**Model 2 Project**

Group member: RUI HUANG; CHENYANG JIANG; HanGyu KANG; ENZE WANG

Main edited by ENZE WANG; xxxxxxxxxxxxxxxxxxxxxxx

**Background**

Bodyfat is the percentage of fat mass in total mass, which is an important index for peoples’ health. But it is not an easy task to estimate bodyfat, by Siri equation, if you want to get the accurate bodyfat you will use underwater submersion way, which is difficult. In our project, we will use 252 observation with 16 variables to get the model to estimate bodyfat with other variables instead of underwater submersion.

**Data Clean**

The first step is data cleaning, here we use three methods: boxplot, Siri equation and BMI equation. In boxplot, we found ID 182 has 0 bodyfat and ID 216, 39, 41 has too many extreme values, we decide to remove them. ID 42 has short height but can be fixed by BMI equation. By Siri equation, Bodyfat=495/Density-450, ID 96, 76 and 48 are outliers, we fix their bodyfat or density with Siri equation. By BMI equation, BMI=703Weight/Heightˆ2, ID 221 and 163 are outliers, we fix their ADIPOSITY(BMI) with BMI equation.

**Variables Selection**

By simple linear regression, there are a lot of variables are not significant, so we use 3 ways to select variables. First, we try use Lasso regression to also deal with multicollinearity, which final model is BODYFAT ~ AGE + HEIGHT + NECK + ABDOMEN + BICEPS + FOREARM + WRIST + THIGH, which is very complex. Next we use subsets method with BIC, Cp and other index. Although by box-cox analysis there is unnecessary for transformation, we still include log and square transformations in our model. Here we get several alternative models and most of them include ABDOMEN and WEIGHT and their transformation. The last way is forward direction search of AIC or BIC. After variables we select 11 alternative models.

**Models Selection**

In 11 alternative models, we use repeated 10-fold cross validation to find the best model. It is not a good idea to use cross validation without repetition because there are about 248 observation and the output will vibrate with test and train data sample. From the part results of cross validation:

|  |  |  |
| --- | --- | --- |
| Model | Root Mean Squared Error | R-squared |
| BODYFAT ~ ABDOMEN + sqWEIGHT + LogWRIST | 3.939160 | 0.7254679 |
| BODYFAT ~ LogABDOMEN + LogWRIST + LogHEIGHT | 3.941639 | 0.7239889 |
| BODYFAT ~ ABDOMEN + WEIGHT + WRIST | 3.962274 | 0.7217524 |

We decide to use BODYFAT ~ ABDOMEN + WEIGHT + WRIST as our final model, it seems that there is no necessary to transformation.

**Model Diagnosis**

We use QQplot for residual, standardized residuals plot, Leverage plot, DIFFIT plot, Cook’s distance plot, and DEBETAS plot for model diagnosis. From the QQ plot for residual the normality assumption is satisﬁed. From standardized residuals plot the equal-variance and linearity assumption are satisﬁed. Leverage plot measure of the distance between each X and mean of X. Cook’s distance measure the influence of ith observation on all fitted values. DFFITS measures the influence of the ith observation on the fitted value. The DFBETAS measures the influence of ith observation on the fit of the regression coefficient. There are no obvious outliers from all these plot above.

Chart, line chart

Description automatically generatedChart, scatter chart

Description automatically generated

**Conclusion**

Our final model is:

BODYFAT=0.87792 ABDOMEN-0.08384 WEIGHT-1.26116 WRIST -24.22925

whose R-square is 0.7168, our model will explain about 72% of all variation, residual standard error is 3.958. All parameters are significant in 0.05 level. The rule of thumb is the sum of 0.87792 times ABDOMEN, - 0.08384 times WEIGHT, -1.26116 times WRIST and -24.22925. From our model we can see that, by increasing one unit in ABDOMEN, BODYFAT will increase 0.87792%, by increasing one unit in WEIGHT, BODYFAT will decrease 0.08384%, by increasing one unit in WRIST, BODYFAT will decrease 1.26116%. Our model has high R square and good performance in diagnosis, but there is three inputs which are a little more.