Analyzing Body Fat Dataset—estimating percentage of body fat using clinically available measurements

Jonquil Liao zliao42@wisc.edu
Runze You ryou3@wisc.edu
Yike Wang wang2557@wisc.edu

1 Introduction JL

Body fat percentage is an important index relating to a person's physical condition. Ideal level of body fat protects your body and helps you restrain energy. The main goal of this project is to generate an accurate body fat calculator, providing a simple method to make people aware of their health condition, thus take actions to fuel and reshape their bodies.

2 Data Cleaning YW

Before modeling, we need to check whether there are any outliers. Based on Siri's equation relating Bodyfat to Density, we detected 5 possible outliers by IQR method: No.48,76,96,182,216. We dropped 182 and 216 with body fat of 1.9% and 45.1%, which were not reliable according to some background research.

However, Siri's equation may not make a perfect prediction. We checked 48,76,96 by bivariate boxplots for variables vs Bodyfat/Density. Points 48,76 remained in the majority in every single plots and showed nothing suspicious, but 96 became an outlier in plots of variables vs Density. So we considered 96 with a wrong Density and kept 48,76,96. Additionally, we dropped points 39,41 that are far away from the majority in many plots of variables vs Bodyfat. Point 42 was far away from the majority in Height and we corrected it by BMI formula.

3 Motivation for Model YW

- **Model Selection**
  Given our cleaned data, there're 14 clinically measurements, and multicollinearity exists among these variables. We first fitted a linear model with all main factors and interaction terms. Then did all possible subsets and stepwise for model selection.
  We also tried Lasso regression for resolving multicollinearity. Compare different models by cross-validation, results were listed in Table 1.
  By comparison, we could see that with interaction terms, the fifth model has the obvious highest prediction accuracy—-with the lowest RMSE and MAE, highest R^2. So we chose it as our final model after trade-off.
- **Final Model**
  $$\begin{aligned}\text{BF} = &-2.728 - 1.143\text{Age} + 2.864\text{Adi} \\ &+ 0.347\text{Chest} + 0.685\text{Abd} - 5.288\text{Wrist} \\ &+ 0.067\text{Age:Wrist} - 0.022\text{Adi:Chest}\end{aligned} \tag{1}$$
- **Model Interpretation**

Criteria	Model	R^2	RMSE	MAE
Ad R^2	BF∼Age+Height+Neck+Chest+Abd+Bic+For+Wrist	0.75	3.71	3.12
C_p	BF∼Age+Height+Neck+Chest+Abd+Bic+Wrist	0.74	3.74	3.13
BIC	BF∼Height+Abd+Wrist	0.73	3.83	3.19
step bic(main)	BF∼Age+Adi+Chest+Abd+Wrist	0.73	3.83	3.18
step bic	BF∼Age+Adi+Chest+Abd+Wrist+Age:Wrist+Adi:Chest	0.76	3.53	2.84
Lasso	BF∼Age+Height+Neck+Abd+Wrist	0.73	3.83	3.20

Table 1: Candidate Models and Prediction Accuracy

To estimate a man's body fat percentage, for each term of variable with fixed other variables, the increase of Body Fat in percentile will be the sum of coefficients of itself and interaction with it.

- **Example Usage**
  For a 28 year-old man with 22.5bmi Adiposity, 93.5cm Chest, 74.4cm Abdomen, and 17.3cm Wrist, his estimated body fat percentage would be 7.1314%

4 Statistical Analysis YW

We conducted some statistical analyses by F test to see whether our predictors are significant. Partial F test: $H_0 : \beta_{age:wri} = \beta_{adi:che} = 0$ the F statistics=6.827 with p-value=0.001307, under $\alpha = 0.05$. Consequently, these two interaction terms were significant.

Seen from the Anova table of final model in Figure1 with partial SS, all predictors are significant with $p - value$<0.05, except for CHEST. But all predictors were kept.

```
Anova Table (Type III tests)

Response: BODYFAT
                Sum Sq  Df F value    Pr(>F)
(Intercept)       0.1   1  0.0097  0.921791
AGE             120.9   1  8.1359  0.004718 **
ADIPOSITY       132.1   1  8.8902  0.003162 **
CHEST            28.0   1  1.8824  0.171340
ABDOMEN        1438.5   1 96.7976 < 2.2e-16 ***
WRIST           365.8   1 24.6131 1.328e-06 ***
AGE:WRIST       139.0   1  9.3526  0.002479 **
ADIPOSITY:CHEST  87.4   1  5.8795  0.016057 *
Residuals      3566.6 240
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figure 1: Anova table

The adjusted R^2 is 0.73, which means the proportion of variation in bodyfat explained by the regression relation is 0.73.

5 Model Diagnostics YW

After model fitting, we performed diagnostics for model assumptions with residual plots and a QQ plot. From Figure 2, normal QQ plot showed the normality is plausible as points were distributed closely to the qq line, and residual plot showed the linearity and homoskedasticity are also reasonable, since points look randomly scattered and no obvious non-linear trends were detected.

Then we'd like to check outliers in Y, high leverage points and influential points after we got the final model. Figures! There's no outliers in Y, but several high leverage points: 252,242,205,241 and so on.

There seems to not exist any influential points based on three detection criteria DFFITS, Cook distance and DFBETAS.

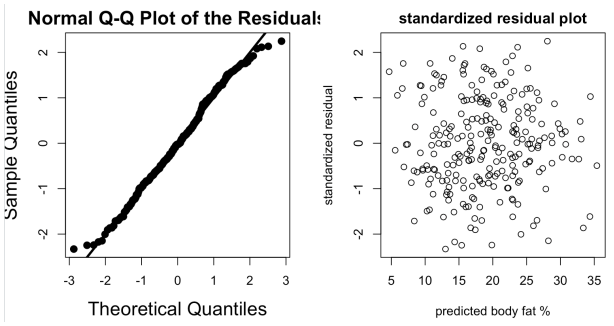

Figure 2: QQ plot & residual plot

6 Model Strengths and Weaknesses JL

1. Stability to abnormal data inputs. We carefully striped abnormal points when constructing model, making it minimally effected by those points.
2. Accuracy of body fat prediction. Our model is constructed by regression on Body fat itself instead of middle variable, which guarantees that the variances and errors won't be multiplied in the process.
3. Simple and interpretable. Linear regression is a simple format for prediction and the inner-relationship between variables and predictors are clear and easy to understand and interpret.
4. Small sample size and limited data resources. Our sample size is only several hundred and basically all from males, which is not suitable for larger range of people.

7 Conclusion and Discussion JL

Overall, we have proposed a creative and useful method for body fat calculation. It would be better if we gather more dynamic data to check the ability of our model. Hope every one a fit body!

JL RY revised report **RY** made slides