MATH 533: Final Project

Due on Monday, November 25, 2013 at 11:59pm

Dr. Abbas Khalili

Sahir Rai Bhatnagar and **Pablo Gonazalez Ginestet**

January 21, 2014

"He uses statistics as a drunken man uses lamp-posts — for support rather than illumination."
Andrew Lang

Abstract

Identifying overweight populations is an important first step in fighting the obesity epidemic. However, accurate measure of body fat are costly and inconvenient. Therefore we are interested in determining predictors of body fat which require only a scale and a measuring tape. We analyze a dataset which contains percentage of body fat, age, weight, height and ten body circumference measurements for 251 men. We model the data using multiple linear regression and perform various model selection techniques. Our results find age, weight, neck, abdomen, hip, thigh, forearm and wrist circumference to be significant predictors of body fat percentage. Model assumptions are assesed and seem reasonable, though residuals have a heavy tail. Using these simple measurements, one can easily and conveniently determine if they are obese, which can in turn help them take action to do more exercise and eat healthier.

I Introduction

Overweight and obesity rates have risen sharply over the past two decades in many industrialized countries, including Canada, leading to an increase in chronic disease rates. There are countless scientific studies that have shown a direct association between obesity and chronic diseases. In the Woman's Health Initiative Observational Study, it was found that body size is an important modifiable risk factor for postmenopausal breast cancer [1]. Obesity contributes to a higher incidence of cardivascular disease and type 2 diabetes, which has been projected to account for almost 75% of all deaths worldwide by 2020 [2]. An important first step in the fight against rising obesity rates is identfying who is overweight so that public health officials can aim their weight loss interventions at the right people. Body fat percentage is most accurately measured by an underwater weighing technique, however this is a costly procedure that requires a subject to go into a lab with the appropriate facilities. Body mass index (BMI), the most commonly used measure of body fat, can actually be innacurate since it does not distinguish between fat mass and lean mass [2]. Therefore we are interested in determining predictors of body fat which 1) are not costly and convenient, i.e., require only a scale and a measuring tape and 2) give a more accurate description of body fat percentage than the current measures. Using multiple linear regression, we analyze a dataset which contains percentage of body fat, age, weight, height and ten body circumference measurements for 251 men. Our hope is that we can find a good model with as few predictors as possible, but at the same time, enough to accurately predict body fat percentage. The remainder of the work is organized as follow: in Section II we describe the data, in Section III we fit a linear regression and we carry out a thorough residual diagnosis. In Section IV we carry out different variable selection techniques. In Section V, we analyze the

multicollinearity issue, and in Section VI we estimate the models and analyze the adequacy of them. Finally, in Section VII we conclude.

II Data

The dataset of body fat contains fourteen variables: Percent body fat using Method 1; Age (yrs); Weight (lbs); Height (inches) ; Neck circumference (cm) ; Chest circumference (cm) ; Abdomen circumference (cm) ; Hip circumference (cm); Thigh circumference (cm) ; Knee circumference (cm) ; Ankle circumference (cm) ; Extended biceps circumference (cm) ; Forearm circumference (cm) and Wrist circumference (cm). Our response or dependent variable is Percent body fat and the sample size is 251. Table 1. shows a descriptive statistics of all the variable in the study. We immediately notice that there may be a problem with one of the observations, as there is someone with a height of only 30 inches. Looking at the dataset, this person (observation number 42) weighs 205 pounds; so clearly there is an issue with this datapoint. The mean of the individual's age in the sample is 45 years-old and this individual has associated percent body fat of 19 and a weight of 179 lbs. From the table we can notice that the variable weight shows more spread than the rest of the variables (std=29). The matrix scatterplot (Figure 1) shows in the diagonal the variable itself and off the diagonal show the relationship between two variables. Although the information conveyed in this plot can be misleading it is an important starting point to have a rough idea of the relationship between the variables. For instance, the first row shows that percent body fat and abdomen circumference are the only pair of variables that seems to follow a linear relationship.

The summary statistics of the Body Fat Data that we will analyze in this report is presented in Table 1

Table 1: Summary Statistics of the Body Fat Data

	pbf1	**age**	**weight**	**height**	**neck**	**chest**	**abdomen**	**hip**	**thigh**	**knee**	**ankle**	**bicep**	**forearm**	**wrist**
min	2	22	125	30	31	83	70	85	49	33	19	25	21	16
max	45	81	363	78	51	136	148	148	87	49	34	45	35	21
range	43	59	238	48	20	53	78	62	38	16	15	20	14	6
median	19	43	177	70	38	100	91	99	59	38	23	32	29	18
mean	19	45	179	70	38	101	93	100	59	39	23	32	29	18
std.dev	8	13	29	4	2	8	11	7	5	2	2	3	2	1

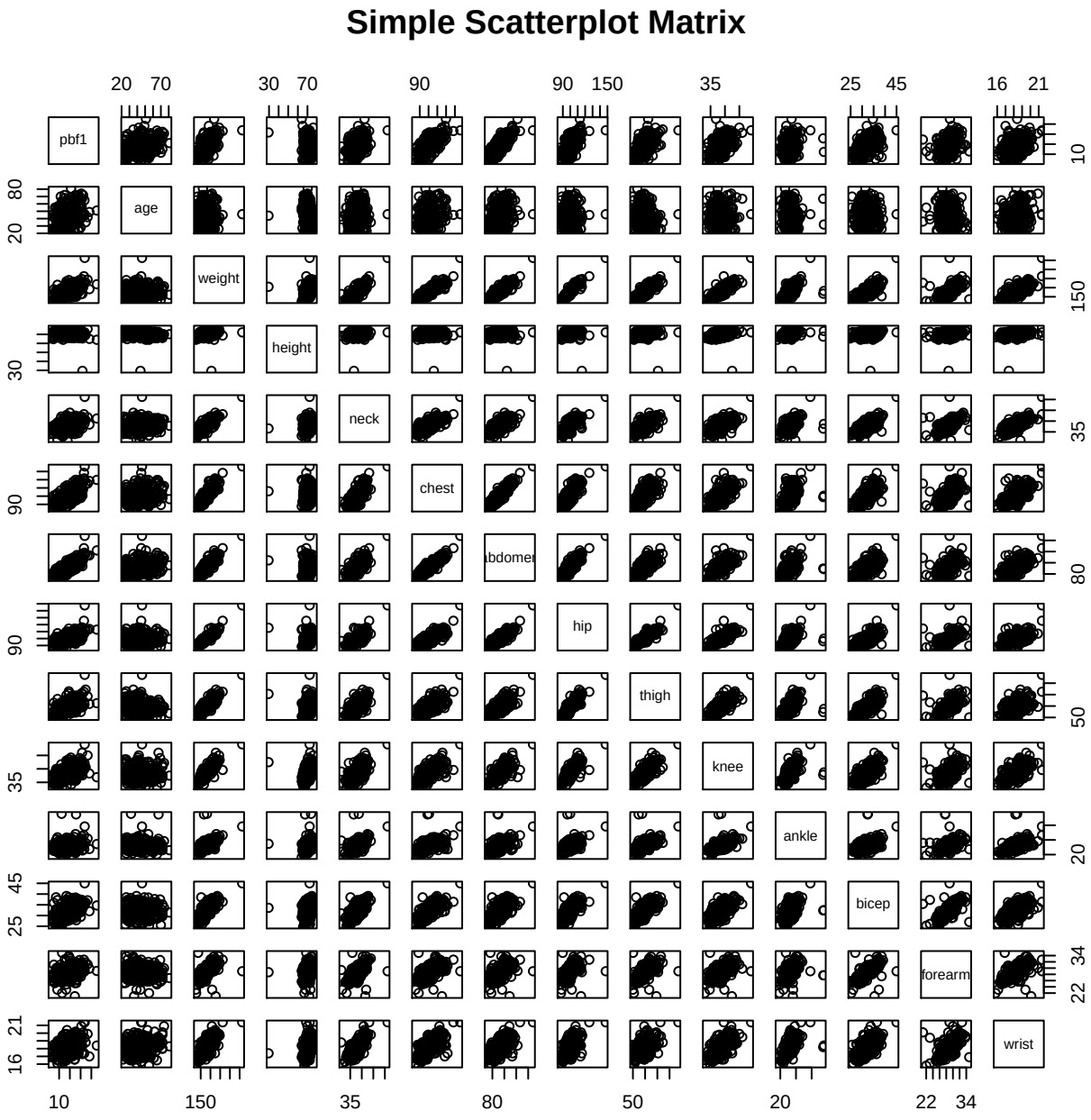

Figure 1: Scatterplot Matrix of Body Fat Data

Finally, another important tool as first starting point for analyzing the relationship among the variables is the correlation matrix. Table 2 shows the correlation matrix of the Body Fat data weight where the cells highlighted in grey represent correlations≥ 0.9. So the variable weight is highly correlated with the circumference measures. Also this high correlation it can be noticed by looking at the above third row of the matrix scatterplot.

Table 2: Correlation Matrix of the Body Fat Data

	age	weight	height	neck	chest	abdomen	hip	thigh	knee	ankle	bicep	forearm	wrist
age	1.0												
weight	-0.0	1.0											
height	-0.2	0.3	1.0										
neck	0.1	0.8	0.3	1.0									
chest	0.2	0.9	0.1	0.8	1.0								
abdomen	0.2	0.9	0.1	0.8	0.9	1.0							
hip	-0.1	0.9	0.2	0.7	0.8	0.9	1.0						
thigh	-0.2	0.9	0.1	0.7	0.7	0.8	0.9	1.0					
knee	0.0	0.9	0.3	0.7	0.7	0.7	0.8	0.8	1.0				
ankle	-0.1	0.6	0.3	0.5	0.5	0.4	0.6	0.5	0.6	1.0			
bicep	-0.0	0.8	0.2	0.7	0.7	0.7	0.7	0.8	0.7	0.5	1.0		
forearm	-0.1	0.6	0.2	0.6	0.6	0.5	0.5	0.6	0.5	0.4	0.7	1.0	
wrist	0.2	0.7	0.3	0.7	0.7	0.6	0.6	0.6	0.7	0.6	0.6	0.6	1.0

III Fitting Multiple Linear Regression

The OLS estimate of the full model is given by (1)

$$\begin{aligned} \hat{pbf1}_i = &-12.39 + 0.06\text{age}_i - 0.07\text{weight}_i - 0.07\text{height}_i - 0.43\text{neck}_i \\ &- 0.04\text{chest} + 0.89\text{abdomen}_i - 0.20\text{hip}_i + 0.21\text{thigh}_i - 0.01\text{knee}_i \\ &+ 0.15\text{ankle}_i + 0.17\text{bicep}_i - 0.02\text{forearm}_i - 0.02\text{wrist}_i, \end{aligned} \quad (1)$$

Table 3: Summary table of the Multiple Linear Regression of the Body Fat Data[a]

Predictor	**Estimate**	**Std. Error**	**t − value**	**P**(> |**t**|)
(Intercept)	-12.39	16.18	-0.77	0.44
age	0.06	0.03	1.86	0.06
weight	-0.07	0.05	-1.49	0.14
height	-0.07	0.09	-0.81	0.42
neck	-0.43	0.21	-2.01	0.05 *
chest	-0.04	0.09	-0.43	0.67
abdomen	0.89	0.08	11.10	0.00 *
hip	-0.20	0.13	-1.50	0.14
thigh	0.21	0.13	1.58	0.11
knee	-0.02	0.22	-0.07	0.95
ankle	0.15	0.20	0.74	0.46
bicep	0.17	0.16	1.05	0.29
forearm	0.42	0.18	2.25	0.03 *
wrist	-1.49	0.49	-3.01	0.00 *

* significant at $\alpha = 0.05$

[a] Residual standard error=3.981 on 237 DF, Multiple $R^2 = 0.7449$, Adjusted $R^2 = 0.7309$, F-Statistic: 53.23 on 13 and 237 DF ($p = 0$)

Since $F_0 = 53.23 > 0.4484$, we see that the null hypothesis is rejected in favour of the

alternative that at least one of the covariates contributes significantly to the model, i.e., a multiple linear regression model seems plausible for this dataset ($p = 0.000$), at $\alpha = 0.05$. Then looking at the contribution of each covariate to the model, in other words looking at t-statistics, we found that the only variables that contribute significantly ($\alpha = 0.05$) to the model are: neck, abdomen, forearm and wrist circomference. However we are a little bit surprised that some variables like age doesnt contribute significantly to the model since it has been shown in the literature that age is a good predictor of measures of adiposity. This may be an indication that the effect of age is confounded by other variables in the model. On the other hand the $R^2_{adj} = 0.73$ meaning that 73% of the variability in the response variable is explained by the linear regression model.

III.I Regression Diagnostics

We present several useful methods for diagnosing violations of the basic regression assumptions. These diagnostic methods are primarily based on study of the model residuals. Figure 2 presents several plots with measures to detect model inadequacy/adequacy. The Normal Probability plot shows that there does not seem to be a problem with the normality assumption since most of the observations are on the straight line. Thus there seems to be no problem with the Normality assumption. The plot of the residuals against the fitted values shows that there is no clear indication that the constant variance assumption is violated because the residuals can be contained in a horizontal band. So for this reason we have carried out a formal test, namely White's Test (1980). In general, White's test is based on running $\hat{\epsilon}^2$ on the cross-product of all the X's in the regression being estimated, computing $n \times R^2$, and comparing it to the critical value of $\chi(2r)$ where r is the number of regressors in this last regression excluding the constant. The output from this test is: $n \times R^2 = 109.5634$ associated with $p - value = 0.3354$. Hence, this statistic is not significant and thus then $\hat{\epsilon}^2$ is not related to the covariates and we can not reject that the variance is constant. Thus this conclusion gives statistical formal support to the plot of the residuals against the fitted values showing no clear indication of violation of the assumption of constant variance.

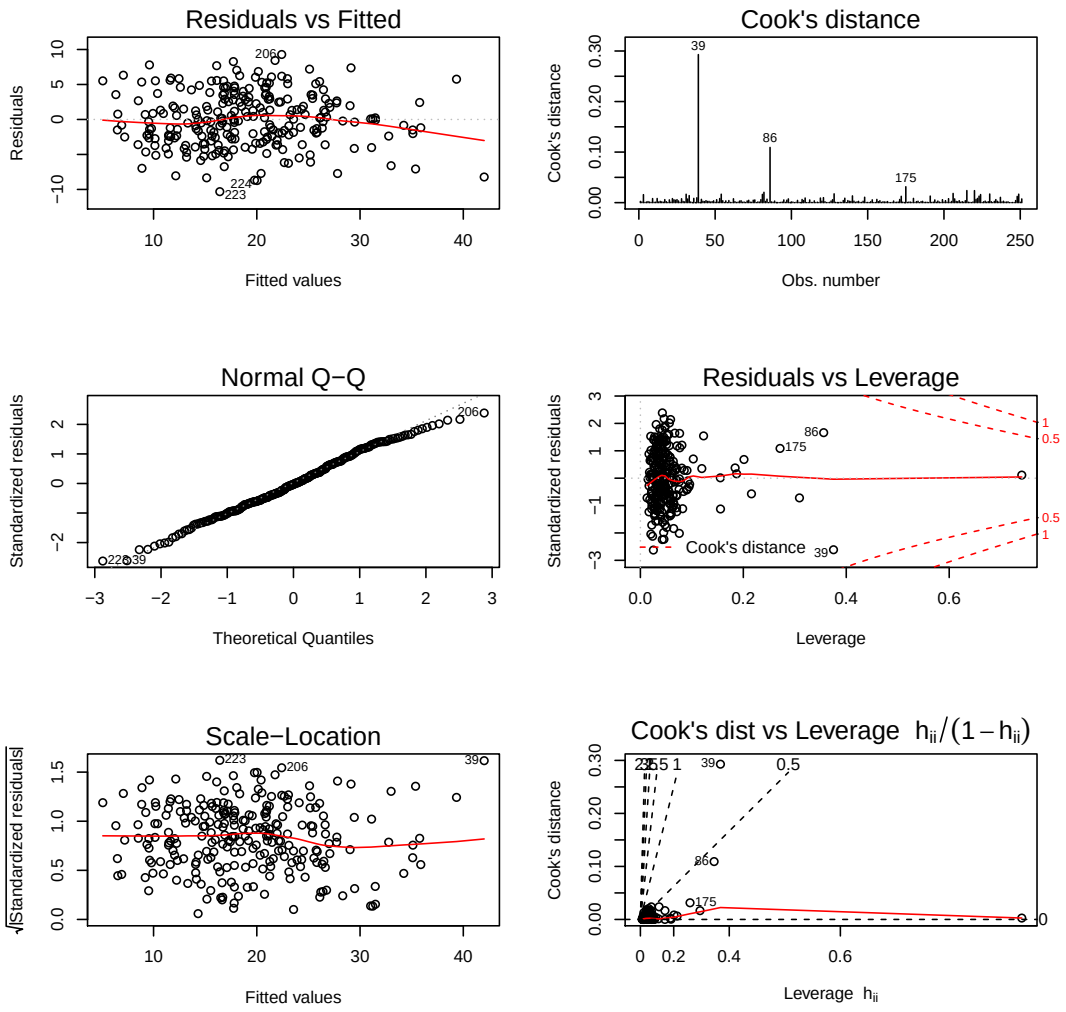

Figure 2: Plot diagnostics of the multiple linear regression model for the Body Fat data

In order to complement the analysis before we plot the residuals against the corresponding values of each regressor. The residuals from the multiple linear regression model vs. each covariate is shown in Figure 3. Neither of these plots reveals any clear indication of a problem with either misspecification of the regressors or inequality of variance. However, some observations show moderately large residual which is apparent on both plots. Moreover, those observations associated with large residuals are shown on the right side of Figure 2 standing out by its large value in the Cook's distance measure or the leverage h_{ii} measure. However, the observation 39, which has the largest Cook's D value, has $CovRatio_{39} = 1.13$ indicating that the inclusion of this observation improves relatively the precision of the estimation. Observations that have $CovRatio_i < 1$ are observations that have a low Cook's D or h_{ii} value. So for this reason we decided to keep the observation 39 as well as 86 and

175.

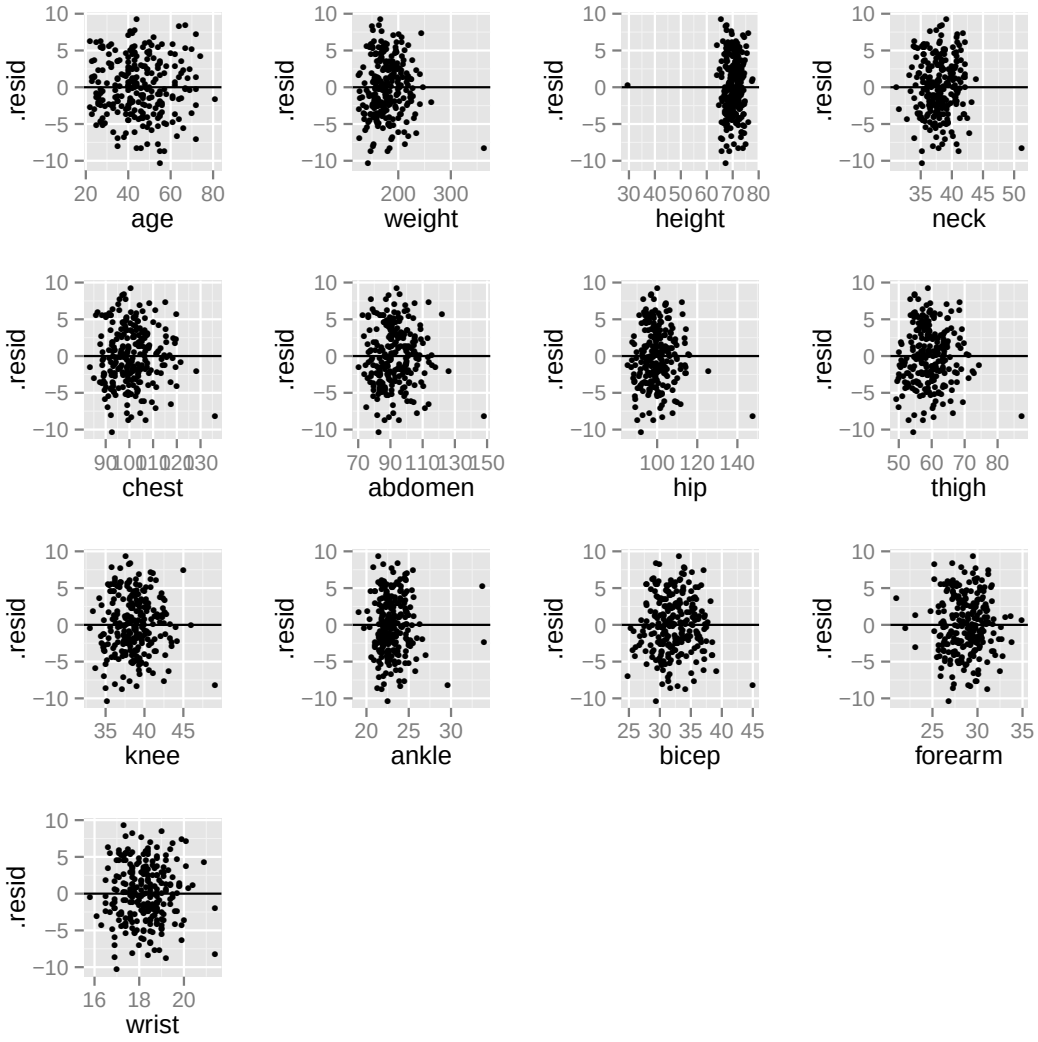

Figure 3: Plot of residuals vs. each covariate in the multiple linear regression model for the Body Fat data

The partial regression plots for the fitted model is given by Figure 4, which is a variation of the plot of residuals versus the predictor that is an enhanced way to study the marginal relationship of a regressor given the other variables that are in the model. This plot can be very useful in evaluating whether we have specified the relationship between the response and the regressor variables correctly. Figure 4 shows that the partial residuals fall along straight line so the variables have relatively significant contribution to the fitted regression model and suggesting that each covariate enters in a linear fashion in the model.However, these plots must be used with caution since as we have noticed before there seem to be strong multicollinearity in the data and thus these plots were giving us incorrect information.

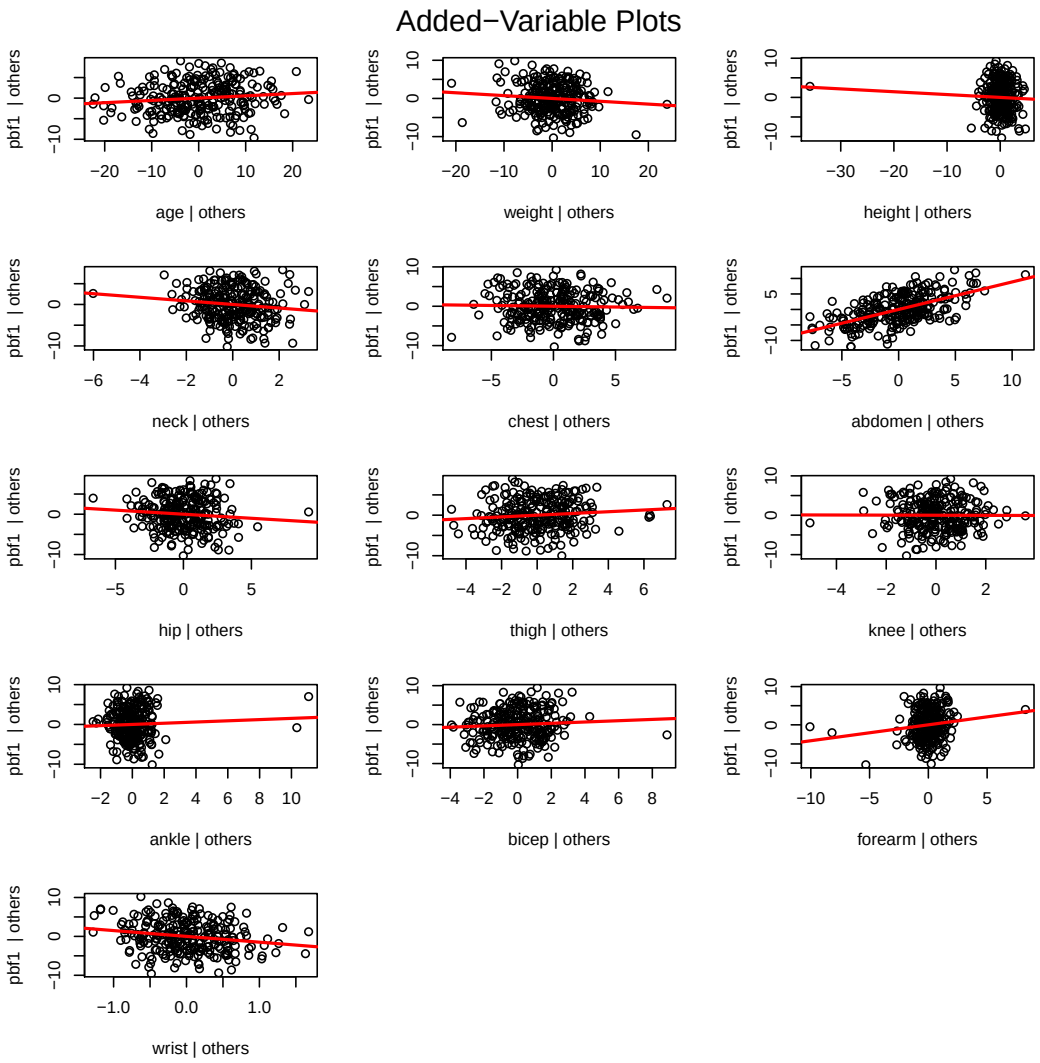

Figure 4: Plot of residuals vs. each covariate in the multiple linear regression model for the Body Fat data

IV Variable Selection

In the previous section, we saw that multicollinearity is a major issue in this dataset, i.e., many . One proposed technique to correct multicollinearity is variable selection. In this section we present several methods to deal with this issue.

IV.I Forwards, Backward and Stepwise Model Selection

These procedures of variable selection have been developed for evaluating only a small number of subset of regression models by either adding or deleting regressors one at a time and for this reason these methods are computationally cheap. The Forwards, Backward and Stepwise Model Selection is summarised in Table 4. At the first glance from Table 4 we note that the Backwards, Forwards and Stepwise algorithms give us the same model in terms of the BIC, however the Forwards procedure gives a different model than the Backwards and Stepwise based on the AIC. In all algorithms, the BIC selects the most parsimonious model. We also note that the R^2, R^2_{adj}, MSE, AIC and BIC are very similar for all the models.

Table 4: Summary table of Forwards, Backward and Stepwise Model Selection of the Body Fat Data

Predictor	**Backwards**		**Forwards**		**Stepwise**	
	AIC	BIC	AIC	BIC	AIC	BIC
(Intercept)	-19	-30.33	-28.95	-30.33	-19	-30.33
age	0.06		0.06		0.06	
weight	-0.08	-0.12	-0.12	-0.12	-0.08	-0.12
height						
neck	-0.42		-0.4		-0.42	
chest						
abdomen	0.87	0.92	0.85	0.92	0.87	0.92
hip	-0.19				-0.19	
thigh	0.28		0.17		0.28	
knee						
ankle						
bicep			0.18			
forearm	0.47	0.43	0.44	0.43	0.47	0.43
wrist	-1.43	-1.41	-1.43	-1.41	-1.43	-1.41
R^2	0.74	0.73	0.74	0.73	0.74	0.73
R^2_{adj}	0.73	0.73	0.73	0.73	0.73	0.73
MSE	15.68	16.1	15.73	16.1	15.68	16.1
AIC	1413.97	1416.7	1414.72	1416.7	1413.97	1416.7
BIC	1449.23	1437.85	1449.97	1437.85	1449.23	1437.85

IV.II Detour: LASSO

Another procedure of variable selection is LASSO. Although it is a much more sophisticated and powerful method of variable selection than the above methods, we decided to included in this subsection because this method is computationally cheap in the sense that it does

not check for every possible subset model. The LASSO provides 14 possible subset models, including the model with an intercept only as well as the full model.

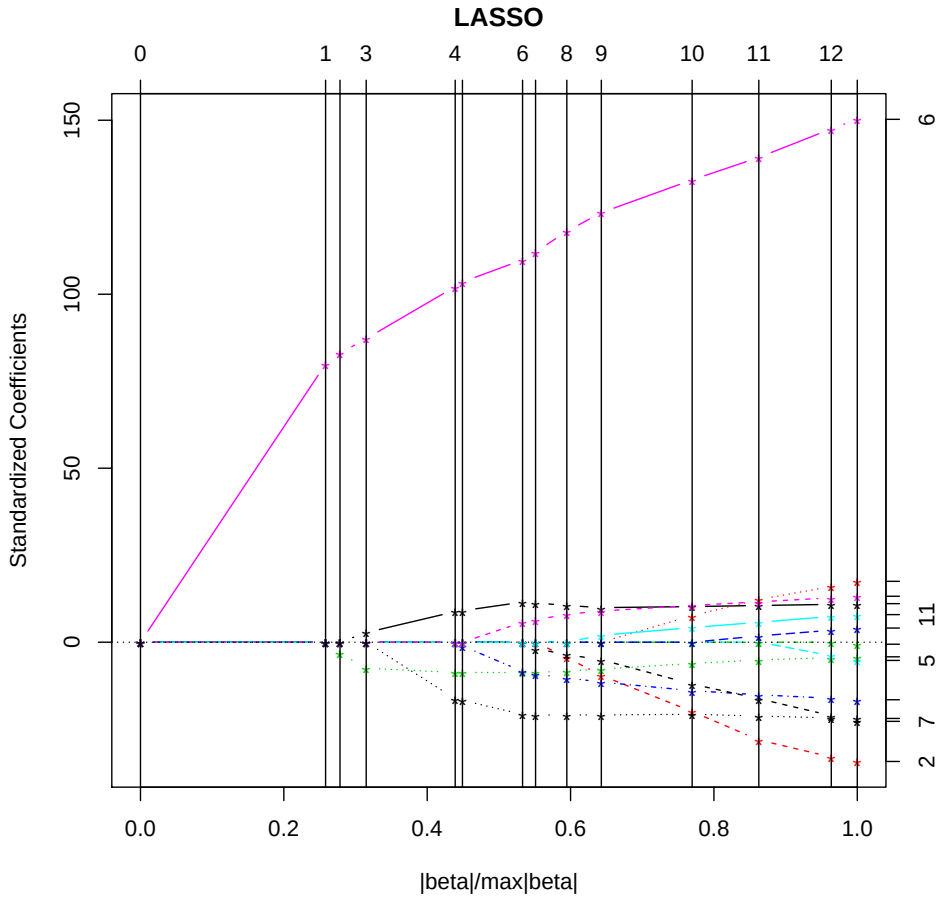

Figure 5: LASSO solution path for the Body Fat data

From Figure 5 we see that the covariates age, height and abdomen circumference are the most selected variables by the LASSO procedure. The criteria C_p and AIC suggest one possible submodel and the criteria BIC suggests another different submodel, a more parsimonious submodel (Table 6).

IV.III Best Subset Selection

In this section we consider all possible regressions and we evaluate those models with different criteria such as Rp^2, Adjusted R^2, $MSRes(p)$, and Cp. Table 5 summarized the results. The first thing we note is that all criteria coincide in the model selected. The best model according

to this criteria would include eight variables (Table 6), where the number of subsets of each size to record was set to 5.

Table 5: Summary table of Best Subset Selection of the Body Fat Data

p	$SS_{res}(p)$	$MSE(p)$	R^2_{adj}	C_p	AIC	BIC	CV	age	weight	height	neck	chest	abdomen	hip	thigh	knee	ankle	bicep	forearm	wrist
1	5065.16	20.34	0.65	72.67	323.67	330.72	521.27	0	0	0	0	0	1	0	0	0	0	0	0	0
1	7603.19	30.54	0.48	232.84	483.85	490.90	770.86	0	0	0	0	1	0	0	0	0	0	0	0	0
1	9095.58	36.53	0.38	327.03	578.03	585.08	936.05	0	0	0	0	0	0	1	0	0	0	0	0	0
1	9324.08	37.45	0.36	341.45	592.45	599.50	952.56	0	1	0	0	0	0	0	0	0	0	0	0	0
1	10248.62	41.16	0.30	399.80	650.80	657.85	1038.33	0	0	0	0	0	0	0	1	0	0	0	0	0
2	4204.58	16.95	0.71	20.36	271.36	281.93	433.53	0	1	0	0	0	1	0	0	0	0	0	0	0
2	4450.69	17.95	0.69	35.89	286.89	297.46	461.63	0	0	0	0	0	1	0	0	0	0	0	0	1
2	4541.34	18.31	0.69	41.61	292.61	303.19	469.45	0	0	0	1	0	1	0	0	0	0	0	0	0
2	4591.06	18.51	0.69	44.75	295.75	306.32	474.98	0	0	0	0	0	1	1	0	0	0	0	0	0
2	4667.62	18.82	0.68	49.58	300.58	311.15	503.83	0	0	1	0	0	1	0	0	0	0	0	0	0
3	4066.01	16.46	0.72	13.61	264.61	278.71	425.14	0	1	0	0	0	1	0	0	0	0	0	0	1
3	4131.29	16.73	0.72	17.73	268.73	282.83	429.61	0	1	0	1	0	1	0	0	0	0	0	0	0
3	4137.32	16.75	0.71	18.11	269.11	283.21	429.89	0	1	0	0	0	1	0	1	0	0	0	0	0
3	4149.88	16.80	0.71	18.90	269.90	284.00	431.29	0	1	0	0	0	1	0	0	0	0	0	1	0
3	4151.62	16.81	0.71	19.01	270.01	284.12	431.77	0	1	0	0	0	1	0	0	0	0	1	0	0
4	3959.91	16.10	0.73	8.91	259.91	277.54	416.49	0	1	0	0	0	1	0	0	0	0	0	1	1
4	3991.60	16.23	0.72	10.91	261.91	279.54	420.72	0	1	0	0	0	1	0	0	0	0	1	0	1
4	4033.95	16.40	0.72	13.59	264.59	282.21	424.95	0	1	0	0	0	1	0	1	0	0	0	0	1
4	4035.23	16.40	0.72	13.67	264.67	282.30	422.10	0	1	0	1	0	1	0	0	0	0	0	1	0
4	4044.84	16.44	0.72	14.27	265.27	282.90	427.34	0	1	1	0	0	1	0	0	0	0	0	0	1
5	3918.27	15.99	0.73	8.29	259.29	280.44	415.00	0	1	0	1	0	1	0	0	0	0	0	1	1
5	3930.02	16.04	0.73	9.03	260.03	281.18	417.54	1	1	0	0	0	1	0	0	0	0	0	1	1
5	3931.43	16.05	0.73	9.12	260.12	281.27	416.79	0	1	0	0	0	1	0	0	0	0	1	1	1
5	3938.18	16.07	0.73	9.54	260.54	281.69	417.33	0	1	0	0	0	1	0	1	0	0	0	1	1
5	3943.47	16.10	0.73	9.88	260.88	282.03	419.48	0	1	1	0	0	1	0	0	0	0	0	1	1
6	3877.36	15.89	0.73	7.70	258.70	283.38	415.33	1	1	0	0	0	1	0	1	0	0	0	1	1
6	3879.90	15.90	0.73	7.86	258.86	283.54	414.24	0	1	0	1	0	1	0	0	0	0	1	1	1
6	3880.76	15.90	0.73	7.92	258.92	283.60	415.22	1	1	0	1	0	1	0	0	0	0	0	1	1
6	3898.13	15.98	0.73	9.02	260.01	284.69	415.81	0	1	0	1	0	1	0	1	0	0	0	1	1
6	3900.91	15.99	0.73	9.19	260.19	284.87	417.74	1	1	0	0	0	1	0	0	0	0	1	1	1
7	3827.05	15.75	0.73	6.53	257.53	285.73	412.54	1	1	0	1	0	1	0	1	0	0	0	1	1
7	3840.60	15.80	0.73	7.38	258.38	286.59	414.27	1	1	0	1	0	1	0	0	0	0	1	1	1
7	3859.67	15.88	0.73	8.59	259.59	287.79	416.33	1	1	0	0	0	1	1	1	0	0	0	1	1
7	3859.97	15.88	0.73	8.61	259.61	287.81	414.29	0	1	0	1	0	1	1	1	0	0	0	1	1
7	3861.04	15.89	0.73	8.67	259.67	287.88	420.36	1	1	0	0	0	1	0	1	0	1	0	1	1
8	3794.28	15.68	0.73	6.46	257.46	289.19	411.60	1	1	0	1	0	1	1	1	0	0	0	1	1
8	3805.59	15.73	0.73	7.17	258.18	289.90	413.79	1	1	0	1	0	1	0	1	0	0	1	1	1
8	3817.40	15.77	0.73	7.92	258.92	290.65	418.31	1	1	0	1	0	1	0	1	0	1	0	1	1
8	3822.59	15.80	0.73	8.25	259.25	290.98	416.32	1	1	1	1	0	1	0	1	0	0	0	1	1
8	3823.29	15.80	0.73	8.29	259.29	291.02	417.31	1	0	1	1	0	1	1	1	0	0	0	1	1
9	3776.99	15.67	0.73	7.37	258.37	293.62	413.59	1	1	0	1	0	1	1	1	0	0	1	1	1
9	3782.77	15.70	0.73	7.74	258.74	293.99	414.22	1	1	1	1	0	1	1	1	0	0	0	1	1
9	3785.56	15.71	0.73	7.91	258.91	294.17	417.58	1	1	0	1	0	1	1	1	0	1	0	1	1
9	3793.91	15.74	0.73	8.44	259.44	294.69	415.89	1	1	0	1	1	1	1	1	0	0	0	1	1
9	3794.17	15.74	0.73	8.45	259.45	294.71	414.91	1	1	0	1	0	1	1	1	1	0	0	1	1
10	3767.01	15.70	0.73	8.74	259.74	298.52	419.00	1	1	0	1	0	1	1	1	0	1	1	1	1
10	3767.43	15.70	0.73	8.77	259.77	298.55	416.18	1	1	1	1	0	1	1	1	0	0	1	1	1
10	3774.78	15.73	0.73	9.23	260.23	299.01	420.08	1	1	1	1	0	1	1	1	0	1	0	1	1
10	3775.84	15.73	0.73	9.30	260.30	299.08	417.98	1	1	0	1	1	1	1	1	0	0	1	1	1
10	3776.99	15.74	0.73	9.37	260.37	299.15	416.77	1	1	0	1	0	1	1	1	1	0	1	1	1
11	3758.23	15.72	0.73	10.19	261.19	303.49	421.56	1	1	1	1	0	1	1	1	0	1	1	1	1
11	3764.08	15.75	0.73	10.55	261.56	303.86	422.31	1	1	1	1	1	1	1	1	0	0	1	1	1
11	3766.07	15.76	0.73	10.68	261.68	303.99	423.48	1	1	0	1	1	1	1	1	0	1	1	1	1
11	3766.82	15.76	0.73	10.73	261.73	304.03	422.59	1	1	0	1	0	1	1	1	1	1	1	1	1
11	3767.27	15.76	0.73	10.76	261.76	304.06	419.12	1	1	1	1	0	1	1	1	1	0	1	1	1
12	3755.36	15.78	0.73	12.01	263.00	308.84	427.72	1	1	1	1	1	1	1	1	0	1	1	1	1
12	3758.21	15.79	0.73	12.19	263.19	309.02	425.03	1	1	1	1	0	1	1	1	1	1	1	1	1
12	3764.03	15.81	0.73	12.55	263.55	309.38	425.53	1	1	1	1	1	1	1	1	1	0	1	1	1
12	3765.77	15.82	0.73	12.66	263.66	309.49	427.41	1	1	0	1	1	1	1	1	1	1	1	1	1
12	3772.74	15.85	0.73	13.10	264.10	309.93	430.19	1	1	1	1	1	1	1	1	1	1	0	1	1
13	3755.29	15.85	0.73	14.00	265.00	314.36	431.43	1	1	1	1	1	1	1	1	1	1	1	1	1

To sum up this section, it is interesting to highlight that from the different selection variable procedures we end up with five clearly distinguished models. The five models appear highlighted in Table 6 and they are denoted in the last row. One of them is a very simple model, it only includes four covariates (Model (1)). And the rest of them go from six to eleven covariates. Most of them differ in the covariates related to circumference measure and one of them does not include the covariate age (Model (1)).

Table 6: Summary table of the Models Selected by Forwards, Backwards, Stepwise, LASSO, and Best Subset algorithms of the Body Fat Data[a]

	Backwards		**Forwards**		**Stepwise**		**LASSO**			**Best Subset**
	AIC	BIC	AIC	BIC	AIC	BIC	AIC	BIC	C_p	C_p
age	✓		✓		✓		✓	✓	✓	✓
weight	✓	✓	✓	✓	✓	✓	✓		✓	✓
height							✓	✓	✓	
neck	✓		✓		✓		✓	✓	✓	✓
chest										
abdomen	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
hip	✓				✓		✓		✓	✓
thigh	✓		✓		✓		✓		✓	✓
knee										
ankle							✓		✓	
bicep			✓				✓		✓	
forearm	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
wrist	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
p	8	4	8	4	8	4	11	6	11	8
Model #[b]		(1)	(2)				(3)	(4)		(5)

[a] Highlighted columns represent the 5 unique models out of the 10 variable selection procedures. These are the 5 candidate models that will be used in the following sections to determine the best possible model.

[b] In the following sections we will refer to each of the 5 models as presented here

V Multicollinearity

In Table 2 we show a very simple measure of detecting multicollinearity, the examination of the correlation matrix. Inspection of the off-diagonal elements in $X'X$ gives us an idea if covariates are nearly linearly dependent. For example, the correlation between the covariate weight and chest circumference is 0.9, very close to unity. Thus weight and chest circumference are highly linearly dependent. But also weight is highly correlated with several measures of circumferences. Also different measures of circumferences are nearly linearly dependent to another measure of circumferences. For instance, the correlation between the

covariate abdomen circumference and chest circumference is 0.9. We know that strong multicollinearity results in large variance and covariance for the OLS estimators of the regression coefficients. The variance of the $\hat{\beta}_{OLS}$ is $Var(\hat{\beta}^i_{OLS}) = \sigma^2(X'X)_{ii} = \sigma^2(1 - R^2_{ii})^{-1}$. Hence, a useful technique to detect multicollinearity is to look at the amount $(1 - R^2_{ii})^{-1}$ which is called Variance Inflation Factor (VIF). Table 7 shows that the maximum VIF for Full Model is $VIF_{weight} = 33.21$ which largely exceeds the value of 10 (which is considered a threshold in practice). However, the covariate weight is not the only covariate which has associated a large value of VIF. Covariates abdomen and chest circumference also have associated VIF's values which exceed the threshold of 10. Another measure to detect multicollinearity is to look at eigenvalues of the matrix $X'X$ because if there are one or more near linear dependences in the data, then one or more of the characteristic roots will be small. Thus we compute the condition number of the matrix $X'X$ which is equal to $k = \frac{\lambda_{max}}{\lambda_{min}}$. The condition number for the Full Model is equal to 335.329 implying moderate to strong multicollinearity. Hence, both measures are indicating that there is a moderate to strong near-linear dependence in the data. So the next natural step would be to apply these techniques to the models obtained by the procedures of variable selection and see if those models overcome the multicollinearity issue and also as first way of rank them.

Table 7: VIF and Condition Number (k) of Matrix $X'X$: Full Model and the Models Selected

		Forward/Back/Step		**LASSO**		**Best Subset**
	Full Model	(1)	(2)	(3)	(4)	(5)
age	2.251		2.065	2.11	1.245	2.059
weight	33.206	6.925	13.871	25.902		18.53
height	1.679			1.579	1.223	
neck	4.275		3.943	4.169	3.556	4.036
chest	9.365					
abdomen	11.555	4.79	7.575	8.887	2.505	8.094
hip	14.546			14.011		13.239
thigh	7.694		5.321	6.824		6.181
knee	4.531					
ankle	1.889			1.832		
bicep	3.601		3.493	3.548		
forearm	2.164	1.771	2.096	2.125	1.845	1.915
wrist	3.333	2.243	3.057	3.312	2.701	3.057
k	335.329	32.569	101.147	222.154	15.553	136.89

From Table 7 we see that the variable selection has done a good job in terms of overcoming the multicollinearity issue, except for Model 3. Model 3 and Model 5 include the covariate

weight which is the covariate with the highest VIF and those models are not able to reduce the VIF significantly. However, Model 5 which also includes the second most covariate with highest VIF, which is the covariate hip, Model 5 is able to reduce the condition number significantly to levels nearby the Model 2. Whereas, Model 2 includes also the covariate weight but that model is able to reduce significantly its measure of VIF for that covariate (goes down from 33.206 to 13.871). However, the condition number of Model 2 is in the threshold 100. Thus in terms of these measures we can give a first ranking of the models selected by the procedures from the previous section. Model 1 and Model 4 would be the best models and the worst model would be Model 3. Hence, we would discard the worst model which is Model 3.

VI Estimation and Model Diagnostics for the 5 Best Models

The OLS estimates of the fives best models are presented in Table 8. In all models the F-ratio is statically significantly meaning that at least one of the covariates contributes significantly to the model, and most of the marginal effects are statistically significant. The marginal effects that are not statistically significant are those related to some measure of the circumference of the body. The marginal effect of age is statistically significant and has the expected sign in all of the models that include this variable. However,the marginal effect of weight is statistically significant but it does not have the expected sign in all the models. The R^2 as well as $AdjR^2$ are very similar in all cases (74% of the variability in the response variable is explained by the linear regression model roughly).

On the other hand, we carry out a residual analysis for the five models. Table 9 shows the results of that analysis. For instance, the normality assumption is violated in model (1) and (2), and models (3)-(5) show that the residuals are coming from a heavy tail distribution. Model (4) shows no evidence of violation of the normality assumption but it seems that the variance depends on the covariates (the p-value associated to the White Test is almost zero). Model (3) and (5) look good as the residuals seem to show no pattern. Finally, all the covariance ratios of the number of influential points in model (2),(3) and (5) fall outside of the threshold region. Thus, in terms of the residual analysis we would discard model (1), model (2) and model (4).

Table 8: Summary table of the 5 reduced fitted regression models compared with the fitted full model

	Full	(1)	(2)	(3)	(4)	(5)
age	0.056^{*}		0.058^{**}	0.055^{*}	0.077^{***}	0.058^{**}
	(0.030)		(0.029)	(0.029)	(0.023)	(0.028)
weight	-0.074	-0.125^{***}	-0.117^{***}	-0.083^{*}		-0.083^{**}
	(0.050)	(0.023)	(0.032)	(0.044)		(0.037)
height	-0.072			-0.064	-0.150^{*}	
	(0.089)			(0.086)	(0.077)	
neck	-0.432^{**}		-0.396^{*}	-0.434^{**}	-0.512^{**}	-0.424^{**}
	(0.215)		(0.206)	(0.211)	(0.199)	(0.208)
chest	-0.040					
	(0.092)					
abdomen	0.888^{***}	0.917^{***}	0.855^{***}	0.872^{***}	0.706^{***}	0.874^{***}
	(0.080)	(0.052)	(0.064)	(0.070)	(0.038)	(0.067)
hip	-0.202			-0.192		-0.185
	(0.135)			(0.132)		(0.128)
thigh	0.213		0.166	0.221^{*}		0.275^{**}
	(0.134)		(0.111)	(0.126)		(0.120)
knee	-0.015					
	(0.224)					
ankle	0.152			0.154		
	(0.205)			(0.201)		
bicep	0.166		0.182	0.161		
	(0.159)		(0.156)	(0.157)		
forearm	0.416^{**}	0.432^{**}	0.437^{**}	0.407^{**}	0.456^{***}	0.470^{***}
	(0.184)	(0.168)	(0.181)	(0.182)	(0.172)	(0.173)
wrist	-1.491^{***}	-1.405^{***}	-1.428^{***}	-1.488^{***}	-1.837^{***}	-1.425^{***}
	(0.495)	(0.409)	(0.472)	(0.491)	(0.451)	(0.471)
Constant	-12.389	-30.333^{***}	-28.945^{***}	-15.532	0.521	-19.003^{*}
	(16.179)	(6.726)	(8.371)	(14.149)	(6.320)	(10.864)
Observations	251	251	251	251	251	251
R^2	0.745	0.731	0.741	0.745	0.731	0.742
Adjusted R^2	0.731	0.727	0.733	0.733	0.724	0.734
Residual Std. Error	3.981 (df = 237)	4.012 (df = 246)	3.966 (df = 242)	3.965 (df = 239)	4.030 (df = 244)	3.960 (df = 242)
F Statistic	53.225^{***} (df = 13; 237)	167.095^{***} (df = 4; 246)	86.748^{***} (df = 8; 242)	63.367^{***} (df = 11; 239)	110.357^{***} (df = 6; 244)	87.097^{***} (df = 8; 242)

Notes: ***Significant at the 1 percent level.
**Significant at the 5 percent level.
*Significant at the 10 percent level.

Table 9: Summary of the Diagnostics for the 5 fitted regression models

	Forward/Back/Step		**LASSO**		**Best Subset**
	(1)	(2)	(3)	(4)	(5)
Normal QQ Plot	Normality Violated	Normality Violated	Heavy Tail	Heavy Tail	Heavy Tail
Influential[a] Points	{39, 223, 224}	{39, 223, 224}	{39, 42, 205, 223}	{39, 42, 205, 223}	{39, 205, 223}
Cov Ratio[b]	14/15 (93%) [0.94,1.06]	14/14 (100%) [0.89,1.11]	15/15 (100%) [0.86,1.14]	13/14 (93%) [0.92,1.08]	18/18 (100%) [0.89,1.11]
Partial Reg. Plots[c]	forearm	thigh, bicep, forearm	ankle,bicep forearm, wrist	height, neck forearm	forearm
Residuals vs. Fitted Values	Double-Bow pattern	No pattern seen	No pattern seen	Double-Bow pattern	No pattern seen
White-Test	Non-constant ($p = 0.05$)	Constant ($p = 0.12$)	Constant ($p = 0.15$)	Non-Constant ($p = 0.00$)	Constant ($p = 0.28$)

[a] Determined by inspecting Cook's distance plots, residuals vs. hat values plots, residuals vs. fitted values plots

[b] The number of influential points whose covariance ratios fall outside of the threshold region $(1 \pm 3p/n)$. The number percentage in brackets=$\frac{\text{\# points outside interval}}{\text{total \# influential points}}$. The threshold region is given in square brackets.

[c] The covariates identified as not having a significant contribution to the regression model.

The diagnostic plots for our final chosen model is shown in Figure 6.

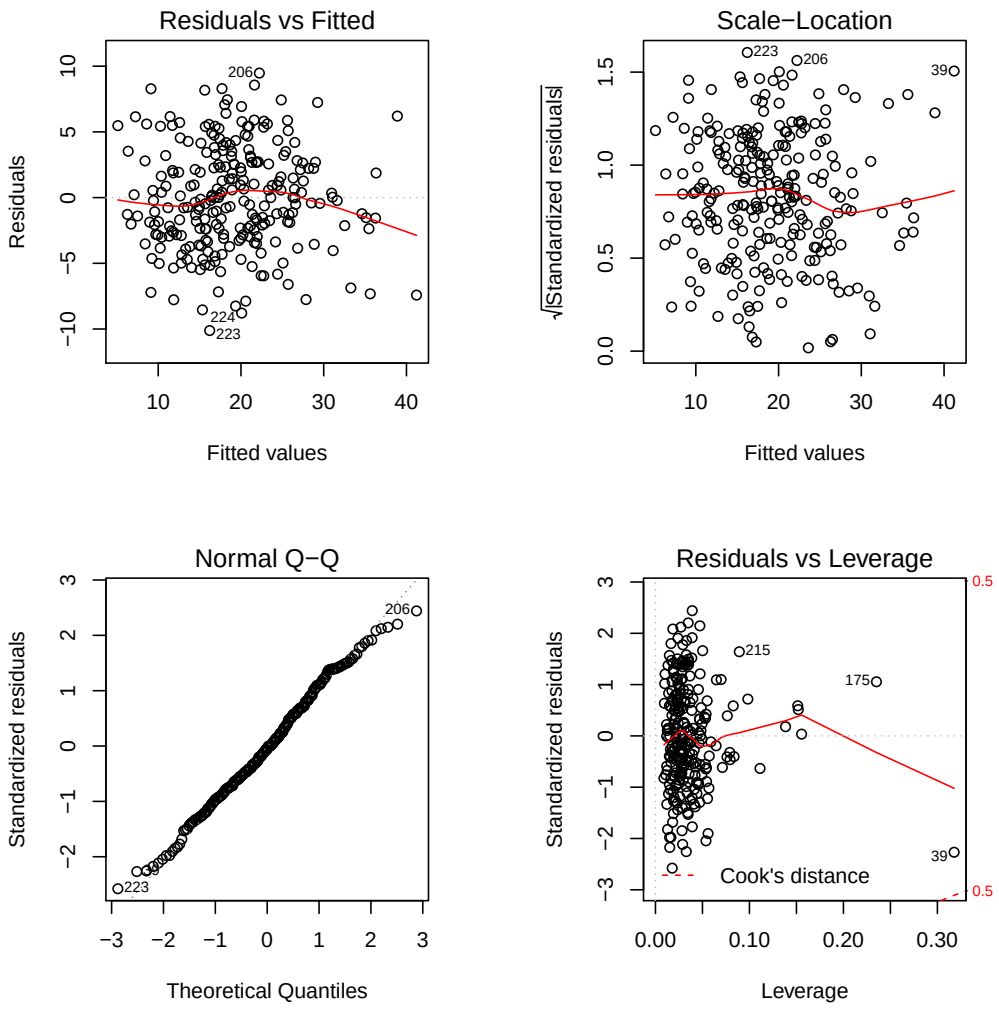

Figure 6: Diagnostics plot of Model (5) : our chosen model for the Body Fat data

VII Conclusion

We have started analyzing the body fat data in very broad terms, considering all possible covariates entering in the model to explain the variability of the response variable percentage body fat. We have identified the main issues such as the multicollinearity problem as well as we shed some light on some influence and outliers observations. Then we take advantage of the different model selection techniques and we end up with five clearly distinguished models. And after applying thorough residual diagnosis to each of them we conclude that the best model is model (5). The model (5) includes eight covariates where only one marginal effect is not statistically significant, and this model is in the middle between the shortest and the

longest model among the five models considered. Finally, one thing that have surprised us is the fact that the covariate weight enters with the opposed expected sign in the five and full model. Another issue is the observation number 42 which needs to be checked with the data collectors, to see if this is a mistake. This fact should trigger further research to to shed light on the relationship between percentage body fat and weight.

References

[1] L. Morimoto, E. White, Z. Chen, R. Chlebowski, J. Hays, L. Kuller, A. Lopez, J. Manson, K. Margolis, P. Muti, M. Stefanick, and A. McTiernan, "Obesity, body size, and risk of postmenopausal breast cancer: the women's health initiative (united states)," *Cancer Causes and Control*, vol. 13, no. 8, pp. 741–751, 2002. 2

[2] M. Snijder, R. van Dam, M. Visser, and J. Seidell, "What aspects of body fat are particularly hazardous and how do we measure them?," *International Journal of Epidemiology*, vol. 35, no. 1, pp. 83–92, 2006. 2

observation	age	weight(lbs)	height(in)
42	44	205	29.5