Logistic Regression Analysis of Pima diabetes Data

Ruoyuan Li (001223313)

Data Description & Concenred Problem

My data set are located at library **faraway** called **pima**, it is diabetes survey on pima Indians.Original data contain 10 variables with binary response variables which indicated liver disease patients or non liver disease patients, and the explanatory variables are all numerical. To make it simpler, therefore in this logistic regression analysis case, only consider 5 out 9 explanatory variable, and also I made the number of pregnant times into 3 groups since there were no categorical variables in the explanatory variables, **Low** referred to pregnant [0,2] times, **Middle** referred to pregnant (2,5] times, **High** referred to pregnant (5,17] times. The detail of the data I used as following:

- $pregnant_f$: explanatory variables, a factor with 3 levels, Low, Middle, and High;
- *age*: explanatory variables, a numerical variables ranged from 21 years old to 81 years old;
- *bmi* (body mass index): explanatory variables, a numerical variables ranged from 0 to 67.1;
- *glucose*: Plasma glucose concentration at 2 hours in an oral glucose tolerance test. explanatory variables, a numerical variables ranged from 0 to 199;
- *triceps*: riceps skin fold thickness (mm). explanatory variables, a numerical variables ranged from 0 to 99;
- *test*: response variables, a factor with 2 level, 0 if negative, 1 if positive;

The question concerned is whether those 5 independent explanatory variables that determine outcome which is "test" in this case, the goal is find the best fitting model to describe their relationship.

Data Analysis with Logistic Regression

First fitting the full model with all the variables together with specify family=binomial("logit"). Then based on summary of full model can see triceps with $p - value$ equals 0.15 which means not significant statistically, and calculating $p - value$ from chi-square distribution based on information from Null deviancec and Residual Deviance ends with 5.149981e-28, which indicates at least one of the "slopes" is significantly different from zero.

Then continuing fitting the model, drop **triceps** and keep all the rest to be same. So the formula used in reduced model is: $test \sim pregnant_f + bmi + age + insulin + triceps$. To compare full model and reduced model, extracting Residual Deviance from reduced model and calculating $p - value$ from chi-square distribution ends with 0.1502215, which indicates reduced model as good as full model, accepts null hypothesis. All other variables are statistically significant with $p - value$ much lower than significant level 0.05, so model fitting stops here. In summary of logsitic regression, the coefficients are log-odds, since we want to focus on odds ration, then we uses *exp* function to switch log-odds coefficients from reduced model to odds ratio.

The output for coefficients after exponential transformation are, $pregnant_f$middle with 1.684713003, $pregnant_f$high with 2.427772749, bmi with 1.101832541, age with 1.031599693 , and insulin with 1.001841596; those can be intrepreted as following:

- having number of pregnant times with **Middle**, versus pregnant times with **Low**, the odds of getting positive result increased by a factor of 1.685; having number of pregnant times with **High**, versus pregnant times with **Low**, the odds of getting positive result increased by a factor of 2.428.
- For every 1 unit increased in bmi, the odds of getting positive results(versus negative) increases by a factor of 1.102; similar interpret ion for pregnant times with age and insulin.

Then divides data set into train set which contain 80% of the total observations and test set, next prediction based of reduced model fitted before with new data to be test set. To make comparison of prediction and actual outcome, first use *table* to build a contingency table of counts at each of factor levels, following with $e1071 :: classAgreement$ function ends with several coefficients. Percentage of data points in the main diagonal of tab is 0.662, rand index is 0.549. The performance for the fitting model does not perform very well, but with limited variables and information, it is an acceptable results.

Finally, after those logistic regression analysis, the number of pregnant times, ages, bmi, and insulin are variables which determine the outcome. Model performance works acceptable.