Kungliga Tekniska Högskolan

SF2930 Regression Analysis

Report I

Isac Karlsson
Ludvig Wärnberg Gerdin

Examiner
Tatjana Pavlenko

March 5, 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 [3]

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

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.

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. [2]

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. 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 [2], 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

2.4 Diagnostics and handling of Multicolinearity

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 characteristics will be interesting later when comparing to 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 come from a distribution with heavier tails than that of a normal distribution. [3]. 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 horizontal line. Hence we conclude that the errors have constant variance.

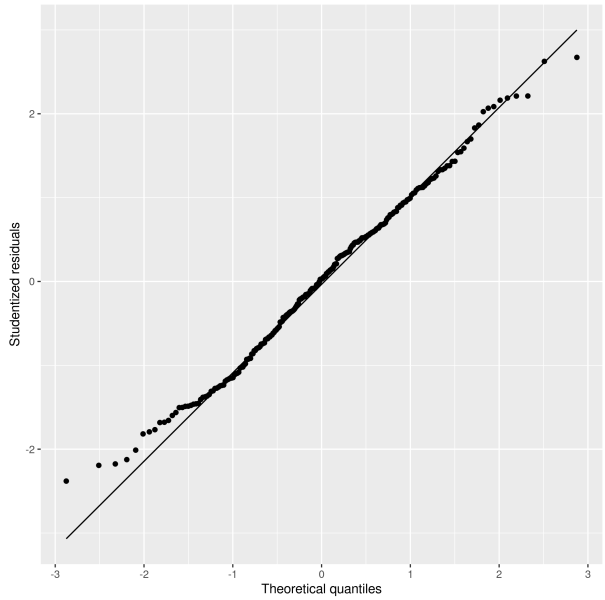

Figure 1: Normality plot of residuals.

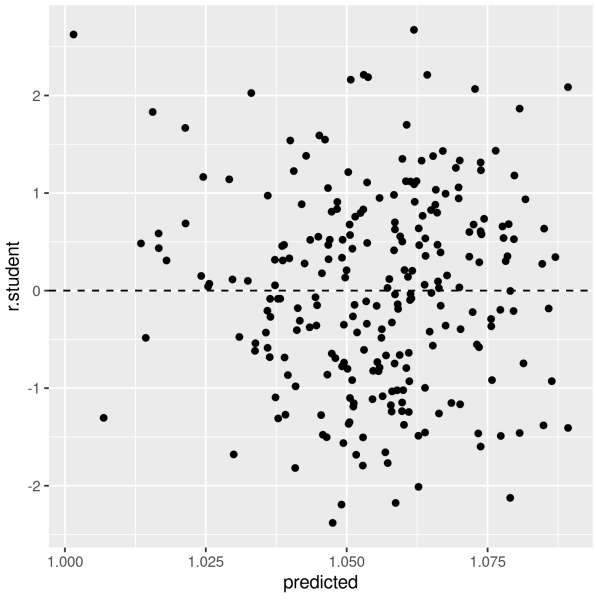

Figure 2: Fitted values against R-student residuals.

3.2.3 Added Variable Analysis

Partial regression plots are found in figure 3, 4, 5, and 6. All figures exhibit potential points that aren't adjacent to majority of the observations and hence their influence on the model fit should be examined further. This will be considered in section 2.2.

Figure 5, and 6 convey important information about the `height`, and `chest` predictors. The `height` regressors exhibit a double-bow pattern, indicating that a transformation on the height regressor could be suitable [3]. This is adjusted for in the upcoming section.

The `chest` regressor follows a horizontal band, suggesting that the predictor does not add any further information to the model. [3] This will be further considered in upcoming sections.

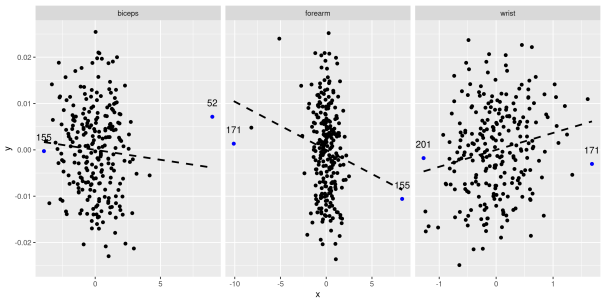

Figure 3: Partial regression plots of regressors `biceps`, `forearm`, and `wrist`.

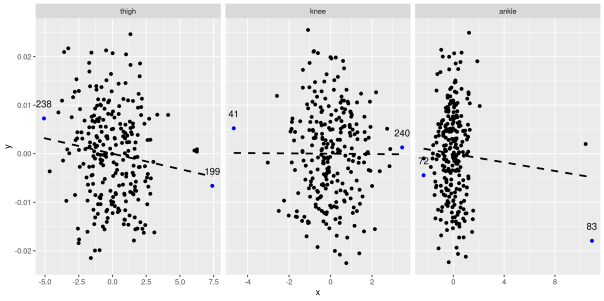

Figure 4: Partial regression plots of regressors `thigh`, `knee`, and `ankle`.

3.3 Significance tests

Table 2 presents the Analysis of Variance table (ANOVA) for the full model. Using a 5% significance level, we see that neither of the predictors `hip`, `knee`, `ankle`, `biceps`, and `forearm` are significant. This indicates that the predictors should be further examined in order to determine whether they should be included in the 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

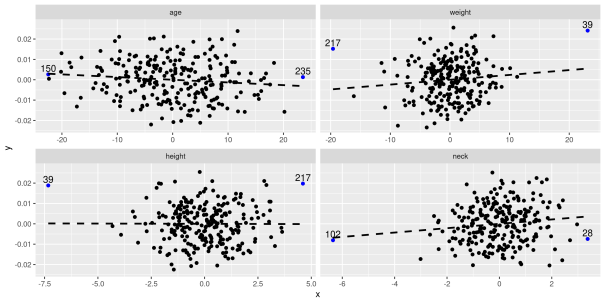

Figure 5: Partial regression plots of regressors `age`, `weight`, `height`, and `neck`.

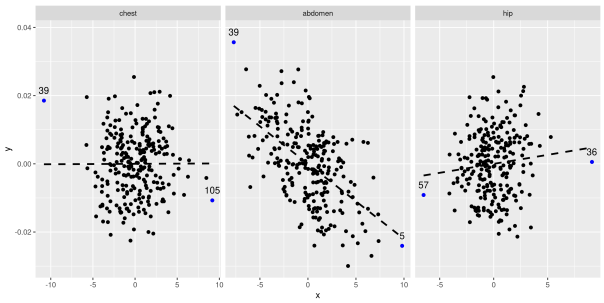

Figure 6: Partial regression plots of regressors `chest`, `abdomen`, and `hip`.

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

skews the results negatively. 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 approximate 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 tails of the distribution are too light for the distribution to be considered normal.

In section 2.1, however, we noted that the relationship between the `height` regressor and the reponse variable was misspecified as indicated by the partial regression plot. In order to correct this we specify the `height` regressor as $1/\texttt{height}^2$. The resulting added-variable plot is shown in figure 8.

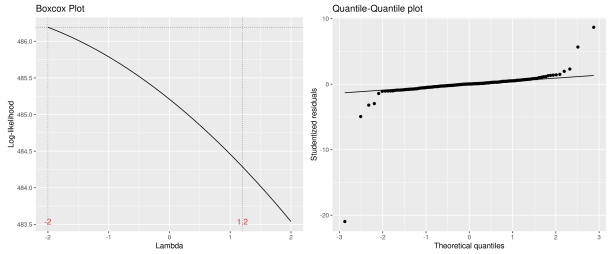

Figure 7: Values for lambda against the log-likelihood of `density` for Box-Cox transformations.

/home/ludvigwgerdin/courses/Regression Analysis/regone/main/height_av.png

Figure 8: The partial regression plot for $1/\texttt{height}^2$

Table 3: Multicolinearity measures.

	age	weight	height	neck	chest	abdomen	hip	thigh	knee	ankle	biceps	forearm	w
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	
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	

3.5 Diagnostics and Handling of Multicolinearity

Table 3 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. Hence, there appears to be multicolinearity amongst the candidate regressors.

A coorelation matrix for the full model is found in section 6. The strong multicolinearity associated with the `weight` regressor is apparent in the correlation matrix in figure 14. 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 involved variables with a summary variable. When starting to replace variables, we noted that the multicolinearity shifted, resulting in a summary variable `combo` that was defined as $\frac{\texttt{hip}\times\texttt{thigh}\times\texttt{abdomen}}{\texttt{weight}}$ The resulting VIF are presented in figure 9. We now note that none of the VIF exceed 10.

Now that we've removed several predictors and replaced in with a summary variable, we have conduct the residual analysis and observe the effect on the analysis. The plots are presented in 7. We note that the effort to reduce multicolinearity did not affect the other diagnostics in a noticeable way. Therefore, we keep the summary variable and move to handling of outliers.

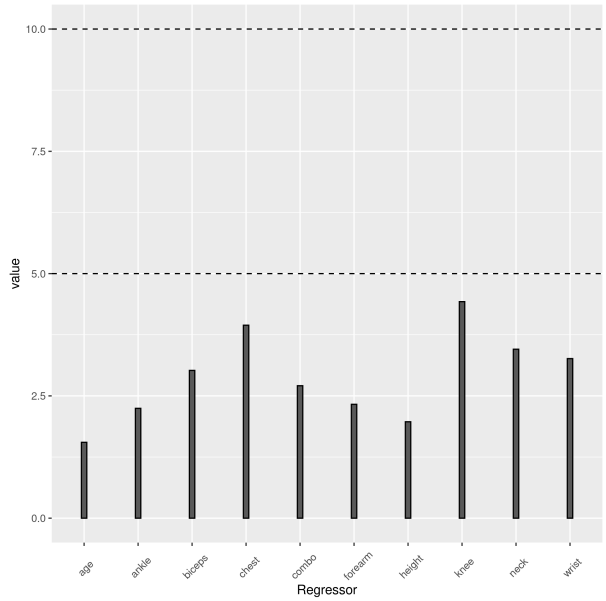

Figure 9: Variance Inflation Factors (VIF) when using the summary variable `combo`.

3.6 Diagnostics and Handling of Outliers

Figure 10 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 [3], 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, which have been mentioned as a diagnostic for further inspection of outliers. [1] The three points are henceforth considered as outliers that may influence our model fit in a considerable way.

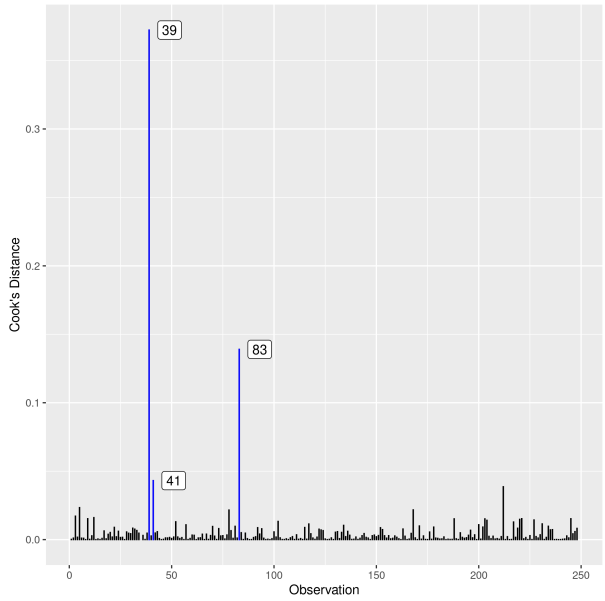

Figure 10: Cook's distance for all observations.

Figure 11 reports the $DFFITS$ values. The recommended cutoff-value mentioned in [3], 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 is labelled. We observe that several points are considered influential points when using that cut-off value. The same values as from the Cook's distance plot appear in the $DFFITS$ plot, but also several new points

Figure 15, 17, 16, and 18 in section 8 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.

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.

We analyse these observations from two perspectives: Cause of outlier tendencies and effect on fit of the model. Looking at the observations, we note that some observations are unlikely but still plausible measurements, for example observation 39. In other words,

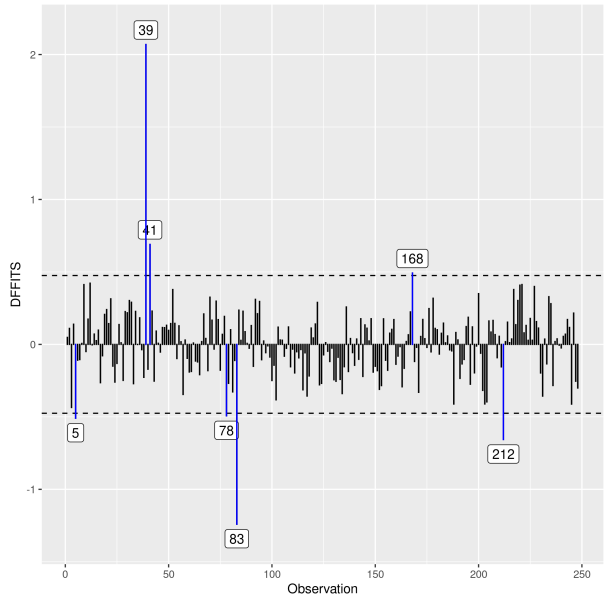

Figure 11: $DFFITS$ for all observations.

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

they are not likely a result of mis-measurements, and hence should not be removed for that reason. The second perspective is handled section 3.7.

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 13. The AIC is not included here. Still, the metric would have conveyed the same information regarding the most well-performing model as did the BIC. We note that the most well-performing model differ depending on the metric.

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. The cross-validated MSE for the full model, the model with a summary variable, and the model with a summary variable without the influential observations

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

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

Predictor	Coefficient (95 %)
(Intercept)	1.1959 (1.1661 to 1.2199)
age	-3e-04 (-4e-04 to -2e-04)
neck	0.0012 (-7e-04 to 0.0014)
chest	-0.0011 (-0.0013 to -7e-04)
forearm	-0.0012 (-0.0017 to 0)
wrist	0.0029 (0.0011 to 0.0059)
combo	0 (0 to 0)

presented in section 3.6, is presented in table 5.

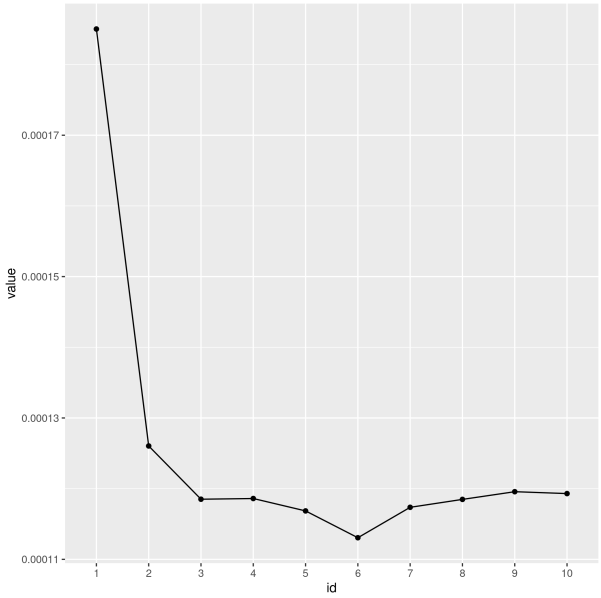

Figure 12: Cross-validated mean squared error for the best subset model and number of regressors.

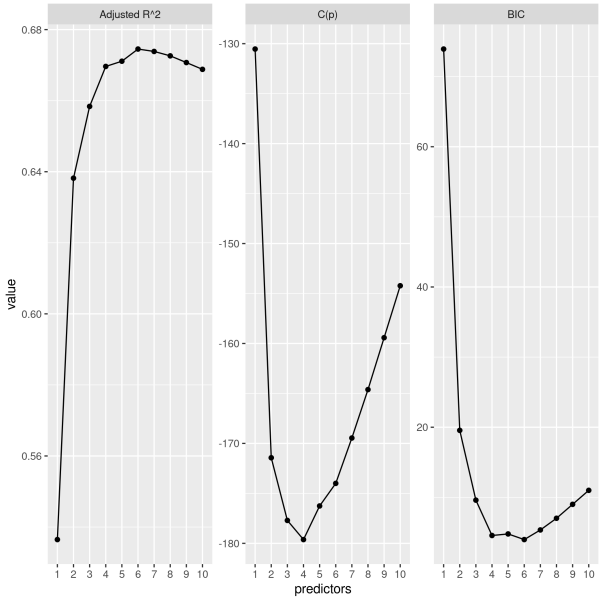

Figure 13: Number of regressors against multiple performance measures for the best subset regression models.

4 Discussion

There are contextual dimensions that are missing in order for us to make an adequate decision on whether the model should be implemented. Firstly, we do not take into the account of the cost of making the measurements for prediction, and thus. Second, we do not know the patients that the model is supposed to be used on. Presumably men, but not if our sample is characterstic for the population.

Since our primary purpose was prediction, one could argue that we should proceeded with the model that minimizes the MSE on the test sample, and not the model that handled multicolinearity better. We would argue, however, that by handling multicolinerity we ensure more stable least-squares estimators for the model, and hence more stable predictions. In doing so, we sacrifice a gain in MSE. One could also argue that we could have handled multicolinearity in a different way, for example by conducting Principal Component Regression (PCR), which combines variables in the direction that minimizes the variance in the data.

5 Conclusion

A preliminary model should include. The included predictors and the corresponding coefficients are

Contextual dimensions are missing for us to make an adequate decision on whether this model should be implemented such as the cost and context of measurement.

6 Appendix A

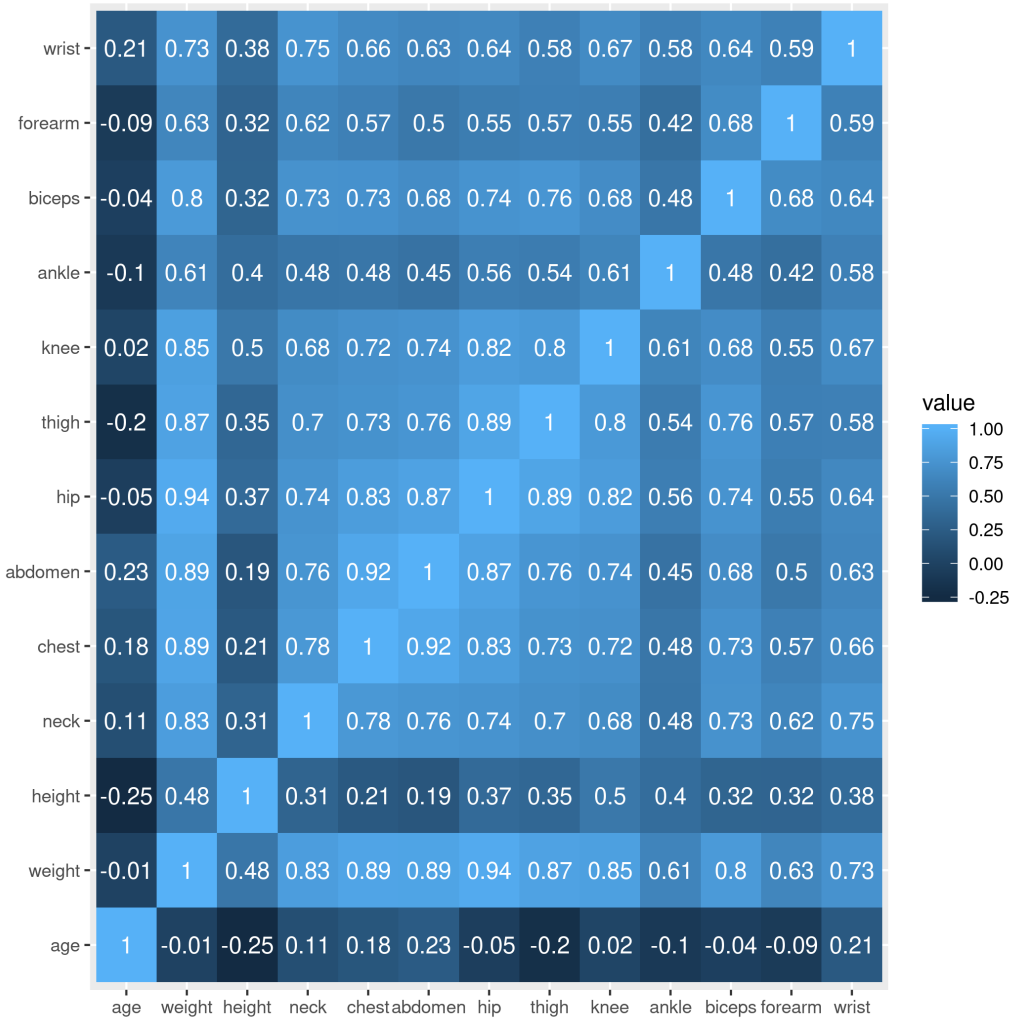

Figure 14: Correlation matrix of the full model

7 Appendix B

8 Appendix C

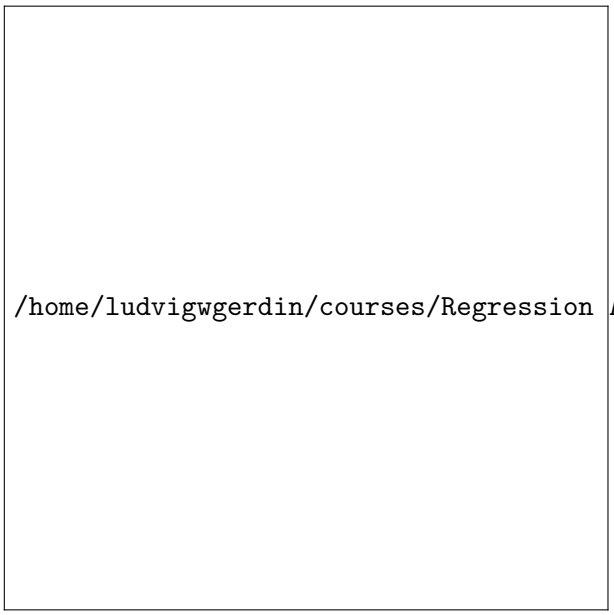

Figure 15: $DFBETA$ for regressors `biceps`, `forearm`, and `wrist`.

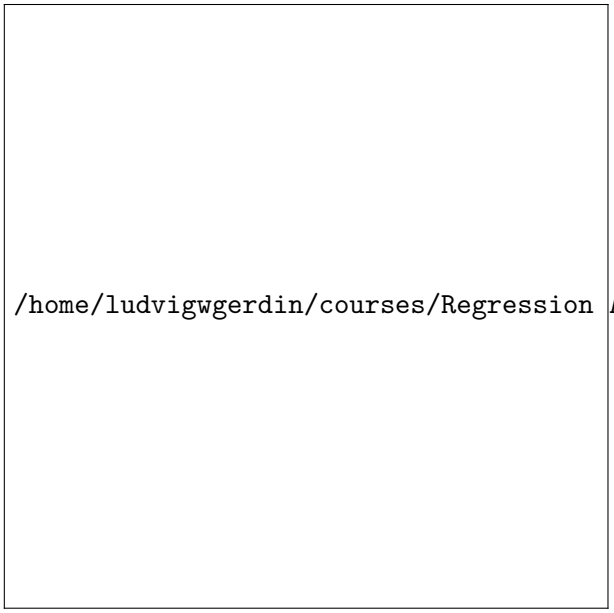

Figure 16: $DFBETA$ for regressors `thigh`, `knee`, and `ankle`.

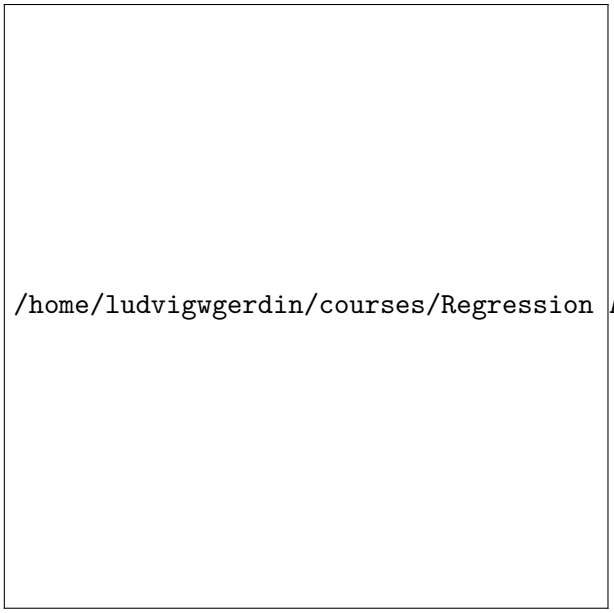

Figure 17: $DFBETA$ for regressors `age`, `weight`, `height` and `neck`.

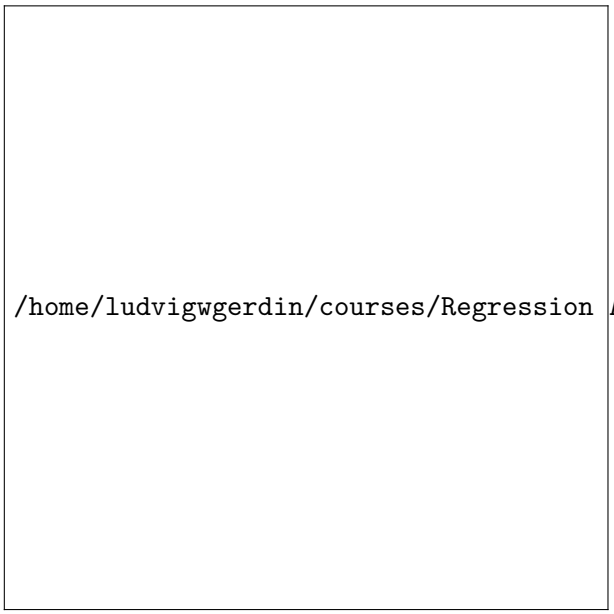

Figure 18: $DFBETA$ for regressors `chest`, `abdomen`, and `hip`.

9 References

References

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

[2] 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).

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