RNR/ENTO 613 – Categorical Responses – Logistic Regression

Logistic Regression for binary response variable

Logistic regression analysis describes how a binary (0 or 1) response variable is associated with a set of explanatory variables (categorical or continuous).

The mean of a binary variable is a probability (π). The logistic regression model relates that probability, through a logistic function, to variation in the explanatory variables.

Equivalently, the logistic regression model relates the log odds [log (π / 1 - π)] in a linear way to variation in the explanatory variables.

Odds

One way to quantify a binary variable is to use *odds* instead of *proportions*. If π is a population *proportion of yes outcomes*, the corresponding *odds* of a yes outcome are:

$$\omega = \frac{\pi}{(1-\pi)} = \frac{p\ of\ a\ yes}{p\ of\ a\ no} = \frac{n\ yes}{n\ \ no}$$

Example: The proportion of cold cases in the vitamin C group (Sleuth Chap. 18) was 0.75, or the odds of getting a cold in the vitamin C group were 3 to 1 (3 = 0.75 / 0.25). The odds of getting a cold were about 3 to 1 for vitamin C consumers; 3 persons got sick for every individual whom did not.

When using odds, it is usual to cite the larger number first.

> Ex: An event with chances of 0.95 has odds of *19 to 1* (0.95 / 0.05 = 19) *in favor* of its occurrence.
>
> An event with chances of 0.05 has odds of *19 to 1* (0.05 / 0.95= 1/19) *against* its occurrence.

Some facts about odds:

1. A proportion of ½ corresponds to odds of 1, in which case we have “equal odds”.
2. Odds vary between 0 and ∞, proportions vary between 0 and 1.
3. Odds are not defined for proportions equal to 0 or 1.
4. If the odds of a yes outcome are ω, the odds of a no are $1/\omega$.
   Ex: ω_{yes} = 3 to 1; ω_{no} = 1 to 3 or 0.33 to 1
5. If the odds of an outcome are ω, then the probability of that outcome is $\pi = \omega / (1+\omega)$.
   Ex: ω_{yes} = 3 to 1; π = probability of a yes = 3 / 3 + 1 = 0.75.

Logistic Regression is a generalized Linear Model

The natural link function for a binary response variable is the *logit* or *log-odds* function, where the *logit link* is $g(\pi) = logit(\pi)$ = **ln[π / (1 - π)] = ln ω**.

When we use the *logit* as a link function, we have the *Logistic regression model*:

$$logit\,(\pi) = \beta_o + \beta_1 X_1 + \ldots . \beta_p X_p$$

The *log-odds* changes *linearly* as a function of variation in the explanatory variables. The practical use and conceptual framework of logistic regression is therefore *closely related* to that of multiple regression.

However, the *logit* is totally defined by π. If logit (π) = ln ω = η, then:

$\pi = \exp(\eta) / [1 + \exp(\eta)]$. [because $\pi = \omega / (1+\omega)$.]

[the above function that transforms *logits* into proportions (π) is called the *logistic function*]

Consequently, the logistic regression model also *differs importantly* from ordinary regression:

1- the variance of π is a function of the mean response
2- the model contains no additional parameter like σ^2

Because logit (π) and π are equivalent, the model:

$$logit\,(\pi) = \beta_o + \beta_1 X_1 + \ldots . \beta_p X_p$$

is used for convenience of interpretation (the response is linear).

The mean and variance specifications of that model are:

$$\mu\{Y \mid X_1, \ldots ., X_p\} = \pi \qquad Var\{Y \mid X_1, \ldots .. X_p\} = \pi(1 - \pi)$$

Maximum Likelihood versus least squares parameter estimation

In the generalized linear model framework, the method of *least squares* parameter estimation is replaced by *maximum likelihood*.

Recall that the method of *least squares* chooses regression coefficients that *minimize the sum of squared residuals*. It minimizes the amount of *unexplained variation* in the response variable.

The method of *maximum likelihood* flips things around: it chooses the regression coefficients that *maximize the joint probability of the predicted values*. It maximizes the amount of *explained variation* in the response variable.

Maximum likelihood estimation of regression coefficients for a binary response

Every observation in the data set can only have a response value of 0 or 1. The regression model predicts for each of these observations a probability $\hat{\pi}_i$. To do this,

a) the value for each level of the explanatory variables (e.g. $X_1 = 1$, $X_2 = 5$, $X_3 = 3$..) are substituted in the regression equation and a *logit* is obtained

b) the *logit* value is transformed to a proportion (π) through the logistic equation.

c) the *probability of obtaining* an observed values is called the ***likelihood*** of a value. Symbol for likelihood is $\hat{\pi}_i$.

- If a given response is 1, its predicted response is $\hat{\pi}_i$, the predicted probability of a *yes* at that level of the X variables (i.e., x_{1i}. x_{2i},….x_{pi}).

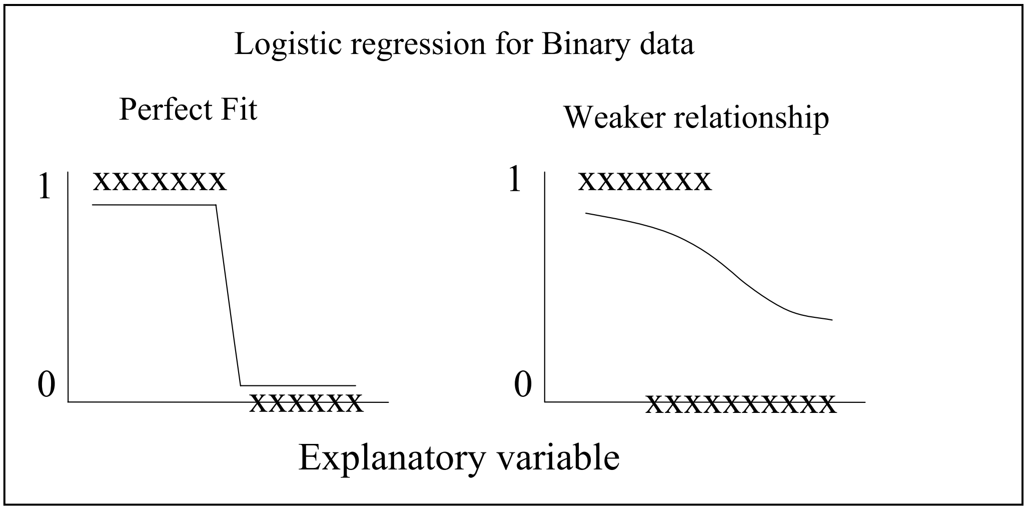
- If a given response is 0, its predicted response is 1 - $\hat{\pi}_i$, the predicted probability of a no.

The model obtained by maximum likelihood specifies the regression coefficients in $logit(\pi) = \beta_o + \beta_1 X_1 + ... \beta_p X_p$ such that the ***product of all likelihoods*** is maximum. The maximum likelihood method fits "the best" possible model given the data at hand.

It is possible to achieve perfect prediction of the data if there is *no overlap* between the values of the response variable (the y = 1 and y = 0 observations) across the values or levels of all explanatory variables. In such a case, JMP would produce a message that the "*regression coefficients are unstable*".

Without any overlap between the values of the response variable across the levels of the explanatory variable, the product of all likelihoods would be 1 x 1 x 1 x 1……..= 1.

In practice however, some 0s and 1s are present together at the same level (s) of the explanatory variable (s), so prediction of the model is never perfect (i.e., some $\hat{\pi}_i$ will be different from 1 or 0).



Interpretation of Odds and Odds Ratio

The logistic regression model is:

$$logit\,(\pi) = \beta_o + \beta_1 X_1 + \ldots\beta_p X_p$$

The *logit* is the log of the *odds* (ln (π / 1- π) : thus e $^{\text{logit}\,(\pi)}$ yields the odds.

So the odds that a response is positive (i.e., Y = 1) at some level of X1,……Xp are:

$$\omega = e^{\left(\beta_o + \beta_1 X_1 + \ldots\ldots\ldots\beta_p \cdot X_p\right)}$$

For example, the odds that Y = 1 at X_1 = 0, X_2 = 0,…..X_p = 0 equals $exp(\beta_o)$ or e^{β_o}

The ratio of the odds (or odds ratio) that Y = 1 when X_1 = A <u>relative to the odds</u> when X_1 = B is:

$$\frac{\omega_A}{\omega_B} = e^{[\beta_1(A\text{-}B)]}$$

when the value of all other explanatory variables in the model are held constant.

[Recall that when 2 populations are the same, then the odds ratio is: $\omega_2 / \omega_1 = 1$. Thus a β near 0 means that there is no association between an explanatory variable and the response ($e^0 = 1$)]

So as *X1 increases by 1 unit* (i.e. A – B =1), the odds of a response (i.e., Y = 1) *changes by a multiplicative factor* of exp (β1) = $e^{\beta 1}$, when other variables are held constant.

Example: Survival in the Donner party. In 1846 the Donner and Reed family travelling by covered wagon got stuck in a snow storm in October in the Sierra Nevada. By the next April when they were rescued, 40 of the 87 people had died from starvation and cold stress. Anthropologists considered mortality in the 45 people aged more than 15 years to investigate whether females are better able to withstand harsh conditions than man.

The specific question is: *For any given age, were the odds of survival different between males and females?*

<Display 20.1>

To assess that question, we use JMP to fit the model:

$$Y = \beta_o + \beta_1 \, female + \beta_2 \, age$$

Response(Y):

Survival, coded as **0 = survived, 1 = died** **{nominal}****

Explanatory (X's):

Sex: *indicator variable*, female =1, male = 0 {continuous}

Age {continuous}

**** NOTE:**

A logistic regression usually model the probability that Y = *yes*.

When you code the response as numbers (e.g. 0 - 1), because you specify the response variable as NOMINAL, JMP always fit the probability of the event corresponding to the SMALLEST number.

For the example above, if you really wanted to code survived = 1, you would have to code died =2, if you want to model the *probability of survival*.

Donner party, JMP analysis:

Coding:	Survived = 0, Died = 1	{nominal}
	Age	{continuous}
	Female = 1, Male = 0	{continuous = INDICATOR variable}

Use the Fit Model platform, with survival as the Y, and Sex and Age as the X:

Parameter Estimates

Term	Estimate	Std Error	ChiSquare	Prob>ChiSq
Intercept	1.63311508	1.1102437	2.16	0.1413
fem ind	1.5972907	0.7555008	4.47	0.0345*
Age	-0.0782039	0.0372874	4.40	0.0360*

Effect Likelihood Ratio Tests

Source	Nparm	DF	L-R ChiSquare	Prob>ChiSq
fem ind	1	1	5.03443702	0.0248*
Age	1	1	6.02999061	0.0141*

Both *age* and *female* explain a significant amount of variation in the probability of *survival* (Y). The regression equation is:

$$\begin{array}{lcccc} \text{Logit}\ (\pi) = & 1.63 & - & 0.078\ \textit{age} + & 1.60\ \textit{female} \\ & (1.11) & & (0.037) & (0.76) \end{array}$$

Each estimated coefficient β_j in the logistic regression has a normal sampling distribution, which implies that the *standard normal distribution* can be used for statistical inferences (i.e. Z-tests and CI estimation with Z-distribution: JMP uses related X^2 tests). Tests provided under "Parameter estimates" are based on a normal approximation are called *Wald's* tests. They produce good approximations *if sample size is large*.

Confidence intervals for regression coefficients can be obtained. They are computed from maximum likelihood iterations.

Likelihood ratio tests (explained later) are better to test the significance of effects included in a model.

****

Note that you can also fit this model with the **Generalized Linear Model approach.**

To do this in the Fit Model platform, enter survival as Y (categorical) and Age and Sex as before. Then chose the *Generalized Linear Model* Personality, *Binomial* Distribution, and *Logit* Link Function.

****

Test and Interpretation of single coefficients

General Age effect (continuous variable):

The estimated coefficient for age is – 0.078, with SE = 0.037.

The 95 % CI is: -0.162, -0.014.

Taking the anti-logarithm of the estimate and CI endpoints, we get the following summary statement:

It is estimated that the odds of survival change by a factor 0.92 (i.e. $e^{-0.078}$) for each one year increase in age (95% CI for this *multiplicative* change is 0.85 to 0.99). So the odds of survival decrease by 0.92, or by 8 %, for every extra year of age (averaging for gender effect).

Specific prediction at different Ages:

For example, compare odds of survival for women 50 years old with women 20 years old. The *odds ratio* is calculated as $\omega_a / \omega_b = \exp[\beta_2 (A - B)]$:

So $\omega_{50} / \omega_{20} =$ exp [- 0.078 (50 – 20)] = 0.096, or about 1 /10. So the odds of survival of 20-year-old women were about 10 times the odds of survival of 50-year-old women.

Effect of Sex (continuous-Coded as DUMMY variable):

The odds-ratio of survival for women (female = 1 = A) compared to men (female = 0 = B) of the same age is:

Exp[1.60 (1 – 0)} = 4.94

Thus the odds of *survival* were about five times greater for women than men of the same age (i.e., after correcting for the effect of Age).

Prediction of a probability of survival (i.e., finding the likelihood of a value, $\hat{\pi}_i$):

We use the logistic function that relates the *logit* to π.

For example, the predicted log-odds survival for a 30-year-old male are:

$$\begin{aligned} \text{Logit}(\pi) &= 1.63 - 0.078 \text{ age} + 1.60 \text{ female} \\ &= 1.63 - 0.078\,(30) + 1.60\,(0) \end{aligned}$$

$= 1.63 – 2.34 + 0$
$= - 0.71$

The inverse of the *logit* function is the *logistic function*:

$\pi = \exp(\eta) / [1 + \exp(\eta)]$.

Where $\eta = \text{logit}(\pi)$. So:

$\hat{\pi} = \exp(- 0.71) / [1 + \exp(- 0.71)] = 0.33$

Indicating that a 30 year-old male had a 0.33 chance of surviving the winter.

<u>Or similarly:</u>

$\text{Logit}(\pi) = \ln(\text{Odds}) = - 0.71$

$\text{Odds} = e^{- 0.71} = 0.492$

$\pi = 0.49 / (0.49 +1) = 0.33$ [because $\pi = \omega / (1+\omega)$]

Test for several coefficients

Recall that to compare the simultaneous effect of many explanatory variables in multiple regression, we compared the difference between the error sums of squares associated with a full and reduced model. {Extra SS test: F = [(ESS reduced – ESS full) / Extra df] / Best estimate of σ^2 }.

In logistic regression the residuals are not homogenous across levels of the explanatory variable, so the Extra SS method cannot be used with *untransformed* residuals.

Transformed residuals are used instead (to <u>normalize</u> their distribution): one type of transformed residuals is called the *Deviance residual.* With such residuals, the *Extra Sum of Square* test becomes a *Deviance* test.

Instead of using $\mathbf{y_i} - \hat{\pi}_i$ to calculate residuals, the deviations are transformed by a function, g(x), which makes the residuals homogenous across the values of X.

The function chosen is g(x) = *twice the logarithm of the likelihood function* (i.e., 2 × log $(\mathbf{y_i} - \hat{\pi}_i)$)

What happens when we take twice the logarithm of the likelihood function?

g(x) for predicted values, $\hat{\pi}_i$:

The likelihood of the ith mean is $\hat{\pi}_i$. We simply take twice the log of that value: g($\hat{\pi}_i$) = 2 log ($\hat{\pi}_i$).

g(x) for the observed values, yi:

The log likelihood for yi = 1 is log yi = log 1 = 0
The log likelihood for yi = 0 is log (1 – yi) = log 1 = 0

Because the log likelihood of yi = 1 or yi = 0 is 0, **g (y_i -** $\hat{\pi}_i$) is:

$$
\begin{aligned}
\mathbf{g}\ (\mathbf{y_i} - \hat{\pi}_i) &= 2\log(yi) - 2\log(\hat{\pi}_i\mathbf{)} \\
&= 0 - 2\log(\hat{\pi}_i) \\
&= -2\log(\hat{\pi}_i)
\end{aligned}
$$

To account for whether a given yi is greater (= 1) or lower (= 0) than its estimated mean ($\hat{\pi}_i$**),** the *deviance residual* is defined as:

$+\sqrt{-2log(\hat{\pi}_i)}$ if yi =1, and
$-\sqrt{-2log(\hat{\pi}_i)}$ if yi =0.

The ***Deviance*** is the sum of the ***squared deviance residuals***:

Deviance = $\sum -2\log(\hat{\pi}_i)$ where $\hat{\pi}_i$ **= π for yi = 1**
$\hat{\pi}_i$ **= 1 - π for yi = 0**

It represents the discrepancy between the *observed* responses and those *predicted* by the fitted model (and is equivalent to the *error SS* in multiple regression).

The *Drop-in-deviance* test is analogous to the *Extra-sum-of-squares F-test* in ordinary regression:

Drop in deviance = Deviance from *reduced* model – Deviance from *full* model

Drop in *df* = Difference in number of parameters (*full* model – *reduced* model)

If the drop in deviance is *small*, the reduced model explains about the same amount of variation in the response (Y) as the full model.

If the drop in deviance is *large*, the reduced model is not adequate as compared to the full model.

The drop in deviance test is important to *assess whether a model fits well* because the residuals from a logistic regression with binary responses (0-1) cannot be used to do this (they fall on 2 parallel lines on each side of the fitted regression line).

Drop in deviance statistics have a χ^2 distribution with *d* df, where *d* is the difference in number of parameters. A *small* p-value indicates that the reduced model is *not* adequate.

Donner Party example

Is there a difference between male and female survival probability after accounting for the effect of age?

Full model: logit (π) = $\beta_o + \beta_1 age + \beta_2\ female$

Reduced model: logit (π) = $\beta_o + \beta_1 age$

Full model

```
Whole Model Test
Model        -LogLikelihood          DF   ChiSquare   Prob>ChiSq
Difference         5.285129           2   10.57026    0.0051*
Full              25.628142
Reduced           30.913271


RSquare (U)                      0.1710
Observations (or Sum             45
Wgts)
```

Reduced model

```
Whole Model Test
Model          -LogLikelihood          DF   ChiSquare   Prob>ChiSq
Difference           2.767910           1     5.53582      0.0186*
Full                28.145361
Reduced             30.913271


RSquare (U)                         0.0895
```

Observations (or Sum Wgts)	45

By default, JMP reports a full and reduced model for each logistic regression as a test for the overall model: this is called the *Whole-Model Test.*

The Whole-Model Test treats the model you specify as a *full* model and compares it to a *reduced* model, which is a model with an intercept only. So the Whole-Model Test is equivalent to a drop in deviance test that investigates whether all coefficients in the model (except the intercept term) are zero.

JMP does not report model *Deviance*. JMP reports the value of the - *Log Likelihood* of a Model:

- *Log Likelihood* = $\sum -\log(\hat{\pi}_i)$,

which is very similar to the deviance [Deviance = $\sum -2\log(\hat{\pi}_i)$].

The – log likelihood is calculated by JMP by summing the negative logarithm of the *predicted probabilities* ($\hat{\pi}_i$ **).**

Maximizing the product of all likelihoods (the maximum value of this product is 1) is the same as *minimizing* the negative sum of the logs of the likelihoods.

The maximum possible value of the product of all likelihoods is 1. This yields - ln 1 = 0;
A smaller product of likelihoods, 0.3, would yield - ln 0.3 = 1.2;
A smaller product of likelihoods, 0.1, would yield - ln 0.1 = 2.3;
The above full model has a – log likelihood of 28.14, which means the product of all likelihoods is < 0.0000001)

So the bigger the difference in –Log likelihood between the full and reduced model, the better is the fit of the full model.

You can therefore use the value of the – log likelihood to perform a drop-in-deviance test, by comparing *twice* the difference in the - log likelihood between *full* and *reduced* models:

χ^2 = (2 * -log likelihood *reduced*) - (2 * -log likelihood *full*)

in the above example,

χ^2 = (2 * 28.14) – (2 * 25.63) = 5.03, with 3 - 2 df.

For a $\chi^2 = 5.03$ with 1 df, P = 0.025, suggesting that there was a difference between male and female survival probabilities after accounting for age.

The drop in deviance test assessing the effect of a single explanatory variable is called the *Likelihood ratio test* in JMP.

*** The likelihood ratio test (always provided in JMP) is **better than** the Wald's test to assess significance of the effect of explanatory variables.***

Logistic Regression for Binomial Responses

The logistic model for binary response variables (0-1) extends to cases when the responses are *proportions* of binary counts (i.e., binomial proportions).

Proportions of *binary counts* take 2 forms:

a) Grouped binary responses, calculated from a sample (proportion of parasitized insects out of samples of 200 insects; proportion of a given number of 50-year-old patients with lung cancer).

b) Proportion based on a count for individual subjects (proportion of 100 cells from each subject with chromosomal aberrations; proportion of times out of 100 that a baseball player hits the ball).

As for logistic regression for *binary* response variables, logistic regression for *binomial proportions* models the population proportion or probability (π) through a *logit link* with a linear function of regression coefficients.

Binomial proportion

A binomial response variable is measured as a *count* of binary events (yes or no) out of a total number of observations. The *binomial denominator m* does not need to be the same for every sample, but it *must be known*.

This is different from a *continuous proportion,* which is the ratio of 2 *continuous variables.* For example, the proportion of fat per unit of weight in ants; the proportion of water by weight in leaves, etc… Continuous proportions do not have a *binomial distribution*. Continuous proportions as response variables *must* be handled with *least squares regression* (e.g., ANOVA, linear regression, multiple regression).

Example: We investigate the number of bird species that got extinct on 18 islands during a 10-year period. Species were monitored on each island in 1949. All the species present were considered "at risk". Presence / absence was monitored again 10 years later: species that were no longer present were considered "extinct". The response variable for each islands takes this form: # extinct / # at risk.

Example of data for the species extinction problem:

ISLAND	EXTINCT / NOT EXTINCT
Ulkokrunni	5/70
Maakrunni	3/64
Ristikari	10/56
Etc…..	Etc….

The interpretation and framework for analysis of logistic regression for binomial proportions is very similar to logistic regression for binary variables.

For binomial proportions we have:

1. Y is a binomial count, where Yi = sum of all 1's out of the m_i binary responses in a sample.

2. $\overline{\pi}_i$ = Y_i / mi is the *observed* response proportion for observation i.

3. We fit the model logit(π) = $\beta_o + \beta_1 X_1 + ... + \beta_p X_p$, which yields predicted responses ($\hat{\pi}_i$) for each level of the explanatory variables.

Example: Island Size and Bird Extinction. Is extinction rate is birds a function of the area available to a species?

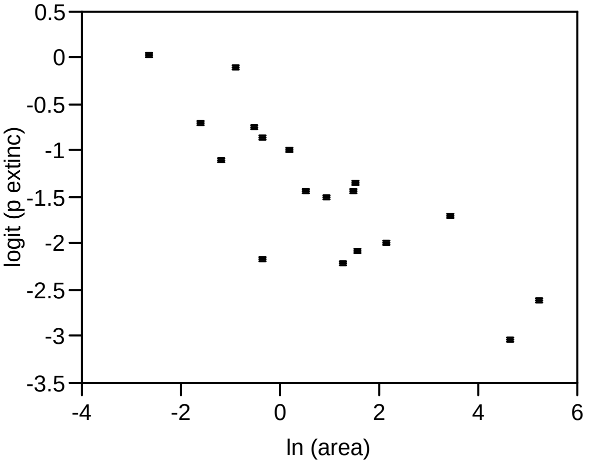
Model Assessment:

1. **Scatterplot of observed logits VS explanatory variable**

Every observation i (sample) has a response proportion: $\overline{\pi}_i = Y_i / m_i$: the observed logit is $log\left[Y_i / (m_i - Y_i)\right] = \log\left[\overline{\pi}_i / (1 - \overline{\pi}_i)\right]$.

Plotting the observed logits vs one or more explanatory variable is useful for visual examination of linearity (in *Fit Y by X or Multivariate platform* in JMP). It parallels the use of scatterplots in ordinary regression.

Example: Island size and *bird extinction* example:



NOTE. If some of the observed proportions are either 0 or 1, the logit function is undefined: to produce a display, add a small amount to the numerator and denominator (e.g., 0.05).

This *is not necessary,* however, to fit an actual logistic regression model in JMP.

2. Examination of residuals

Two types of residuals are used in logistic regression for binomial counts: *Deviance residuals* and *Pearson residuals* (see Sleuth Chapter 21). Plotting those as a function of predicted values is useful to check whether the assumptions of the model are met.

These residuals are available in the Generalized Linear Model platform.

Fitting the Logistic regression Model in JMP

To fit the model using the logistic regression platform (in Fit Model), you need to get a data table that looks like this (this could involve using the “Stack” option under Tables):

ISLAND	AREA	EXTINCTION	COUNT
A	100	Not Extinct	70
A	100	Extinct	5
B	145	Not Extinct	36
B	145	Extinct	7
Etc…			

Logistic Binomial Analysis:

Response: **Extinction (extinct = 0)** **{nominal}**
Explanatory: Log(Area) {continuous}
Count {continuous, Frequency}

In Fit Model, choose Extinction as Y, Log (area) as X, Count as frequency.

```
Whole Model Test
Model        -LogLikelihood          DF   ChiSquare   Prob>ChiSq
Difference         11.73149           1    23.46298       <.0001
Full              295.82138
Reduced           307.55287


RSquare (U)                      0.0381
Observations (or Sum                740
Wgts)


Parameter Estimates
Term            Estimate   Std Error   ChiSquare   Prob>ChiSq
Intercept     -1.4786275   0.1139677      168.33       <.0001
ln area       -0.2432824   0.0527819       21.24       <.0001

For log odds of Ext/Not Ext

Effect Likelihood Ratio Tests
                                          L-R
Source      Nparm       DF   ChiSquare   Prob>ChiSq
ln area         1        1  23.4629788       <.0001
```

Tests of individual regression coefficients under “Parameter Estimates” are *Wald’s* tests.

Likelihood ratio tests are drop-in-deviance test.

You may also fit a logistic regression model using the Generalized Linear Model platform.

Data should look like this:

ISLAND	AREA	EXTINCT	TOTAL
Ulkokrunni	185	5	80
Maakruni	105	3	70
Ristikari	31	10	76
Isonkivenletto	9	6	57
Etc….			

Personality: Generalized Linear Model
Distribution: Binomial
Link function: Logit

Then enter EXTINCT and TOTAL as Y and Log (area) as X.

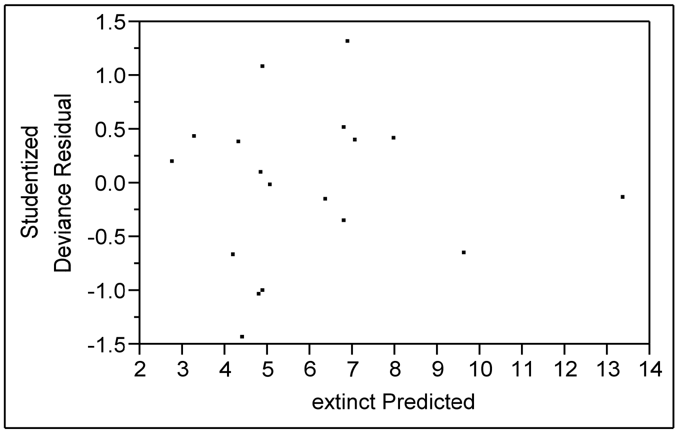
```
Whole Model Test
Model          -LogLikelihood   L-R ChiSquare   DF   Prob>ChiSq
Difference         11.7314895         23.4630    1      <.0001*
Full               34.2593747
Reduced            45.9908642

Goodness Of Fit      ChiSquare     DF   Prob>ChiSq
Statistic
Pearson                 7.7603     16       0.9557
Deviance                8.0302     16       0.9480

          AICc
       73.3187

Parameter Estimates
Term             Estimate   Std Error   L-R ChiSquare   Prob>ChiSq    Lower CL    Upper CL
Intercept       -1.478634   0.1139678       209.89025      <.0001*   -1.707377   -1.260017
log(Area)       -0.243294   0.0527826       23.462979      <.0001*   -0.349505   -0.142179
```

Studentized Deviance Residual by Predicted



Output is the same as before (95% CI are provided directly here: they must be requested if using Nominal Logistic Regression).

Deviance Goodness-of-fit test

Before drawing conclusions based on the analysis above, we must decide whether model fit is adequate. There are many potential reasons why this would not be so:

1) The effects of the explanatory variables are not linear on the *logit* scale. In such case the model would need extra terms, like quadratic terms (βX^2) or interaction terms,

to fit the data adequately. (This could be detected by inspection of scatterplots of Xs vs Y)

2) The response counts do not conform to a binomial distribution: There are some outliers, or there is *Extra-binomial* variation in the response (e.g., because the binary variables are not independent, important explanatory variables are not included in the model, etc…).

A single test, *the Deviance Goodness-of-fit test*, can be used to address these issues.

Recall that in linear regression with replication, the lack-of-fit test compared the separate-mean model (one-way ANOVA) to the simpler linear regression model.

In logistic regression for binomial counts, we can compare the observed proportions (which are sample means) to the proportions predicted by the reduced model. To do this, the *log likelihood* of the model of interest is compared to the *log likelihood* of a model fitted to each *observed* population proportion. This test is valid only if $m_i > 5$ (denominator of the binomial counts) for every observations.

The *null hypothesis* is that the *model of interest fits well*. The *alternative hypothesis* depends on the model of interest:

1) If the model *contains all terms that might be important, including polynomials (i.e. quadratic) and interactions*, and there are no obvious *outliers*, then the *alternative hypothesis* is that there is *extra-binomial variation* in the data. In such case, data should be analyzed with a modified method: the *quasi-likelihood approach* provides an easy alternative (the test is done by hand after adjusting SE to take into account the extra variation; see Sleuth p.622).

   Alternatively in the Generalized Regression Platform, ask for "overdispersion tests and intervals" and the logistic regression procedure will automatically correct for overdispersion!

2) If the model of interest is too simple, then *more structure* (extra terms) is needed to fit the data adequately.

Model of interest: $logit\left(\pi_i\right) = \beta_o + \beta_i X_{i1} + \ldots.. X_{ip}$

Saturated model: $logit\left(\pi_i\right) = \alpha_i$ (i = 1, 2,…n, where n is the number of observations, and α are the observed population proportions)

In JMP the Deviance Goodness-of-fit test is called the lack of fit test if the Logistic Regression platform is used. For the *Island size* and *bird extinction* example:

Lack Of Fit

Source	DF	-LogLikelihood	ChiSquare
Lack Of Fit	15	3.06127	6.122547
Saturated	16	292.76011	Prob>ChiSq
Fitted	1	295.82138	0.9776

The Degrees of Freedom (DF) for the Saturated model is the number of unique binomial responses (here 16 because 2 islands had the same area). The DF for the fitted model is the number of parameters (not counting the intercept). Here, these are 16 and 1 DF, respectively. The Lack of Fit DF is the difference between the Saturated and Fitted models, 15. The lack of Fit chi-square is not significant ($\chi^2_{15} = 6.12$, P = 0.98), which indicates that *no extra terms are needed* in the model, and that no *extra-binomial variation* is present in those data.

An equivalent test (called Goodness of Fit) is provided in the Generalized Linear Model platform.

Final Interpretation (using output from the Logistic Regression platform):

A one unit change in island area was associated with a change in the odds of extinction of exp(β) = exp (-0.243). Because the explanatory variable was logged, we conclude that for each doubling of island area, the odds of extinction changed by 2^{β} (i.e. $2^{-0.243}$), or 0.84. In other words, the odds of extinction for a species on an island of size 2A were 84 % of the odds of extinction on an island of size A. The 95% CI for this multiplicative change was ($2^{-0.349} – 2^{-0.142}$, i.e. 0.78 – 0.90)

**Results for logistic regression for binomial counts in the last version of JMP differ slightly from analyses presented in Sleuth.