

Probits

Catalina Stefanescu ^{*}, Vance W. Berger [†]

Scott Hershberger [‡]

Abstract

Probit models belong to the class of latent variable threshold models for analyzing binary data. They arise by assuming that the binary response is the indicator of the event that an unobserved latent variable exceeds a given threshold. Estimation can be done either in a likelihood or a Bayesian framework. The probit models can be generalized for the analysis of a variety of qualitative and limited dependent variables, as well as to the analysis of correlated data.

Key Words: Binary data; Latent variables; Threshold models.

Probit models have arisen in the context of analysis of dichotomous data. Let Y_1, \dots, Y_n be n binary variables and let $\mathbf{x}_1, \dots, \mathbf{x}_n \in \mathfrak{R}^p$ denote corresponding vectors of covariates. The flexible class of Probit models may be obtained by assuming that the response Y_i ($1 \leq i \leq n$) is an indicator of the

^{*}London Business School. Email: cstefanescu@london.edu

[†]University of Maryland Baltimore County. Email: vance917@comcast.net

[‡]California State University, Department of Psychology. Email: scotth@csulb.edu

event that some unobserved continuous variable, Z_i say, exceeds a threshold, which can be taken to be zero, without loss of generality. Specifically, let Z_1, \dots, Z_n be latent continuous variables and assume that

$$Y_i = I_{\{Z_i > 0\}}, \quad \text{for } i = 1, \dots, n, \quad Z_i = \mathbf{x}_i \boldsymbol{\beta} + \varepsilon_i, \quad \varepsilon_i \sim N(0, \sigma^2), \quad (1)$$

where $\boldsymbol{\beta} \in \mathfrak{R}^p$ is the vector of regression parameters. In this formulation, $\mathbf{x}_i \boldsymbol{\beta}$ is sometimes called the index function [11]. The marginal probability of a positive response with covariate vector \mathbf{x} is given by

$$p(\mathbf{x}) = \Pr(Y = 1; \mathbf{x}) = \Pr(\mathbf{x} \boldsymbol{\beta} + \varepsilon > 0) = 1 - \Phi(-\mathbf{x} \boldsymbol{\beta}), \quad (2)$$

where $\Phi(x)$ is the standard normal cumulative distribution function. Also,

$$\text{Var}(Y; \mathbf{x}) = p(\mathbf{x})\{1 - p(\mathbf{x})\} = \{1 - \Phi(-\mathbf{x} \boldsymbol{\beta})\}\Phi(-\mathbf{x} \boldsymbol{\beta}).$$

As a way of relating stimulus and response, the Probit model is a natural choice in situations in which an interpretation for a threshold approach is readily available. Examples include attitude measurement, assigning pass/fail gradings for an examination based on a mark cut-off, and categorization of illness severity based on an underlying continuous scale [10]. The Probit models first arose in connection with bioassay [4]—in toxicology experiments, for example, sets of test animals are subjected to different levels x of a toxin. The proportion $p(x)$ of animals surviving at dose x can then be modelled as a function of x , following (2). The surviving proportion is increasing in the dose when $\beta > 0$ and it is decreasing in the dose when $\beta < 0$. Surveys of the

toxicology literature on Probit modelling are included in [7] and [9].

Probit models belong to the wider class of generalized linear models [13]. This class also includes the logit models, arising when the random errors ε_i in (1) have a logistic distribution. Since the logistic distribution is similar to the normal except in the tails, whenever the binary response probability p belongs to $(0.1, 0.9)$ it is difficult to discriminate between the logit and Probit functions solely on the grounds of goodness-of-fit. As Greene [11] remarks, "it is difficult to justify the choice of one distribution or another on theoretical grounds... in most applications, it seems not to make much difference."

Estimation of the Probit model is usually based on maximum likelihood methods. The nonlinear likelihood equations require an iterative solution; the Hessian is always negative definite, so the log-likelihood is globally concave. The asymptotic covariance matrix of the maximum likelihood estimator can be estimated by using an estimate of the expected Hessian [2], or with the estimator developed by Berndt, Hall, Hall and Hausman [3]. Windmeijer [18] provides a survey of the many goodness of fit measures developed for binary choice models, and in particular for Probits.

The following data example has been first offered by Bliss [4]. Table 1 reports the number of beetles killed after five hours of exposure to carbon disulfide at various concentrations. A probit model fitted with maximum likelihood gives

$$\Phi^{-1}[\hat{p}(x)] = -34.96 + 19.74 \cdot x.$$

The table also reports the fitted values from the probit model corresponding to different dose levels x .

Table 1: Beetles killed after exposure to carbon disulfide

Log Dose x	Number of beetles	Number killed	Fitted Values Probit
1.691	59	6	3.4
1.724	60	13	10.7
1.755	62	18	23.4
1.784	56	28	33.8
1.811	63	52	49.6
1.837	59	53	53.4
1.861	62	61	59.7
1.884	60	60	59.2

The maximum likelihood estimator in a Probit model is sometimes called a quasi-maximum likelihood estimator (QMLE) since the normal probability model may be misspecified. The QMLE is not consistent when the model exhibits any form of heteroscedasticity, nonlinear covariate effects, unmeasured heterogeneity or omitted variables [11]. In this setting, White [17] proposed a robust "sandwich" estimator for the asymptotic covariance matrix of the QMLE.

As an alternative to maximum likelihood estimation, Albert and Chib [1] developed a framework for estimation of latent threshold models for binary data, using data augmentation. The univariate Probit is a special case of this class of models, and data augmentation can be implemented by means of Gibbs sampling. Under this framework, the class of Probit regression models can be extended by using mixtures of normal distributions to model the latent data.

There is a large literature on the generalizations of the Probit model to the analysis of a variety of qualitative and limited dependent variables. For

example, McKelvey and Zavoina [14] extend the Probit model to the analysis of ordinal dependent variables, while Tobin [16] discusses a class of models in which the dependent variable is limited in range. In particular, the Probit model specified in (1) can be generalized by allowing the error terms ε_i to be correlated. This leads to a multivariate Probit model, useful for the analysis of clustered binary data. The multivariate Probit focuses on the conditional expectation given the cluster-level random effect, and thus it belongs to the class of cluster-specific approaches for modelling correlated data, as opposed to population-average approaches of which the most common example are the GEE-type methods [19].

The multivariate Probit model has several attractive features which make it particularly suitable for the analysis of correlated binary data. First, the connection to the Gaussian distribution allows for flexible modelling of the association structure and straightforward interpretation of the parameters. For example, the model is particularly attractive in marketing research of consumer choice, because the latent correlations capture the cross-dependencies in latent utilities across different items. Also, within the class of cluster-specific approaches, the exchangeable multivariate Probit model is more flexible than other fully specified models (such as the beta-binomial) which use compound distributions to account for overdispersion in the data. This is due to the fact that both underdispersion and overdispersion can be accommodated in the multivariate Probit model through the flexible underlying covariance structure. Finally, due to the underlying threshold approach, the multivariate Probit model has the potential of extensions to the analysis of clustered mixed binary and continuous data, or of multivariate binary data

([12], [15]).

Likelihood methods are one option for inference in the multivariate Probit model (see, e.g. [5]), but they are computationally difficult due to the intractability of the expressions obtained by integrating out the latent variables. As an alternative, estimation can be done in a Bayesian framework ([6], [8]) where generic prior distributions may be employed to incorporate prior information. Implementation is usually done with Markov chain Monte Carlo methods — in particular the Gibbs sampler is useful in models where some structure is imposed on the covariance matrix (e.g. exchangeability).

References

- [1] Albert, J.H. and Chib, S. (1997) Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association*, **88**, 669–679.
- [2] Amemiya, T. (1981) Qualitative response models: A survey. *Journal of Economic Literature*, **19**, 481–536.
- [3] Berndt, E., Hall, B., Hall, R., and Hausman J. (1974) Estimation and inference in nonlinear structural models. *Annals of Economic and Social Measurement*, **3/4**, 653–665.
- [4] Bliss, C.I. (1935) The calculation of the dosage–mortality curve. *Annals of Applied Biology*, **22**, 134–167.

- [5] Chan, J.S.K. and Kuk, A.Y.C. (1997) Maximum likelihood estimation for probit–linear mixed models with correlated random effects. *Biometrics*, **53**, 86–97.
- [6] Chib, S. and Greenberg, E. (1998) Analysis of multivariate probit models. *Biometrika*, **85**, 347–361.
- [7] Cox, D. (1970) *Analysis of Binary Data*. London: Methuen.
- [8] Edwards, Y.D. and Allenby, G.M. (2003) Multivariate analysis of multiple response data. *Journal of Marketing Research*, **40**, 321–334.
- [9] Finney, D. (1971) *Probit Analysis*. Cambridge: Cambridge University Press.
- [10] Goldstein, H. (2003) *Multilevel Statistical Models*. 3rd Edition, London: Arnold.
- [11] Greene, W.H. (2000) *Econometric Analysis*. 4th Edition, Englewood Cliffs, NJ: Prentice Hall.
- [12] Gueorguieva, R.V. and Agresti, A. (2001) A correlated probit model for joint modelling of clustered binary and continuous responses. *Journal of the American Statistical Association*, **96**, 1102–1112.
- [13] McCullagh, P. and Nelder, J.A. (1989) *Generalised Linear Models*. 2nd Edition, London: Chapman and Hall.
- [14] McKelvey, R.D., and Zavoina, W. (1976) A statistical model for the analysis of ordinal level dependent variables. *Journal of Mathematical Sociology*, **4**, 103–120.

- [15] Regan, M.M. and Catalano, P.J. (1999) Likelihood models for clustered binary and continuous outcomes: Application to developmental toxicology. *Biometrics*, **55**, 760–768.
- [16] Tobin, J. (1958) Estimation of relationships for limited dependent variables. *Econometrica*, **26**, 24–36.
- [17] White, H. (1982) Maximum likelihood estimation of misspecified models. *Econometrica*, **53**, 1–16.
- [18] Windmeijer, F. (1995) Goodness of fit measures in binary choice models. *Econometric Reviews*, **14**, 101–116.
- [19] Zeger, S.L., Liang, K.Y. and Albert, P.S. (1988) Models for longitudinal data: A generalized estimating equations approach. *Biometrics*, **44**, 1049–1060.