Logistic regression

1. data structure

There are two common ways in which data can be entered for logistic regression, either as individual observations or as grouped counts.

1) If individual data points are entered, each line of the data file corresponds to a single individual. The columns will corresponds to the predictors (X) that can be continuous (interval or ratio scales) or classification variables (nominal or ordinal). The response (Y) must be a classification variable with any two possible outcomes. Most packages will arbitrarily choose one of these classes to be the success –often this is the first category when sorted alphabetically. I would recommend that you do NOT code the response variable as 0/1 – it is far to easy to forget that the 0/1 correspond to nominally or ordinally scaled variables and not to continuous variables. As an example, suppose you wish to predict if an egg will hatch given the height in a tree. The data structure for individuals would look something like:

Egg	Height	Outcome
1	10	hatch
2	15	not hatch
3	5	hatch
4	10	hatch
5	10	not hatch
...		

2) In grouped counts, each line in the data file corresponds to a group of events with the same predictor (X) variables. Often researchers record the number of events and the number of successes in two separate columns, or the number of success and the number of failures in two separate columns.

   a) if the number of events and the number of successes in two separate columns, then we don't need to do any change in data format, but need to be careful when we input the Y variables in the Y box using "Fit model" and "generalized linear model "with "binomial distribution".

Height	total	Hatch
10	3	2

15	1	0
5	1	0
…		

b) If the number of success and the number of failures in two separate columns then we can make the data table like a) using formula to generate a new column.

2. Logistic regression

For two types of data format there are two ways to do logistic regression, respectively.

1) Individual data:

Dose Response Example

In the Dose Response.jmp sample data table, the dose varies between 1 and 12.

1. Open the Dose Response.jmp sample data table.
2. Select **Analyze > Fit Model**.
3. Select response and click **Y**.
4. Select dose and click **Add**.
5. Click **Run**.
6. From the red triangle next to Nominal Logistic Fit, select **Odds Ratio**.

2) Grouped data

— Check the necessary of transformation on x: calculate logit, then plot logit vs each x. Do transformation if need.

— use **Analyze->Fit model platform** and then check "**generalized linear model**" and "**binomial distribution**". We get similar results with more goodies under the red-triangles such as confidence intervals for the MEAN probability of success that can be saved to the data table, residual plots, and more. Note: you need put both "hatch" and "total" in Y box and "hatch" go first.

— Use the dataset ex2116.jmp

Practice:

I. Dataset ex2018.jmp:

   a. Check whether the tire-related fatal accident depends on the vehicle age.

   b. Check whether the odds that a fatal accident is tire-related depend on whether the vehicle is a Ford, after accounting for age of the car and number of passengers.

II. Use the dataset case2101.jmp to find the relationship between the species extinction and the island size.