Module 2 Summary

Contributor : Xu Zou, Yunqing Shao, Qizhou Huang

1 Introduction and Rule of Thumb

As a group of data scientists, our goal is to figure out a reasonable and precise model to predict body fat percentage of males based on BodyFat Dataset. After data cleaning, we first build a simple linear model with single covariate “Abdomen”. Second, we use added variable plot to see whether any other covariates can be added to improve the accuracy of model. Then we compare the value of R square of different models to find the best one, which is $BodyFat \sim Abdomen + Height$. We diagnose the normality and influential points using Q-Q plot and cook’s distance to show the final model is appropriate.

Based on final model, the rule of thumb is “multiply your Abdomen(cm) by 0.64 and Height(cm) by -0.21, sum them and minus 2.8”. So, a man’s body fat (%) with Abdomen (85cm) and Height (172cm) is about (15.48%). Moreover, if his age ranges from 20 to 40, we have 90% accurate rate that the body fat is among 15.48 $\pm$ 5 %.

2 Data Cleaning and Model Building

We use the equation in data description to calculate body fat by body density and remove 5 points that do not satisfy the equation. We also search the equation to calculate BMI by height and weight, remove 3 points. Then, we transferred the unit of height to cm.

We hope to build a simple linear model between BodyFat and one or two explainable body measurements such as age, weight and chest circumference.

The first step is to choose an initial model that contains only one body measurement as covariate. After drawing the scatter graph between each body measurement and BodyFat, we choose Abdomen as the covariate since it has a strong linear relationship with BodyFat and the model R square value is 0.6484.

$$Model_1: \quad BodyFat = \beta_0 + \beta_1 \times Abdomen + \epsilon$$

When we draw the scatter graph between Abdomen and BodyFat, we find there is an outlier, which may strongly affect the regression line. So we remove this point and use the new dataset to reconstruct model 1:

$$Model_2: \quad BodyFat = \beta_0 + \beta_1 \times Abdomen + \epsilon \quad \textbf{without outlier}$$

The R square value is 0.6664, larger than model 1’s, which implies the rationality of removing outlier. In the later model, we will keep using the new dataset.

Next, we consider to add another covariate to model 2 to improve the accuracy of model. We use added variable plot. The added variable plot is used to check whether a new variable should be included. By using this method, we plot scatter graph between residuals of $Y \sim X_1$ and $X_2 \sim X_1$ to see whether it demonstrates a linear relationship. Finally, we find two covariates likely to be added into the model 2: Age and Height. Their added variable plots are shown below:

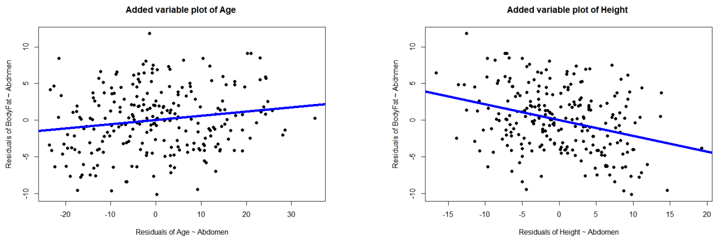


Figure 1: Added Variable Plot

The two models are:

$$Model_3 : \quad BodyFat = \beta_0 + \beta_1 \times Abdomen + \beta_2 \times Age + \epsilon$$

$$Model_4 : \quad BodyFat = \beta_0 + \beta_1 \times Abdomen + \beta_2 \times Height + \epsilon$$

3 Model Comparison

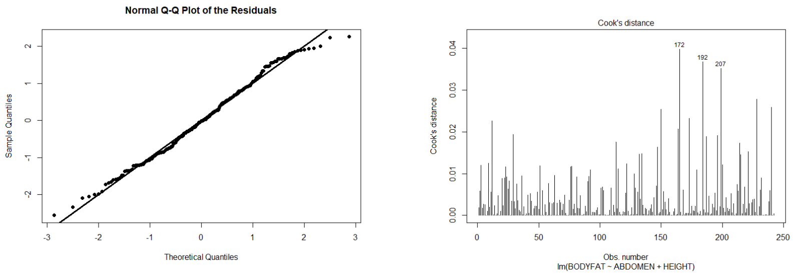
We use R square value to compare models. From the table, Model 4 has the highest R square value. Although it has two covariates, it performs much better than Model 2 and it is meaningful to add Height to initial model. So our final model is Model 4.

Table 1: Model Comparison

Model		R^2
$Model_1$:	$BodyFat \sim Abdomen$	0.6484
$Model_2$:	$BodyFat \sim Abdomen$ (without outlier)	0.6664
$Model_3$:	$BodyFat \sim Abdomen + Age$	0.6760
$Model_4$:	$BodyFat \sim Abdomen + Height$	0.6978

4 Model Diagnostics

For diagnosis to the final model, we firstly check the assumption of normality. We use Q-Q plot to diagnose, see Figure 2(a). The graph shows that the residuals of the final model could be assumed to normal distributed. Then we plot the cook's distance to check whether there still exist influential points, see Figure 2(b). It shows the largest cook's distance is less than 0.04, implying no influential point.



(a) Q-Q plot

(b) Cook's distance

Figure 2: Diagnostic

5 Model Strengths and Weaknesses

It's a linear regression model which contains only two covariates that can be measured easily.The R square value is about 0.7, means we can explain 70% variance. Since the model is simple, it would not over-fit .

However, after doing more analysis. we find that, we only have 82% predictions within +/- 5% of true value when individual's age ranges from 40 to 60, although 90% predictions when age ranges from 20 to 40. So we have higher accuracy to predict the younger than the elder. The reason why this happens requires more discussion.

6 Conclusion

The final model is $BodyFat = -2.8040 + 0.6363 \times Abdomen - 0.2073 \times Height$ .This model has 70% prediction accuracy and it's easy to understand and explain.

7 Contribution

We did all parts of group work with collaboration and have discussed for over 10 times before we get the final version.

Qizhou Huang wrote the initial code, including data cleaning, model building and model diagnosis. He also worked on Shinyapp, giving an initial demo. He wrote Data Cleaning of the summary.

Xu Zou created the github repository. He simplified the code and gave some annotations. He corrected some error on code and wrote code including model comparison and model weakness. He wrote summary from Model Building to Model Strengths part. He also worked on Shinyapp, adding more if/conditional statement code to ensure the robust of Shinyapp.

Yunqing Shao wrote the remaining part of the summary, including Introduction and conclusion. She also wrote the whole slides.