Probit Analysis

Summary

The **Probit Analysis** 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 the probit function. The *Logistic Regression* 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.

Since the *Probit Analysis* procedure closely parallels that of *Logistic Regression*, you should refer to that documentation for a detailed description of the various options. This document highlights the difference in the two models and covers a simple example.

Sample StatFolio: *probit.sgp*

Sample Data:

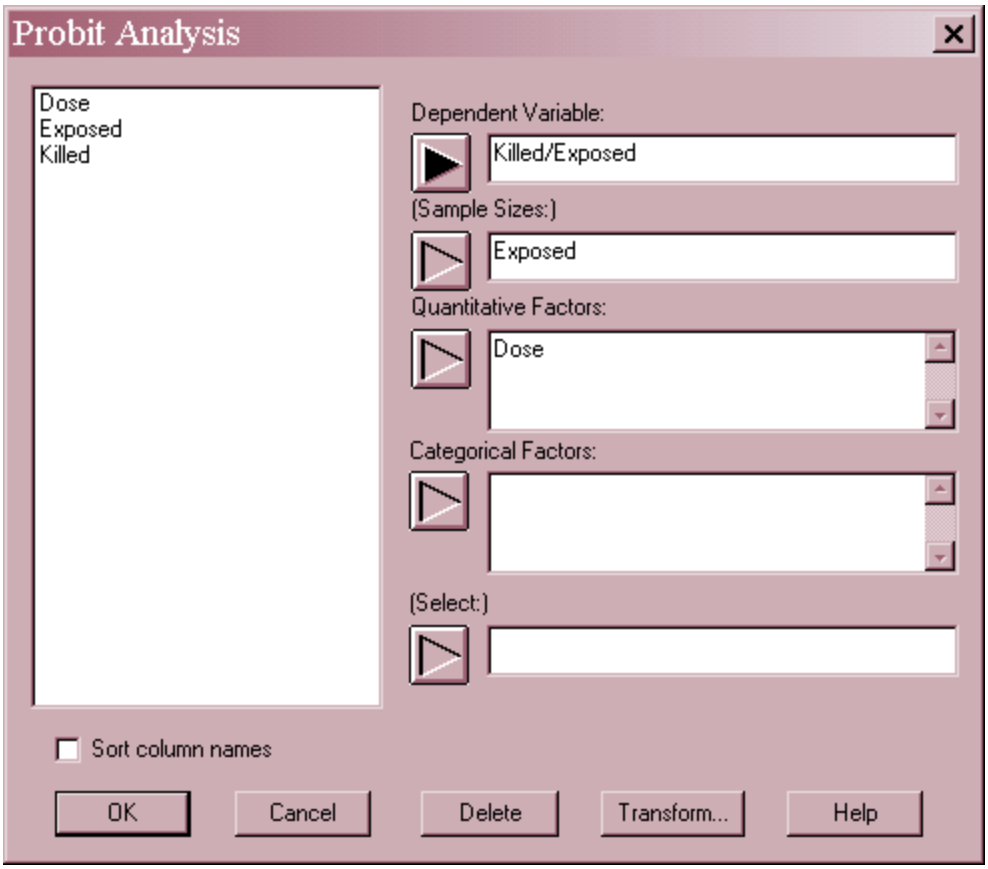
The file *beetles.sf3* contains a set of well-known data from Bliss (1935) showing the results of experiments in which beetles were exposed to different concentrations of carbon disulphide. The data file shows the dose, the number of beetles exposed, and the number of beetles killed.

Dose	*Exposed*	*Killed*
1.6907	59	6
1.7242	60	13
1.7552	62	18
1.7842	56	28
1.8113	63	52
1.8369	59	53
1.861	62	61
1.8839	60	60

For this data, the dependent variable *Y* is the proportion of exposed beetles at each dose that died, calculated by *Y = Exposed / Killed*. There is a single predictor variable *X = Dose*. There are a total of $n = 481$ subjects.

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.

Statistical Model

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

$$P(X) = \Phi(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_k X_k) \qquad (1)$$

where Φ(Z) is the standard normal cumulative distribution function.

Analysis Summary

The *Analysis Summary* displays a table showing the estimated model and tests of significance for the model coefficients. Typical output is shown below.

Probit Analysis - Killed/Exposed
Dependent variable: Killed/Exposed
Sample sizes: Exposed
Factors:
  Dose

Estimated Regression Model (Maximum Likelihood)

		Standard
Parameter	*Estimate*	*Error*
CONSTANT	-34.9349	2.65395
Dose	19.7277	1.49062

Analysis of Deviance

Source	*Deviance*	*Df*	*P-Value*
Model	274.083	1	0.0000
Residual	10.1198	6	0.1197
Total (corr.)	284.202	7	

Percentage of deviance explained by model = 96.4392
Adjusted percentage = 95.0318

Likelihood Ratio Tests

Factor	*Chi-Squared*	*Df*	*P-Value*
Dose	274.083	1	0.0000

Residual Analysis

	Estimation	*Validation*
n	8	
MSE	0.131797	
MAE	0.0562163	
MAPE	17.4188	
ME	-0.0211148	
MPE	-3.25668	

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.

- **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

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

    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

$$R^2_{adj} = \frac{\lambda(\beta_1, \beta_2, ..., \beta_k \mid \beta_0) - 2p}{\lambda(\beta_0)} \tag{3}$$

    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

$$e_i = y_i - \hat{P}(X_i) \tag{4}$$

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

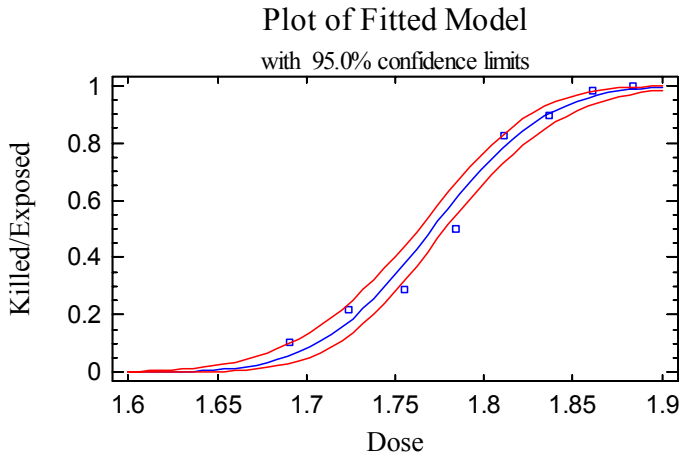
$$P(\mathit{failure}) = \Phi(-34.9349 + 19.7277\ \mathrm{Dose}) \qquad (5)$$

The regression explains about 96.4% of the deviance of a model without *Dose*. The P-value for *Dose* is very small, indicating that it is a statistically significant predictor for the proportion of beetles *Killed*.

Note that the P-value for the *Residuals* is not significant, indicating no significant lack-of-fit remaining unexplained.

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.

Probit Plot

The P*robit 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.