Lecture 8

Econ241c, Spring 2015
Lecture 8: Binary Response Models

Next lectures will be focused on limited dependent variables:

Binary response models: \( y_i = \{0, 1\} \)
Count / multinomial data: \( y_i = \{1, 2, 3\} \)
Censored regression models: \( y_i = 0 \) or \( y_i > 0 \) and continuous
Duration analysis

Today: Binary response models

Examples
Linear probability model
Probit and logit model
Latent variables representation
ML estimation
Models with binary dependent variables:

Binary: \( y_i = \{0, 1\} \)

Examples:

\( y_i = 1 (\text{person } i \text{ is currently employed}) \)

\( y_i = 1 (\text{person } i \text{ loan application is approved}) \)

Issue: Because of the binary nature of \( y_i \), a linear model for \( E[y_i|x_i] \) may be inappropriate

Probit and logit are alternatives
Application: determinants of loan approvals

- Data are from Boston in 1990

- Focus on $\Pr(\text{loan is rejected} \mid \text{Debt to income ratio})$

- `summ reject di white pubrec mortg_cs cons_cs married`

<table>
<thead>
<tr>
<th>Variable</th>
<th>Obs</th>
<th>Mean</th>
<th>Std. Dev.</th>
<th>Min</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>reject</td>
<td>1989</td>
<td>0.1226747</td>
<td>0.3281459</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>di_ratio</td>
<td>1989</td>
<td>32.38903</td>
<td>8.26302</td>
<td>0</td>
<td>95</td>
</tr>
<tr>
<td>loan_value</td>
<td>1989</td>
<td>73.83875</td>
<td>18.75115</td>
<td>2</td>
<td>257.1429</td>
</tr>
<tr>
<td>white</td>
<td>1989</td>
<td>0.8451483</td>
<td>0.3618542</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>pubrec</td>
<td>1989</td>
<td>0.0688788</td>
<td>0.2533117</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>mortg_cs</td>
<td>1989</td>
<td>1.708396</td>
<td>0.5552338</td>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td>cons_cs</td>
<td>1989</td>
<td>2.109603</td>
<td>1.663024</td>
<td>1</td>
<td>6</td>
</tr>
<tr>
<td>married</td>
<td>1986</td>
<td>0.6586103</td>
<td>0.4742953</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>
Distribution of loan approval data and D/I Ratio

Note: data are “jittered” around 1 and 0 for better visual display.
Response probability

In binary response models, the interest is in estimating the probability that \( y_i \) takes the value 1, conditional on some covariates

This is sometimes called “response probability” or “probability of success” – even though the event may not represent “success”:

\[
p(x_i) = \Pr(y_i = 1 | x_i)
\]

Since \( y_i = \{0, 1\} \), \( y_i | x_i \) is a Bernoulli random variable, so \( p(x_i) \) fully characterizes the probability distribution:

\[
\Pr(y_i = 1 | x_i) = f(1 | x_i) = p(x_i)
\]
\[
\Pr(y_i = 0 | x_i) = f(0 | x_i) = 1 - p(x_i)
\]
\[
E(y_i | x_i) = p(x_i)
\]
\[
\text{Var}(y_i | x_i) = p(x_i)[1 - p(x_i)]
\]
Linear probability model

The linear probability model (LPM) for a binary response model is:

\[ \Pr(y_i = 1|x_i) = E(y_i|x_i) = x_i'\beta \]

Thus we can write the regression equation:

\[ y_i = E(y_i|x_i) + \epsilon_i = x_i'\beta + \epsilon_i \]

LPM is a linear regression model when the dependent variable is \( \{0, 1\} \)

The marginal effect of a small change in \( x_{ij} \) (if continuous) is:

\[ \frac{\partial}{\partial x_{ij}} \Pr(y_i = 1|x_i) = \beta_j \]

\( \beta_j \) measures the increase in the response probability associated with a 1-unit increase in \( x_{ij} \), holding all other \( x \)'s constant. Note that the marginal effect is constant: it does not change with the value of \( x_{ij} \) and this can cause problems...
LPM on loan application data:

- In Stata:

- `reg reject di_ratio, robust;`

Regression with robust standard errors

|                | Coef.  | Std. Err. | t     | P>|t|  | [95% Conf. Interval] |
|----------------|--------|-----------|-------|------|---------------------|
| reject di_ratio| 0.0069501 | 0.00112  | 6.21  | 0.000 | 0.0047536 - 0.0091466 |
| _cons          | -0.1024322 | 0.03521  | -2.91 | 0.004 | -0.1714846 - 0.0333798 |
Advantage of LPM

The LPM is the best linear predictor (BLP) of \( y_i \) given \( x_i \) (in MSE sense), so weak functional form assumptions are required.

No distributional assumption on \( \varepsilon_i | x_i \) (beside having mean 0).

Easy to implement and interpret with more complicated data structures (ex: panel data).

Limiting distribution is known, and simple to compute:

\[
\sqrt{n}(\hat{\beta} - \beta) \rightarrow N(0, H^{-1}JH^{-1})
\]

\[
H = E[x_i x'_i]
\]

\[
J = E[\varepsilon_i^2 x_i x'_i]
\]
Disadvantage of LPM (1)

LPM may predict response probabilities greater than 1 or smaller than 0

The main cause of this problem comes from the linearity of the CEF, which implies that the marginal effects are constant

A 1-unit increase in $x_{ij}$ always change $Pr(y_i = 1|x_i)$ by the same amount (i.e. by $\beta_j$) regardless of the initial value of $x_{ij}$

However: If the model is completely saturated (1 dummy variable for every possible permutation of the vector $x_i$), then the LPM is completely general. As an intermediate case, we can include nonlinear effects in the elements of $x_i$ to possible alleviate the problem of predicted probabilities that are greater/ smaller than 1 or 0.
Example of overfitted probabilities

With $d_{i\text{ ratio}}$ small enough, predicted $Pr(\text{reject})$ is negative
Aside: saturated model

If the model is completely saturated (1 dummy variable for possible permutation of the vector \(x_i\)), then the LPM is completely general.

In that case, the fitted probabilities are simply the average of \(y_i\) in each of the “cells” defined by the vector \(x_i\), which are always between 0 and 1 by definition (since \(y_i = \{0, 1\}\)).

Example: \(x_i = \{1, 2, 3\}\). Corresponding saturated model is:

\[
y_i = \beta_0 + \beta_1 D_{1i} + \beta_2 D_{2i}
\]

\[
D_{1i} = 1(x_i = 1)
\]
\[
D_{2i} = 1(x_i = 2)
\]
Example of saturated model
With larger sample you would increase the number of “steps” until the fitted curve is completely general.
Disadvantage of LPM (2)

Model is heteroskedastic by construction since the variance of $y_i|x_i$ is $p(x_i)[1 - p(x_i)]$

This affects the inference, but not the consistency of the estimator

Two solutions:

1. Use “heteroskedasticity-robust” inference (customary practice now)

2. Correct the heteroskedasticity by using weighted least squares (not done much anymore in this context). Weight given by:

$$\sqrt{p(x_i)[1 - p(x_i)]}$$
Solution to LPM problem (1)

Use index models that restrict the functional form of the response probability such that $0 < \Pr(y_i = 1|x_i) < 1$

Put in another way, index models have the form:

$$\Pr(y_i = 1|x_i) = F(x_i'\beta) \quad 0 < F(z) < 1, \quad \forall z$$

These are called index models because they restrict the way in which the response probability depends on $x_i$. $\Pr(y_i = 1|x_i)$ depends on $x_i$ only through the index $x_i'\beta$

In most applications, $F()$ will be a cumulative distribution function (which always map real numbers to a number between 0 and 1)
Probit and Logit models

For index models we have \( \Pr(y_i = 1|x_i) = F(x'_i\beta) \)

The probit model is given by:

\[
\Pr(y_i = 1|x_i) = \Phi(x'_i\beta)
\]

Where \( \Phi(x'_i\beta) \) is in the CDF of the standard normal distribution

Logit model:

\[
\Pr(y_i = 1|x_i) = \Lambda(x'_i\beta) = \frac{\exp(x'_i\beta)}{1 + \exp(x'_i\beta)}
\]

Where \( \Lambda(x'_i\beta) \) is the CDF of the standard logistic distribution
Graph of probit and logit CDF
Latent variable interpretation

The specific functional form of $F()$ can be derived from a latent variable model:

$$y_i^* = x_i' \beta + \epsilon_i$$
$$y_i = 1(y_i^* > c)$$

Note: latent here means “partially observed”

Where $\epsilon_i$ is a continuously distributed iid random variable (with distribution symmetric around 0) and independent of $x_i$.

The threshold “c” is a number. We do not need to know this number, but since it only changes the interpretation of the intercept of the model, so we normalize it to 0.
Latent variable representation for probit and logit models

**Probit model**: Assume $\varepsilon_i|x_i \sim N(0, 1)$

Note the normalization $\sigma^2 = 1$. Proceeding:

\[ y_i^* = x'_i \beta + \varepsilon_i \]
\[ y_i = 1(y_i^* > 0) \]

\[ \Pr(y_i = 1|x_i) = \Pr(\varepsilon_i > -x'_i \beta|x_i) = 1 - \Phi(-x'_i \beta) = \Phi(x'_i \beta) \]

**Logit model**: Assume $\varepsilon_i|x_i \sim \text{Logistic}(0, 1)$

Normalize $Var(\varepsilon_i|x_i) = \pi^2 / 3$. Proceeding:

\[ \Pr(y_i = 1|x_i) = \Pr(\varepsilon_i > -x'_i \beta|x_i) = 1 - \Lambda(-x'_i \beta) = \Lambda(x'_i \beta) \]
Comparison of standard normal and logistic

![Graph comparing standard normal and standard logistic distributions.](image-url)
Economic interpretation of latent variables

1. Random utility model

\[ U_{1i} = x_i' \gamma_1 + \nu_{1i} \quad \text{“utility if choose action 1”} \]
\[ U_{2i} = x_i' \gamma_2 + \nu_{2i} \quad \text{“utility if choose action 2”} \]

The system can be represented with a latent variable framework:

\[ y_i^* = U_{1i} - U_{2i} \]
\[ \beta = \gamma_1 - \gamma_2 \]
\[ \epsilon_i = \nu_{1i} - \nu_{2i} \]

\[ y_i^* = x_i' \beta + \epsilon_i \]
\[ y_i = 1(y_i^* > 0) \]

So \( y_i = 1\) (choose action 1)
Economic interpretation of latent variables

2. ‘Scoring’

Suppose that person i, with characteristics $x_i$ applies for a mortgage loan ($x_i$ includes characteristics of the house buyer (e.g., income) and house (e.g., price, location, etc))

The banker then creates a score representing the “credit-notworthyness” of the loan applicant:

$$y_i^* = x_i' \beta + \varepsilon_i$$

The higher the score, the less likely it is that the bank will get its money back, and thus the less likely the bank will approve it.

Suppose the bank rejects all loans such that $y^* > 0$

We observe in the data $y_i$ (=1 if loan is rejected) and $x_i$. 
Marginal effects in probit and logit models

For continuous regressors, the marginal effects in the probit model are computed as follows:

\[
\frac{\partial \Pr(y_i = 1|\mathbf{x}_i)}{\partial x_{ij}} = \frac{\partial \Phi(x_i'\beta)}{\partial x_{ij}} = \beta_j \phi(x_i'\beta)
\]

For continuous regressors, the marginal effects in the logit model are computed as follows:

\[
\frac{\partial \Pr(y_i = 1|\mathbf{x}_i)}{\partial x_{ij}} = \frac{\partial \Lambda(x_i'\beta)}{\partial x_{ij}} = \beta_j \frac{\exp(x_i'\beta)}{[1 + \exp(x_i'\beta)]^2}
\]

Note the similarity: \(\beta_j\) multiplied by the pdf at \(x_i'\beta\)
More on marginal effects

Marginal effect changes as $x_{ij}$ changes since the pdf is a nonlinear function (unlike LPM)

Since by definition $f() > 0$ (it is a pdf), the sign of $\beta_j$ will determine the sign of the marginal effect. That is generally true in binary and multinomial response models (when y takes a fixed set of integer values)

That is, the estimated coefficients $\beta$ — i.e. the maximizers of the likelihood function are only informative about the sign of the marginal effect

You should report the average marginal effect (or it’s value evaluated at some value for the vector $x$ such as $\bar{x}$) instead of the ML estimates of $\beta$
Estimation

Binary response models are typically estimated using ML techniques (since we have already specified a distribution in the latent variable representation)

Suppose we have data on \( w_i = (y_i, x_i) : i = 1, \ldots, n \) drawn iid from the model below

Each observation on \( y_i | x_i \) is drawn from the Bernoulli distribution \( f(0|x_i) = 1 - p(x_i) \) and \( f(1|x_i) = p(x_i) \)

Thus the pdf of \( y_i | x_i \) is:

\[
\begin{align*}
  f(y_i|x_i; \theta_0) &= \Pr(y_i = 1|x_i)^y_i \Pr(y_i = 0|x_i)]^{1-y_i} \\
  &= [F(x_i'; \theta_0)]^{y_i} [1 - F(x_i'; \theta_0)]^{1-y_i}
\end{align*}
\]
The joint pdf of the sample is given by the product of the marginal pdfs:

\[ f(w_1, w_2, ..., w_n; \theta_0) = \prod_{i=1}^{n} [F(x'_i \theta_0)]^{y_i} [1 - F(x'_i \theta_0)]^{1-y_i} \]

Thus, the log-likelihood function is given by:

\[ \log L(\theta) = \sum_{i=1}^{n} y_i \log F(x'_i \theta) + (1 - y_i) \log [1 - F(x'_i \theta)] \]
Large-sample properties

The extremum estimator is defined as follows:

\[ \hat{\theta} = \arg\max_{\theta \in \Theta} \frac{1}{n} \sum_{i=1}^{n} y_i \log F(x'_i \theta) + (1 - y_i) \log [1 - F(x'_i \theta)] \]

The probit and logit models estimators of \( \theta_0 \) are obtained by plugging in the appropriate form for \( F(x'_i \theta) \)

Provided we make the right choice for the distribution of \( y_i | x_i \), the estimator is identified (recall the KL identification result)

The estimator is consistent given the usual regularity conditions (i.e. continuity, uniform convergence etc – Thm 2.5)
Asymptotic distribution

Assuming we specify the correct distribution (in addition to the other regularity conditions – Thm 3.3), the ML estimator with binary dependent variable is asymptotically normal, with limiting distribution:

\[ \sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{d} N(0, J^{-1}) \]

Since we proceed with ML, the information matrix equality holds (which explains the simpler form for \( \text{Avar}(\hat{\theta}) \)). The \( J \) matrix is given by:

\[
J = E \left[ \frac{f(x_i'\theta_0)^2 x_i x_i'}{F(x_i'\theta_0)[1 - F(x_i'\theta_0)]} \right]
\]
Probit estimates in loan application example

In Stata:

```
.probit reject di_ratio, robust;
```

Iteration 0:  log likelihood = -740.34659
Iteration 1:  log likelihood = -712.97982
Iteration 2:  log likelihood = -712.84527
Iteration 3:  log likelihood = -712.84526

Probit estimates

|                   | Coef. | Std. Err. | z     | P>|z|   | [95% Conf. Interval] |
|-------------------|-------|-----------|-------|-------|----------------------|
| reject            |       |           |       |       |                      |
| di_ratio          | 0.0313371 | 0.0052126 | 6.01  | 0.000 | 0.0211205, 0.0415536  |
| _cons             | -2.213679  | 0.1833123 | -12.08| 0.000 | -2.572965, -1.854394  |

Number of obs = 1989
Wald chi2(1) = 36.14
Prob > chi2 = 0.0000
Pseudo R2 = 0.0371
Probit marginal effects in loan application example

- In Stata:

  . dprobit reject di_ratio, robust;

Iteration 0:   log likelihood = -740.34659
Iteration 1:   log likelihood = -712.97982
Iteration 2:   log likelihood = -712.84527
Iteration 3:   log likelihood = -712.84526

Probit estimates

|             | dF/dx   | Std. Err. | z    | P>|z| | x-bar  | [    95% C.I.   |  
|-------------|---------|-----------|------|------|-------|----------------|
| reject      |         |           |      |      |       |                |
| di_ratio    | .0060947| .0009734  | 6.01 | 0.000| 32.389 | .004187 .008002|
| obs. P      | .1226747|           |      |      |       |                |
| pred. P     | .115322 | (at x-bar)|      |      |       |                |

z and P>|z| are the test of the underlying coefficient being 0
Probit estimates in loan application example