Logistic Regression:

Overview:

When we have a binary outcome variable, we often want to test whether there are any associations between outcome and certain features, such as patient race/ethnicity, treatment group, or expression levels of a specific gene. For example, if we are interested in looking at 1 year survival (yes or no) after treatment, we may hypothesize that patients treated with a novel therapy may have a higher probability of 1 year survival compared to patients treated with standard therapy. We can test for such bivariate associations using **chi-square or Fisher's exact tests** however this does not allow us to examine multiple predictors of outcome simultaneously in the same model. To do this, we need to use logistic regression.

What is a logistic regression model?

As usual, the form of the outcome variable (in this example, whether someone survives for at least a year or not after treatment) dictates the statistical model you will use. Linear regression is used when the outcome variable is continuous. Logistic regression is used specifically for outcome variables that are binary in nature.

The form of the model:

Suppose we want to model the outcome as a function of two predictors – patient age (in years) and treatment group (1=novel therapy, 0=standard therapy). The model equation would be:

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_{TRT}Treatment + \beta_{age}Age$$

If we look at the left side of the equation, we see we are not modeling the p directly, we instead are modeling the log odds of p. There are a few reasons for this, and if you want more detail you can read about it **here**. We are thus relating the log odds of p to a linear combination of an intercept plus the treatment the patient had multiplied by some weight, plus the age of the patient multiplied by some weight. Clearly if the weight for the treatment (β_{TRT}) is very large, treatment has a lot of influence over whether we predict a patient will survive for at least one year with high or low probability. And if the weight is close to zero, what treatment a patient has does not influence the prediction very much at all. Similarly, if β_{AGE} is large and positive, increasing age is going to increase the survival probability.

When we talk about *estimating* this model, this means that we are using our data and the values of treatment, age, and outcome (0/1) to obtain the estimates for β_0 (the intercept), β_{TRT} and β_{AGE} that maximize the likelihood of our observed data. Of course if we observe in our data set that most one-year survivors are on the novel therapy with few survivors on standard therapy, an estimate of zero for β_{TRT} would not be the most likely given our data. The statistical package that we use automatically computes these best estimates for us once we specify the predictors to be used in the model.

Because we are relating the log odds of p (left hand side of the equation) to the linear combination (right-hand side of the equation), the estimates we obtain are giving the effect of treatment and age on the log odds of p. We can transform these betas to give us the effect of treatment and age on the odds of surviving for one year by exponentiating each estimate. Exp(β_{TRT}) thus gives us the effect of the novel treatment (vs. standard) on the odds of surviving for at least one year.

For example, let's assume we obtain the following model estimates:

Treatment: $\beta_{TRT} = 0.21$ $\exp(\beta_{TRT}) = 1.23$

Age: $\beta_{AGE} = -0.03$ $\exp(\beta_{AGE}) = 0.97$

The way we interpret these estimates is that for treatment, being on the novel therapy increases the odds of surviving for at least one year by 23% compared to standard therapy. For patient age, patients who are one year older have decreased odds (0.97 times that of a patient one year younger) of surviving.

Of course, we cannot fully interpret these estimates without also computing 95% confidence intervals and p-values to see whether they are statistically significant. We will see a much more detailed example when we look at an **example data set** and output from the **R code** used to estimate a logistic regression model.

Where can I read more about logistic regression?

There are many **excellent online resources** that will allow you to more fully understand logistic regression.

How do I estimate a logistic regression model?

Here are some **basic instructions** for estimating and interpreting your own logistic regression model using the R software package .