Chapter 22

Logistic Regression

Contents

22.1 Introduction

22.1.1 Difference between standard and logistic regression

In regular multiple-regression problems, the Y variable is assumed to have a continuous distribution with the vertical deviations around the regression line being independently normally distributed with a mean of 0 and a constant variance σ^2. The X variables are either continuous or indicator variables.

In some cases, the Y variable is a categorical variable, often with two distinct classes. The X variables can be either continuous or indicator variables. The object is now to predict the CATEGORY in which a particular observation will lie.

For example:

- The Y variable is over-winter survival of a deer (yes or no) as a function of the body mass, condition factor, and winter severity index.
- The Y variable is fledging (yes or no) of birds as a function of distance from the edge of a field, food availability, and predation index.
- The Y variable is breeding (yes or no) of birds as a function of nest density, predators, and temperature.

Consequently, the linear regression model with normally distributed vertical deviations really doesn't make much sense – the response variable is a category and does NOT follow a normal distribution. In these cases, a popular methodology that is used is *logistic regression*.

There are a number of good books on the use of logistic regression:

- Agresti, A. (2002). Categorical Data Analysis. Wiley: New York.
- Hosmer, D.W. and Lemeshow, S. (2000). Applied Logistic Regression. Wiley: New York.

These should be consulted for all the gory details on the use of logistic regression.

22.1.2 The Binomial Distribution

A common probability model for outcomes that come in only two states (e.g. alive or dead, success or failure, breeding or not breeding) is the *Binomial* distribution. The *Binomial* distribution counts the number of times that a particular event will occur in a sequence of observations.[1] The binomial distribution is used when a researcher is interested in the occurrence of an event, not in its magnitude. For instance, in a clinical trial,

[1] The Poisson distribution is a close cousin of the Binomial distribution and is discussed in other chapters.

a patient may survive or die. The researcher studies the number of survivors, and not how long the patient survives after treatment. In a study of bird nests, the number in the clutch that hatch is measured, not the length of time to hatch.

In general the binomial distribution counts the number of events in a set of trials, e.g. the number of deaths in a cohort of patients, the number of broken eggs in a box of eggs, or the number of eggs that hatch from a clutch. Other situations in which binomial distributions arise are quality control, public opinion surveys, medical research, and insurance problems.

It is important to examine the assumptions being made before a Binomial distribution is used. The conditions for a Binomial Distribution are:

- n identical trials (n could be 1);
- all trials are independent of each other;
- each trial has only one outcome, success or failure;
- the probability of success is constant for the set of n trials. Some books use p to represent the probability of success; other books use π to represent the probability of success;[2]
- the response variable Y is the the number of successes[3] in the set of n trials.

However, not all experiments, that on the surface look like binomial experiments, satisfy all the assumptions required. Typically failure of assumptions include non-independence (e.g. the first bird that hatches destroys remaining eggs in the nest), or changing p within a set of trials (e.g. measuring genetic abnormalities for a particular mother as a function of her age; for many species, older mothers have a higher probability of genetic defects in their offspring as they age).

The probability of observing Y successes in n trials if each success has a probability p of occurring can be computed using:

$$p(Y = y|n, p) = \left(\begin{array}{c} n \\ y \end{array} \right) p^y \left(1 - p\right)^{n-y}$$

where the binomial coefficient is computed as

$$\left(\begin{array}{c} n \\ y \end{array} \right) = \frac{n!}{y!(n-y)!}$$

and where $n! = n(n-1)(n-2)\ldots(2)(1)$.

[2]Following the convention that Greek letters refer to the population parameters just like μ refers to the population mean.

[3]There is great flexibility in defining what is a *success*. For example, you could count either the number of successful eggs that hatch or the number of eggs that failed to hatch in a clutch. You will get the same answers from the analysis after making the appropriate substitutions.

For example, the probability of observing $Y = 3$ eggs hatch from a nest with $n = 5$ eggs in the clutch if the probability of success $p = .2$ is

$$p(Y = 3|n = 5, p = .2) = \left(\begin{array}{c} 5 \\ 3 \end{array} \right) (.2)^3 (1 - .2)^{5-3} = .0512$$

Fortunately, we will have little need for these probability computations. There are many tables that tabulate the probabilities for various combinations of n and p – check the web.

There are two important properties of a binomial distribution that will serve us in the future. If Y is Binomial(n, p), then:

- $E[Y] = np$
- $V[Y] = np(1 - p)$ and standard deviation of Y is $\sqrt{np(1 - p)}$

For example, if $n = 20$ and $p = .4$, then the average number of successes in these 20 trials is $E[Y] = np = 20(.4) = 8$.

If an experiment is observed, and a certain number of successes is observed, then the estimator for the success probability is found as:

$$\widehat{p} = \frac{Y}{n}$$

For example, if a clutch of 5 eggs is observed (the set of trials) and 3 successfully hatch, then the estimated proportion of eggs that hatch is $\widehat{p} = \frac{3}{5} = .60$. This is exactly analogous to the case where a sample is drawn from a population and the sample average $\overline{Y}$ is used to estimate the population mean μ.

22.1.3 Odds, risk, odds-ratio, and probability

The *odds* of an event and the *odds ratio* of events are very common terms in logistic contexts. Consequently, it is important to understand exactly what these say and don't say.

The odds of an event are defined as:

$$Odds(event) = \frac{P(event)}{P(not\ event)} = \frac{P(event)}{1 - P(event)}$$

The notation used is often a colon separating the odds values. Some sample values are tabulated below:

Probability	Odds
.01	1:99
.1	1:9
.5	1:1
.6	6:4 or 3:2 or 1.5
.9	9:1
.99	99:1

For very small or very large odds, the probability of the event is *approximately* equal to the odds. For example if the odds are 1:99, then the probability of the event is 1/100 which is roughly equal to 1/99.

The *odds ratio (OR)* is by definition, the ratio of two odds:

$$OR_{A\ vs.\ B} = \frac{odds(A)}{odds(B)} = \frac{\frac{P(A)}{1-P(A)}}{\frac{P(B)}{1-P(B)}}$$

For example, of the probability of an egg hatching under condition A is 1/10 and the probability of an egg hatching under condition B is 1/20, then the odds ratio is $OR = (1:9)/(1:19) = 2.1:1$. Again for very small or very larger odds, the odds ratio is *approximately* equal to the ratio of the probabilities.

An odds ratio of 1, would indicate that the probability of the two events is equal.

In many studies, you will hear reports that the odds of an event have doubled. This give NO information about the base rate. For example, did the odds increase from 1:million to 2:million or from 1:10 to 2:10.

It turns out that it is convenient to model probabilities on the *log-odds* scale. The log-odds (LO), also known as the *logit*, is defined as:

$$logit(A) = log_e(odds(A)) = log_e\left(\frac{P(A)}{1-P(A)}\right)$$

We can extend the previous table, to compute the log-odds:

Probability	Odds	Logit
.01	1:99	-4.59
.1	1:9	-2.20
.5	1:1	0
.6	6:4 or 3:2 or 1.5	.41
.9	9:1	2.20
.99	99:1	4.59

Notice that the log-odds is zero when the probability is .5 and that the log-odds of .01 is symmetric with the log-odds of .99.

It is also easy to go back from the log-odds scale to the regular probability scale in two equivalent ways:

$$p = \frac{e^{\textit{log-odds}}}{1 + e^{\textit{log-odds}}} = \frac{1}{1 + e^{-\textit{log-odds}}}$$

Notice the minus sign in the second back-translation. For example, a $LO = 10$, translates to $p = .9999$; a $LO = 4$ translates to $p = .98$; a $LO = 1$ translates to $p = .73$; etc.

22.1.4 Modeling the probability of success

Now if the probability of success was the same for all sets of trials, the analysis would be trivial: simply tabulate the total number of successes and divide by the total number of trials to estimate the probability of success. However, what we are really interested in is the relationship of the probability of success to some covariate X such as temperature, or condition factor.

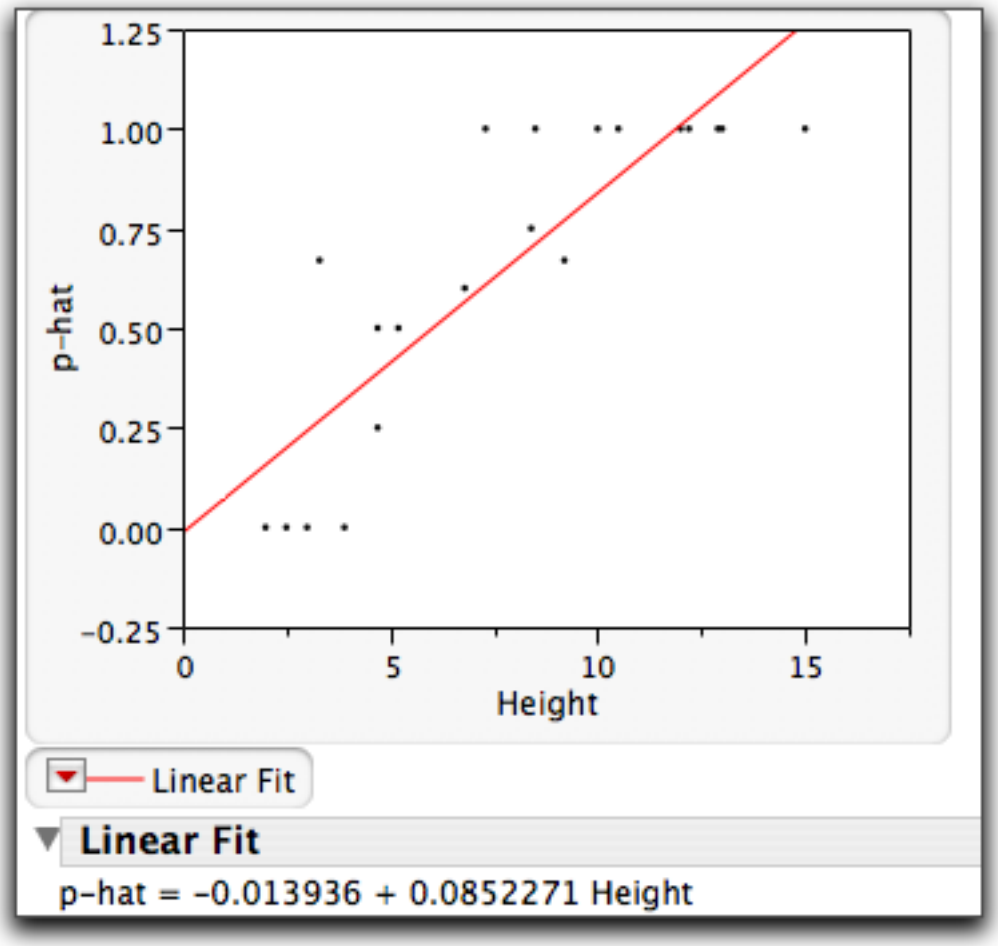
For example, consider the following (hypothetical) example of an experiment where various clutches of bird eggs were found, and the number of eggs that hatched and fledged were measured along with the height the nest was above the ground:

Height	Clutch Size	Fledged	$\widehat{p}$
2.0	4	0	0.00
3.0	3	0	0.00
2.5	5	0	0.00
3.3	3	2	0.67
4.7	4	1	0.25
3.9	2	0	0.00
5.2	4	2	0.50
10.5	5	5	1.00
4.7	4	2	0.50
6.8	5	3	0.60
7.3	3	3	1.00
8.4	4	3	0.75
9.2	3	2	0.67
8.5	4	4	1.00
10.0	3	3	1.00
12.0	6	6	1.00
15.0	4	4	1.00
12.2	3	3	1.00
13.0	5	5	1.00
12.9	4	4	1.00

Notice that the probability of a fledging seems to increase with height above the grounds (potentially reflecting distance from predators?).

We would like to model the probability of success as a function of height. As a first attempt, suppose that we plot the estimated probability of success ($\widehat{p}$) as a function of height and try and fit a straight line to the plotted points.

The *Analyze->Fit Y-by-X* platform was used, and $\widehat{p}$ was treated as the Y variable and *Height* as the X variable:



This procedure is not entirely satisfactory for a number of reasons:

- The data points seem to follow an S-shaped relationship with probabilities of success near 0 at lower heights and near 1 at higher heights.
- The fitted line gives predictions for the probability of success that are more than 1 and less than 0 which is impossible.
- The fitted line cannot deal properly with the fact that the probability of success is likely close to 0% for a wide range of small heights and essentially close to 100% for a wide range of taller heights.
- The assumption of a normal distribution for the deviations from the fitted line is not tenable as the $\widehat{p}$ are essentially discrete for the small clutch sizes found in this experiment.
- While not apparent from this graph, the variability of the response changes over the different parts of the regression line. For example, when the true probability of success is very low (say 0.1), the

standard deviation in the number fledged for a clutch with 5 eggs is found as $\sqrt{5(.1)(.9)} = .67$ while the standard deviation of the number of fledges in a clutch with 5 eggs and the probability of success of 0.5 is $\sqrt{5(.5)(.5)} = 1.1$ which is almost twice as large as the previous standard deviation.

For these (and other reasons), the analysis of this type of data are commonly done on the *log-odds* (also called the *logit*) scale. The odds of an event is computed as:

$$ODDS = \frac{p}{1-p}$$

and the log-odds is found as the (natural) logarithm of the odds:

$$LO = \log\left(\frac{p}{1-p}\right)$$

This transformation converts the 0-1 scale of probability to a $-\infty \rightarrow \infty$ scale as illustrated below:

p	LO
0.001	-6.91
0.01	-4.60
0.05	-2.94
0.1	-2.20
0.2	-1.39
0.3	-0.85
0.4	-0.41
0.5	0.00
0.6	0.41
0.7	0.85
0.8	1.39
0.9	2.20
0.95	2.94
0.99	4.60
0.999	6.91

Notice that the log-odds scale is symmetrical about 0, and that for moderate values of p, changes on the p-scale have nearly constant changes on the log-odds scale. For example, going from .5 $\rightarrow$ .6 $\rightarrow$ .7 on the p-scale corresponds to moving from 0 $\rightarrow$ .41 $\rightarrow$ .85 on the log-odds scale.

It is also easy to go back from the log-odds scale to the regular probability scale:

$$p = \frac{e^{LO}}{1+e^{LO}} = \frac{1}{1+e^{-LO}}$$

For example, a $LO = 10$, translates to $p = .9999$; a $LO = 4$ translates to $p = .98$; a $LO = 1$ translates to $p = .73$; etc.

We can now return back to the previous data. At first glance, it would seem that the estimated log-odds is simply estimated as:

$$\widehat{LO} = \log\left(\frac{\widehat{p}}{1-\widehat{p}}\right)$$

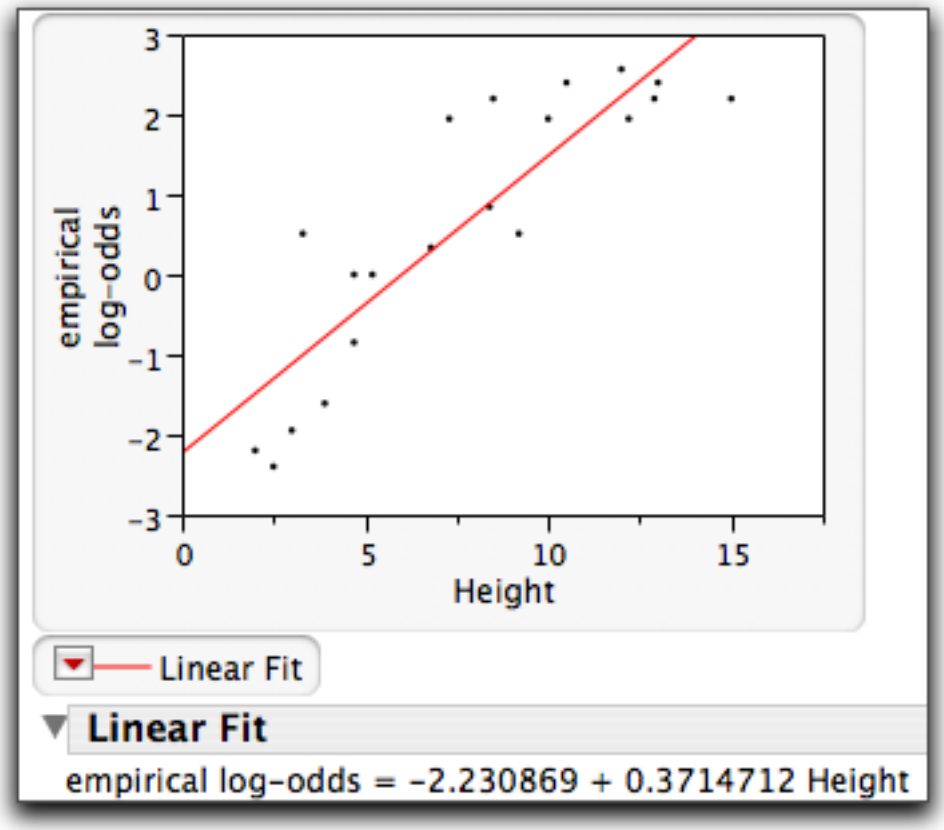
but this doesn't work well with small sample sizes (it can be shown that the simple logit function is biased) or when values of $\widehat{p}$ close to 0 or 1 (the simple logit function hits $\pm\infty$). Consequently, in small samples or when the observed probability of success is close to 0 or 1, the empirical log-odds is often computed as:

$$\widehat{LO}_{empirical} = \log\left(\frac{n\widehat{p}+.5}{n(1-\widehat{p})+.5}\right) = \log\left(\frac{\widehat{p}+.5/n}{1-\widehat{p}+.5/n}\right)$$

We compute the empirical log-odds for the hatching data:

Height	Clutch	Fledged	$\widehat{p}$	$\widehat{LO}_{emp}$
2.0	4	0	0.00	-2.20
3.0	3	0	0.00	-1.95
2.5	5	0	0.00	-2.40
3.3	3	2	0.67	0.51
4.7	4	1	0.25	-0.85
3.9	2	0	0.00	-1.61
5.2	4	2	0.50	0.00
10.5	5	5	1.00	2.40
4.7	4	2	0.50	0.00
6.8	5	3	0.60	0.34
7.3	3	3	1.00	1.95
8.4	4	3	0.75	0.85
9.2	3	2	0.67	0.51
8.5	4	4	1.00	2.20
10.0	3	3	1.00	1.95
12.0	6	6	1.00	2.56
15.0	4	4	1.00	2.20
12.2	3	3	1.00	1.95
13.0	5	5	1.00	2.40
12.9	4	4	1.00	2.20

and now plot the empirical log-odds against height:



The fit is much nicer, the relationship has been linearized, and now, no matter what the prediction, it can always be translated back to a probability between 0 and 1 using the inverse transform seen earlier.

22.1.5 Logistic regression

But this is still not enough. Even on the log-odds scale the data points are not normally distributed around the regression line. Consequently, rather than using ordinary least-squares to fit the line, a technique called *generalized linear modeling* is used to fit the line.

In generalized linear models a method called *maximum likelihood* is used to find the parameters of the model (in this case, the intercept and the regression coefficient of height) that gives the best fit to the data. While details of *maximum likelihood estimation* are beyond the scope of this course, they are closely related to *weighted least squares* in this class of problems. *Maximum Likelihood Estimators* (often abbreviated as MLEs) are, under fairly general conditions, guaranteed to be the "best" (in the sense of having smallest standard errors) in large samples. In small samples, there is no guarantee that MLEs are optimal, but in practice, MLEs seem to work well. In most cases, the calculations must be done numerically – there are no

simple formulae as in simple linear regression.[4]

In order to fit a logistic regression using maximum likelihood estimation, the data must be in a standard format. In particular, both success and failures must be recorded along with a classification variable that is nominally scaled. For example, the first clutch (at 2.0 m) will generate two lines of data – one for the successful fledges and one for the unsuccessful fledges. If the count for a particular outcome is zero, it can be omitted from the data table, but I prefer to record a value of 0 so that there is no doubt that all eggs were examined and none of this outcome were observed.

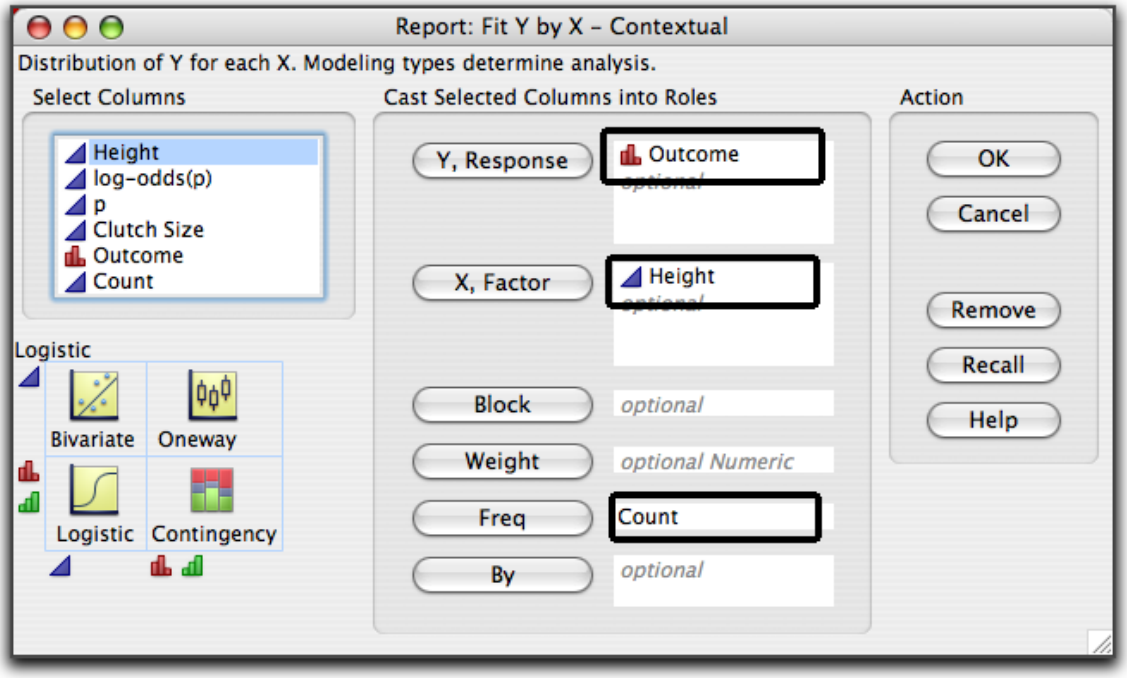
A new column was created in *JMP* for the number of eggs that failed to fledge, and after stacking the revised dataset, the dataset in *JMP* that can be used for logistic regression looks like:[5]

[4]Other methods that are qute popular are non-iterative weighted least squares and discriminant function analysis. These are beyond the scope of this course.

[5] This stacked data is available in the *eggsfledge2.jmp* dataset available from the Sample Program Library at `http://www.stat.sfu.ca/~cschwarz/Stat-650/Notes/MyPrograms`.

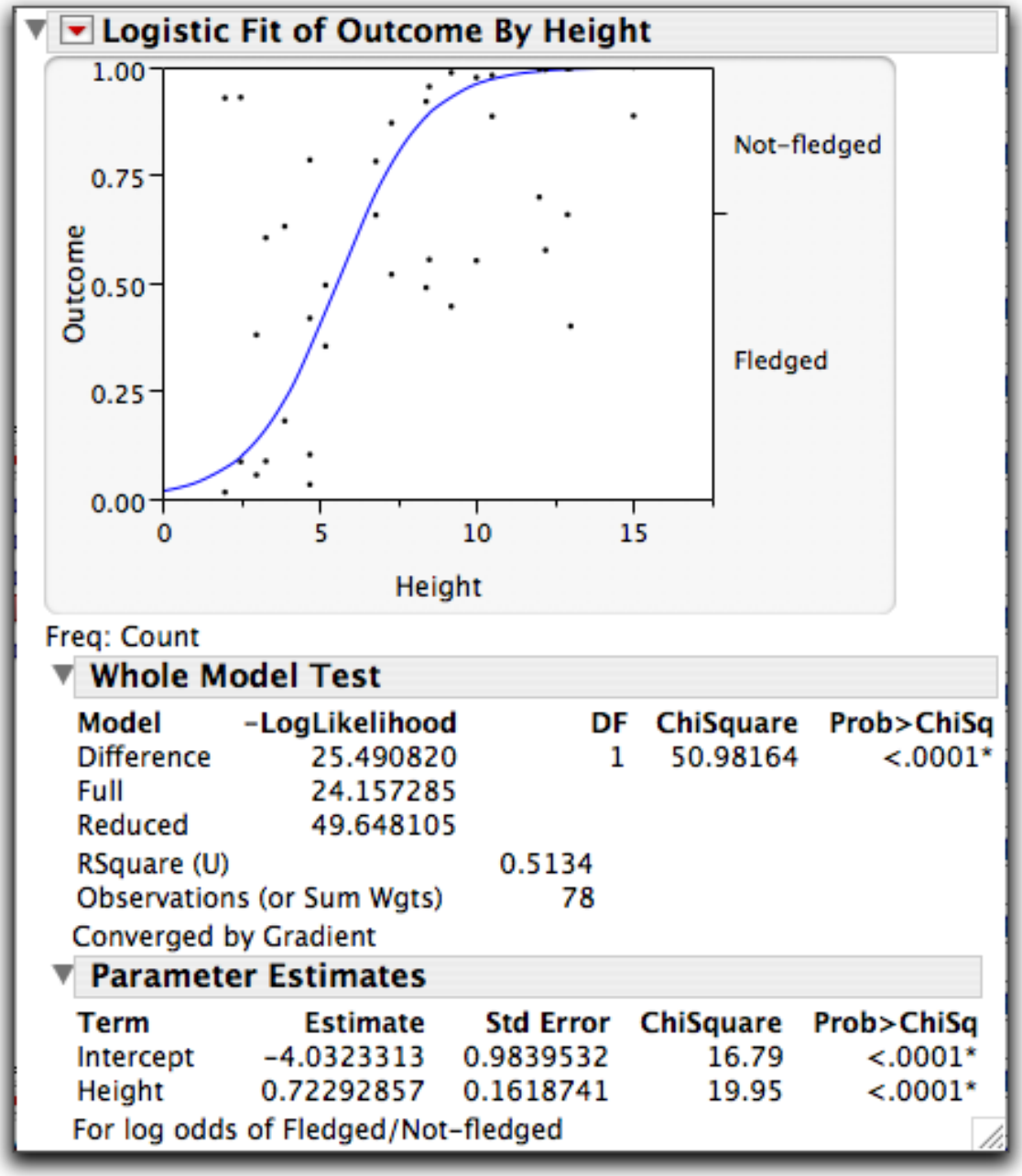
ols	Height	Clutch Size	Outcome	Count
1	2.0	4	Fledged	0
2	2.0	4	not fledged	4
3	3.0	3	Fledged	0
4	3.0	3	not fledged	3
5	2.5	5	Fledged	0
6	2.5	5	not fledged	5
7	3.3	3	Fledged	2
8	3.3	3	not fledged	1
9	4.7	4	Fledged	1
10	4.7	4	not fledged	3
11	3.9	2	Fledged	0
12	3.9	2	not fledged	2
13	5.2	4	Fledged	2
14	5.2	4	not fledged	2
15	10.5	5	Fledged	5
16	10.5	5	not fledged	0
17	4.7	4	Fledged	2
18	4.7	4	not fledged	2
19	6.8	5	Fledged	5
20	6.8	5	not fledged	0
21	7.3	3	Fledged	3
22	7.3	3	not fledged	0
23	8.4	4	Fledged	4
24	8.4	4	not fledged	0
25	9.2	3	Fledged	3
26	9.2	3	not fledged	0
27	8.5	4	Fledged	4
28	8.5	4	not fledged	0
29	10.0	3	Fledged	3
30	10.0	3	not fledged	0
31	12.0	6	Fledged	6
32	12.0	6	not fledged	0
33	15.0	4	Fledged	4
34	15.0	4	not fledged	0
35	12.2	3	Fledged	3
36	12.2	3	not fledged	0
37	13.0	5	Fledged	5
38	13.0	5	not fledged	0
39	12.9	4	Fledged	4
40	12.9	4	not fledged	0

The *Analyze->Fit Y-by-X* platform is used to launch simple logistic regression:



Note that the *Outcome* is the actual Y variable (and is nominally scaled) while the *Count* column simply indicates how many of this outcome were observed. The X variable is *Height* as before. *JMP* knows this is a logistic regression by the combination of a nominally or ordinally scaled variable for the Y variable, and a continuously scaled variable for the X variable as seen by the reminder at the left of the platform dialogue box.

This gives the output:



The first point to note is that most computer packages make arbitrary decisions on what is a "success" and what is a "failure" when fitting the logistic regression. It is important to always look at the output carefully to see what has been defined as a success. In this case, at the bottom of the output, *JMP* has indicated that *fledged* is considered as a "success" and *not fledged* as a "failure". If it had reversed the roles of these two

categories, everything would be "identical" except reversed appropriately.

Second, rather bizarrely, **the actual data points plotted by *JMP* really don't any meaning!** According the *JMP* help screens:

> Markers for the data are drawn at their x-coordinate, with the y position jittered randomly within the range corresponding to the response category for that row.

So if you do the analysis on the exact same data, the data points are jittered and will look different even though the fit is the same. The explanation in the *JMP* support pages on the web state:[6]

> The exact vertical placement of points in the logistic regression plots (for instance, on pages 308 and 309 of the *JMP* User's Guide, Version 2, and pages 114 and 115 of the *JMP* Statistics and Graphics Guide, Version 3) has no particular interpretation. The points are placed midway between curves so as to assure their visibility. However, the location of a point between a particular set of curves is important. All points between a particular set of curves have the same observed value for the dependent variable. Of course, the horizontal placement of each point is meaningful with respect to the horizontal axis.

This is rather unfortunate, to say the least! This means that the user must create nice plot by hand. This plot should plot the estimated proportions as a function of height with the fitted curve then overdrawn.

Fortunately, the fitted curves are correct (whew). The curves presented doesn't look linear only because *JMP* has transformed back from the log-odds scale to the regular probability scale. A linear curve on the log-odds scale has a characteristic "S" shape on the regular probability scale with the ends of the curve flattening out a 0 and 1. Using the *Cross Hairs* tool, you can see that a height of 5 m gives a predicted probability of success (*fledged*) of .57; by 7 m the estimated probability of success is almost 100%.

The table of parameter estimates gives the estimated fit on the log-odds scale:

$$\widehat{LO} = -4.03 + .72(Height)$$

Substituting in the value for $Height = 5$, gives an estimated log-odds of $-.43$ which on the regular probability scale corresponds to $.394$ as seen before from using the cross hairs.

The coefficient associated with *height* is interpreted as the increase in log-odds of fledging when height is increased by 1 m.

As in simple regression, the precision of the estimates is given by the standard error. An approximate 95% confidence interval for the coefficient associated with height is found in the usual fashion, i.e. $estimate \pm 2se$.[7] This confidence interval does NOT include 0; therefore there is good evidence that the probability of fledging is not constant over the various heights.

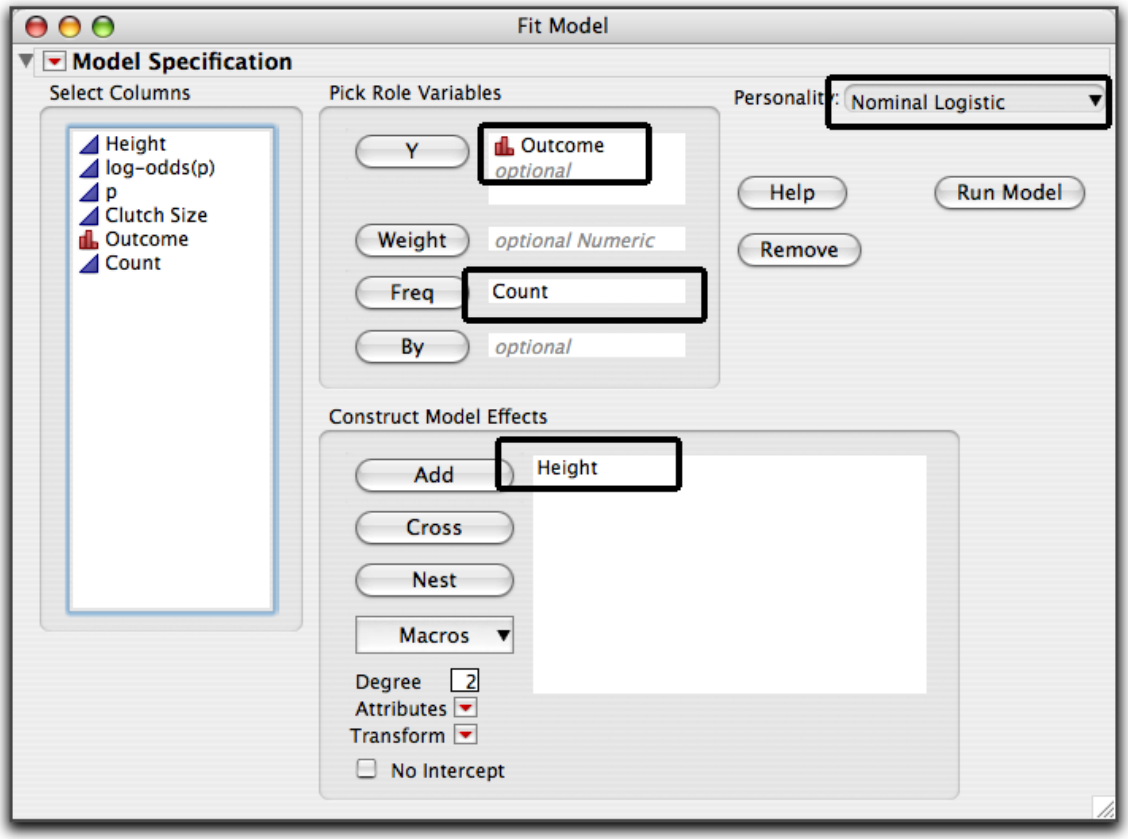
[6] `http://www.jmp.com/support/techsup/notes/001897.html`

[7] It is not possible to display the 95% confidence intervals in the *Analyze->Fit Y-by-X* platform output by right clicking in the table (don't ask me why not). However, if the *Analyze->Fit Model* platform is used to fit the model, then right-clicking in the Estimates table does make the 95% confidence intervals available.

Similarly, the p-value is interpreted in the same way – how consistent is the data with the hypothesis of NO effect of height upon the survival rate. Rather than the *t-test* seen in linear regression, maximum likelihood methods often constructs the test statistics in a different fashion (called χ^2 likelihood ratio tests). The test statistic is not particularly of interest – only the final p-value matters. In this case, it is well below $\alpha = .05$, so there is good evidence that the probability of success is not constant across heights. As in all cases, statistical significance is no guarantee of biological relevance.

In theory, it is possible to obtain prediction intervals and confidence intervals for the MEAN probability of success at new values of X – *JMP* does not provide these in the *Analyze->Fit Y-by-X* platform with logistic regression. It does do *Inverse Predictions* and can give confidence bounds on the inverse prediction which require the confidence bounds to be computed, so it is a mystery to me why the confidence intervals for the mean probability of success at future X values are not provided.

The *Analyze->Fit Model* platform can also be used to fit a logistic regression in the same way:



Be sure to specify the Y variable as a nominally or ordinally scaled variable; the *count* as the frequency variable; and the X variables in the usual fashion. The *Analyze->Fit Model* platform automatically switches to indicate logistic regression will be run.

The same information as previously seen is shown again. But, you can now obtain 95% confidence

intervals for the parameter estimates and there are additional options under the red-triangle pop-down menu. These features will be explored in more detail in further examples.

Lastly, the *Analyze->Fit Model* platform using the *Generalized Linear Model* option in the personality box in the upper right corner, also can be used to fit this model. Specify a binomial distribution with the logit link. You 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. Again, these will be explored in more details in the examples.

22.2 Data Structures

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

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[8]. 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
...		

Notice that even though three eggs were all at 10 m height, separate data lines for each of the three eggs appear in the data file.

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. This data must be converted to two rows per group - one for the success and one for the failures with one variable representing

[8] In more advanced classes this restriction can be relaxed.

the outcome and a second variable representing the frequency of this event. The outcome will be the Y variable, while the count will be the *frequency* variable.[9]

For example, the above data could be originally entered as:

Height	Hatch	Not Hatch
10	2	1
15	0	1
5	1	0
...		

but must be translated (e.g. using the *Tables* $\rightarrow$ *Stack* command) to:

Height	Outcome	Count
10	Hatch	2
10	Not Hatch	1
15	Hatch	0
15	Not Hatch	1
5	Hatch	1
5	Not Hatch	0
...		

While it is not required that counts of zero have data lines present, it is good statistical practice to remind yourself that you did look for failures, but failed to find any of them.

22.3 Assumptions made in logistic regression

Many of the assumptions made for logistic regression parallel those made for ordinary regression with obvious modifications.

1. **Check sampling design.** In these course notes it is implicitly assumed that the data are collected either as simple random sample or under a completely randomized design experiment. This implies that the units selected must be a random sample (with equal probability) from the relevant populations or complete randomization during the assignment of treatments to experimental units. The experimental unit must equal the observational unit (no pseudo-replication), and there must be no pairing, blocking, or stratification.

   It is possible to generalize logistic regression to cases where pairing, blocking, or stratification took place (for example, in case-control studies), but these are not covered during this course.

[9] Refer to the section on Poisson regression for an alternate way to analyze this type of data where the count is the response variable.

Common ways in which assumption are violated include:

- Collecting data under a cluster design. For example, class rooms are selected at random from a school district and individuals within a class room are then measured. Or herds or schools of animals are selected and all individuals within the herd or school are measured.
- Quota samples are used to select individuals with certain classifications. For example, exactly 100 males and 100 females are sampled and you are trying to predict sex as the outcome measure.

2. **No outliers.** This is usually pretty easy to check. A logistic regression only allows two categories within the response variables. If there are more than two categories of responses, this may represent a typographical error and should be corrected. Or, categories should be combined into larger categories.

   It is possible to generalize logistic regression to the case of more than two possible outcomes. Please contact a statistician for assistance.

3. **Missing values are MCAR**. The usual assumption as listed in earlier chapters.

4. **Binomial distribution.** This is a crucial assumption. A binomial distribution is appropriate when there is a fixed number of trials at a given set of covariates (could be 1 trial); there is constant probability of "success" within that set of trials; each trial is independent; and the number of trials in the n successes is measured.

   Common ways in which this assumption is violated are:

   - Items within a set of trials do not operate independently of each other. For example, subjects could be litter mates, twins, or share environmental variables. This can lead to over- or under-dispersion.
   - The probability of success within the set of trials is not constant. For example, suppose a set of trials is defined by weight class. Not everyone in the weight class is exactly the same weight and so their probability of "success" could vary. Animals all don't have exactly the same survival rates.
   - The number of trials is not fixed. For example, sampling could occur until a certain number of success occur. In this case, a negative binomial distribution would be more appropriate.

5. **Independence among subjects.** See above.

22.4 Example: Space Shuttle - Single continuous predictor

In January 1986, the space shuttle *Challenger* was destroyed on launch. Subsequent investigations showed that an O-ring, a piece of rubber used to seal two segments of the booster rocket, failed, allowing highly flammable fuel to leak, light, and destroy the ship.[10]

As part of the investigation, the following chart of previous launches and the temperature at which the shuttle was launched was presented:

[10] Refer to `http://en.wikipedia.org/wiki/Space_Shuttle_Challenger_disaster`.

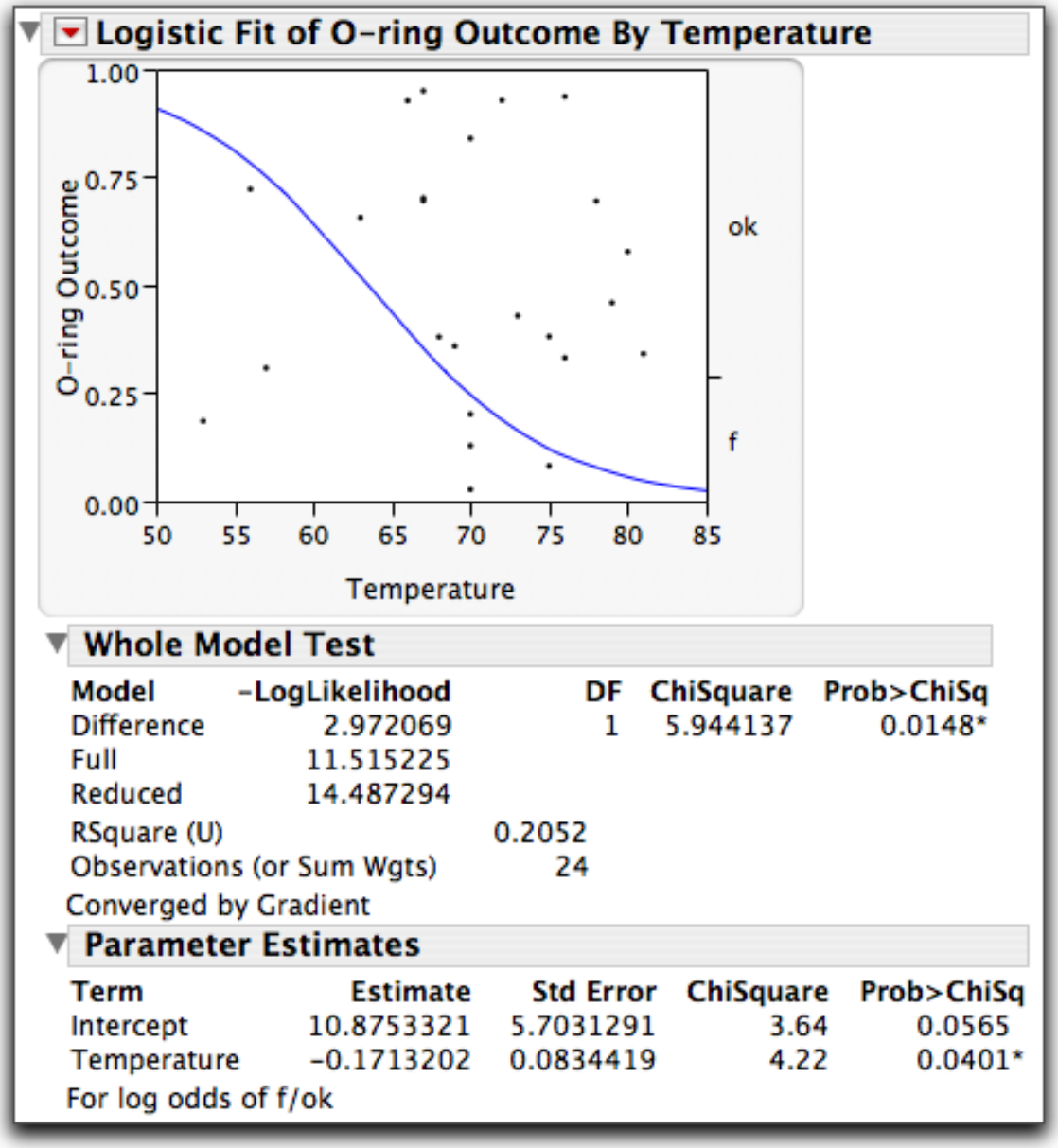
	Temperature	O-ring Outcome
1	53	f
2	56	f
3	57	f
4	63	ok
5	66	ok
6	67	ok
7	67	ok
8	67	ok
9	68	ok
10	69	ok
11	70	ok
12	70	f
13	70	f
14	70	f
15	72	ok
16	73	ok
17	75	ok
18	75	f
19	76	ok
20	76	ok
21	78	ok
22	79	ok
23	80	ok
24	81	ok

The raw data is available in the *JMP* file *spaceshuttleoring.jmp* available from the Sample Program Library at `http://www.stat.sfu.ca/~cschwarz/Stat-650/Notes/MyPrograms`.

Notice that the raw data has a single line for each previous launch even though there are multiple launches at some temperatures. The X variable is temperature and the Y variable is the outcome – either *f* for failure of the O-ring, or *OK* for a launch where the O-ring did not fail.

With the data in a single observation mode, it is impossible to make a simple plot of the empirical logistic function. If some of the temperatures were pooled, you might be able to do a simple plot.

The *Analyze->Fit Y-by-X* platform was used and gave the following results:



First notice that *JMP* treats a failure *f* as a "success", and will model the probability of failure as a function

of temperature. This is why it is important that you examine computer output carefully to see exactly what a package is doing.

The graph showing the fitted logistic curve must be interpreted carefully. While the plotted curve is correct, the actual data points are randomly placed - groan – see the notes in the previous section.

The estimated model is:

$$\widehat{logit}(failure) = 10.875 - .17(temperature)$$

So, the log-odds of failure decrease by .17 (se .083) units for every degree (°F) increase in launch temperature. Conversely, the log-odds of failure increase by .17 by every degree (°F) decrease in temperature.

The *p*-value for no effect of temperature is just below $\alpha = .05$.

Using the same reasoning as was done for ordinary regression, the odds of failure increase by a factor of $e^{.17} = 1.18$, i.e. almost a 18% increase per degree drop.

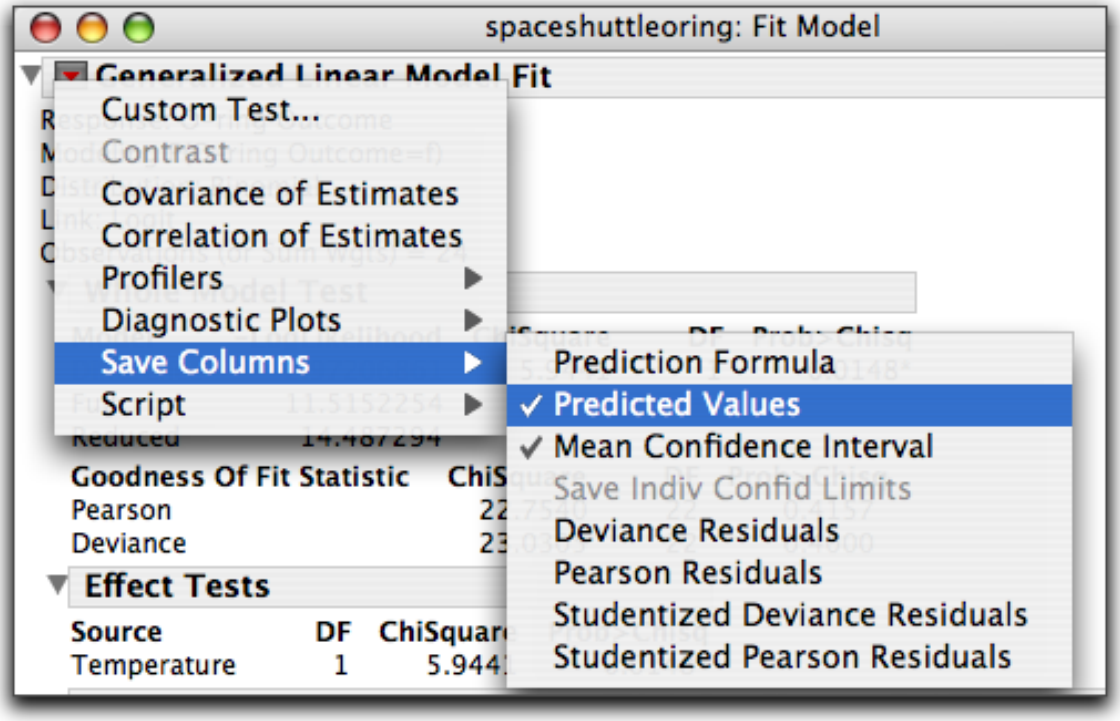
To predict the failure rate at a given temperature, a two stage-process is required. First, estimate the log-odds by substituting in the X values of interest. Second, convert the estimated log-odds to a probability using $p(x) = \frac{e^{LO(x)}}{1+e^{LO(x)}} = \frac{1}{1+e^{-LO(x)}}$.

The actual launch was at 32°F. While it is extremely dangerous to try and predict outside the range of observed data, the estimated log-odds of failure of the O-ring are $10.875 - .17(32) = 5.43$ and then $p(failure) = \frac{e^{5.43}}{1+e^{5.43}} = .99+$, i.e. well over 99%!

It is possible to find confidence bounds for these predictions – the easiest way is to create some "dummy" rows in the data table corresponding to the future predictions with the response variable left blank. Use *JMP*'s *Exclude Rows* feature to exclude these rows from the model fit. Then use the red-triangle to same predictions and confidence bounds back to the data table.

The *Analyze->Fit Model* platform gives the same results with additional analysis options that we will examine in future examples.

The *Analyze->Fit Model* platform using the *Generalized Linear Model* option also gives the same results with additional analysis options. For example, it is possible to compute confidence intervals for the predicted probability of success at the new X. Use the pop-down menu beside the red-triangle:



The predicted values and 95% confidence intervals for the predicted probability are stored in the data table:

	Temperature	O-ring Outcome	Pred O-ring	Lower 95% Mean O-	Upper 95% Mean O-
2	56	f	0.78	0.28	0.97
3	57	f	0.75	0.28	0.96
4	63	ok	0.52	0.23	0.80
5	66	ok	0.39	0.18	0.65
6	67	ok	0.35	0.16	0.60
7	67	ok	0.35	0.16	0.60
8	67	ok	0.35	0.16	0.60
9	68	ok	0.32	0.14	0.56
10	69	ok	0.28	0.12	0.52
11	70	ok	0.25	0.10	0.49
12	70	f	0.25	0.10	0.49
13	70	f	0.25	0.10	0.49
14	70	f	0.25	0.10	0.49
15	72	ok	0.19	0.07	0.44
16	73	ok	0.16	0.05	0.42
17	75	ok	0.12	0.03	0.39
18	75	f	0.12	0.03	0.39
19	76	ok	0.10	0.02	0.38
20	76	ok	0.10	0.02	0.38
21	78	ok	0.08	0.01	0.36
22	79	ok	0.07	0.01	0.35
23	80	ok	0.06	0.01	0.35
24	81	ok	0.05	0.00	0.34

These are found by finding the predicted log-odds and a 95% confidence interval for the predicted log-odds and then inverting the confidence interval endpoints in the same way as the predicted probabilities are obtained from the predicted log-odds.

While the predicted value and the 95% confidence interval are available, for some odd reason the *se* of the predicted probability is not presented – this is odd as it is easily computed. The confidence intervals are quite wide given that there were only 24 data values and only a few failures.

It should be noted that only predictions of the probability of success and confidence intervals for the

probability of success are computed. These intervals would apply to all future subjects that have the particular value of the covariates. Unlike the case of linear regression, it really doesn't make sense to predict individual outcomes as these are categories. It is sensible to look at which category is most probable and then use this as a "guess" for the individual response, but that is about it. This area of predicting categories for individuals is called *discriminant* analysis and has a long history in statistics. There are many excellent books on this topic.

22.5 Example: Predicting Sex from physical measurements - Multiple continuous predictors

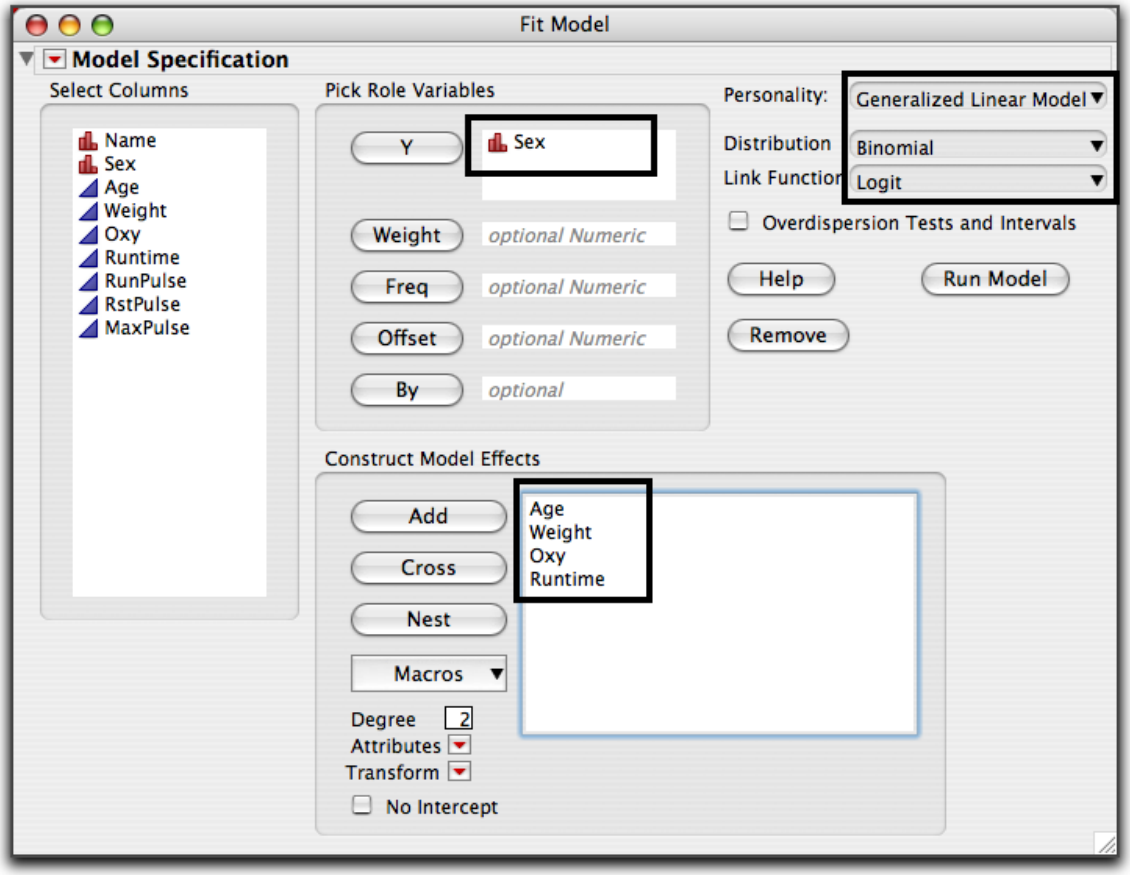
The extension to multiple continuous X variables is immediate. As before there are now several predictors. It is usually highly unlikely to have multiple observations with exactly the same set of X values, so the data sets usually consist of individual observations.

Let us proceed by example using the *Fitness* data set available in the *JMP* sample data library. This dataset has variables on age, weight, and measurements of performance while performing a fitness assessment. In this case we will try and predict the sex of the subject given the various attributes.

As usual, before doing any computations, examine the data for unusual points. Look at pairwise plots, the pattern of missing values, etc.

It is important that the data be collected under a completely randomized design or simple random sample. If your data are collected under a different design, e.g. a cluster design, please see suitable assistance.

Use the *Analyze->Fit Model* platform to fit a logistic regression trying to predict sex from the age, weight, oxygen consumption and run time:



This gives the summary output:

Generalized Linear Model Fit

Response: Sex
Modeling P(Sex=F)
Distribution: Binomial
Link: Logit
Observations (or Sum Wgts) = 31

Whole Model Test

Model	-LogLikelihood	ChiSquare	DF	Prob>Chisq
Difference	13.3694658	26.7389	4	<.0001*
Full	8.10196497			
Reduced	21.4714308			

Goodness Of Fit Statistic	ChiSquare	DF	Prob>Chisq
Pearson	12.7361	26	0.9862
Deviance	16.2039	26	0.9312

Effect Tests

Parameter Estimates

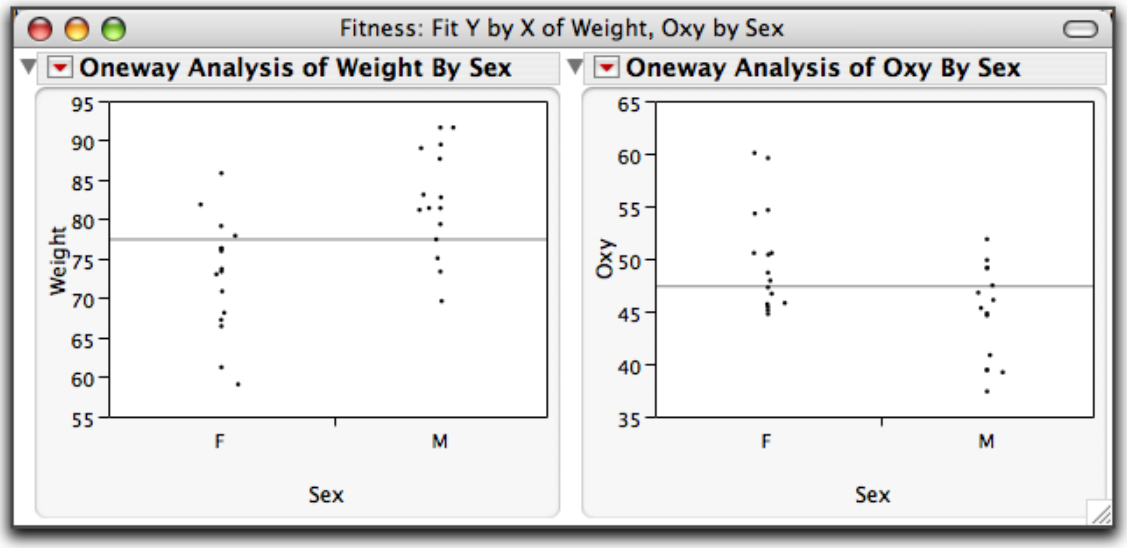
Term	Estimate	Std Error	ChiSquare	Prob>Chisq	Lower CL	Upper CL
Intercept	17.1846119	20.147174	0.7490	0.3868	-22.56774	62.354625
Age	-0.0263299	0.1300906	0.0411	0.8393	-0.305515	0.2417392
Weight	-0.7344983	0.4282811	17.9015	<.0001*	-1.573929	0.1049327
Oxy	0.92503062	0.5786904	5.8206	0.0158*	0.1249072	2.0592637
Runtime	-0.2195561	0.8181341	0.0718	0.7887	-1.90565	1.4727363

First determine which category is being predicted. In this case, the $sex = f$ category will be predicted.

The *Whole Model Test* examines if there is evidence of any predictive ability in the 4 predictor variable. The *p*-value is very small indicating that there is predictive ability.

Because we have NO categorical predictors, the *Effect Tests* can be ignored for now. The *Parameter Estimates* look for the marginal contribution of each predictor to predicting the probability of being a Female. Just like in regular regression, these are MARGINAL contributions, i.e. how much would the log-odds for the probability of being female change if this variable changed by one unit and all other variables remained in the model and did not change. In this case, there is good evidence that weight is a good predictor (not surprisingly), but also some evidence that oxygen consumption may be useful.[11] If you look at the dot plots for the weight for the two sexes and for the oxygen consumption for the two sexes, the two groups seem to be separated on these variables:

[11] The output above actually appears to be a bit contradictory. The chi-square value for the effect of weight is 17 with a *p*-value $< .0001$. Yet the 95% confidence interval for the coefficient associated with weight ranges from $(-1.57 \rightarrow .105)$ which INCLUDES zero, and so whould not be statistically significant! It turns out that *JMP* has mixed two (asymptotically) equivalent methods in this one output. The chi-square value and *p*-value are computed using a *likelihood ratio* test (a model with and without this variable is fit and the difference in fit is measured), while the confidence intervals are computed using a *Wald* approximation ($estimate \pm 2(se)$). In small samples, the sampling distribution for an estimate may not be very symmetric or close to normally shaped and so the Wald intervals may not perform well.

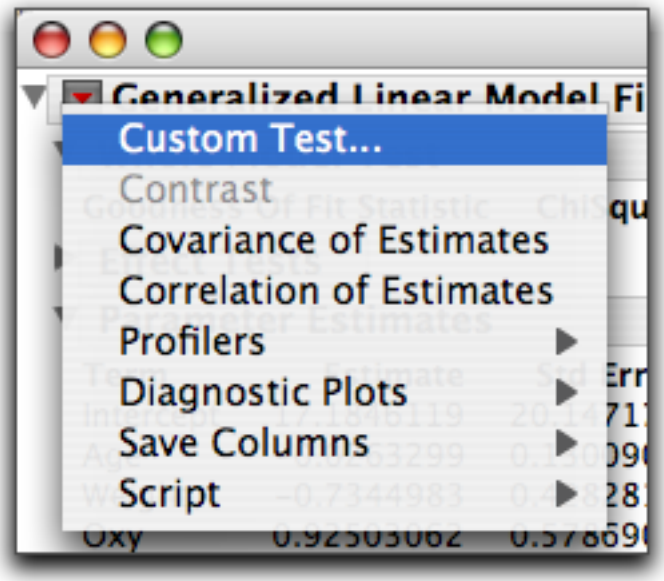


The estimated coefficient for weight is $-.73$. This indicates that the log-odds of being female decrease by .73 for every additional unit of weight, all other variables held fixed. This often appears in scientific reports as the adjusted effect of weight – the *adjusted* term implies that it is the marginal contribution.

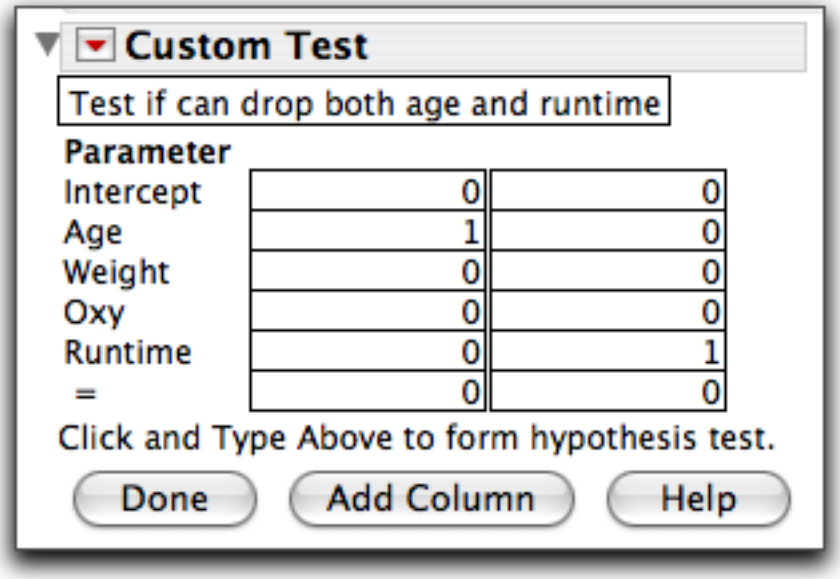
Confidence intervals for the individual coefficient (for predicting the log-odds of being female) are interpreted in the same way.

Just like in regular regression, collinearity can be a problem in the X values. There is no easy test for collinearity in logistic regression in *JMP*, but similar diagnostics as in in ordinary regression are becoming available.

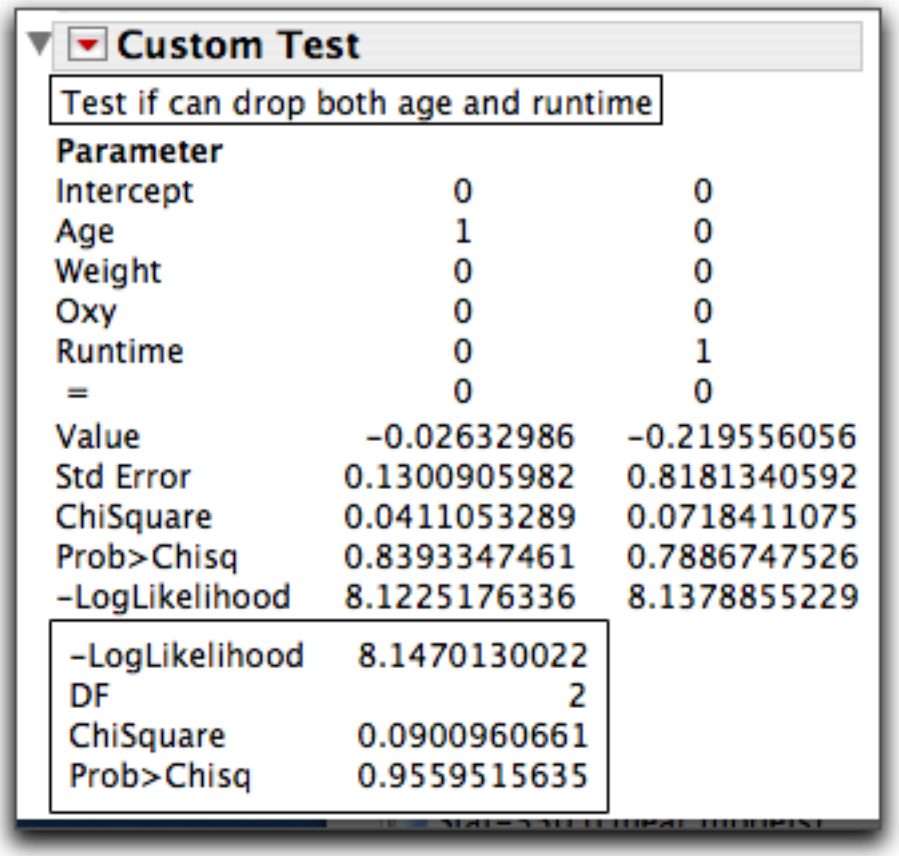
Before dropping more than one variable, it is possible to test if two or more variables can be dropped. Use the *Custom Test* options from the drop-down menu:



Complete the boxes in a similar way as in ordinary linear regression. For example, to test if both *age* and *runtime* can be dropped:



which gives:



Custom Test

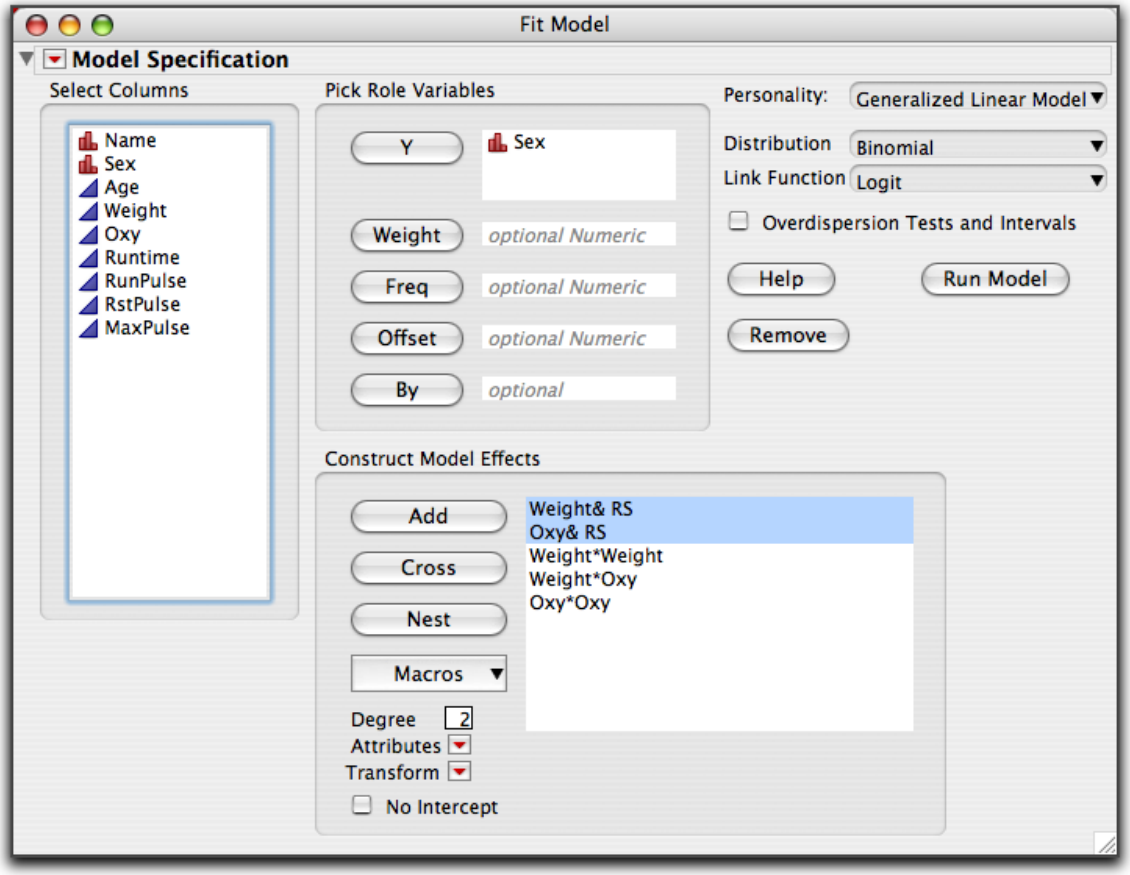
Test if can drop both age and runtime

Parameter		
Intercept	0	0
Age	1	0
Weight	0	0
Oxy	0	0
Runtime	0	1
=	0	0
Value	-0.02632986	-0.219556056
Std Error	0.1300905982	0.8181340592
ChiSquare	0.0411053289	0.0718411075
Prob>Chisq	0.8393347461	0.7886747526
-LogLikelihood	8.1225176336	8.1378855229

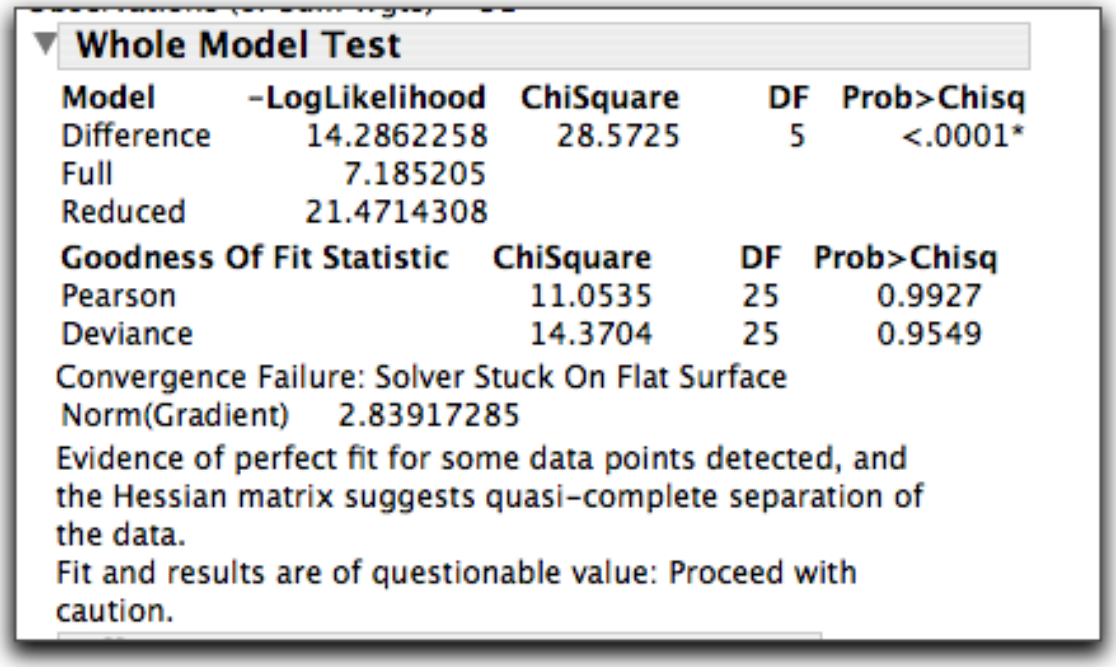
-LogLikelihood	8.1470130022
DF	2
ChiSquare	0.0900960661
Prob>Chisq	0.9559515635

It appears safe to drop both variables.

Just as in regular regression, you can fit quadratic and product terms to try and capture some non-linearity in the log-odds. This affects the interpretation of the estimated coefficients in the same way as in ordinary regression. The simpler model involving *weight* and *oxygen consumption*, their quadratic terms and cross product term was fit using the *Analyze->Fit Model* platform:



Surprisingly, the model has problems:



Whole Model Test

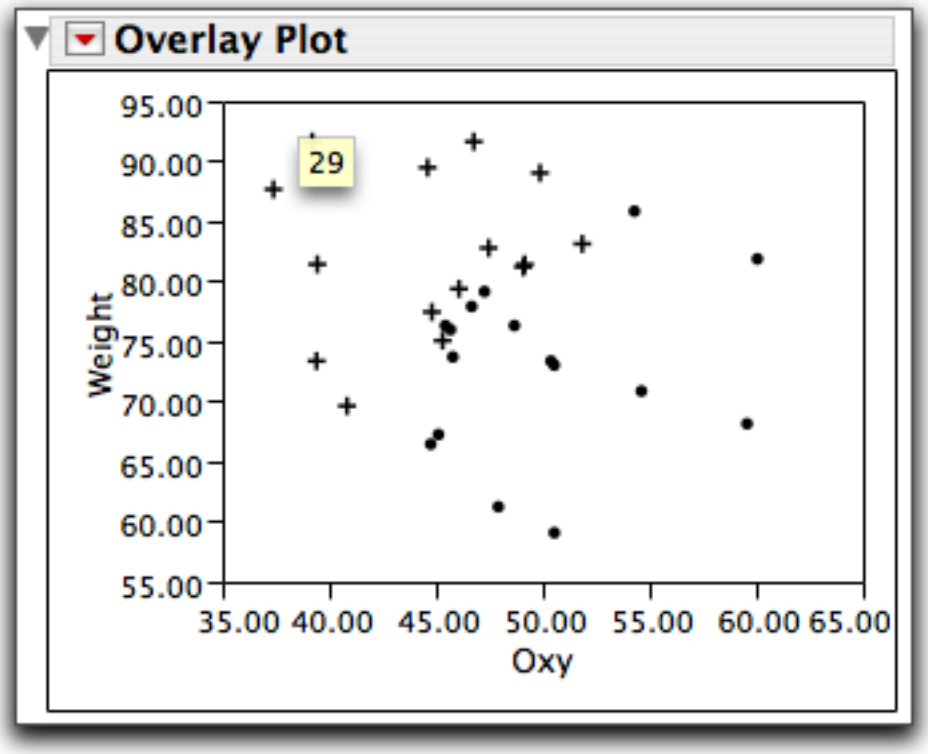
Model	-LogLikelihood	ChiSquare	DF	Prob>Chisq
Difference	14.2862258	28.5725	5	<.0001*
Full	7.185205			
Reduced	21.4714308			

Goodness Of Fit Statistic	ChiSquare	DF	Prob>Chisq
Pearson	11.0535	25	0.9927
Deviance	14.3704	25	0.9549

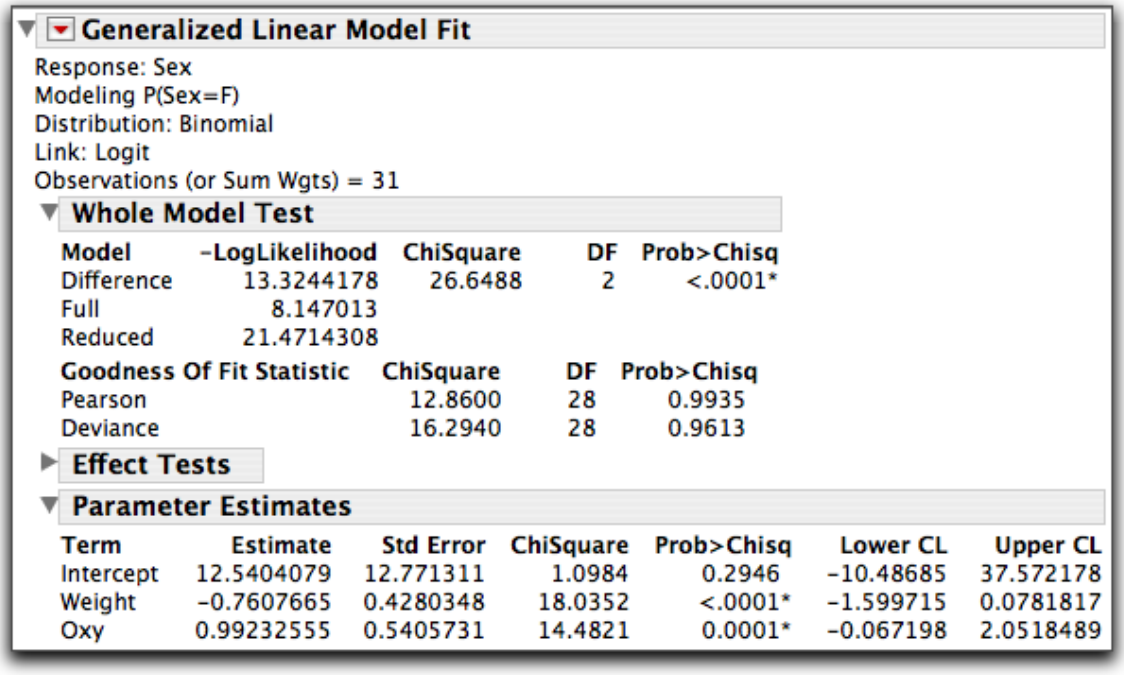
Convergence Failure: Solver Stuck On Flat Surface
Norm(Gradient) 2.83917285
Evidence of perfect fit for some data points detected, and the Hessian matrix suggests quasi-complete separation of the data.
Fit and results are of questionable value: Proceed with caution.

Ironically, it is because the model is too good of a fit. It appears that you can discriminate perfectly between men and women by fitting this model. Why does a perfect fit cause problems. The reason is that if the $p(sex = f) = 1$, the log-odds is then $+\infty$ and it is hard to get a predicted value of ∞ from an equation without some terms also being infinite.

If you plot the *weight* against *oxygen consumption* using different symbols for males and females, you can see the near complete separation based on simply looking at oxygen consumption and weight without the need for quadratic and cross products:



I'll continue by fitting just a model with linear effects of weight and oxygen consumption as an illustration. Use the *Analyze->Fit Model* platform to fit this model with just the two covariates:



Generalized Linear Model Fit

Response: Sex
Modeling P(Sex=F)
Distribution: Binomial
Link: Logit
Observations (or Sum Wgts) = 31

Whole Model Test

Model	-LogLikelihood	ChiSquare	DF	Prob>Chisq
Difference	13.3244178	26.6488	2	<.0001*
Full	8.147013			
Reduced	21.4714308			

Goodness Of Fit Statistic	ChiSquare	DF	Prob>Chisq
Pearson	12.8600	28	0.9935
Deviance	16.2940	28	0.9613

Effect Tests

Parameter Estimates

Term	Estimate	Std Error	ChiSquare	Prob>Chisq	Lower CL	Upper CL
Intercept	12.5404079	12.771311	1.0984	0.2946	-10.48685	37.572178
Weight	-0.7607665	0.4280348	18.0352	<.0001*	-1.599715	0.0781817
Oxy	0.99232555	0.5405731	14.4821	0.0001*	-0.067198	2.0518489

Both covariates are now statistically significant and cannot be dropped.

The *Goodness-of-fit* statistic is computed in two ways (which are asymptotically equivalent), but both are tedious to compute by hand. The *Deviance* of a model is a measure of how well a model performs. As there are 31 data points, you could get a perfect fit by fitting a model with 31 parameters – this is exactly what happens if you try and fit a line through 2 points where 2 parameters (the slope and intercept) will fit exactly two data points. A measure of goodness of fit is then found for the model in question based on the fitted parameters of this model. In both cases, the measure of fit is called the *deviance* which is simply twice the negative of the log-likelihood which in turn is related to the probability of observing this data given the parameter values. The difference in deviances is the deviance goodness-of-fit statistic. If the current model is a good model, the difference in deviance should be small (this is the column labeled *chi-square*). There is no simple calibration of deviances[12], so a *p*-value must be found which say how large is this difference. The *p*-value of .96 which indicates that the difference is actually quite small, almost 96% of the time you would get a larger difference in deviances.

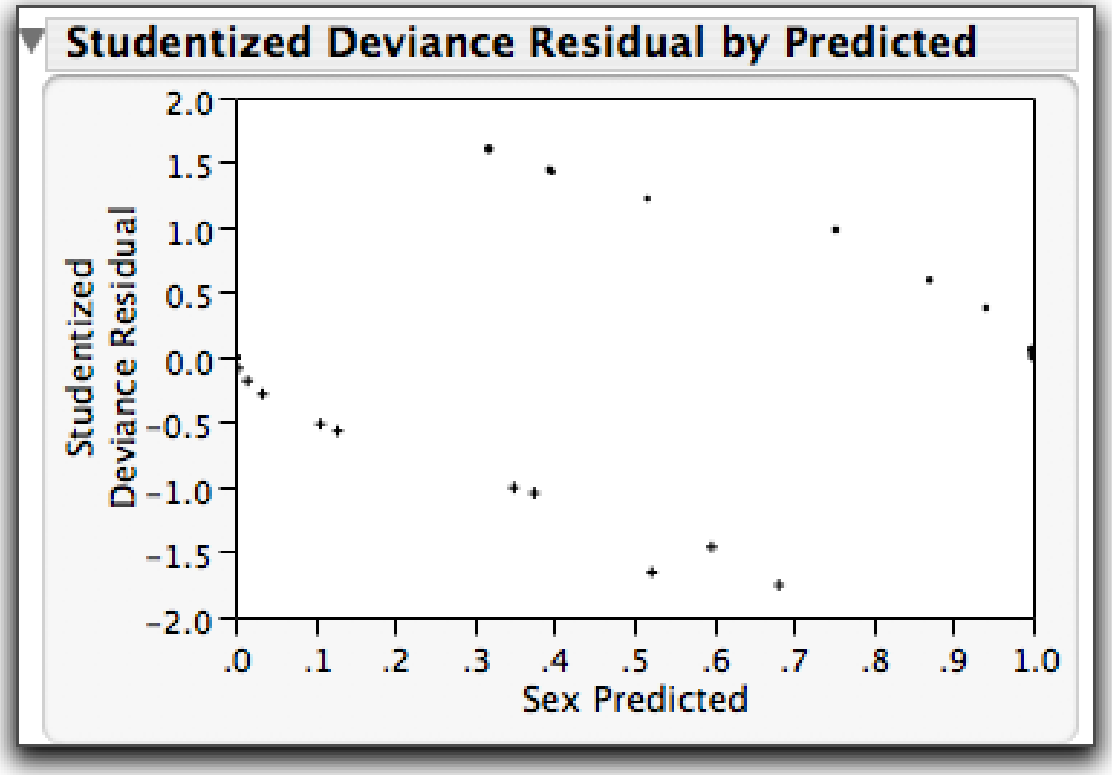
Similarly, the row labeled the Pearson goodness-of-fit is based on the same idea. A perfect fit is obtained with a model of 31 parameters. A comparison of the observed and predicted values is found for the model with 3 parameters. How big is the difference in fit? How unusual is it?

NOTE that for goodness-of-fit tests, you DO NOT WANT TO REJECT the null hypothesis. Hence *p*-values for a goodness-of-fit test that are small (e.g. less than $\alpha = .05$) are NOT good!

[12] The df=31-3

So for this model, there is no reason to be upset with the fit.

The residual plots look strange, but this is an artifact of the data:



Along the bottom axis is the predicted probability of being female. Now consider a male subject. If the predicted probability of being female is small (e.g. close to 0 because the subject is quite heavy), then there is an almost perfect agreement of the observed response with the predicted probability. If you compute a residual by defining a male=0 and female=1, then the residual here would be computed as $(obs - predicted)/se(predicted)$ = $(0 - 0)/blah = 0$. This corresponds to points near the (0,0) area of the plots.

What about males whose predicted probability of being female is almost .7 (which corresponds to observation 15). This is a poor prediction. and the residual is computed as $(0 - .7)/se(predicted)$ which is approximately equal to $(0 - .7)/\sqrt{.7(.3)} \approx -1.52$ with some further adjustment to compute the se of the predicted value. This corresponds to the point near (.7, -1.5).

On the other hand, a female with a predicted probability of being female will have a residual equal to approximately $(1 - .7)/\sqrt{.7(.3)} = .65$.

Hence the two lines on the graph correspond to males and female respectively. What you want to see is

this two parallel line system, particularly with few males near the probability of being female close to 1, and few females with probability of being female close to 0.

There are four possible residual plots available in *JMP* – they are all based on a similar procedure with minor adjustments in the way they compute a standard error. Usually, all four plots are virtually the same – anomalies among the plots should be investigated carefully.

22.6 Examples: Lung Cancer vs. Smoking; Marijuana use of students based on parental usage - Single categorical predictor

22.6.1 Retrospect and Prospective odds-ratio

In this section, the case where the predictor (X) variable is also a categorical variable will be examined. As seen in multiple linear regression, categorical X variables are handled by the creation of indicator variables. A categorical variable with k classes will generate $k-1$ indicator variables. As before, there are many ways to define these indicator variables and the user must examine the computer software carefully before using any of the raw estimated coefficients associated with a particular indicator variable.

It turns out that there are multiple ways to analyze such data – all of which are asymptotically equivalent. Also, this particular topic is usually divided into two sub-categories - problems where there are only two levels of the predictor variables and cases where there are three or more levels of the predictor variables. This division actually has a good reason – it turns out that in the case of 2 levels for the predictor and 2 levels for the response variable (the classic 2×2 contingency table), it is possible to use a retrospective study and actually get valid estimates of the prospective odds ratio.

For example, suppose you were interested in the looking at the relationship between smoking and lung cancer. In a prospective study, you could randomly select 1000 smokers and 1000 non-smokers for their relevant populations and follow them over time to see how many developed lung cancer. Suppose you obtained the following results:

Cohort	Lung Cancer	No lung cancer
Smokers	100	900
Non-smoker	10	990

Because this is a prospective study, it is quite valid to say that the probability of developing lung cancer if you are a smoker is $100/1000$ and the probability of developing lung cancer if you are not a smoker is $10/1000$. The odds of developing cancer if you are smoker are 100:900 and the odds of developing cancer if you are non-smoker are 10:990. The odds ratio of developing cancer of a smoker vs. a non-smoker is then

$$OR(LC)_{S\ vs.\ NS} = \frac{100:900}{10:990} = 11:1$$

But a prospective study takes too long, so an alternate way of studying the problem is to do a retrospective study. Here samples of 1000 people with lung cancer, and 1000 people without lung cancer are selected at random from their respective populations. For each subject, you determine if they smoked in the past. Suppose you get the following results:

Lung Cancer	Smoker	Non-smoker
yes	810	190
no	280	720

Now you can't directly find the probability of lung cancer if you are smoker. It is NOT simply $810/(810+280)$ because you selected equal number of smokers and non-smokers while less than 30% of the population generally smokes. Unless that proportion is known, it is impossible to compute the probability of getting lung cancer if you are a smoker or non-smoker directly, and so it would seem that finding the odds of lung cancer would be impossible.

However, not all is lost. Let $P(smoker)$ represent the probability that a randomly chosen person is a smoker; then $P(\textit{non-smoker}) = 1 - P(smoker)$. Bayes' Rule[13]

$$
\begin{aligned}
P\left(\textit{lung cancer} \mid \textit{smoker}\right) &= \frac{P(\textit{smoker} \mid \textit{lung cancer})P(\textit{lung cancer})}{P(\textit{smoker})} \\
P\left(\textit{no lung cancer} \mid \textit{smoker}\right) &= \frac{P(\textit{smoker} \mid \textit{no lung cancer})P(\textit{no lung cancer})}{P(\textit{smoker})} \\
P\left(\textit{lung cancer} \mid \textit{non-smoker}\right) &= \frac{P(\textit{non-smoker} \mid \textit{lung cancer})P(\textit{lung cancer})}{P(\textit{non-smoker})} \\
P\left(\textit{no lung cancer} \mid \textit{non-smoker}\right) &= \frac{P(\textit{non-smoker} \mid \textit{no lung cancer})P(\textit{no lung cancer})}{P(\textit{non-smoker})}
\end{aligned}
$$

This doesn't appear to helpful, as P(smoker) or P(non-smoker) is unknown. But, look at the odds-ratio of getting lung cancer of a smoker vs. a non-smoker:

$$
\begin{aligned}
OR(LC)_{S \text{ vs. } NS} &= \frac{ODDS(\textit{lung cancer if smoker})}{ODDS(\textit{lung cancer if non-smoker})} \\
&= \frac{\frac{P(\textit{lung cancer} \mid \textit{smoker})}{P(\textit{no lung cancer} \mid \textit{smoker})}}{\frac{P(\textit{lung cancer} \mid \textit{non-smoker})}{P(\textit{no lung cancer} \mid \textit{non-smoker})}}
\end{aligned}
$$

If you substitute in the above expressions, you find that:

$$
OR(LC)_{S \text{ vs } NS} = \frac{\frac{P(\textit{smoker} \mid \textit{lung cancer})}{P(\textit{smoker} \mid \textit{no lung cancer})}}{\frac{P(\textit{non-smoker} \mid \textit{lung cancer})}{P(\textit{non-smoker} \mid \textit{no lung cancer})}}
$$

which can be computed from the retro-spective study. Based on the above table, we obtain

$$
OR(LC)_{S \text{ vs } NS} = \frac{\frac{.810}{.280}}{\frac{.190}{.720}} = 11:1
$$

This symmetric in odds-ratios between prospective and retrospective studies only works in the 2x2 case for simple random sampling.

[13] See `http://en.wikipedia.org/wiki/Bayes_rule`

22.6.2 Example: Parental and student usage of recreational drugs

A study was conducted where students at a college were asked about their personal use of marijuana and if their parents used alcohol and/or marijuana.[14] The following data is a collapsed version of the table that appears in the report:

Parental	Student Usage	
Usage	Yes	No
Yes	125	85
No	94	141

This is a retrospective analysis as the students are interviewed and past behavior of parents is recorded.

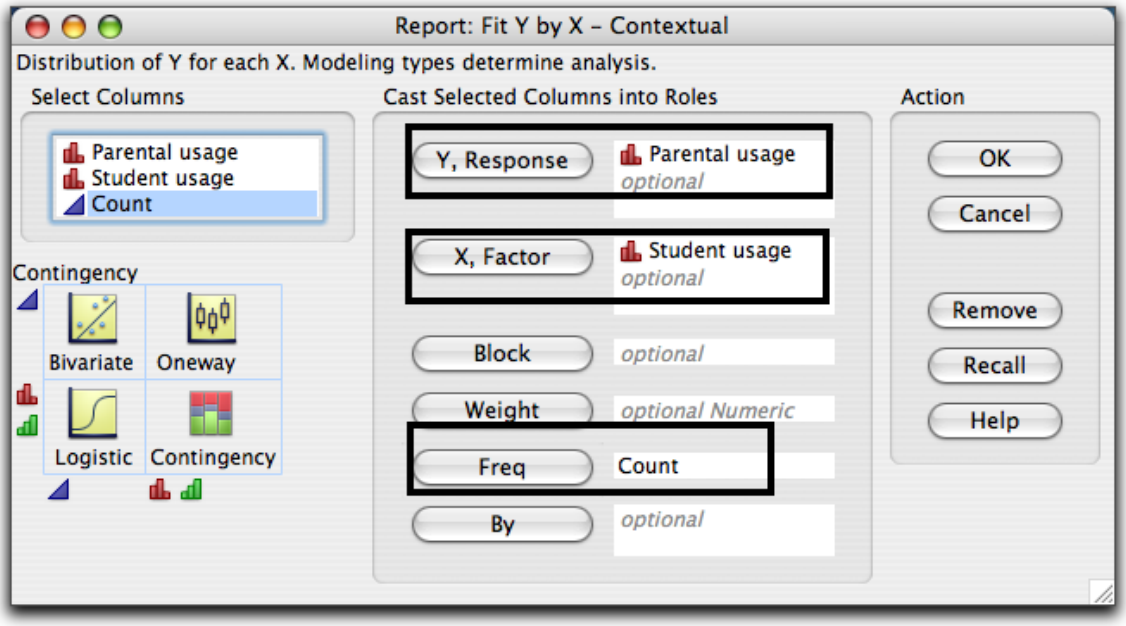
The data are entered in *JMP* in the usual format. There will be four lines, and three variables corresponding to parental usage, student usage, and the count.

marijuana-short.jmp

	Parental usage	Student usage	Count
1	yes	no	85
2	no	no	141
3	yes	yes	125
4	no	yes	94

Start using the *Analyze->Fit Y-by-X* platform:

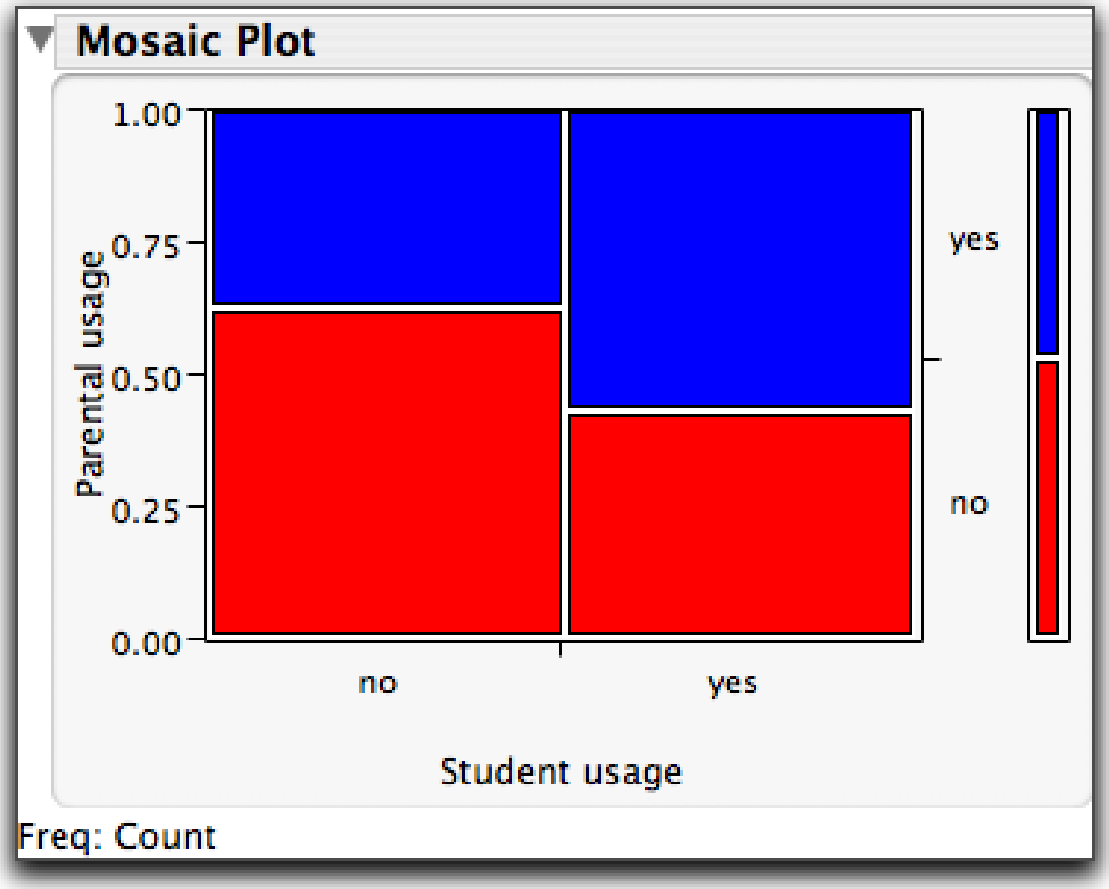
[14] "Marijuana Use in College, Youth and Society, 1979, 323-334.



but don't forget to specify the *Count* as the frequency variable. It doesn't matter which variable is entered as the X or Y variable. Note that *JMP* actually will switch from the logistic platform to the contingency platform[15] as noted by the diagram at the lower left of the dialogue box.

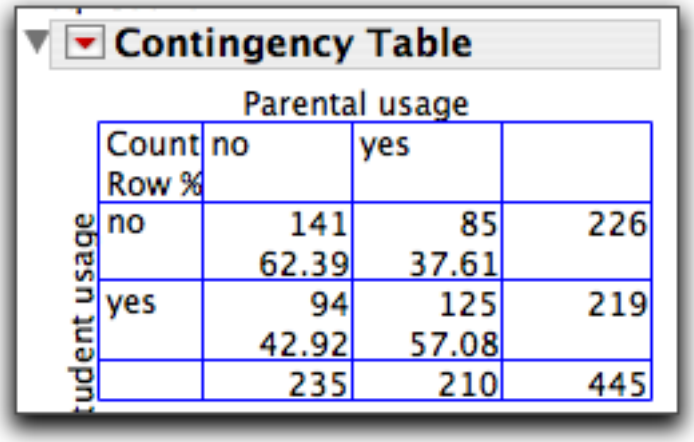
The mosaic plot shows the relative percentages in each of the student usage groups:

[15] Refer to the chapter on Chi-square tests.



The contingency table (after selecting the appropriate percentages for display from the red-triangle pop-down menu)[16]

[16] In my opinion, I would never display percentages to more than integer values. Displays such as 42.92% are just silly as they imply a precision of 1 part in 10,000 but you only have 219 subjects in the first row.

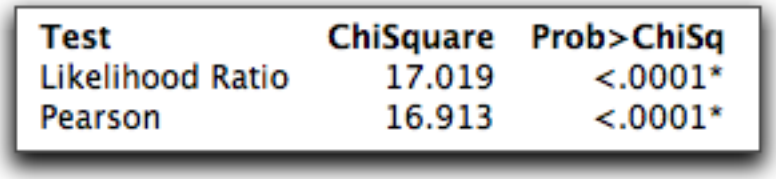


Contingency Table

Parental usage

Student usage	Count / Row %	no	yes	
	no	141 62.39	85 37.61	226
	yes	94 42.92	125 57.08	219
		235	210	445

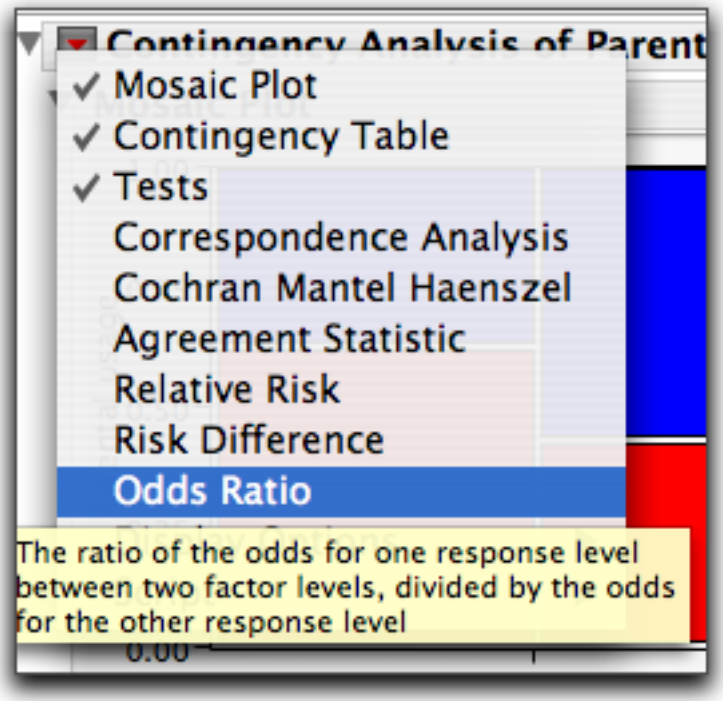
The contingency table approach tests the hypothesis of independence between the X and Y variable, i.e. is the proportion of parents who use marijuana the same for the two groups of students:



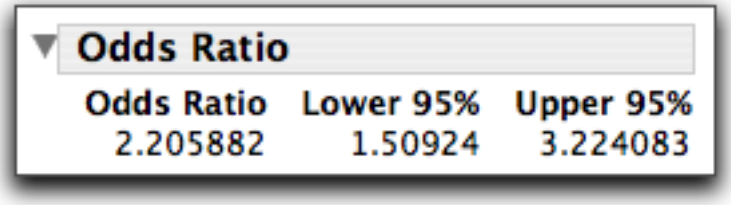
Test	ChiSquare	Prob>ChiSq
Likelihood Ratio	17.019	<.0001*
Pearson	16.913	<.0001*

As explained in the chapter on chi-square tests, there are two (asymptotically) equivalent ways to test this hypothesis – the Pearson chi-square statistic and the likelihood ratio statistic. In this case, you would come to the same conclusion.

The odds-ratio is obtained from the red-triangle at the top of the display:

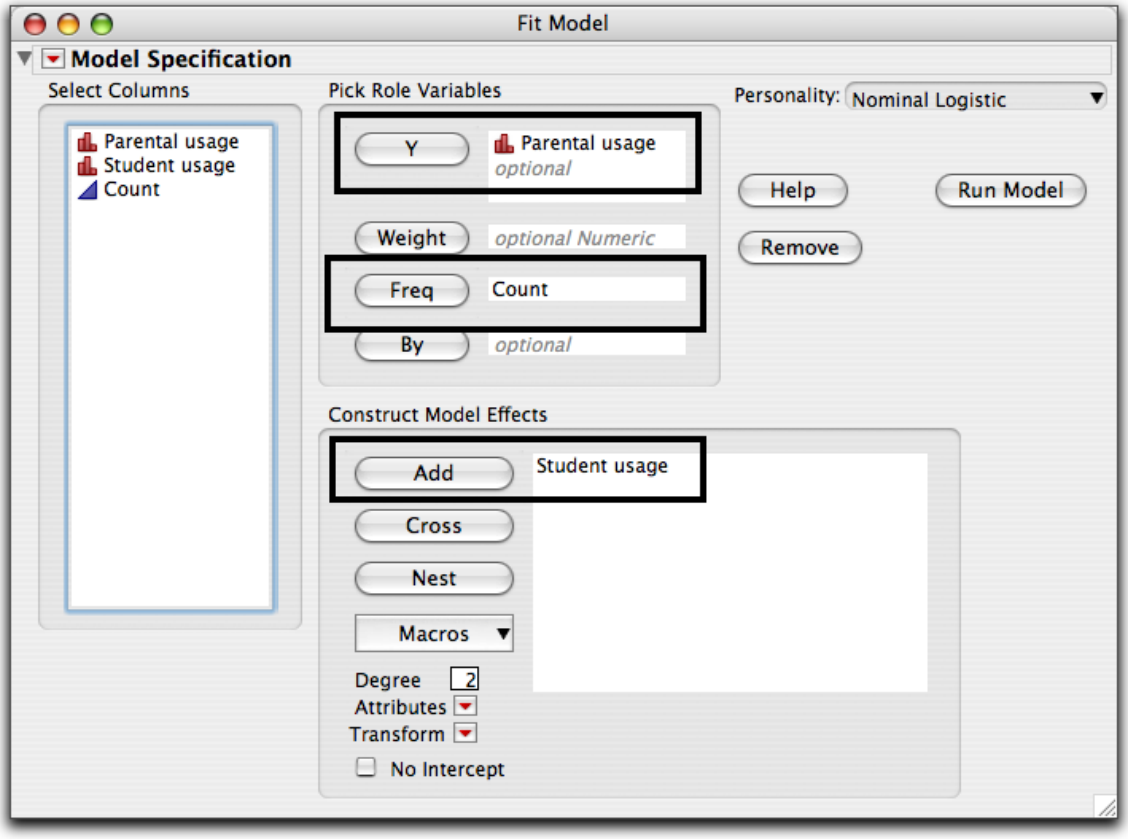


and gives:

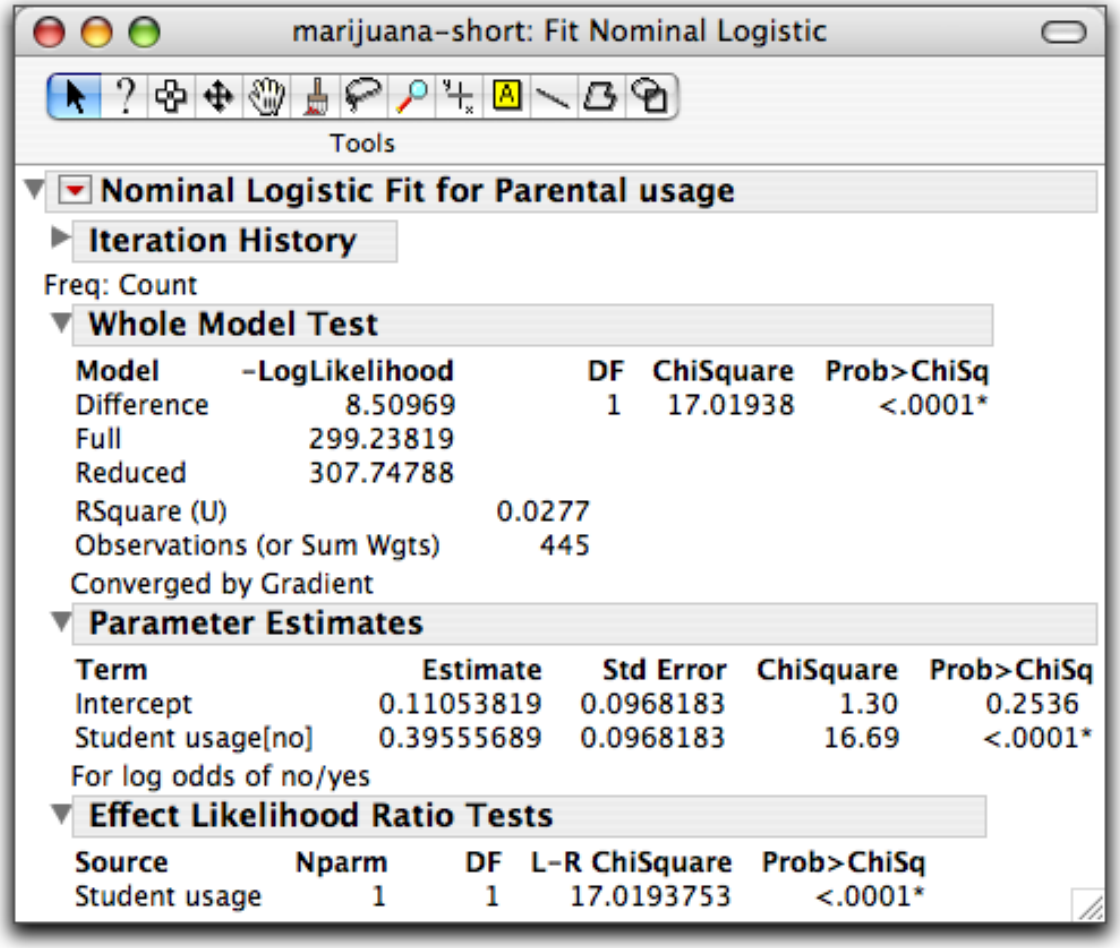


It is estimated that the odds of children using marijuana if their parents use marijuana or alcohol are about 2.2 times that of the odds of a child using marijuana for parents who don't use marijuana or alcohol. The 95% confidence interval for the odds-ratio is between 1.51 and 3.22. In this case, you would examine if the confidence interval for the odds-ratio includes the value of 1 (why?) to see if anything interesting is happening.

If the *Analyze->Fit Model* platform is used and a logistic regression is fit:

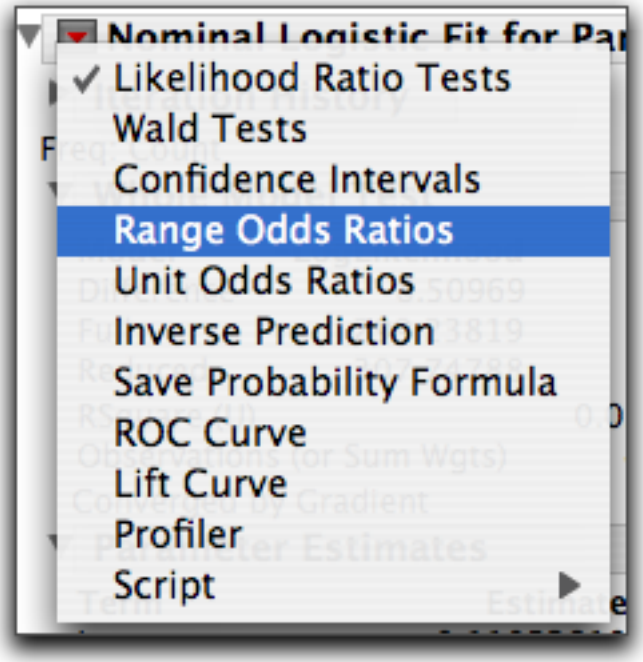


This gives the output:



The coefficient of interest is the effect of student usage on the no/yes log-odds for parental usage. The test for the effect of student usage has chi-square test value of 17.02 with a small p-value which matches the likelihood ratio test from the contingency table approach. Many packages use different codings for categorical X variables (as seen in the section on multiple regression) so you need to check the computer manual carefully to understand exactly what the coefficient measures.

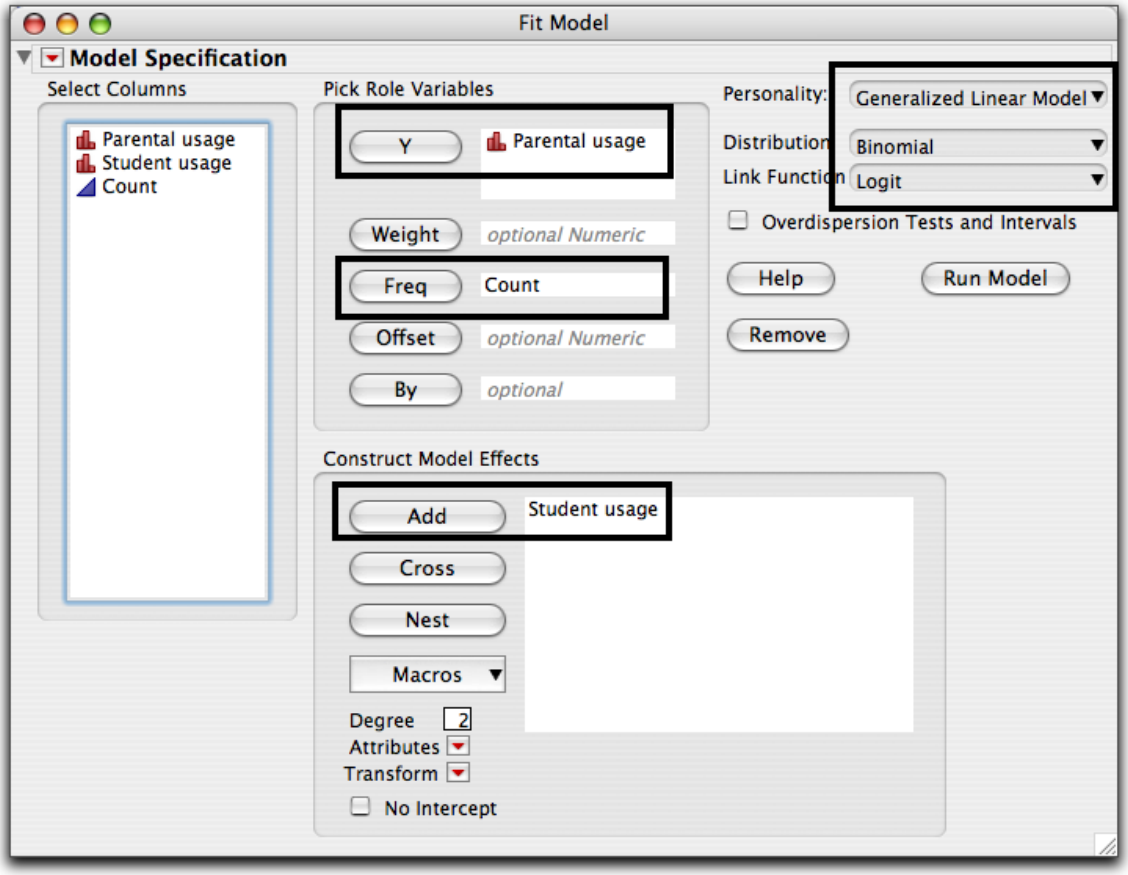
However, the odds-ratio can be found from the red-triangle pop-down menu:



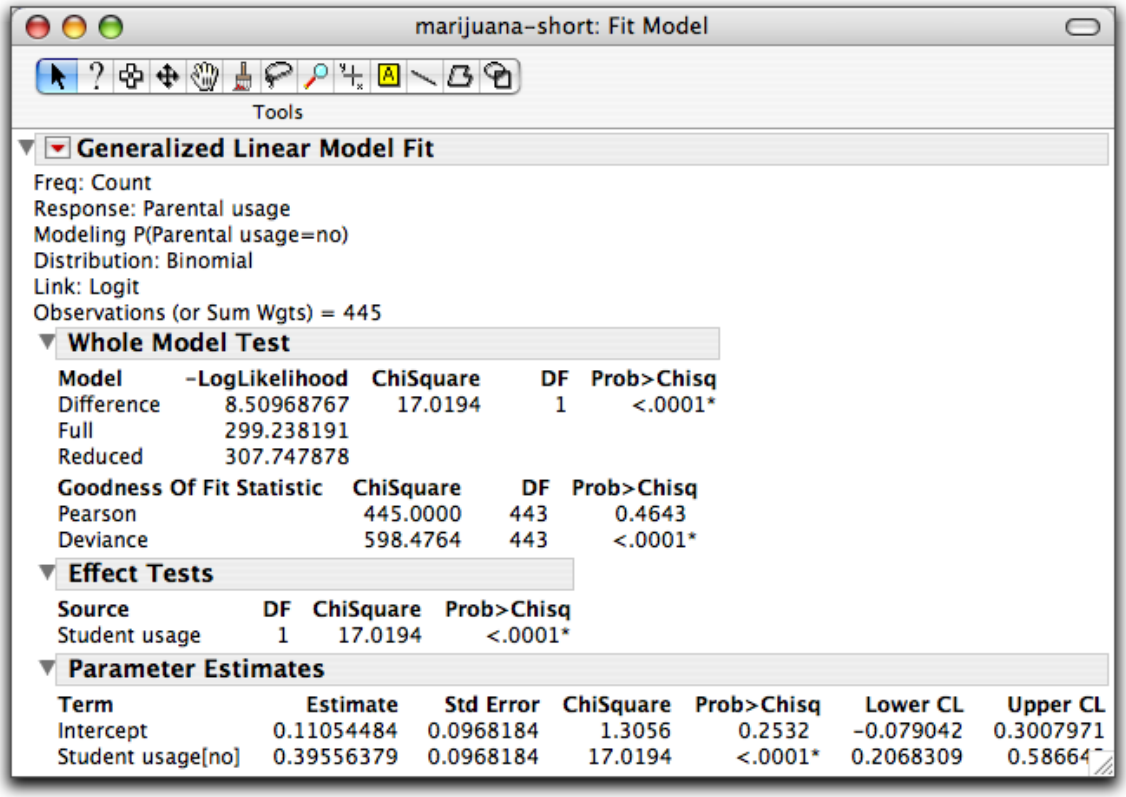
Odds Ratio	Odds Lower	Odds Upper
.	.	.
2.20585189	1.5123456	3.23259738

and matches what was seen earlier.

Finally, the *Analyze->Fit Model* platform can be used with the *Generalized Linear Model* option:


Fit Model
Model Specification
Select Columns
Parental usage
Student usage
Count
Pick Role Variables
Y
Parental usage
Weight
optional Numeric
Freq
Count
Offset
optional Numeric
By
optional
Personality:
Generalized Linear Model
Distribution
Binomial
Link Function
Logit
Overdispersion Tests and Intervals
Help
Run Model
Remove
Construct Model Effects
Add
Student usage
Cross
Nest
Macros
Degree
2
Attributes
Transform
No Intercept

This gives:



The test for a student effect has the same results as seen previously. But, ironically, gives no easy easy to compute the odds ratio. It turns out that given the parameterization used by *JMP*, the log-odds ratio is twice the coefficient of the student-usage, i.e. twice of -.3955. The odds-ratio would be found as the anti-log of this value, i.e. $e^{2\times -.3955} = .4522$ and the confidence interval for the odds-ratio can be found by anti-logging twice the confidence intervals for this coefficient, i.e. ranging from $(e^{2\times -.5866} = .31 \rightarrow e^{2\times -.2068} = .66)$.[17] These values are the inverse of the value seen earlier but this is an artefact of which category is modelled. For example, the odds ratio of $Parents_{Y\ vs.N}(student_{Y\ vs.N})$ = $\frac{1}{Parents_{N\ vs.Y}(student_{Y\ vs.N})}$ = $\frac{1}{Parents_{Y\ vs.N}(student_{N\ vs.Y})}$ = $Parents_{N\ vs.Y}(student_{N\ vs.Y})$

22.6.3 Example: Effect of selenium on tadpoles deformities

The generalization of the above to more than two levels of the X variable is straight forward and parallels the analysis of a single factor CRD ANOVA. Again, we will assume that the experimental design is a completely randomized design or simple random sample.

Selenium (Se) is an essential element required for the health of humans, animals and plants, but be-

[17] This simple relationship may not be true with other computer packages. YMMV.

comes a toxicant at elevated concentrations. The most sensitive species to selenium toxicity are oviparous (egg-laying) animals. Ecological impacts in aquatic systems are usually associated with teratogenic effects (deformities) in early life stages of oviparous biota as a result of maternal sequestering of selenium in eggs. In aquatic environments, inorganic selenium, found in water or in sediments is converted to organic selenium at the base of the food chain (e.g., bacteria and algae) and then transferred through dietary pathways to other aquatic organisms (invertebrates, fish). Selenium also tends to biomagnify up the food chain, meaning that it accumulates to higher tissue concentrations among organisms higher in the food web.

Selenium often occurs naturally in ores and can leach from mine tailings. This leached selenium can make its way to waterways and potentially contaminate organisms.

As a preliminary survey, samples of tadpoles were selected from a control site and three sites identified as low, medium, and high concentrations of selenium based on hydrologic maps and expert opinion. These tadpoles were examined, and the number that had deformities were counted.

Here is the raw data:

Site	Tadpoles	Deformed	% deformed
Control	208	56	27%
low	687	243	35%
medium	832	329	40%
high	597	283	47%

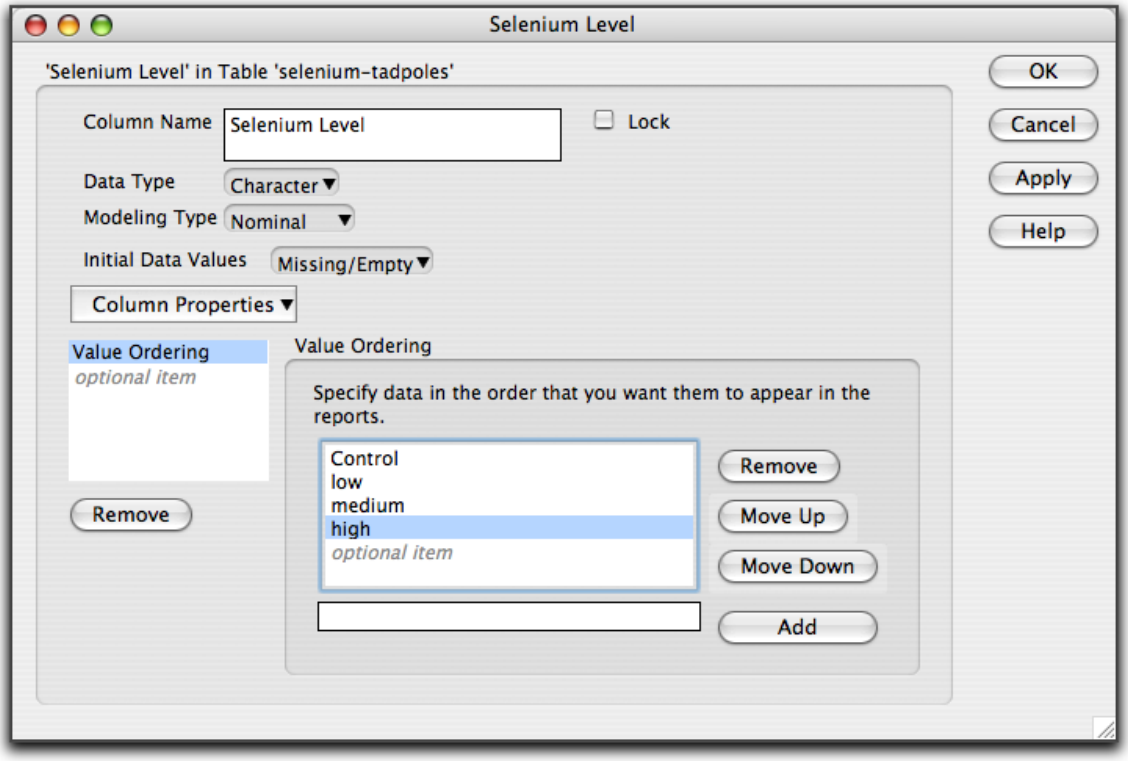
The data are entered in *JMP* in the usual fashion:

selenium-tadpoles

	Selenium Level	Total Tadpoles	Status	Count
1	low	687	Deformed	243
2	low	687	Not Deformed	444
3	high	597	Deformed	283
4	high	597	Not Deformed	314
5	medium	832	Deformed	329
6	medium	832	Not Deformed	503
7	Control	208	Deformed	56
8	Control	208	Not Deformed	152

Notice that the status of the tadpoles as deformed or not deformed is entered along with the count of each status.

As the selenium level has an ordering, it should be declared as an ordinal scale and the ordering of the values for the selenium levels should be specified using the *Column Information* → *Column Properties* → *Value Ordering* dialogue box



The hypothesis to be tested can be written in a number of equivalent ways:

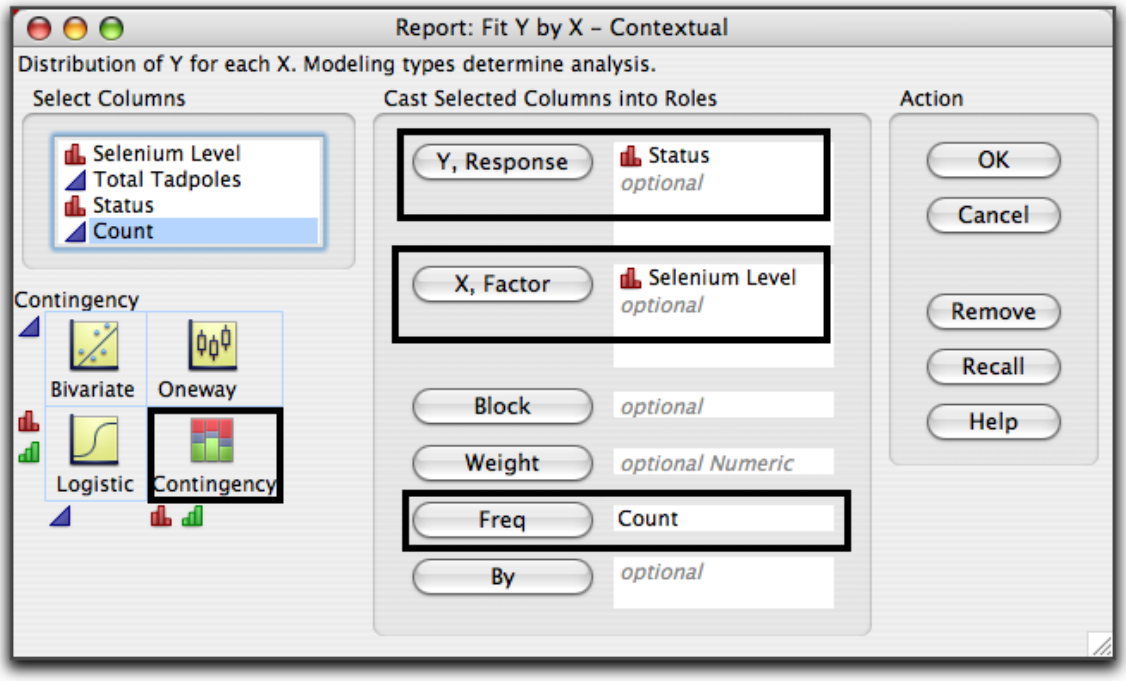
- H: p(deformity) is the same for all levels of selenium.
- H: odds(deformity) is the same for all levels of selenium.
- H: log-odds(deformity) is the same for all levels of selenium.
- H: p(deformity) is independent of the level of selenium.[18]
- H: odds(deformity) is independent of the level of selenium.
- H: log-odds(deformity) is independent of the level of selenium.

[18]The use of *independent* in the hypothesis is a bit old-fashioned and not the same as statistical independence.

- H: $p_c(D) = p_L(D) = p_M(D) = p_H(D)$ where $p_L(D)$ is the probability of deformities at low doses, etc.

There are again several ways in which this data can be analyzed.

Start with the *Analyze->Fit Y-by-X* platform:



This will give a standard contingency table analysis (see chapter on chi-square tests).

The mosaic plot:

seems to show an increasing trend in deformities with increasing selenium levels. It is a pity that *JMP* doesn't display any measure of precision (such *se* bars or confidence intervals) on this plot.

The contingency table (with suitable percentages shown[19])

[19] I would display percentages to the nearest integer. Unfortunately, there doesn't appear to be an easy way to control this in *JMP*.

Contingency Table

Selenium Level	Count Row %	Status: Deformed	Status: Not Deformed	
	Control	56 26.92	152 73.08	208
	low	243 35.37	444 64.63	687
	medium	329 39.54	503 60.46	832
	high	283 47.40	314 52.60	597
		911	1413	2324

also gives the same impression.

A formal test for equality of proportion of deformations across all levels of the factor gives the following test statistics and *p*-values:

Test	ChiSquare	Prob>ChiSq
Likelihood Ratio	34.684	<.0001*
Pearson	34.279	<.0001*

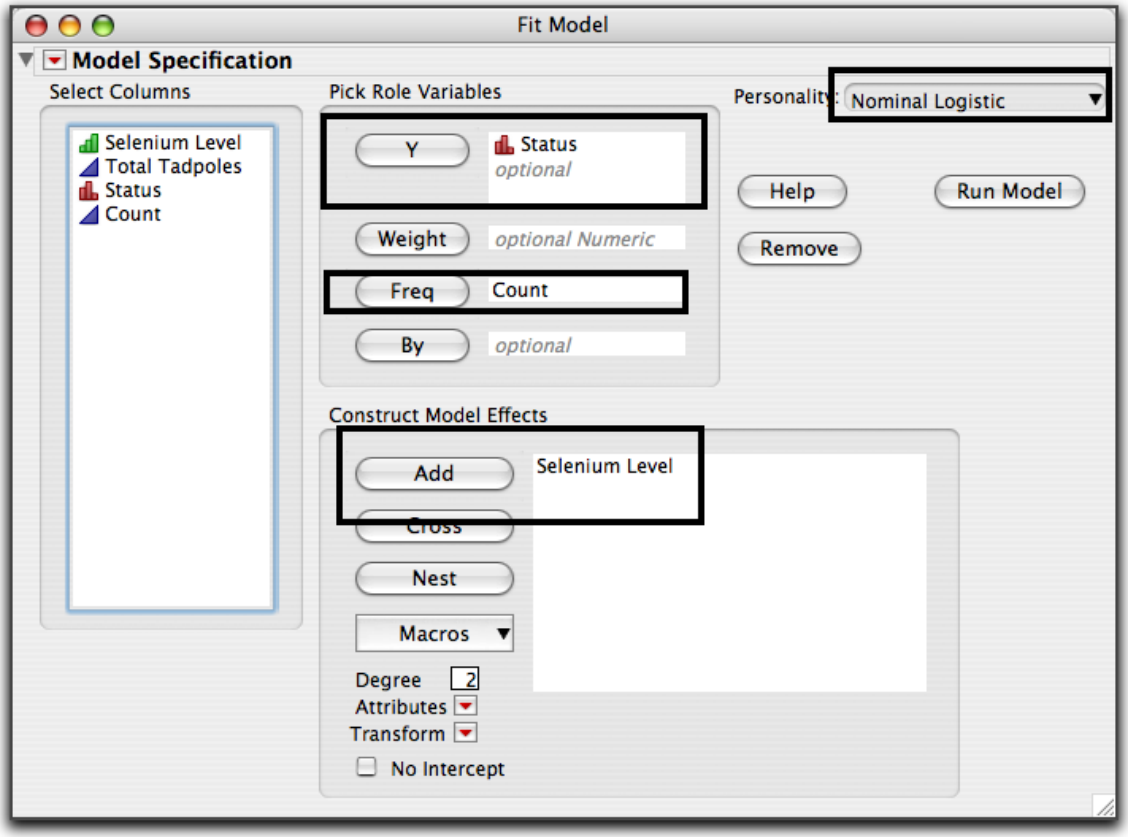
There are two common test-statistics. The *Pearson chi-square* test-statistic which examines the difference between observed and expected counts (see chapter on chi-square tests), and the likelihood-ratio test which compares the model when the hypothesis is true vs. the model when the hypothesis is false. Both are asymptotically equivalent. There is strong evidence against the hypothesis of equal proportions of deformities.

Unfortunately, most contingency table analyses stop here. A naked *p*-value which indicates that there is evidence of a difference but does not tell you where the differences might lie, is not very informative! In the same way that ANOVA must be followed by a comparison of the mean among the treatment levels, this test should be followed by a comparison of the proportion of deformities among the factor levels.

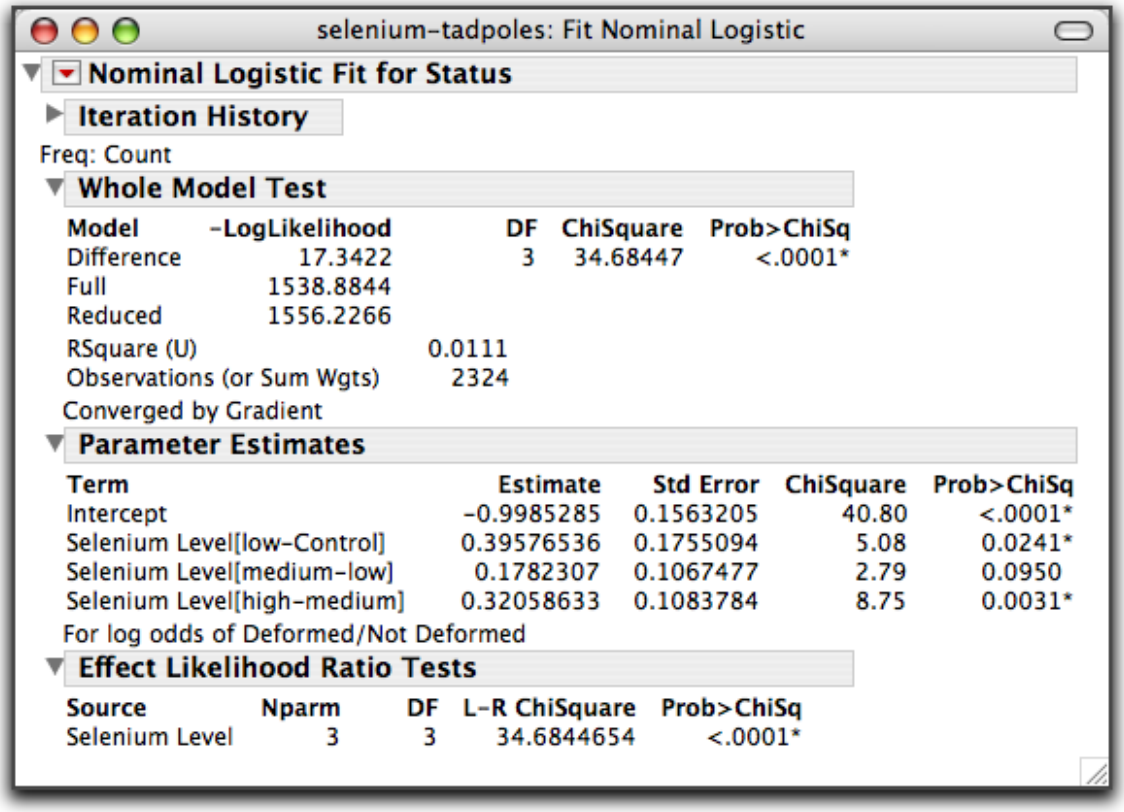
Logistic regression methods will enable us to estimate the relative odds of deformities among the various

classes.

Start with the *Analyze->Fit Model* platform:



This gives the output:



selenium-tadpoles: Fit Nominal Logistic

Nominal Logistic Fit for Status

Iteration History

Freq: Count

Whole Model Test

Model	-LogLikelihood	DF	ChiSquare	Prob>ChiSq
Difference	17.3422	3	34.68447	<.0001*
Full	1538.8844			
Reduced	1556.2266			

RSquare (U)	0.0111
Observations (or Sum Wgts)	2324

Converged by Gradient

Parameter Estimates

Term	Estimate	Std Error	ChiSquare	Prob>ChiSq
Intercept	-0.9985285	0.1563205	40.80	<.0001*
Selenium Level[low-Control]	0.39576536	0.1755094	5.08	0.0241*
Selenium Level[medium-low]	0.1782307	0.1067477	2.79	0.0950
Selenium Level[high-medium]	0.32058633	0.1083784	8.75	0.0031*

For log odds of Deformed/Not Deformed

Effect Likelihood Ratio Tests

Source	Nparm	DF	L-R ChiSquare	Prob>ChiSq
Selenium Level	3	3	34.6844654	<.0001*

First, the *Effect Tests* tests the hypothesis of equality of the proportion of defectives among the four levels of selenium. The test-statistic and *p*-value match that seen earlier, so there is good evidence of a difference among the deformity proportions among the various levels.

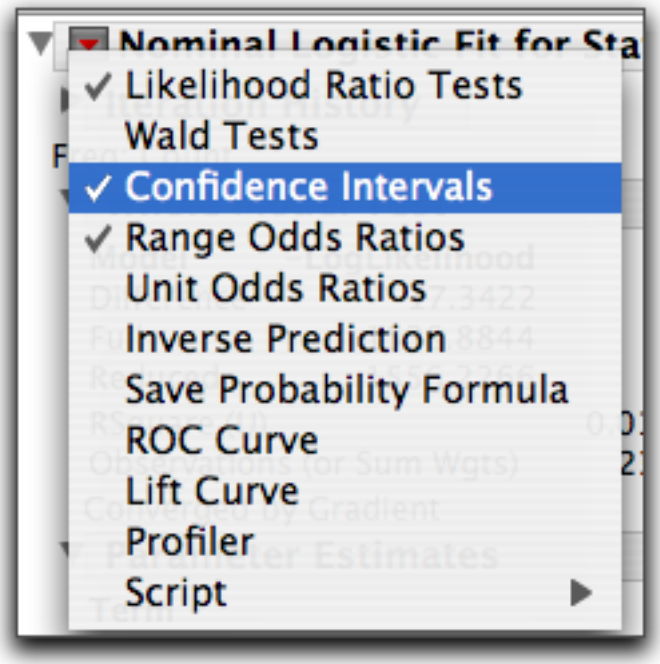
At this point in a ANOVA, a multiple comparison procedure (such a Tukey's HSD) would be used to examine which levels may have different means from the other levels. There is no simple equivalent for logistic regression implemented in *JMP*.[20] It would be possible to use a simple Bonferonni correction if the number of groups is small.

JMP provides some information on comparison among the levels. In the *Parameter Estimates* section, it presents comparisons of the proportion of defectives among the successive levels of selenium.[21] The estimated difference in the log-odds of deformed for the *low* vs. *control* group is .39 (*se* .18). The associated *p*-value for no difference in the proportion of deformed is .02 which is less than the $\alpha = .05$ levels so there is evidence of a difference in the proportion of deformed between these two levels.

By requesting the confidence interval and the odds-ratio these can be transformed to the odds-scale (rather than the log-odds) scale.

[20] This is somewhat puzzling as the theory should be straight forward.

[21] This is purely a function of the internal coding used by *JMP*. Other packages may use different coding. YMMV.



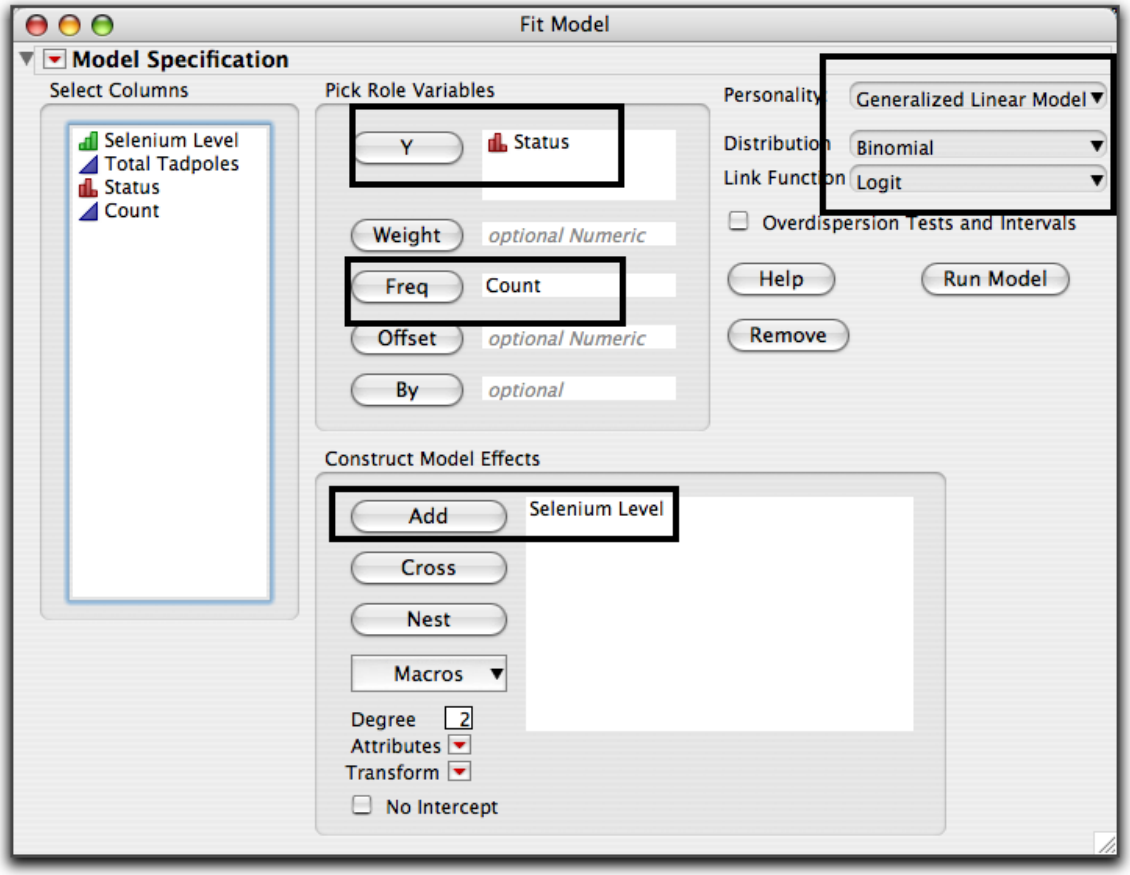
Parameter Estimates

Term	Estimate	Std Error	ChiSquare	Prob>ChiSq	Odds Ratio	Lower 95%	Upper 95%	Odds Lower	Odds Upper
Intercept	-0.9985285	0.1563205	40.80	<.0001*	.	-1.3127906	-0.6987416	.	.
Selenium Level[low-Control]	0.39576536	0.1755094	5.08	0.0241*	1.48552071	0.05674711	0.74578998	1.05838812	2.10810613
Selenium Level[medium-low]	0.1782307	0.1067477	2.79	0.0950	1.195101	-0.0309909	0.38745228	0.96948442	1.47322265
Selenium Level[high-medium]	0.32058633	0.1083784	8.75	0.0031*	1.37793545	0.10816866	0.533004	1.11423566	1.70404358

For log odds of Deformed/Not Deformed

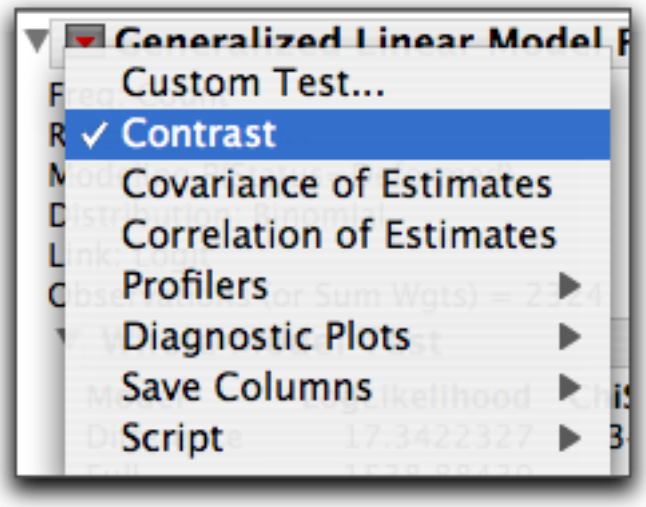
Unfortunately, there is no simple mechanism to do a more general contrasts in this variant of the *Analyze->Fit Model* platform.

The *Generalized Linear Model* platform in the *Analyze->Fit Model* platform gives more options:



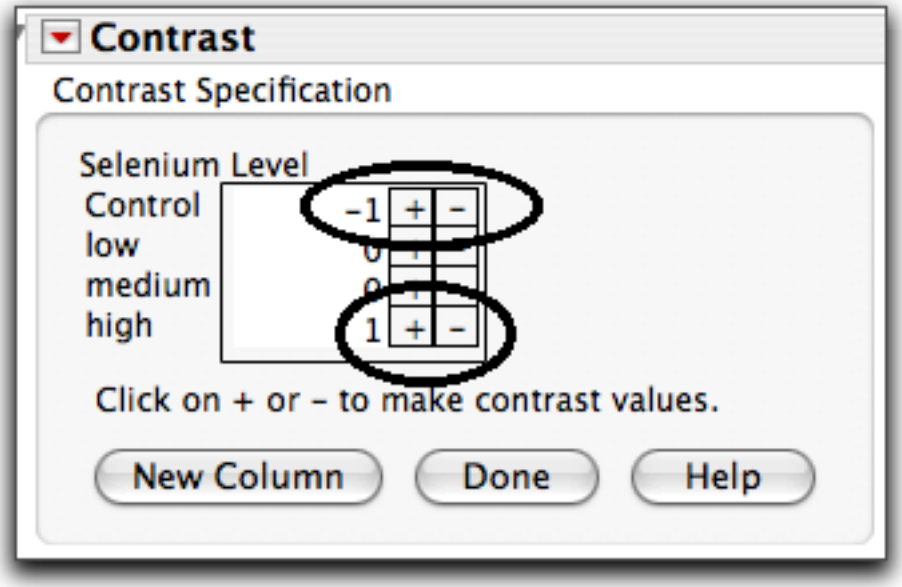
The output you get is very similar to what was seen previously. Suppose that a comparison between the proportions of deformities between the *high* and *control* levels of selenium are wanted.

Use the red-triangle pop-down menu to select the *Contrast* options:

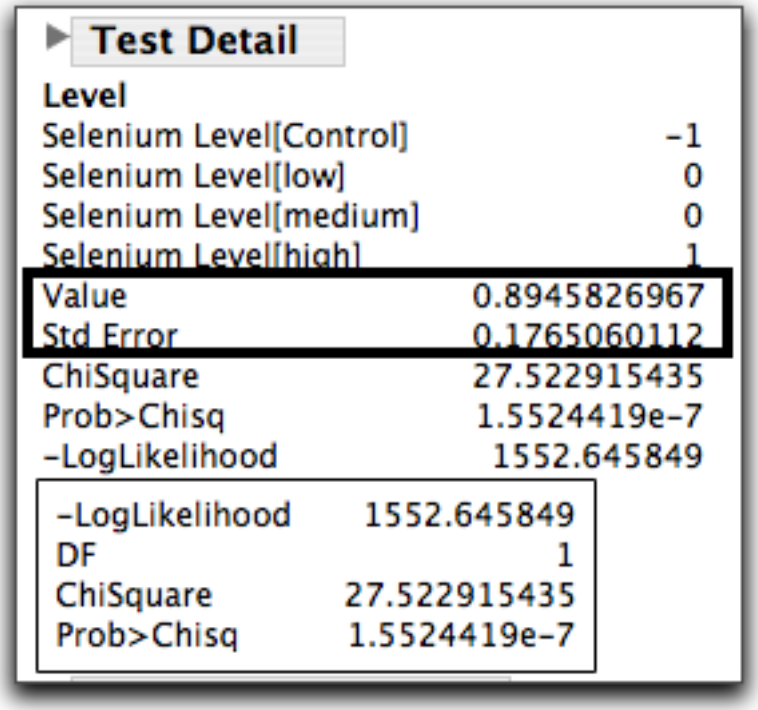


Then select the radio button for comparisons among selenium levels:

Click on the $+$ and $-$ to form the contrast. Here you are interested in $LO_{high} - LO_{control}$ where the LO are the log-odds for a deformity.


Contrast
Contrast Specification
Selenium Level
Control
low
medium
high
-1
+
-
1
+
-
Click on + or - to make contrast values.
New Column
Done
Help

This gives:


Test Detail
Level
Selenium Level[Control] -1
Selenium Level[low] 0
Selenium Level[medium] 0
Selenium Level[high] 1
Value 0.8945826967
Std Error 0.1765060112
ChiSquare 27.522915435
Prob>Chisq 1.5524419e-7
-LogLikelihood 1552.645849
-LogLikelihood 1552.645849
DF 1
ChiSquare 27.522915435
Prob>Chisq 1.5524419e-7

The estimated log-odds ratio is .89 (*se* .18). This implies that the odds-ratio of deformity is $e^{.89} = 2.43$, i.e. the odds of deformity are 2.43 greater in the high selenium site than the control site. The *p*-value is well below $\alpha = .05$ so there is strong evidence that this effect is real. It is possible to compute the *se* of the odds-ratio using the Delta method – pity that *JMP* doesn't do this directly.[22] An approximate 95% confidence interval for the log-odds ratio could be found using the usual rule of $estimate \pm 2se$. The 95% confidence interval for the odd-ratio would be found by taking anti-logs of the end points.

This procedure could then be repeated for any contrast of interest.

22.7 Example: Pet fish survival as function of covariates - Multiple categorical predictors

There is no conceptual problem in having multiple categorical X variables. Unlike the case of a single categorical X variable, there is no simple contingency table approach. However, in more advanced classes, you will learn about a technique called log-linear modeling that can often be used for these types of tables.

Again, before analyzing any dataset, ensure that you understand the experimental design. In these notes, it is assumed that the design is completely randomized design or a simple random sample. If your design is more complex, please seek suitable help.

A fish is a popular pet for young children – yet the survival rate of many of these fish is likely poor. What factors seem to influence the survival probabilities of pet fish?

A large pet store conducted a customer follow-up survey of purchasers of pet fish. A number of customers were called and asked about the hardness of the water used for the fish (soft, medium, or hard), where the fish was kept which was then classified into cool or hot locations within the living dwelling, if they had previous experience with pet fish (yes or no), and if the pet fish was alive six months after purchase (yes or no).

Here is the raw data[23]:

[22] For those so inclined, if $\widehat{\theta}$ is the estimator with associated *se*, then the *se* of $e^{\widehat{\theta}}$ is found as $se(e^{\widehat{\theta}}) = se(\widehat{\theta}) \times e^{\widehat{\theta}}$. In this case, the *se* of the odd-ratio would be $.18 \times e^{.89} = .44$.

[23] Taken from Cox and Snell, Analysis of Binary Data

Softness	Temp	PrevPet	N	Alive
h	c	n	89	37
h	h	n	67	24
m	c	n	102	47
m	h	n	70	23
s	c	n	106	57
s	h	n	48	19
h	c	y	110	68
h	h	y	72	42
m	c	y	116	66
m	h	y	56	33
s	c	y	116	63
s	h	y	56	29

There are three factors in this study:

- **Softness** with three levels (*h, m* or *s*);
- **Temperature** with two levels (*c* or *h*);
- **Previous ownership** with two levels (*y* or *n*).

This a factorial experiment because all 12 treatment combinations appear in the experiment.

The experimental unit is the household. The observational unit is also the household. There is no pseudo-replication.

The randomization structure is likely complete. It seems unlikely that people would pick particular individual fish depending on their water hardness, temperature, or previous history of pet ownership.

The response variable is the Alive/Dead status at the end of six months. This is a discrete binary outcome. For example, in the first row of the data table, there were 37 households where the fish was still alive after 6 months and therefore $89 - 37 = 52$ households where the fish had died somewhere in the 6 month interval.

One way to analyze this data would be to compute the proportion of households that had fish alive after six months, and then use a three-factor CRD ANOVA on the estimated proportions. Because each treatment combination is based on a different number of trial (ranging from 48 to 116) which implies that the variance of the estimated proportion is not constant. This violates (but not likely too badly) one of the assumptions of ANOVA – that of constant variance in each treatment combination. Also, this seems to throw away data, as these 1000 observations are basically collapsed into 12 cells.

Because the outcome is a discrete binary response and each trial within each treatment is independent, a logistic regression (or generalized linear model) approach can be used.

The data is available in the *JMP* data file *fishsurvive.jmp* available in the Sample Program Library at `http://www.stat.sfu.ca/~cschwarz/Stat-650/Notes/MyPrograms`. Here is the data file:

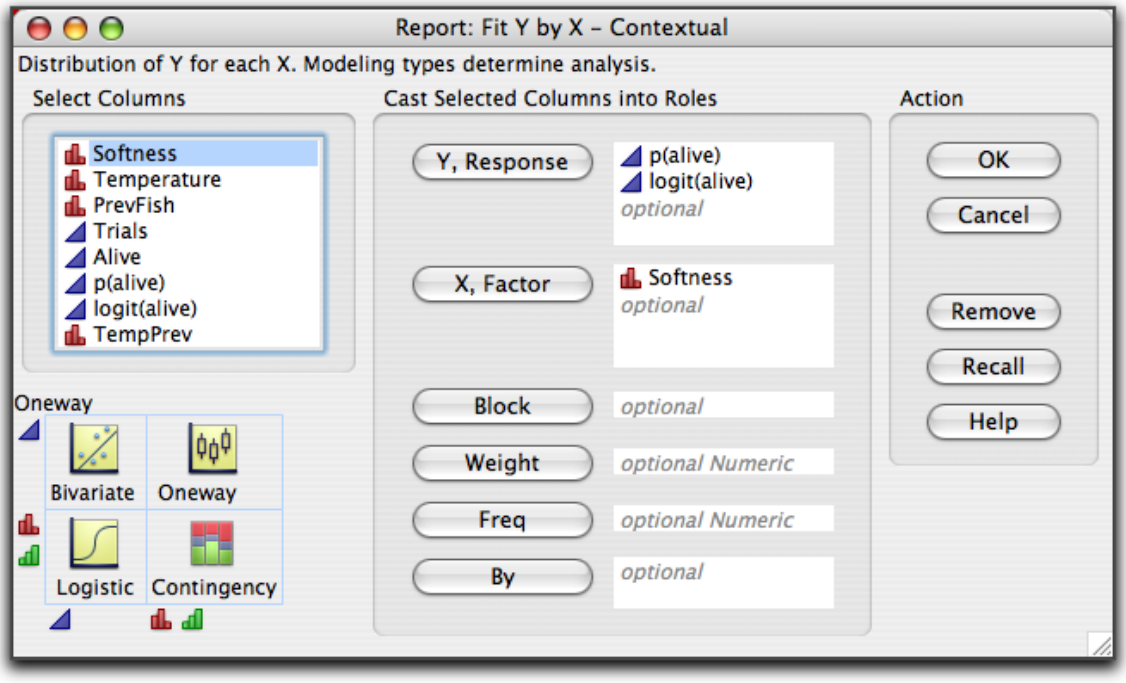
	Softness	Temperatu	Prev Fish	Trials	Alive
1	h	c	n	89	37
2	h	h	n	67	24
3	m	c	n	102	47
4	m	h	n	70	23
5	s	c	n	106	57
6	s	h	n	48	19
7	h	c	y	110	68
8	h	h	y	72	42
9	m	c	y	116	66
10	m	h	y	56	33
11	s	c	y	116	63
12	s	h	y	56	29

To begin with, construct some profile plots to get a feel for what is happening. Create new variables corresponding to the proportion of fish alive and its logit[24]. These are created using the formula editor of *JMP* in the usual fashion. Also, for reasons which will become apparent in a few minutes, create a variable which is the concatenation of the *Temperature* and *Previous Ownership* factor levels. This gives:

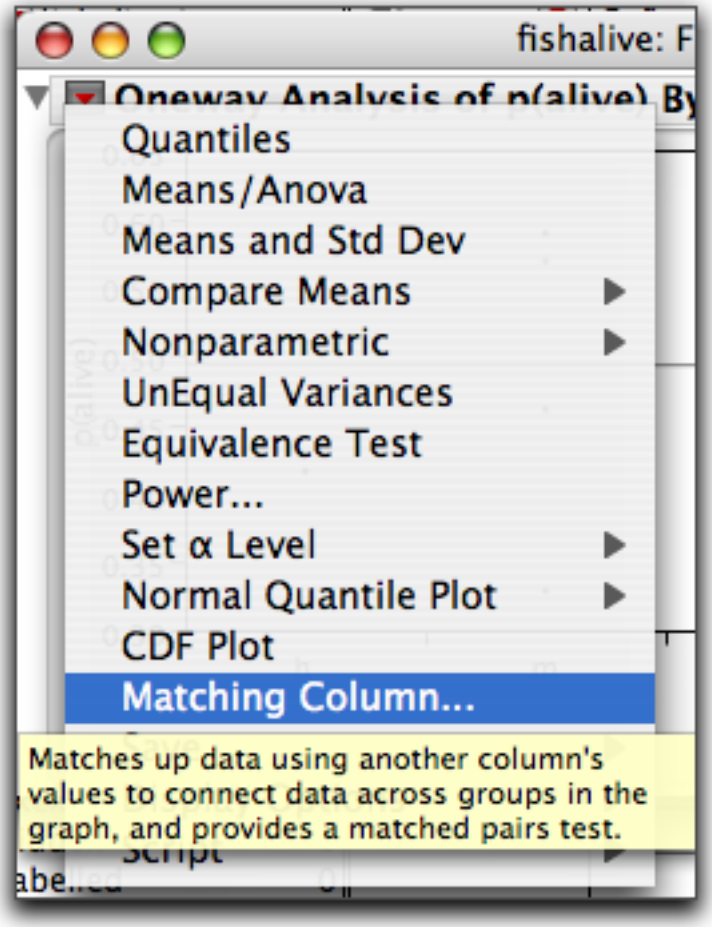
[24] Recall that $logit(p) = \log\left(\frac{p}{1-p}\right)$.

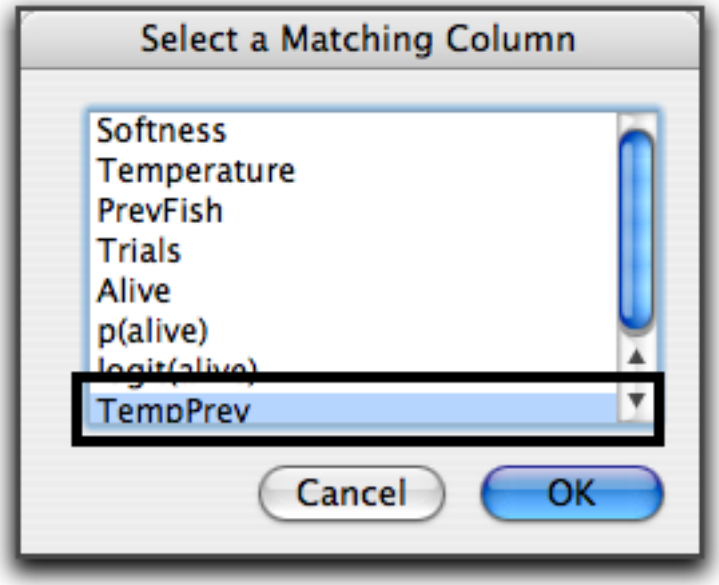
	Softness	Temperatu	Prev Fish	Trials	Alive	p(alive)	logit (alive)	Temp Prev
1	h	c	n	89	37	0.42	-0.34	c-n
2	h	h	n	67	24	0.36	-0.58	h-n
3	m	c	n	102	47	0.46	-0.16	c-n
4	m	h	n	70	23	0.33	-0.71	h-n
5	s	c	n	106	57	0.54	0.15	c-n
6	s	h	n	48	19	0.40	-0.42	h-n
7	h	c	y	110	68	0.62	0.48	c-y
8	h	h	y	72	42	0.58	0.34	h-y
9	m	c	y	116	66	0.57	0.28	c-y
10	m	h	y	56	33	0.59	0.36	h-y
11	s	c	y	116	63	0.54	0.17	c-y
12	s	h	y	56	29	0.52	0.07	h-y

Now use the *Analyze->Fit Y-by-X* platform and specify that the $p(alive)$ or $logit(alive)$ is the response variable, with the *WaterSoftness* as the factor.

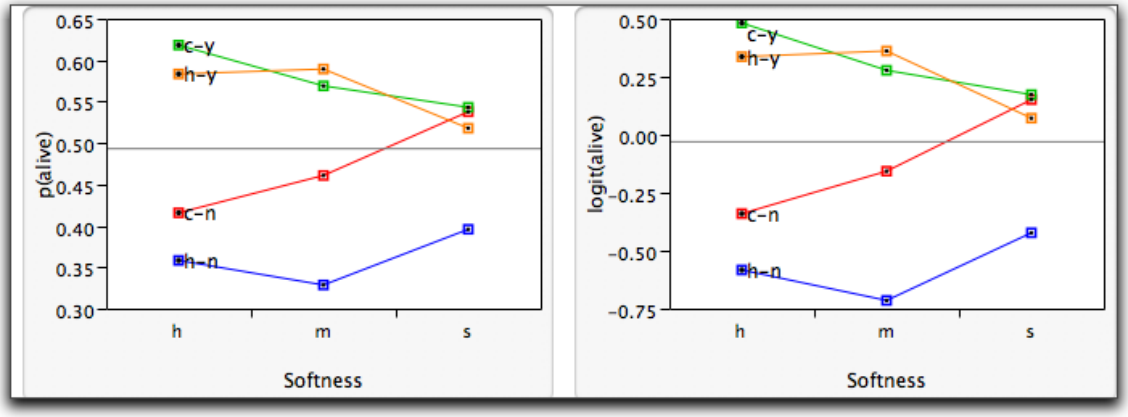


Then specify a matching column for the plot (do this on both plots) using the concatenated variable defined above.





This creates the two profile plots[25]:



The profile plots seem to indicate that the $p(alive)$ tends to increase with water softness if this is a first time pet owner, and (ironically) tends to decrease if a previous pet owner. Of course without standard error bars, it is difficult to tell if these trends are real or not. The sample sizes in each group are around 100 households. If $p(alive) = .5$, then the approximate size of a standard error is $se = \sqrt{\frac{.5(.5)}{100}} = .05$ or the approximate 95% confidence intervals are $\pm .1$. It looks as if any trends will be hard to detect with the sample sizes used in this experiment.

[25] To get the labels on the graph, set the concatenated variable to be a label variable and the rows corresponding to the *h* softness level to be labeled rows.

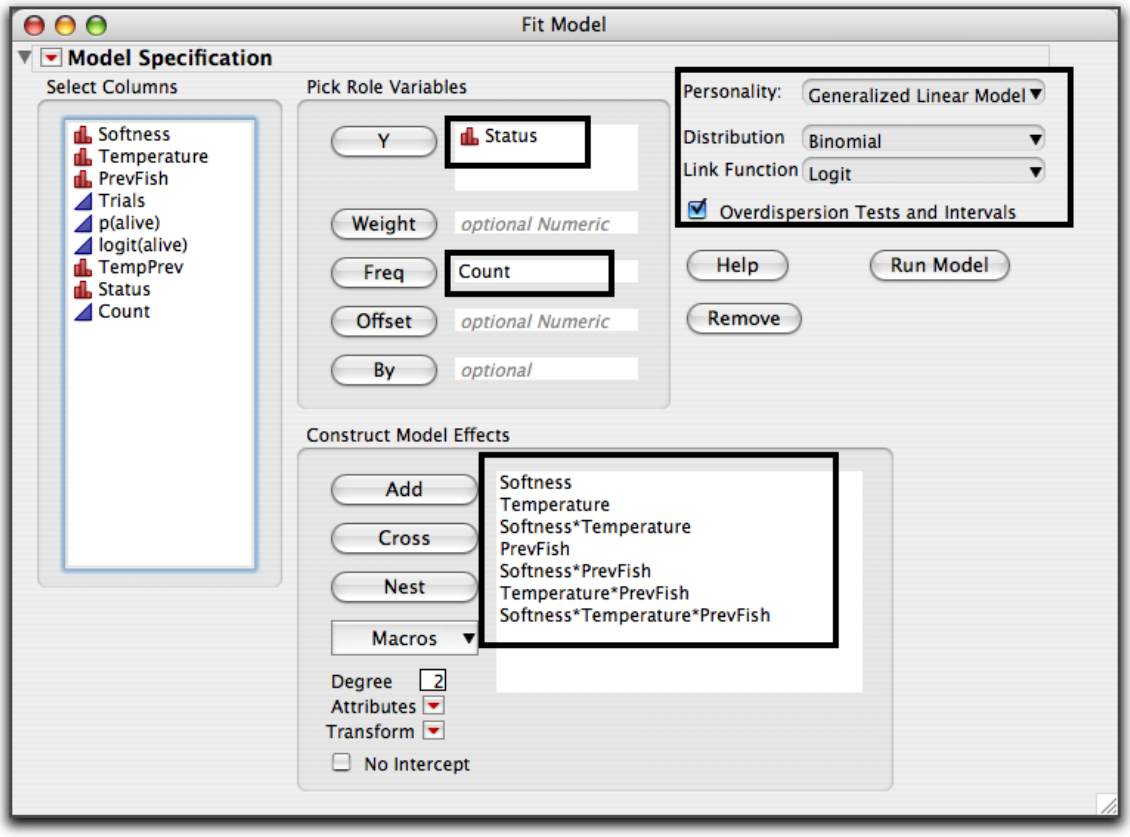
In order to fit a logistic-regression model, you must first create new variable representing the number *Dead* in each trial[26], and then stack[27] the *Alive* and *Dead* variables, label the columns as *Status* and the *Count* of each *Status* to give the final table:

	Softnes	Temperatu	PrevFi	Trials	p (alive)	logit (alive)	TempPrev	Status	Count
1	h	c	n	89	0.42	-0.34	c-n	Alive	37
2	h	c	n	89	0.42	-0.34	c-n	Dead	52
3	h	h	n	67	0.36	-0.58	h-n	Alive	24
4	h	h	n	67	0.36	-0.58	h-n	Dead	43
5	m	c	n	102	0.46	-0.16	c-n	Alive	47
6	m	c	n	102	0.46	-0.16	c-n	Dead	55
7	m	h	n	70	0.33	-0.71	h-n	Alive	23
8	m	h	n	70	0.33	-0.71	h-n	Dead	47
9	s	c	n	106	0.54	0.15	c-n	Alive	57
10	s	c	n	106	0.54	0.15	c-n	Dead	49
11	s	h	n	48	0.40	-0.42	h-n	Alive	19
12	s	h	n	48	0.40	-0.42	h-n	Dead	29
13	h	c	y	110	0.62	0.48	c-y	Alive	68
14	h	c	y	110	0.62	0.48	c-y	Dead	42
15	h	h	y	72	0.58	0.34	h-y	Alive	42
16	h	h	y	72	0.58	0.34	h-y	Dead	30
17	m	c	y	116	0.57	0.28	c-y	Alive	66
18	m	c	y	116	0.57	0.28	c-y	Dead	50
19	m	h	y	56	0.59	0.36	h-y	Alive	33
20	m	h	y	56	0.59	0.36	h-y	Dead	23
21	s	c	y	116	0.54	0.17	c-y	Alive	63
22	s	c	y	116	0.54	0.17	c-y	Dead	53
23	s	h	y	56	0.52	0.07	h-y	Alive	29
24	s	h	y	56	0.52	0.07	h-y	Dead	27

Whew! Now we can finally fit a model to the data and test for various effects. In *JMP* 6.0 and later, there are two ways to proceed (both give the same answers, but the generalized linear model platform gives a richer set of outputs). Use the *Analyze->Fit Model* platform:

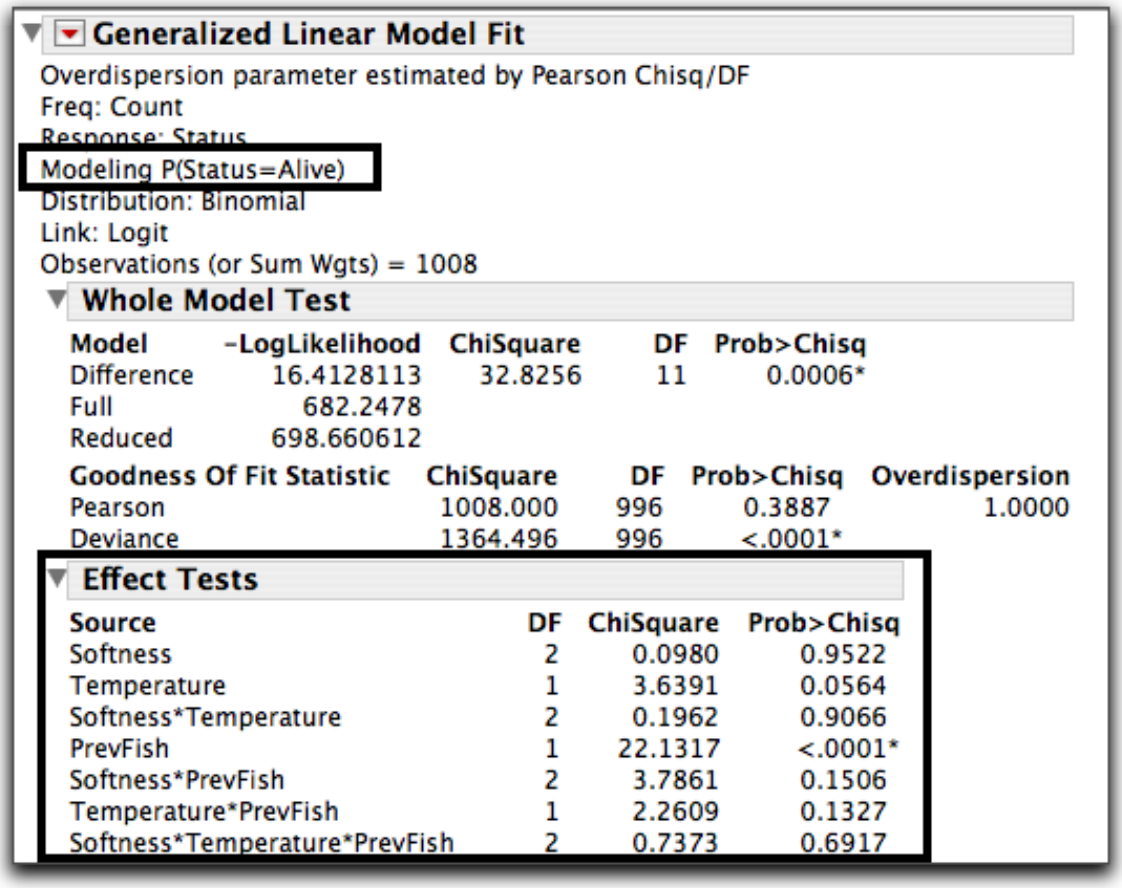
[26] Use a formula to subtract the number alive from the number of trials.

[27] Use the *Tables->Stack* command.



Notice that the response variable is *Status* and that the frequency variable is the *Count* of the number of times each status occurs. The model effects box is filled with each factors effect, and the second and third order interactions.

This gives the following output:



Generalized Linear Model Fit

Overdispersion parameter estimated by Pearson Chisq/DF
Freq: Count
Response: Status
Modeling P(Status=Alive)
Distribution: Binomial
Link: Logit
Observations (or Sum Wgts) = 1008

Whole Model Test

Model	-LogLikelihood	ChiSquare	DF	Prob>Chisq
Difference	16.4128113	32.8256	11	0.0006*
Full	682.2478			
Reduced	698.660612			

Goodness Of Fit Statistic	ChiSquare	DF	Prob>Chisq	Overdispersion
Pearson	1008.000	996	0.3887	1.0000
Deviance	1364.496	996	<.0001*	

Effect Tests

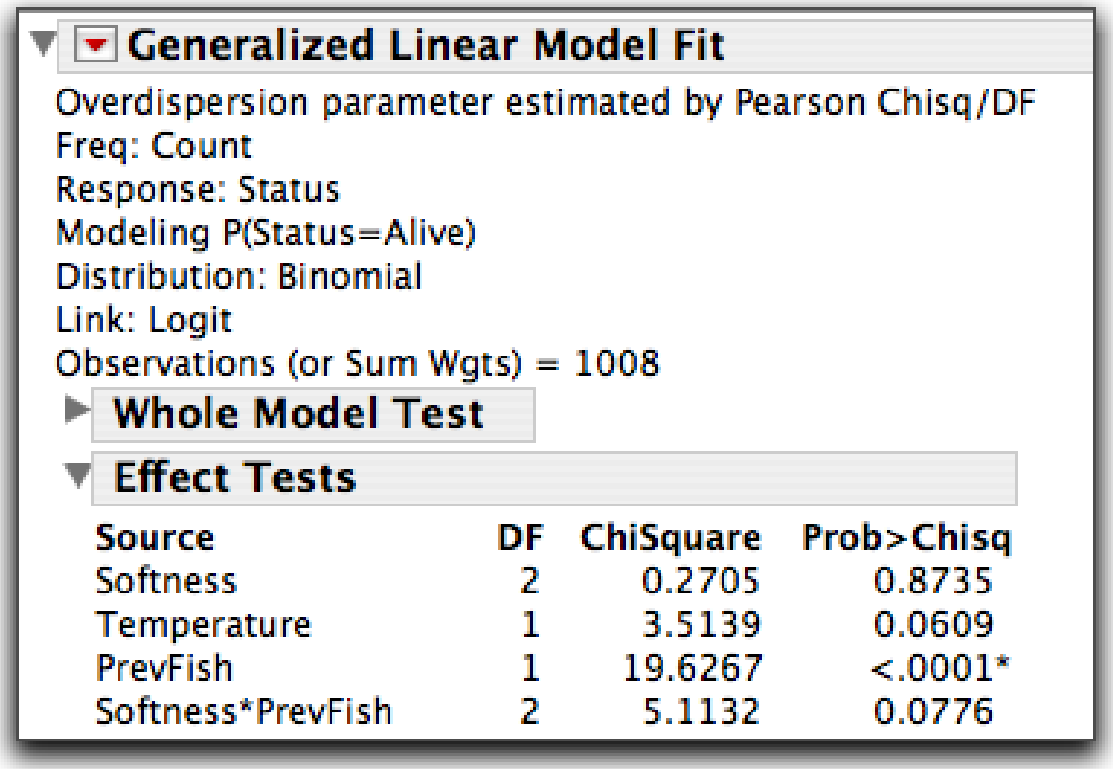
Source	DF	ChiSquare	Prob>Chisq
Softness	2	0.0980	0.9522
Temperature	1	3.6391	0.0564
Softness*Temperature	2	0.1962	0.9066
PrevFish	1	22.1317	<.0001*
Softness*PrevFish	2	3.7861	0.1506
Temperature*PrevFish	1	2.2609	0.1327
Softness*Temperature*PrevFish	2	0.7373	0.6917

Check to see exactly what is being modeled. In this case, it is the probability of the **first** level of the responses, $logit(alive)$.

Then examine the effect tests. Just as in ordinary ANOVA modeling, start with the most complex term, and work backwards successively eliminating terms until nothing more can be eliminated. The third-order interaction is not statistically significant. Eliminate this term from the *Analyze->Fit Model* dialog box, and refit using only main effects and two factor interactions.[28]

Successive terms were dropped to give the final model:

[28] Just like regular ANOVA, you can't examine the *p*-values of lower order interaction terms if a higher order interaction is present. In this case, you can't look at the *p*-values for the second order interaction when the third order interaction is present in the model. You must first refit the model after the third order interaction is dropped.



Generalized Linear Model Fit

Overdispersion parameter estimated by Pearson Chisq/DF
Freq: Count
Response: Status
Modeling P(Status=Alive)
Distribution: Binomial
Link: Logit
Observations (or Sum Wgts) = 1008

Whole Model Test

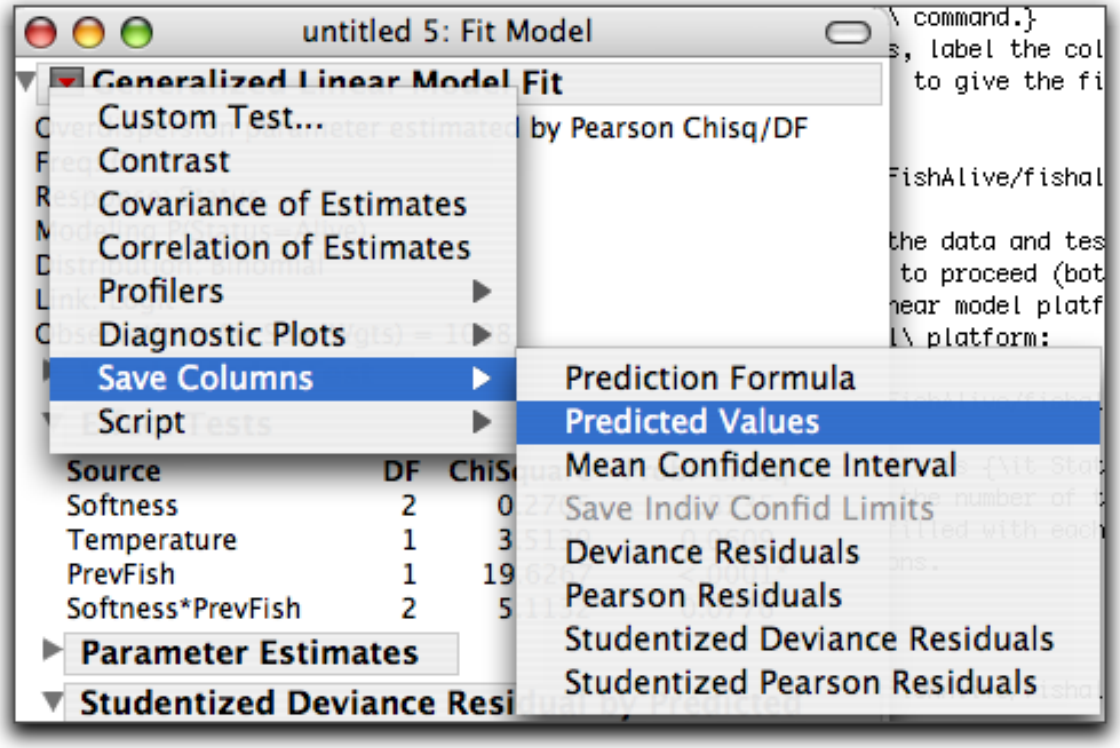
Effect Tests

Source	DF	ChiSquare	Prob>Chisq
Softness	2	0.2705	0.8735
Temperature	1	3.5139	0.0609
PrevFish	1	19.6267	<.0001*
Softness*PrevFish	2	5.1132	0.0776

It appears that there is good evidence of *Previous Ownership*, marginal evidence of an effect of *Temperature* and an interaction between water softness and previous ownership. [Because the two factor interaction was retained, the main effects of softness and previous ownership must be retained in the model even though it looks as if there is no main effect of softness. Refer to the previous notes on two-factor ANOVA for details.]

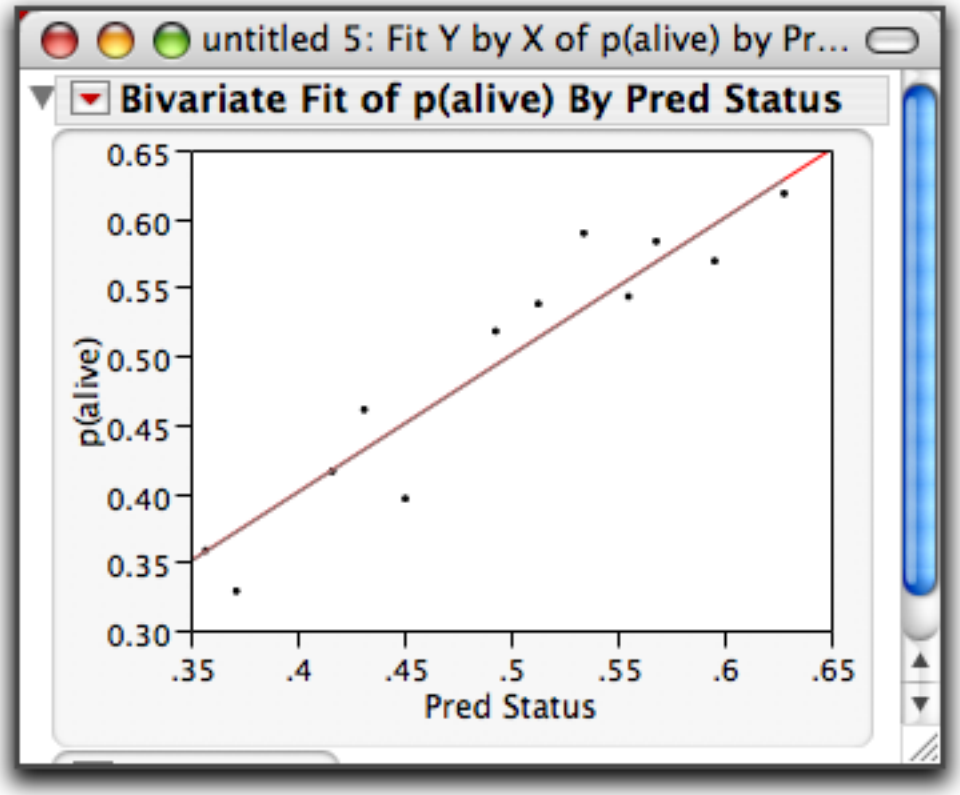
Save the predicted $p(alive)$ to the data table[29]

[29] CAUTION: the predicted $p(alive)$ is saved to the data line even if the actual status is *dead*.

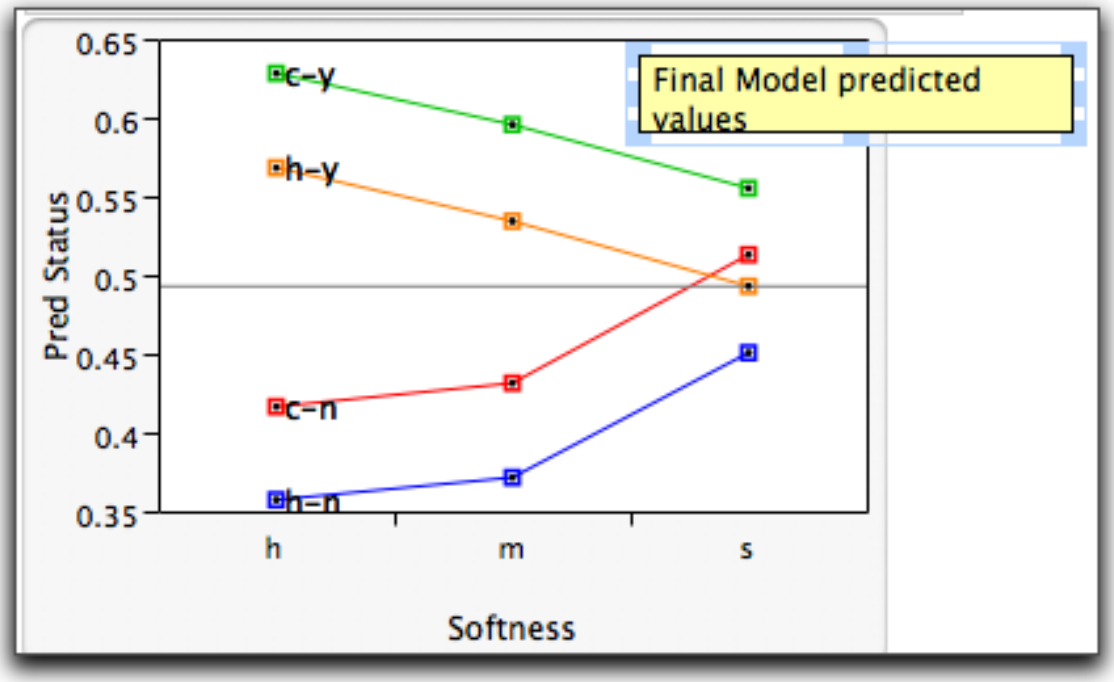


and plot the observed proportions against the predicted values as seen in regression examples earlier.[30]

[30] Use the *Analyze->Fit Y-by-X* platform, and then the *Fit Special* option to draw a line with slope=1 on the plot.



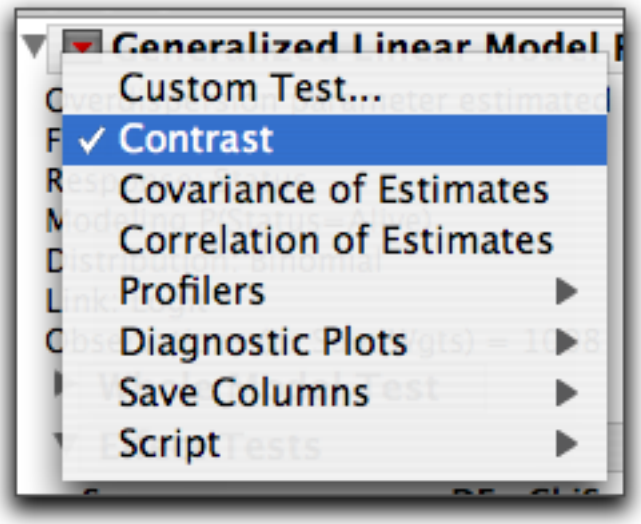
The plot isn't bad and seems to have captured most of what is happening. Use the *Analyze->Fit Y-by-X* platform, with the *Matching Column* as before to create the profile plot of the predicted values:



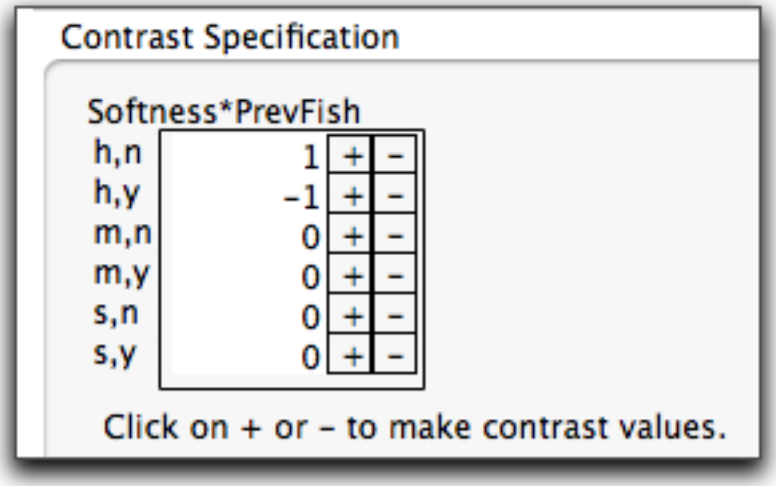
It is a pity that *JMP* gives you no easy way to annotate the standard error or confidence intervals for the predicted mean $p(alive)$, but the confidence bounds can be saved to the data table.

Unlike regular regression, it makes no sense to make predictions for individual fish.

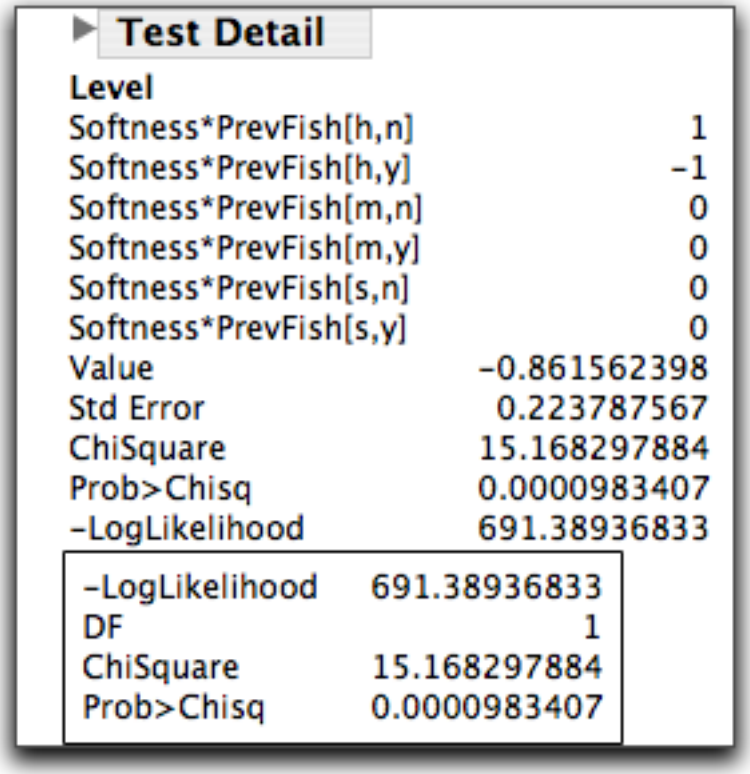
By using the *Contrast* pop-down menu, you can estimate the difference in survival rates (but, unfortunately, on the logit scale) as needed. For example, suppose that you wished to estimate the difference in survival rates between fish raised in hard water and no previous experience and hard water with previous experience. Use the *Contrast* pop-down menu:



The contrast is specified by pressing the - and + boxes as needed:


Contrast Specification
Softness*PrevFish
h,n 1 + -
h,y -1 + -
m,n 0 + -
m,y 0 + -
s,n 0 + -
s,y 0 + -
Click on + or - to make contrast values.

This gives:


Test Detail
Level
Softness*PrevFish[h,n] 1
Softness*PrevFish[h,y] -1
Softness*PrevFish[m,n] 0
Softness*PrevFish[m,y] 0
Softness*PrevFish[s,n] 0
Softness*PrevFish[s,y] 0
Value -0.861562398
Std Error 0.223787567
ChiSquare 15.168297884
Prob>Chisq 0.0000983407
-LogLikelihood 691.38936833
-LogLikelihood 691.38936833
DF 1
ChiSquare 15.168297884
Prob>Chisq 0.0000983407

Again this is on the logit scale and implies that the $logit(p(alive))_{h,n} - logit(p(alive))_{h,y} = -.86$ (*se* .22). This is highly statistically significant. But, what does this mean? Working backwards, we get:

$$\begin{aligned}
&logit\left(p\left(alive\right)_{hn}\right) - logit\left(p\left(alive\right)_{hy}\right) = -.86\\
&\log\left[\frac{p(alive)_{hn}}{1-p(alive)_{hn}}\right] - \log\left[\frac{p(alive)_{hy}}{1-p(alive)_{hy}}\right] = -.86\\
&\log\left[\frac{odds(alive)_{hn}}{odds(alive)_{hy}}\right] = -.86\\
&\frac{odds(alive)_{hn}}{odds(alive)_{hy}} = e^{-.86} = .423
\end{aligned}$$

Or, the odds of a fish being alive from a non-owner in hard water are about 1/2 of the odds of a fish being alive from a previous owner in hard water. If you look at the previous graphs, this indeed does match. It is possible to compute a *se* for this odds ratio, but is beyond the scope of this course.

22.8 Example: Horseshoe crabs - Continuous and categorical predictors.

As to be expected, combinations of continuous and categorical X variables can also be fit using similar reasoning as ANCOVA models discussed in the chapter on multiple regression.

If the categorical X variable has k categories, $k-1$ indicator variables will be created using an appropriate coding. Different computer packages use different codings, so you must read the package documentation carefully in order to interpret the estimated coefficients. However, the different codings, must, in the end, arrive at the same final estimates of effects.

Unlike the ANCOVA model with continuous responses, there are no simple plots in logistic regression to examine visually the parallelism of the response or the equality of intercepts.[31] Preliminary plots where data are pooled into various classes so that empirical logistic plots can be made seem to be the best that can be done.

As in the ANCOVA model, there are three models that are usually fit. Let X represent the continuous predictor, let Cat represent the categorical predictor, and p the probability of success. The three models are:

- $logit(p) = X \;\; Cat \;\; X * Cat$ - different intercepts and slopes for each group;
- $logit(p) = X \;\; Cat$ - different intercepts but common slope (on the logit scale);
- $logit(p) = X$ - same slope and intercept for all groups - coincident lines.

The choice among these models is made by examining the *Effect Tests* for the various terms. For example, to select between the first and second model, look at the *p*-value of the $X * Cat$ term; to select between the second and third model, examine the *p*-value for the Cat term.

[31] This is a general problem in logistic regression because the responses are one of two discrete categories.

These concepts will be illustrated using a dataset on nesting horseshoe crabs[32] that is analyzed in Agresti's book.[33]

The design of the study is given in Brockmann H.J. (1996). Satellite male groups in horseshoe crabs, *Limulus polyphemus*. Ethology, 102, 1-21. Again it is important to check that the design is a completely randomized design or a simple random sampling. As in regression models, you do have some flexibility in the choice of the X settings, but for a particular weight and color, the data must be selected at random from that relevant population.

Each female horseshoe crab had a male resident in her nest. The study investigated other factors affecting whether the female had any other males, called *satellites* residing nearby. These other factors includes:

- crab color where 2=light medium, 3=medium, 4=dark medium, 5=dark.
- spine condition where 1=both good, 2=one worn or broken, or 3=both worn or broken.
- weight
- carapace width

The number of satellites was measured; for this example we will convert the number of satellite males into a presence (number at least 1) or absence (no satellites).

A *JMP* dataset *crabsatellites.jmp* is available from the Sample Program Library at `http://www.stat.sfu.ca/~cschwarz/Stat-650/Notes/MyPrograms`. A portion of the datafile is shown below:

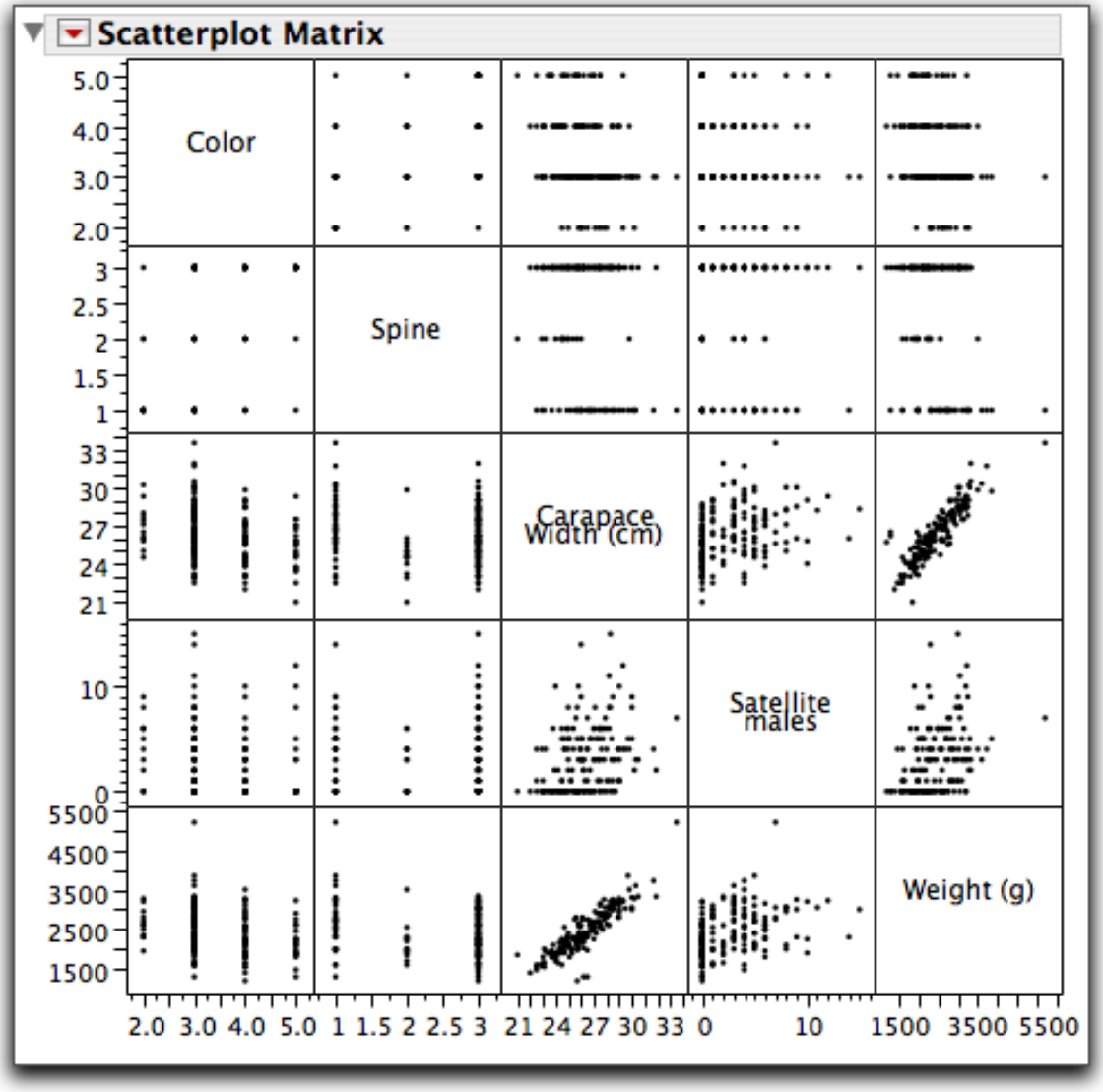
crabsatellites.jmp

	Color	Spine	Carapace Width	Satellite males	Weight (g)	Satellite Males Present
1	3	3	28.3	8	3050	yes
2	4	3	22.5	0	1550	no
3	2	1	26.0	9	2300	yes
4	4	3	24.8	0	2100	no
5	4	3	26.0	4	2600	yes
6	3	3	23.8	0	2100	no
7	2	1	26.5	0	2350	no

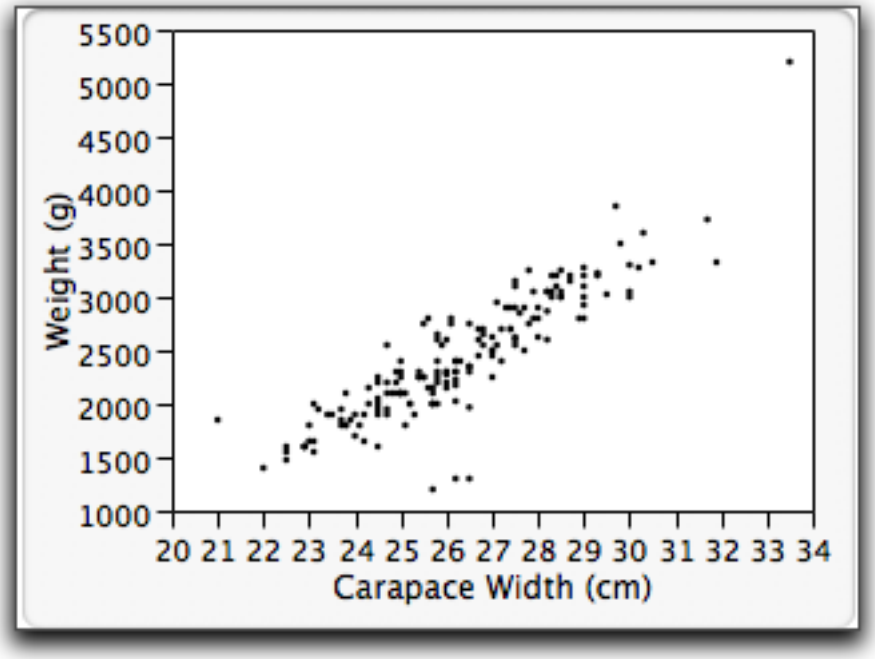
[32] See `http://en.wikipedia.org/wiki/Horseshoe_crab`.
[33] These are available from Agresti's web site at `http://www.stat.ufl.edu/~aa/cda/sas/sas.html`.

Note that the *color* and *spine condition* variables should be declared with an **ordinal** scale despite having numerical codes. The number of satellite males was converted to a presence/absence value using the *JMP* formula editor.

A preliminary scatter plot of the variables shows some interesting features.



There is a very high positive relationship between carapace width and weight, but there are few anomalous crabs that should be investigated further as shown in this magnified plot:



There are three points with weights in the 1200-1300 g range whose carapace widths suggest that the weights should be in the 2200-2300 g range, i.e. a typographical error in the first digit. There is a single crab whose weight suggests a width of 24 cm rather than 21 cm – perhaps a typo in the last digit. Finally, there is one crab which is extremely large compared to the rest of the group. In the analysis that follows, I've excluded these five crabs.

The final point also appears to have an unusual number of satellite males compared to the other crabs in the dataset.

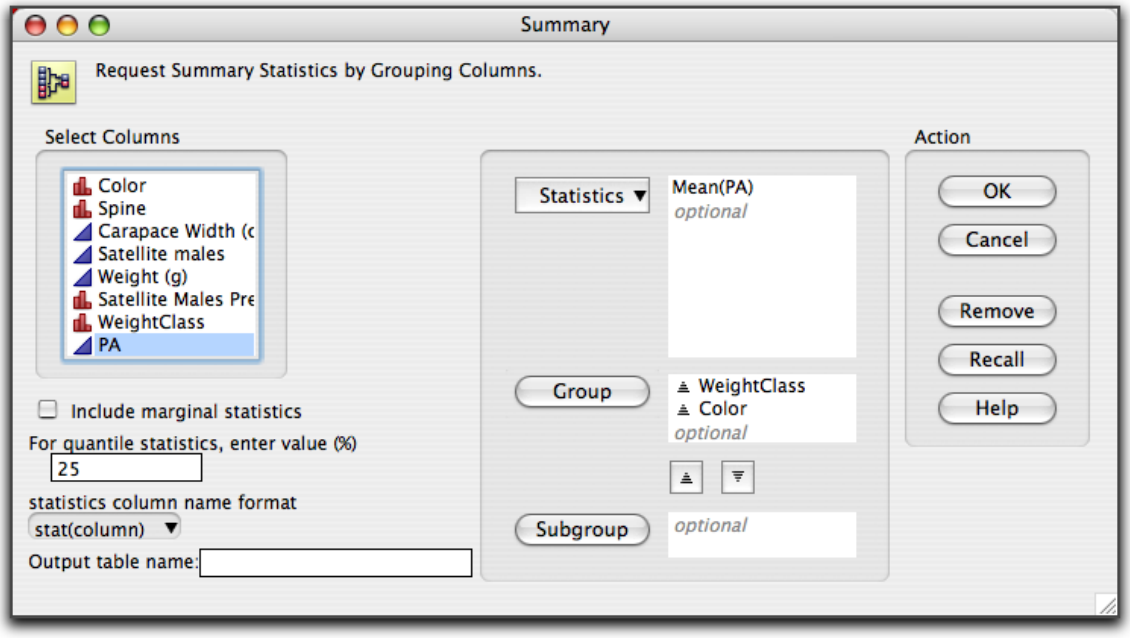
The *Analyze->Fit Y-by-X* platform was then used to examine the differences in mean or proportions in the other variables when grouped by the presence/absence score. These are not shown in these notes, but generally demonstrate some separation in the means or proportions between the two groups, but there is considerable overlap in the individual values between the two groups. The group with no satellite males tends to have darker colors than the presence group; while the distinction between the spine condition is not clear cut.

Because of the high correlation between carapace size and weight, the weight variable was used as the continuous covariate and the color variable was used as the discrete covariate.

A preliminary analysis divided weight into four classes (up to 2000g; 2000-2500 g; 2500-3000 g; and over 3000 g).[34] Similarly, a new variable (PA) was created to be 0 (for absence) or 1 (for presence) for the presence/absence of satellite males. The *Tables->Summary* was used to compute the mean PA (which then

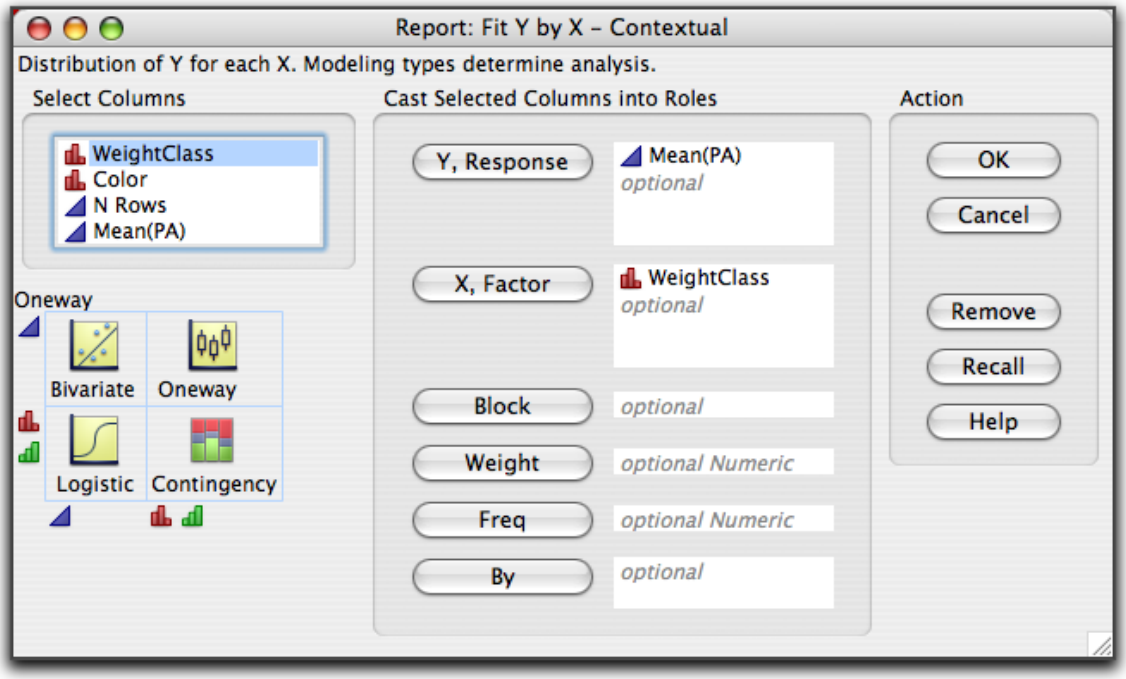
[34]The formula commands of *JMP* were used.

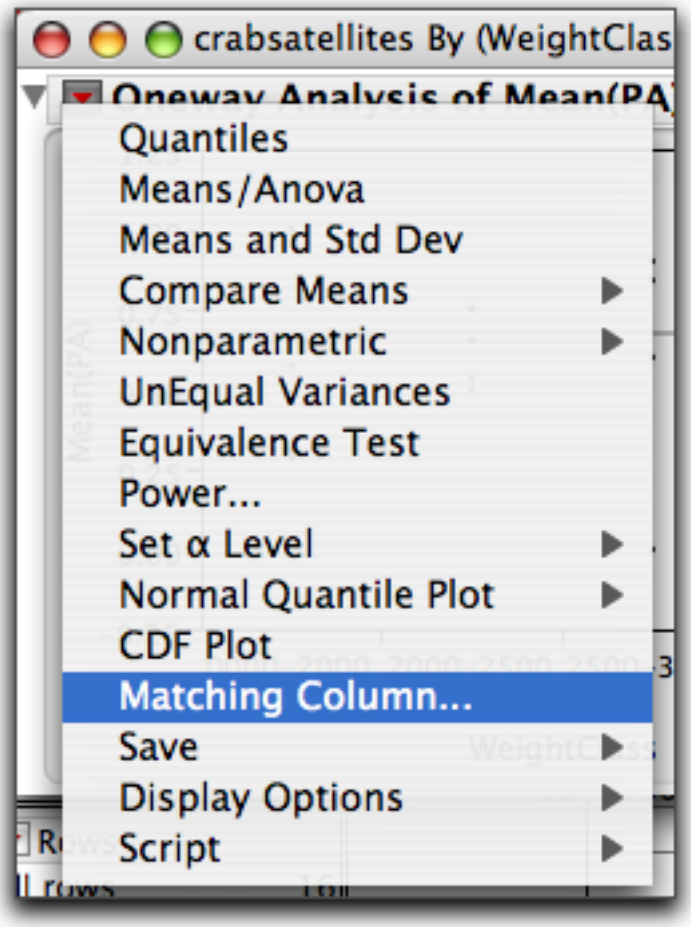
corresponds to the estimated probability of presence) for each combination of weight class and color:

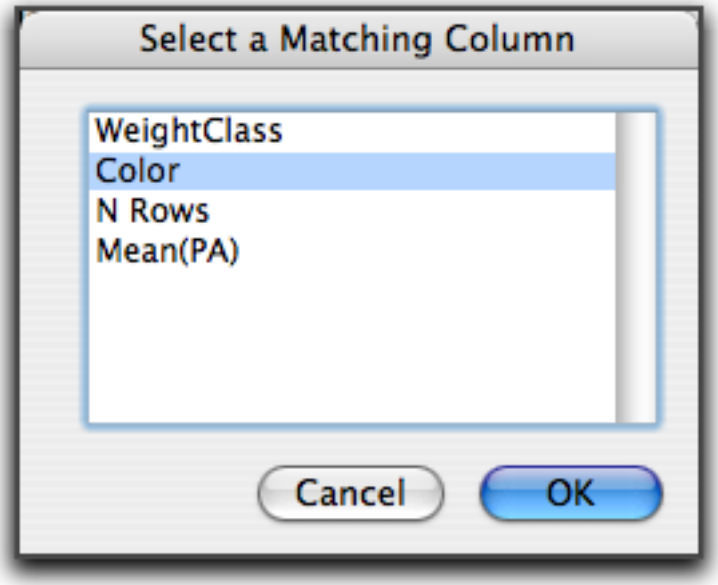


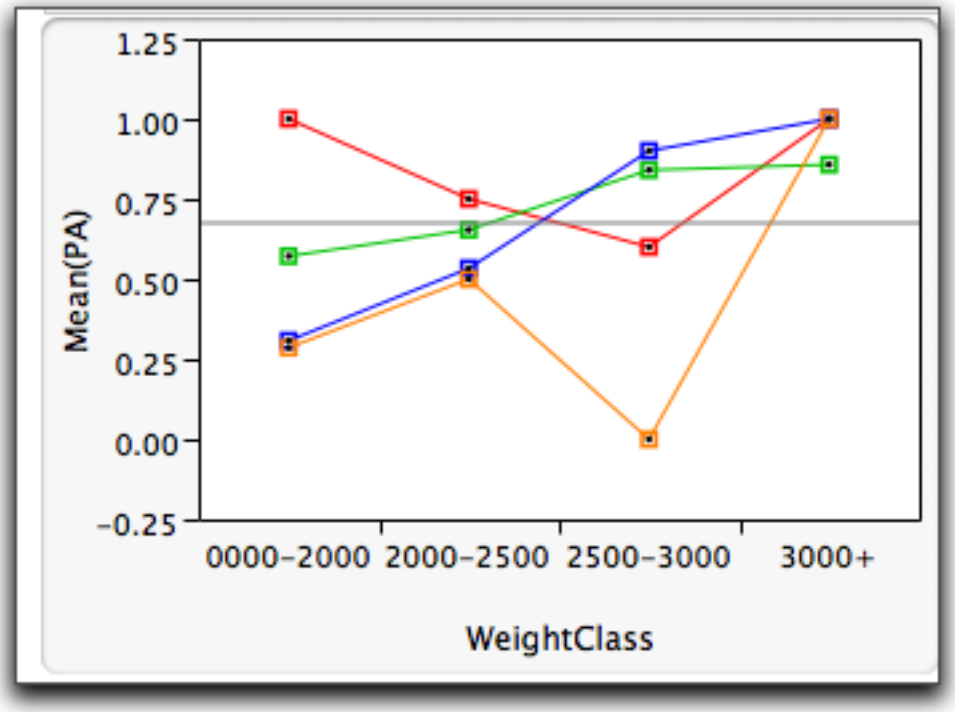
	WeightClass	Color	N Rows	Mean(PA)
1	0000-2000	2	1	1.00
2	0000-2000	3	21	0.57
3	0000-2000	4	13	0.31
4	0000-2000	5	7	0.29
5	2000-2500	2	4	0.75
6	2000-2500	3	26	0.65
7	2000-2500	4	15	0.53
8	2000-2500	5	8	0.50
9	2500-3000	2	5	0.60
10	2500-3000	3	25	0.84
11	2500-3000	4	10	0.90
12	2500-3000	5	4	0.00
13	3000+	2	2	1.00
14	3000+	3	21	0.86
15	3000+	4	5	1.00
16	3000+	5	1	1.00

Finally, the *Analyze->Fit Y-by-X* platform was used to plot the probability of presence by weight class, using the *Matching Column* to joint lines of the same color:


Report: Fit Y by X – Contextual
Distribution of Y for each X. Modeling types determine analysis.
Select Columns
WeightClass
Color
N Rows
Mean(PA)
Cast Selected Columns into Roles
Y, Response
Mean(PA)
optional
X, Factor
WeightClass
optional
Block
optional
Weight
optional Numeric
Freq
optional Numeric
By
optional
Action
OK
Cancel
Remove
Recall
Help
Oneway
Bivariate
Oneway
Logistic
Contingency


crabsatellites By (WeightClas
Quantiles
Means/Anova
Means and Std Dev
Compare Means
Nonparametric
UnEqual Variances
Equivalence Test
Power...
Set α Level
Normal Quantile Plot
CDF Plot
Matching Column...
Save
Display Options
Script

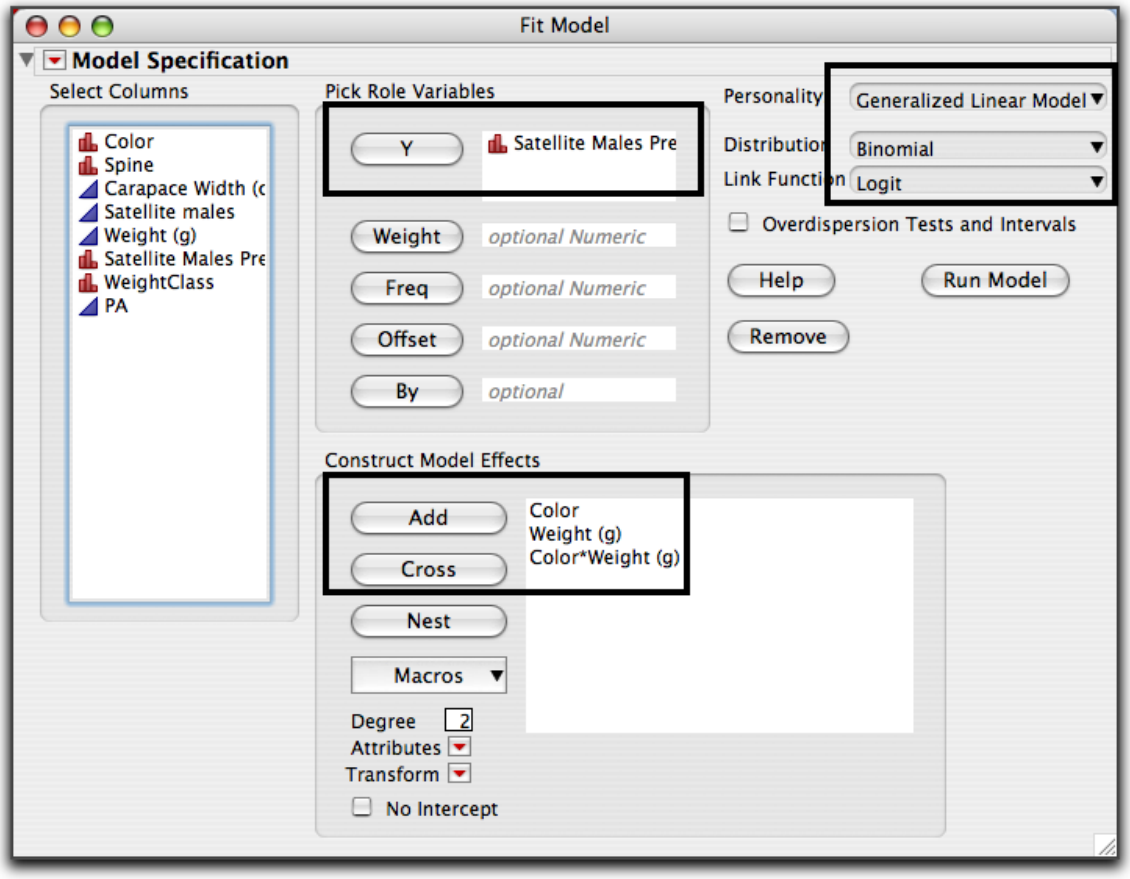




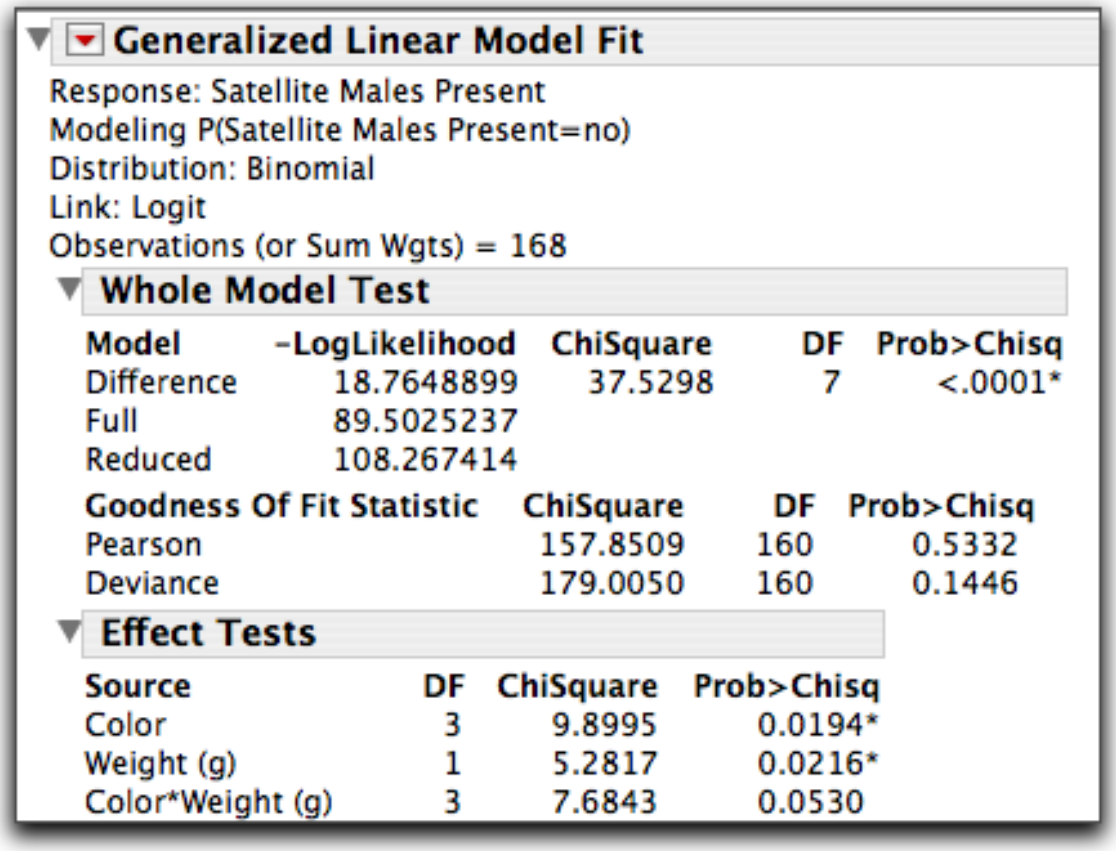
Note despite the appearance of non-parallelism for the bottom line, the point in the 2500-3000 gram category is only based on 4 crabs and so has very poor precision. Similarly, the point near 100% in the 0000-2000 g

category is based on 1 data point! The parallelism hypothesis may be appropriate.

A generalized linear model using the *Analyze->Fit Y-by-X* platform was used to fit the most general model using the raw data:



This gives the results:



Generalized Linear Model Fit

Response: Satellite Males Present
Modeling P(Satellite Males Present=no)
Distribution: Binomial
Link: Logit
Observations (or Sum Wgts) = 168

Whole Model Test

Model	-LogLikelihood	ChiSquare	DF	Prob>Chisq
Difference	18.7648899	37.5298	7	<.0001*
Full	89.5025237			
Reduced	108.267414			

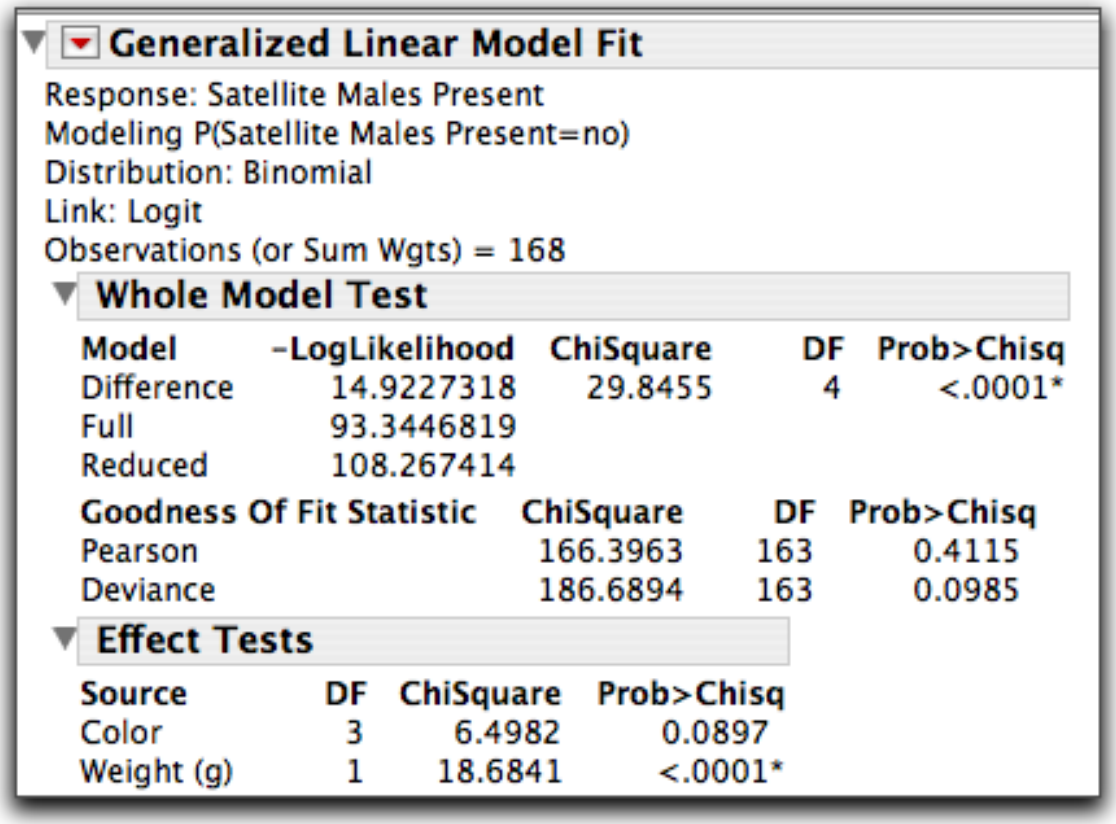
Goodness Of Fit Statistic	ChiSquare	DF	Prob>Chisq
Pearson	157.8509	160	0.5332
Deviance	179.0050	160	0.1446

Effect Tests

Source	DF	ChiSquare	Prob>Chisq
Color	3	9.8995	0.0194*
Weight (g)	1	5.2817	0.0216*
Color*Weight (g)	3	7.6843	0.0530

The *p*-value for non-parallelism (refer to the line corresponding to the *Color*Weight* term) is just over $\alpha = .05$ so there is some evidence that perhaps the lines are not parallel. The parameter estimates are not interpretable without understanding the coding scheme used for the indicator variables. The goodness-of-fit test does not indicate any problems.

Let us continue with a the parallel slopes model by dropping the interaction term. This gives the following results:



Generalized Linear Model Fit

Response: Satellite Males Present
Modeling P(Satellite Males Present=no)
Distribution: Binomial
Link: Logit
Observations (or Sum Wgts) = 168

Whole Model Test

Model	-LogLikelihood	ChiSquare	DF	Prob>Chisq
Difference	14.9227318	29.8455	4	<.0001*
Full	93.3446819			
Reduced	108.267414			

Goodness Of Fit Statistic	ChiSquare	DF	Prob>Chisq
Pearson	166.3963	163	0.4115
Deviance	186.6894	163	0.0985

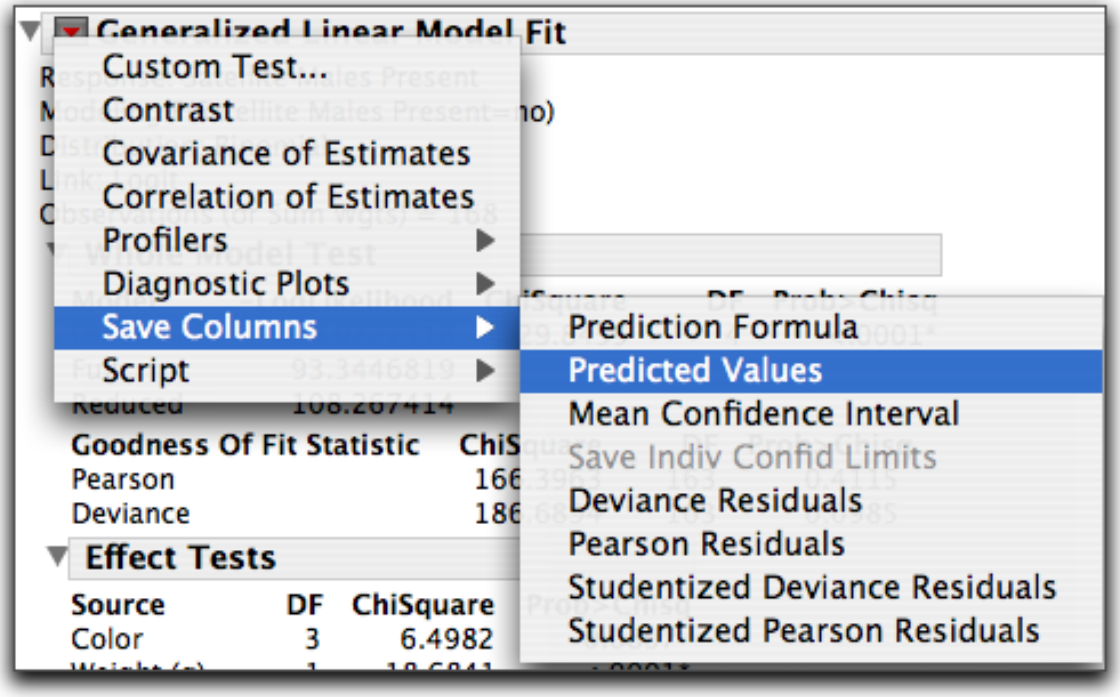
Effect Tests

Source	DF	ChiSquare	Prob>Chisq
Color	3	6.4982	0.0897
Weight (g)	1	18.6841	<.0001*

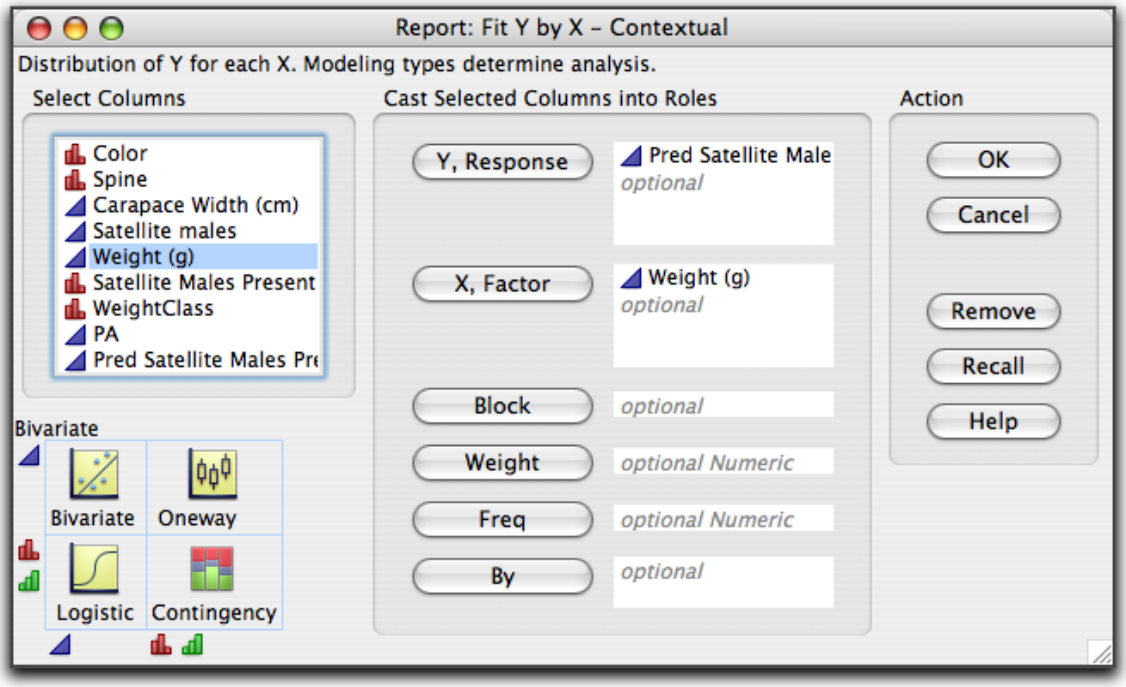
There is good evidence that the log-odds of NO males present decrease as weight increases (i.e. the log-odds of a male being present increases as weight increases), with an estimated increase of .0016 in the log-odds per gram increase in weight. There is very weak evidence that the intercepts are different as the p-value is just under 10%.

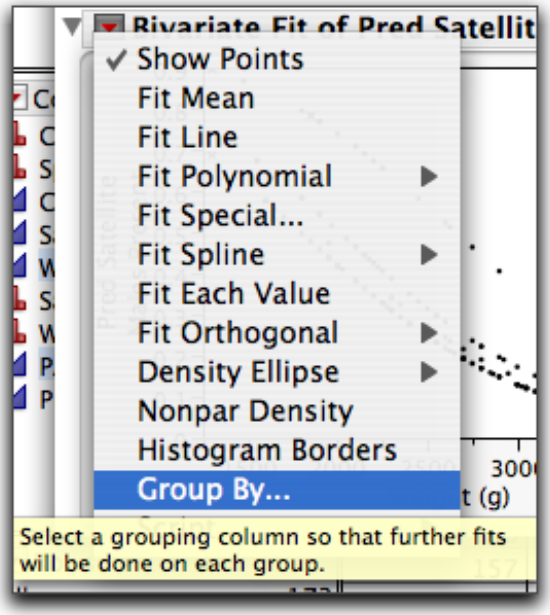
The goodness-of-fit test seems to indicate no problem. The residual plot must be interpreted carefully, but its appearance was explained in a previous section.

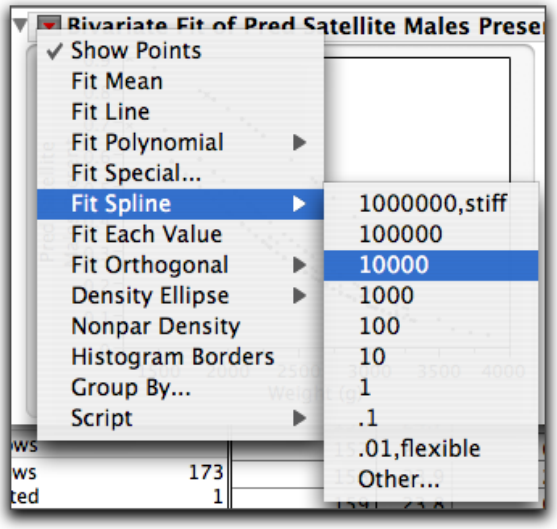
The different intercepts will be retained to illustrate how to graph the final model. Use the red-triangle to save the predicted probabilities to the data table. Note that you may wish to rename the predicted column to remind yourself that the probability of NO male is being predicted.



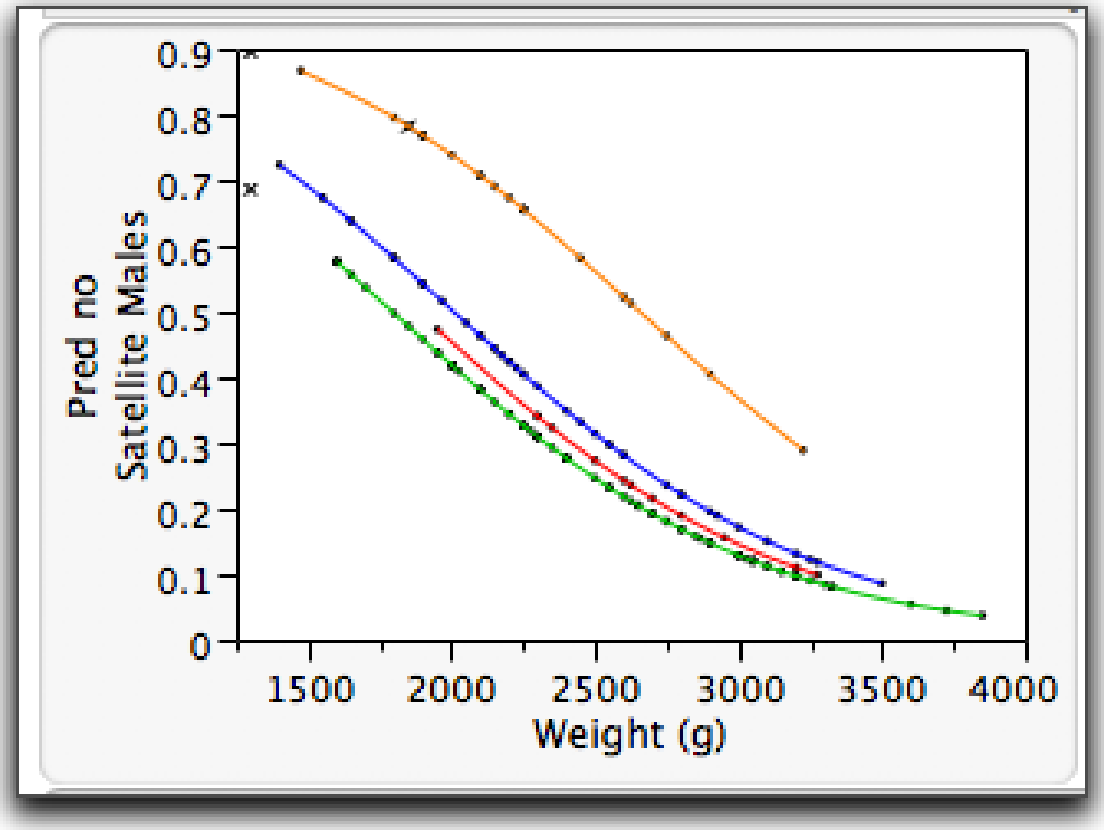
Use the *Analyze->Fit Y-by-X* platform to plot the predicted probability of absence against weight, use the group by option to separate by color, and then fit a spline (a smooth flexible curve) to draw the four curves:


Report: Fit Y by X – Contextual
Distribution of Y for each X. Modeling types determine analysis.
Select Columns
Color
Spine
Carapace Width (cm)
Satellite males
Weight (g)
Satellite Males Present
WeightClass
PA
Pred Satellite Males Pre
Cast Selected Columns into Roles
Y, Response
Pred Satellite Male
optional
X, Factor
Weight (g)
optional
Block
optional
Weight
optional Numeric
Freq
optional Numeric
By
optional
Action
OK
Cancel
Remove
Recall
Help
Bivariate
Bivariate
Oneway
Logistic
Contingency


Bivariate Fit of Pred Satellit
Show Points
Fit Mean
Fit Line
Fit Polynomial
Fit Special...
Fit Spline
Fit Each Value
Fit Orthogonal
Density Ellipse
Nonpar Density
Histogram Borders
Group By...
Select a grouping column so that further fits will be done on each group.

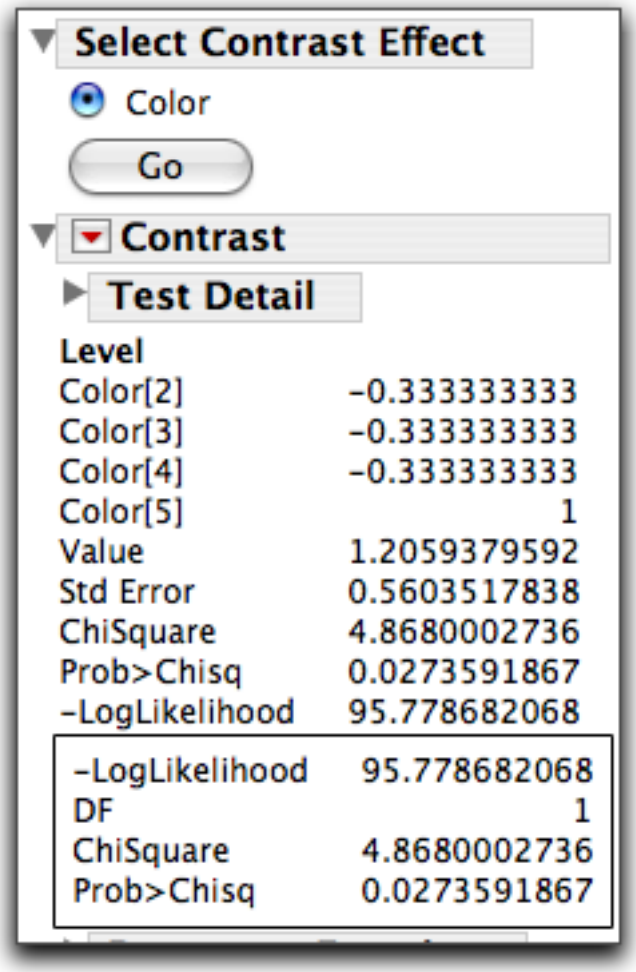


to give the final plot:



Notice that while the models are linear on the log-odds scale, they plots will show a non-linear shape on the regular scale.

It appears that the color=5 group appears to be different from the rest. If you do a contrast among the intercepts (not really a good idea as this could be considered data dredging), you indeed find evidence that the intercept (on the log-odds scale) for color 5 may be different than the average of the intercepts for the other three colors:



22.9 Assessing goodness of fit

As is the case in all model fitting in Statistics, it is important that the model provides an adequate fit to the data at hand. Without such an analysis, the inferences drawn from the model may be misleading or even totally wrong!

One of the "flaws" of many published papers is a lack to detail on how the fit of the model to the data was assessed. The logistic regression model is a powerful statistical tool, but it must be used with caution.

Goodness-of-fit for logistic regression models are more difficult than similar methods for multiple regression because of the binary (success/failure) nature of the response variable. Nevertheless, many of the

methods used in multiple regression have been extended to the logistic regression case.

A nice review paper of the methods of assessing fit is given by

Hosmer, D. W., Tabler, S., and Lameshow, S. (1991).
The importance of assessing the fit of logistic regression models: a case study.
American Journal of Public Health, 81, 1630âĂŞ1635.
http://dx.doi.org/10.2105/AJPH.81.12.1630

In any statistical model, there are two components – the structural portion (e.g. the fitted curve) and the residual (or noise) (e.g. the deviation of the actual values from the fitted curve). The process of building a model focuses on the structural portion. Which variables are important in predicting value? Is the correct scale (e.g. should x or x^2 be used?) After the structural model is fit, the analyst should assess the degree fit.

Assessing goodness-of-fit (GOF) usually entails two stages. First, computing a statistic that summarizes the general fit of the model to the data. Second, computing statistics for individual observations that assess the (lack of) fit of the model to individual observations and their leverage in the fit. This may indentify particular observations that are outliers or have undue influence or leverage on the fit. These points need to be inspected carefully, but it is important to remember that data should not be arbitrarily deleted based solely on a statistical measure.

Let $\widehat{\pi}_i$ represent the predicted probability for case i whose response is either 0 (for failure) or 1 (for success). The deviance of a point is defined as

$$d_i = \sqrt{2|\ln(\widehat{\pi}_i^{y_i}(1-\widehat{\pi}_i)^{1-y_i})|}$$

and is basically a function of the log-likelihood for that observation.

The total deviance is defined as:

$$D = \sum d_i^2$$

Another statistics, the Pearson residual, is defined as:

$$f_i = \frac{y_i - \widehat{\pi}_i}{\sqrt{\widehat{\pi}_i(1-\widehat{\pi}_i)}}$$

and the Pearson chi-square statistic is defined as

$$\chi^2 = \sum r_i^2$$

The summary statistics D and χ^2 each have degrees of freedom approximately equal to $n-(p+1)$ where p is the number of predictor variables, but don't have any nice distributional forms (i.e. you can't assume that they follow a chi-square distribution). This is because the individual components are essentially fromed from $n \times 2$ contingency table with all counts 1 or 0 so the problem of small expected counts found in

chi-square tests is quite serious. So any *p*-value reported for these overall goodness-of-fit measures are not very reliable, and about the only thing that is useful is to compare these statistics to their degrees of freedom to compute an approximate variance inflation factor as seen earlier in the Fitness example.

One strategy for sparse tables is to pool. The Lemeshow test divides the data into 10 groups of equal sizes based on the deciles of the fitted values. The observed and expected counts are computed by summing the estimated probabilities and the observed values in the usual fashion, and then computing a standard chi-square goodness-of-fit statistic. It is compared to a chi-square distribution with 8 df.

Any assessment of goodness of fit should then start with the examination of the D, χ^2 and Lemeshow statistics. Then do a careful evauation of the individual terms d_i and r_i.

To start with, examine the residual plots. Suppose we wish to predict membership in a category as a function of a continuous covariate. For example, can we predict the sex of an individual based on their weight? This is known as logistic regression and is discussed in another chapter in this series of notes.

Again refer to the *Fitness* dataset. The (Generalized Linear) model is:

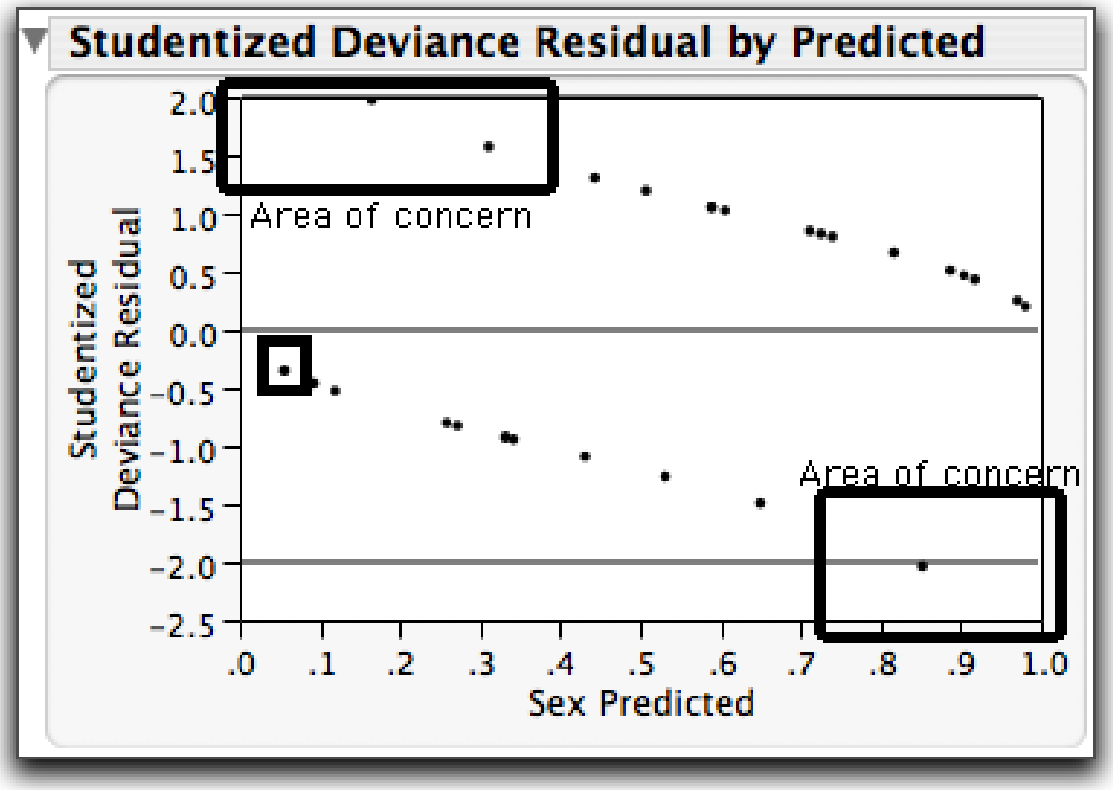
$$Y_i \text{ distributed as } Binomial(p_i)$$

$$\phi_i = logit(p_i)$$

$$\phi_i = Weight$$

The residual plot is produced automatically from the Generalized Linear Model option of the *Analyze->Fit Model* platform and looks like[35]:

[35] I added reference lines at zero, 2, and -2 by clicking on the Y axis of the plot



This plot looks a bit strange!

Along the bottom of the plot, is the predicted probability of being female[36] This is found by substituting in the weight of each person into the estimated linear part, and then back-transforming from the logit scale to the ordinary probability scale. The first point on the plot, identified by a square box, is from a male who weighs over 90 kg. The predicted probability of being female is very small, about 5%.

The first question is exactly how is a residual defined when the Y variable is a category? For example, how would the residual for this point be computed - it makes no sense to simply take the observed (male) minus the predicted probability (.05)?

Many computer packages redefine the categories using 0 and 1 labels. Because *JMP* was modeling the probability of being female, all males are assigned the value of 0, and all females assigned the value of 1. Hence the residual for this point is 0-.05-0.05 which after studentization, is plots as shown.

The bottom line in the residual plot corresponds to the male subjects, The top line corresponds to the female subjects. Where are areas of concern? You would be concerned about females who have a very small probability of prediction for being female, and males who have a large probability of prediction of being

[36] The first part of the output from the platform states that the probability of being female is being modeled.

female. These are located in the plot in the circled areas.

The residual plot's strange appearance is an artifact of the modeling process.

What happens if the predictors in a logistic regression are also categorical. Based on what what seen for the ordinary regression case, you can expect to see a set of vertical lines. But, there are only two possible responses, so the plot reduces to a (non-informative) set of lattice points.

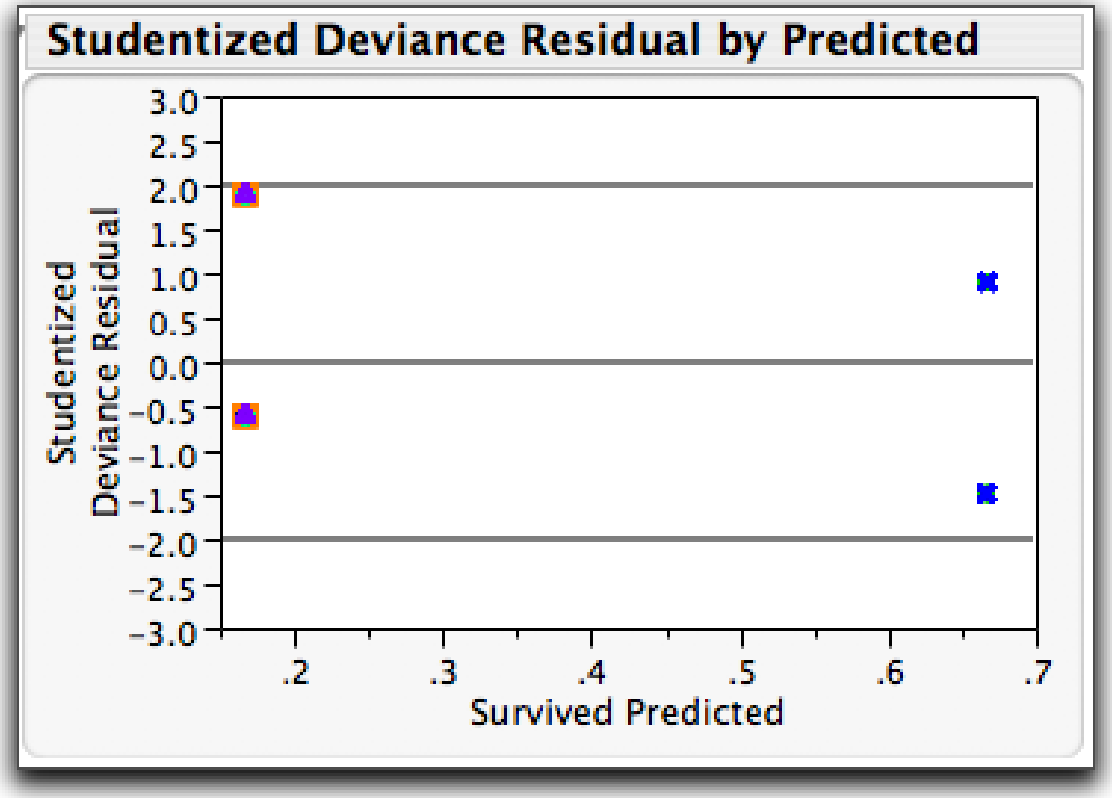
For example, consider predicting survival rates of Titanic passengers as a function of their sex. This model is:

$$Y_i \text{ distributed as } Binomial(p_i)$$

$$\phi_i = logit(p_i)$$

$$\phi_i = Sex$$

The residual plot is produced automatically from the Generalized Linear Model option of the *Analyze->Fit Model* platform and looks like[37]:



[37] I added reference lines at zero, 2, and -2 by clicking on the Y axis of the plot

The same logic applies as in the previous sections. Because Sex is a discrete predictor with two possible values, there are only two possible predicted probability of survival corresponding to the two vertical lines in the plot. Because the response variable is categorical, it is converted to a 0 or 1 values, and the residuals computed which then correspond to the two dots in each vertical line. Note that each dot represents several hundred data values!

This residual plot is rarely informative – after all, if there are only two outcomes and only two categories for the predictors, some people have to lie in the two outcomes for each of the two categories of predictors.

The leverage of a point measures how extreme the set of predictors is relative to the rest of the predictors in the study. Leverage in logistic regression depends no only this distance, but also the weight in predictions which is a function of $\pi(1-\pi)$. Consequently, points with very small predicted (i.e. $\widehat{\pi}_i < 0.15$) or very larger predicted (i.e. $\widehat{\pi}_i > 0.85$) actually have little weight on the fit and the maximum leverage occurs with points where the predicted probability is close to 0.15 or 0.85.

Hosmer et al. (1991) suggest plotting the leverage of each point vs. $\widehat{\pi}_i$ to determine the regions where the leverage is highest. These values may not be available in your package of choice.

Hosmer et al. (1991) also suggest computing the Cook's distance – how much does the regression coefficient change if a case is dropped from the model. These values may not be available in your package of choice.

22.10 Variable selection methods

22.10.1 Introduction

In the previous examples, there were only a few predictor variables and generally, there was only model really of interest. In many cases, the form of the model is unknown, and some sort of variable selection methods are required to build realistic model.

As in ordinary regression, these variable selection methods are NO substitute for intelligent thought, experience, and common sense.

As always, before starting any analysis, check the sample or experimental design. This chapter only deals with data collected under a simple random sample or completely randomized design. If the sample or experimental design is more complex, please consult with a friendly statistician.

Epidemiologists often advise that all clinically relevant variables should be included regardless if statistically significant or not. The rationale for this approach is to provide as complete control of confounding as possible – we saw in regular regression that collinearity among variables can mask statistical significance. The major problem with this approach is over-fitting. Over-fitted models have too many variables relative to the number of observations, leading to numerically unstable estimates with large standard errors.

I prefer a more subdued approach rather than this shotgun approach and would follow these steps to find a reasonable model:

- Start with a multi-variate scatter-plot matrix to investigate pairwise relationships among variables. Are there pairs of variables that appear to be highly correlated? Are there any points that don't seem to follow the pattern seen with the other points?
- Examine each variable separately using the *Analyze->Distribution* platform to check for anomalous values, etc.
- Start with simple univariate logistic regression with each variable in turn.

  For continuous variables, there are two suggested analyses. First, use the binary variable as the X variable and do a simple two-sample t-test to look for differences among the means of the potential predictors. The dot plots should show some separation of the two groups. Second, try a simple univariate logistic-regression using the binary variable as the Y variable with each individual predictor. Third, although it seems odd to do so, convert the binary response variable to a 0/1 continuous response and try some of the standard smoothing methods, such a spline fit to investigate the general form of the response. Does it look logistic? Are quadratic terms needed?

  For nominal or ordinal variables, the two above analyses often start with a contingency table. Particular attention should be paid to problem cases – cells in a contingency table which have a zero count. For example, if an experiment was testing different doses of a drug for the LD50[38] and no deaths occurred at a particular dose. In these situations, the log-odds of success are either $\pm\infty$ which is impossible to model properly using virtually any standard statistical package.[39] If there are cells with 0 counts, some pooling is often required.

  Looking at all the variables, which variables appear to be statistically significant? Approximately how large are these simple effects – can the predictor variables be ranked in approximate order of univariate importance?
- Based upon the above results, start with a model that includes what appear to be the most important variables. As a rule of thumb[40] include variables that have a p-value under .25 rather relying on a stricter criteria. At this stage of the game, building a good starting model is of primary importance.
- Use standard variable selection methods, such as stepwise selection (forward, backward, combined) or all subset regression to investigate potential models. These mechanical methods are not to be used as a substitute for thinking! Remember that highly collinear variables can mask the importance of each other.

  If categorical variables are to be included then some care must be used on how the various indicator variables are included. The reason for this is that the coding of the indicator variables is arbitrary and the selection of a particular indicator variable may be artifact of the coding used. One strategy is that all the indicator variables should be included or excluded as a set, rather than individually selecting separate indicator variables. As you will see in the example, *JMP* has four different rules that could be used.

[38] LD50=Lethal Dose 50th percentile – that dose which kills 50% of the subjects

[39] However, refer to Hosmer and Lemeshow (2000) for details on alternate approaches.

[40] Hosmer and Lemeshow (2000), p. 95

- Once main effects have be identified, look at quadratic, interaction, and crossproduct terms.
- Verify the final model. Look for collinearity, high leverage, etc. Check if the response to the selected variables are linear on the logistic scale. For example, break a continuous variable into 4 classes, and refit the same model with these discretized classes. The estimates of the effects for each class should then follow an approximate linear pattern.
- Cross validate the model so that artifacts of that particular dataset are not highlighted.

22.10.2 Example: Predicting credit worthiness

In credit business, banks are interested in information whether prospective consumers will pay back their credit or not. The aim of credit-scoring is to model or predict the probability that a consumer with certain covariates is to be considered as a potential risk.

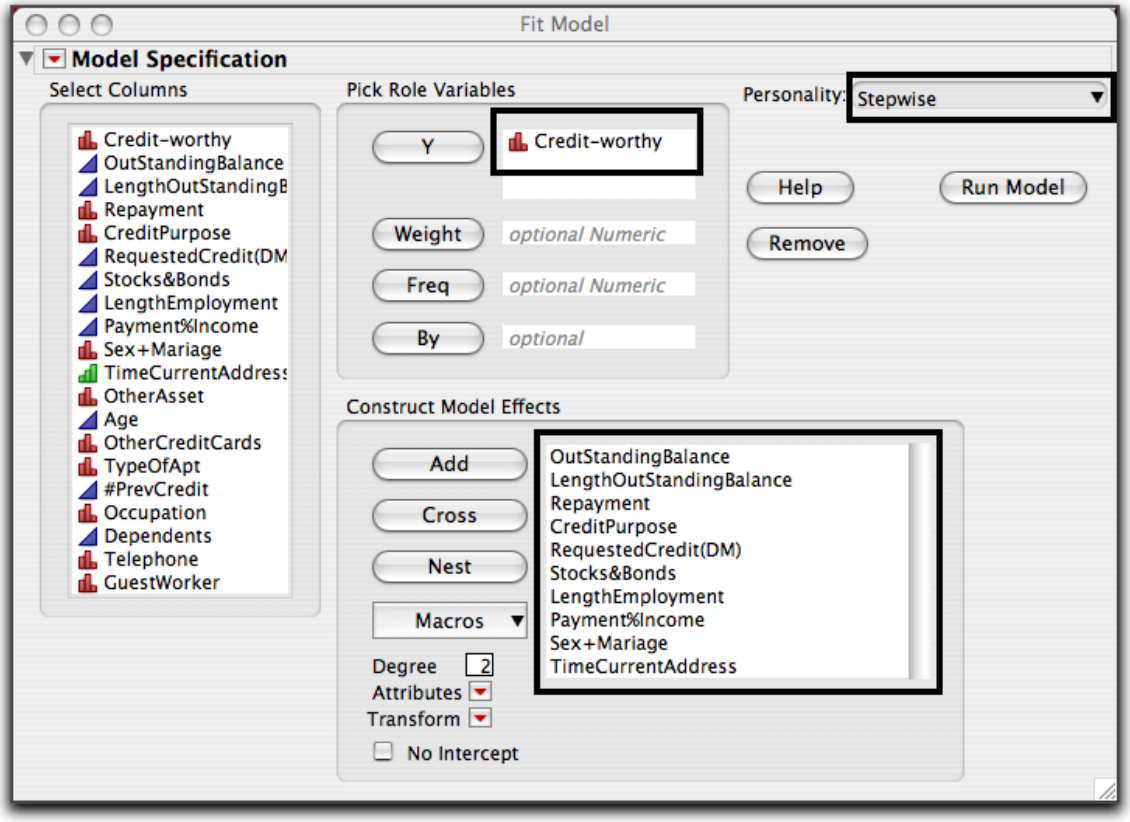
If you visit `http://www.stat.uni-muenchen.de/service/datenarchiv/welcome_e.html` you will find a dataset consisting of 1000 consumer credits from a German bank. For each consumer the binary response variable "creditability" is available. In addition, 20 covariates that are assumed to influence creditability were recorded. The dataset is available in the *creditcheck.jmp* datafile from the Sample Program Library at `http://www.stat.sfu.ca/~cschwarz/Stat-650/Notes/MyPrograms`.

The variable descriptions are available at `http://www.stat.uni-muenchen.de/service/datenarchiv/kredit/kreditvar_e.html` and in the Sample Program Library.

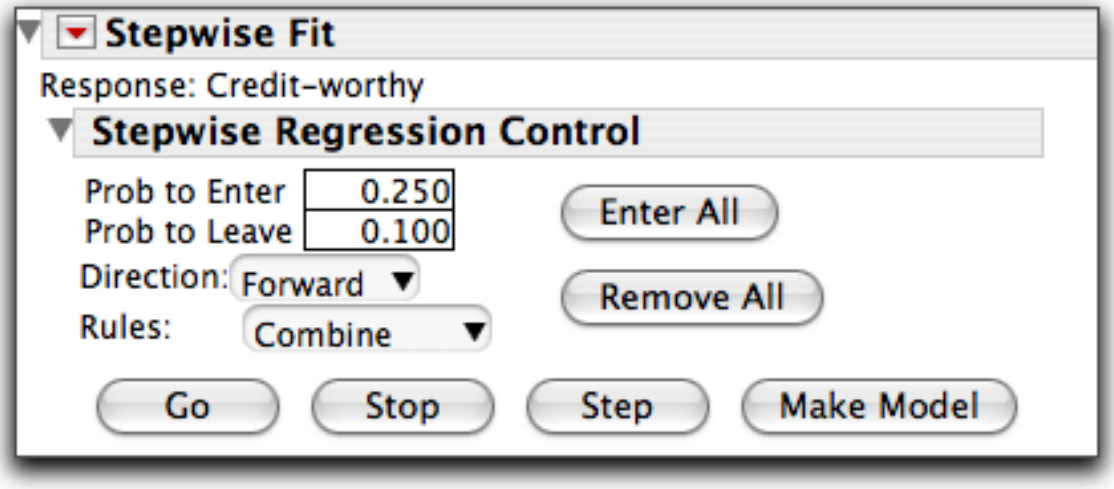
I will assume that the initial steps in variable selection have been done such as scatter-plots, looking for outliers etc.

This dataset has a mixture of continuous variables (such as length of time an account has been paid in full), nominal scaled variables (such as sex, or the purpose of the credit request), and ordinal scaled variables (such as length of employment). Some of the ordinal variables may even be close enough to interval or ratio scaled to be usable as a continuous variables (such as length of employment). Both approaches should be tried, particularly if the estimates for the individual categories appear to be increasing in a linear fashion.

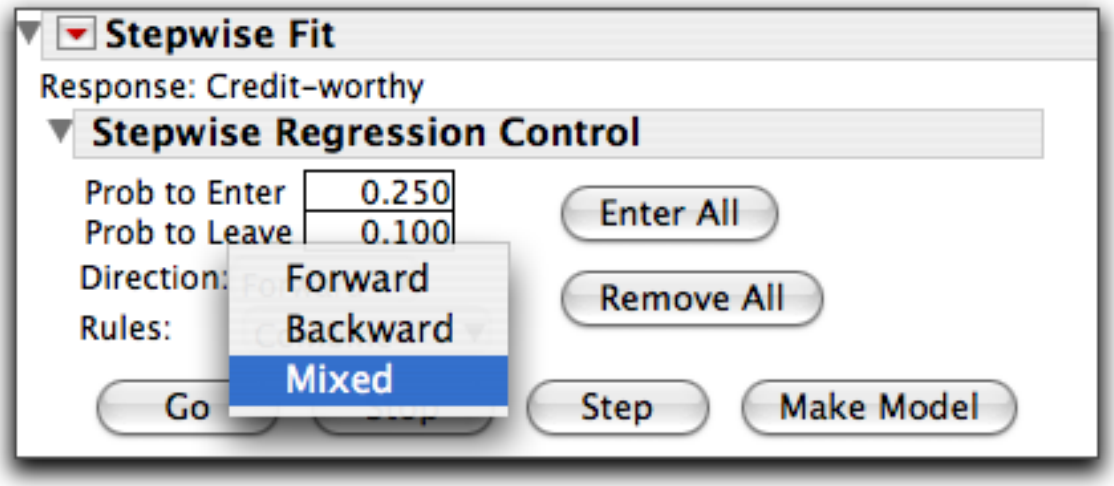
The *Analyze->Fit Model* platform was used to specify the response variable, the potential covariates, and that a variable selection method will be used:



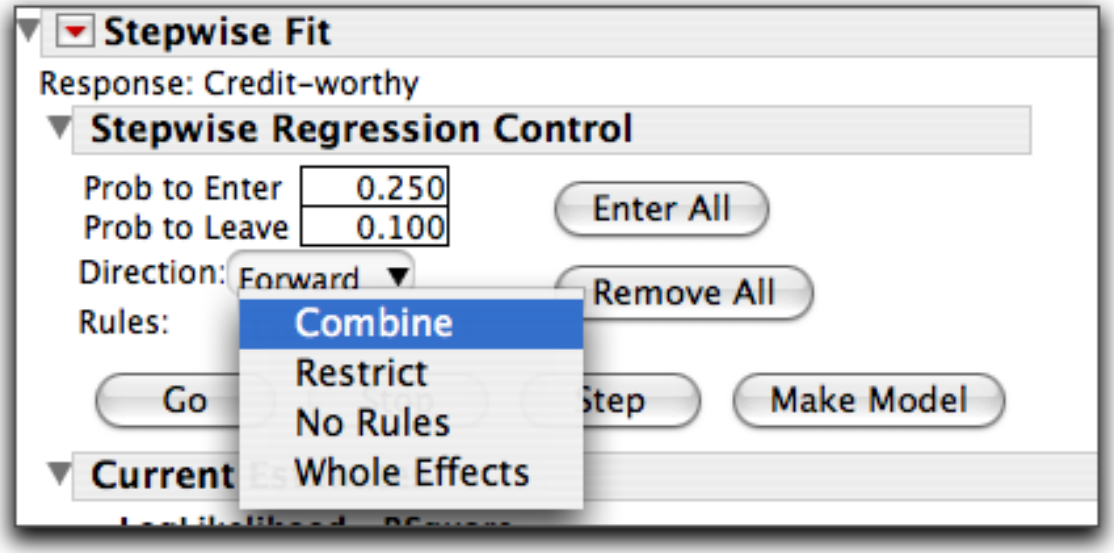
This brings up the standard dialogue box for step-wise and other variable selection methods.



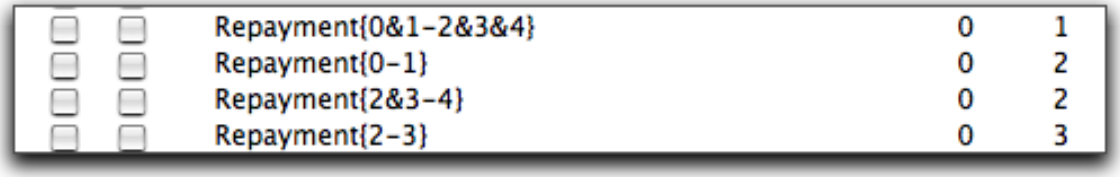
In the stepwise paradigm, the usual *forward*, *backwards*, and *mixed* (i.e. forward followed by a backwards step at each iteration):



In cases where variables are nominally or ordinally scales (and discrete), *JMP* provides a number of way to include/exclude the individual indicator variables:



For example, consider the variable *Repayment* that had levels 0 to 4 corresponding from 0=repayment problems in the past, to 4=completely satisfactory repayment of past credit. *JMP* will create 4 indicator variables to represent these 5 categories. These indicator variables are derived in a hierarchical fashion:



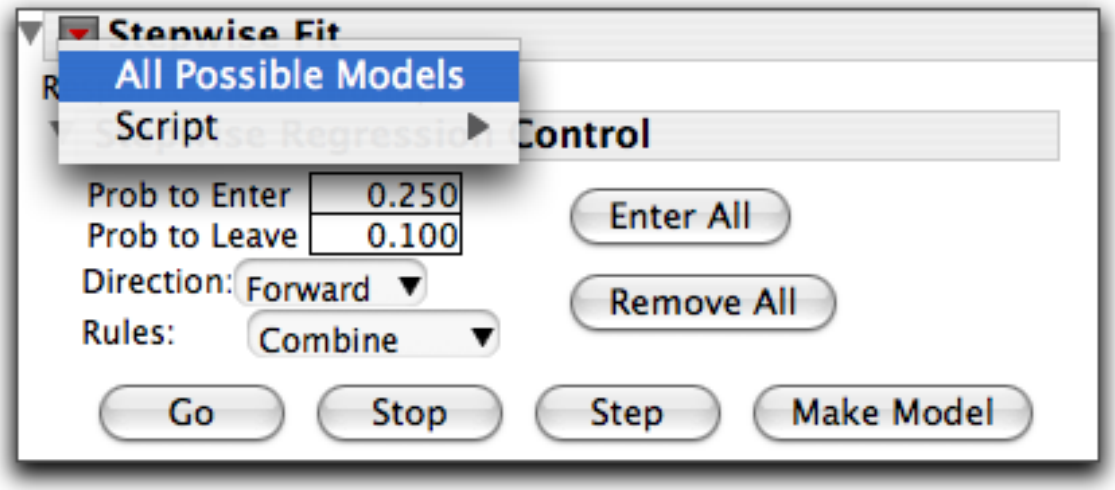
The first indicator variable, splits the classes in such a way to maximize the difference between the proportion of credit worthiness between the two parts of the split. This corresponds to grouping levels 0 and 1 vs. levels 2, 3, and 4. The next indicator variables then split the splits, again, if possible, to maximize the difference in the credit worthiness between the two parts of the split. [If the split is of a pair of variables, there is no choice in the split.] This corresponds to splitting the 0&1 categories into another indicator variable that distinguishes category 0 from 1. The 2&3&4 class is split into two sub-splits corresponding to categories 2&3 vs. category 4. Finally, the 2&3 class is split into an indicator variable differentiating categories 2 and 3.

Now the rules for entering effects correspond to :

- **Combined** When terms enter the model, they are combined with all higher terms in the hierarchy and tested as a group to enter or leave.
- **Restrict** Terms cannot be entered into the model unless terms higher in the hierarchy are already entered. Hence the indicator variable that distinguishes categories 0 and 1 in the *repayment* variable cannot enter before the indicator variable that contrasts 0&1 and 2&3&4.
- **No Rules** Each indicator variable is free to enter or leave the model regardless of the presence or absence of other variables in the set.
- **Whole Effects** All indicator variable in a set must enter or leave together as a set.

The *Combined* or *Whole Effects* are the two most common choices.

This platform also supports all possible subset regressions:



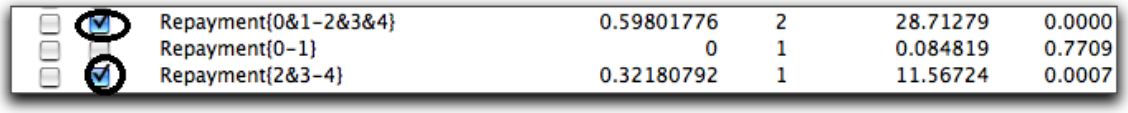
This should be used cautiously with a large number of variables.

Because it is computationally difficult to fit thousands of models using maximum likelihood methods for each of the potential new variables that enter the model, a computationally simpler (but asymptotically equivalent) test procedure (called the Wald or score test) is used in the table of variables to enter or leave. In a forward selection, the variable with the smallest p-value or the largest Wald test-statistic is chosen:

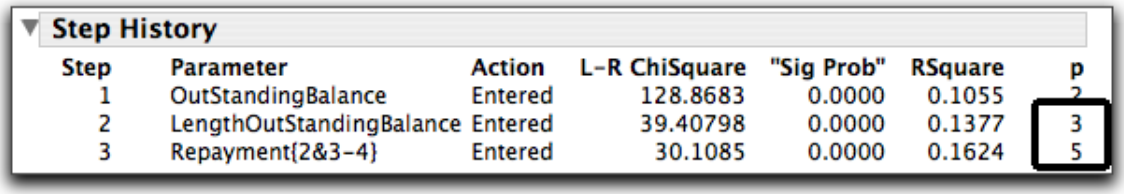
Once this variable is chosen, the current model is refit using maximum likelihood, so the report in the *Step History* may show a slightly different test statistics (the *L-R ChiSquare*) than the score statistic and the p-value may be different.

The stepwise selection continues.

In a few steps, the next variable to enter is the indicator variable that distinguishes categories 2&3 and 4. Because of the restriction on entering terms, if this indicator variable is entered, the first cut must also be entered. Hence, this step actually enters 2 variables and the number of predictors jumps from 3 to 5:



☑	Repayment{0&1-2&3&4}	0.59801776	2	28.71279	0.0000
☐	Repayment{0-1}	0	1	0.084819	0.7709
☑	Repayment{2&3-4}	0.32180792	1	11.56724	0.0007



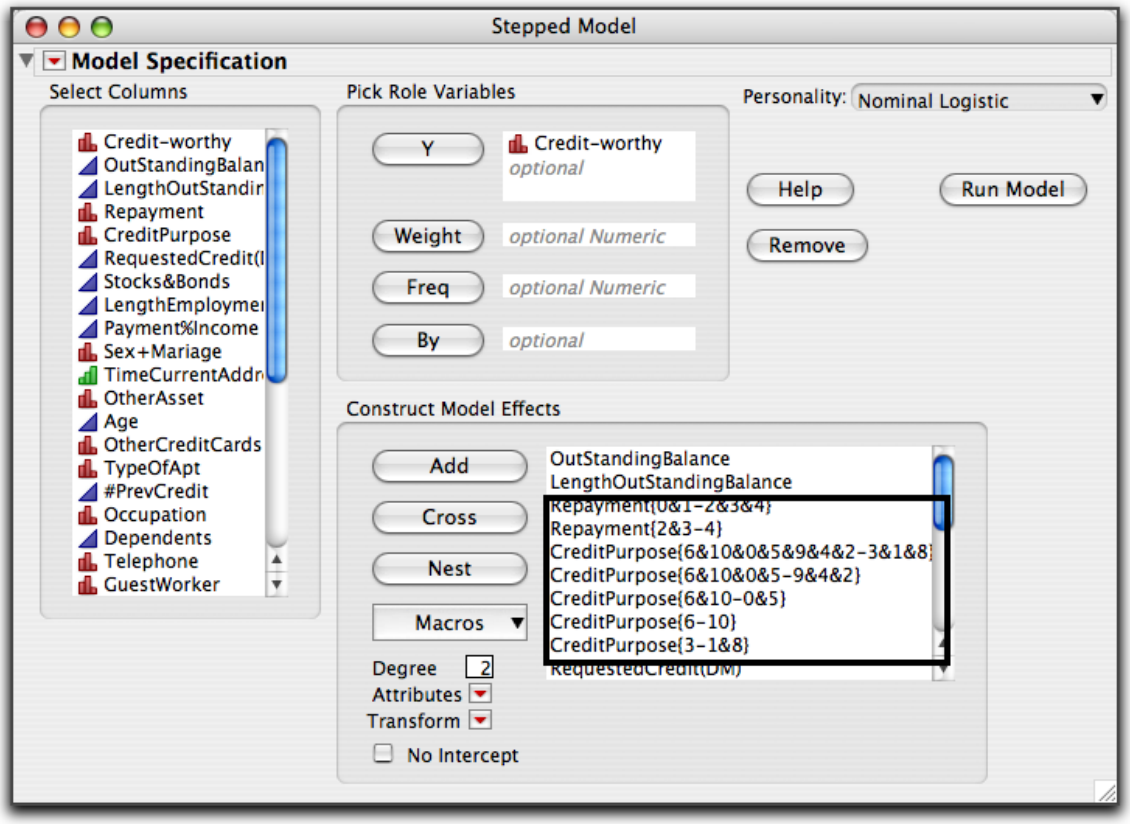
Step History

Step	Parameter	Action	L-R ChiSquare	"Sig Prob"	RSquare	p
1	OutStandingBalance	Entered	128.8683	0.0000	0.1055	2
2	LengthOutStandingBalance	Entered	39.40798	0.0000	0.1377	3
3	Repayment{2&3-4}	Entered	30.1085	0.0000	0.1624	5

In a few more steps, some of the credit purpose variables are entered, again as a pair.

The stepwise continues for a total of 18 steps.

As before, once you have identified a candidate model, it must be fit and examined in more detail. Use the *Make Model* button to fit the final model. Note that *JMP* must add new columns to the data tables corresponding to the indicator variables created during the stepwise report. These can be confusing to the novice, but just keep in mind that any set of indicator variables is somewhat arbitrary.



The model fit then has separate variables used for each indicator variable created:

Parameter Estimates

Term	Estimate	Std Error	ChiSquare	Prob>ChiSq
Intercept	-0.8269838	0.5541132	2.23	0.1356
OutStandingBalance	-0.5698144	0.0721025	62.45	<.0001*
LengthOutStandingBalance	0.02594712	0.0087179	8.86	0.0029*
Repayment{0&1-2&3&4}	0.55823628	0.1440614	15.02	0.0001*
Repayment{2&3-4}	0.31503929	0.1034913	9.27	0.0023*
CreditPurpose{6&10&0&5&9&4&2-3&1&8}	0.40119633	0.1077189	13.87	0.0002*
CreditPurpose{6&10&0&5-9&4&2}	0.20511218	0.1303104	2.48	0.1155
CreditPurpose{6&10-0&5}	-0.2479297	0.2132879	1.35	0.2451
CreditPurpose{6-10}	0.68290115	0.3980495	2.94	0.0862
CreditPurpose{3-1&8}	0.36017088	0.1788359	4.06	0.0440*
RequestedCredit(DM)	0.00012792	4.1535e-5	9.48	0.0021*
Stocks&Bonds	-0.2515152	0.0603091	17.39	<.0001*
LengthEmployment	-0.0979457	0.0739886	1.75	0.1856
Payment%Income	0.31828886	0.0838418	14.41	0.0001*
Sex+Mariage{1&2-4&3}	0.18877483	0.0984243	3.68	0.0551
Sex+Mariage{4-3}	0.18660451	0.148697	1.57	0.2095
TimeCurrentAddress{1-2&3&4}	-0.2716612	0.1278759	4.51	0.0336*
TimeCurrentAddress{2-3&4}	0.14831206	0.0985035	2.27	0.1322
Age	-0.0096947	0.0080579	1.45	0.2289
OtherCreditCards{1&2-3}	0.22450976	0.1057489	4.51	0.0337*
TypeOfApt{3&1-2}	0.21262373	0.0944488	5.07	0.0244*
Telephone[1]	0.12500096	0.0913749	1.87	0.1713
GuestWorker[1]	0.74866893	0.3093471	5.86	0.0155*

For log odds of 0/1

The log-odds of NOT repaying the loan is computed (see the bottom of the estimates table). Do the coefficient make sense?

Can some variables be dropped?

Pay attention to how the indicator variables have been split. For example, do you understand what terms are used if the borrower intends to use the credit to do repairs (*CreditPurpose* value =6)?

Models that are similar to this one should also be explored.

Again, just like in the case of ordinary regression, model validation using other data sets or hold-out samples should be explored.

22.11 Model comparison using AIC

Sorry, to be added later.

22.12 Final Words

22.12.1 Two common problems

Two common problems can be encountered with logistic regression.

Zero counts

As noted earlier, zero counts for one category of a nominal or ordinal predictor (X) variable are problematic as the log-odds of that category then approach $\pm\infty$ which is somewhat difficult to model.

One simplistic approach is that similar to the computation of the empirical logistic estimate – add $1/2n$ to each cell so that the counts are no longer-integers, but most packages will deal with non-integer counts without problems.

If the zero counts arise from spreading the data over too many cells, perhaps some pooling of adjacent cells is warranted. If the data are sufficiently dense that pooling is not needed, perhaps this level of the variable can be dropped.

Complete separation

Ironically, this is a problem because the logistic models are performing too well! We saw an example of this earlier, when the fitness data could predict perfectly the sex of the subject.

This is a problem, because not the predicted log-odds for the groups must again be $\pm\infty$. This can only happen if some of the estimated coefficients are also infinite which is difficult to deal with numerically. Theoretical considerations show that in the case of complete separation, maximum likelihood estimates do not exist!

Sometimes this complete separation is an artifact of too many variables and not enough observations. Furthermore, it is not so much a problem of total observations, but also the division of observations between the two binary outcomes. If you have 1000 observations, but only 1 "success", then any model with more than a few variables will be 100% efficient in capturing the single success – however, it is almost certain to be an artifact of the particular dataset.

22.12.2 Extensions

Choice of link function

The *logit* link function is the most common choice for the link function between the probability of an outcome and the scale on which the predictors operate in a linear fashion.

However, other link functions have been used in different situations. For example, a log-link ($\log(p)$), the log-log link ($\log(-log(p))$), the complementary log-link ($\log(-\log(1-p))$), the probit function (the inverse normal distribution), the identity link (p) have all been proposed for various special cases. Please consult a statistician for details.

More than two response categories

Logistic regression traditionally has two response categories that are classified as "success" or "failure". It is possible to extend this modelling framework to cases where the response variable has more than two categories.

This is known as *multinomial logistic regression*, *discrete choice*, *polychotomous logistic* or *polytomous logistic* model, depending upon your field of expertise.

There is a difference in the analysis if the responses can be ordered (i.e. the response variable takes an ordinal scale), or remain unordered (i.e. the response variable takes an nominal scale).

The basic idea is to compute a logistic regression of each category against a reference category. So a response variable with three categories is translated into two logistic regressions where, for example, the first regression is category 1 vs. category 0 and the second regression is category 2 vs. category 0. These can be used to derive the results of category 2 vs. category 1. What is of particular interest is the role of the predictor variables in each of the possible comparison, e.g. does weight have the same effect upon mortality for three different disease outcomes.

Consult one of the many book on logistic regression for details.

Exact logistic regression with very small datasets

The methods presented in this chapter rely upon maximum likelihood methods and asympototic arguments. In very small datasets, these large sample approximations may not perform well.

There are several statistical packages which perform exact logistic regression and do not rely upon asymptotic arguments.

A simple search of the web brings up several such packages.

More complex experimental designs

The results of this chapter have all assumed that the sampling design was a simple random sample or that the experiment design was a completely randomized design.

Logistic regression can be extended to many more complex designs.

In matched pair designs, each "success" in the outcome is matched with a randomly chosen "failure" along as many covariates as possible. For example, lung cancer patients could be matched with healthy patients with common age, weight, occupation and other covariates. These designs are very common in health studies. There are many good books on the analysis of such design.

Clustered designs are also very common where groups of subjects all receive a common treatment. For example, classrooms may be randomly assigned to different reading programs, and the success or failure of individual students within the classrooms in obtaining reading goals is assessed. Here the experimental unit is the classroom, not the individual student and the methods of this chapter are not directly applicable. Several extensions have been proposed for this type of "correlated" binary data (students within the same classroom are all exposed to exactly the same set of experimenal and non-experimental factors). The most common is known as *Generalized Estimating Equations* and is described in many books.

More complex experimental designs (e.g. split-plot designs) can also be run with binary outcomes. These complex designs require high power computational machinery to analyze.

22.12.3 Yet to do

- examples - dov's example used in a comprehensive exam in previous years This is the end of the chapter