Logistic Regression

Summary

The **Logistic Regression** procedure is designed to fit a regression model in which the dependent variable *Y* characterizes an event with only two possible outcomes. Two types of data may be modeled:

1. Data in which *Y* consists of a set of 0's and 1's, where 1 represents the occurrence of one of the 2 outcomes.
2. Data in which *Y* represents the proportion of time that one of the 2 outcomes occurred.

The fitted regression model relates *Y* to one or more predictor variables *X*, which may be either quantitative or categorical. In this procedure, it is assumed that the probability of an event is related to the predictors through a logistic function. The *Probit Analysis* procedure can be used to fit the same type of data but uses a different functional form.

The procedure fits a model using either maximum likelihood or weighted least squares. Stepwise selection of variables is an option. Likelihood ratio tests are performed to test the significance of the model coefficients. The fitted model may be plotted and predictions generated from it. Unusual residuals are identified and plotted.

Sample StatFolio: *logistic.sgp*

Sample Data:

Two examples will be considered. The first example, from Myers (1990), is contained in the file *fabric.sf3*. It describes the failure of specimens of a fabric subjected to different loads.

Load	*Specimens*	*Failures*
5	600	13
35	500	95
70	600	189
80	300	95
90	300	130

For this data, the dependent variable *Y* is the proportion of specimens at a given load that failed, calculated by *Y = failures / specimens*. There is a single predictor variable *X = Load*. There are a total of *n* = 2,300 specimens.

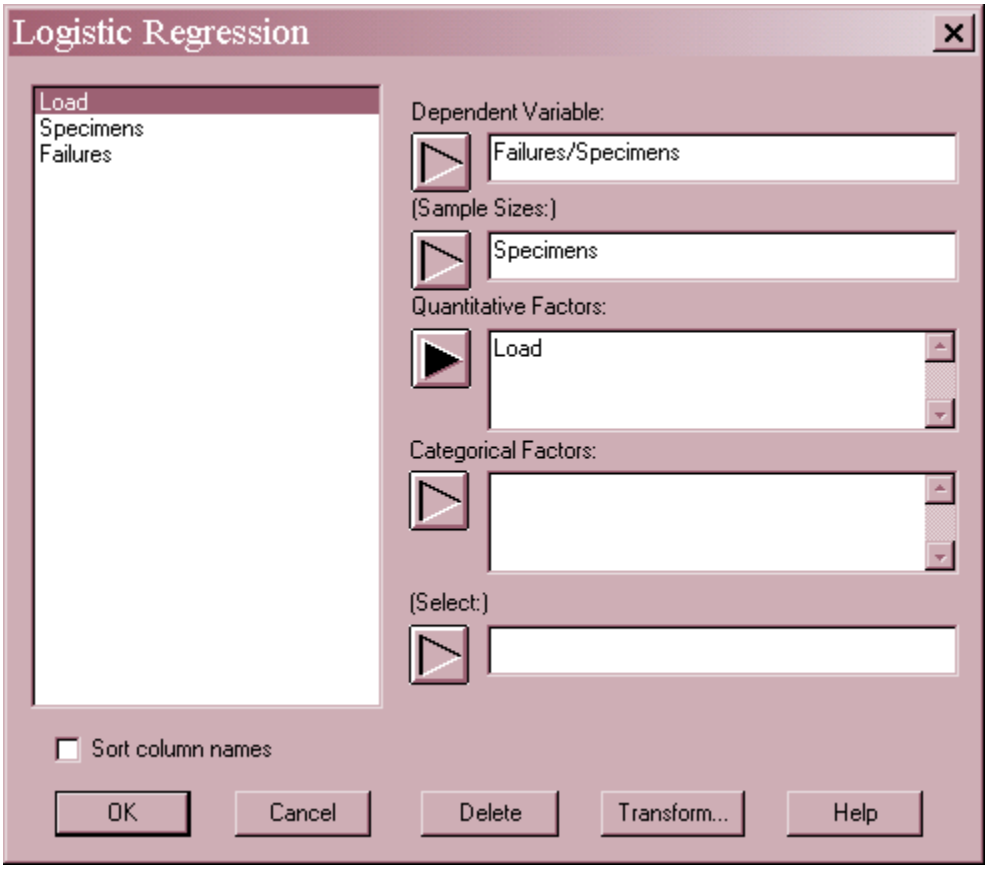
The second data file, *collisions.sf6*, is from Härdle and Stoker (1989). It describes *n* = 58 side impact collisions of automobiles. The response variable *Y* is binary, quantifying whether or not the collision resulted in a fatality. A portion of the file is shown below:

Age	*Acceleration*	*Velocity*	*Fatality*
22	50	98	0
21	49	160	0
40	50	134	1
43	50	142	1
23	51	118	0
58	51	143	1
29	51	77	0
29	51	184	0
47	51	100	1
…	…	…	…

The dependent variable Y = *Fatality* equals 1 if a fatality occurred and 0 otherwise. The predictor variables are the *Age* of the person involved and the *Acceleration* and *Velocity* of the object that hit that person's automobile.

Data Input

The data input dialog box requests information about the input variables:



- **Dependent Variable**: a numeric variable containing the dependent variable Y. Y may consist of either a set of s proportions, each between 0 and 1, or a set of n binary 0's and 1's representing the occurrence or non-occurrence of an outcome.

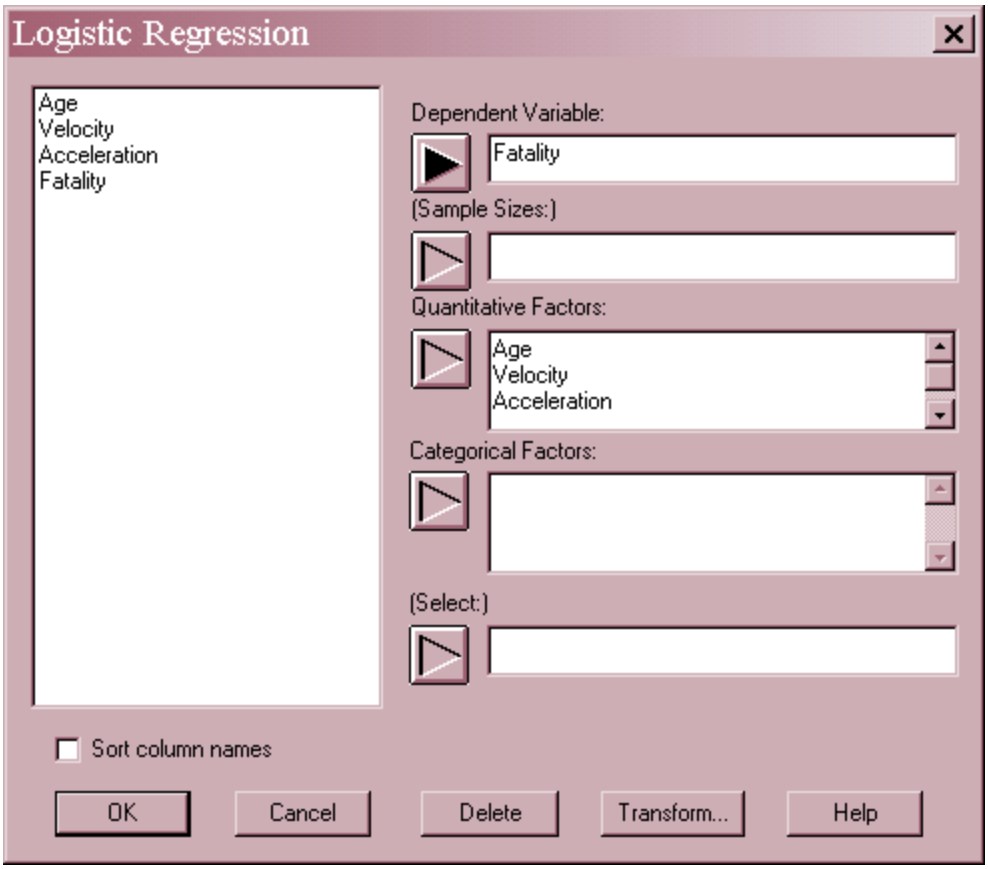
- **(Sample Sizes)**: If *Y* contains a set of proportions, enter a column with the sample sizes corresponding to each proportion. If Y contains a set of 0's and 1's, leave this field blank.

- **Quantitative Factors**: numeric columns containing the values of any quantitative factors to be included in the model.

- **Categorical Factors**: numeric or non-numeric columns containing the levels of any categorical factors to be included in the model.

- **Select**: subset selection.

For the *collisions.sf6* file, where the data is binary, the data input dialog box is shown below:



Statistical Model

The logistic model relates the probability of occurrence *P* of the outcome counted by *Y* to the predictor variables *X*. The model takes the form

$$P(X) = \frac{1}{1+\exp[-(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_k X_k)]} \qquad (1)$$

Alternatively, the model can be written in the form

$$\log\left(\frac{P(X)}{1-P(X)}\right) = \exp(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_k X_k) \qquad (2)$$

where the left hand side of the above equation is referred to as the *logit transformation*.

Analysis Summary

The *Analysis Summary* displays a table showing the estimated model and test of significance for the model coefficients. The output depends on the method used to estimate the model.

Maximum Likelihood Estimation

Maximum likelihood estimation may be used whether Y is binary or contains proportions. Typical output when maximum likelihood is used is shown below:

Logistic Regression - Failures/Specimens

Dependent variable: Failures/Specimens
Sample sizes: Specimens
Factors:
  Load

Estimated Regression Model (Maximum Likelihood)

		Standard	*Estimated*
Parameter	*Estimate*	*Error*	*Odds Ratio*
CONSTANT	-2.9949	0.145939	
Load	0.0307699	0.00209432	1.03125

Analysis of Deviance

Source	*Deviance*	*Df*	*P-Value*
Model	283.056	1	0.0000
Residual	36.2181	3	0.0000
Total (corr.)	319.274	4	

Percentage of deviance explained by model = 88.6561
Adjusted percentage = 87.4033

Likelihood Ratio Tests

Factor	*Chi-Squared*	*Df*	*P-Value*
Load	283.056	1	0.0000

Residual Analysis

	Estimation	*Validation*
n	5	
MSE	0.159284	
MAE	0.0299959	
MAPE	23.9252	
ME	-0.000979783	
MPE	-10.6729	

The output includes:

- **Data Summary:** a summary of the input data.

- **Estimated Regression Model:** estimates of the coefficients in the regression model, with standard errors and estimated odds ratios. The odds ratios are calculated from the model coefficients $\hat{\beta}_j$ by

$$\text{odds ratio} = \exp\left(\hat{\beta}_j\right) \tag{3}$$

The odds ratio represents the percentage increase in the odds of an outcome for each unit increase in X.

- **Analysis of Deviance:** decomposition of the deviance of the data into an explained (*Model*) component and an unexplained (*Residual*) component. *Deviance* compares the likelihood function for a model to the largest value that the likelihood function could achieve, in a manner such that a perfect model would have a deviance equal to 0. There are 3 lines in the table:

1. **Total (corr.)** – the deviance of a model containing only a constant term, $\lambda(\beta_0)$.

2. **Residual** – the deviance remaining after the model has been fit.

3. **Model** – the reduction in the deviance due to the predictor variables, $\lambda(\beta_1,\beta_2,\ldots,\beta_k|\beta_0)$, equal to the difference between the other two components.

The P-Value for the *Model* tests whether the addition of the predictor variables significantly reduces the deviance compared to a model containing only a constant term. A small P-Value (less than 0.05 if operating at the 5% significance level) indicates that the model has significantly reduced the deviance and is thus useful for predicting the probability of the studied outcome. The P-Value for the *Residual* term tests whether there is significant lack-of-fit, i.e., whether a better model may be possible. A small P-value indicates that significant deviance remains in the residuals, so that a better model might be possible.

- **Percentage of Deviance** – the percentage of deviance explained by the model, calculated by

$$R^2 = \frac{\lambda(\beta_1,\beta_2,...,\beta_k \mid \beta_0)}{\lambda(\beta_0)} \tag{4}$$

It is similar to an R-squared statistic in multiple regression, in that it can range from 0% to 100%. An adjusted deviance is also computed from

$$R^2_{adj} = \frac{\lambda(\beta_1,\beta_2,...,\beta_k \mid \beta_0) - 2p}{\lambda(\beta_0)} \tag{5}$$

where *p* equals the number of coefficients in the fitted model, including the constant term. It is similar to the adjusted R-squared statistic in that it compensates for the number of variables in the model.

- **Likelihood Ratio Tests** – a test of significance for each effect in the fitted model. These tests compare the likelihood function of the full model to that of the model in which only the

indicated effect has been dropped. Small P-values indicate that the model has been improved significantly by the corresponding effect.

- **Residual Analysis** – if a subset of the rows in the datasheet have been excluded from the analysis using the *Select* field on the data input dialog box, the fitted model is used to make predictions of the *Y* values for those rows. This table shows statistics on the prediction errors, defined by

$$e_i = y_i - \hat{P}(X_i) \tag{6}$$

Included are the mean squared error (MSE), the mean absolute error (MAE), the mean absolute percentage error (MAPE), the mean error (ME), and the mean percentage error (MPE). These validation statistics can be compared to the statistics for the fitted model to determine how well that model predicts observations outside of the data used to fit it.

The fitted model for the sample data is

$$P(failure) = \frac{1}{1+\exp[-(-2.9949+0.0307699 Load)]} \tag{7}$$

The regression explains about 88.7% of the deviance of a model without *Load*. The P-value for *Load* is very small, indicating that it is a statistically significant predictor for the proportion of *Failures*. The odds ratio is approximately 1.03, indicating a 3% increase in the odds of a failure for each unit increase in *Load*.

Note that the P-value for the *Residuals* is also significant, indicating that significant lack-of-fit remains unexplained. This can be rectified by returning to the data input dialog box and entering *LOG(Load)* as the predictor variable rather than *Load*. The result is a loglogistic model, as shown below:

Logistic Regression - Failures/Specimens
Dependent variable: Failures/Specimens
Sample sizes: Specimens
Factors:
LOG(Load)

Estimated Regression Model (Maximum Likelihood)

		Standard	*Estimated*
Parameter	*Estimate*	*Error*	*Odds Ratio*
CONSTANT	-5.5784	0.368202	
LOG(Load)	1.13997	0.0892554	3.12667

Analysis of Deviance

Source	*Deviance*	*Df*	*P-Value*
Model	313.886	1	0.0000
Residual	5.38828	3	0.1455
Total (corr.)	319.274	4	

Percentage of deviance explained by model = 98.3123
Adjusted percentage = 97.0595

Likelihood Ratio Tests

Factor	*Chi-Squared*	*Df*	*P-Value*
LOG(Load)	313.886	1	0.0000

Notice the increase in the percentage of deviance explained to over 98%. In addition, the P-Value for the *Residuals* no longer shows significant lack of fit.

Weighted Least Squares Regression

When the input data *Y* consists of a set of proportions, the model may be estimated using weighted least squares rather than maximum likelihood. The output then takes the following form:

Logistic Regression - Failures/Specimens

Dependent variable: Failures/Specimens
Sample sizes: Specimens
Factors:
  Load

Estimated Regression Model (Weighted Least Squares)

		Standard	*T*		*Estimated*
Parameter	*Estimate*	*Error*	*Statistic*	*P-Value*	*Odds Ratio*
CONSTANT	-2.72665	0.525557	-5.18811	0.0139	
Load	0.0272839	0.00753311	3.62186	0.0362	1.02766

Analysis of Variance

Source	*Sum of Squares*	*Df*	*Mean Square*	*F-Ratio*	*P-Value*
Model	132.876	1	132.876	13.12	0.0362
Residual	30.3881	3	10.1294		
Total (Corr.)	163.264	4			

R-Squared = 81.3871 percent
R-Squared (adjusted for d.f.) = 75.1828 percent
Standard Error of Est. = 3.18267
Mean absolute error = 0.168476
Durbin-Watson statistic = 2.15796
Lag 1 residual autocorrelation = -0.390383

Type III Sums of Squares

Source	*Sum of Squares*	*Df*	*Mean Square*	*F-Ratio*	*P-Value*
Load	132.876	1	132.876	13.12	0.0362
Residual	30.3881	3	10.1294		

Residual Analysis

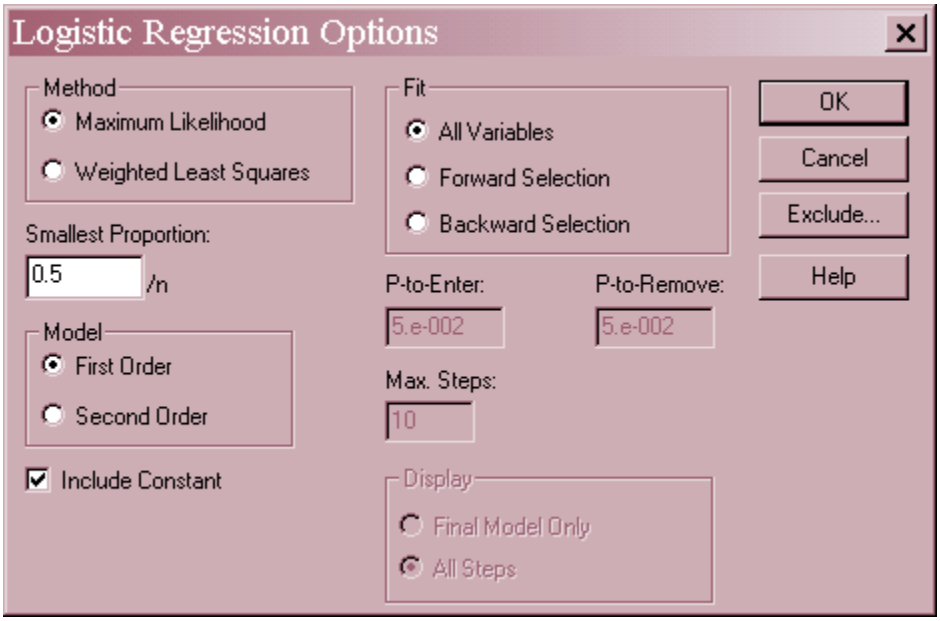
	Estimation	*Validation*
n	5	
MSE	10.1294	
MAE	0.168476	
MAPE	254.223	
ME	3.19675E-17	
MPE	-171.933	

The table differs from the output of the MLE option in several ways:

1. Each coefficient is shown together with a t-statistic and associated P-value, which tests whether a specified coefficient may be equal to 0.
2. The analysis of deviance is replaced by a standard *analysis of variance*. The *F-Ratio* tests the statistical significance of the model as a whole.
3. The percentage of deviance is replaced by a standard *R-Squared* statistic.
4. The likelihood ratio tests of the effects are replaced by *F tests* based on Type III sums of squares. The same interpretation of the P-values applies, however, with small P-values corresponding to significant effects.

For more explanation of the regression statistics, see the *General Linear Models* documentation.

Analysis Options



- **Method**: method used to estimate the model coefficients. For binary *Y*, *Maximum Likelihood* is the only choice.

- **Smallest Proportion**: For data *Y* consisting of proportions, the smallest allowable proportion P_{min}. All observations less than P_{min} are set equal to P_{min}, while all observations greater than 1- P_{min} are set equal to 1- P_{min}.

- **Model**: order of the model to be fit. First order models include only main effects. Second order models include quadratic effects for quantitative factors and two-factor interactions amongst all variables.

- **Include Constant**: If this option is not checked, the constant term β_0 will be omitted from the model.

- **Fit**: specifies whether all independent variables specified on the data input dialog box should be included in the final model, or whether a stepwise selection of variables should be applied. Stepwise selection attempts to find a parsimonious model that contains only statistically significant variables. A *Forward Stepwise* fit begins with no variables in the model. A *Backward Stepwise* fit begins with all variables in the model.

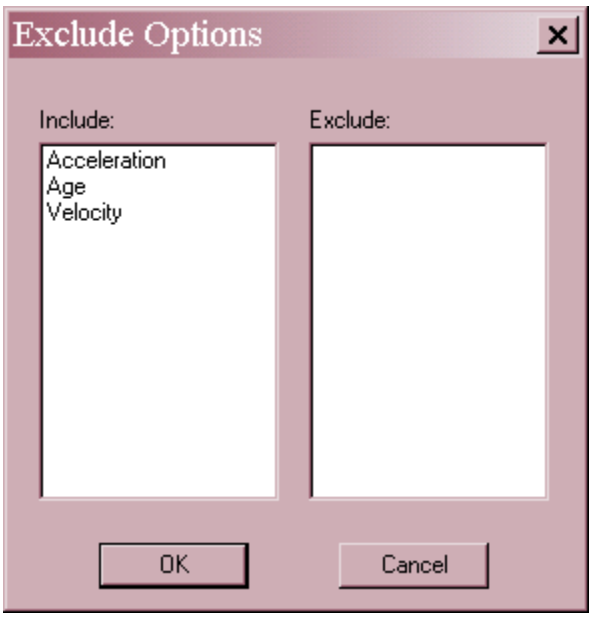
- **P-to-Enter** - In a stepwise fit, variables will be entered into the model at a given step if their P-values are less than or equal to the *P-to-Enter* value specified.

- **P-to-remove** - In a stepwise fit, variables will be removed from the model at a given step if their P-values are greater than the *P-to-Remove* value specified.

- **Max Steps:** maximum number of steps permitted when doing a stepwise fit.

- **Display**: whether to display the results at each step when doing a stepwise fit.

- **Exclude**: Press this button to exclude effects from the model. A dialog box will be displayed:



Double click on an effect to move it from the *Include* field to the *Exclude* field or back again.

Example: Stepwise Fit Using Binary Data

The data on automobile collisions contains 3 possible predictor variables: *Age*, *Velocity*, and *Acceleration*. To select a model containing only significant predictors, a stepwise fit could be used. Two algorithms are available:

- **Forward selection** – Begins with a model involving only a constant term and enters one variable at a time based on its statistical significance if added to the current model. At each step, the algorithm brings into the model the variable that will be the most statistically significant if entered. As long as the most significant variable has a P-value less than or equal to that specified on the *Analysis Summary* dialog box, it will be brought into the model. When no variable has a small enough -value, variable selection stops. In addition, variables brought into the model early in the procedure may be removed later if their P-value falls below the *P-to-remove* criterion.

- **Backward selection** – Begins with a model involving all the variables specified on the data input dialog box and removes one variable at a time based on its statistical significance in the current model. At each step, the algorithm removes from the model the variable that is the least statistically significant. If the least significant variable has a P-value greater than that specified on the *Analysis Summary* dialog box, it will be removed from the model. When all remaining variables have small P-values, the procedure stops. In addition, variables removed from the model early in the procedure may be re-entered later if their P-values reach the *P-to-enter* criterion.

The following output shows the result of a backwards stepwise fit:

Logistic Regression - Fatality

Dependent variable: Fatality
Factors:
  Age
  Velocity
  Acceleration

Estimated Regression Model (Maximum Likelihood)

		Standard	*Estimated*
Parameter	*Estimate*	*Error*	*Odds Ratio*
CONSTANT	-16.9845	5.14861	
Age	0.162501	0.041448	1.17645
Velocity	0.233906	0.0862681	1.26353

Analysis of Deviance

Source	*Deviance*	*Df*	*P-Value*
Model	33.3408	2	0.0000
Residual	45.3315	55	0.8206
Total (corr.)	78.6723	57	

Percentage of deviance explained by model = 42.3793
Adjusted percentage = 34.7527

Likelihood Ratio Tests

Factor	*Chi-Squared*	*Df*	*P-Value*
Age	29.9333	1	0.0000
Velocity	10.0497	1	0.0015

Residual Analysis

	Estimation	*Validation*
n	58	
MSE	0.0221508	
MAE	0.340955	
MAPE		
ME	0.00127246	
MPE		

Stepwise factor selection
Method: backward selection
P-to-enter: 0.05
P-to-remove: 0.05

Step 0:
  3 factors in the model. 54 d.f. for error.
  Percentage of deviance explained = 44.10% Adjusted percentage = 33.93%

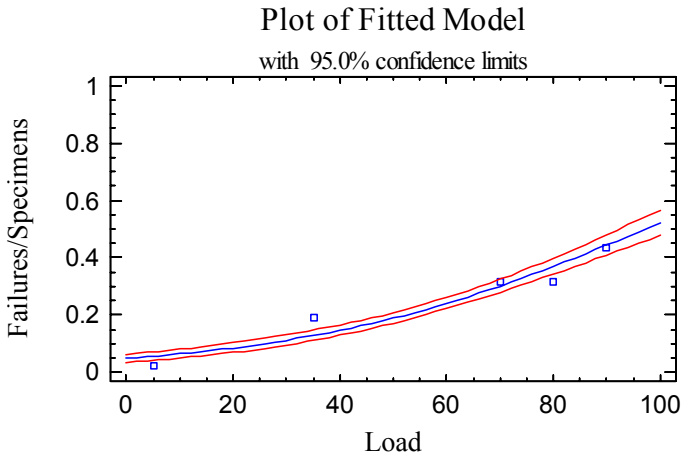
Step 1:
  Removing factor Acceleration with P-to-remove = 0.244299
  2 factors in the model. 55 d.f. for error.
  Percentage of deviance explained = 42.38% Adjusted percentage = 34.75%

Final model selected.

The algorithm begins with a model containing all three predictors. It then removes *Acceleration*, since its P-value is large. The final model involves only *Age* and *Velocity*, each of which has a P-value at or below 0.05.

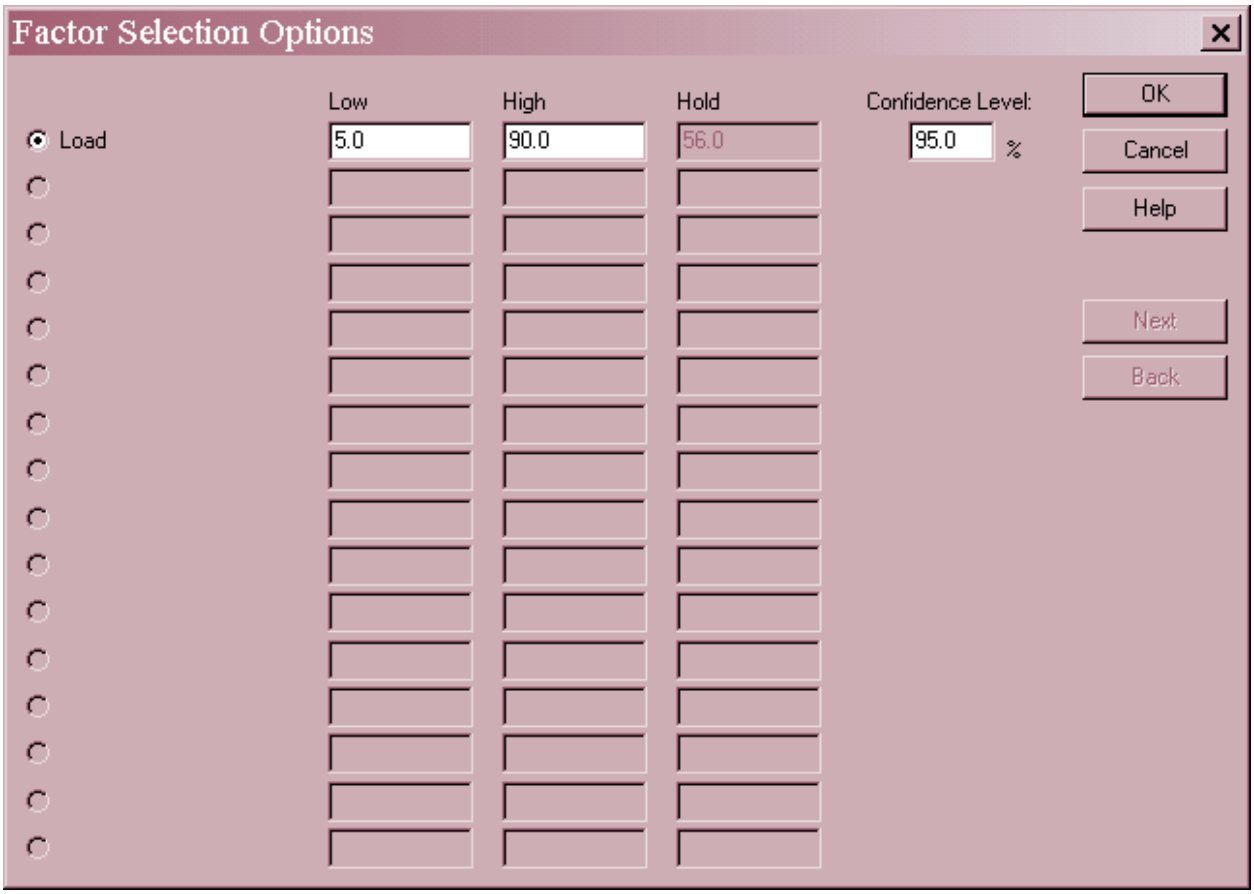
Plot of Fitted Model

The *Plot of Fitted Model* displays the estimated probability of an outcome $\hat{P}(X)$ versus any single predictor variable, with the other variables held constant.



Confidence limits for *P(X)* are included on the plot.

Pane Options



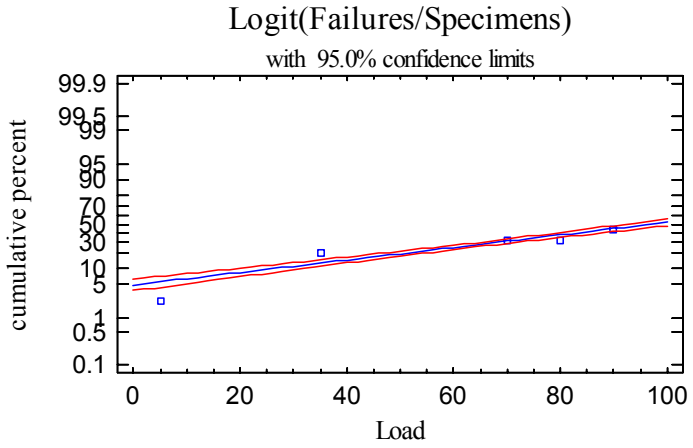
- **Factor:** select the factor to plot on the horizontal axis.

- **Low and High:** specify the range of values for the selected factor.
- **Hold:** select values to hold the unselected factors at.
- **Confidence Level:** percentage used for the confidence limits. Set to 0 to suppress the limits.
- **Next** and **Back:** used to display other factors when more than 16 are present.

The estimated probability of a failure increases from approximately 5% at low loads to nearly 50% when *Load* = 100.

Logit Plot

The *Logit Plot* is similar to the *Plot of Fitted Model*, except that the vertical axis is scaled so that the fitted model will be a straight line.

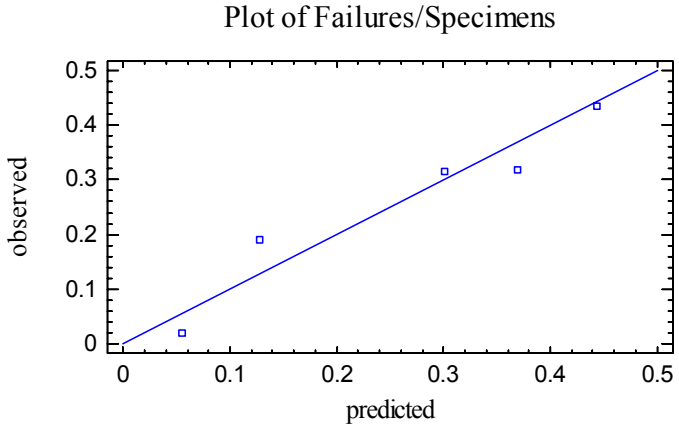


Pane Options

The options are the same as those for the *Plot of the Fitted Model*.

Observed Versus Predicted

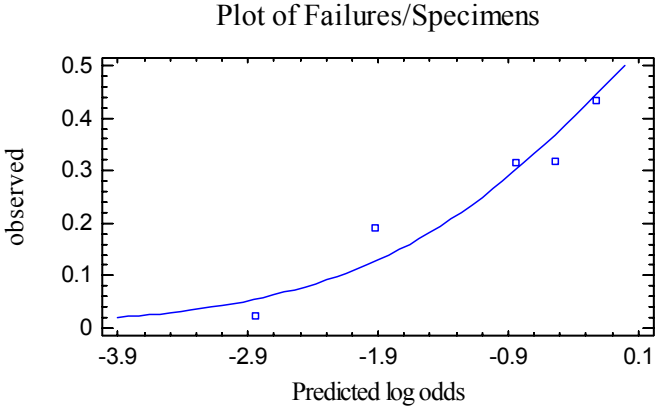
The *Observed versus Predicted* plot shows the observed values of Y on the vertical axis and the predicted values $\hat{P}(X)$ on the horizontal axis.



If the model fits well, the points should be randomly scattered around the diagonal line.

Observed versus Log Odds

The *Observed versus Log Odds* pane plots the observed values of Y versus the predicted log odds, given by $\log\left(\frac{\hat{P}(X)}{1-\hat{P}(X)}\right)$.



The log odds equals the logistic transformation, which is an exponential function of the predictor variables.

Inverse Predictions

The *Inverse Predictions* table displays estimated values of a selected variable X at which the probability $\hat{P}(X)$ equals set percentages. All other variables in the model are fixed at user-specified values.

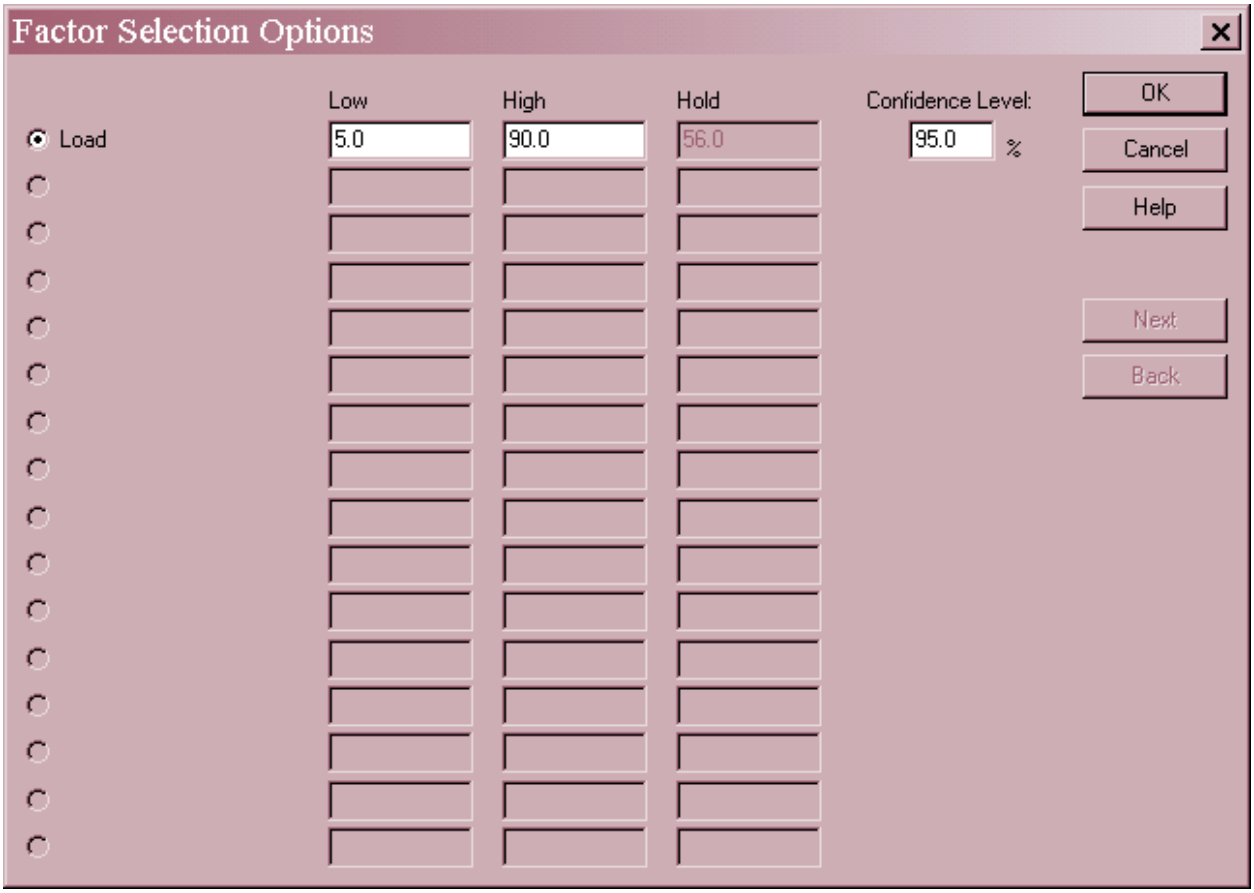
Table of Inverse Predictions for Load

		Lower 95.0%	*Upper 95.0%*
Percent	*Load*	*Conf. Limit*	*Conf. Limit*
0.1	-127.132	-156.921	-104.304
0.5	-74.6964	-96.4948	-57.9582
1.0	-52.0059	-70.3688	-37.8809
2.0	-29.1492	-44.0783	-17.6297
3.0	-15.6385	-28.5593	-5.6379
4.0	-5.95227	-17.4486	2.97493
5.0	1.64004	-8.75274	9.73882
6.0	7.90927	-1.58389	15.3356
7.0	13.2666	4.53136	20.1292
8.0	17.9577	9.87545	24.3372
9.0	22.1407	14.6304	28.0999
10.0	25.924	18.9204	31.5135
15.0	40.9589	35.8156	45.2329
20.0	52.2786	48.2389	55.8593
25.0	61.6281	58.1385	64.9974
30.0	69.7956	66.4205	73.3465
35.0	77.2138	73.6499	81.2223
40.0	84.1548	80.2174	88.7884
45.0	90.8105	86.3904	96.1679
50.0	97.3322	92.3594	103.479
55.0	103.854	98.275	110.843
60.0	110.509	104.274	118.396
65.0	117.45	110.502	126.302
70.0	124.869	117.134	134.775
75.0	133.036	124.417	144.123
80.0	142.386	132.735	154.843
85.0	153.705	142.788	167.839
90.0	168.74	156.119	185.123
91.0	172.524	159.471	189.474
92.0	176.707	163.176	194.287
93.0	181.398	167.33	199.686
94.0	186.755	172.072	205.852
95.0	193.024	177.62	213.07
96.0	200.617	184.337	221.813
97.0	210.303	192.903	232.97
98.0	223.813	204.848	248.536
99.0	246.67	225.048	274.878
99.5	269.361	245.093	301.036
99.9	321.797	291.4	361.501

Fiducial confidence levels for the values of X are also included.

For example, the probability of failure for the fabric example is estimated to reach p = 50% at *Load* = 97.33. The 95% confidence limits range from 92.36 to 103.48.

Pane Options



- **Factor:** select the factor for which to calculate the inverse predictions.
- **Low and High:** ignored.
- **Hold:** select values to hold the unselected factors at.
- **Confidence Level:** percentage used for the confidence limits.
- **Next** and **Back:** used to display other factors when more than 16 are present.

Goodness-of-Fit

The Goodness-of-Fit pane performs a chi-squared test to determine whether the fitted model adequately describes the observed data. It does so by dividing the fitted logit values into classes (groups) and performing a chi-squared test to compare the observed versus fitted values in each interval.

Chi-Squared Goodness of Fit Test

	Logit		*TRUE*	*TRUE*	*FALSE*	*FALSE*
Class	*Interval*	*n*	*Observed*	*Expected*	*Observed*	*Expected*
1	less than -2.84105	600	13.0	33.0874	587.0	566.913
2	-2.84105 to -1.91796	500	95.0	64.0449	405.0	435.955
3	-1.91796 to -0.841008	600	189.0	180.793	411.0	419.207
4	-0.841008 or greater	600	225.0	244.074	375.0	355.926
Total		2300	522.0		1778.0	

Chi-squared = 33.1125 with 2 d.f. P-value = 6.45217E-8

In creating the classes, the program attempts to create groups of approximately equal size.

The table shows the following information for each class:

1. *Logit interval* - the range of logit values $\log\left(\frac{\hat{P}(X)}{1-\hat{P}(X)}\right)$ corresponding to that class.

2. *n* - the total number of samples with fitted values within that class.

3. *TRUE Observed* – of the number of samples in that interval, how many were observed to be TRUE (1).

4. *TRUE Predicted* – of the number of samples in that interval, how many were predicted by the fitted model to be TRUE.

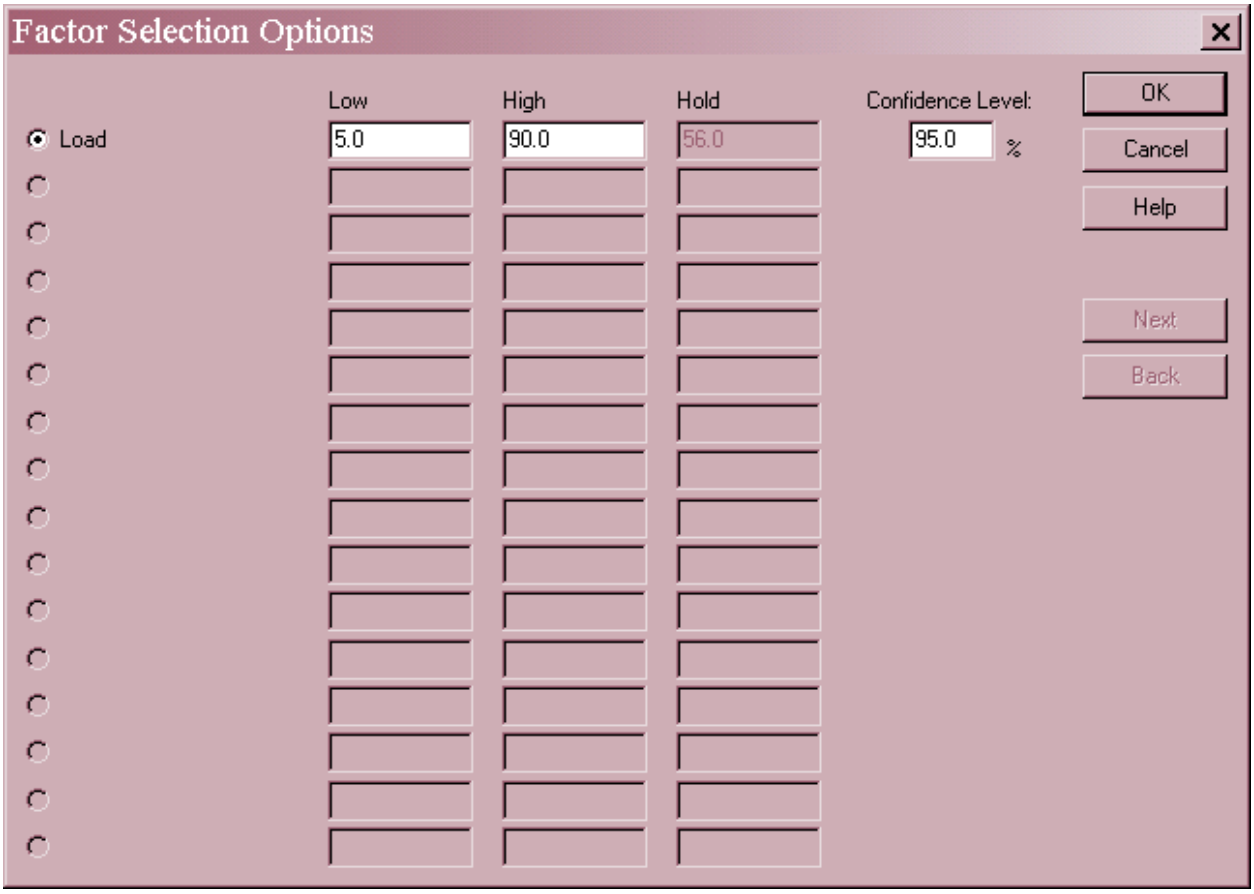
5. *FALSE Observed* – of the number of samples in that interval, how many were observed to be FALSE (0).

6. *FALSE Predicted* – of the number of samples in that interval, how many were predicted by the fitted model to be FALSE.

For example, a total of 600 samples of fabric have predicted logit values less than –2.84105 (corresponding to row #1 of the data file). 13 failures were observed, as observed to a predicted value of approximately 33.

To compared the observed counts to the expected counts, a chi-squared goodness-of-fit test is performed. A small P-Value (less than 0.05 if operating at the 5% significance level) leads to the conclusion that the fitted model does not adequately match the data. In the example, the P-Value is very small, indicating a poor fit of the logistic model.

For comparison purposes, note the test for the loglogistic model with *X = LOG(Load):*



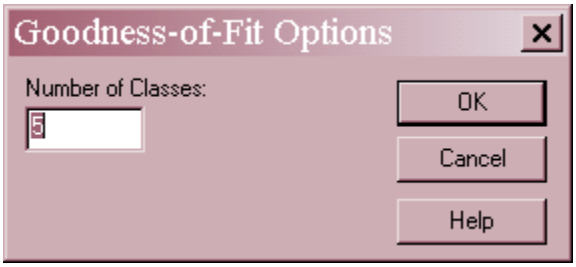
Chi-Squared Goodness of Fit Test

	Logit		*TRUE*	*TRUE*	*FALSE*	*FALSE*
Class	*Interval*	*n*	*Observed*	*Expected*	*Observed*	*Expected*
1	less than -3.74369	600	13.0	13.8717	587.0	586.128
2	-3.74369 to -1.52541	500	95.0	89.333	405.0	410.667
3	-1.52541 to -0.735245	600	189.0	194.427	411.0	405.573
4	-0.735245 or greater	600	225.0	224.368	375.0	375.632
Total		2300	522.0		1778.0	

Chi-squared = 0.720702 with 2 d.f. P-value = 0.697432

In that case, the P-Value does not show significant lack-of-fit.

Pane Options



Number of Classes: maximum number of classes into which to group the data.

Predictions

The fitted logistic model may be used to predict the outcome of new samples whose predictor variables are given. For example, suppose a new sample is collected at a *Load* equal to 50. If one wanted to predict whether or not the item would fail, the fitted model could be evaluated for the new sample and a failure predicted if

$$\hat{P}(X_{new}) > c$$

for some cutoff value *c*. The value of c would affect the probability of obtaining a false positive or false negative result.

The top section of the *Predictions* table shows the percentage of items correctly classified as a function of *c*.

Prediction Performance - Percent Correct

Cutoff	*TRUE*	*FALSE*	*Total*
0.0	100.00	0.00	22.70
0.05	100.00	0.00	22.70
0.1	97.51	33.01	47.65
0.15	79.31	55.79	61.13
0.2	79.31	55.79	61.13
0.25	79.31	55.79	61.13
0.3	79.31	55.79	61.13
0.35	43.10	78.91	70.78
0.4	24.90	90.44	75.57
0.45	0.00	100.00	77.30
0.5	0.00	100.00	77.30
0.55	0.00	100.00	77.30
0.6	0.00	100.00	77.30
0.65	0.00	100.00	77.30
0.7	0.00	100.00	77.30
0.75	0.00	100.00	77.30
0.8	0.00	100.00	77.30
0.85	0.00	100.00	77.30
0.9	0.00	100.00	77.30
0.95	0.00	100.00	77.30
1.0	0.00	100.00	77.30

.

Included in the table are:

- *Cutoff* – the value of c.

- *TRUE* - using the indicated value of *c*, the percent of observed failures that would have been correctly predicted.

- *FALSE* - using the indicated value of *c*, the percent of observed non-failures that would have been correctly predicted.

- *Total* - using the indicated value of *c*, the percent of all samples that would have been correctly predicted.

For example, using a cutoff of $c = 0.45$ would have predicted correctly the largest percentage of total samples (77.3%). Unfortunately, it would have predicted FALSE for all samples (meaning a non-failure), which classifies all non-failures correctly but misses all failures! In order to predict the failures with a high probability would require a value of $c = 0.1$, which also results in misclassifying 33% of the non-failures. If the model is to be used to screen samples, the best value of *c* would depend on the relative cost of missed failures versus the cost of misclassified non-failures.

The second table in the output pane evaluates the fitted model for selected rows in the datasheet. Predictions can be made for all rows that have complete information on the X variables or only those rows that have missing values for Y. The latter option is useful for making predictions at values of X not used to fit the model.

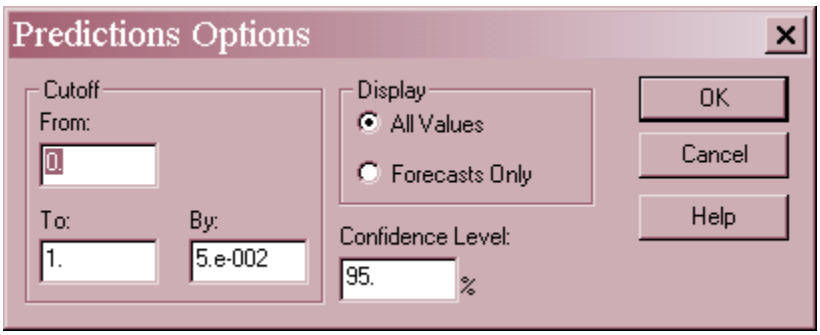
For example, a sixth row could be added to the datasheet with *Load* = 50, leaving the *Failures* column blank.

Predictions for Failures/Specimens

	Observed	*Fitted*	*Lower 95.0%*	*Upper 95.0%*
Row	*Value*	*Value*	*Conf. Limit*	*Conf. Limit*
6		0.189018	0.171239	0.208178

The table predicts a mean failure rate of approximately 18.9% at that load, with a 95% confidence interval for the mean rate ranging between 17.1% and 20.8%.

Pane Options



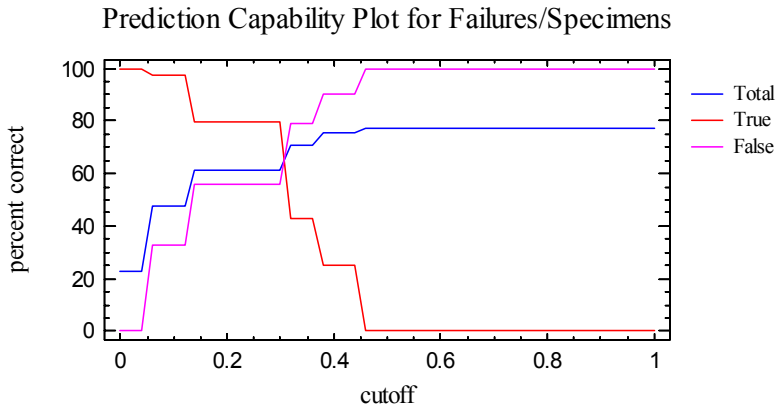
- **Cutoff***:* the range of values and increment for *c* in the table of prediction percentages.

- **Display:** whether to display predictions for *All Values* (rows) in the datasheet or *Forecasts Only* (rows with a missing value for *Y*).

- **Confidence Level:** percentage of confidence for the confidence limits.

Prediction Capability

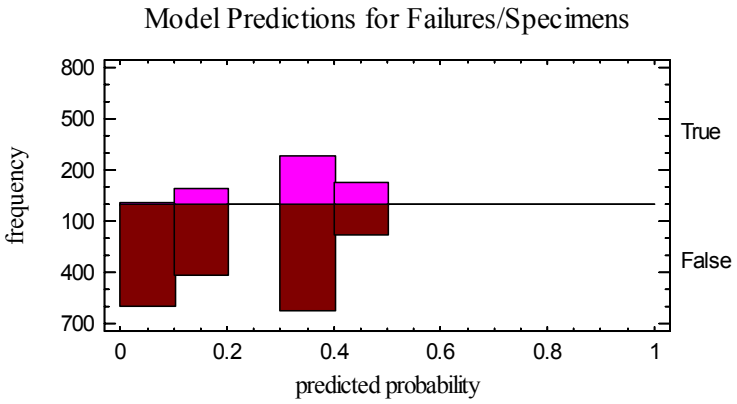
The *Prediction Capability* plot displays the same information as in the *Predictions* table.



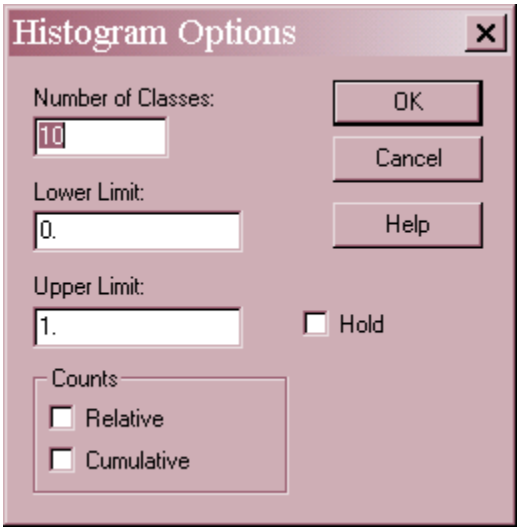
It plots the correct prediction percentages as a function of the cutoff value *c*.

Prediction Histogram

The *Prediction Histogram* illustrates the predicted number of total samples that will fail (TRUE) and not fail (FALSE), versus the predicted probability $\hat{P}(X)$.



Pane Options



- **Number of Classes**: the total number of classes into which the horizontal axis will be divided.
- **Lower and Upper Limit**: the limits of the horizontal axis.
- **Hold:** check to prevent the histogram scaling from changing if the data changes.
- **Counts**: Select *Relative* to plot proportions on the vertical axis rather than counts. Select *Cumulative* to plot cumulative counts from left to right.

Confidence Intervals

The *Confidence Intervals* pane shows the potential estimation error associated with each coefficient in the model, as well as for the odds ratios.

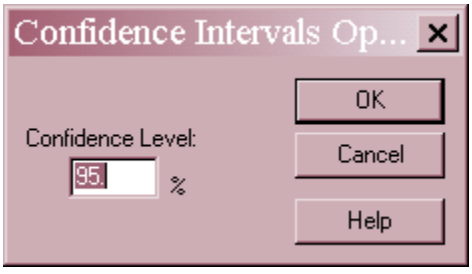
95.0% confidence intervals for coefficient estimates

		Standard		
Parameter	*Estimate*	*Error*	*Lower Limit*	*Upper Limit*
CONSTANT	-2.9949	0.145939	-3.45934	-2.53046
Load	0.0307699	0.00209432	0.0241049	0.037435

95.0% confidence intervals for odds ratios

Parameter	*Estimate*	*Lower Limit*	*Upper Limit*
Load	1.03125	1.0244	1.03814

Pane Options



- **Confidence Level:** percentage level for the confidence intervals.

Correlation Matrix

The *Correlation Matrix* displays estimates of the correlation between the estimated coefficients.

Correlation matrix for coefficient estimates

	CONSTANT	Load
CONSTANT	1.0000	-0.9320
Load	-0.9320	1.0000

This table can be helpful in determining how well the effects of different independent variables have been separated from each other.

Unusual Residuals

Once the model has been fit, it is useful to study the residuals to determine whether any outliers exist that should be removed from the data. The *Unusual Residuals* pane lists all observations that have unusually large residuals

Unusual Residuals for Failures/Specimens

		Predicted		*Pearson*	*Deviance*
Row	*Y*	*Y*	*Residual*	*Residual*	*Residual*
1	0.0216667	0.0551456	-0.033479	-3.59	-4.07
2	0.19	0.12809	0.0619103	4.14	3.91

The table displays:

- **Row** – the row number in the datasheet.
- **Y** – the observed value of Y.
- **Predicted Y** – the fitted value $\hat{P}(X)$.
- **Residual** – the difference between the observed and predicted values defined by

$$e_i = Y - \hat{P}(X) \tag{8}$$

- **Pearson Residual** – a standardized residual in which each residual is divided by an estimate of its standard error:

$$r_i = \frac{e_i}{\sqrt{\frac{\hat{P}(X_i)\left(1-\hat{P}(X_i)\right)}{n_i}}} \tag{9}$$

- **Deviance Residual –** a residual that measures each observation's contribution to the residual deviance:

$$d_i = sign(e_i)\sqrt{2\left\{n_i y_i \ln\left(\frac{y_i}{\hat{P}(X_i)}\right) + n_i(1-y_i)\ln\left(\frac{1-y_i}{1-\hat{P}(X_i)}\right)\right\}} \qquad (10)$$

The sum of squared deviance residuals equals the deviance on the *Residuals* line of the analysis of deviance table.

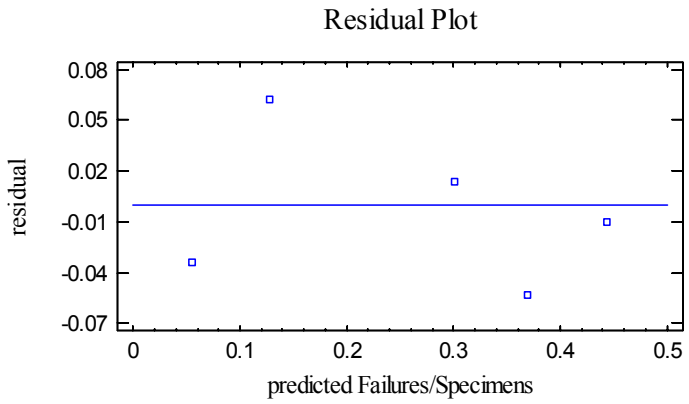
The table includes all rows for which the absolute value of the Pearson residual is greater than 2.0. The current example shows 2 very large residuals.

Residual Plots

As with all statistical models, it is good practice to examine the residuals. The *Logistic Regression* procedure various type of residual plots, depending on *Pane Options*.

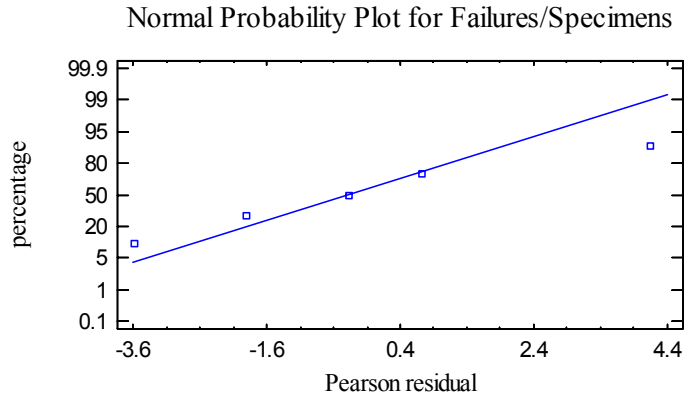
Scatterplot versus Predicted Value

This plot is helpful in visualizing whether the variability of he residuals is constant or depends on the predicted value.



Normal Probability Plot

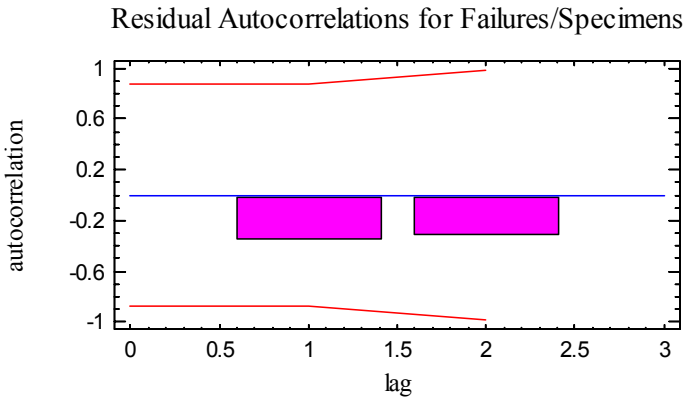
This plot can be used to determine whether or not the deviations around the line follow a normal distribution.



If the deviations follow a normal distribution, they should fall approximately along a straight line.

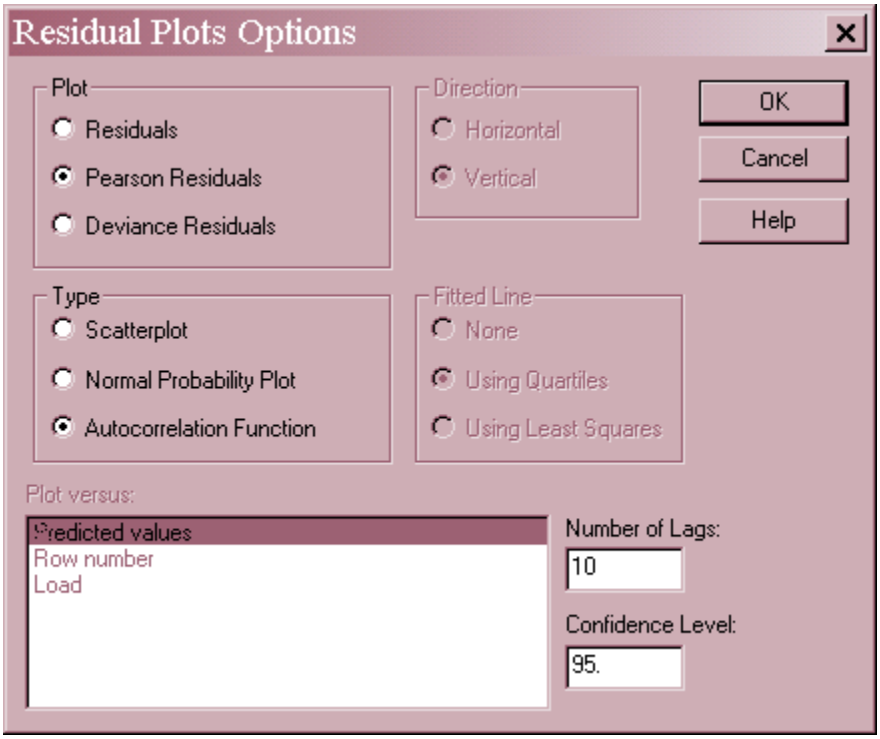
Residual Autocorrelations

This plot calculates the autocorrelation between residuals as a function of the number of rows between them in the datasheet.



It is only relevant if the data have been collected sequentially. Any bars extending beyond the probability limits would indicate significant dependence between residuals separated by the indicated "lag".

Pane Options



- **Plot:** the type of residuals to plot:

    1. *Residuals* – the observed values minus the fitted values.

2. *Studentized residuals* – the residuals divided by their estimated standard errors.
3. *Deviance Residuals* – residuals scaled so that their sum of squares equals the residual deviance.

- **Type:** the type of plot to be created. A *Scatterplot* is used to test for curvature. A *Normal Probability Plot* is used to determine whether the model residuals come from a normal distribution. An *Autocorrelation Function* is used to test for dependence between consecutive residuals.

- **Plot Versus**: for a *Scatterplot*, the quantity to plot on the horizontal axis.

- **Number of Lags**: for an *Autocorrelation Function*, the maximum number of lags. For small data sets, the number of lags plotted may be less than this value.

- **Confidence Level:** for an *Autocorrelation Function*, the level used to create the probability limits.

Save Results

The following results may be saved to the datasheet:

1. *Predicted Values* – the fitted values $\hat{P}(X_i)$ corresponding to each row of the datasheet.
2. *Lower Limits* – the lower confidence limits for $\hat{P}(X_i)$.
3. *Upper Limits* – the upper confidence limits for $\hat{P}(X_i)$.
4. *Residuals* – the ordinary residuals.
5. *Pearson Residuals* – the standardized Pearson residuals.
6. *Deviance Residuals* – the deviance residuals.
7. *Leverages* – if the model was fit using weighted least squares, the leverages for each row.
8. *Percentages* – the percentages at which inverse predictions were made.
9. *Inverse Predictions* – the inverse predictions.
10. *Lower Fiducial Limits* – the lower confidence limits for the inverse predictions.
11. *Upper Fiducial Limits* - the upper confidence limits for the inverse predictions.

Calculations

Likelihood Function

For *Y* consisting of proportions: $L = \prod_{i=1}^{s} [P(X_i)]^{r_i} [1 - P(X_i)]^{n_i - r_i}$ where $r_i = n_i p_i$ (11)

For binary Y: $L = \frac{\prod_{i=1}^{n} P(X_i)^{Y_i}}{\prod_{i=1}^{n} [1 + P(X_i)]}$ (12)

Weights for Weighted Least Squares

$$w_i = \frac{1}{y_i(1 - y_i)} \quad (13)$$

Deviance

For *Y* consisting of proportions: $\lambda(\beta) = -2 \ln \left[\frac{L(\hat{\beta})}{\prod_{i=1}^{s} \left(\frac{r_i}{n_i} \right)^{r_i} \left(\frac{n_i - r_i}{n_i} \right)^{n_i - r_i}} \right]$ (14)

For binary Y: $\lambda(\beta) = -2 \ln \left[\frac{L(\hat{\beta})}{\prod_{i=1}^{s} (y_i)^{y_i} (1 - y_i)^{(1 - y_i)}} \right]$ (15)

