628 Module 2

Group 7

October 2022

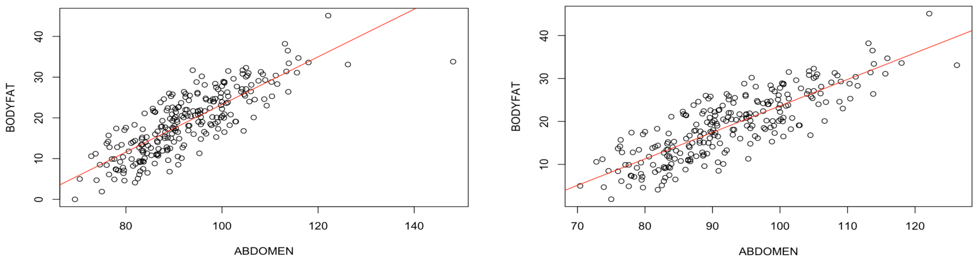
Introduction: One person's body fat can be reflected by many aspects such as the BMI and the fat of arms and legs. It will also be affected by someone's age, height and weight. Therefore, we want to select the most essential factor that can reflect the body fat accurately so that we will simplify the way of determining someone's body fat.
Background and Data Cleaning: The response variable Y we focus on is the body fat percentage with mean 18.93% which indicates 18.93% of a male's body in our sample is fat on average. And the corresponding predictors reflect a male's body fat through his age, height, weight and the body fat in all parts of the body including abdomen, chest, hip, etc. We remove two individuals with body fat less than 4% since it's too low for a normal male and one individual with height less than 30 which is different from other heights ranging from 60 to 80.
Shiny App: We use shiny app to systematically explore and understand the relationship between body fat and other variables. Fristly, we divide the age variable into 3 different strata. By comparing the different plot, the relationship between age and body fat is not obvious since the distribution of y-value(body fat) doesn't have dramatic difference. From the first plot, we could check there is a positive relationship between weight and body fat in all age strata. In the second plot, we found that the height doesn't have a significant relationship with body fat, since the trend is not clear. In the third plot, there is absolutely positive trend between abdomen and body fat. It is meaningful for us to analyze this relation since it's a suprising result that body fat is associated with a specific part of human body. The shiny app link is:ShinyApp
Choosing Model: Our final model is the simple linear regression model with the response variable BodyFat and the predictor Abdomen:

$$\text{BodyFat} = \beta_0 + \beta_1 \text{Abdomen} + \epsilon,\ \epsilon \sim N(0, \sigma^2)$$

After removing the outliers, our estimated intercept $\hat{\beta_0}$ is -37.45 and the estimated coefficient of Abdomen $\hat{\beta_1}$ is 0.61, so our estimated regression line is:

$$\hat{\text{BodyFat}} = -37.45 + 0.61\hat{\text{Abdomen}},$$

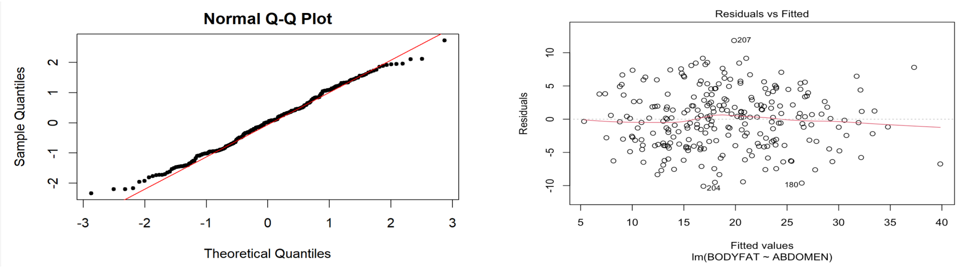
which indicates every 1-cm increase in Abdomen will cause 0.61% increase in BodyFat. Therefore, the body fat percentage of a man with 80 cm abdomen 2 circumference is likely to be 11.42% with the confidence interval between 10.55% and 12.29%. The reason why we choose this model and select Abdomen to be our predictor is this model has the lowest RMSE (Root-mean-square deviation$\sqrt{\frac{\Sigma_1^n (y_i - \hat{y_i})^2}{n}}$= 4.48) when we build the one-factor linear regression models with the response variable BodyFat and each related attribute. And the reason why we use RMSE to compare each model is that it has the same unit as the dependent variable.
After we determine the predictor Abdomen, we further improve our model by removing other outliers. We remove 39th individual with 148.1 cm Abdomen 2 circumference which brings some deviation of the regression line (See Figure 1-(a)) and we also remove 180th, 204th and 207th individuals which have relatively larger residuals to get the final model(See Figure 1-(b)).
Statistical Analysis: We use ANOVA F-test to test whether the null hypothesis $H_0 : \hat{\beta_1} = 0$ is true or the alternative hypothesis $H_1 : \hat{\beta_1} \neq 0$ is true. The result shows that the p-value is quite small ($< 2.2 * 10^{-16}$) which indicates we should reject H_0 at significance level $\alpha = 0.05$. Therefore, $\hat{\beta_1} \neq 0$ and our predictor Abdomen is significant at $\alpha = 0.05$. In other words, abdomen 2 circumference can reflect the body fat effectively and adequately.
Model Diagnostics: First, we check the normality of the residuals in our final model through Q-Q plot (See Figure 2-(a)). It's obvious that the residuals follow the red straight line well indicating the residuals are normally distributed which meets the assumption of a linear model (Assumption: residuals $\epsilon \sim N(0, \sigma^2)$).
Then, we check the equality of variance. The plot of Residuals VS. Fitted Values shows that the residuals equally spread around a horizontal line without distinct patterns (See Figure 2-(b)). So, there is non-linear patterns in residuals which also meets the assumption of residuals in the linear model (Assumption: the variance of the errors does not depend on the values of the predictor variables).
Model Strengths and Weaknesses:



(a) Initial Model

(b) Final Model

Figure 1: Model Comparison



(a) Normal Q-Q Plot

(b) Residuals VS. Fitted Values

Figure 2: Model Diagnostics

Strengths:

· Our model is simple and robust that follows the rule of thumb since we want to figure out the most effective variable to predict the body fat.

· Our model is very easy to interpret, that is, we can directly see the linear trend between BodyFat and Abdomen, which provides insightful information to users.

Weaknesses:

· Simple linear model may not capture a very comprehensive picture of all the relevant independent variables related to body fat.

Conclusion: We want to figure out the most important and relevant variable of body fat so that after we remove the abnormal data we build the linear regression models between each attribute and the body fat and choose the model with the smallest RMSE. The model with the smallest RMSE is the one with independent variable Abdomen and dependent variable BodyFat. We improve the model by removing the outliers and some points with relatively larger residuals. And finally we generate a shinyApp to show the relationship between abdomen 2 circumference and body fat, weight and body fat and height and body fat under different ranges of age.

Contributions:

1. Yuqian Chen wrote the Choosing Model, Statistical Analysis and Model Diagnostics parts in the two-page summary. YC did the data cleaning, revised the data used in the code and made some plots of the model and data.
2. Duohan Zhang wrote the codes for data cleaning, model choosing and model diagnostics. He also wrote model strengths and weaknesses in the summary, and model diagnostics in the final slides.
3. Jianzhuo Liu wrote the codes of shiny app. He also wrote the shiny app part in the report, and the shiny app part in the final slides.

References

[Chr02] Ronald Christensen. *Plane answers to complex questions*, volume 35. Springer, 2002.