Logistic Regression

1 Introduction

The relationship between a dichotomous variable and a set of risk factors becomes more difficult to explore as the number of factors to be investigated becomes larger. Risk models allow the use of the variables without the necessity of categorizing variables.

One immediate restriction of the traditional risk models is that the outcome must be dichotomous, i.e., death or survival, disease or no disease. [Generalizations are possible, although we will not discuss them] Also, the investigator must keep in mind that any model imposes a structure of its own upon the possible results, and that if the model is incorrect bias may result. In addition interpretation of the results based on risk models may be less intuitively clear than SLR models.

The logit linear (logistic model)

$$log\left[\frac{p(x)}{1-p(x)}\right] = \beta_0 + \beta_1 X \quad or \quad p(x) = \frac{exp(\beta_0 + \beta_1 x)}{1 + exp(\beta_0 + \beta_1 x)} \tag{1}$$

where $\frac{p(x)}{1-p(x)}$ is the odds.

Note that for a dichotomous independent variable:

$$log\left[\frac{p(0)}{1-p(0)}\right] = \beta_0 \quad and \quad log\left[\frac{p(1)}{1-p(1)}\right] = \beta_0 + \beta_1 \tag{2}$$

So this means that:

$$\beta_1 = log\left[\frac{p(1)}{1-p(1)}\right] - log\left[\frac{p(0)}{1-p(0)}\right] \tag{3}$$

Raising both sides to the exponential power we get that (with OR=Odds Ratio):

$$e^{\beta_1} = \frac{\frac{p(1)}{1-p(1)}}{\frac{p(0)}{1-p(0)}} = OR \quad or \quad \beta_1 = log(OR) \tag{4}$$

If we have a continuous independent variable we get that $e^{\beta_1 \Delta}$ = Odds ratio per Δ units increase in X, regardless of the baseline value for X. Since:

$$OR = \frac{\frac{p(x+\Delta)}{1-p(x+\Delta)}}{\frac{p(x)}{1-p(x)}} = e^{\beta_1 \Delta} = e^{\beta_1 (X^* - X)} \tag{5}$$

Where $\Delta = X^* - X$ (i.e. two values of the independent variable that we are interested in comparing).

The odds ratio (approximately the relative risk) of X^* compared to reference level X depends only on the distance Δ and β_1 which measures the strength of the association between the risk factor and the distance.

Relative risk (RR=$\frac{p_{exposed}}{p_{control}}$)- Risk of an event relative to exposure.

Example: If the probability of developing cancer is 20% for smoders and 10% for non-smokers the relative risk of cancer associated with smoking is 2.

Here are some Model Options you can specify the following options after a slash (/):

1. CLODDS=(PL — WALD — BOTH)
2. CLPARM=(PL — WALD — BOTH)
3. LACKFIT
4. LINK=keyword
5. RSQUARE

Lets do an example:

2 Interaction Models

Interaction will be done similarly to MLR. In the logistic model, interaction can be represented by the product of the risk factor and another adjustment variable. In deciding whether to use an interaction term, remember that the logistic model already specifies a multiplicative relationship between risk factors. Here is the model we'll be working with:

$$log\left[\frac{p(x)}{1-p(x)}\right] = \beta_0+\beta_1 X_1+\beta_2 X_2+\beta_3 X_1 X_2 \quad or \quad p(x) = \frac{exp(\beta_0 + X_1 + \beta_2 X_2 + \beta_3 X_1 X_2)}{1 + exp(\beta_0 + X_1 + \beta_2 X_2 + \beta_3 X_1 X_2)} \tag{6}$$

Here will have for a Δ change in X_1, and a given value of X_2 that:

$$OR = exp\{\beta_1 \Delta + \beta_3 \Delta X_2\} \tag{7}$$