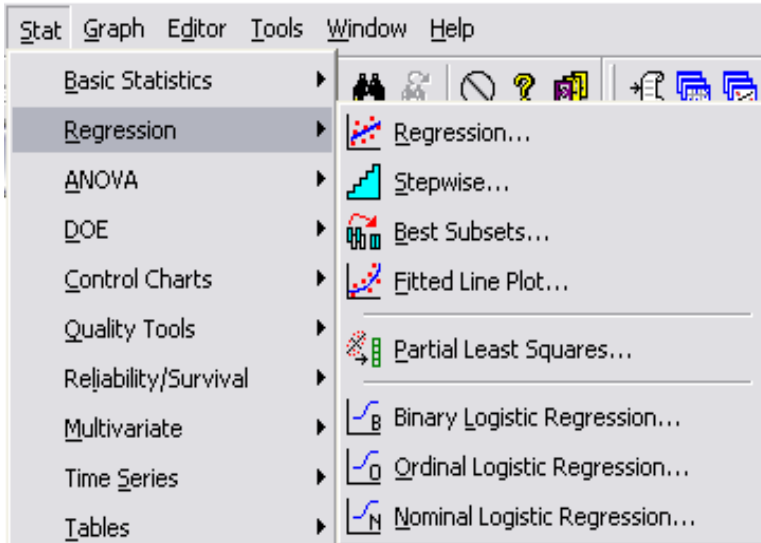


Lab 12

Logistic Regression

Logistic regression relates a binary response variable (0-failure/1-success) to predictor variables. In Minitab it is run under Stat/Regression/Binary Logistic Regression.



Example 1: Vaso-Constriction Data

(http://ftp.sas.com/techsup/download/sample/samp_lib/statsampDocumentation_Examples_f00000061.html)

Finney (1947) lists data on a controlled experiment to study the effect of the rate and volume of air on a transient reflex vaso-constriction in the skin of the digits. 39 tests under various combinations of rate and volume of air inspired were obtained. The end point of each test is whether or not vaso-constriction occurred.

```

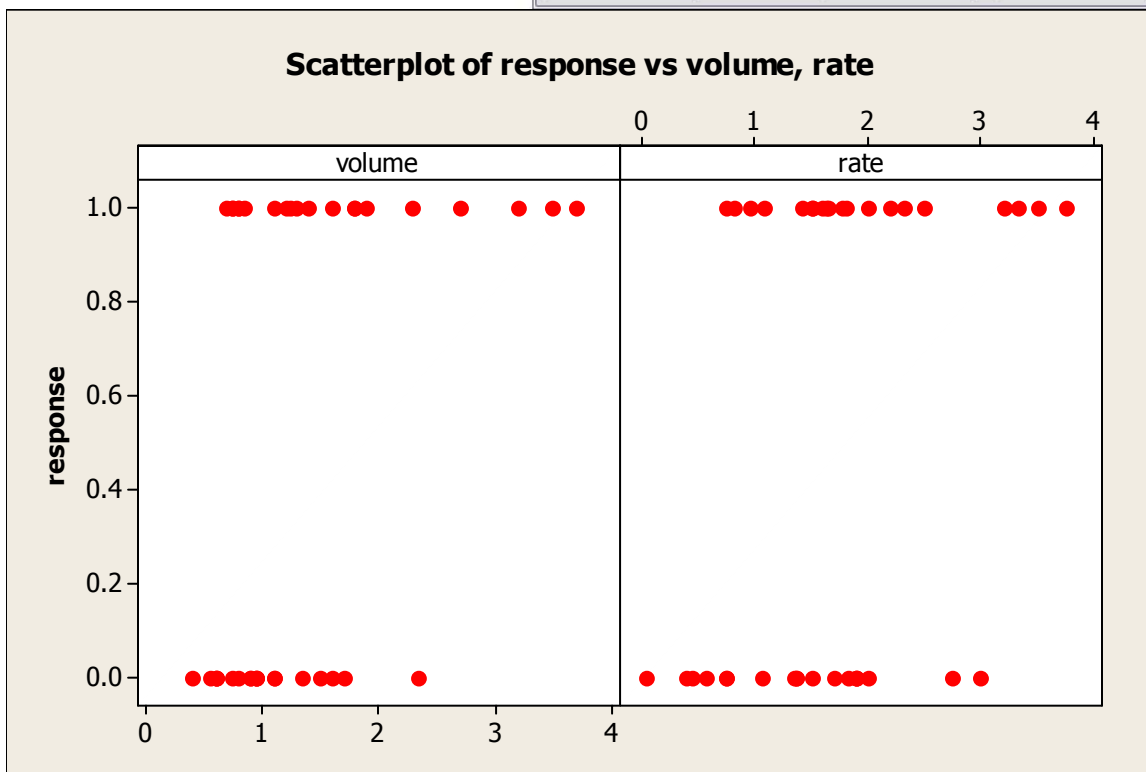
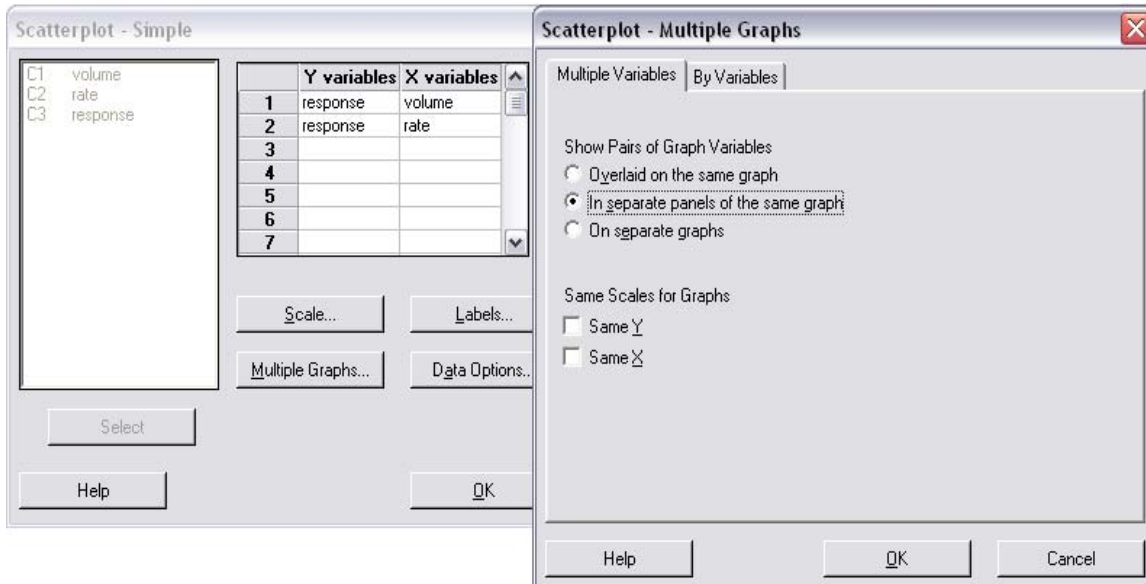
volume  rate  response
3.7    .825  1
3.5    1.09  1
1.25   2.5   1
.75    1.5   1
.8     3.2   1
.7     3.5   1
.6     .75   0
1.1    1.7   0
.9     .75   0
.9     .45   0
.8     .57   0
.55    2.75  0
.6     3.0   0
1.4    2.33  1
.75    3.75  1
2.3    1.64  1
3.2    1.6   1
.85    1.415 1
1.7    1.06  0
1.8    1.8   1
.4     2     0
.95    1.36  0
1.35   1.35  0
1.5    1.36  0
1.6    1.78  1
.6     1.5   0
1.8    1.5   1
.95    1.9   0
1.9    .95   1
1.6    .4    0
2.7    .75   1
2.35   .03   0
1.1    1.83  0
1.1    2.2   1
1.2    2.0   1
.8     3.33  1
.95    1.9   0
.75    1.9   0
1.3    1.625 1
  
```

Example 2: nx2 Categorical Table

Angina Treatment data Example 10.11 p.402

	Timolol	Placebo
Angina free	44	19
Not angina free	116	128
Total	160	147

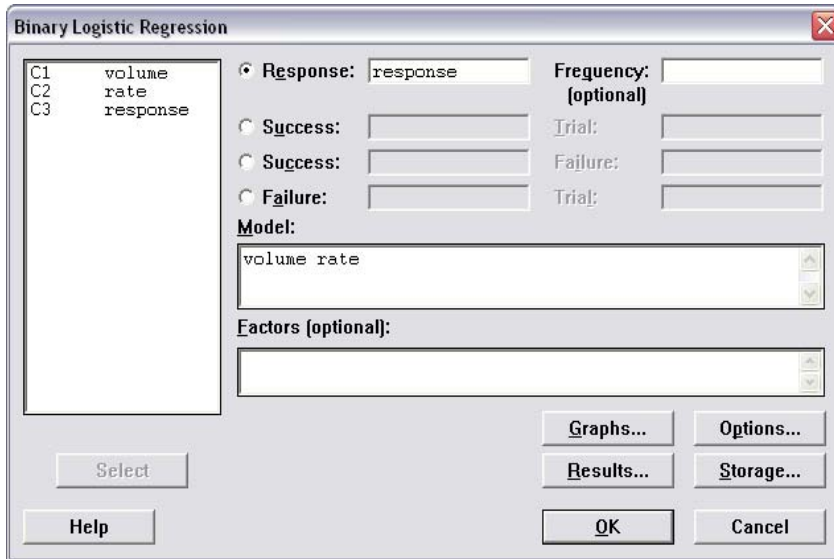
What do we do first? Always **plot the data**.
Create scatterplots to see how the binary response variable relates to the predictor variables.



This is what typical logistic data look like. The response is either 0 or 1 (sometimes called “failure” and “success”), which is why the data points are either low or high. We can get an idea of the relationship between the predictor variables and the response. For example, above, volume is slightly greater when the response is 1 than when 0. Similarly, rate is greater when the response is 1 than when 0. We can expect to see these relationships in the output of our logistic regression fit.

Logistic regression is under Stat/Regression/Binary Logistic Regression.

Input the response variable (response) in response, and the predictor variables (volume and rate) in model.



Here is abbreviated output from the logistic regression. Below the output covered are

1. assessing model fit,
2. significance of predictor variables and their effect on the response, and
3. interpretation of the odds ratio.

Binary Logistic Regression: response versus volume, rate

Link Function: Logit

Response Information

Variable	Value	Count
response	1	20 (Event)
	0	19
	Total	39

Logistic Regression Table

Predictor	Coef	SE Coef	Z	P	Odds Ratio	95% CI	
						Lower	Upper
Constant	-9.52959	3.23308	-2.95	0.003			
volume	3.88215	1.42856	2.72	0.007	48.53	2.95	798.02
rate	2.64912	0.914191	2.90	0.004	14.14	2.36	84.85

Goodness-of-Fit Tests

Method	Chi-Square	DF	P
Pearson	39.0128	35	0.294
Deviance	29.7723	35	0.718
Hosmer-Lemeshow	7.7996	8	0.453

1. Assessing model fit using the *Goodness-of-Fit* test.

How well the model fits the observed data is assessed by a number of ways; we will use the Deviance, but all three statistical tests (Pearson; Deviance; Hosmer-Lemeshow) are similar. A low p-value indicates that the predicted probabilities deviate from the observed probabilities in a way that the binary distribution does not predict. When the Deviance test is significant (p-value<0.05) it means the model does not describe the data well. This means that the null hypothesis of data fitting the model can be rejected. In other words the goodness-of-fit is not that good.

H_0 : model does fit the data

H_A : model does not fit the data

Pearson and Deviance are both types of residuals. **The larger the p-value the better is the fit of the model to the data.** When the model does not fit (reject H_0) it would be best to try alternative models and opt for the one that produces the largest p-values.

Note well: No model has an exact fit. With enough data the goodness of fit test will *always* reject the model. However, what one is looking for is whether the model is good enough for analysis purposes. One can still make inferences from a model not fitting well but caution is needed.

In our output, the Deviance p-value of 0.718 does not give us significant evidence that our model is not fitting our data. That is, the model adequately describes the relationship between the response and predictor variables in the data. Analysis can continue.

Goodness-of-Fit Tests			
Method	Chi-Square	DF	P
Pearson	39.0128	35	0.294
Deviance	29.7723	35	0.718
Hosmer-Lemeshow	7.7996	8	0.453

2. Significance of predictor variables and their effect on the response

As in any regression, we can look at the p-values to see whether the predictor variables each have a significant relationship with the response variable. The coefficient will tell us what that relationship is (positive or negative). In Logistic regression, interpretation is typically done in terms of the odds ratio, which we cover next.

In our model, both predictor variables rate and volume are significant (p-values < 0.01), and have a positive effect on the response (coefficients are both positive). That is, as volume increases, the response is more likely to be 1. Similarly for rate.

Logistic Regression Table

Predictor	Coef	SE Coef	Z	P	Odds	95% CI	
					Ratio	Lower	Upper
Constant	-9.52959	3.23308	-2.95	0.003			
volume	3.88215	1.42856	2.72	0.007	48.53	2.95	798.02
rate	2.64912	0.914191	2.90	0.004	14.14	2.36	84.85

3. Interpretation of the odds ratio.

How do I interpret odds ratios in logistic regression?

(Borrowed graciously from UCLA Academic Technology Services at <http://www.ats.ucla.edu/stat/sas/faq/oratio.htm>)

Let's begin with probability. Let's say that the probability of success is .8, thus

$$p = .8$$

Then the probability of failure is

$$q = 1 - p = .2$$

The odds of success are defined as

$$\text{odds(success)} = p/q = .8/.2 = 4,$$

that is, the odds of success are 4 to 1. The odds of failure would be

$$\text{odds(failure)} = q/p = .2/.8 = .25,$$

that is, the odds of failure are 1 to 4. Next, let's compute the odds ratio by

$$\text{OR} = \text{odds(success)}/\text{odds(failure)} = 4/.25 = 16$$

The interpretation of this odds ratio would be that the odds of success are 16 times greater than for failure. Now if we had formed the odds ratio the other way around with odds of failure in the numerator, we would have gotten something like this,

$$\text{OR} = \text{odds(failure)}/\text{odds(success)} = .25/4 = .0625$$

Interestingly enough, the interpretation of this odds ratio is nearly the same as the one above. Here the interpretation is that the odds of failure are one-sixteenth the odds of success. In fact, if you take the reciprocal of the first odds ratio you get

$$1/16 = .0625$$

Another example

This example is adapted from Pedhazur (1997). Suppose that seven out of 10 males are admitted to an engineering school while three of 10 females are admitted. The probabilities for admitting a male are,

$$p = 7/10 = .7 \quad q = 1 - .7 = .3$$

Here are the same probabilities for females,

$$p = 3/10 = .3 \quad q = 1 - .3 = .7$$

Now we can use the probabilities to compute the admission odds for both males and females,

$$\text{odds(male)} = .7/.3 = 2.33333$$

$$\text{odds(female)} = .3/.7 = .42857$$

Next, we compute the odds ratio for admission,

$$\text{OR} = 2.3333/.42857 = 5.44$$

Thus, the odds of a male being admitted are 5.44 times greater than for a female.

Interpretation of the odds ratio in Logistic regression.

Our odds ratio for volume is 48.53. With everything else held constant (other predictor variables are fixed), for a one unit increase in volume the model predicts an increase of 48.53 in the odds of the response being a 1 to being a 0. That is, when volume is increased one unit, the response is 48 times more likely to be a 1 than a 0.

Similarly, with everything else held constant, for a one unit increase in rate the model predicts an increase of 14.14 in the odds of the response being a 1 to being a 0. That is, when rate is increased one unit, the response is 14 times more likely to be a 1 than a 0.

Logistic Regression Table

Predictor	Coef	SE Coef	Z	P	Odds Ratio	95% CI Lower	95% CI Upper
Constant	-9.52959	3.23308	-2.95	0.003			
volume	3.88215	1.42856	2.72	0.007	48.53	2.95	798.02
rate	2.64912	0.914191	2.90	0.004	14.14	2.36	84.85

Note that the odds ratio is the exponential of the coefficient. That is, $48.53 = e^{3.88215}$. Alternatively, the coefficient is the natural log of the odds ratio, or $3.88215 = \log(48.53)$.

Odds ratio for an $nx2$ table

To construct a naïve estimate of the odds ratio, first we need to calculate the odds of success for each group, then take the ratio of those.

Consider “Angina free” as a “success”. The odds is the probability of success over the probability of failure.

For the Timolol group, the odds are $(44/160)/(116/160) = 0.3793$.

For the Placebo group, the odds are $(19/147)/(128/147) = 0.1484$.

The odds ratio of the Timolol group over the Placebo group is

$$((44/160)/(116/160)) / ((19/147)/(128/147)) = 0.3793/0.1484 = 2.5554.$$

This means that the Timolol group is 2.5 times more likely to be Angina free than the Placebo group.

Alternatively, the odds ratio of the Placebo group over the Timolol group is

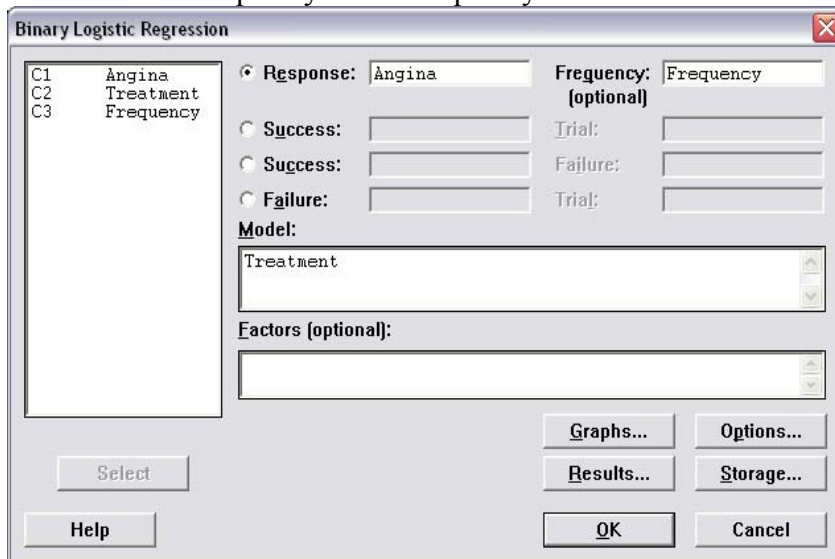
$$((19/147)/(128/147)) / ((44/160)/(116/160)) = 0.1484/0.3793 = 0.3913.$$

This means that the Placebo group is 0.39 times more likely to be Angina free than the Timolol group.

Input the data into Minitab in this way. Angina = 1 is Angina free, the success. The treatment = 1 is the drug Timolol. They are interested in whether the drug improves Angina free probability.

Angina	Treatment	Frequency
1	1	44
1	0	19
0	1	116
0	0	128

In the regression dialog, consider Angina the response, where Treatment is a predictor variable. Put Frequency as the frequency.



Abbreviated output is below. Notice that the treatment odds ratio is 2.56, the same as the one we calculated by hand above. We have more information here, though. We see that treatment is significant in its relationship to the probability of being Angina free.

Binary Logistic Regression: Angina versus Treatment

Logistic Regression Table

Predictor	Coef	SE Coef	Z	P	Odds Ratio	95% CI	
						Lower	Upper
Constant	-1.90759	0.245854	-7.76	0.000			
Treatment	0.938191	0.302972	3.10	0.002	2.56	1.41	4.63