STAT 628: Module 2

Xinyue Wang, Chenhao Fang, Milica Cvetkovic

2020-10-26

Introduction

Written by Chenhao Fang and edited by Xinyue and Milica. In this project, we first cleaned the data, then, we considered several models and found the best one by AIC in a stepwise algorithm and did some statistical tests on it. Next, we presented the model diagnostics and talked about the strengths and weaknesses of this model. Finally, we proposed our Rule of Thumbs.

Data Cleaning

Written by Milica and edited by Xinyue. Before any data cleaning, I researched how to calculate body fat. I read the description of this specific data set and how the data was collected. It was important to note that all the subjects were men. This gave me some ideas about averages in body measurements. I started the data cleaning with unit conversions. Next, after a comprehensive reading of the data summary, I spotted some strange behaviors. For example, the minimum body fat was 0, and there were atypical occurrences with height and weight. I recalculated body fat using the Siri's equation and replaced those values. It turned out that the body density had a typo. To correct this, I made some calculated deductions. I did similar processes with other features. In addition, I visualized data using interactive data plots. Not only this helped me see the outliers, but also the interactivity enables me to know exactly which data points were outliers.
Xinyue Wang's Data Cleaning – Before building the model. After recalculating BMI, I compared two BMI results and detected that the IDNO 42, 163, and 221 might have wrong weights and heights. After the analysis, I added 100cm to the 42nd height, added 20lbs to the 221st weight(lbs), and subtracted 20lbs from the 163rd weight(lbs). I removed the 172nd and 182nd data points according to Milica's suggestion.

Choosing Model

Written by Chenhao Fang and edited by Xinyue and Milica. Before we chose the model, we selected variables for the best subset of predictors. We wanted to explain the data and the variance most simply and remove redundant variables. The goal was to also address multicollinearity. In our model selection part, we used backward elimination. We first started with all the variables and removed some variables that have relatively low p-values. When we chose the final model between these models, we used the Akaike Information Criterion for our criteria. The AIC is a metric that can determine which model is most likely to be the best model among others for a given data set. Note that it estimates models relatively. We would choose the model with the lowest AIC. **The final model:**

$$\begin{aligned}BodyFat = 7.77622 - 0.12630 * HEIGHT\_CM + 0.05329 * AGE - 0.37239 * NECK \\ + 0.72955 * ABDOMEN + 0.27822 * FOREARM - 1.6408 * WRIST\end{aligned}$$

A sample use: a man with height of 178 cm, 44 years old, neck circumference 38 cm, abdomen circumference 93 cm, forearm 29 cm and wrist 18 cm, has a predicted body fat percentage 41.3%.

Statistical Analysis

Written by Chenhao Fang, edited by Milica We conducted a t-test and F-test for our model. By the result of our t-test, most of our coefficients' p-values were much smaller than 0.05, which showed that our model's variables are statistically significant.
Also, we conducted the F-test, which tests if at least one feature has a significant effect. Our model's F-statistic is large, and the p-value is almost 0. Therefore, we should reject the null hypothesis that all of our model's coefficients are 0.

Model Diagnostics

Written by Xinyue, edited by Milica. First, we checked the outliers. I calculated the hat matrix, Cook's distance, and influential points and plotted Residuals vs. Fitted and Leverage plots, among others. I deleted IDNO 39 and kept the other seemingly outliers because of the small dataset. Then I fitted the model again. The following results are from the "new" model. (More details can be found in the slides.)

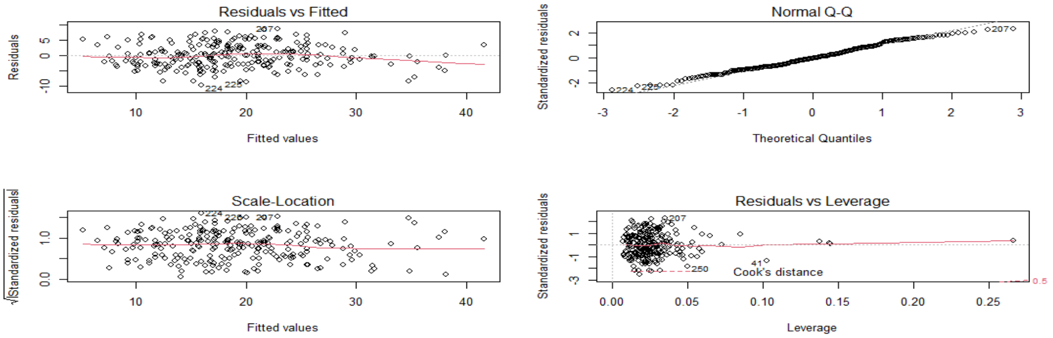

Figure 1: Model Diagnostics

Next, I assessed the linear model assumptions using the global test with 4 degrees of freedom with the "gvlma" function. With under 0.05 level of significance, all decisions were "Assumptions Acceptable".
To check the normality, I used QQplot with simulation. Although the tail is slightly off the line, we can assume that it does not violate the normal assumption.
To check the homoscedasticity, I used the "ncvTest" function. It computes a score test of the hypothesis of constant error variance. The alternative hypothesis is that the error variance changes with the level of the response (in our case, the body fat) or with a linear combination of predictors. The p-value is 0.96059, which means we can believe the constant variance hypothesis.
To check the multicollinearity, I used the "vif" function to calculate the variance inflation factors. Since correlations between some variables were high, specifically between weight(kg) and THIGH, NECK, HIP, and ABDOMEN, we wanted to ensure that they will not appear in the model at the same time. The model we selected does not show any problem with multicollinearity, which makes it more reliable. Finally, I used, for example, the Durbin-Watson test to check the independence of errors, and I also used "crPlots" to check the linearity between variables and body fat. Our model performed very well in all of these aspects.

Model Strengths and Weaknesses

Written by Xinyue. **Strengths**: Our model is simple and can be interpreted easily because it does not contain any cross terms or higher order terms which can't improve the model significantly. Our model does not violate any assumptions of linear regression, especially multicollinearity. Moreover, our model can explain more than 73% of variation in body fat. Besides, since we used AIC as the criterion, we have largely avoided the problem of overfitting. **Weaknesses**: The R^2 or adjusted R^2 of this model is not the largest among our candidates. But because all of them are very similar, it doesn't weaken our confidence in this model.

Conclusion

Written by Chenhao Fang, Xinyue and Milica. In conclusion, we created a elegant multivariable linear model with statistical significance which explained most of variation of the body fat. We believe that the simplicity of our model and design of the app will aid the body fat calculations and the research on body fat.