Binary Response Models

A **binary dependent variable** can take on only two values, which for practical reasons are usually coded as 0 and 1.

For example, a person may be in or out of the labor force, a commuter may drive to work or take public transit, a household may own or rent the home it resides in, and so on.

A **binary response model** tries to explain the probability that an agent chooses alternative 1 as a function of observed explanatory variables.

Let P_i denote $\Pr(y_i = 1 \mid \Omega_i)$, where Ω_i denotes an information set. A binary response model attempts to model this conditional probability.

A binary response model can also be thought of as modeling a conditional expectation, since

$$P_i \equiv \Pr(y_i = 1 \mid \Omega_i) = \mathrm{E}(y_i \mid \Omega_i). \tag{1}$$

Any reasonable binary response model must ensure that $\mathrm{E}(y_i \mid \Omega_i)$ lies in the 0-1 interval.

Commonly used models ensure that $0 < P_i < 1$ by specifying that

$$P_i \equiv \mathrm{E}(y_i \mid \Omega_i) = F(\boldsymbol{X}_i\boldsymbol{\beta}). \tag{2}$$

Here $\boldsymbol{X}_i\boldsymbol{\beta}$ is an **index function**, which maps from $\boldsymbol{X}_i$ and $\boldsymbol{\beta}$ to a scalar index, and $F(x)$ is a **transformation function**, for which

$$F(-\infty) = 0, \quad F(\infty) = 1, \quad \text{and} \quad f(x) \equiv \frac{dF(x)}{dx} > 0. \tag{3}$$

These are the properties of a CDF. Although $\boldsymbol{X}_i\boldsymbol{\beta}$ can take any value on the real line, the value of $F(\boldsymbol{X}_i\boldsymbol{\beta})$ must lie between 0 and 1.

Because $F(x)$ has to be nonlinear, changes in the values of the x_{ij}, the elements of $\boldsymbol{X}_i$, necessarily affect $\mathrm{E}(y_i \mid \Omega_i)$ in a nonlinear fashion.

Specifically, when P_i is given by (2), its derivative with respect to x_{ij} is

$$\frac{\partial P_i}{\partial x_{ij}} = \frac{\partial F(\boldsymbol{X}_i\boldsymbol{\beta})}{\partial x_{ij}} = f(\boldsymbol{X}_i\boldsymbol{\beta})\beta_j, \tag{4}$$

where β_j is the j^{th} element of $\boldsymbol{\beta}$.

- For the usual transformation functions, $f(\boldsymbol{X}_i\boldsymbol{\beta})$ achieves a maximum at $\boldsymbol{X}_i\boldsymbol{\beta} = 0$ and then falls as $|\boldsymbol{X}_i\boldsymbol{\beta}|$ increases.
- Thus the effect on P_i of a change in one of the independent variables is greatest when $P_i = 0.5$ and very small when P_i is close to 0 or 1.

This makes sense. Consider the probability of owning a house when one of the regressors is income, or even log income.

- The effect of moving from \$100,000 to \$120,000 is surely much greater than the effect of moving from \$500,000 to \$520,000.

However, assuming that $F(x)$ takes any particular functional form is quite a strong assumption.

The Probit Model

One of two widely-used choices for $F(x)$ is the cumulative standard normal distribution function,

$$\Phi(x) \equiv \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} \exp\left(-\frac{1}{2}X^2\right) dX. \tag{5}$$

When $F(\boldsymbol{X}_i\boldsymbol{\beta}) = \Phi(\boldsymbol{X}_i\boldsymbol{\beta})$, (2) is called the **probit model**.

There is no closed-form expression for $\Phi(x)$, but it is easily evaluated numerically. Its first derivative is $\phi(x)$, the standard normal density.

The probit model can be derived from a model involving an unobserved, or **latent**, variable y_i°. Suppose that

$$y_i^\circ = \boldsymbol{X}_i\boldsymbol{\beta} + u_i, \quad u_i \sim \text{NID}(0,1). \tag{6}$$

We observe only the sign of y_i°, which determines the value of the observed binary variable y_i.

The latent variable y_i° determines y_i by the equations

$$y_i = 1 \text{ if } y_i^\circ > 0; \quad y_i = 0 \text{ if } y_i^\circ \leq 0. \tag{7}$$

Think of y_i° as an index of the utility associated with some action. If the action yields positive utility, it is undertaken; otherwise, it is not.

We set $\text{Var}(u_i) = 1$ because it is not identified. If $\text{Var}(u_i)$ were some other value, say σ^2, we could divide $\boldsymbol{\beta}$, y_i°, and u_i by σ. Then $\text{Var}(u_i/\sigma) = 1$, but the value of y_i would be unchanged.

We can now compute P_i, the probability that $y_i = 1$. It is

$$\begin{aligned} \Pr(y_i = 1) &= \Pr(y_i^\circ > 0) = \Pr(\boldsymbol{X}_i\boldsymbol{\beta} + u_i > 0) && (8) \\ &= \Pr(u_i > -\boldsymbol{X}_i\boldsymbol{\beta}) = \Pr(u_i \leq \boldsymbol{X}_i\boldsymbol{\beta}) = \Phi(\boldsymbol{X}_i\boldsymbol{\beta}). && (9) \end{aligned}$$

This uses the symmetry of the normal distribution.

The final result is what we get by substituting $\Phi(\boldsymbol{X}_i\boldsymbol{\beta})$ for $F(\boldsymbol{X}_i\boldsymbol{\beta})$ in (2). We did not really need the latent variable model.

The Logit Model

The **logit model**, sometimes called **logistic regression**, is very similar to the probit model.

The function $F(x)$ is now the **logistic function**

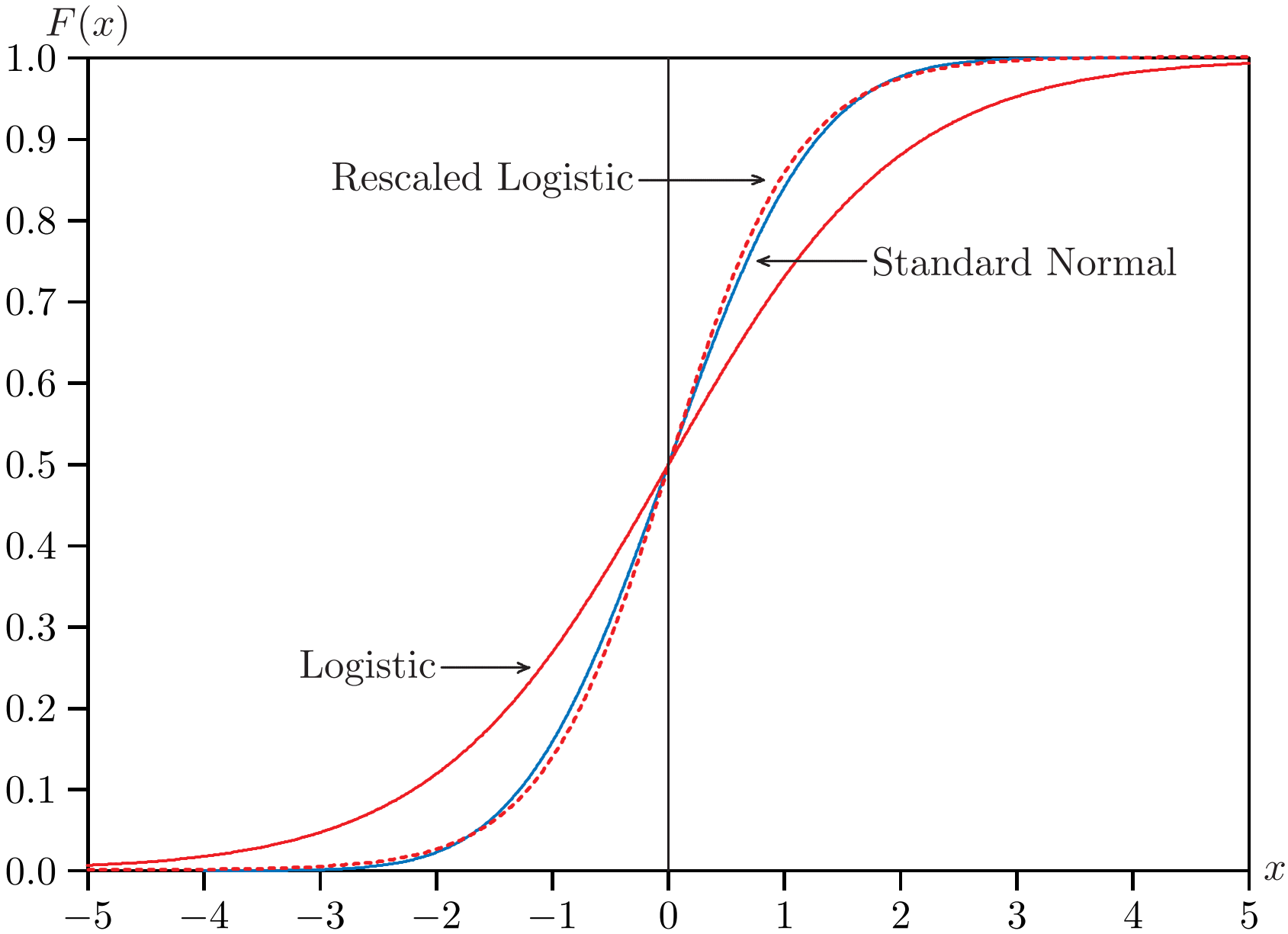
$$\Lambda(x) \equiv \frac{1}{1+e^{-x}} = \frac{e^x}{1+e^x}, \tag{10}$$

which has first derivative

$$\lambda(x) \equiv \frac{e^x}{(1+e^x)^2} = \Lambda(x)\Lambda(-x). \tag{11}$$

Because this first derivative is symmetric around zero, $\Lambda(-x) = 1 - \Lambda(x)$.

The figure shows the logistic and standard normal distribution functions, plus logistic rescaled to have variance 1 instead of $\pi^2/3$.


$F(x)$
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
Rescaled Logistic
Standard Normal
Logistic
x
−5
−4
−3
−2
−1
0
1
2
3
4
5

The logit model is most easily derived by assuming that

$$\log\left(\frac{P_i}{1-P_i}\right) = \boldsymbol{X}_i\boldsymbol{\beta}, \tag{12}$$

which says that the logarithm of the **odds** (that is, the ratio of the two probabilities) is equal to $\boldsymbol{X}_i\boldsymbol{\beta}$. Solving for P_i, we find that

$$P_i = \frac{\exp(\boldsymbol{X}_i\boldsymbol{\beta})}{1+\exp(\boldsymbol{X}_i\boldsymbol{\beta})} = \frac{1}{1+\exp(-\boldsymbol{X}_i\boldsymbol{\beta})} = \Lambda(\boldsymbol{X}_i\boldsymbol{\beta}). \tag{13}$$

This result is what we would get by letting $\Lambda(\boldsymbol{X}_i\boldsymbol{\beta})$ play the role of the transformation function $F(\boldsymbol{X}_i\boldsymbol{\beta})$ in (2).

We can also derive the logit model from a latent variable model like (7) using an extreme value distribution.

Logit and probit models generally yield very similar results.

Logit estimates generally have the same signs as probit ones, but are larger in absolute value; t statistics tend to be very similar.

Maximum Likelihood Estimation

Loglikelihood function for any linear-index binary response model:

$$\ell(\boldsymbol{y}, \boldsymbol{\beta}) = \sum_{i=1}^{N} \Big(y_i \log F(\boldsymbol{X}_i \boldsymbol{\beta}) + (1 - y_i) \log\big(1 - F(\boldsymbol{X}_i \boldsymbol{\beta})\big) \Big), \tag{14}$$

because $\Pr(y_i = 1) = F(\boldsymbol{X}_i \boldsymbol{\beta})$ and $\Pr(y_i = 0) = 1 - F(\boldsymbol{X}_i \boldsymbol{\beta})$. The maximum possible value of (14) is 0.

For the logit and probit models, symmetry implies that we can replace $1 - F(\boldsymbol{X}_i \boldsymbol{\beta})$ by $F(-\boldsymbol{X}_i \boldsymbol{\beta})$ in the second term of (14). Always do this!

In the case of the logit model, (14) can also be written as

$$\sum_{i=1}^{N} \Big(y_i (\boldsymbol{X}_i \boldsymbol{\beta}) - \log\big(1 + \exp(\boldsymbol{X}_i \boldsymbol{\beta})\big) \Big). \tag{15}$$

This version is popular in the statistical learning literature.

The first-order conditions, or **likelihood equations** are

$$\sum_{i=1}^{N} \frac{\big(y_i - F(\boldsymbol{X}_i\boldsymbol{\beta})\big) f(\boldsymbol{X}_i\boldsymbol{\beta}) x_{ij}}{F(\boldsymbol{X}_i\boldsymbol{\beta}) F(-\boldsymbol{X}_i\boldsymbol{\beta})} = 0, \;\; j = 1, \ldots, k. \tag{16}$$

It is not hard to find $\hat{\boldsymbol{\beta}}$ that satisfies these conditions, because (14) is globally concave in $\boldsymbol{\beta}$.

Conditions (16) look just like the first-order conditions for weighted least-squares estimation of the nonlinear regression model

$$y_i = F(\boldsymbol{X}_i\boldsymbol{\beta}) + v_i, \tag{17}$$

where the weight for observation i is

$$\Big(F(\boldsymbol{X}_i\boldsymbol{\beta})\, F(-\boldsymbol{X}_i\boldsymbol{\beta})\Big)^{-1/2}. \tag{18}$$

This weight is one over the square root of the variance of $v_i \equiv y_i - F(\boldsymbol{X}_i\boldsymbol{\beta})$, which is a binary random variable.

Perfect Classifiers

ML estimation of a binary response model fails when it fits too well.

Suppose there is some linear combination of the independent variables, say $\boldsymbol{X}_i\boldsymbol{\beta}^\bullet$, such that

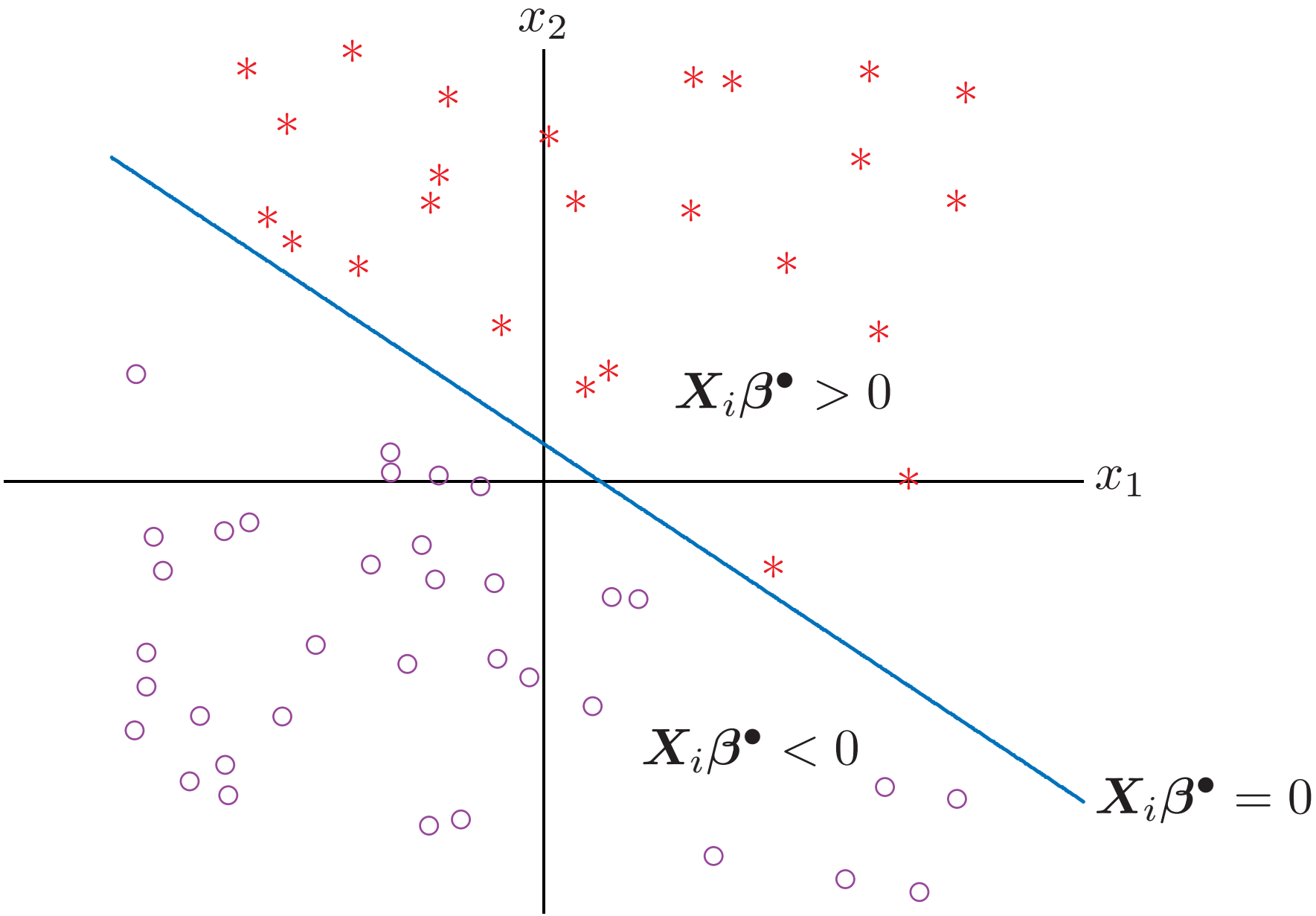
$$y_i = 0 \text{ whenever } \boldsymbol{X}_i\boldsymbol{\beta}^\bullet < 0, \text{ and} \tag{19}$$

$$y_i = 1 \text{ whenever } \boldsymbol{X}_i\boldsymbol{\beta}^\bullet > 0. \tag{20}$$

If so, it is possible to make $\ell(\boldsymbol{y}, \boldsymbol{\beta})$ arbitrarily close to 0 by setting $\boldsymbol{\beta} = \gamma\boldsymbol{\beta}^\bullet$ and letting $\gamma \to \infty$. Then $\boldsymbol{X}_i\boldsymbol{\beta}^\bullet$ is a **perfect classifier**.

Numerical optimization methods fail when there is a perfect classifier. The algorithm typically terminates at $\ell(\boldsymbol{y}, \ddot{\boldsymbol{\beta}}) \cong 0$ with all elements of $\ddot{\boldsymbol{\beta}}$ large in absolute value.

Geometrically, a perfect classifier exists if there is a **separating hyperplane**; see the figure. Note that $\boldsymbol{\beta}^\bullet$ is not unique.


x_2
x_1
$\boldsymbol{X}_i\boldsymbol{\beta}^\bullet > 0$
$\boldsymbol{X}_i\boldsymbol{\beta}^\bullet < 0$
$\boldsymbol{X}_i\boldsymbol{\beta}^\bullet = 0$

Inference

It can be shown that

$$\text{Var}\left(N^{1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)\right) = \underset{N\to\infty}{\text{plim}}\left(\frac{1}{N}\boldsymbol{X}^\top \mathbf{Y}(\boldsymbol{\beta}_0)\boldsymbol{X}\right)^{-1}, \tag{21}$$

where $\boldsymbol{X}$ is an $N \times k$ matrix with typical row $\boldsymbol{X}_i$, $\boldsymbol{\beta}_0$ is the true value of $\boldsymbol{\beta}$, and the $N \times N$ diagonal matrix $\mathbf{Y}(\boldsymbol{\beta})$ has typical diagonal element

$$\text{Y}_i(\boldsymbol{\beta}) \equiv \frac{f^2(\boldsymbol{X}_i\boldsymbol{\beta})}{F(\boldsymbol{X}_i\boldsymbol{\beta})F(-\boldsymbol{X}_i\boldsymbol{\beta})}. \tag{22}$$

In practice, we use the covariance matrix estimator

$$\widehat{\text{Var}}(\hat{\boldsymbol{\beta}}) = \left(\boldsymbol{X}^\top \mathbf{Y}(\hat{\boldsymbol{\beta}})\boldsymbol{X}\right)^{-1}. \tag{23}$$

This gives us asymptotically valid standard errors, t statistics, Wald statistics, and confidence intervals.

Consider the GNR

$$y_i - F(\boldsymbol{X}_i\boldsymbol{\beta}) = f(\boldsymbol{X}_i\boldsymbol{\beta})\boldsymbol{X}_i\boldsymbol{b} + \text{residual}. \tag{24}$$

It is not valid here, because the disturbances are not homoskedastic.

The variance of $y_i - F(\boldsymbol{X}_i\boldsymbol{\beta})$ is $V_i(\boldsymbol{\beta}) = F(\boldsymbol{X}_i\boldsymbol{\beta})F(-\boldsymbol{X}_i\boldsymbol{\beta})$.

Dividing (24) by the square root of $V_i(\boldsymbol{\beta})$ yields the **binary response model regression**, or **BRMR**. It is

$$V_i^{-1/2}(\boldsymbol{\beta})\big(y_i - F(\boldsymbol{X}_i\boldsymbol{\beta})\big) = V_i^{-1/2}(\boldsymbol{\beta})f(\boldsymbol{X}_i\boldsymbol{\beta})\boldsymbol{X}_i\boldsymbol{b} + \text{residual}. \tag{25}$$

Running this at $\hat{\boldsymbol{\beta}}$ yields the covariance matrix (23), multiplied by s^2, the squared standard error of the BRMR. Since $s^2 \to 1$, this is valid, but it is better to divide by s^2.

We can also use the BRMR within a nonlinear optimization procedure or to test restrictions on $\boldsymbol{\beta}$.

Recall that OLS regressions fit too well, causing $||\hat{\boldsymbol{u}}|| < ||\boldsymbol{u}||$. Binary response models also tend to fit too well. The $F(\boldsymbol{X}_i\hat{\boldsymbol{\beta}})$ tend to be closer to 0 and 1 than the $F(\boldsymbol{X}_i\boldsymbol{\beta}_0)$.

Overfitting causes the elements of $\hat{\boldsymbol{\beta}}$ to be biased away from zero.

One way to reduce bias is to use a **parametric bootstrap**. If we generate B bootstrap samples using $\hat{\boldsymbol{\beta}}$, we can estimate the bias as

$$\text{Bias}^*(\hat{\boldsymbol{\beta}}) = \frac{1}{B}\sum_{b=1}^{B}\hat{\boldsymbol{\beta}}_b^* - \hat{\boldsymbol{\beta}}, \tag{26}$$

where $\hat{\boldsymbol{\beta}}_b^*$ is the estimate of $\boldsymbol{\beta}$ using the b^{th} bootstrap sample. Therefore, a bias-corrected estimate is

$$\hat{\boldsymbol{\beta}}_{\text{bc}} \equiv \hat{\boldsymbol{\beta}} - \text{Bias}^*(\hat{\boldsymbol{\beta}}) = 2\hat{\boldsymbol{\beta}} - \frac{1}{B}\sum_{b=1}^{B}\hat{\boldsymbol{\beta}}_b^*. \tag{27}$$

This result applies in many cases; see MacKinnon and Smith (1998).

The Linear Probability Model

Many applied works estimate the **linear probability model**, or **LPM**,

$$y_i = \boldsymbol{X}_i\boldsymbol{\gamma} + u_i. \tag{28}$$

This model seems to make no sense, because $\mathrm{E}(y_i \mid \boldsymbol{X}_i)$ is a probability, and probabilities must lie between 0 and 1.

But there is nothing in (28) to prevent $\boldsymbol{X}_i\hat{\boldsymbol{\gamma}} < 0$ or $\boldsymbol{X}_i\hat{\boldsymbol{\gamma}} > 1$. If any element of $\boldsymbol{X}_i$ takes on an extreme value, this is likely to happen.

Thus the linear probability model is not a sensible way to model conditional probabilities if regressors can take on extreme values.

Note that u_i in (28) is heteroskedastic. We must use HCCME or CRVE.

Using (28) is probably not harmful if all regressors are dummy variables. Suppose that $\boldsymbol{X}_i = [1 \;\; d_i]$ where $d_i = 0$ or 1. Then $\mathrm{E}(y_i \mid \boldsymbol{X}_i)$ can take only two values.

For the linear probability model,

$$\mathrm{E}(y_i \,|\, d_i = 0) = \gamma_1 \quad \text{and} \quad \mathrm{E}(y_i \,|\, d_i = 1) = \gamma_1 + \gamma_2. \tag{29}$$

The OLS estimate $\hat{\gamma}_1$ is just the average of the y_i for $d_i = 0$, and the OLS estimate $\hat{\gamma}_1 + \hat{\gamma}_2$ is just the average of the y_i for $d_i = 1$.

For the probit model,

$$\mathrm{E}(y_i \,|\, d_i = 0) = \Phi(-\beta_1) \quad \text{and} \quad \mathrm{E}(y_i \,|\, d_i = 1) = \Phi(\beta_1 + \beta_2). \tag{30}$$

What we will probably find if we estimate both models is that

$$\hat{\gamma}_1 \cong \Phi(-\hat{\beta}_1) \quad \text{and} \quad \hat{\gamma}_1 + \hat{\gamma}_2 \cong \Phi(\hat{\beta}_1 + \hat{\beta}_2). \tag{31}$$

The parameter estimates will look different, but the fitted values will probably be very similar.

This is true even with many dummies. But the LPM can behave badly when there are continuous regressors that can take on extreme values.

Cluster-Robust Inference

For logit and probit models with clustered disturbances, the CRVE is

$$\widehat{\text{Var}}(\hat{\boldsymbol{\beta}}) = \left(\boldsymbol{X}^\top \mathbf{Y}(\hat{\boldsymbol{\beta}})\boldsymbol{X}\right)^{-1} \left(\sum_{g=1}^{G} \boldsymbol{s}_g(\hat{\boldsymbol{\beta}}) \boldsymbol{s}_g^\top(\hat{\boldsymbol{\beta}}) \right) \left(\boldsymbol{X}^\top \mathbf{Y}(\hat{\boldsymbol{\beta}})\boldsymbol{X}\right)^{-1}, \tag{32}$$

where $\boldsymbol{s}_g$ is the score vector for cluster g, of which a typical element is

$$s_{gj} = \sum_{i=1}^{N_g} \left(\frac{y_{gi}}{F(\boldsymbol{X}_{gi}\hat{\boldsymbol{\beta}})} + \frac{y_{gi}-1}{F(-\boldsymbol{X}_{gi}\hat{\boldsymbol{\beta}})} \right) f(\boldsymbol{X}_{gi}\hat{\boldsymbol{\beta}}) X_{gij}, \quad j = 1, \ldots, k, \tag{33}$$

and $\mathbf{Y}(\boldsymbol{\beta})$ is an $N \times N$ diagonal matrix with typical diagonal element

$$\Upsilon_i(\boldsymbol{\beta}) \equiv \frac{f^2(\boldsymbol{X}_i\boldsymbol{\beta})}{F(\boldsymbol{X}_i\boldsymbol{\beta})F(-\boldsymbol{X}_i\boldsymbol{\beta})}. \tag{34}$$

Without clustering, the covariance matrix would be $\left(\boldsymbol{X}'\mathbf{Y}(\hat{\boldsymbol{\beta}})\boldsymbol{X}\right)^{-1}$.

Stata replaces $\boldsymbol{X}'\mathbf{Y}(\hat{\boldsymbol{\beta}})\boldsymbol{X}$ in (32) by the Hessian of the loglikelihood function. This estimator is identical for logit but not for probit.

Instrumental Variables

One or more of the regressors in $\boldsymbol{X}$ may be endogenous. If so, probit and logit estimates are inconsistent.

The control function approach is an easy way to correct for endogeneity in this case.

Suppose that some columns of $\boldsymbol{X}$, say $\boldsymbol{Y}$, are endogenous, and that $\boldsymbol{W}$ is a matrix of instruments.

Simply add $\boldsymbol{M}_{\boldsymbol{W}}\boldsymbol{Y}$ as additional regressors in a binary response model to obtain consistent estimates of $\boldsymbol{\beta}$; see Rivers and Vuong (1988).

We cannot just use reported variance matrix, because the control functions are generated regressors.

We can either use analytical results or employ bootstrap methods.

For other methods, see Lewbel, Dong, and Yang (2012) and Lewbel and Dong (2015).