STAT 628: Module 2- Body Fat Prediction

Group 18: Midhun Satheesh, Jiaqi Xia, Yulong Zhao

1 Introduction

Bodyfat percentage is an important measure of the risk of weight-related diseases. Based on a comprehensive dataset that lists estimates of the percentage of body fat determined by underwater weighing and various body circumference measurements for 252 men, we are trying to build a simple model to estimate the bodyfat percentage as accurately as possible.

2 Background Information & Data Cleaning

Some columns are highly skewed [reference] like HEIGHT, ADIPOSITY, and ANKLE ; from the correlation matrix heatmap, multicollinearity is present in the data, such as WEIGHT is highly correlated with HIP and THIGH. And we removed outliers outside the 75th percentile (Q3) and the 25th percentile (Q1)[reference].

3 Model Fitting

- **Our final model: BODYFAT = -43.6366 -0.1211 * WEIGHT + 0.9110 * ABDOMEN.**
- **Possible rule of thumb: BODYFAT = -44 + ABDOMEN -0.1*WEIGHT.**
- **Example usage:** A man weighing 180 lbs and measuring 90 cm of abdomen circumference is expected to have a body fat % of 16.55 based on our model, which is healthy according to Bodyfat Percentage Chart[reference]. His 95% prediction interval is between 8.513 and 24.595.
- **Interpret the model:** Our estimated coefficients are -0.1211 and 0.9110, which are in the units of lbs, and cm. This means that for every one unit of WEIGHT decrease in lbs, the body fat % will increase, on average by 0.1211 based on our model; a change in one cm in ABDOMEN circumference will bring 0.9110 units to change in body fat %.
- **Why chose the final model:** First, all variables are easy to obtain. Second, we checked the relationships between variables and found that bodyfat is linearly correlated with variables we chose, so we thought multiple linear regression is reasonable. Third, other models with more variables, THIGH for example, have similar performance with the final model. Considering the model simplicity, we choose our final model with only WEIGHT and ABDOMEN.

4 Statistical Analysis

We conducted the following tests to see whether the predictors we chose are significant in predicting the outcome, including **t-test, F-test and R-square after modeling**. Suppose our null hypothesis is that the slope is equal to 0. Our estimated slopes are -0.1211 and 0.9110 and estimated intercept is -43.6366, all with 95% confidence. We found that all p-values are smaller than 0.05, which indicates that our model is significant. Our R-square is 0.7195, which implies that 71.95% of the data fits well with the final model. Considering the volume of the dataset is about 200, the R-square is acceptable.

Model	WEIGHT + THIGH + ABDOMEN	WEIGHT + ABDOMEN
AIC	1402.94	1405.32
R-Squared	0.7245	.7195

Table 1: AIC and R-square of candidate and selected model

5 Model Diagnostics

After model fitting, we diagnosed the MLR assumptions with a residual plot and a QQ plot. We also checked for leverage and influential points.

- We checked linearity and homoscedasticity (see figure 1 below) assumptions using residual plot. Linearity seems reasonable because there are no obvious non-linear trends in the residual plot; the points look randomly scattered around the X axis. Homoscedasticity is also plausible, since the spread of points did not follow a funnel shape.
- We checked normality using QQ plot (see figure 2 below). Normality also looks reasonable because the points in the QQ plot follow a roughly straight line.

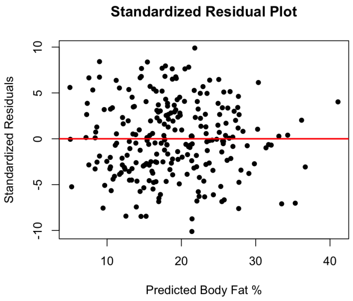

Figure 1: Residual plot

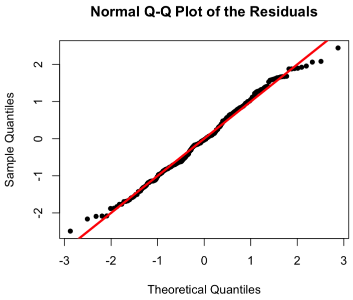

Figure 2: QQ Plot

- We also looked at three types of outliers in regression models: we checked outliers in BODYFAT using studentized residual plot, checked leverage points using pii measures, and checked influential points using both the Cook's distance and the pii measures. From studentized residual plot, since all points are within the horizontal line of t3 and -1, there does not seem to be any outliers in BODYFAT. There does not seem to be any leverage point. Also, there does not seem to be any influential points by comparing both the Cook's distance and the leverage values.

6 Model Strengths/Weaknesses

Some strengths of our model include the model simplicity, model assumptions. In particular, our model only includes 2 easily accessible variables, which is convenient enough to gather information of customers. Also, our model satisfies the linear regression assumptions of linearity, homoscedasticity and normality, bringing credence to our results in prediction. Some weaknesses of our model include constant effects, relatively low R-square. In particular, people of different age group have different lifestyles, the slope coefficients may be different for these two subgroups of men. Compared with other models using polynomial methods or machine learning methods, our model is not very accurate.

7 Conclusion

Overall, our model provides a very simple way of estimating the body fat % purely based on only weight and abdomen circumference. MLR assumptions (linearity, homoscedasticity and normality) are plausible. After removing the outliers, there are not any leverage points and influential points. When interpreting the linear model, the constant effect assumption is likely violated. Further experiments are needed to customize slope coefficients for different age groups and improve the model accuracy.

8 Contributions

- **MIDHUN** did data cleaning (removing the outliers),residual analysis, and modeling, co-wrote two-page summary on latex, finished shiny app and publish the Github repository.

- **YULONG** finished presentation slides, chose model, co-wrote two-page summary, co-edited R code.

- **JIAQI** wrote two-page executive summary, did EDA and chose model, final editing of R code.

9 References

- Skew and Kurtosis: 2 Important Statistics terms you need to know in Data Science

- How to Remove Outliers from Data in R

- deal Body Fat Percentage Chart: How Lean Should You Be?