1 Group Members: Mengqi Li, Xiangyu Wang, Yiqun Xiao, and Zijun Feng.

2 Introduction (YX)

Body fat is an important indicator of physical health, but it is not easy to measure directly and conveniently. With a data set (252 observations, 16 variables) including body fat percentage for adult men, we tried to build up a regression model to predict body fat percentage with other features (e.g. weight, abdomen circumference). We did model selection and get the best model with 3 predictors. We also constructed a body fat calculator on Shiny App server based on our model.

3 Data Cleaning (ZF)

There are several abnormal data points in the origin data set:

No.48, 76, 96 don't follow the linear relationship between body fat and 1/density. The body fat of **No.172, 182** are 1.9 and 0, which are too low. The body fat of **No.216** is 45.1, which is too high. The body fat of these data points seems to be incorrect, so we delete them. (See figure 1)

The height of **No.42** is 29.5, which is too low and seems to be misrecorded, so we use Adiposity (BMI) and weight to recalculate it (we use the corrected height of No.42 in figure 2). The Adiposity (BMI) of **No.163, 221** are inconsistent with weight and height, but we don't know which value is misrecorded, so we have to delete them. The Adiposity (BMI) of **No.39** is 48.9, which is too large; We also checked the weight and some skinfold measurements of No.39, and they are also too large; It seems to be an outlier, so we delete it. (see figure 2).

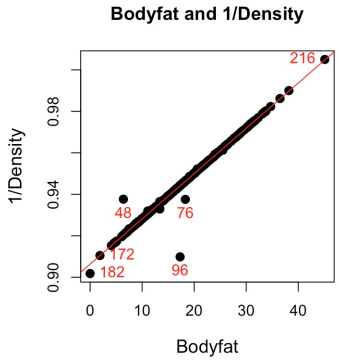

Figure 1: Bodyfat

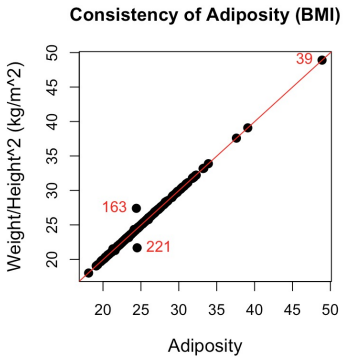

Figure 2: Adiposity (BMI)

In total, we deleted 9 data points and impute 1 data point. After data cleaning, the mean and standard deviation of body fat is 19.03% and 7.41%.

4 Model Choosing (ZF)

We decide to use linear regression to fit the Body Fat prediction model. There are 14 available features. First, we want to check whether we need to add higher order terms (like x_1^2, x_1x_2) into our model. We use ANOVA test to compare model 1 (with all 14 features) and model 2 (with all 14 features and second order terms). The null hypothesis is that the coefficient of all second order terms are 0. The result ($p = 0.28 > 0.05 = \alpha$) shows that second order terms are not significant, so it's unnecessary to add higher order terms into our model and we can just focus on 14 main effects. Since the data set is quite small, we can do best subsets regression and use some criteria to find the best model. The best model under $AdjR^2$, C_p and AIC criteria contain 8, 6 and 7 features respectively, while the best model under BIC criterion only contains 3 features. Then, we use leave one out cross validation (LOOCV) to get RMSE for these four models: 3.95, 3.93, 3.93, 3.97. The accuracy of them are almost the same, so we can choose the simplest model with 3 features (Weight, Abdomen and Wrist) under BIC criteria.

5 Statistical Analysis (XW)

We checked the scatter plots and found there are linear relationships between these 3 features and Body Fat, and there is no extreme outlier. Then we fit the model using R.

According to our model, if we want to predict Body Fat %, we can use the equation:
$\text{Body Fat}(\%) = -24.4911 - 0.0913\text{Weight(lbs)} + 0.8826\text{Abdomen(cm)} - 1.1954\text{Wrist(cm)}$
Given abdomen and wrist circumferences, For every 1 lb increase in weight of an adult man, the model predicts that body fat % will decrease by 0.09% on average. That may because the density of fat tissue is lower than lean tissue. We can interpret Abdomen and Wrist in the same way. It's reasonable that people who have larger abdomen circumference tend to have higher body fat and a larger wrist circumference also means more muscle which represents lower body fat to some extend. These 3 features explain about 73.41% of the variation in body fat.

Next we use t-test the relationship between each feature and Body Fat in our model. According to the results of hypothesis testing, we can declare that all 3 features are significant. What's more, the 95% confidence interval for Weight coefficient is $(-0.14, -0.05)$, for Abdomen coefficient is $(0.79, 0.99)$ and for Wrist coefficient is $(-2.02, -0.42)$.

6 Rule of thumb (XW)

Your body fat percentage can be estimated by: -24.5 minus your weight multiply by 0.1 plus your abdomen multiply by 0.9 minus your wrist multiply by 1.2.

7 Model Diagnostics (ML)

First we check the linearity and equal variance assumption using residuals vs fitted values plot. The plot shows no distinct pattern, which indicates the linearity assumption is satisfied. The points are randomly scattered, which means that the equal variance assumption is appropriate. Then we check the normality using Q-Q plot. As the Q-Q plot shows, the normality assumption is reasonable. Furthermore, we using p_{ii} values to check the leverage points and using cook's distance, DFFITS and DFBETAS to check the influential points. There are no influential points by comparing those measures. We also check multicollinearity by calculating VIF values. All the VIF values for each predictors are smaller than 10, which indicates there are no significant multicollinearity problems.

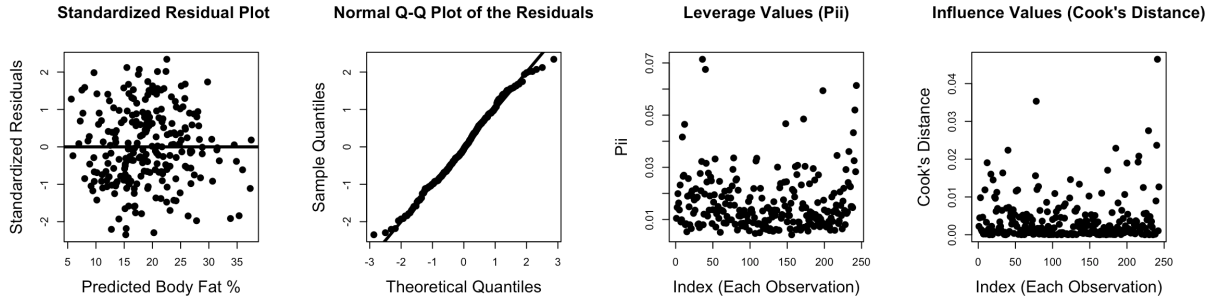

Figure 3: Diagnostic plots

8 Strengths & Weaknesses (ML)

We use best subsets regression which fits all possible models and find the best model based on BIC. Our model only include three predictors which is simple and easy to interpret. But we should mention that this model is only appropriate for calculating male bodyfat percentage. Meanwhile, since the model is linear, the prediction may be unreliable with extreme input values.

9 Conclusion (YX)

Due to the low capacity of the data set, it is not reasonable to adapt algorithms involving randomness because a slight change of the parameters may cause a significant difference in the final model. But the relatively low number of observations also makes exhaustive search possible. We confirmed that it is unnecessary to introduce interaction terms, and then selected 3 features to built up a linear model after searching through all possible subsets of explanatory features. The equation is also intuitive and interpretable. For instance, a higher circumference of abdomen indicates a higher body fat percentage, because most of body fat inside a man's body is concentrated in the abdomen. In addition, we want to emphasize that the model and calculator is only applicable to adult men because of the limitation of the given data set.