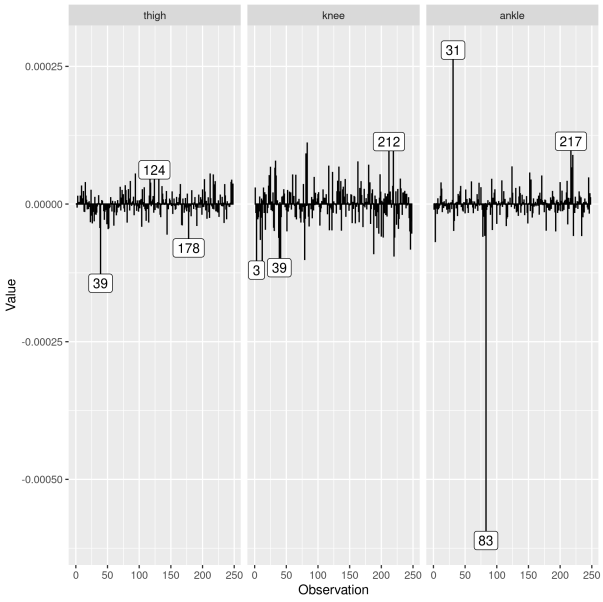

Figure 12: $DFBETA$ for regressors `thigh`, `knee`, and `ankle`.

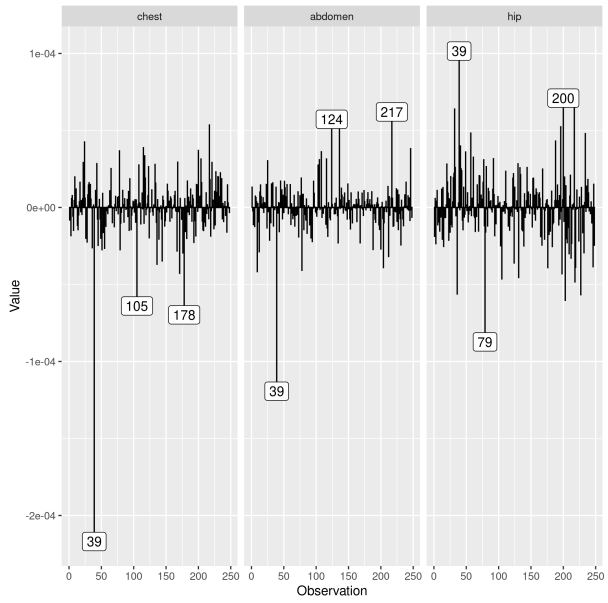

Figure 13: $DFBETA$ for regressors `chest`, `abdomen`, and `hip`.

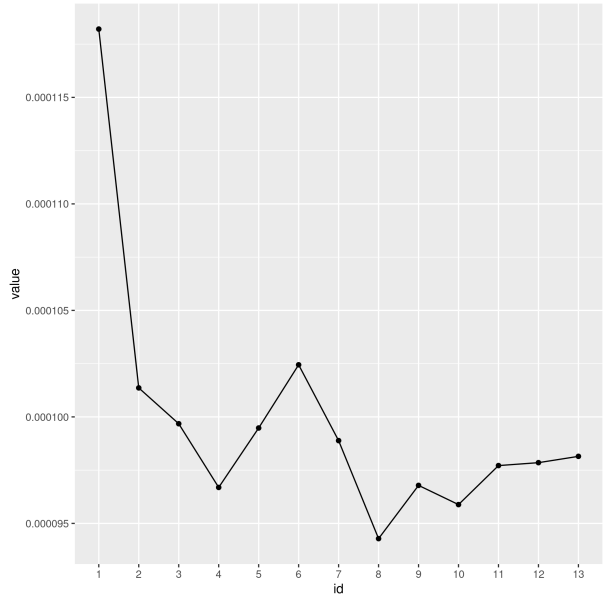

Figure 14: Cross-validated mean squared error for the best subset model and number of regressors.

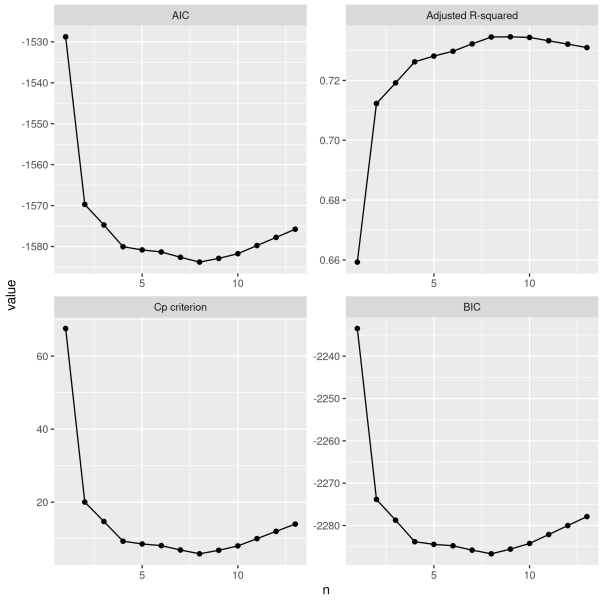

Figure 15: Number of regressors against multiple performance measures for the best subset regression models.

Table 5: Subset model corresponding to each model index in Table 3.

Model index	Regressors
1	abdomen
2	weight abdomen
3	weight abdomen wrist
4	weight abdomen forearm wrist
5	weight neck abdomen forearm wrist
6	weight neck abdomen biceps forearm wrist
7	age weight neck abdomen thigh forearm wrist
8	age weight neck abdomen hip thigh forearm wrist
9	age weight neck abdomen hip thigh biceps forearm wrist
10	age weight neck abdomen hip thigh ankle biceps forearm wrist
11	age weight height neck abdomen hip thigh ankle biceps forearm wrist
12	age weight height neck abdomen hip thigh knee ankle biceps forearm wrist
13	age weight height neck chest abdomen hip thigh knee ankle biceps forearm wrist

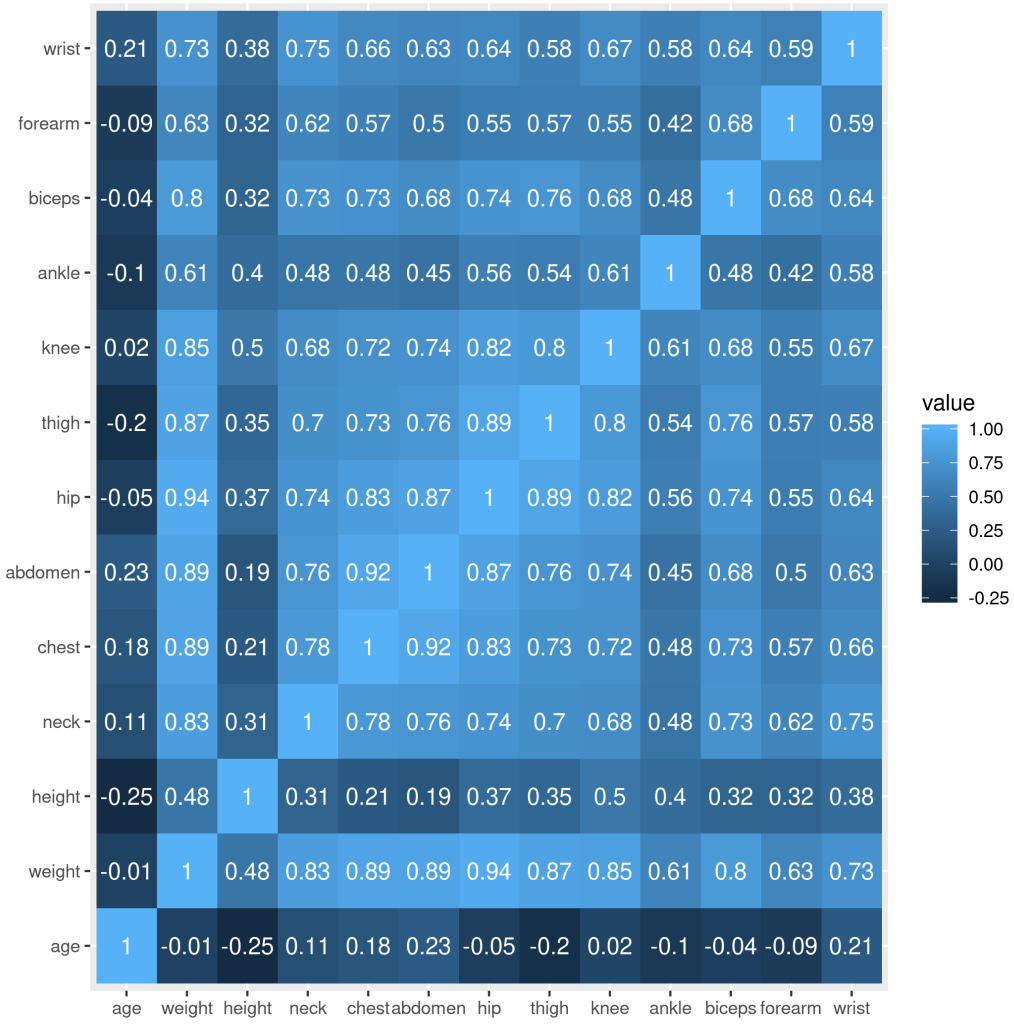

Figure 16: Correlation matrix of the full model