STAT 628 Module 2 Summary

Group members: RUI HUANG, CHENYANG JIANG, HanGyu KANG, ENZE WANG

Introduction and Background

Body fat is the percentage of fat mass in total mass, which is an important index for peoples' health. In our project, we will use 252 men observation with 16 variables to get the model to estimate body fat conveniently instead of underwater submersion. Welcome to our main code on GitHub: `https://github.com/moslandwez/Module_2`

Data Clean

From box plots and histograms, observations have a normal range in every variable unless it's from extreme obesity and thinness. The first step to detect outliers is from histograms and box plots. We decide to remove ID 182 because its initial BODYFAT is 0 and the new one from Siri equation is -3.6. Other extreme obesity outliers such as ID 39, we decide to keep them because there are only 252 observations and obesity becomes more common in modern society. Meanwhile, there exists a connection between BODYFAT and DENSITY which is called Siri equation and among ADIPOSITY, WEIGHT and HEIGHT which is called BMI equation. By linear regression or comparing new outputs with initial BODYFAT and ADIPOSITY, ID 96, 76 and 48 are outliers, we fix ID 76 and 48's BODYFAT with 14.09 and 14.13 and decide to keep ID 96 initial BODYFAT for the new one is too small. For ID 163 and 221, outliers in BMI equation, we fix their ADIPOSITY(BMI) with BMI equation by 27.4 and 21.68.

Model Build

By simple linear regression of all variables, although a lot of variables are not significant and serious multicollinearity, the model has good R-squared, therefore we decide to use linear regression model which is simple and easily explained and fixed for the unprofessional. Our model build including two steps, variables and model selection.

In variables selection, we use tree regression, subsets method and forward, backward and both direction search with AIC, BIC and other index. We design four full models, which is full model with all raw variables, with log, with square transformation and that including all variables before to find any improvement room. The results of variables selection show that ABDOMEN and its transformation is the most important, followed by WRIST and WEIGHT. In short, we found 19 alternative models. For which we use 30-repeated 10-fold cross-validation for models selection. In cross-validation, log and square transformation or adding variables from 3 to 4 or more adds only 0.01 in R-squared, which is not necessary to sacrifice simplicity for this level of precision. On the other hand, models with 2 or 3 variables have great improvement compared with 1 variable and models with ABDOMEN, WEIGHT, WRIST have the best performance. Compared with 3 variables one, we choose the 2 variables one because they have only 0.003 difference in R-squared. ID 39 is removed in our final model for its stand out in Cook's distance. Therefore our final model is simple linear regression between BODYFAT with ABDOMEN and WEIGHT with data except ID 182 as outliers and ID 39 as influential observation.

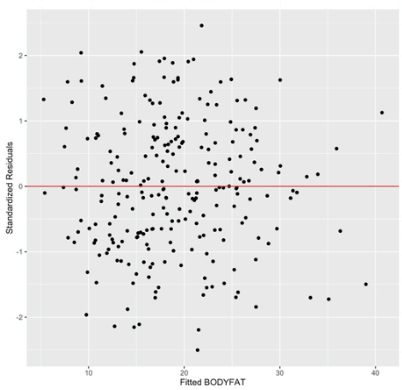
Coefficients	Estimate	Std. Error	p value	95% CI Lwr	95% CI Upr
(Intercept)	-42.26886	2.44647	$< 2e-16$	-47.08746	-37.45026
ABDOMEN	0.89944	0.05158	$< 2e-16$	0.79783	1.00103
WEIGHT	-0.12270	0.01948	1.37e-09	-0.16108	-0.08433
Standard error:	4.041	R-squared:	0.7197	Adjusted R-squared:	0.7174

Rule of thumb: BODYFAT(%)=abdomen circumference(cm)*0.899 minus weight(lbs)*0.123 minus 42.3. For example, a man with 154 lbs weight, 85 cm abdomen, his body fat percentage will be 15%, with 95% CI is from 7.45% to 23.42%. For every abdomen circumference increase in 1 cm, body fat(%) will increase average 0.899%; every weight increase in 1 lbs, body fat(%) will decrease average 0.123%. All three coefficients are significant. For test H0: coefficient of ABDOMEN (WEIGHT) is 0, H1: coefficient of ABDOMEN (WEIGHT) is not 0, for the p- value is smaller than 0.05, we reject H0 at 0.05 level and there exist clear linear relationships. This model is simple with only two variables, whose R-squared is 0.72, which can explain about 72% variation in body fat. The residual standard error is 4.04, which is small. In medicine, our model shows that with weight holds, the increase of abdomen circumference reflect more body fat accumulates in your abdomen and body fat percentage increases. With abdomen circumference holds, the increase of weight reflect you becomes stronger instead of fatter, your body fat percentage may decrease.

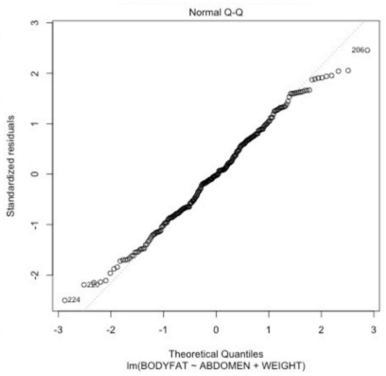
Model Diagnosis

The following are standardized residual, residual QQ and Cook's distance plots, there are no obvious outliers and special pattern in standardized residuals plot and qq plot meets 45-degree line except few outliers. There are no outliers in Cooks distance because ID 39 is removed. Therefore our model meets linearity, normality and constant variance assumptions.

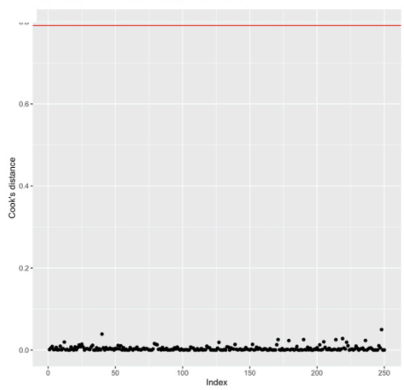
Standardized Residual Plot



Normal Q-Q Plot of the Residuals



Cook's Distance Plot



We also plots DFFITS, DFBETAS plots in GitHub to find influential observation, there is no obvious influential observation and our model performs well.

Summary, Advantages and Disadvantages

Our model includes ABDOMEN and WEIGHT to estimate body fat, performs well in diagnosis, there is no serious multicollinearity (mean VIF=4) and all linear regression assumptions hold. The rule is simple with two inputs with small residual standard error. But we sacrifice precision for simplicity and we still think that there is room for improvement in R-squared with further study with larger and detailed data and advanced methods.

Acknowledgement

RUI HUANG, edits model selection part in summary and presentation.
CHENYANG JIANG, edits model build part in summary and presentation
HanGyu KANG, edits tree regression part in summary, main editor of video.
ENZE WANG, main editor of summary and presentation file.