SF2930 Regression Analysis

Project 1

Emil Book

ebook@kth.se

940122-0034

Deepika Anantha Padmanaban

deap@kth.se

950303-4408

11.03.2019

1.1 INTRODUCTION

The aim of this project is to develop a modern regression model for one of the scenarios presented. We have chosen the **Scenario 1 - Large -Sample Regression (p < n)** to fit a regression model to find the Body Fat Mass (BFM) for the **bodyfatwomen** dataset.

The given dataset bodyfatwomen contains measurements of the BFM (DEXfat) of 71 women assessed by DXA. The possible explanatory variables included in the dataset are age, waist circumference, hip circumference, elbow breadth, knee breadth, anthro3a, anthro3b, anthro3c and anthro4. In this project, we have tried to find the best explanatory variables and develop a simple yet well-fit model to explain the dataset. This project includes residual analysis, transformation of variables, variable selection, multicollinearity diagnostics, leverage and influential points treatment, bootstrapping and various model evaluations.

1.2 MODEL BUILDING:

FULL MODEL FIT AND RESIDUAL ANALYSIS:

The primary aim was to build a model that explains the given dataset with better generalization capabilities. We started with an initial model which contained all the variables available in the dataset and fit a model. The statistics of the full model is;

R-Squared	Adjusted R-Squared	AIC	BIC
0.9231	0.9117	381.4125	406.302

Figure 1: Statistics of the full model with 9 regressors

Since, the R-Squared shows the deviation of the model fit from the actual data, this approximately shows that 92.3% of the variance is explained by the model. But, the AIC and BIC values also seem to be too high. But, R-squared also shows the realization of the error(ϵ) in the model with the following assumptions;

- The relationship between the response y and the regressors is approximately linear.
- The error term ϵ has zero mean and a constant variance.
- The errors are uncorrelated and are normally distributed.

This model was used for further analysis of residual plots, where we could analyze the adequacy of the model to fulfill the considered assumptions about the model. The plots in figure[2] were analyzed. Though, the residuals in figure[2(a) and 2(c)] show residuals equally distributed on either side of the zero-axis, it is seen that there is variance is not linear which implies that the relationship between the regressors and response variable may not be linear. Similarly, with figure[2(b)], the Normal Q-Q plot seems to be positively skewed, showing some deviations in the normality assumption rooting for the presence of possible outliers. The figure[2(d)] also shows the presence of possible outliers found using the cook's distance.

TRANSFORMATION TO CORRECT MODEL INADEQUACIES:

With the idea that there was a non-linear variance in the residual plots, we had come to a conclusion that there is a need for transformation. To see what kind of transformation we should use, we chose the analytical method box cox transformation to produce a linear relationship between the regressors and the response variable. The box-cox transformation with a confidence interval yielded a lambda value, λ =0.1515 as shown in figure[3].

The general rule is to round off to nearest whole number, which in our case would be 0. The zero value of λ implies a log transformation of the response variable. The statistics of the model fit using log-transformation of the response variable is as shown in figure[4]. The improvement in the R-squared values and reduction

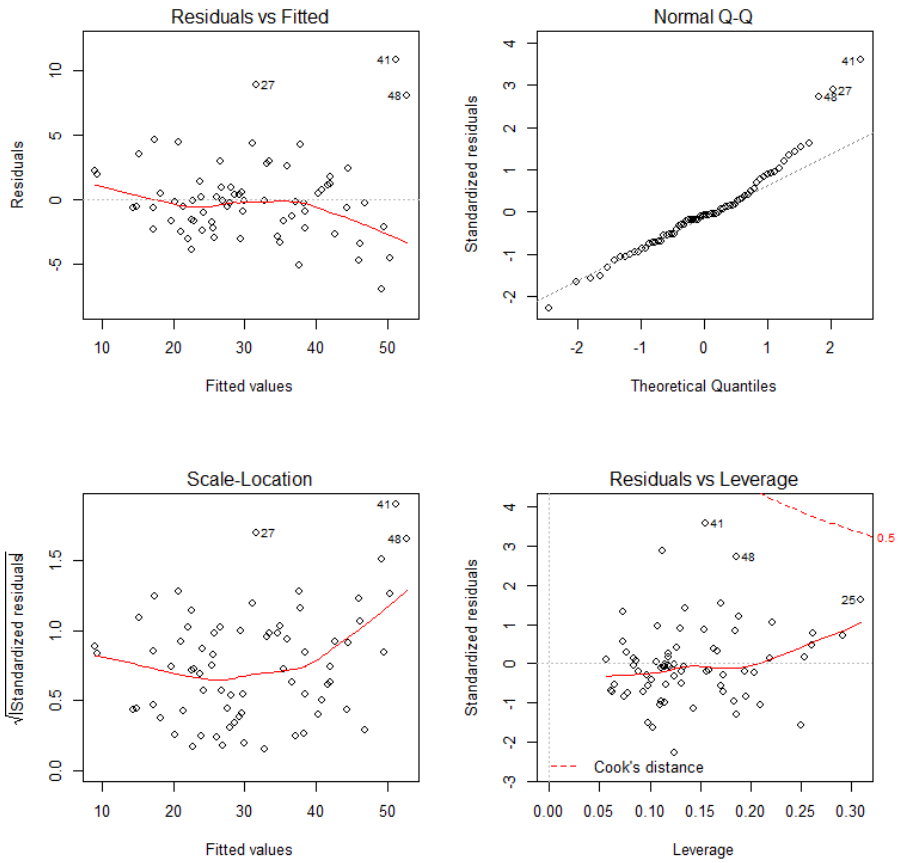

Figure 2: from top left: a)Residuals of the datapoints Vs Fitted values of the model b)Normal Q-Q plot for standardized residuals c)Standardized residuals Vs Fitted values d)Residuals Vs Leverage for all data points

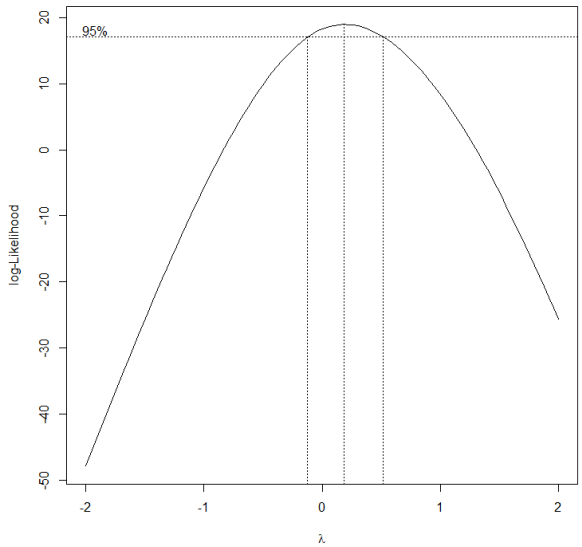

Figure 3: Log-Likelihood of the power parameter, λ with 95% confidence intervals

R-Squared	Adjusted R-Squared	AIC	BIC
0.9408	0.932	-115.714	-90.825

Figure 4: Statistics of the log transformed model

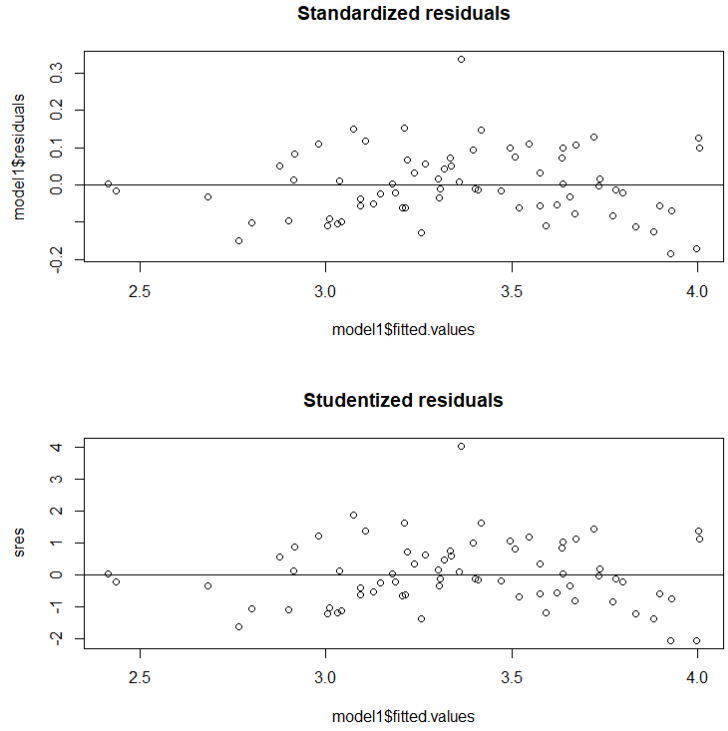

Figure 5: Standardized and Studentized residuals of the model after the log transformation

in AIC and BIC values show a successful transformation. We also tried plotting the residuals to check the variance and it seemed random as expected,figure[5].

MULTICOLLINEARITY DIAGNOSTICS AND TREATMENT:

The model that was previously fitted seemed to explain the variance in data but was complex since it included all the explanatory variables. A problem that occurs when fitting a full model is that there may be near linear dependencies between the regressors. Body Fat mass data for women has a lot of explanatory variables, it is logical that some might say the same thing. We tried plotting the correlation matrix to give us an indication if there was some redundancy, figure[6]

From the plots its visible that there is a high collinearity between the variables **anthro3a, anthro3b, anthro3c and anthro4**. For example we could see a correlation of 0.9509316 between anthro3a and anthro3b, 0.9838199 between anthro4 and anthro3b. Thus, we performed Variance Inflation Factor(VIF) test which gives us combined effects of dependences among different regressors and the results are presented in figure[7].

age	waistcirc	hipcirc	elbowbreadth	kneebreadth	anthro3a	anthro3b	anthro3c	anthro4
1.207	5.771	5.121	1.417	2.853	39.471	49.101	9.038	111.261

Figure 7: Variation Inflation Factor for all the regressors in the full model

VIF values over 10 should be removed one at the time and over 5 need further analysis. As is seen in the table above, the VIF value for anthro3a, anthro3b, anthro3c and anthro4 are high, confirming our predictions from the correlation matrix. Thus, we tried removing one at a time and re-estimated VIF. On removal of anthro3a with highest VIF, we still could find a high VIF(>10) for a few regressors, figure[8].

Figure 6: Correlation among all the variables in the dataset

age	waistcirc	hipcirc	elbowbreadth	kneebreadth	anthro3b	anthro3c	anthro4
1.207	5.759	5.065	1.416	2.746	43.232	8.938	37.655

Figure 8: Variation Inflation Factor for all the regressors in the model after removing anthro3a

Next, we tried anthro3b and performed VIF test, which yielded the expected result, where the VIF values where lesser than 10 for the remaining regressors, figure[9]. Thus, we concluded with the remval of two variables with high correlation to the other regressors, namely, **anthro3a and anthro3b**.

age	waistcirc	hipcirc	elbowbreadth	kneebreadth	anthro3c	anthro4
1.207	5.493	5.047	1.404	2.743	6.899	7.218

Figure 9: Variation Inflation Factor for all the regressors in the model after removing anthro3a and anthro3b

We then performed ridge- regression to further evaluate the problem with multicollinearity and checked if our model has poor estimates of the coefficients. The corresponding lambda for ridge regression was $\lambda = 0.1$ and we can see that the coefficients become stable, so no further action regarding multicollinearity had to be taken, figure[10].

Regressors	Coefficients without ridge	Coefficients with Ridge
Intercept	-0.0909	-0.0887
age	0.0015	0.0015
waistcirc	0.0037	0.0037
hipcirc	0.0105	0.0105
elbowbreadth	0.0084	0.0084
kneebreadth	0.0377	0.0379
anthro3c	0.1771	0.1768
anthro4	0.1548	0.1547

Figure 10: Coefficients of all the regressor before and after ridge regression with $\lambda = 0.1$

1.3 MODEL SELECTION:

VARIABLE SELECTION AND CROSS-VALIDATION:

Overfitting is a serious problem that occur in multiple linear regression, it often occurs when we have to many explanatory variables for a limited set of data points. This may result in an overly complex model, and can cause the coefficients in our OLS to be misleading. That would make our model a bad predictor for other samples of the same population, thus leading to poor generalization. Also, not variables included in the full model is necessarily explanatory of the variance in data and the Occam Razor rule also states that a simple model is always preferred over an overly complex model. Thus, we try to select variables that can explain the variation in data to the best possible extent, thus reducing the complexity.

Backward Variable Selection and Evaluation:

One way to simplify our model would be by iterative backward elimination. The model evaluation criteria are R-squared and AIC. The problem with R-squared is that the more variables used, generally better is the R-squared for the model, therefore we need to look at criteria that penalize the model with the amount of explanatory variables. An alternative for that would be Adjusted R-squared. Similarly, AIC also favours model with high complexity and we also consider the BIC for the models as selection criteria. R-squared should always increase whilst AIC and BIC should decrease with the better model. Akaike information criterion and Bayesian information criterion are both a penalized log-likelihood measurement. They are calculated as;

$$AIC = -2ln(L) + 2p$$

$$BIC = -2ln(L) + pln(n)$$

We start with our full model without the variables that were removed during multicollinearity treatment,anthro3a and anthro3b. We started removing variable with the lowest significance value(p-value) and ensured that selection criteria (R-squared, Adjusted R-squared, AIC, BIC, MSE and Mallows Cp) were fulfilled.

We also performed a cross-validation test on all of the above criteria. The variables were removed iteratively until the evaluation criteria were obeyed and the cross validation error remained the lowest. For the **cross-validation**, we used a **4-fold CV**, as the dataset had a total of 71 samples and there was a need to ensure that the size of the training data was large enough to cover all the functionalities of the dataset.

Mallow's Cp:

The best subset was selected using Mallow's Cp on the Cross-Validation dataset using Backward-selection. It was found that, for three/four trials, the Mallow's Cp was lowest for the model with 4 variables and the tabulations and plot in figures[11 and 12] show the same. From the plot and the graph, it is evident that the Mallow's Cp prefers the model with 4 variables, namely, **waistcirc, hipcirc, anthro3c and anthro4**.

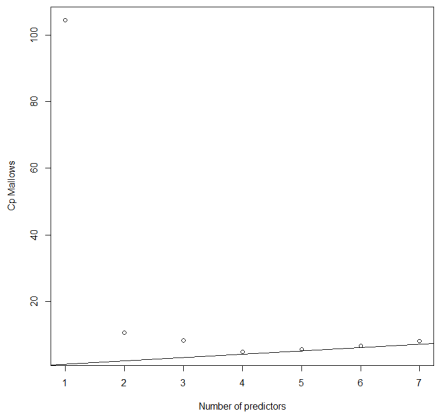

Figure 11: Mallow's Cp Vs number of predictors

Model used	Mallows Cp
anthro4	100.59
hipcirc + anthro4	17.88
waistcirc + hipcirc + anthro4	7.93
waistcirc + hipcirc + anthro3c + anthro4	4.84
waistcirc + hipcirc + kneebreadth + anthro3c + anthro4	5.19
age + waistcirc + hipcirc + kneebreadth + anthro3c + anthro4	6.06
age + waistcirc + hipcirc + elbowbreadth + kneebreadth + anthro3c + anthro4	8.00

Model used	Mallows Cp
anthro3c	107.17
hipcirc + anthro3c	20.35
age + hipcirc + anthro3c	10.89
age + hipcirc + kneebreadth + anthro3c	7.32
age + hipcirc + kneebreadth + anthro3c + anthro4	5.76
age + waistcirc + hipcirc + kneebreadth + anthro3c + anthro4	6.43
age + waistcirc + hipcirc + elbowbreadth + kneebreadth + anthro3c + anthro4	8.00

Model used	Mallows Cp
hipcirc	130.22
hipcirc + anthro3c	17.75
hipcirc + anthro3c + anthro4	9.86
waistcirc + hipcirc + anthro3c + anthro4	4.99
age + waistcirc + hipcirc + anthro3c + anthro4	5.24
age + waistcirc + hipcirc + kneebreadth + anthro3c + anthro4	6.12
age + waistcirc + hipcirc + elbowbreadth + kneebreadth + anthro3c + anthro4	8.00

Model used	Mallows Cp
anthro4	104.51
hipcirc + anthro4	10.46
kneebreadth + hipcirc + anthro4	8.22
kneebreadth + hipcirc + anthro3c + anthro4	4.71
kneebreadth + hipcirc + elbowbreadth + anthro3c + anthro4	5.48
age + kneebreadth + hipcirc + elbowbreadth + anthro3c + anthro4	6.58
age + waistcirc + hipcirc + elbowbreadth + kneebreadth + anthro3c + anthro4	8.00

Figure 12: Mallow's Cp for the different folds of 4-fold CV

AIC, BIC, MSE, R-squared and Adjusted R-squared:

We also conducted the AIC,BIC, R-squared and adjusted R-squared tests on the cross-validation sets using the different models and the results are tabulated as shown below, figure[13 and 14]. As is expected, the AIC and R-squared favour the model with the highest number of explanatoy variables. But, BIC seems to favour the model with 4 variables (same as Mallow's Cp) and seems to reduce with further increase in the number of variables. Also, Adjusted R-squared also increases, but after the 4 variable model, the increase in the Adjusted R-square is not very significant. The MSE also seems to be low starting from the 4-variable model with **waistcirc, hipcirc, anthro3c and anthro4**, with lowest value for complex models.

Model used	AIC	BIC	R2	Adjsted R2	MSE
anthro4	-273.91	-69.59	0.7874	0.7829	0.0283
hipcirc + anthro4	-324.31	-101.17	0.8954	0.8911	0.0087
kneebreadth + hipcirc + anthro4	-336.39	-105.77	0.9118	0.9061	0.0088
kneebreadth + hipcirc + anthro3c + anthro4	-342.74	-106.34	0.9193	0.9122	0.0099
kneebreadth + hipcirc + elbowbreadth + anthro3c + anthro4	-346.48	-105.06	0.9235	0.9148	0.0086
age + kneebreadth + hipcirc + elbowbreadth + anthro3c + anthro4	-349.85	-103.52	0.9271	0.9168	0.0072
age + waistcirc + hipcirc + elbowbreadth + kneebreadth + anthro3c + anthro4	-349.88	-99.63	0.9271	0.9149	0.0073

Model used	AIC	BIC	R2	Adjsted R2	MSE
anthro4	-257.25	-75.75	0.7998	0.7958	0.0247
hipcirc + anthro4	-331.63	-126.27	0.9298	0.9269	0.0099
kneebreadth + hipcirc + anthro4	-342.32	-130.14	0.9396	0.9358	0.0094
kneebreadth + hipcirc + anthro3c + anthro4	-349.01	-131.09	0.9451	0.9404	0.0066
kneebreadth + hipcirc + elbowbreadth + anthro3c + anthro4	-353.99	-130.79	0.9488	0.9432	0.0056
age + kneebreadth + hipcirc + elbowbreadth + anthro3c + anthro4	-356.14	-128.41	0.9503	0.9437	0.0061
age + waistcirc + hipcirc + elbowbreadth + kneebreadth + anthro3c + anthro4	-356.33	-124.59	0.9504	0.9425	0.006

Figure 13: Model Evaluation for different subsets using 4-fold cross validation

Model used	AIC	BIC	R2	Adjsted R2	MSE
anthro4	-251.65	-81.01	0.8018	0.7981	0.0459
hipcirc + anthro4	-318.19	-128.55	0.9224	0.9194	0.0216
kneebreadth + hipcirc + anthro4	-328.77	-132.74	0.9331	0.9292	0.0177
kneebreadth + hipcirc + anthro3c + anthro4	-336.02	-134.34	0.9396	0.9348	0.0167
kneebreadth + hipcirc + elbowbreadth + anthro3c + anthro4	-338.37	-132.16	0.9416	0.9356	0.0202
age + kneebreadth + hipcirc + elbowbreadth + anthro3c + anthro4	-339.69	-129.17	0.9426	0.9355	0.0198
age + waistcirc + hipcirc + elbowbreadth + kneebreadth + anthro3c + anthro4	-339.75	-125.22	0.9427	0.9342	0.0198

Model used	AIC	BIC	R2	Adjsted R2	MSE
anthro4	-251.31	-79.79	0.7917	0.7878	0.0249
hipcirc + anthro4	-315.56	-126.46	0.9157	0.9126	0.0104
kneebreadth + hipcirc + anthro4	-325.98	-130.65	0.9272	0.9231	0.0071
kneebreadth + hipcirc + anthro3c + anthro4	-335.21	-133.91	0.9361	0.9311	0.0061
kneebreadth + hipcirc + elbowbreadth + anthro3c + anthro4	-336.93	-131.24	0.9376	0.9314	0.0052
age + kneebreadth + hipcirc + elbowbreadth + anthro3c + anthro4	-337.21	-127.43	0.9379	0.9303	0.0046
age + waistcirc + hipcirc + elbowbreadth + kneebreadth + anthro3c + anthro4	-337.35	-123.52	0.9381	0.9289	0.0046

Figure 14: Model Evaluation for different subsets using 4-fold cross validation

OUTLIERS, LEVERAGE AND INFLUENTIAL POINT:

The Normal Q-Q plot shows certain deviation, with a positive skew. Thus, there is a high chance of having outliers in the model. Thus, we analyze the presence of outliers, leverage points and find ways of handling them in case they are influential.

Leverage Diagnostics:

Leverage Diagnostic is to identify high hat-matrix diagonals h_{ii}. All the points that exceed the cut-off (2p/n), where p is the number of features and n is the number of datapoints. In our case, p=4 and n = 71, thus, the cut-off is 0.1126761. The points violating this cut-off are listed in the table below, figure[15].

Datapoints	Leverage	Cook's D	Covar-Ratio
18	0.1647	0.1052	1.0512
20	0.1221	0.0035	1.1181
36	0.1249	0.0004	1.2318
43	0.1313	0.0004	1.2412
45	0.1199	0.1486	0.7968
54	0.1443	0.0121	1.2273
66	0.2361	0.000009	1.4129
67	0.2361	0.00009	1.4129

Figure 15: Influential Point analysis

Cooks Distance analysis:

Cook's distance measures the distance between the least-square estimates of the regression coefficients $\hat{\beta}$ and the regression coefficients $\hat{\beta}_{(i)}$ estimated when the i^{th} datapoint is deleted from the data sample. Points with large cook's distance can have significant influence on the least-squares estimates of the coefficients. The points above the threshold can be seen from figure[16], which are 18,23,27 and 45, can be potential influential points.

Covariance Ratio Test:

The covariance test included finding the covariance values and finding if the value fell far-off from the cut-off, which was >1+3p/n and <1-3p/n. It is seen that only point 27 had a huge deviation from the cut-off value, figure[17]. So, that requires further analysis.

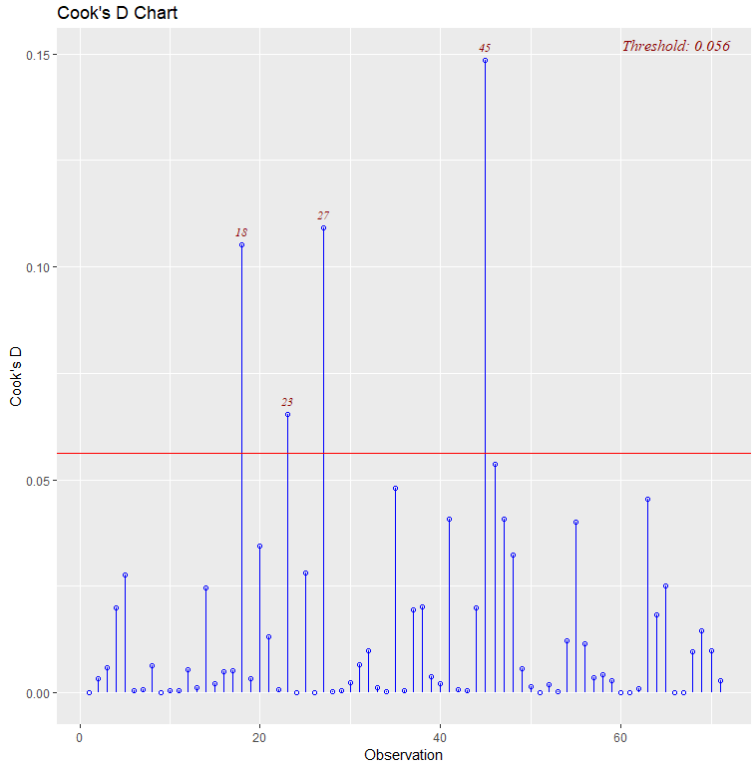

Figure 16: Cook's Distance plot for the dataset

Data point	9	22	36	43	54	66	67	27	45	63
Covar ratio	1.186	1.169	1.231	1.241	1.227	1.413	1.413	0.496	0.797	0.824

Figure 17: Covariance Ratio for all the influential points

Handling Influential Points:

The potential influential points in the datasample are 18,23,27 and 45. We tried checking if there is any improvement in the model evaluation criteria with the removal of these points and there weren't much improvements, so we concluded with not removing any of these points, figure[18].

Removed point	R-squared	Adjusted R2
27	0.9425	0.9389
18	0.934	0.9299
23	0.9379	0.934
45	0.9419	0.9383
18,23,27,45	0.9511	0.9479
No point removed	0.9329	0.9288

Figure 18: Model Evaluation after removal of certain points

BOOTSTRAP BASED CONFIDENCE INTERVALS:

Bootstrapping is a computer - intensive procedure that was developed to allow us to determine reliable estimates of the standard errors of regression estimates of coefficients in terms of their confidence intervals. The bootstrap method requires us to select a random sample of size n=71 with replacement from this original sample, called the bootstrap sample. Since it is selected with replacement, the bootstrap sample will contain observations from the original sample, with some of them duplicated and some of them omitted. We obtained 500 such bootstrapped samples and found the 2.5% to 97.5% confidence interval for the estimates of the regression coefficient. The values have been listed in figure[19] and the histograms are plotted in figure[20]

Variable	2.50%	97.50%
Intercept	-0.1759	0.3937
waistcirc	0.0011	0.0087
anthro3c	0.0535	0.2743
anthro4	0.0765	0.2721

Figure 19: Coefficient estimates for 95% confidence interval

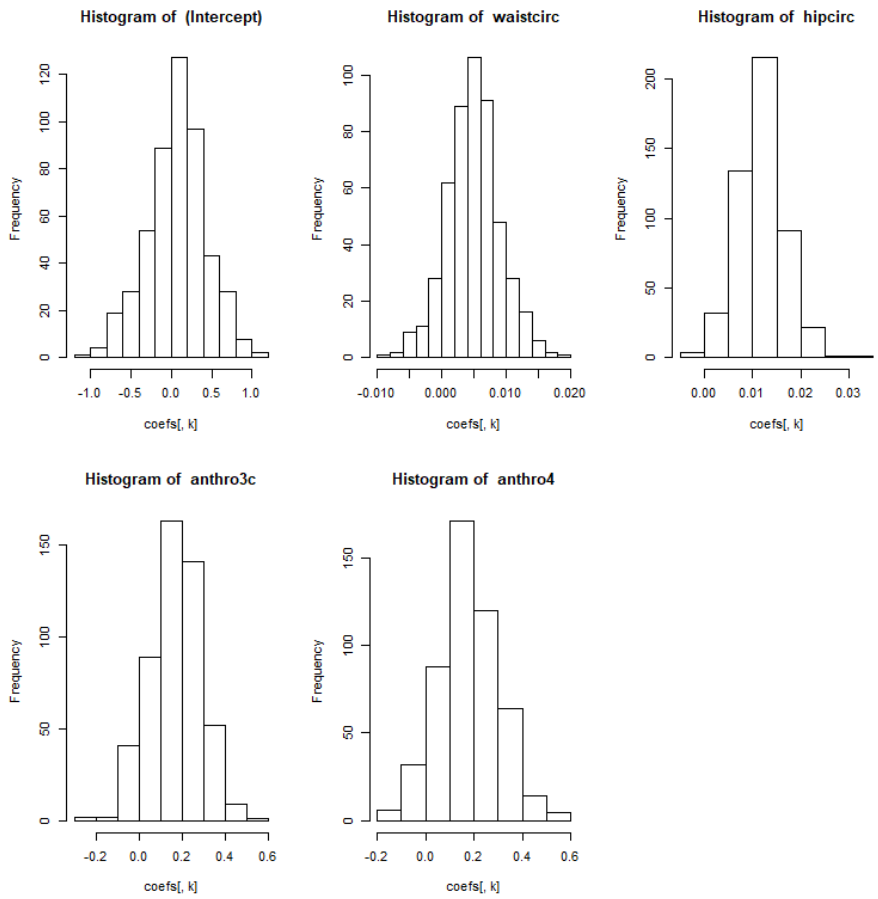

Figure 20: Histogram of Coefficients after bootstrap sampling of the model

1.4 CONCLUSION:

The final model that we prefer is the one with the log transformation of the response variable and the regressors we included in the model are waistcrc, hipcirc, anthro3c and anthro4. The model was finalized based on backward variable selection using evaluation criteria such as AIC,BIC, Adjusted R2, Mallow's Cp and MSE. We also removed highly correlated variables using VIF tests. Outliers in the model were diagosed, but were not removed as they were not affecting the model greatly. Residual analysis was also performed to ensure that the assumptions of the error are not violated and finally, we did bootstrapping of the model to obtain the confidence interval for the co-efficients. Thus, the regression model developed in this project is;

$$log(\textbf{DEXfat}) = 0.108926 + 0.004888 * \textbf{waistcirc} + 0.011831 * \textbf{hipcirc} + 0.163928 * \textbf{anthro3c} + 0.174321 * \textbf{anthro4}$$

whose,

R-squared = **0.9329**

Adjusted R-squared = **0.9288**

AIC = **-116.8352** and BIC = **-103.2591**

REFERENCES

[1] P. Bühlmann, M. Kalisch, and L. Meier. High-dimensional statistics with a view toward applications in biology. Annual Review of Statistics and Its Application, 1(1):255– 278, 2014. doi: 10.1146/annurev-statistics-022513-115545. https://doi.org/10.1146/annurev-statistics-022513-115545

[2] R. Dezeure, P. Bühlmann, L. Meier, N. Meinshausen, et al. High-dimensional inference: Confidence intervals, p-values and r-software hdi. Statistical science, 30(4): 533–558, 2015

[3] L. Garcia, K. Wagner, T. Hothorn, C. Koebnick, H.-J. F. Zunft, and U. Trippo. Improved prediction of body fat by measuring skinfold thickness, circumferences, and bone breadths. Obesity Research, 13(3):626–634, 2005. ISSN 1550-8528.

[4] D. Montgomery, E. Peck, and G. Vining. Introduction to Linear Regression Analysis. Wiley Series in Probability and Statistics. Wiley, 2012. ISBN 9780470542811.