Introduction

• In some situations, the object of interest is a binary variable:
  – Do married women have a paid job or not?
  – Which families own their own car?
  – Which students will pass the exam?

• The binary variable is usually coded as
  \[ y_i = \begin{cases} 
    1 & \text{if the woman works} \\
    0 & \text{if she doesn’t} 
  \end{cases} \]

  We want to explain \( y_i \) \((i = 1, 2, \ldots, N)\) given \( K \) characteristics in \( x_i \).

• Often applications in micro-econometrics:
  Explain the choices of individuals and firms.
Outline of the Lecture

1. Interpretation of a linear regression model, \( y_i = x'_i \beta + \epsilon_i \), where \( y_i \) is binary.


4. Likelihood analysis.


The Simple Linear Probability Model

- Consider a linear regression
  \[ y_i = x'_i \beta + \epsilon_i, \quad i = 1, 2, ..., N, \]
  where \( y_i \) is binary, and \( x_i \) is a \( K \times 1 \) vector of binary or quantitative characteristics.

- We assume that \( E[\epsilon_i \mid x_i] = 0 \) such that \( E[y_i \mid x_i] = x'_i \beta \).
  Since \( y_i \) is binary, the conditional expectation can be found as
  \[
  x'_i \beta = E[y_i \mid x_i] = 1 \cdot \text{Prob}(y_i = 1 \mid x_i) + 0 \cdot \text{Prob}(y_i = 0 \mid x_i)
  = \text{Prob}(y_i = 1 \mid x_i).
  \]
  Therefore \( x'_i \beta \) is interpretable as the probability that \( y_i = 1 \).

- The derivative, \( \beta \), is interpretable as the change in the probability
  \[
  \Delta \text{Prob}(y_i = 1 \mid x_i) = \beta \cdot \Delta x_i,
  \]
  and not the change in \( y_i \).
Complications...

(1) Since $x_i^\prime \beta$ is a probability, it should be bounded:

$$0 \leq x_i^\prime \beta \leq 1.$$ 

This requires that $x_i$ is bounded or that $\beta$ is restricted in some way.

(2) The error term is highly non-normal and heteroskedastic.

The error term, $\epsilon_i$, can take two values (conditional on $x_i$)

<table>
<thead>
<tr>
<th>$y_i$</th>
<th>$\epsilon_i$</th>
<th>With Prob.</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>$1 - x_i^\prime \beta$</td>
<td>$x_i^\prime \beta$</td>
</tr>
<tr>
<td>0</td>
<td>$-x_i^\prime \beta$</td>
<td>$1 - x_i^\prime \beta$</td>
</tr>
</tbody>
</table>

The mean is given by

$$E[\epsilon_i \mid x_i] = (1 - x_i^\prime \beta) \cdot x_i^\prime \beta + (-x_i^\prime \beta) \cdot (1 - x_i^\prime \beta) = 0.$$ 

The variance is

$$V[\epsilon_i \mid x_i] = (1 - x_i^\prime \beta)^2 \cdot x_i^\prime \beta + (-x_i^\prime \beta)^2 \cdot (1 - x_i^\prime \beta) = x_i^\prime \beta \cdot (1 - x_i^\prime \beta),$$

which depends on $x_i$ and $\beta$.

Binary Choice Models

- As an alternative, consider a class of models

$$\text{Prob}(y_i = 1 \mid x_i) = F(x_i^\prime \beta)$$

for some link function $F(\cdot)$ that always lies between 0 and 1.
Choice of Link Functions

- All distribution functions are valid as link functions. Common choices are

1. The normal distribution leading to the so-called probit model

\[ F(w) = \int_{-\infty}^{w} \frac{1}{\sqrt{2\pi}} \exp \left\{ -\frac{t^2}{2} \right\} dt. \]

2. The logistic distribution leading to the logit model

\[ F(w) = \frac{e^w}{1 + e^w}. \]

Interpretation of Coefficients

- The signs of the coefficients are directly interpretable, but not the magnitudes.

- To interpret the model we can look at the marginal effects or the slopes:

\[ \frac{\partial \text{Prob}(y_i = 1 \mid x_i)}{\partial x_{ik}} = \frac{\partial F(x'_i \beta)}{\partial x_{ik}} = F'(x'_i \beta) \beta_k. \]

They depend on the values of \( x_i \). Often evaluated in

(a) the sample mean, or
(b) a standard individual defined by characteristics \( x^* \).

- The marginal effects depend on \( F'(x'_i \beta) \).

To have the same marginal effects in a logit and probit model, the coefficients have to be different. Note that

\[ F'_{\text{probit}}(0) = \frac{1}{\sqrt{2\pi}} \approx 0.4 \]

\[ F'_{\text{logit}}(0) = \frac{\exp(0)}{(1 + \exp(0))^2} = 0.25 \]

so there is a scale difference of \( 0.4/0.25 = 1.6 \).
• The derivatives are only approximations if $x_k$ is a dummy variable. An alternative is to calculate the discrete effect of a specific change:
\[
\text{Prob}(y_i = 1 \mid x) - \text{Prob}(y_i = 1 \mid x + \delta) = F(x'\beta) - F((x + \delta)'\beta).
\]
Here $x$ could be the mean or a standard person $x^*$.

• For a specific variable $x_1$ this could be
\[
F(\beta_0 + x_1\beta_1 + x_2\beta_2 + ... x_k\beta_k) - F(\beta_0 + (x_1 + 1)\beta_1 + x_2\beta_2 + ... x_k\beta_k).
\]
E.g. a dummy taking values $x_1 = 0$ versus $x_1 = 1$.

---

**Underlying Latent Model**

• A particular way to motivate the binary choice model is the so-called latent variable representation.

• Consider a woman, $i$, and consider her decision to work or not to work. Assume that the utility gain from starting to work, $y_i^*$, is a function of observed characteristics, $x_i$, and unobserved characteristics, $\epsilon_i$, i.e.
\[
y_i^* = x_i'\beta + \epsilon_i.
\]
Note that $y_i^*$ is unobservable. It is called a latent variable. Assume that the distribution of $\epsilon_i \mid x_i$ is symmetric, i.e. $F(\epsilon) = 1 - F(-\epsilon)$.

• We can always normalize so that she goes to work if $y_i^* > 0$. 
We only observe the decision: \( y_i = \begin{cases} 1 & \text{if she goes to work} \\ 0 & \text{otherwise} \end{cases} \)

We can calculate the probability:

\[
\begin{align*}
\text{Prob}(y_i = 1 \mid x_i) &= \text{Prob}(y_i^* > 0 \mid x_i) \\
&= \text{Prob}(x_i \beta + \epsilon_i > 0 \mid x_i) \\
&= \text{Prob}(\epsilon_i > -x_i \beta \mid x_i) \\
&= 1 - \text{Prob}(\epsilon_i \leq -x_i \beta \mid x_i) \\
&= \text{Prob}(\epsilon_i \leq x_i \beta \mid x_i) \\
&= F(x_i \beta).
\end{align*}
\]

The binary choice model can therefore be written in the latent variable representation. For the probit model that is

\[
y_i^* = x_i \beta + \epsilon_i, \quad \epsilon_i \sim N(0, 1), \quad y_i = \begin{cases} 1 & \text{if } y_i^* > 0 \\ 0 & \text{if } y_i^* \leq 0 \end{cases}
\]

For the logit model \( N(0, 1) \) is replaced by a standard logistic distribution.

### Likelihood Analysis

We have a fully specified model and a natural estimator is maximum likelihood. The density for the binary variable is the probability

\[
f(y_i \mid x_i; \beta) = \text{Prob}(y_i = 1 \mid x_i; \beta)^{y_i} \cdot \text{Prob}(y_i = 0 \mid x_i; \beta)^{1-y_i}.
\]

Assuming that the individuals are independent, we have a likelihood function:

\[
L(\beta) = f(y_1, ..., y_N \mid \beta) \]

\[
= \prod_{i=1}^{N} \text{Prob}(y_i = 1 \mid x_i; \beta)^{y_i} \cdot \text{Prob}(y_i = 0 \mid x_i; \beta)^{1-y_i}
\]

\[
= \prod_{i=1}^{N} F(x_i \beta)^{y_i} \cdot (1 - F(x_i \beta))^{1-y_i}.
\]

The log-likelihood function is given by

\[
\log L(\beta) = \sum_{i=1}^{N} [y_i \cdot \log (F(x_i \beta)) + (1 - y_i) \cdot \log (1 - F(x_i \beta))].
\]

The f.o.c. is \( s(\beta) = 0 \), but the likelihood equations have no analytical solution.
Measuring Goodness-of-fit

- The idea of $R^2$ is to compare with a model with only a constant. Let
  \[ \log L_1 \text{ maximized log-likelihood function} \]
  \[ \log L_0 \text{ maximized value in a model with only a constant term.} \]

- Two possible goodness-of-fit measures are
  \[
  \text{McFadden} - R^2 = 1 - \frac{\log L_1}{\log L_0} \\
  \text{Pseudo} - R^2 = 1 - \frac{1}{1 + \frac{2}{N} (\log L_1 - \log L_0)}.
  \]

- Note that in the model with only a constant, $\text{Prob}(y_i = 1) = p$.
  This is a simple binomial model and $\hat{p}_{ML} = N_1/N$, where $N_1 = \sum y_i$.
  In this case $\log L_0$ is given by
  \[
  \log L_0 = N_1 \cdot \log \left( \frac{N_1}{N} \right) + (N - N_1) \cdot \log \left( 1 - \frac{N_1}{N} \right).
  \]

Predicted Values

- We can construct predictions from the probabilities:
  \[
  \hat{y}_i = 1 \text{ if } F(x_i'\hat{\beta}) > 1/2 \\
  \hat{y}_i = 0 \text{ if } F(x_i'\hat{\beta}) \leq 1/2.
  \]

- To measure the goodness-of-fit we can compare $y_i$ and $\hat{y}_i$ in a simple cross table:
  \[
  \begin{array}{ccc}
  & \hat{y}_i & \\
  y_i & 0 & 1 \\
  0 & n_{00} & n_{01} & N_0 \\
  1 & n_{10} & n_{11} & N_1 \\
  \hline
  n_0 & n_1 & N
  \end{array}
  \]

- Let $p = N_1/N > 1/2$.
  Then a simple model saying $\hat{y}_i = 1$ for all $i$ will have $N_1/N$ correct predictions.
  To beat that we should have at least
  \[
  \frac{n_{00} + n_{11}}{N} > \frac{N_1}{N}.
  \]
Empirical Example: Teaching Economics!

- We analyze whether a new way to teach economics (called PSI) is effective or not. We have data for 32 classes of which 14 was exposed to PSI.

- The data set includes

<table>
<thead>
<tr>
<th>Variable</th>
<th>Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>GRADE</td>
<td>0.344</td>
</tr>
<tr>
<td>GPA</td>
<td>3.117</td>
</tr>
<tr>
<td>TUCE</td>
<td>21.938</td>
</tr>
<tr>
<td>PSI</td>
<td>0.438</td>
</tr>
</tbody>
</table>

- We want to estimate a binary choice model, i.e.

$$\text{Prob}(\text{GRADE}_i = 1 \mid x_i) = F(x_i'\beta),$$

where $F(\cdot)$ is a logit or probit function and $x_i = (1, \text{GPA}, \text{TUCE}, \text{PSI})'$.

Results of Linear regression, Logit, and Probit

<table>
<thead>
<tr>
<th>Variable</th>
<th>Linear</th>
<th>Logit</th>
<th>Probit</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>-1.498 (2.75)</td>
<td>...</td>
<td>-13.021 (2.64)</td>
</tr>
<tr>
<td>GPA</td>
<td>0.464 (2.78)</td>
<td>0.464</td>
<td>2.826 (2.24)</td>
</tr>
<tr>
<td>TUCE</td>
<td>0.010 (0.55)</td>
<td>0.010</td>
<td>0.095 (0.67)</td>
</tr>
<tr>
<td>PSI</td>
<td>0.379 (2.38)</td>
<td>0.379</td>
<td>2.379 (2.23)</td>
</tr>
</tbody>
</table>

- Note that the results are very close (in terms of derivative).
- Note the scale differences for $\beta_4$ to PSI

$$\frac{\hat{\beta}_{\text{Logit}}}{\hat{\beta}_{\text{Probit}}} = \frac{2.379}{1.426} = 1.668.$$
### Results for 3 Logit Estimations

<table>
<thead>
<tr>
<th>Variable</th>
<th>Model (A)</th>
<th>Model (B)</th>
<th>Model (C)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Coeff.</td>
<td>Coeff.</td>
<td>Coeff.</td>
</tr>
<tr>
<td></td>
<td>$t$–ratio</td>
<td>$t$–ratio</td>
<td>$t$–ratio</td>
</tr>
<tr>
<td>Constant</td>
<td>$-13.021$</td>
<td>$-10.656$</td>
<td>$-0.647$</td>
</tr>
<tr>
<td></td>
<td>$-2.64$</td>
<td>$-2.63$</td>
<td>$-1.74$</td>
</tr>
<tr>
<td>GPA</td>
<td>$2.826$</td>
<td>$2.538$</td>
<td>...</td>
</tr>
<tr>
<td></td>
<td>$2.24$</td>
<td>$2.15$</td>
<td>...</td>
</tr>
<tr>
<td>TUCE</td>
<td>$0.095$</td>
<td>$0.086$</td>
<td>...</td>
</tr>
<tr>
<td></td>
<td>$0.67$</td>
<td>$0.64$</td>
<td>...</td>
</tr>
<tr>
<td>PSI</td>
<td>$2.379$</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td></td>
<td>$2.23$</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>Log-likelihood</td>
<td>$-12.890$</td>
<td>$-15.991$</td>
<td>$-20.592$</td>
</tr>
<tr>
<td>$N$</td>
<td>32</td>
<td>32</td>
<td>32</td>
</tr>
<tr>
<td>Pseudo$–R^2$</td>
<td>0.325</td>
<td>0.223</td>
<td>...</td>
</tr>
<tr>
<td>McFadden$–R^2$</td>
<td>0.374</td>
<td>0.223</td>
<td>...</td>
</tr>
</tbody>
</table>

- PSI is significant.
  - The $t$–ratio (Wald test) is 2.23 with a $p$–value of 0.025 in $N(0, 1)$.
  - The likelihood ratio test is
    \[
    LR = -2 \cdot (-15.991 - (-12.890)) = 6.202,
    \]
    with a $p$–value of 0.013 in a $\chi^2(1)$.

- To see the failure rate, look at a cross tabulation:

<table>
<thead>
<tr>
<th>$\hat{y}_i$</th>
<th>0</th>
<th>1</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>$y_i$</td>
<td>0</td>
<td>18</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>3</td>
<td>8</td>
</tr>
</tbody>
</table>

  A naive model ($\hat{y}_i = 0$) would have $21/32 = 0.656$ correct predictions.
  The logit model has $26/32 = 0.813$.

- To find the numerical effect of PSI we look at an average person.
  I.e. with GPA = 3.117 and TUCE = 21.938.

- The probabilities of an improvement in test scores are $F(x'\hat{\beta})$:
  \[
  \text{Prob}(\text{GRADE} = 1 \mid \text{GPA} = 3.117, \text{TUCE} = 21.938, \text{PSI} = 0) = 0.1068
  \]
  \[
  \text{Prob}(\text{GRADE} = 1 \mid \text{GPA} = 3.117, \text{TUCE} = 21.938, \text{PSI} = 1) = 0.5633
  \]
  The effect seems numerically important.