STAT 628 Data Science Practicum HW2

Zhihao Zhao, Shikun Liu, Yuxiao Li

Motivation: The report concerns on determining body fat percentage of male based on collected body data in a simple and accurate way. Since "IDNO" has nothing to do with body fat and "DENSITY" is only used to calculate "BODYFAT" via Siri's equation, these two variables are removed from our model construction.

Data Cleaning: The mean value of our goal variable "body fat percentage for males" is 18.94 and the standard deviation is 7.75. There are some abnormal data in the data set that drew our attention, No.39 reaches the maximum of measurement in most variables so we categorize him as an outlier and remove him. There are two individuals with ankle circumference >30cm, which is far away from most individuals but since other indexes are quite reasonable, we decide to keep them. The height of No.42 is 29.5 inches, which is much lower than any other individual so we use the bmi formula: $BMI = Weight(kg)/Height^2(m^2)$ after converting Weight and Height into kilograms and meters. The calculated result is 69.5 inches, which is plausible, so we assume 29.5 is a typo and fix it with 69.5. As for percentage of body fat, it is suggested that the value is less likely to be under 2 so we use Siri's equation to test if data under 2 are typo but unfortunately they are not. So No.172 and No.182 are viewed as outliers and dropped as well.

Model Selection: Our final model is

$$Body\ fat(\%) = -9.04 + 0.07 * AGE(years) + 0.71 * Abdomen(cm) - 2.24 * Wrist(cm)$$

which means a 40 years old man with 90cm abdomen circumference and 18cm wrist circumference is expected to have a body fat percentage of 17.41% based on our model and his 95% prediction interval is between 9.55% and 25.28%.

The parameter estimate 0.07 for age in unit of years represents that the body fat percentage will increase, on average, by 0.07 for every 1 year increase in age. The estimated coefficient for wrist circumference is -2.24 in unit of centimeter meaning that as men's waist increases by one inch, he is expected to loss 2.24% in body fat. As men's abdomen increases by 1 cm, he is expected to gain 0.71% in body fat.

We choose this model since all variables in it show significance in full model under $\alpha = 0.05$. Besides, the R^2 for this model is 0.726, while the full model is just 0.748 in comparison, not a big improvement but a lot more measurements to be done. Third, we also use step-wise regression to construct a new model that $Bodyfat \sim Age + Abdomen + Wrist + Height + Chest + Forearm + Neck$, but F-statistic for the comparison of our selected model and this one is greater than 0.05,which means that our simpler model holds and that we have 95% confidence that there is no difference between these two.

Statistical Analysis: Suppose that age is linearly unrelated to body fat percentage, i.e. $H_0 : \beta_{age} = 0$, and the alternative is that it isn't, i.e. $H_0 : \beta_{age} \neq 0$. The associated p-value for t-statistic is $6.7 * 10^{-4}$, which is much less than 0.05, so we conclude that there is a linear relationship between age and body fat percentage. However, the conclusion may carry 5% error rate where we may have falsely declared that there is a relationship even if there actually isn't a relationship. Correspondingly, the 95% confidence interval for age coefficient is $(0.031, 0.112)$. In other words, we assure 95% that the interval $(0.031, 0.112)$

contains the true age coefficient value β_{age}. The conclusion for variables Abdomen and Wrist can be drawn similarly.

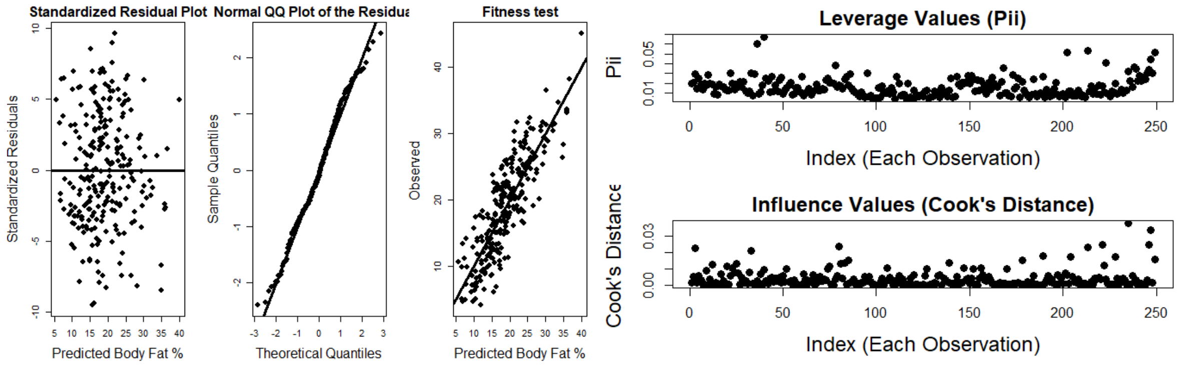
The R^2 for our model is 0.726, which means that it's able to explain 72.6% of variation in body fat percentage.

Model Diagnosis: We checked **lack of fit**, **linearity** and **homoscedasticity** in the left plot of Fig 1(a) and the conclusion is there is no evidence of lack of fit since we don't detect any systematic pattern. As for linearity, it seems be fine since there are no obvious non-linear trends like a curve or quadratic tape. Homoscedasticity is plausible since there is just a slight horn-shape pattern open to the right.

Normality is checked in the middle part of Fig 1(a) and we conclude that it's tolerable since there is only slightly trend of light-tail distribution, not a severe one.

We also check **goodness of fit** in the right panel of Fig 1(a), the predicted values versus observed values are close to the line y=x, which means the predictions are very close to the true values since for perfect prediction, Pred = True exactly, i.e. y=x. So the goodness of fit is also fine.

We check **leverage values** and **influence values** in Fig 1(b). There might be two leverage points at around the 40th-ish observation but no influential points.



(a) Residual plot, QQ plot, Fitness plot

(b) Leverage values and Influence values

Model Strengths and Weaknesses: We agree the final model is reasonable and can explain most of variation of body fat percentage, despite some caveats. The homoscedasticity and normality are not very satisfying, so we suggest to transform the data via Box-Cox Transformation then fit another model. But such model is more sophisticated and hard to explain for people with little statistical training and of course, hard to calculate. Besides, we view the whole data as one big group, but in fact, we should split people into more groups and utilize piece-wise regression for a better prediction.

Conclusion: Our model provides a simple way of evaluating body fat percentage via age, abdomen and wrist. The rate of metabolism will slow down as people getting aged so older people are more likely to gain weight. Abdomen circumference can reflect the quantity of visceral fat which is found in the abdomen amongst the major organs. Wrist circumference, however, may reflect the quantity of muscle, and of course, the higher level of muscle means lower level of body fat. When interpreting the model, linearity and goodness of fit are reasonable while normality and homoscedasticity may be slightly violated.

Contributions: Yuxiao wrote the app; Shikun wrote for presentation; Zhihao wrote the model code and wrote the report. We all together review and edit app, ppt as well as report. Overall, we met 4 times online and make countless conversations.