Review of Logistic Regression Analysis

- It is a regression model only suitable for a response variable that is measured in two categories, yes (success) and no (fail).
- In data, measurements of the response variable may appear in binary, i.e. 0 or 1, or appear in counts from a set of integers, say $1, \ldots, 5$. Therefore they are termed as dichotomous and group data, respectively.
- In the context of regression analysis, a set of explanatory variables are also observed together with Y.
- A logistic model consists of two components:

  (1) Random component: refers to the sampling distribution, namely the sampling distribution – Y follows a binomial distribution $B(n, \pi)$, where π is the probability of success.

  (2) Systematic component: refers to the model that specifies the relationship between the probability of success and a set of explanatory variables:

$$Logit(\pi) = b_0 + b_1 X_1 + \cdots + b_k X_k$$

- The key concept is the logarithm of odds of success, also called logit. Here the logit is modeled as a linear function of explanatory variables.
- As a result, the odds ratio is then modeled as a function of explanatory variables. For example, at two different levels of $X_1 = a$ or b, while the other variables $X_2, \ldots, X_k$ are withheld, the log odd ratio is $b_1(a-b)$, so the odds ratio is $e^{(b_1(a-b)}$. It is known that odds ratio reflects the strength of association between the Y and the explanatory variable X_1 of interest. Similar arguments apply for each of the other explanatory variables.
- Modeling odds ratios as functions of explanatory variables is particularly useful when explanatory variables are categorical. In this situation, the linear correlation coefficient is not suitable to depict the association between the categorical Y and categorical X. In other words, the regression coefficient in the logistic regression model has no interpretation of the linear correlation, but effectively is interpreted in the language of odds ratio, a measure of association between two categorical e variables.
- So, when the logistic model is applied to analyzed categorical explanatory variables such as contingency table data, interpreting results of the modeling in the language of odds ratio is especially important, because that is indeed the goal of the data analysis.
- In contrast to the logistic model-based analysis, for contingency table data, there is a set of well-established tools to carry out non-model-based analysis, including Pearson X^2 or likelihood ratio G^2 test for the independence. One major disadvantage of non-model-based analysis is it is awkward in the case of many explanatory variables. We have seen that even in the case of 3-way tables, extending the odds ratio association becomes complicated. We have to deal with conditional analysis (CMH test for conditional independence, Breslow-Day test for the homogeneity of odds ratio and MH estimate of the common odds ratio).

- In contrast, the model-based analysis appears to be naturally extended to deal with multiple explanatory variables. We do not need to develop new concepts when the number of explanatory variables increases. In addition, most of concepts established in the classical linear regression analysis can be re-developed in the context of the logistic regression model. For example,

Type of Analysis	Linear Regression	Logistic Regression
Goodness of fit	R^2	Deviance Γ
Significance of model	F-test	Likelihood ratio chi-squared test
Testing individual significance	t-test	Wald chi-squared test
Model selection	Stepwise	Stepwise
Influential analysis	Outliers/influential cases	Outliers/influential cases
Residual analysis	Normality	Overdispersion

- Extension of the logistic regression model to handle multi-categorical response variable.

  (1) Baseline-category logistic regression model: assigning one category (SAS default the last category) as the reference category and then run logistic regression for each of the other categories with the reference category. As a result the probability for each category can be modeled as a function of explanatory variables.

  (2) Proportional odds model: utilize the natural ordering among the categories of an ordinal response variable, and collapse the categories in the fashion of cumulative probabilities. Logistic regression models are then built on the cumulative logits. As a result, the probability for each category can be modeled as a function of explanatory variables.

$$\begin{aligned} Prob(Marked) &= \gamma_{1|Trt=i,Sex=k} \\ Prob(Some) &= \gamma_{2|Trt=i,Sex=k} - \gamma_{1|Trt=i,Sex=k} \\ Prob(None) &= 1 - \gamma_{2|Trt=i,Sex=k}. \end{aligned}$$