**Stat 401 - lab 13**

Goals:

 How to fit logistic regression to yes/no data

We will use the Donner Party survival data discussed in lecture. This has a yes/no (actually, 1/0) response (did the individual survive where 1 = yes) and two potential predictor variables: age (in years) and fem (coded as 1 = female, 0 = male). The data are in donner.csv. Load the data

**How to fit a logistic regression to yes/no data**

The concept is very similar to fitting a multiple regression, because you use Analyze / Fit Model.

The difference is the type of the response variable:

 Multiple regression: continuous (blue ramp) variable

 Logistic regression: categorical (red bar) variable

When the response is recorded as a character variable (yes/no, Y/N, survived/died), the variable will be stored as a red bar. When the response is recorded as a numeric variable (0/1), the variable will be stored as a blue ramp and you need to change it to a red bar.

Change survive to a red bar variable.

Then Analyze / Fit Model. Put survive into the Y box. You will see the Personality change to Nominal Logistic (instead of the Standard Least Squares that you're used to seeing). Put the desired model terms in the construct model effects box. Our example will use an additive model (no interactions) with age and male, so put age and male into the Construct Model Effects box.

Remember that logistic regression is modeling the odds of an event and you get to choose what "event" is. Do you want to model the odds (and hence the probability) of a 0 (death) or the odds (and hence the probability) of a 1 (survive)? This changes the signs of the coefficients in the model, but tests of = 0 are unaffected. You get to choose.

Look for the Target level box: right side, upper middle. That indicates the "event". The default is the smallest value (or first in an alphabetic sort, so A before B, no before yes). If you want to model the other "event", change the target level. The drop-down menu gives you your choices. Change the target level to 1. Then click Run.

Just prior to running the model, your Fit Model dialog should look like this (next page): Notice that survival is red bar, the personality is Nominal Logistic, and Target level is 1.



The results box has a lot of information about the fit. The results I find most useful are:

The estimated logistic regression coefficients are in the Parameter Estimates box. You should have a regression coefficient for age of -0.078 and a regression coefficient for fem of 1.597. Standard errors and Wald tests of true coefficient = 0 are provided for each coefficient. Below that box is a note saying what odds are modeled. You should see "For log odds of 1/0". If your estimated coefficients are +0.078 and -1.597 and the table is annotated "For log odds of 0/1", you are modeling the odds (and probability) of a 0 response. Either refit the model, or remember to rephrase all your results.

Likelihood ratio tests of each coefficient = 0 are in the Effect Likelihood Ratio tests box. These usually give very similar p-values unless the sample size is small. In general, I (and lots of others) prefer the Likelihood Ratio tests.

The "overall" test of the model (i.e. of all coefficients = 0, not including the intercept) is in the Whole Model Test box. This is a *k* degree of freedom test, where *k* is the number of parameters. Below the test information are the AICc and BIC values for this model. The RSquare value is an attempt to create a number analogous to RSquare for a least squares model for a situation where there are no squared errors. I ignore that.

Further information about the model and its predictions is obtained by clicking the red triangle by nominal logistic fit for XXX. Some useful options are:

 Confidence intervals: add confidence intervals for each regression coefficient to the Param Ests box

 JMP provides likelihood-based confidence intervals.

 Odds ratio: convert the parameter estimates (on log odds scale) to odds ratios

You get two sets of odds ratios. The first are for a 1 unit change in each variable (so Male (0) to Female (1) , or Age 30 to Age 31). The second are for changing the X variable over the entire range of the variable (so Male to Female or Age 15 (youngest) to Age 65 (oldest) ).

 For each you get the odds ratio, a 95% confidence interval, and the reciprocal odds.

Save Probability Formula: adds columns to the data window with the linear predictor, probability of the event and probability of the "not" event, and the most likely event (larger probability). You can add rows to the data set to calculate probabilities for situations of interest that might not be in the data set, e.g. female age 60.

There are also options for other useful results that we won't talk about, such as the Receiver Operating Characteristic curve (ROC curve), Confusion Matrix, and a Profiler that allows you to plot contributions of each variable.