STAT 628 Module 2 Summary

Xingrong Chen, Ziao Zhang, Chongxuan Bi

1 Introduction

Body fat percentage is an effective measure of fitness level. However, it is not straightforward to measure body fat, and we need to infer it from other characteristics of a person. In this project, we used a dataset of 252 men's measurements of body fat percentage and various other features to derive a simple and robust statistical model for predicting body fat.

2 Data Cleaning

First, we identified columns that could be recomputed using others. Then we utilized box plots to detect each outlier, and recomputed its value using information about other columns. If the new value is distant from the original data, then it's replaced.

3 Modeling

We start with a simple regression model within which the variable selection is based on the correlation. The model shows below:

$$\widehat{BODYFAT} = -35.19 + 0.58 * ABDOMEN$$

For example, a man with abdomen circumference 100cm is expected to have a body fact of 23.4 and the 95% prediction interval is [14.4, 32.3]. Furthermore, it means that if abdomen circumference increases 10cm, body fat increases 5.8% on average. Meanwhile, the R^2 of the model is 0.66, which implies 66% of the variation in body fat% is explained by abdomen.

Besides, we also conducted a t-test for the hypothesis $H_0 : \beta_1 = 0$ vs $H_1 : \beta_1 \neq 0$. P-value is close to zero, hence, abdomen circumference shows great significance in predicting body fat. Therefore there is a strong correlation between them.

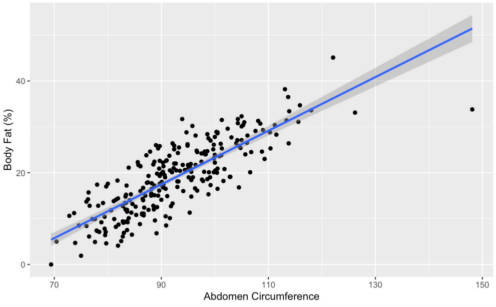

Figure 1: Regression

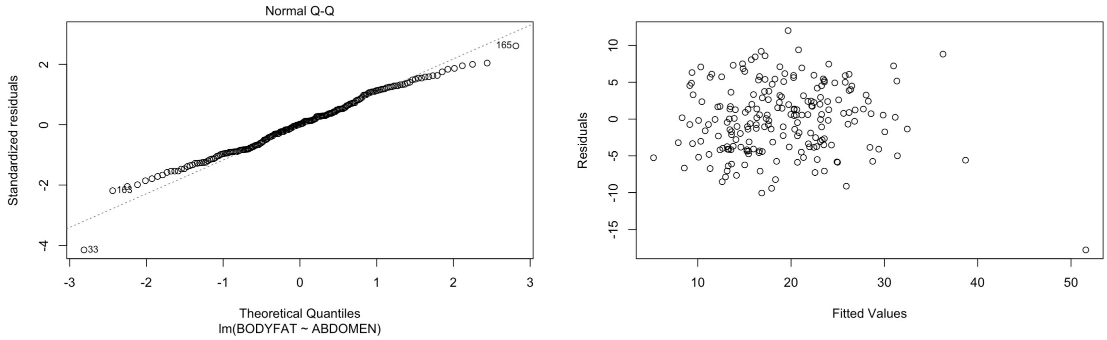

Figure 2: Model Diagnostics

3.1 Diagnostics

1. Since there are no obvious trends in the residual plot
(Figure 2), the linearity assumption and the constant variance assumption are satisfied.
2. The QQ-plot shows a straight 45 degree line, meaning that errors are normally distributed.
3. Although there is an outlier in the dataset, we decided to keep it because there is only one outlier and removing it causes RMSE in the testings set to decrease.

4 Strengths & Weaknesses

4.1 Strengths

The model is outstanding in its simplicity. Using abdomen circumference as the only predictor in the linear model makes predicting body fat easy and fast.

The model is valid by satisfying all observable linear regression assumptions. We are able to demonstrate the linear relationship between the predictor and the response, constant variance in residuals across fitted values, and normally distributed residuals. The estimates of the intercept and the predictor coefficients are also statistically significant.

The model is accurate and robust. While it achieves an R^2 close to more complex models, it proves to be more robust and not overfitting. After we fit models on the training set and made predictions on the testing set, the final model has the lowest RMSE.

The model has strong interpretability. Due to its simplicity, it's free from the multicollinearity issue. Therefore the interpretation is straightforward and unaffected by other variables. Meanwhile, since a single predictor is unlikely to overfit the data, the predictors are more probable to describe the real relationships rather than noises.

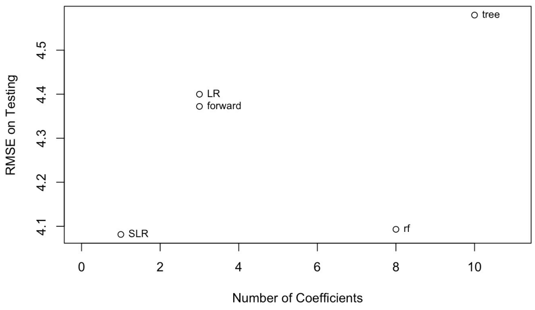

Figure 3: Model Comparison

4.2 Weaknesses

The data cleaning is inaccurate. We only identified outliers in predictors that are re-computable from others, and ignored those in other predictors, which could affect how the heatmap is drawn, and then how the models perform.

The simple linear regression model yields an R^2 that is 0.07 smaller than forward selection model, meaning that the latter model explains extra 7% of the variation in body fat.

The large estimated standard error causes prediction intervals to be wide, introducing more uncertainty. While the standard deviation of the sample is 7.7, 95% prediction interval has width 18, and even 50% interval has width more than 6.

5 Conclusion

In summary, the simple linear model has a good fit on the data. It is simple, satisfies all assumptions, fits the data well, and has strong generalizability without overfitting.

6 Contribution

6.1 Chongxuan Bi

CB wrote strengths weaknesses and conclusion of the summary. created analysis and shinyapp code, and worked on page 6-7 on slides.

6.2 Xingrong Chen

XC wrote introduction of the summary. maintained GitHub repository, and worked on page 3 on slides.

6.3 Ziao Zhang

ZZ Wrote the modeling of the summary. Tried some alternative models. Worked on page 4-5 on slides.