Kurt Schmidheiny
University of Basel

# Binary Response Models

*matrix-free*

## 1 Introduction

Many dependent variables of interest in economics and other social sciences can only take two values. The two possible outcomes are usually denoted by 0 and 1. Such variables are called *dummy* variables or *dichotomous* variables. Some examples:

- The labor market status of a person. The variable takes the value 1 if a person is employed and 0 if he is unemployed. The values 1 and 0 can be assigned arbitrarily.
- Voting behavior of a person. The variable takes 1 if the person votes in favor of a new policy and 0 otherwise. Again the values 1 and 0 are arbitrary.

The expected value of a dichotomous variable $y_i \in \{0,1\}$ is the probability that it takes the value 1:

$$E(y_i) = 0 \cdot P(y_i = 0) + 1 \cdot P(y_i = 1) = P(y_i = 1)\,.$$

The linear regression model, e.g. for one explanatory variable

$$y_i = \beta_0 + \beta_1 x_i + v_i, \quad E(v_i) = 0$$

is called *linear probability model* in this context. This linear model is not an adequate statistical model as the expected value $E(y_i|x_i) = \beta_0 + \beta_1 x_i$ can lie outside [0,1] and does not represent a probability. In addition, the error term is heteroscedastic as $V(v_i|x_i) = (\beta_0 + \beta_1 x_i)(1-\beta_0-\beta_1 x_i)$ depends on $x_i$.





## 2 The Econometric Model: Probit and Logit

Binary response models directly describe the response probabilities $P(y_i = 1)$ of the dependent variable $y_i$.

Consider a sample of $N$ independently and identically distributed (i.i.d.) observations $i = 1, \ldots, N$ of the dependent dummy variable $y_i$ and $K$ explanatory variables $x_{i1}, ..., x_{iK}$. The probability that the dependent variable takes value 1 is modeled as

$$P(y_i = 1|x_{i1}, ..., x_{iK}) = F(z_i) = F(\beta_0 + \beta_1 x_{i1} + ... + \beta_K x_{iK})$$

where $\beta_0$ to $\beta_k$ are $K+1$ parameters and

$$z_i = \beta_0 + \beta_1 x_{i1} + ... + \beta_K x_{iK}$$

is a *single linear index*. The function $F$ maps the single index into [0,1] and satisfies in general

$$F(-\infty) = 0, \quad F(\infty) = 1, \quad \partial F(z)/\partial z > 0.$$

The *probit* model assumes that the transformation function $F$ is the cumulative distribution function (cdf) of the standard normal distribution. The response probabilities are then

$$P(y_i = 1|x_{i1}, ..., x_{iK}) = \Phi\left(z_i\right) = \int_{-\infty}^{z_i} \phi(t)\, dt = \int_{-\infty}^{z_i} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}t^2}\, dt$$

where $\phi(.)$ is the pdf and $\Phi(.)$ the cdf of the standard normal distribution.

In the *logit* model, the transformation function $F$ is the logistic function. The response probabilities are then

$$P(y_i = 1|x_{i1}, ..., x_{iK}) = \frac{e^{z_i}}{1 + e^{z_i}} = \frac{1}{1 + e^{-z_i}}$$

Figure 1 shows the transformation function $F$ for the two models.

Note: The Logit and Probit model are almost identical and the choice of the model is usually arbitrary. However, the parameters $\beta$ of the two

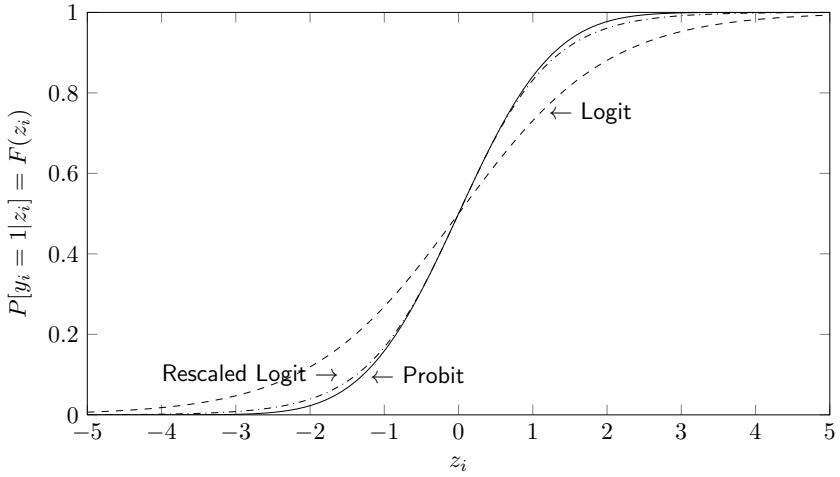


Figure 1: Mapping of the linear index $z_i$ in the probit model, the logit model and the rescaled logit model (factor 1.6).

models are scaled differently. Multiplying the parameters in the probit model by 1.6 are approximately the same as the logit estimates.[1]

## 3 Latent Variable Model

There is an alternative interpretation that gives rise to the probit (and analogously the logit) model. Consider a *latent* variable which is not observed by the researcher and linearly depends on $x_{i1}, ..., x_{iK}$

$$y_i^* = \beta_0 + \beta_1 x_{i1} + ... + \beta_K x_{iK} + u_i, \quad E(u_i|x_{i1}, ..., x_{iK}) = 0$$

[1]The factor 1.6 is derived from equating the first derivative of $F$ in the probit model with the one in the rescaled logit model. This is the appropriate rescaling for the marginal effect of the average type. An alternative approach is to equate the standard deviation of the distribution for which F is the cdf. For the probit model the standard deviation is 1 and for the logit model $\pi/\sqrt{3} \cong 1.81$.



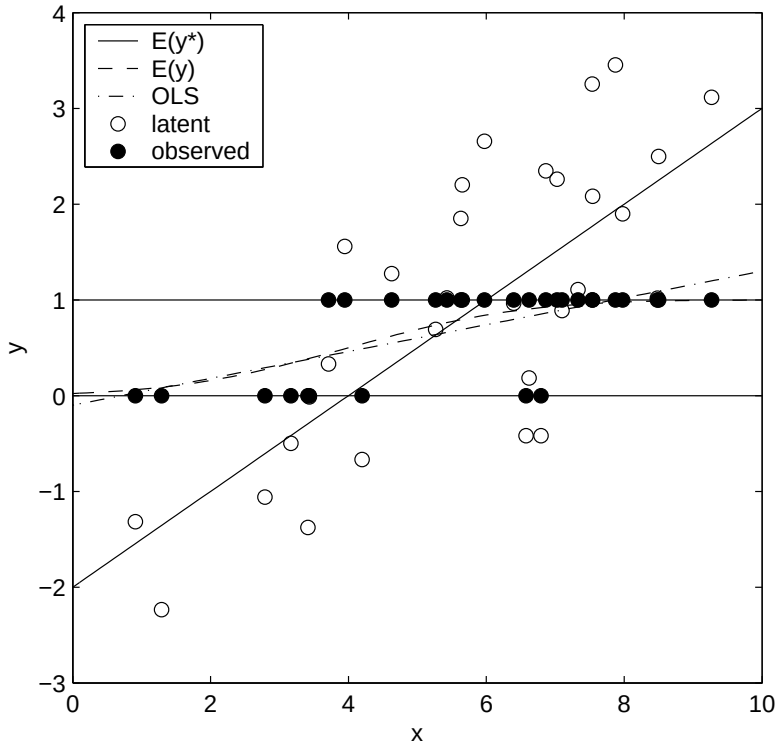


Figure 2: The probit model with a latent variable. $N = 30$, $K = 2$, $\beta_0 = -2$ and $\beta_1 = 0.5$.

The latent variable $y_i^*$ can be interpreted as the utility difference between choosing $y_i = 1$ and 0. In is then called a *random utility model*.

Only the choice $y_i$ is observed by the researcher. An individual chooses $y_i = 1$ if the latent variable is positive and 0 otherwise, hence the observed variable is

$$y_i = \begin{cases} 1 & \text{if} \quad y_i^* > 0 \\ 0 & \text{if} \quad y_i^* \leq 0 \end{cases}$$

Furthermore, assume that the individual observations $(x_{i1}, ..., x_{iK}, y_i)$ are i.i.d., that the explanatory variables are *exogenous* and that the error term is normally distributed and homoskedastsic

$$u_i | x_{i1}, ..., x_{iK} \sim N(0, \sigma^2)$$

The probability that individual $i$ chooses $y_i = 1$ can now be derived from the latent variable and the decision rule, i.e.

$$\begin{aligned} P(y_i = 1|x_{i1}, ..., x_{iK}) &= P(y_i^* > 0|x_{i1}, ..., x_{iK}) = P(z_i + u_i > 0|x_{i1}, ..., x_{iK}) \\ &= P(u_i > -z_i|x_{i1}, ..., x_{iK}) = 1 - \Phi(-z_i/\sigma) \\ &= \Phi\left(\frac{z_i}{\sigma}\right) = \Phi\left(\frac{\beta_0}{\sigma} + \frac{\beta_1}{\sigma}x_{i1} + ... + \frac{\beta_K}{\sigma}x_{iK}\right). \end{aligned}$$

The probit model arises when $\sigma^2$ is set to unity.

Note: $\beta_k$ and $\sigma$ are *not separately identified* as only the ratio $\beta_k/\sigma$ can be estimated. Figure 2 visualizes the latent variable model.

## 4 Interpretation of the Parameters

Different from the linear regression model, the parameters $\beta$ cannot directly be interpreted as marginal effects on the dependent variable $y_i$. In some situations, the index function $z_i = \beta_0 + \beta_1 x_{i1} + ... + \beta_K x_{iK}$ has a clear interpretation in a theoretical model and the marginal effect $\beta_k$ of a change in the independent variable $x_{ik}$ on $y_i^*$ is meaningful. Even then, the marginal effect is only identified if there is reason to set $\sigma^2$ to unity.

In general, we are interested in the marginal effect of a change in $x_{ik}$



on the expected value of the observed variable $y_i$, i.e.

$$\begin{aligned} \text{Probit:} \quad & \frac{\partial E(y_i|x_{i1}, ..., x_{iK})}{\partial x_{ik}} = \frac{\partial P(y_i = 1|x_{i1}, ..., x_{iK})}{\partial x_{ik}} = \phi(z_i)\beta_k \\ \text{Logit:} \quad & \frac{\partial E(y_i|x_{i1}, ..., x_{iK})}{\partial x_{ik}} = \frac{\partial P(y_i = 1|x_{i1}, ..., x_{iK})}{\partial x_{ik}} = \frac{e^{z_i}}{(1 + e^{z_i})^2}\beta_k \end{aligned}$$

This marginal effect depends on the values of all explanatory variables $x_{ik}$ for observation $i$. Therefore, any individual has a different marginal effect. There are several ways to summarize and report the information in the model. A first possibility is to present the marginal effects for the "mean type", i.e. $x_{ik} = \bar{x}_{ik}$ for all $k$ , the "median type", or some interesting extreme types. A second approach is to calculate the marginal effects for all observations in the sample and report the mean of the effects.

The estimated model can also be used for predictions

$$\begin{aligned} \text{Probit:} \quad & \widehat{P}(y_i = 1|x_{i1}, ..., x_{iK}) = \Phi(\widehat{z}_i) \\ \text{Logit:} \quad & \widehat{P}(y_i = 1|x_{i1}, ..., x_{iK}) = \frac{e^{\widehat{z}_i}}{1 + e^{\widehat{z}_i}} \end{aligned}$$

where $\widehat{z}_i = \widehat{\beta}_0 + \widehat{\beta}_1 x_{i1} + ... + \widehat{\beta}_K x_{iK}$.

For a discrete explanatory variable $x_{ik}$ it is more accurate to report the effect of a discrete change $\Delta x_{ik}$. The discrete effect of a dummy variable $x_{ik}$ changing from 0 to 1 is estimated as

$$\widehat{\Delta P} = \widehat{P}(y_i = 1|..., x_{ik} = 1, ...) - \widehat{P}(y_i = 1|..., x_{ik} = 0, ...)$$

and depends on the values of all other explanatory variables $x_{i\ell}$, $\ell \neq k$.

Predictions can also be aggregated to, for example, the predicted number of observations with $y_i = 1$. There are two prediction methods for this aggregate: (1) assume $\widehat{y}_i = 1$ if $\widehat{P}_i > 0.5$ and calculate $\sum_i \widehat{y}_i$ or (2) sum the predicted choice probabilities $\sum_i \widehat{P}(y_i = 1|x_{i1}, ..., x_{iK})$. The two measures can be contrasted to the actual numbers. Method 1 also allows to compare actual and predicted outcomes for any observation. It is also often interesting to report and contrast predicted numbers for certain types of individuals.

## 5 Estimation with Maximum Likelihood

The probit and logit models are estimated by maximum likelihood (ML). Assuming independence across observations, the likelihood function is

$$\begin{aligned}\mathcal{L} &= \prod_{\{i|y_i=0\}} P(y_i=0|x_i) \prod_{\{i|y_i=1\}} P(y_i=1|x_i) \\ &= \prod_{i=1}^{N} [1-F(z_i)]^{1-y_i} F(z_i)^{y_i}\end{aligned}$$

where $P(y_i = 1|x_{i1}, ..., x_{iK}) = F(z_i) = \Phi(z_i)$ in the probit model and $P(y_i = 1|x_{i1}, ..., x_{iK}) = F(z_i) = e^{z_i}/(1 + e^{z_i})$ in the logit model. The corresponding log likelihood function is

$$\log \mathcal{L} = \sum_{i=1}^{N} [(1-y_i) \log (1-F(z_i)) + y_i \log F(z_i)]$$

The first order conditions for an optimum are in general, for all $k$ including a constant $x_{i0} = 1$

$$\frac{\partial \log \mathcal{L}}{\partial \beta_k} = \sum_{i=1}^{N} \left[ (1-y_i) \frac{-f(z_i)}{1-F(z_i)} + y_i \frac{f(z_i)}{F(z_i)} \right] x_{ik} = 0$$

where $f(z) \equiv \partial F(z)/\partial z$. This simplifies in the probit model to

$$\frac{\partial \log \mathcal{L}}{\partial \beta_k} = \sum_{\{i|y_i=0\}} \frac{-\phi(z_i)}{1-\Phi(z_i)} x_{ik} + \sum_{\{i|y_i=1\}} \frac{\phi(z_i)}{\Phi(z_i)} x_{ik} = 0$$

and in the logit model to

$$\frac{\partial \log \mathcal{L}}{\partial \beta_k} = \sum_{i=1}^{N} \left( y_i - \frac{e^{z_i}}{1+e^{z_i}} \right) x_{ik} = 0.$$

There is no analytical solution to these FOCs and numerical optimization routines are used. The log likelihood function can be shown to be globally concave for both models and numerical routines converge well to the unique global maximum.



The ML estimator of $\beta$ is consistent and asymptotically normally distributed. The approximate distribution in large samples is

$$\widehat{\beta} \overset{A}{\sim} N(\beta, Avar(\beta))$$

where $Avar(\beta)$ is estimated by one of the standard ML procedures (inverse expected H, inverse Hessian, BHHH, or Eicker-Huber-White-Sandwich). Asymptotic hypothesis tests are performed as Wald, likelihood ratio or lagrange multiplier tests.

The ML estimation of the probit model (and analogously the logit model) rests on the strong assumption that the latent error term is normally distributed and homoscedastic. The ML estimator is inconsistent in the presence of heteroscedasticity and robust (sandwich) covariance estimators cannot solve this. Several *semi-parametric* estimation strategies have been proposed that relax the distributional assumption about the error term. See Horowitz and Savin (2001) for an introduction and Gerfin (1996) for a nice comparison of different estimators.

## 6 Estimation with OLS

Despite the logical inconsistency of the linear probability model, OLS can be used to estimate binary choice models. OLS is then called the *linear probability model* (LPM). The estimated OLS slope coefficients are estimates for the average marginal effects of the true non-linear model. In practice, the OLS slope coefficients will be very similar to the average marginal effects calculated after probit or logit estimation. However, it is very important to report robust (Eicker-Huber-White) standard errors because of the intrinsic heteroscedasticity of the linear probability model.

The linear probability model has in practice several advantages over probit or logit estimation: it is easier to calculate, the parameters are directly interpretable, fixed effects and instrumental variables estimators can easily be implemented. Note that adding fixed effects as dummy variables in the probit or logit model will yield biased estimates.

## 7 Implementation in Stata 17

The probit and logit model are estimated with the `probit` and, respectively, `logit` command. For example, load data

```
webuse auto.dta
```

and estimate the effect of the explanatory variables `weight` and `mpg` on the dependent dummy variable `foreign` with the probit model

```
probit foreign weight mpg
```

or the logit model

```
logit foreign weight mpg
```

Stata reports the inverse hessian matrix as default covariance estimator. The sandwich covariance estimator is reported with

```
probit foreign weight mpg, vce(robust)
```

Response probabilities are estimated for each observation with the post-estimation command `predict`:

```
predict p_foreign, pr
```

Marginal effects for specific types are calculated with the post-estimation command `margins`. For example, the marginal effects for a car with weight of 2000 lbs. and 40 mpg is reported by

```
margins, dydx(*) at(weight = 2000 mpg = 40)
```

The marginal effects for the mean type, e.g. a car with average weight and mpg, are calculated with

```
margins, dydx(*) atmeans
```

If explanatory dummy variables are defined as factor variables, Stata reports exact discrete effect.
Average marginal effects in the estimation sample are calculated by

```
margins, dydx(*)
```



## 8 Implementation in R 4.3.1

The probit and logit model are estimated with the command `glm` which fits generalized linear models. For example, load data

```
library(haven)
auto <- read_dta("https://www.stata-press.com/data/r17/auto.dta")
```

and estimate the effect of the explanatory variables `weight` and `mpg` on the dependent dummy variable `foreign` with the probit model

```
probit <- glm(foreign~weight+mpg, family=binomial(link=probit),
              data=auto)
summary(probit)
```

or the logit model

```
logit <- glm(foreign~weight+mpg, family=binomial(link = "logit"),
              data=auto)
summary(logit)
```

R reports the inverse hessian matrix as default covariance estimator. The sandwich covariance estimator is reported with

```
library(sandwich)
coeftest(probit, vcov=sandwich)
```

Response probabilities are estimated for each observation with `predict`:

```
p_foreign <- predict(probit,type=c("response"))
```

The R package `margins` offers a convenient way to calculate marginal effects. For example, the marginal effects for a car with weight of 2000 lbs. and 40 mpg are calculated with

```
library(margins)
margins(probit, at = list(weight=2000, mpg=40))
```

Tests and confidence bounds for maginal effects are reported with

```
mfx <- margins(probit, at = list(weight=2000, mpg=40))
summary(mfx)
```

Average marginal effects in the estimation sample are calculated by

```
margins(probit)
```

## References

### Introductory textbooks

Stock, James H. and Mark W. Watson (2020), Introduction to Econometrics, 4th Global ed., Pearson. Chapter 11.

Wooldridge, Jeffrey M. (2009), *Introductory Econometrics: A Modern Approach*, 4th ed., Cengage Learning. Chapters 17.1.

Aldrich, John and Forrest D. Nelson (1984), *Linear Probability, Logit and Probit Models*, Sage University Press.

### Advanced textbooks

Cameron, A. Colin and Pravin K. Trivedi (2005), *Microeconometrics: Methods and Applications*, Cambridge University Press. Chapter 14.

Wooldridge, Jeffrey M. (2010), *Econometric Analysis of Cross Section and Panel Data*, MIT Press. Chapter 15.

Maddala, G.S. (1983), *Limited-Dependent and Qualitative Variables in Econometrics*, Cambridge: Cambridge University Press. Chapter 2.

### Articles

Gerfin Michael (1996), Parametric and Semi-Parametric Estimation of the Binary Response Model of Labour Market Participation, *Journal of Applied Econometrics*, 11, 321-339.

Horowitz, Joel and N. Savin (2001), Binary Response Models: Logits, Probits and Semiparametrics, *Journal of Economic Perspectives*, 15(4), 43-56.