STAT 628 Module 2 Group 7

Motivation

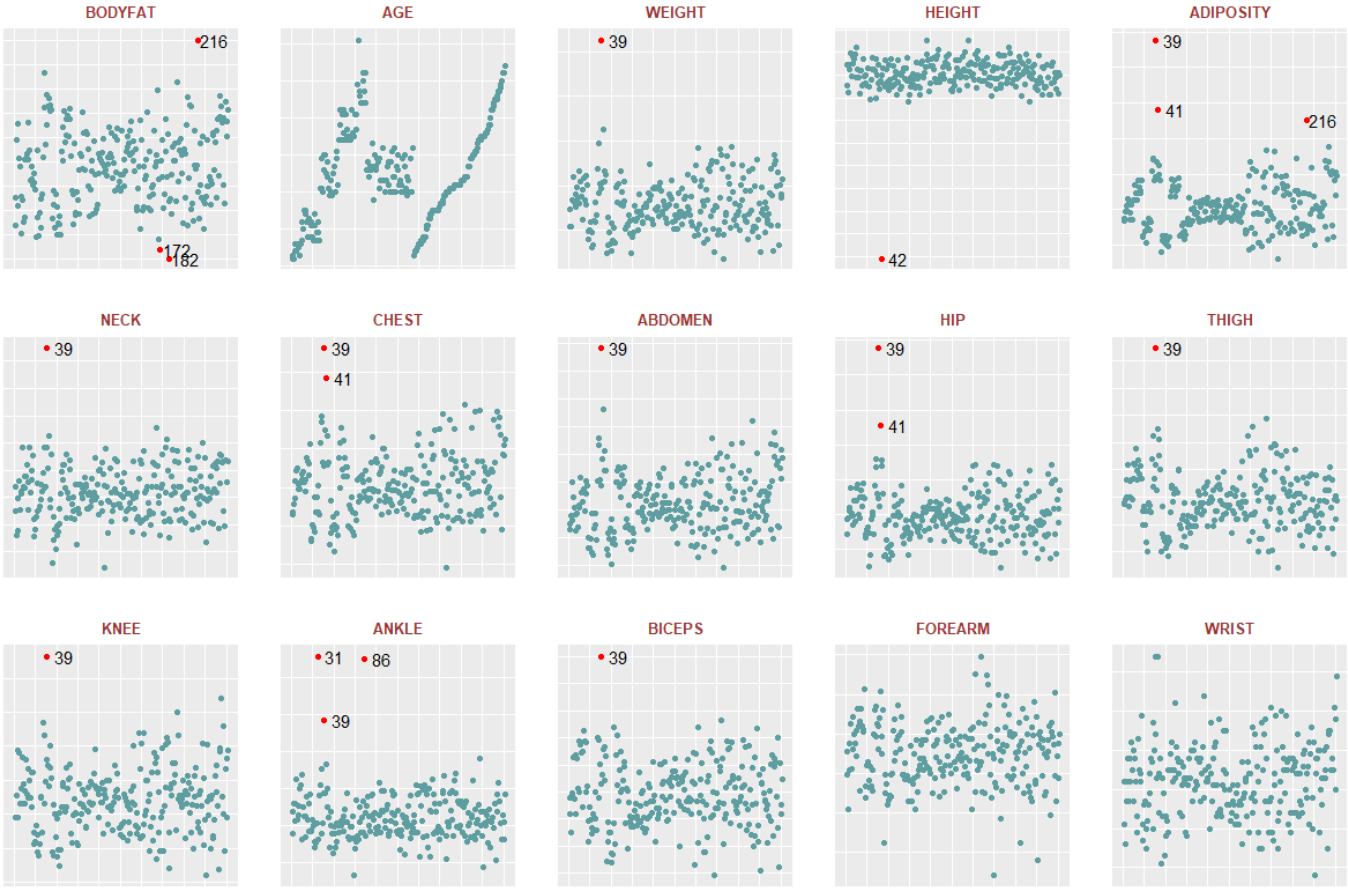
In our project, we want to use the BODYFAT data to train a linear model in order to predict the body fat by selected features. To let it be convienient for people to calculate their body fat, we want to make our model involve in as less variables as possible.To sum up, our project aims to give people a quick and simple way to get an estimation of their body fat value. Detailed codes are presented in our [Code Notebook (https://github.com/OliverXUZY/STAT628/blob/master/code/Code_Notebook.ipynb)](https://github.com/OliverXUZY/STAT628/blob/master/code/Code_Notebook.ipynb).

Data description

We use the body fat data with 252 observations and 17 variables. The 17 variables are: ID number, Body fat, density, age, weight, height, Adiposity, Neck, Chest, Abdomen, Hip, Thigh, Knee, Ankel, Biceps, Forearm, Wrist. In the 17 variables, we use the Body fat as the reponse variables. Since body fat is calculated by the density according to Folk, in the data, the density can explain over 90% variation of the body fat and, in addition, it is hard for people to measure their density, we exclude this variable from our data analysis. Finally, we have a objective variable and 15 candidate variables.

Data Processing

We first plot the scatter plots of the original data:



We may notice some potential outliers in each plots. We found some problems in the original data and processed as below:

- NO.31 and NO.86 has much thicker ankle than others. NO.39 seems to be extreme fat compared with others. We delete NO.31, NO.39 and NO.86.
- NO.41 is fatter than normal obviously but not unacceptable. We keep this point.

- NO.216 has extremely high bodyfat, NO.172 and NO.182 have extremely low bodyfat. So we delete these three points.
- NO.42 typically has wrong height. We recalculated its height with BMI and weight, which is 69.43.
- We replace the BMI of NO.163 and NO.221 with recalculated BMI because their values in the original data is quite different (difference larger than 1 unit) from the recalculated value.

Then we did a simple linear regression on the whole dataset after data processing for further investigation. The residuals did not have obvious patterns and QQ-plot suggests normality. Their leverage values and DFFITS values are not unacceptable. We decided not to further remove any new points.

Model selection

After removing the outliers and replacing the wrong record with approximately estimated one, we got the appropriate data that suitable for data analysis. The next step comes naturally: selecting the variables and construct the model. Since there is linear relationship between BodyFat and some other variables, we consider constructing a linear model to predict the BodyFat. According to our previous linear naive model, the linear combination of certain variables can actually account for most of the information of BodyFat. Thus, we regard linear model as a reasonable choice.

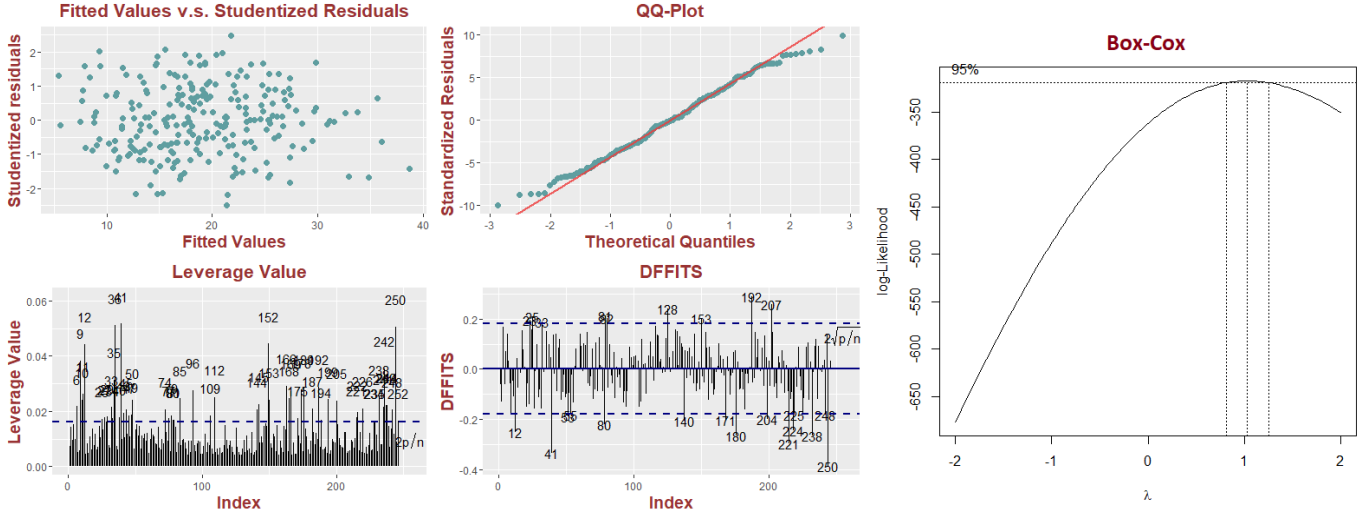
By observing directly, we found there is strong colinearity among several predictors, the Variance Influential Factor also bolster this conclusion. We need select a relatively small set of variables as predictors in the model. Considering the severe colinearity in the model, and wanting our application as simple as possible (a.e. a off the shelf method for a non-statistician required a little inputs applied on the application), we use forward(empirically select less variables than backward) step-wise selection with AIC criterion which balance the fidelity and model capacity.

Forward AIC selection

After selection, we have 4 variables: ABDOMEN + WEIGHT + WRIST + BICEPS. Then we plot the figure to depict the correlation between them.

There is still colinearity among these 4 variables, we use **All Possible Subsets Methods** to select the best subset. We search all the possible subset of variables as predictors, and compare different subsets with R^2, adjusted R^2 and C_p criterion.

The result shows the best 2 variables are **weight** and **abdomen**, the best 3 variables are **weight** and **abdomen** and **wrist**. The additional **wrist** variable can only cause increasing of R^2 of no more than 0.01. We keep **weight** and **abdomen** as our final predictors in our linear model. We also search all the possible subsets among all (total 14) variables, the result verify our current conclusion.



As we can see, the residual are randomly scattered around both side of x axis. The scatter plot didn't show any specific pattern. The QQplot shows the data has a heavy tail. We might need to construct boxcox transformation to meet the normality assumption.

Since the boxcox didn't suggest any transformation on the data, we use our raw data and the model we got right now to analysis.

Model Summary

The summary of our final model is shown as follows:

Coefficients	Estimate	Standard Error	p-value	CI.lower	CI.upper
Intercept	-41.45	2.50	<2e-16	-46.372	-36.524
Weight	-0.12	0.02	3.13e-09	-0.159	-0.082
Abdomen	0.89	0.05	<2e-16	0.783	0.990

From the summary we can see that all the coefficients of the variables are significant and the standard error of the coefficients are small. The lower bound and the upper bound of the 95% confidence interval is shown in the last two columns of the table, the CI's are relatively narrow. The R^2 is 0.708, which means our model explains 71% of the variation and we think this is a desirable number.

Conclusion

We find that the best way to predict one's bodyfat percentage is to use his weight and abdomen. If we denote bodyfat, weight and abdomen as b,w and a, the formula for calculating bodyfat percentage is $\hat{b}$ (%) = -42.63 -0.12 w (lbs) +0.89 a (cm).

Advantages and Disadvantages

The basic advantage of our algorithm is that it's really simple and easy to calculate. Also, the two data, weight and abdomen, are easy to obtain, so it is of practical use, especially for doctors and normal users. Another advantage is that the standard error of the residuals are small, which means the variation of our prediction will be small.

However, this model also has its demerits which is the strong multicollinearty in the model. If we draw a plot about the two explanatory variables abdomen and weight we can find a high positive correlation between them. As a result, the normal inference procedure no longer makes sense and it's difficult to interpret the coefficients as the effect of weight and abdomen on bodyfat, since while we change one of them, the other can't be held constant. For doctors and normal users, they still can use the formula get the prediction of the bodyfat. Moreover, they maybe do not need or care about the details behind the algorithm. So this formula is at least accurate and useful for the doctors and the public.

Contribution

- **Zhao Li:** Data processing and related images and codes. Data processing part in presentation.
- **Yujie Zhang:** Summary and model evaluation, related images and codes, example in presentation.
- **Yaobin Ling:** Shiny. Simple linear regression model, model evaluation and related presentation.
- **Zhuoyan Xu** Model selection and related images and codes and related part in presentation.