STAT 628 Module 2 Summary

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

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 has a normal range in every variable unless it's from extreme obesity and thinness. The first steps to detect outliers is from histograms and box plots. We decide to remove ID 182 because its initial BODYFAT is 0 and new one from Siri equation is -3.6. Meanwhile, some extreme obesity outliers such as ID 39, we decide to keep them because there is only 252 observations and obesity becomes more common in modern society. Meanwhile, there exist 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 for the unprofessional. Some model such as neural networks is too complex and difficult to rebuild. Our model build including two steps, variables and models selection.

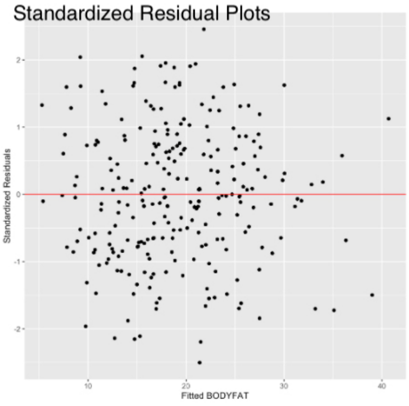
In variables selection, we use tree regression, subsets method and forward, backward and stepwise direction search with AIC, BIC and other index. We design four full models, which is full model with all raw variables, with log, with square transformation and that including all variables before to find any improvement room. Our rule of model is simple and precise in case of multicollinearity and overfitting. 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, we decide to use ABDOMEN, WEIGHT as our final model predictors. In cross validation, transformation and adding variables from 3 to 4 or more adds only 0.01 in Rsquared, 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 and our final model have the best performance. Where we choose the 2 variables one because they have almost no difference. ID 39 is removed in our final model for its stand out in Cook's distance.

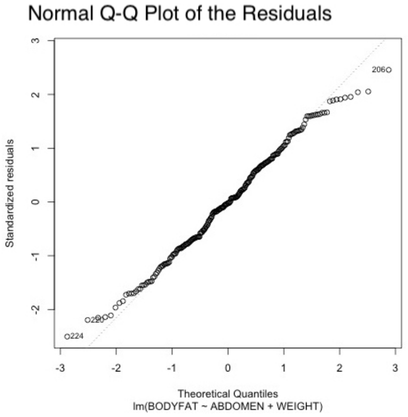
Coefficients	Estimate	Std. Error	p value	95% CI Lwr	95% CI Upr
(Intercept)	-42.26886	2.44647	$< 2e-16$	-47.08746	-37.45026
ABDOMEN	0.89944	0.05158	$< 2e-16$	0.79783	1.00103
WEIGHT	-0.12270	0.01948	1.37e-09	-0.16108	-0.08433
Standard error:	4.041	R-squared:	0.7197	Adjusted R-squared:	0.7174

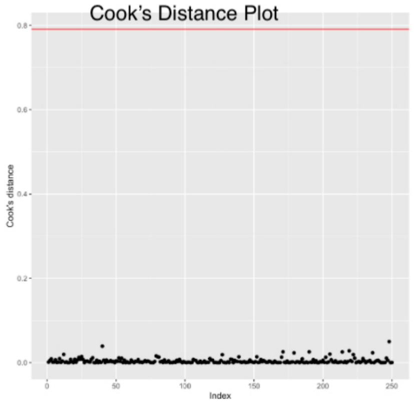
Rule of thumb: BODYFAT(%)=abdomen circumference(cm)*0.899 minus weight(lbs)*0.123 minus 42.3. For example, man with 154 lbs weight, 85 cm abdomen, his body fat percentage will be 15%, his 95%CI is from 7.45% to 23.42%. For every abdomen increase in 1 cm, body fat(%) will increase mean by 0.899%; every weight increase in 1 lbs, body fat(%) will decrease mean by 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 and there exist clear linear relationships. This model is simple with only two variables, whose Rsquared 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 show that with weight holds, the increase of abdomen circumference reflect more body fat accumulate 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 decreases.

Model Diagnosis

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







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 of improvement in Rsquared 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.