Logistic Regression

March 14, 2005

Logistic regression is a technique for modeling 0/1 data – this means data for which the response variable is a 0 (individual does not have the condition) or 1 (individual does).

Although the data have 0 and 1 values, we're interested in the probability of being a 1. For example, given a certain blood-test value, what's the probability the patient has the condition?

Before, we modelled the mean function as

$$E(Y) = \beta_0 + \beta_1 x$$

Note that if Y is now a binomial random variable, then $E(Y) = p$, and so what we want to do is model the probabilities.

Note that the striaght-line model above won't work—for large values of x it will predict a probability bigger than 1, and for small values it will predict negative probabilities. For this reason, a tranformation (called the logit) transformation, works best:

$$\log \frac{p}{1-p} = \beta_0 + \beta_1 x$$

The LHS is called the "log odds ratio". An odds ratio is the probability of an event occurring over the probability it does not occur. So if this ratio is 0 (which means the log is 0) then the events are equally likely. A ratio ¿ 1 (log is positive) means the event is more likely to occur than not, and so negative log odds means it is less likely to occur than not.

There are some other differences. For example, instead of choosing the line that minimizes least sqaures, we choose the line that minimizes something called the deviance:

Today, we're going to show how to fit a logistic model and how to interpret the coefficients.

First, let's look at some graphical summaries. Of course you can make a plot of the predictor (x axis) against the 0/1 response variable. The problem is that it's hard to see any trend.

The lowess line is helpful.

Another approach is to "jitter" the response variable; add a small, random number so that we no longer have 0's and 1's, but 0's and 1's with a bit of noise added. Plotting the jittered data gives us a better undrestanding of the pattern.

Some care must be taken if the response variable is categorical. You should convert it to numerical first.

With the gender/height data we see the (not too surprising) fact that men seem to be taller than women, and that the probability of a randomly selected person being male increases fairly sharply with increasing height.

We can fit the model using the GLM feature of R. General Linear Models are beyond our ability to cover in two lectures (indeed, difficult to cover in a quarter), but suffice it to say that these are a class of models that include linear regression and also include logistic regression.

The model fitting techniques are slightly different. In linear regression, we choose the line that minimizes the RSS. In logistic regression we minimize the "deviance" G. The definition is a bit awkward.

For each value of x_i, let y_i be the number of 1's observed (the number of successes) and let m_i be the number of observations at that x_i (the number of trials). $\hat{y}_i$ is the number of successes the model predicts: $m_i\hat{p}_i$.

$$G^2 = 2\sum y_i \log(\frac{y_i}{\hat{y}_i}) + (m_i - y_i)\log(\frac{m_i - y_i)}{m_i - \hat{y}_i})$$

This measures the difference between the log liklihood obtained by fitting a the model with the log likelihood obtained by fitting a separate estimate of the probability for each value of x. We want deviance to be small.

We can use the deviance to compare models. So if we want to add a predictor (weight?) to the model, we would look to see if the difference between G^2 without weight minus G^2 with weight included, and see if this difference was large. Large differences would mean that the new variable was useful. To determine "large", we compare this to a chi-squared distribution whose degrees of freedom is equal to the difference in the number of parameters in the two models (which would be 1 in this example.)

For example, fitting the model using height to predict weight results in

$$\log\frac{\hat{p}}{1-\hat{p}} = -32.4 + .49x$$

Exponentiating both sides gives us that the odds-ratio is

$$e^{-32.4}e^{.49x}$$

This means that a one-inch difference in height is associated with an increase in the odds of $e^{.49} = 1.6$. A person one inch taller is 1.6 times more likely to be male than female.

The deviance is 101.8 with 106 degrees of freedom. Note that we also get the AIC, and this too can be used to compare models.