ApEc 8213: Econometric Analysis III -- Lecture #2

Binary Choice Models
Hansen, Chapter 25

I. Introduction (25.1, 25.2)

In many situations your ***Y* variable** is a **dummy variable**. Examples are employment and unemployment, firm exit, and purchase of a major durable good (e.g. car).

In these situations, the **relationship of interest** is the **probability**, conditional on *X*, **that *Y* equals 1**: **Prob[*Y* = 1| *X*]**. Often we want to estimate the **marginal effect** of some *X* variable on this probability. Common examples of these models are probit, logit and linear probability models.

The **response probability** of *Y* with respect to variables *X* is:

$$P(x) = \text{Prob}[Y = 1 | X = x] = \text{E}[Y | X = x]$$

Since *Y* takes only 2 values, $P(x)$ fully describes the distribution of *Y*. Also of interest is the **marginal effect**:

$$\partial P(x)/\partial x = \partial \text{Prob}[Y = 1 | X = x]/\partial x = \partial \text{E}[Y | X = x]/\partial x$$

We can write this as a regression model as follows:

$$Y = P(X) + e, \quad \text{where } \mathrm{E}[e|\, X] = 0$$

Note that the **error term is also binary** (takes only 2 values):

$$e = 1 - P(X), \text{ with probability } P(X)$$
$$= - P(X), \quad \text{with probability } 1 - P(X)$$

You should be able to show that the conditional variance of e is:

$$\mathrm{Var}[e|\, X] = P(X)(1 - P(X))$$

II. Models for the Response Probability (25.3)

The three most common regression models for binary variables are:

Linear Probability Model (LPM): $P(x) = x'\beta$.

The LPM can be easily **estimated using OLS**, and it also allows for estimating more complex models, such as panel data models and IV models. It has the **disadvantage** that **predicted probabilities can be < 0 or > 1**.

Probit Model: $P(x) = \Phi(x'\beta)$, where Φ = standard normal cdf.

Logit Model: $P(x) = \Lambda(x'\beta)$, where, $\Lambda(u) = (1 + e^{-u})^{-1}$.

Both the probit and the logit are special cases of …

Index Model: $P(x) = G(x'\beta)$, where $G(\ \ )$ is a distribution (cdf). Note that $x'\beta$ is linear in x, and is called a **linear index** function. **In most cases**, including probit and logit, ***G*(*u*) is symmetric around 0**, which implies that $G(-u) = 1 - G(u)$.

A final characteristic of $G(x'\beta)$ is that:

$$\partial P(x)/\partial x = \beta g(x'\beta)$$

where $g(u)$ is the density function of $G(u)$: $\partial G(u)/\partial u = g(u)$.

On pages 831-32 **Hansen also presents** the **Linear Series** and **Index Series models**. The linear series model is an LPM that includes transformations of x, such as squared terms and interaction terms, and the index series model is an index model that includes transformations of x. Hansen denotes the expanded set of variables by x_K. I think that **most economists just think of these as straightforward modifications of the LPM and Index models**. Hansen shows on page 832 that such models can have better fits than "strictly linear" (no interaction/squared terms) models.

III. Latent Variable Representation (25.4)

Index models can be interpreted as latent variable models:

$$Y^* = X'\beta + e$$

$$e \sim G(e)$$

$$Y = 1[Y^* > 0]$$

The *continuous* latent variable *Y is not observed. *Y* is observed.** $Y = 1$ if $Y^* > 0$, and $Y = 0$ if $Y^* \leq 0$. The error term e is distributed as $G(e)$, which is symmetric around zero.

If $Y = 1$, then $Y^* > 0$, which implies:

$$X'\beta + e > 0$$

This in turn implies that the **response probability** is:

$$P(x) = \text{Prob}[e > -x'\beta] = 1 - G(-x'\beta) = G(x'\beta)$$

(The last = hold because $G(u)$ is symmetric around zero.

This latent variable representation **can be interpreted as the *relative* utility of choosing between *Y* = 1 and *Y* = 0**. More specifically, Y^* = (utility of $Y = 1$) – (utility of $Y = 0$). $Y = 1$ implies $Y^* > 0$, so (utility of $Y = 1$) > (utility of $Y = 0$).

Note: If $G(u)$ is the standard normal distribution, this is a probit model, and if $G(u)$ is logistic, this is a logit model.

It turns out that $\boldsymbol{\beta}$ **is not identified**. However, $\boldsymbol{\beta^* = \beta/\sigma}$ **is identified**, where σ is the standard deviation of u. To see this, suppose that $e = \sigma\varepsilon$, where ε is distributed as $G(\ )$ with a variance of 1. Then:

$$\text{Prob}[Y = 1 | X = x] = \text{Prob}[\sigma\varepsilon > -x'\beta] = G(x'\beta/\sigma) = G(x'\beta^*)$$

This means that **we can estimate $\boldsymbol{\beta^*}$**, that is β/σ, but we **cannot estimate $\boldsymbol{\beta}$ and $\boldsymbol{\sigma}$ separately**. For probit models we usually set $\sigma = 1$, and for logit we set $\sigma = \pi/\sqrt{3} = 1.814$. **In practice, probits and logits yield similar results**, except that logit coefficients are about 1.8 times larger than probit coefficients.

Note finally, that although we cannot estimate β, **for any two variables $\boldsymbol{x_1}$ and $\boldsymbol{x_2}$ we can estimate $\boldsymbol{\beta_1/\beta_2}$**, since this equals β_1^*/β_2^*. Thus we can estimate the **relative impacts of different variables**. **We can also estimate marginal effects of any $\boldsymbol{X}$ variable** since marginal effects depend only on β^*, not β. That is, for a specific x, $\partial P(x)/\partial x = (\beta/\sigma)g(x'(\beta/\sigma)) = \beta^* g(x'\beta^*)$.

IV. Maximum Likelihood Estimation (25.5 and 25.6)

Probit and logit models are **almost always estimated using maximum likelihood estimation**. Recall from Lecture 1 (p. 7) that, for a binary (Bernoulli) variable, $\text{Prob}[Y = 1] = p$ and $\text{Prob}[Y = 0] = 1 - p$. Thus for observed Y we have:

$$\pi(y) = p^y(1-p)^{y-1}$$

Both probits and logits are index models, and conditional on X, we have Prob$[Y = 1|X] = G(x'\beta)$. Thus we have:

$$\pi(Y|X) = G(X'\beta)^Y(1 - G(X'\beta))^{1-Y} = G(X'\beta)^Y(G(-X'\beta))^{1-Y} = G(Z'\beta)$$

where $Z = X$ if $Y = 1$ and $Z = -X$ if $Y = 0$.

To obtain the log-likelihood function, take the log of $\pi(Y|X)$ and sum over all of the observations:

$$\ell_n(\beta) = \sum_{i=1}^{n} \log\left(G(Z_i'\beta)\right)$$

For the logit and probit models we have:

$$\ell_n^{\text{probit}}(\beta) = \sum_{i=1}^{n} \log\left(\Phi(Z_i'\beta)\right)$$

$$\ell_n^{\text{logit}}(\beta) = \sum_{i=1}^{n} \log\left(\Lambda(Z_i'\beta)\right)$$

Define the **first derivative** of $\log(G(x))$ as $h(x) = \partial\log(G(x))/\partial x$ and the (negative) **second derivative** as $H(x) = \partial^2\log(G(x))/\partial x^2$. For the **logit model**, these are:

$$h_{\text{logit}}(x) = 1 - \Lambda(x)$$

$$H_{\text{logit}}(x) = \Lambda(x)(1 - \Lambda(x))$$

For the **probit model**, these are:

$$h_{\text{probit}}(x) = \phi(x)/\Phi(x), \text{ which is defined as } \lambda(x)$$

$$H_{\text{probit}}(x) = \lambda(x)(x + \lambda(x))$$

The function $\lambda(x)$ is called the **inverse Mills ratio**.

What we really need are the **derivatives with respect to $\boldsymbol{\beta}$**. The general result for index models are:

$$S_n(\beta) = \partial \ell_n(\beta)/\partial\beta = \sum_{i=1}^{n} Z_i h(Z_i'\beta)) \quad \text{(score)}$$

$$\mathscr{H}_n(\beta) = -\partial^2 \ell_n(\beta)/\partial\beta\partial\beta' = \sum_{i=1}^{n} X_i X_i' H(Z_i'\beta)) \quad \text{(hessian)}$$

Hansen explains on pp.834-35 that **both $\ell_n^{\text{probit}}(\boldsymbol{\beta})$ and $\ell_n^{\text{logit}}(\boldsymbol{\beta})$ are globally concave in $\boldsymbol{\beta}$**, which means that there is **only one value of $\boldsymbol{\beta}$ for which $\boldsymbol{S_n(\beta) = 0}$**, which is the unique maximum. The practical consequence of this is that almost any numerical method for maximizing these likelihood function will quickly reach that value of β. That is, probit and logit maximum likelihood estimation is quite fast and almost never fails to converge.

Finally, we can **define the ML estimate of $\boldsymbol{\beta}$ as the value that maximizes the empirical likelihood function**:

$$\hat{\beta}^{\text{probit}} = \arg\max_{\beta} \ell_n^{\text{probit}}(\beta)$$

$$\hat{\beta}^{\text{logit}} = \arg\max_{\beta} \ell_n^{\text{logit}}(\beta)$$

There is **no explicit solution for $\widehat{\boldsymbol{\beta}}^{\text{probit}}$ or $\widehat{\boldsymbol{\beta}}^{\text{logit}}$**, the **solutions must be found by** choosing a random starting value and then **iterating** until a maximum is reached.

V. Pseudo-true Values, Correctly Specified vs. Misspecified

The expectation of the log likelihood of index models, $\ell(\beta)$, is:

$$\ell(\beta) = \mathrm{E}[\log(G(Z'\beta))]$$

The model (probit or logit) is **correctly specified if there exists a coefficient $\boldsymbol{\beta_0}$ such that Prob[$Y = 1| X$] = $G(X'\boldsymbol{\beta})$**. When this holds, then β_0 maximizes $\ell(\beta)$:

$$\beta_0 = \arg\max_{\beta} \ell(\beta)$$

Alternatively, the model is **misspecified if there is no $\boldsymbol{\beta}$ such that Prob[$Y = 1| X$] = $G(X'\boldsymbol{\beta})$**. In this case, the model $G(X'\beta)$ is an **approximation**, and we can **define the pseudo-true coefficient $\boldsymbol{\beta_0}$** as the **value that maximizes $\boldsymbol{\ell(\beta)}$** (even though, for this β_0, Prob[$Y = 1| X$] $\neq G(X'\beta_0)$).

You can think of this coefficient as the "best fit" β for Prob[$Y = 1|X$] when it is approximated by $G(X'\beta)$.

In either case (correctly specified or misspecified), if the log of $G(\ )$ is concave, then $\ell(\beta)$ is globally concave. Note: the log of $G(\ )$ is concave for both probit and logit.

This can be shown using the following expression:

$$\boldsymbol{Q}(\beta) = -\partial^2\ell(\beta)/\partial\beta\partial\beta' = \mathrm{E}[XX'H(Z'\beta)], \text{ where } H(x) = \partial^2\log(G(x))/\partial x^2$$

Log concavity of $G(\ )$ implies that $H(Z'\beta) > 0$ ($H(Z'\beta)$ is positive definite), which in turn implies that $\boldsymbol{Q}(\beta) \geq 0$ ($\boldsymbol{Q}(\beta)$ is positive semi-definite). The slightly stronger condition that $\boldsymbol{Q}(\beta) > 0$ ($\boldsymbol{Q}(\beta)$ is positive definite) implies that β_0 is the unique value that maximizes $\ell(\beta)$, even if the model is missepcified.

For completeness, the probit and logit versions of $\ell(\beta)$ are:

$$\ell^{\text{probit}}(\beta) = \mathrm{E}[\log(\Phi(Z'\beta))]$$

$$\ell^{\text{logit}}(\beta) = \mathrm{E}[\log(\Lambda(Z'\beta))]$$

The pseudo-true values (which are "true" values if the model is correctly specified) are:

$$\beta^{\text{probit}} = \arg\max_{\beta} \ell^{\text{probit}}(\beta)$$

$$\beta^{\text{logit}} = \arg\max_{\beta} \ell^{\text{logit}}(\beta)$$

The expectation of $-\partial^2 \ell(\beta)/\partial\beta\partial\beta'$:

$$\boldsymbol{Q}_{\text{probit}} = \mathrm{E}[XX' H_{\text{probit}}(Z'\beta^{\text{probit}})]$$

$$\boldsymbol{Q}_{\text{logit}} = \mathrm{E}[XX' \Lambda(X'\beta^{\text{logit}})(1 - \Lambda(X'\beta^{\text{logit}}))]$$

VI. Consistency and Asymptotic Distribution

Maximum likelihood estimates of logit and probit models converge to the true values of β (or to the pseudo-true values as defined on the previous page, if the model is misspecified, as long as $\boldsymbol{Q}(\beta) > 0$):

Theorem 25.1. Consistency of Logit Estimation.
If (Y_i, X_i) are i.i.d, $\mathrm{E}[||X||] < \infty$, and $\boldsymbol{Q}_{\text{logit}} > 0$ (i.e. $\boldsymbol{Q}_{\text{logit}}$ is positive definite), then $\hat{\beta}^{\text{logit}} \underset{p}{\to} \beta^{\text{logit}}$ as $n \to \infty$.

Theorem 25.2. Consistency of Probit Estimation.
If (Y_i, X_i) are i.i.d, $\mathrm{E}[||X||^2] < \infty$, and $\boldsymbol{Q}_{\text{probit}} > 0$ (i.e. $\boldsymbol{Q}_{\text{probit}}$ is positive definite), then $\hat{\beta}^{\text{probit}} \underset{p}{\to} \beta^{\text{probit}}$ as $n \to \infty$.

The following two theorems show that $\hat{\beta}^{\text{logit}}$ and $\hat{\beta}^{\text{probit}}$ are asymptotically normally distributed:

Theorem 25.3. If the conditions for Theorem 25.1 hold, $E[\|X\|^4] < \infty$, and β^{logit} is in the interior of B (possible values for β), then as $n \to \infty$:

$$\sqrt{n}(\hat{\beta}^{\text{logit}} - \beta^{\text{logit}}) \xrightarrow[d]{} N(0, \boldsymbol{V}_{\text{logit}})$$

where $\boldsymbol{V}_{\text{logit}} = \boldsymbol{Q}_{\text{logit}}^{-1}\boldsymbol{\Omega}_{\text{logit}}\boldsymbol{Q}_{\text{logit}}^{-1}$ and $\boldsymbol{\Omega}_{\text{logit}} = E[X'X(Y - \Lambda(X'\beta^{\text{logit}}))^2]$

Theorem 25.4. If the conditions for Theorem 25.2 hold, $E[\|X\|^4] < \infty$, and β^{probit} is in the interior of B (possible values for β, then as $n \to \infty$:

$$\sqrt{n}(\hat{\beta}^{\text{probit}} - \beta^{\text{probit}}) \xrightarrow[d]{} N(0, \boldsymbol{V}_{\text{probit}})$$

where $\boldsymbol{V}_{\text{probit}} = \boldsymbol{Q}_{\text{probit}}^{-1}\boldsymbol{\Omega}_{\text{probit}}\boldsymbol{Q}_{\text{probit}}^{-1}$ and $\boldsymbol{\Omega}_{\text{probit}} = E[X'X\lambda(Z'\beta^{\text{probit}})^2]$ and $\lambda(\ )$ is defined at the top of page 7.

You can think of these $\boldsymbol{V}$ matrices as "robust" estimates that work even if the model specifications are not correct but are only approximations. However, if the model specifications are correct, then these matrices simplify to $\boldsymbol{V}_{\text{logit}} = \boldsymbol{Q}_{\text{logit}}^{-1}$ and $\boldsymbol{V}_{\text{probit}} = \boldsymbol{Q}_{\text{probit}}^{-1}$, and there is a simplication for $\boldsymbol{\Omega}_{\text{probit}}$: $\boldsymbol{\Omega}_{\text{probit}} = \boldsymbol{Q}_{\text{probit}} = E[X'X\lambda(X'\beta^{\text{probit}})\lambda(-X'\beta^{\text{probit}})]$.

VII. Covariance Matrix Estimation

Theorems 25.3 and 25.4 can be used to estimate the covariance matrices for the logit and probit specifications.

For the logit, define $\widehat{\Lambda}_i = \Lambda(X_i\hat{\beta}^{\text{logit}})$, and define:

$$\widehat{\boldsymbol{Q}}_{\text{logit}} = (1/n)\sum_{i=1}^{n} X_i X_i' \widehat{\Lambda}_i(1 - \widehat{\Lambda}_i)$$

$$\widehat{\boldsymbol{\Omega}}_{\text{logit}} = (1/n)\sum_{i=1}^{n} X_i X_i' (Y_i - \widehat{\Lambda}_i)^2$$

The "robust" (sandwich) estimate for $\boldsymbol{V}_{\text{logit}}$ is $\widehat{\boldsymbol{V}}_{\text{logit}} = \widehat{\boldsymbol{Q}}_{\text{logit}}^{-1}\widehat{\boldsymbol{\Omega}}_{\text{logit}}\widehat{\boldsymbol{Q}}_{\text{logit}}^{-1}$. Under the assumption of correct specification this simplifies to $\widehat{\boldsymbol{V}}^{0}_{\text{logit}} = \widehat{\boldsymbol{Q}}_{\text{logit}}^{-1}$.

For the probit, define $\hat{\mu}_i = Z_i'\hat{\beta}^{\text{probit}}$, $\hat{\lambda}_i = \lambda(\hat{\mu}_i)$, and:

$$\widehat{\boldsymbol{Q}}_{\text{probit}} = (1/n)\sum_{i=1}^{n} X_i X_i' \hat{\lambda}_i(\hat{\mu}_i - \hat{\lambda}_i)$$

$$\widehat{\boldsymbol{\Omega}}_{\text{probit}} = (1/n)\sum_{i=1}^{n} X_i X_i' \hat{\lambda}_i^{\,2}$$

The "robust" (sandwich) estimate for $\boldsymbol{V}_{\text{probit}}$ is $\widehat{\boldsymbol{V}}_{\text{probit}} = \widehat{\boldsymbol{Q}}_{\text{probit}}^{-1}\widehat{\boldsymbol{\Omega}}_{\text{probit}}\widehat{\boldsymbol{Q}}_{\text{probit}}^{-1}$. Under the assumption of correct specification this simplifies to:

$$\widehat{\boldsymbol{V}}^{0}_{\text{probit}} = \widehat{\boldsymbol{Q}}^{0}_{\text{probit}} = (1/n)\sum_{i=1}^{n} X_i X_i' \lambda(X_i\hat{\beta}^{\text{probit}})\lambda(-X_i\hat{\beta}^{\text{probit}})$$

In Stata, the default matrices are $\widehat{\boldsymbol{V}}^{0}_{\text{logit}}$ and $\widehat{\boldsymbol{V}}^{0}_{\text{probit}}$. To get the robust/sandwich matrices use "vce(robust)" option.

VIII. Marginal Effects

Unlike the linear probability model, the estimated logit and probit coefficients are not the marginal effects of the X variables. It is useful to calculate the marginal effects. For the general index model, where Prob[$Y = 1 | X = x$] = $G(x'\beta)$, the vector of marginal effects, denoted by $\delta(x)$, is:

$\delta(x) = \partial P(x)/\partial x = \beta g(x'\beta)$, where $g(x'\beta)$ = density of $G(x'\beta)$

Since this is a function of x, it is useful to calculate the average over the distribution of X. The **average marginal effect** (which is a vector) is:

$$\text{AME} = \text{E}[\delta(X)] = \beta \text{E}[g(X'\beta)]$$

To calculate this, estimate $\delta(x)$ by $\hat{\delta}(x) = \hat{\beta} g(x'\hat{\beta})$. Then we have:

$$\widehat{\text{AME}} = \hat{\beta}(1/n) \sum_{i=1}^{n} g(X_i'\hat{\beta})$$

This is for the simple case where X does not have squared terms, interaction effects, etc. Calculating marginal

effects when you have such terms is more complicated. See page 839 of Hansen for an example.

Hansen discusses three topics for which we do not have time to cover: semiparametric models (Section 25.11), IV probit (Section 25.12) and panel data (Section 25.13).