Logistic Regression

(gustavoc@msu.edu)

Binary Outcomes

Many outcomes of interest are binary, implying that they can take two values (say, 0/1). Disease is a typical example of this. Binary random variables follow Bernoulli distributions: $p(Y_i = 1) = \theta^{Y_i}(1-\theta)^{1-Y_i}$ or $p(Y_i = 1) = \theta \;\; ; p(Y_i = 0) = 1 - \theta$.

Logistic Regression

We are often interested on learning the effects of some factors (e.g., sex) and covariates (e.g., age) on the probability of a binary outcome (e.g., disease). In logistic regression, we make θ a function of covariates. Since $\theta \in [0,1]$ we cannot model θ directly using linear regression because a linear function can take any value in the real line. To deal with this problem we introduce a "link" function (e.g., probit, logit). A link function maps from the real line onto the $[0,1]$. The most commonly used link is the logit which is the logarithm of the odds of success, that is: $log\left(\frac{\theta_i}{1-\theta_i}\right)$. This function can take values in the real line, thus, we can model the logit using linear methods

$$\log\left(\frac{\theta_i}{1-\theta_i}\right) = \mu + X_{i1}\beta_1 + \cdots + X_{ip}\beta_p. \qquad [1]$$

Note that the above regression is a regression for the probability, not for the data, thus, it typically does not include an error term (in some over-dispersed models it may contain an error).

From regression to probabilities

Solving [1] for θ_i gives

$$\theta_i = \frac{\exp\{\mu + X_{i1}\beta_1 + \cdots + X_{ip}\beta_p\}}{1 + \exp\{\mu + X_{i1}\beta_1 + \cdots + X_{ip}\beta_p\}}. \qquad [2]$$

Odds and Odds Ratios

The odds of success is the ratio between the success and failure probabilities, that is $\frac{\theta_i}{1-\theta_i}$. The odds ratio (OR) is the ratio between the odds of two groups. Suppose we have θ_F and θ_M representing the success probabilities for male and female, then, the female:male odds ratio is $OR\left(\frac{F}{M}\right) = \frac{\frac{\theta_F}{1-\theta_F}}{\frac{\theta_M}{1-\theta_M}}$.

From regression coefficients to odds-ratio

Suppose our logistic regression takes the form

$$\log\left(\frac{\theta_i}{1-\theta_i}\right) = \mu + F_i\beta_1 + X_i\beta_2 \qquad [3]$$

where $F_i = \{1\ if\ female; 0\ otherwise\}$ is a female dummy variable and X_i is a covariate of interest (say age). Using [2] and [3] we have that the success probabilities for male and female are

$$\log\left(\frac{\theta_i}{1-\theta_i}|female\right) = \mu + \beta_1 + X_i\beta_2 \text{ and } \log\left(\frac{\theta_i}{1-\theta_i}|male\right) = \mu + X_i\beta_2$$

Thus, the logarithm of the odds ratio is

$$\log\left\{\frac{\frac{\theta_i}{1-\theta_i}\ |female}{\frac{\theta_i}{1-\theta_i}\ |male}\right\} = \log\left(\frac{\theta_i}{1-\theta_i}|female\right) - \log\left(\frac{\theta_i}{1-\theta_i}|male\right) = \beta_1$$

Then, the odds ratio becomes

$$OR\left(\frac{F}{M}\right) = Exp\{\beta_1\} \qquad [4]$$

A nice property of the odds ratios is that they do not depend on the value that other covariates (age in our example take).

Relative Risk (RR)

A perhaps more intuitive metric is the relative risk between two groups, that is: the ratio of the probability of developing disease between the two groups (e.g., female:male RR). Using [2] the female:male relative risk for the model we discuss above is

$$RR(female{:}\,male|X_i) = \frac{\frac{\exp\{\mu+\beta_1+X_i\beta_2\}}{1+\exp\{\mu+\beta_1+X_i\beta_2\}}}{\frac{\exp\{\mu+X_i\beta_2\}}{1+\exp\{\mu+X_i\beta_2\}}} \qquad [4]$$

Note that unlike OR, RRs depend on the values of other covariates.