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 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

All variables are positive, and most of them are uni modal in histogram. 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 found 5 outliers, ID 182 has 0 body fat and ID 216, 39, 41 has too many extreme values, we decide to remove them. ID 42 has extreme HEIGHT but can be fixed by BMI equation by 69 inches. Meanwhile, there exist connection between BODYFAT and DENSITY which is called Siri equation and among ADIPOSITY, WEIGHT and HEIGHT which is called BMI equation. In linear regression between BODYFAT and 1/DENSITY, 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 his new BODYFAT is too small. In linear regression among ADIPOSITY, WEIGHT and HEIGHT, ID 163 and 221 are outliers, 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 clear. Our model build including two steps, variables selection and models selection. Our groups also do PCA and factor analysis to find any improvement space for linear regression on new variables, but the results is bad because PCA and factor analysis consider the relationship among predictor variables instead of that between BODYFAT.

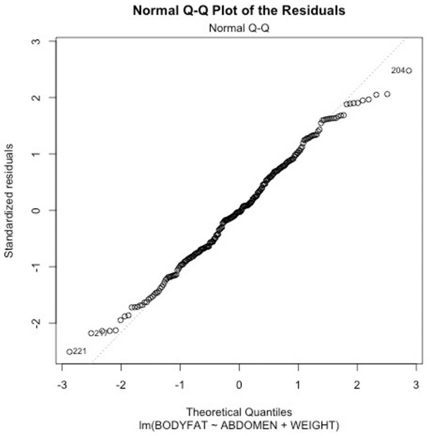
In variables selection, we use Lasso regression, subsets method and forward Direction Search with AIC, BIC and other index. We design four full model, which is full model with all raw variables, with log, with square transformation and that including all variables before to find any improvement room. Our rule of model is simple and precise in case of multicollinearity. The results of variables selection show that ABDOMEN and its transformation is the most important, followed by WRIST and WEIGHT. In short, we found 12 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. The following is results of our model, which performs well.

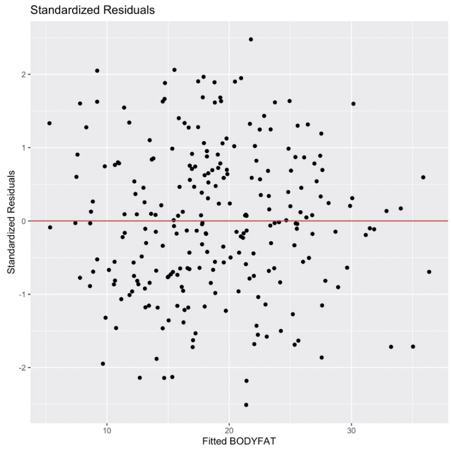
Coefficients	Estimate	Std. Error	p value	CI Lwr	CI Upr
(Intercept)	-42.39091	2.54235	$< 2e-16$	-47.3985640	-37.3832595
ABDOMEN	0.89355	0.05280	$< 2e-16$	0.7895427	0.9975474
WEIGHT	-0.11893	0.01962	5.07e-09	-0.1575816	-0.0802828
Standard error:	4.03	R-squared:	0.7053	Adjusted R-squared:	0.7029

Rule of thumb: BODYFAT(%)=abdomen circumference(cm)*0.89 minus weight(lbs)*0.12 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 7.42% to 23.36%. For every abdomen increase in 1 cm, body fat(%) will increase mean by 0.89%; every weight increase in 1 lbs, body fat(%) will decrease mean by -0.112%. All three coefficients are significant and exist clear linear relationships. This model is simple with only two variables compared with another alternatives, Although from ANOVA results of WRIST, which is significant, our groups hold that it is unnecessary to add WRIST into our model with only 0.01 increase in R squared.

Model Diagnosis

We plot qq plot for residual to check normality assumption and standardized residuals plot to check constant variance, linearity assumption and any outliers. The following are the plots, there is no outliers and both constant variance, linearity assumption hold in standardized residuals plot and normality holds in qq plot except few outliers. But there is a lot of outliers in Leverage plots, considering there is almost no change in new model without these outliers, we still keep the initial model.





We plots Cook's distance, DFFITS and DFBETAS plots to find influential observation, there is no obvious influential observation and our model perform well. More pictures can be viewed in our GitHub.

Summary, Advantages and Disadvantages

Our model performs well in diagnosis, there is no serious multicollinearity and all linear regression assumptions hold. The rule is simple with two inputs with small residual standard error. But we still think that there is room of improvement in R squared with further study with larger and detailed data and advanced methods.

Acknowledgement

RUI HUANG, main editor of shiny app, model selection part in model, fix presentation file and summary, attend in presentation.
CHENYANG JIANG, main editor of model build, fix presentation file and summary, attend in presentation.
HanGyu KANG, main editor of PCA, FA part in model and presentation video, attend in presentation.
ENZE WANG, main editor of main code, GitHub, summary and presentation file, fix shiny app, attend in presentation.