PROBIT[1]

The Probit procedure is used to estimate the effects of one or more independent variables on a dichotomous dependent variable. The program is designed for dose-response analyses and related models, but Probit can also estimate logistic regression models.

Notation

The following notation is used throughout this chapter unless otherwise stated:

m	Number of covariate patterns
n_i	Number of subjects for *i*th covariate pattern
r_i	Number of responses for *i*th covariate pattern
p	Number of independent variables
q	Number of levels of the grouping variable. $q = 0$ when there is no grouping variable
c	Natural response rate
$\mathbf{X}$	$n \times (p+q)$ matrix with element x_{ij}, which represents the *j*th covariate for the *i*th covariate pattern
γ	$p \times 1$ vector with element γ_j, which represents the slope parameter of the *j*th independent variable
α	$q \times 1$ vector with element α_j, which represents the parameter for the *j*th level of the grouping variable
β	$(p+q) \times 1$ vector which is a composite of γ and α
s	Total number of parameters in the model, equal to $p+q$ if the natural response rate is set to a constant, $p+q+1$ if the natural response rate is to be estimated by the model

[1] This algorithm applies to SPSS 5.0 and later releases. To learn about algorithms for previous releases, call SPSS Technical Support.

Model

The model assumes a dichotomous dependent variable with probability P for the event of interest. Since the procedure assumes aggregated data for every covariate pattern, the random variable y_i takes a binomial distribution.

$$P(y_i = r_i) = \binom{n_i}{r_i} P_i^{r_i} (1 - P_i)^{n_i - r_i} \qquad i = 1, \ldots, m$$

Hence, the log likelihood, L, for m observations after ignoring the constant factor can be written as

$$L = \sum_{i=1}^{m} r_i \ln P_i + (n_i - r_i) \ln(1 - P_i)$$

For dose-response models, it is further assumed that

$$P_i = c + (1 - c) F\left(\mathbf{X}_i' \beta\right) \qquad \textbf{(1)}$$

where $\mathbf{X}_i$ is the vector of covariates for the ith covariate pattern and $F\left(\mathbf{X}_i' \beta\right)$ has two forms:

$$F\left(\mathbf{X}_i' \beta\right) = \begin{cases} \dfrac{e^{\mathbf{X}_i' \beta}}{1 + e^{\mathbf{X}_i' \beta}} & \text{if logit model} \\ \displaystyle\int_{-\infty}^{\mathbf{X}_i' \beta} \frac{1}{\sqrt{2\pi}} e^{-z^2/2} \, dz & \text{if probit model} \end{cases} \qquad \textbf{(2)}$$

When there is no grouping variable, x_{ij} is simply the observed value of the jth independent variable for the ith covariate pattern, and $\beta = \gamma$. When there is a grouping variable, a set of indicator variables is constructed. There will be q indicator variables $l_{i1}, \ldots, l_{iq}$ added to the $\mathbf{X}$ matrix and q parameters $\alpha_1, \ldots, \alpha_q$ added to the β vector.

$$l_{ij} = \begin{cases} 1 & \text{if the } i\text{th covariate pattern is in the } j\text{th level} \\ 0 & \text{otherwise} \end{cases}$$

Hence, the $\mathbf{X}_i$ vector has $p+q$ elements and the associated parameter vector β is expanded to $\left(\beta_1,\ldots,\beta_p,\beta_{p+1},\ldots,\beta_{p+q}\right)$, where $\alpha_j = \beta_{p+j}$.

Maximum-Likelihood Estimates (MLE)

To obtain the maximum likelihood estimates for c, and $\beta_1,\ldots,\beta_{p+q}$, set the following equations, (3) and (4), equal to 0:

$$L_c^* = \sum_{i=1}^{m} \frac{r_i - n_i P_i}{P_i\left(1-P_i\right)}\left[1 - F\left(\mathbf{X}_i'\beta\right)\right] \qquad \textbf{(3)}$$

$$L_{\beta_j}^* = \begin{cases} \left(1-c\right)\sum_{i=1}^{m} \frac{r_i - n_i P_i}{P_i\left(1-P_i\right)} x_{ij} F\left(\mathbf{X}_i'\beta\right)\left(1 - F\left(\mathbf{X}_i'\beta\right)\right) & \text{if logit model} \\ \left(1-c\right)\sum_{i=1}^{m} \frac{r_i - n_i P_i}{P_i\left(1-P_i\right)} x_{ij} \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\left(\mathbf{X}_i'\beta\right)^2\right\} & \text{if probit model} \end{cases} \qquad \textbf{(4)}$$

where $L_{\beta_j}^*$ is the derivative of L with respect to β_j , and $F\left(\mathbf{X}_i'\beta\right)$ and P_i are defined by equations (1) and (2).

Algorithm

Probit uses the algorithms proposed and implemented in NPSOL by Gill, Murray, Saunders, and Wright. The loss function for this procedure is the negative of the log-likelihood described in the model. The derivatives for the parameters are described in equations (3) and (4). The only bound for the parameters is $0 < c < 1$. For more details of the NPSOL algorithms, see CNLR (constrained nonlinear regression).

Natural Response Rate

When the user specifies a fixed number for the natural response rate, L_c^* is set to 0 for iterations and the bound for c is set equal to the fixed number.

Initial Values

The initial value for each β is set to 0. If there is a control group, the initial value of c, designated by c_0, is set to the ratio of the response to the number of subjects for the control group. If there is no control group, then c_0 is set to the minimum ratio of the response to the number of subjects, over all covariate patterns.

Criteria

Users can control two criteria, ITER and CONV. ITER is the maximum number of iterations allowed. The default value is $\max(50,3(s+1))$. CONV (criterion of convergence) is the same as the OPTOLERANCE criterion in CNLR.

Asymptotic Covariance Matrix

The asymptotic covariance matrix for the MLE $(\hat{c},\hat{\beta}_1,\ldots,\hat{\beta}_{p+q})$ is estimated by $\mathbf{I}^{-1}$, where $\mathbf{I}$ is the information matrix containing the negatives of the second partial derivatives of L.

$$\frac{\partial^2 L}{\partial c^2}=\sum_{i=1}^{m}\left[\frac{r_i-n_iP_i}{P_i(1-P_i)^2}-\frac{r_i}{P_i^2(1-P_i)}\right]\left(1-F\left(\mathbf{X}_i'\beta\right)\right)^2$$

$$\frac{\partial^2 L}{\partial c\partial\beta_j}=\sum_{i=1}^{m}x_{ij}\left((1-c)\left(1-F\left(\mathbf{X}_i'\beta\right)\right)\left[\frac{r_i-n_iP_i}{P_i(1-P_i)^2}-\frac{r_i}{P_i^2(1-P_i)}\right]-\frac{r_i-n_iP_i}{P_i(1-P_i)}\right)\frac{dF\left(\mathbf{X}_i'\beta\right)}{d\,\mathbf{X}_i'\beta}$$

where

$$\frac{dF\left(\mathbf{X}_i'\beta\right)}{d\,\mathbf{X}_i'\beta} = \begin{cases} F\left(\mathbf{X}_i'\beta\right)\left(1-F\left(\mathbf{X}_i'\beta\right)\right) & \text{if logit model} \\ \frac{1}{\sqrt{2\pi}}\exp\left\{-\frac{1}{2}\left(\mathbf{X}_i'\beta\right)^2\right\} & \text{if probit model} \end{cases}$$

$$\frac{\partial^2 L}{\partial\beta_j\partial\beta_h} = \left(1-c^2\right)\sum_{i=1}^{m}\left[\frac{r_i - n_i P_i}{P_i\left(1-P_i\right)^2} - \frac{r_i}{P_i^2\left(1-P_i\right)}\right]x_{ij}x_{ih}\left(\frac{dF\left(\mathbf{X}_i'\beta\right)}{d\mathbf{X}_i'\beta}\right)^2$$
$$+\left(1-c\right)\sum_{i=1}^{m}\left[\frac{r_i - n_i P_i}{P_i\left(1-P_i\right)}\right]x_{ij}x_{ih}\frac{d^2F\left(\mathbf{X}_i'\beta\right)}{d^2\,\mathbf{X}_i'\beta}$$

where

$$\frac{d^2F\left(\mathbf{X}_i'\beta\right)}{d^2\,\mathbf{X}_i'\beta} = \begin{cases} F\left(\mathbf{X}_i'\beta\right)\left(1-F\left(\mathbf{X}_i'\beta\right)\right)\left(1-2F\left(\mathbf{X}_i'\beta\right)\right) & \text{if logit model} \\ \frac{1}{\sqrt{2\pi}}\left(-\mathbf{X}_i'\beta\right)\exp\left(-\frac{1}{2}\left(\mathbf{X}_i'\beta\right)^2\right) & \text{if probit model} \end{cases}$$

Frequency Table and Goodness of Fit

For every covariate pattern i, $i = 1,\ldots,m$, compute

$$\hat{F}_i = \begin{cases} \dfrac{e^{\mathbf{X}_i'\beta}}{1+e^{\mathbf{X}_i'\beta}} & \text{if logit model} \\ \dfrac{1}{\sqrt{2\pi}}\displaystyle\int_{-\infty}^{\mathbf{X}_i'\beta} e^{-z^2/2}\,dz & \text{if probit model} \end{cases}$$

$$\hat{P}_i = \hat{c} + \left(1-\hat{c}\right)\hat{F}_i$$

Then the expected frequency is equal to

$$\hat{E}_i = n_i\hat{P}_i$$

The Pearson chi-square statistic is defined by

$$\chi^2 = \sum_{i=1}^{m} \frac{\left(r_i - \hat{E}_i\right)^2}{\hat{E}_i\left(1-\hat{P}_i\right)}$$

and the degrees of freedom (*df)* is

$$df = \begin{cases} (q-1)m - s & \text{if } q \geq 2 \\ m - s & \text{if } q = 1 \end{cases}$$

Fiducial Limits, RMP, and Parallelism

The parallelism test statistic, fiducial limits, and relative median potency are available when there is only one covariate (predictor variable). Assuming that $\hat{\alpha}_1,\ldots,\hat{\alpha}_q$ are the MLE's for $\alpha_1,\ldots,\alpha_q$ and $\hat{\gamma}$ is the MLE for γ, $v(\hat{\alpha}_j)$ is the asymptotic variance for $\hat{\alpha}_j$, $v(\hat{\gamma})$ is the asymptotic variance for $\hat{\gamma}$, and cov$(\hat{\alpha}_j,\hat{\gamma})$ is the asymptotic covariance for $\hat{\alpha}_j$ and $\hat{\gamma}$.

Fiducial Limits for Effective dose x

For level of the grouping variable j and P = 0.01 through 0.09, 0.10 through 0.90 (by 0.05), and 0.91 through 0.99, compute

$$y = \begin{cases} \ln\left(P / (1 - P)\right) & \text{if logit model} \\ \text{probit}(P) & \text{if probit model} \end{cases}$$

Then the effective dose x_j to obtain probability P of response for level j is defined by

$$x_j = \left(\left(y - \hat{\alpha}_j\right) / \hat{\gamma}\right)$$

and the 95% fiducial limit for effective dose x_j is computed by

$$x_j + \frac{g}{1-g}\left(x + \frac{\operatorname{cov}(\hat{\alpha}_j, \hat{\gamma})}{v(\hat{\gamma})}\right)$$

$$\pm \frac{t}{\hat{\gamma}(1-g)} \sqrt{\left\{ v(\hat{\alpha}_j) + 2x_j \operatorname{cov}(\hat{\alpha}_j, \hat{\gamma}) + x_j^2 v(\hat{\gamma}) - g\left(v(\hat{\alpha}_j) - \frac{\left(\operatorname{cov}(\hat{\alpha}_j, \hat{\gamma})\right)^2}{v(\hat{\gamma})} \right) \right\} h^*}$$

where

$$g = \begin{cases} \dfrac{t^2 v(\hat{\gamma})}{\hat{\gamma}^2} & \text{without heterogeneity factor} \\ \dfrac{t^2 v(\hat{\gamma})}{\hat{\gamma}^2} h & \text{with heterogeneity factor} \end{cases}$$

$$t = \begin{cases} 1.96 & \text{without heterogeneity factor} \\ t_{(0.025, df)} & \text{with heterogeneity factor} \end{cases}$$

$$h = x_j^2 / (df)$$

$$h^* = \begin{cases} 1 & \text{without heterogeneity factor} \\ h & \text{with heterogeneity factor} \end{cases}$$

The heterogeneity factor is used if the Pearson chi-square statistic is significant.

Note: If the covariate (predictor variable) x is transformed, transform it back to the original metrics for the estimate and its two limits. For example, if $\log_{10}$ is applied to the predictor for the analysis and $\hat{x}_L, \hat{x}$, and $\hat{x}_U$ are the lower limit, the estimate, and the upper limit on the $\log_{10}$ scale, then $10^{\hat{x}_L}$ and $10^{\hat{x}_U}$ are the lower and upper limits on the original scale.

Relative Median Potency

The relative median potency is available when there is a factor variable and the covariate is transformed. It is not available if there is no factor variable or if there is more than one covariate.

The estimate of relative median potency for group j versus group k is

$$M_{jk} = \left(\hat{\alpha}_k - \hat{\alpha}_j\right) / \hat{\gamma}$$

and its 95% confidence limit is

$$\frac{g}{1-g}\left(M_{jk} - \frac{v_{12}}{v_{22}}\right) \pm \frac{t}{\hat{\gamma}(1-g)}\sqrt{\left\{v_{11} - 2M_{jk}v_{12} + M_{jk}^2 - g\left(v_{11} - \frac{v_{12}^2}{v_{22}}\right)\right\}h^*}$$

where

$$v_{11} = v(\hat{\alpha}_j) + v\left(\hat{\alpha}_k\right) - 2\,\mathrm{cov}(\hat{\alpha}_j, \hat{\alpha}_k)$$
$$v_{12} = \mathrm{cov}(\hat{\alpha}_j, \hat{\gamma}) - \mathrm{cov}(\hat{\alpha}_k, \hat{\gamma})$$
$$v_{22} = v(\hat{\gamma})$$

Note: If the covariate (predictor variable) *x* is transformed, transform it back to the original metrics for the relative median potency.

Parallelism Test Chi-Square Statistic

The parallelism test is available only if there is a factor variable.

$$\chi^2 = \chi_0^2 - \sum_{j=1}^{q} \chi_j^2$$

where χ_0^2 is the Pearson chi-square statistic, assuming that the group variable is in the model and χ_j^2 is the Pearson chi-square for the *j*th group and the degrees of freedom for χ^2 is $q-1$.

References

Finney, D. J. 1971. *Probit analysis*. Cambridge: Cambridge University Press.

Gill, P. E., Murray, W. M., Saunders, M. A., and Wright, M. H. 1986. *User's guide for NPSOL* (*version* 4.0): *A fortran package for nonlinear programming*. Technical Report SOL 86–2, Department of Operations Research, Stanford University.