UNIT 1. Discrete Choice Models

Instructor: Henry Redondo

From: *Econometric Analysis of Cross-Section and Panel Data* by J.M. Wooldridge

Dpto. de Fundamentos del Análisis Económico.
Universidad de Alicante

Econometrics II. Academic Year 2023/2024

Contents

1. Introduction
2. The linear probability model
3. Index models: Probit and Logit models
   - Partial effects
   - Maximum likelihood estimation
   - Estimated partial effects
   - Goodness of fit measures
   - Comparing probit, logit and LPM estimation results
4. Multinomial response models
   - Multinomial logit
   - Probabilistic choice models
   - Independence from irrelevant alternatives
5. Ordered response models
   - Ordered probit and ordered logit

Introduction

- Discrete choice models are models where the variable to be explained, y, is a **random variable taking on a finite number of outcomes**; in practice, the number of outcomes is usually small.
- The leading case is when y is binary indicating whether or not a certain event has occurred. We model y as taking values 0 and 1. Examples:
  - **Labor Economics:** Female labor force participation; Self-employment or wage work; Retirement; Migration.
  - **Population and Family Economics:** Marriage; Divorce; Number of children; Contraceptive choices.
  - **Industrial Organization:** Localization decisions of firms; entry/exit into/from a market.
  - **Economics of Education:** High school drop out; go to college.
  - **Public choice:** Voting.
  - **Consumers behavior:** brand choice in a differentiated product market; purchase of durable goods.

Introduction

It is convenient to distinguish between:

- **Binary choices:** the dependent variable can take two values (1 when the decision is taken, 0 otherwise)
- **Multiple choices:** the dependent variable can take more than two values

Regardless of the definition of y, it is traditional to refer to $y = 1$ as a success and $y = 0$ as a failure.

Introduction

- As in linear models, we often call y the explained variable, the response variable, the dependent variable, or the endogenous variable; $\mathbf{x} = (x_1, x_2, .., x_k)'$ is the vector of explanatory variables, regressors, independent variables, exogenous variables, or covariates.
- In binary response models, interest lies primarily in the response probability,

  $$P(y = 1 \,|\, \mathbf{x}) = P(y = 1 \,|\, x_1, x_2, .., x_k)$$

  For example, when y is an employment indicator, $\mathbf{x}$ might contain various individual characteristics such as education, age, marital status, etc.

Theoretical foundation from Utility (or profit maximization)

Let y_i =Indicator of event "individual i goes to college". Let U_{0i} and U_{1i} be the utilities for individual i associated with choosing $y = 0$ (not college) and $y = 1$ (college), respectively.

Consider the following specification of these utility functions:

$$
\begin{aligned}
U_{0i} &= x_i'\beta_0 + \epsilon_{0i} \\
U_{1i} &= x_i'\beta_1 + \epsilon_{1i}
\end{aligned}
$$

where ϵ_{0i} and ϵ_{1i} represent different factors that affect the utilities of the choice alternatives 0 and 1, that are observable to the individual taking the decision, but are unobservable to the econometrician.

If the individual maximizes her utility, then:

$$
y_i = \begin{cases} 1 & \text{if } U_{1i} \geq U_{01} \iff x_i'\left(\beta_1 - \beta_0\right) - \left(\varepsilon_{0i} - \varepsilon_{1i}\right) \geq 0 \\ 0 & \text{if } U_{1i} \leq U_{01} \iff x_i'\left(\beta_1 - \beta_0\right) - \left(\varepsilon_{0i} - \varepsilon_{1i}\right) < 0 \end{cases}
$$

And therefore we have the model $y_i = \mathbb{I}\{x_i'\beta - \epsilon_i \geq 0\}$ where $\beta = \beta_1 - \beta_0$, and $\epsilon_i = \epsilon_{0i} - \epsilon_{1i}$.

Lets define $P_i(x_i) \equiv Pr(y_i = 1|x_i)$. Note that $E(y_i|x_i) = Pr(y_i = 1|x_i)$ and $Var(y_i|x_i) = Pr(y_i = 1|x_i)(1 - Pr(y_i = 1|x_i))$
Assuming ϵ_i are iid: $\Pr(y_i = 1 \mid x_i) = \Pr(\varepsilon_i \leq x_i'\beta \mid x_i) \equiv P(x_i)$
In principle, given a random sample of (y, x) for N individuals we can estimate $P(x)$ for different values of x without making any parametric assumption about the form of the function $P(\cdot)$. This is particularly simple when x is a vector of discrete random variables.
The log-likelihood function is given by

$$l(\{P(x)\}) = \sum_{i=1}^{N} 1\{x_i = x\}\left(y_i \ln[P(x)] + (1 - y_i)\ln[1 - P(x)]\right)$$

$$\widehat{P(x)} = \frac{\sum_{i=1}^{N} 1\{x_i = x\} y_i}{\sum_{i=1}^{N} 1\{x_i = x\}} = \frac{\#\{y_i = 1, x_i = x\}}{\#\{x_i = x\}}$$

- However, there are several reasons why we can be interested in the estimation of a parametric model for P(x)
  - Efficiency: more precise estimates
  - Parsimony: a few parameters can summarize the relationship between y and x.
  - Out of sample predictions
- Different parametric models have been proposed:
  - Linear probability model
  - Index models

Partial effects

- For a continuous variable, x_j, the **partial effect** of x_j on the response probability is

$$\frac{\partial p(y = 1 \,|\, x_1, x_2, .., x_k)}{\partial x_j}$$

- If x_j is a discrete variable, the partial effect is

$$p(y = 1 \,|\, x_1, ..x_{j-1}, x_j + 1, x_{j+1}.., x_k) - p(y = 1 \,|\, x_1, ..x_{j-1}, x_j, x_{j+1}.., x_k)$$

  - In particular, if x_j is a binary variable, interest lies in

$$p(y = 1 \,|\, x_1, ..x_{j-1}, 1, x_{j+1}.., x_k) - p(y = 1 \,|\, x_1, ..x_{j-1}, 0, x_{j+1}.., x_k)$$

    which is the difference in response probabilities when $x_j = 1$ and $x_j = 0$.

- For most of the models we consider the partial effect of x_j on $p(y = 1 \,|\, \mathbf{x})$ depends on $\mathbf{x}$ (both when x_j is continuous or discrete)

Bernoulli distribution

- In studying binary response models, we need to recall some basic facts about Bernoulli (zero-one) random variables.
- The only difference between the setup here and that in Statistics is the conditioning on $\mathbf{x}$.
- If we denote by $p(\mathbf{x}) = p(y = 1 \,|\, \mathbf{x})$ then

Bernoulli distribution

- In studying binary response models, we need to recall some basic facts about Bernoulli (zero-one) random variables.
- The only difference between the setup here and that in Statistics is the conditioning on $\mathbf{x}$.
- If we denote by $p(\mathbf{x}) = p(y = 1 \,|\, \mathbf{x})$ then

$$\begin{aligned}
&p(y = 0 \,|\, \mathbf{x}) = 1 - p(\mathbf{x}) \\
&E(y \,|\, \mathbf{x}) = p(\mathbf{x}) \\
&Var(y \,|\, \mathbf{x}) = p(\mathbf{x})(1 - p(\mathbf{x}))
\end{aligned} \tag{1}$$

- We are now going to consider various functional forms for $p(\mathbf{x})$.

The linear probability model

- The linear probability model (LPM) for binary response y is specified as

$$p(y = 1 \,|\, x_1, x_2, .., x_k) = \beta_0 + \beta_1 x_1 + .. + \beta_k x_k = \mathbf{x}'\boldsymbol{\beta}$$

- As usual, the x_j can be functions of underlying explanatory variables, which would simply change the interpretations of the β_j.
- Assuming that x_j is continuous and it is not functionally related to the other explanatory variables, the partial effect of x_j on the response probability is

$$\frac{\partial p(y = 1 \,|\, x_1, x_2, .., x_k)}{\partial x_j} = \beta_j$$

Therefore, β_j is the change in the probability of success given a one-unit increase in x_j, holding the other explanatory variables fixed.

The linear probability model

- If x_j is binary, β_j is just the difference in the probability of success when $x_j = 1$ and $x_j = 0$, holding the other explanatory variables fixed.
- Since y is a Bernoulli random variable, using (1) we have

$$\begin{aligned} E(y\,|\,\mathbf{x}) &= \beta_0 + \beta_1 x_1 + .. + \beta_k x_k \qquad (2) \\ Var(y\,|\,\mathbf{x}) &= \mathbf{x}'\boldsymbol{\beta}(1 - \mathbf{x}'\boldsymbol{\beta}) \qquad (3) \end{aligned}$$

- Equation (2) implies that the OLS regression of y on $1, x_1, x_2, .., x_k$ produces consistent and even unbiased estimators of the β_j
- Equation (3) means that heteroskedasticity is present unless all of the slope coefficients are zero.
  - A way to deal with this issue is to use heteroskedasticity-robust standard errors and t statistics.
  - Further, robust tests of multiple restrictions should also be used.

The linear probability model

- If x_j is binary, β_j is just the difference in the probability of success when $x_j = 1$ and $x_j = 0$, holding the other explanatory variables fixed.
- Since y is a Bernoulli random variable, using (1) we have

$$\begin{aligned} E(y\,|\,\mathbf{x}) &= \beta_0 + \beta_1 x_1 + .. + \beta_k x_k && (2) \\ Var(y\,|\,\mathbf{x}) &= \mathbf{x}'\boldsymbol{\beta}(1 - \mathbf{x}'\boldsymbol{\beta}) && (3) \end{aligned}$$

- Equation (2) implies that the OLS regression of y on $1, x_1, x_2, .., x_k$ produces consistent and even unbiased estimators of the β_j
- Equation (3) means that heteroskedasticity is present unless all of the slope coefficients are zero.
  - A way to deal with this issue is to use heteroskedasticity-robust standard errors and t statistics.
  - Further, robust tests of multiple restrictions should also be used.
    - ⋆ There is one case where the non-robust F statistic can be used, and that is to test for joint significance of all variables (leaving the constant unrestricted). This test is asymptotically valid because $Var(y\,|\,\mathbf{x})$ is constant under this particular null hypothesis.

The linear probability model

- Since the form of the variance is determined by the model for $p(y = 1 \,|\, x)$, an asymptotically more efficient method is Weighted Least Squares (WLS).

The linear probability model

- Since the form of the variance is determined by the model for $p(y = 1 \mid x)$, an asymptotically more efficient method is Weighted Least Squares (WLS).
  - Let $\widehat{\beta}$ be the OLS estimator, and let $\widehat{y}_i = x_i' \widehat{\beta}$ be the OLS fitted values.
  - Then, provided $0 < \widehat{y}_i < 1$ for all observations, define the estimated standard deviation as $\widehat{\sigma}_i = \sqrt{\widehat{y}_i(1 - \widehat{y}_i)}$.
  - Then the WLS estimator, $\widehat{\beta}_{WLS}$, is obtained from the OLS regression of $y_i/\widehat{\sigma}_i$ on $1/\widehat{\sigma}_i, x_{1i}/\widehat{\sigma}_i, .., x_{ki}/\widehat{\sigma}_i$, $i = 1, 2, .., N$.
- If some of the OLS fitted values are not between zero and one, WLS analysis is not possible without ad hoc adjustments to bring deviant fitted values into the unit interval.
- Further, since the OLS fitted value $\widehat{y}_i$ is an estimate of the conditional probability $p(y = 1 \mid \mathbf{x})$, it is somewhat awkward if the predicted probability is negative or above unity.

The linear probability model

- Aside from the issue of fitted values being outside the unit interval, the LPM implies that a ceteris paribus unit increase in x_j always changes $p(y = 1 \,|\, x_1, x_2, .., x_k)$ by the same amount, regardless of the initial value of x_j. This implication cannot literally be true because continually increasing one of the x_j would eventually drive $p(y = 1 \,|\, x_1, x_2, .., x_k)$ to be less than zero or greater than one.
- Even with these weaknesses, the LPM often seems to give good estimates of the partial effects on the response probability near the center of the distribution of $\mathbf{x}$.

- **Advantages**
  - Computational simplicity.
  - In some cases it can be a reasonable first order approximation to the marginal effect (not to β).
- **Disadvantages**
  - Given the nonlinearity nature of the problem, it is not a consistent estimator of the parameters.
  - Imposes constant Marginal Effects.
  - The disturbance terms are heteroskedastic.
  - The conditional expectation is not bounded between zero and one

Example 1: Married Women's Labor Force Participation

- We use the data from MROZ.DTA to estimate a linear probability model for labor force participation of married women.
- The variables we use to explain labor force participation (*inlf*) are age, education, experience, nonwife income in thousands of dollars (*nwifeinc*), number of children less than six years of age (*kidslt*6), and number of kids between 6 and 18 inclusive (*kidsge*6).
- The usual OLS standard errors are in parentheses, while the heteroskedasticity-robust standard errors are in brackets:

$$
\begin{aligned}
inlf &= \underset{\substack{(0.154)\\ [0.151]}}{0.586} - \underset{\substack{(0.0014)\\ [0.0015]}}{0.0034}\,nwifeinc + \underset{\substack{(0.007)\\ [0.007]}}{0.038}\,educ + \underset{\substack{(0.006)\\ [0.006]}}{0.039}\,exper \\
&\quad - \underset{\substack{(0.00018)\\ [0.00019]}}{0.00060}\,exper^2 - \underset{\substack{(0.002)\\ [0.002]}}{0.016}\,age - \underset{\substack{(0.034)\\ [0.032]}}{0.262}\,kidslt6 + \underset{\substack{(0.013)\\ [0.013]}}{0.013}\,kidsge6 \\
n &= 753, \qquad R^2 = 0.264
\end{aligned}
$$

Example 1: Married Women's Labor Force Participation

- With the exception of *kidsge*6, all coefficients have sensible signs and are statistically significant; *kidsge*6 is neither statistically significant nor practically important.
- The coefficient on *nwifeinc* means that if nonwife income increases by 10 ($10,000), the probability of being in the labor force is predicted to fall by 0.034.
  - This is a small effect given that an increase in income by $10,000 in 1975 dollars is very large in this sample (the average of *nwifeinc* is about $20,129 with standard deviation $11,635)

Example 1: Married Women's Labor Force Participation

- Having one more small child is estimated to reduce the probability of participating in the labor force by about 0.262, which is a fairly large effect.
- Of the 753 fitted probabilities, 33 are outside the unit interval.
  - Rather than using some adjustment to those 33 fitted values and applying weighted least squares, we just use OLS and report heteroskedasticity-robust standard errors.
  - Interestingly, heteroskedasticity-robust standard errors differ in practically unimportant ways from the usual OLS standard errors.

Index models

- We now study binary response models of the form

$$p(y = 1 \mid \mathbf{x}) = G(\mathbf{x}'\boldsymbol{\beta}) \tag{4}$$

  where $\mathbf{x} = (x_1, x_2, .., x_k)'$, $\boldsymbol{\beta} = (\beta_1, \beta_2, .., \beta_k)'$ and we take the first element of $\mathbf{x}$ to be unity.

- We assume that $G(\cdot)$ takes on values in the open unit interval: $0 < G(z) < 1$ for all $z \in \Re$.
- These type of models are called **index models** because $p(\mathbf{x})$ is a function of $\mathbf{x}$ only through the index $\mathbf{x}'\boldsymbol{\beta}$.
- In most applications, G is a cumulative distribution function (cdf)

Index models

- Index models where G is a cdf can be derived from an underlying latent variable model,

$$y^* = \mathbf{x}'\boldsymbol{\beta} + \varepsilon, \qquad y = 1(y^* > 0)$$

where ε is a continuously distributed random variable independent of $\mathbf{x}$ and the distribution of ε is symmetric about zero. $1(\cdot)$ is the indicator function.

Index models

- Index models where G is a cdf can be derived from an underlying latent variable model,

$$y^* = \mathbf{x}'\boldsymbol{\beta} + \varepsilon, \qquad y = 1(y^* > 0)$$

where ε is a continuously distributed random variable independent of $\mathbf{x}$ and the distribution of ε is symmetric about zero. $1(\cdot)$ is the indicator function.

- If G is the cdf of ε,

$$\begin{aligned} & p(y = 1 \mid \mathbf{x}) = p(y^* > 0 \mid \mathbf{x}) = p(\varepsilon > -\mathbf{x}'\boldsymbol{\beta} \mid \mathbf{x}) \\ & = p(\varepsilon < \mathbf{x}'\boldsymbol{\beta} \mid \mathbf{x}) = G(\mathbf{x}'\boldsymbol{\beta}) \end{aligned}$$

which is equation (4).

- Identification requires a restriction on the variance of ε, as the single-index model can only identify $\boldsymbol{\beta}$ up to scale since

$$y^* > 0 \Leftrightarrow \mathbf{x}'\boldsymbol{\beta} + \varepsilon > 0 \Leftrightarrow \mathbf{x}'\boldsymbol{\beta}^* + \varepsilon^* > 0, \text{ where } \boldsymbol{\beta}^* = \lambda\boldsymbol{\beta} \text{ and } \varepsilon^* = \lambda\varepsilon$$

Probit and Logit models

The most common choices for G are

- G is the cdf of the **standard normal distribution** ⇒ **Probit model**
  - The cdf of the standard normal distribution is denoted by Φ

    $$\Phi(z) = \int_{-\infty}^{z} \phi(v)dv$$

    where $\phi(z)$ is the pdf of the standard normal distribution

    $$\phi(z) = \frac{1}{\sqrt{2\pi}} \exp(-\frac{z^2}{2})$$

- G is the cdf of the **standard logistic distribution** ⇒ **Logit model**
  - The cdf of the standard logistic distribution is denoted by Λ

    $$\Lambda(z) = \frac{\exp(z)}{1 + \exp(z)}$$

  - The mean is zero and the variance is $\pi^2/3$.
- Notice that β is scaled differently in the two models due to different $Var(\varepsilon)$.

Partial effects

- If x_j is continuous, the partial effect of x_j on $p(\mathbf{x})$, holding all other variables fixed, is

$$\frac{\partial p(\mathbf{x})}{\partial x_j} = g(\mathbf{x}'\boldsymbol{\beta})\beta_j, \qquad \text{where } g(z) = G'(z)$$

  - The partial effect of x_j on $p(\mathbf{x})$ depends on $\mathbf{x}$ through $g(\mathbf{x}'\boldsymbol{\beta})$.
  - If $G(\cdot)$ is a strictly increasing cdf, as in the probit and logit cases, $g(z) > 0$ for all z. Therefore, the sign of the effect is given by the sign of β_j.
  - For continuous variables x_j and x_h, the relative effects do not depend on $\mathbf{x}$

$$\frac{\frac{\partial p(\mathbf{x})}{\partial x_j}}{\frac{\partial p(\mathbf{x})}{\partial x_h}} = \frac{\beta_j}{\beta_h}$$

  - For the probit and logit models the partial effect of x_j on $p(\mathbf{x})$ is

$$\begin{array}{ll} \text{Probit model} & \frac{\partial p(\mathbf{x})}{\partial x_j} = \phi(\mathbf{x}'\boldsymbol{\beta})\beta_j \\ \text{Logit model} & \frac{\partial p(\mathbf{x})}{\partial x_j} = \Lambda(\mathbf{x}'\boldsymbol{\beta})(1-\Lambda(\mathbf{x}'\boldsymbol{\beta}))\beta_j \end{array} \tag{5}$$

    Notice that $\Lambda(\mathbf{x}'\boldsymbol{\beta})(1-\Lambda(\mathbf{x}'\boldsymbol{\beta}))$ is the pdf of the logistic distribution.

Partial effects

- If x_k is discrete, the partial effect of x_k on $p(\mathbf{x})$, holding all other variables fixed, is

$$G(\beta_1 x_1 + .. + \beta_{k-1} x_{k-1} + \beta_k (x_k + 1)) - G(\beta_1 x_1 + .. + \beta_{k-1} x_{k-1} + \beta_k x_k)$$

  - The partial effect depends on $\mathbf{x}$.
  - If $G(\cdot)$ is a strictly increasing cdf, as in the probit and logit cases, the sign of the effect is given by the sign of β_k.

- In particular, if x_k is binary, the partial effect from changing x_k from zero to one, holding all other variables fixed, is

$$G(\beta_1 x_1 + .. + \beta_{k-1} x_{k-1} + \beta_k) - G(\beta_1 x_1 + .. + \beta_{k-1} x_{k-1})$$

  - This expression depends on the values of the other x_j.

Partial effects

- It is straightforward to include standard functional forms among the explanatory variables.
- For example, in the model

$$P(y=1\,|\,\mathbf{z}) = G(\beta_0 + \beta_1 z_1 + \beta_2 z_1^2 + \beta_3 \log(z_2))$$

  - The partial effect of z_1 on $P(y=1\,|\,\mathbf{z})$ is

$$\frac{\partial P(y=1\,|\,\mathbf{z})}{\partial z_1} = g(\mathbf{x}'\boldsymbol{\beta})(\beta_1 + 2\beta_2 z_1)$$

    where $x = (1, z_1, z_1^2, \log(z_2))'$. It follows that if the quadratic in z_1 has a hump shape or a U shape, the turning point in the response probability is $-\beta_1/2\beta_2$.
  - The partial effect of $\log(z_2)$ on $P(y=1\,|\,\mathbf{z})$ is

$$\frac{\partial P(y=1\,|\,\mathbf{z})}{\partial \log(z_2)} = g(\mathbf{x}'\boldsymbol{\beta})\beta_3$$

    and so $g(\mathbf{x}'\boldsymbol{\beta})\beta_3/100$ is the approximate change in $P(y=1\,|\,\mathbf{z})$ given a 1 percent increase in z_2, holding z_1 fixed.

Maximum likelihood estimation

- Assume we have N *iid* observations following model (4). To estimate the model by (conditional) maximum likelihood, we need the log-likelihood function of the sample.
- Since the distribution of y_i given $\mathbf{x}_i$ is Bernouilli and $p(\mathbf{x}) = G(\mathbf{x}_i'\beta)$, the probability mass function of y_i given $\mathbf{x}_i$ is

$$f(y_i \mid \mathbf{x}_i) = G(\mathbf{x}_i'\beta)^{y_i}(1 - G(\mathbf{x}_i'\beta))^{1-y_i}$$

- Then, the likelihood function is given by

$$L(\beta; y_1, ..y_N \mid \mathbf{x}_1, .., \mathbf{x}_N) = \prod_{i=1}^{N} f(y_i \mid \mathbf{x}_i) = \prod_{i=1}^{N} G(\mathbf{x}_i'\beta)^{y_i}(1-G(\mathbf{x}_i'\beta))^{1-y_i}$$

and the log-likelihood function is

$$\log L(\beta; y_1, ..y_N \mid \mathbf{x}_1, .., \mathbf{x}_N) = \sum_{i=1}^{N} \log f(y_i \mid \mathbf{x}_i)$$
$$= \sum_{i=1}^{N} [y_i \log G(\mathbf{x}_i'\beta) + (1 - y_i) \log(1 - G(\mathbf{x}_i'\beta))]$$

Maximum likelihood estimation

- The Maximum Likelihood Estimator (MLE) of $\boldsymbol{\beta}$, denoted $\widehat{\boldsymbol{\beta}}$, maximizes this log likelihood. If $G(\cdot)$ is the standard normal cdf, then $\widehat{\boldsymbol{\beta}}$ is the probit estimator; if $G(\cdot)$ is the logistic cdf, then $\widehat{\boldsymbol{\beta}}$ is the logit estimator.
- For most choices of $G(\cdot)$, there is no explicit solution for the MLE but the estimator can be found by numerical methods. This is the case of the probit and logit estimators.
- From the general maximum likelihood results, the MLE is consistent and asymptotically normal if the conditional density of y given $\mathbf{x}$ is correctly specified. Since the density here must be Bernoulli, the only possible misspecification is that the Bernoulli probability is misspecified. So the MLE is consistent if $p(y = 1 \,|\, \mathbf{x}) = G(\mathbf{x}'\boldsymbol{\beta})$.
- Any of the three tests from general MLE analysis—the Wald, LR, or LM test—can be used to test hypotheses in binary response contexts.

Estimated partial effects

- In index models, the $\widehat{\beta}_j$ give the signs of the partial effects of each x_j on the response probability, and the statistical significance of x_j is determined by whether we can reject $H_0 : \beta_j = 0$.
- Often we want to estimate the effects of the variables x_j on the response probabilities $P(y = 1 \,|\, \mathbf{x})$. If x_j is (roughly) continuous then the estimated partial effect of x_j on the response probabilities, evaluated at $\mathbf{x}$, is
$$\frac{\partial P(\widehat{y = 1 \,|\, \mathbf{x}})}{\partial x_j} = g(\mathbf{x}'\widehat{\boldsymbol{\beta}})\widehat{\beta}_j \tag{6}$$
- Since $g(\mathbf{x}'\widehat{\boldsymbol{\beta}})$ depends on $\mathbf{x}$, we must compute $g(\mathbf{x}'\widehat{\boldsymbol{\beta}})$ at interesting values of $\mathbf{x}$. Often the sample averages of the x_j's are plugged in to get $g(\overline{\mathbf{x}}'\widehat{\boldsymbol{\beta}})$, with $\overline{x}_1 \equiv 1$ because we include a constant. We call the resulting partial effect the **partial effect at the average**.

Estimated partial effects

- If $\mathbf{x}$ contains nonlinear functions of some explanatory variables, such as natural logs, there is the issue of using the log of the average versus the average of the log.
    - To get the effect for the "average" person, it makes more sense to plug the averages into the nonlinear functions, rather than average the nonlinear functions.
- If two or more elements of $\mathbf{x}$ are functionally related, such as quadratics or interactions, the estimated partial effect in (6) has, in general, no meaning. For instance, suppose that $x_{k-1} = age$ and $x_k = age^2$. Then, the estimated partial effect of *age* on the response probabilities, evaluated at $\mathbf{x}$, is not $g(\mathbf{x}'\widehat{\boldsymbol{\beta}})\widehat{\boldsymbol{\beta}}_{k-1}$ but

$$\frac{\partial P(\widehat{y=1}\,|\,\mathbf{x})}{\partial age} = g(\mathbf{x}'\widehat{\boldsymbol{\beta}})\left(\widehat{\beta}_{k-1} + 2\widehat{\beta}_k age\right)$$

Now we might be interested in evaluating this partial effect at the mean values, but that would entail using $\overline{age}^2$ inside $g(\cdot)$.

Estimated partial effects

- An alternative way to summarize the estimated marginal effects is to estimate the **average partial effect (APE)**.
  - If x_j is a continuous variable the average partial effect is

$$E\left[\frac{\partial P(y=1\,|\,\mathbf{x})}{\partial x_j}\right] = E\left[g(\mathbf{x}'\boldsymbol{\beta})\right]\beta_j$$

    and a consistent estimator of the average partial effect is

$$\frac{1}{N}\sum_{i=1}^{N} g(\mathbf{x}_i'\widehat{\boldsymbol{\beta}})\widehat{\beta}_j$$

  - If x_k is binary, the average partial effect is

$$E\left[G(\,\beta_1 x_1 + .. + \beta_{k-1}x_{k-1} + \beta_k) - G(\beta_1 x_1 + .. + \beta_{k-1}x_{k-1})\right]$$

    and a consistent estimator of the average partial effect is

$$\frac{1}{N}\sum_{i=1}^{N} G(\,\widehat{\beta}_1 x_{1i} + .. + \widehat{\beta}_{k-1}x_{k-1,i} + \widehat{\beta}_k) - \frac{1}{N}\sum_{i=1}^{N} G(\widehat{\beta}_1 x_{1i} + .. + \widehat{\beta}_{k-1}x_{k-1,i})$$

- The delta method can be used to obtain standard errors of estimated partial effects.

Goodness of fit measures

- One measure of goodness of fit that is usually reported is the **percent correctly predicted**.
    - For each i, we compute the predicted probability that $y_i = 1$, given the explanatory variables, $\mathbf{x}_i$.
    - If $G(\mathbf{x}_i'\widehat{\beta}) > 0.5$, we predict y_i to be unity; if $G(\mathbf{x}_i'\widehat{\beta}) \leq 0.5$, y_i is predicted to be zero.
    - The percentage of times the predicted y_i matches the actual y_i is the percent correctly predicted.
    - In many cases it is easy to predict one of the outcomes and much harder to predict another outcome, in which case the percent correctly predicted can be misleading as a goodness-of-fit statistic. More informative is to compute the percent correctly predicted for each outcome, $y_i = 0$ and $y_i = 1$.

Goodness of fit measures

- Various pseudo R-squared measures have been proposed for binary response. McFadden (1974) suggests the measure $1 - \log L_{ur} / \log L_0$, where $\log L_{ur}$ is the log-likelihood function for the estimated model and $\log L_0$ is the log-likelihood function in the model with only an intercept. Because the log likelihood for a binary response model is always negative, $|\log L_{ur}| < |\log L_0|$, and so the pseudo R-squared is always between zero and one.
- Several other measures have been suggested, but goodness of fit is not as important as statistical and economic significance of the explanatory variables.

Comparing probit, logit and LPM estimation results

- In general, the results based on these models are similar.
  - It is rare to find, say, logit and probit estimates having different signs unless the coefficients are estimated imprecisely.
  - There is often little difference between the predicted probabilities from these models. The difference is greater in the tails where probabilities are closed to 0 or 1.
  - The difference is also small if interest lies APE
- The different models do yield different estimates of the $\widehat{\beta}_j$ but this is due to the fact that logit, probit and LPM use different scale factors. If we compare the partial effects in (5) we have that
  - Since $\phi(z)$ achieve the maximum at $z = 0$ and $\phi(0) = \frac{1}{\sqrt{2\pi}} \simeq 0.4$, for the probit model $\frac{\partial \widehat{P(y=1 \mid \mathbf{x})}}{\partial x_j} \leq 0.4\widehat{\beta}_j$
  - Since $\Lambda(z)(1-\Lambda(z))$ achieve the maximum at $z = 0$ and $\Lambda(0)(1-\Lambda(0)) = 0.25$, for the logit model $\frac{\partial \widehat{P(y=1 \mid \mathbf{x})}}{\partial x_j} \leq 0.25\widehat{\beta}_j$
  - For the LPM $\frac{\partial \widehat{P(y=1 \mid \mathbf{x})}}{\partial x_j} = \widehat{\beta}_j$

Comparing probit, logit and LPM estimation results

- This bounds suggest the following rule of thumb to compare the estimated coefficients of these three models

$$\widehat{\beta}_{Logit} \simeq \frac{1}{0.25}\widehat{\beta}_{LPM} = 4\widehat{\beta}_{LPM}$$
$$\widehat{\beta}_{Probit} \simeq \frac{1}{0.4}\widehat{\beta}_{LPM} = 2.5\widehat{\beta}_{LPM}$$
$$\widehat{\beta}_{Logit} \simeq \frac{0.4}{0.25}\widehat{\beta}_{Probit} = 1.6\widehat{\beta}_{Probit}$$

Example 1: Married Women's Labor Force Participation

LPM, logit and probit estimates of female labor force participation (Mroz)

Dependent variable: *inlf*			
Explanatory variables	LPM	Logit	Probit
nwifeinc	−0.0034 (0.0015)	−0.021 (0.008)	−0.012 (0.005)
educ	0.038 (0.007)	0.221 (0.043)	0.131 (0.025)
exper	0.039 (0.006)	0.206 (0.032)	0.123 (0.019)
$exper^2$	−0.00060 (0.00019)	−0.0032 (0.0010)	−0.0019 (0.0006)
age	−0.016 (0.002)	−0.088 (0.015)	−0.053 (0.008)
*kidslt*6	−0.262 (0.032)	−1.443 (0.204)	−0.868 (0.119)
*kidsge*6	0.013 (0.013)	0.060 (0.075)	0.036 (0.043)
Observations	753	753	753
Percent correctly predicted	73.4	73.6	73.4
Log-likelihood		−401.77	−401.30
Pseudo R^2	0.264	0.220	0.221

Example 1: Married Women's Labor Force Participation

- The estimates from the three models tell a consistent story. The signs of the coefficients are the same across models, and the same variables are statistically significant in each model.
- The pseudo R-squared for the LPM is just the usual R-squared reported for OLS; for logit and probit the pseudo R-squared is the measure based on the log likelihoods described previously.
- In terms of overall percent correctly predicted, the models do equally well.
- They also do equally well in terms of the percent of 0 and 1 correctly predicted

	0	1
LPM	62.46	81.78
Probit	63.08	81.31
Logit	63.69	81.07

Example 1: Married Women's Labor Force Participation

- As we emphasized earlier, the magnitudes of the coefficients are not directly comparable across the models. Using the rough rule of thumb discussed earlier, we can divide the logit estimates by four and the probit estimates by 2.5 to make all estimates comparable to the LPM estimates. For example, for the coefficients on *kidslt*6, the scaled logit estimate is about 0.361, and the scaled probit estimate is about 0.347. These are larger in magnitude than the LPM estimate.

Example 1: Married Women's Labor Force Participation

- The biggest difference between the LPM model on one hand, and the logit and probit models on the other, is that the LPM assumes constant marginal effects, while the logit and probit models imply diminishing marginal magnitudes of the partial effects.
- For a women with average values for the covariates
  - the estimated fall in the probability of working in going from zero to one small child is 0.334 probability points using the probit estimates (0.344 using logit)
  - and in going from one to two young children, the fall is 0.255 probability points using the probit estimates (0.245 using logit)
- Using the LPM estimates, one more small child is estimated to reduce the probability of labor force participation by about 0.262 probability points, regardless of how many young children the woman already has (and regardless of the levels of the other covariates).

Example 1: Married Women's Labor Force Participation

- We now compare the estimated partial effect from the LPM for a roughly continuous variable like education, with the estimated average partial effect from the probit and logit models.
  - The estimated partial effect of education is 0.0380 in the LPM. In the probit model, the estimated average partial effect is 0.0394, and in the logit model is 0.0395. As expected, these partial effects are very similar. One additional year of education increases the probability of working by about 0.04 probability points.

Multinomial response models

- We will now consider discrete response models with more than two outcomes.
- In this section we consider **unordered responses** where the values attached to different outcomes are arbitrary and have no effect on estimation, inference or interpretation.
- Examples: occupational choice, transportation mode for commuting to work, etc.
  - Suppose there are three type of transportation: bus, train and private car. It doesn't matter whether we can label them as 0, 1 and 2 or as 100, 500, 1000. It doesn't matter either which type of transportation we assign to each number.

Multinomial logit

- This model applies when a unit's response or choice depends on individual characteristics of the unit but not on attributes of the choices.
- Let y denote a random variable taking on the values $\{0, 1, .., J\}$ for J a positive integer, and let $\mathbf{x}$ denote a set of conditioning variables.
  - For example, if y denotes occupational choice, $\mathbf{x}$ can contain things like education, age, gender, race, marital status, etc.
- As usual, $(\mathbf{x}_i', y_i)'$ is a random draw from the population.
- As in the binary response case, we are interested in how ceteris paribus changes in the elements of $\mathbf{x}$ affect the response probabilities, $P(y = j \,|\, \mathbf{x})$, $j = 0, 1, .., J$.
- Since the probabilities must sum to unity, $P(y = 0 \,|\, \mathbf{x})$ is determined once we know the probabilities for $j = 1, .., J$.

Multinomial logit

- Let $\mathbf{x}$ be a $k \times 1$ vector with first-element unity. The multinomial logit (MNL) model has response probabilities

$$P(y = j \,|\, \mathbf{x}) = \frac{\exp(\mathbf{x}'\boldsymbol{\beta}_j)}{1 + \sum_{h=1}^{J} \exp(\mathbf{x}'\boldsymbol{\beta}_h)} \tag{7}$$

where $\boldsymbol{\beta}_j$, $j = 1, .., J$, are $k \times 1$ vectors of unknown parameters

- Because the response probabilities must sum to unity,

$$P(y = 0 \,|\, \mathbf{x}) = \frac{1}{1 + \sum_{h=1}^{J} \exp(\mathbf{x}'\boldsymbol{\beta}_h)}$$

- When $J = 1$ we get the binary logit model.

Multinomial logit

- The partial effects for this model are complicated. For continuous x_l, we can write

$$\frac{\partial P(y=j\,|\,\mathbf{x})}{\partial x_l} = P(y=j\,|\,\mathbf{x}) \left(\beta_{jl} - \frac{\sum_{h=1}^{J} \beta_{hl} \exp(\mathbf{x}'\boldsymbol{\beta}_h)}{1+\sum_{h=1}^{J} \exp(\mathbf{x}'\boldsymbol{\beta}_h)} \right) \tag{8}$$

  where β_{hl} is the l-th element of $\boldsymbol{\beta}_h$, $h = 1, .., J$.
- This equation shows that even the direction of the effect is not determined entirely by β_{jl}.

Multinomial logit

- Since we have fully specified the distribution of y given $\mathbf{x}$, estimation of the MNL model is best carried out by maximum likelihood. For each i the conditional log-likelihood can be written as

$$\log L(\beta; y_1, ..y_N \,|\, \mathbf{x}_1, .., \mathbf{x}_N) = \sum_{i=1}^{N} \sum_{j=0}^{J} 1(y_i = j) \log P(y = j \,|\, \mathbf{x})$$

- There is no explicit solution for the MNL estimator but it can be found by numerical methods.
- The log-likelihood function is globally concave, and this fact makes the maximization problem straightforward.
- The conditions required for consistency and asymptotic normality are broadly applicable.

Example 2: School and Employment Decisions for Young Men

- The data KEANE.DTA (a subset from Keane and Wolpin, 1997) contains employment and schooling history for a sample of men for the years 1981 to 1987. We use the data for 1987.
- The three possible outcomes are enrolled in school ($\mathit{status} = 1$), not in school and not working ($\mathit{status} = 2$), and working ($\mathit{status} = 3$). The base category is enrolled in school.
- The explanatory variables are education, a quadratic in past work experience, and a black binary indicator.
- Out of 1,717 observations, 99 are enrolled in school, 332 are at home, and 1,286 are working. The results are given below.

Example 2: School and Employment Decisions for Young Men

Multinomial Logit Estimates of School and Labor Market Decisions

	Dependent variable: *status*	
Explanatory variables	Home ($status = 2$)	Work ($status = 3$)
educ	-0.674 (0.070)	-0.315 (0.065)
exper	-0.106 (0.173)	0.849 (0.157)
exper^2	-0.013 (0.025)	-0.077 (0.023)
black	0.813 (0.303)	0.311 (0.282)
constant	10.28 (1.13)	5.54 (1.09)
Observations	1717	
Percent correctly predicted	79.6	
Log-likelihood	-907.86	
Pseudo R^2	0.243	

Example 2: School and Employment Decisions for Young Men

- The magnitudes of the coefficients are difficult to interpret. Instead, we can either compute partial effects, using equation (8), or compute differences in probabilities.
  - For example, consider two black men, each with five years of experience. A black man with 16 years of education has an employment probability that is 0.042 higher than a man with 12 years of education, and the at-home probability is 0.072 lower. These results are easily obtained by comparing fitted probabilities after multinomial logit estimation.

Example 2: School and Employment Decisions for Young Men

- The experience terms are each insignificant in the home column, but the Wald test for joint significance of *exper* and *exper*2 gives p-value=0.047, and so they are jointly significant at the 5 percent level.
- The fitted probabilities can be used for prediction purposes: for each observation *i*, the outcome with the highest estimated probability is the predicted outcome.
  - This can be used to obtain a percent correctly predicted, by category if desired.
  - For the example, the overall percent correctly predicted is almost 80%, but the model does a much better job of predicting that a man is employed (95.2% correct) than in school (12.1%) or at home (39.2%).

Probabilistic choice models

- Suppose that, for a random draw i from the underlying population (usually, but not necessarily, individuals), the utility from choosing alternative j is

  $$y_{ij}^{*} = \mathbf{x}_{ij}'\boldsymbol{\beta} + a_{ij}, \quad j = 0, 1, .., J$$

  where $a_{ij}, j = 0, 1, .., J$, are unobservables affecting tastes.
- Here, $\mathbf{x}_{ij}$ is a $k \times 1$ vector that differs across alternatives and possibly across individuals as well. For example, $\mathbf{x}_{ij}$ might contain the commute time for individual i using transportation mode j, or the co-payment required by health insurance plan j (which may or may not differ by individual).
- For reasons we will see, $\mathbf{x}_{ij}$ cannot contain elements that vary only across i and not j; in particular, $\mathbf{x}_{ij}$ does not contain unity.
- We assume that the $(J+1)$-vector $\mathbf{a}_i = (a_{i0}, a_{i1}, .., a_{iJ})'$ is independent of $\mathbf{x}_i = (\mathbf{x}_{i0}', \mathbf{x}_{i1}', .., \mathbf{x}_{iJ}')'$.

Probabilistic choice models

- Let y_i denote the alternative chosen by individual i to maximize utility

  $$\max\{y^*_{i0}, y^*_{i1}, .., y^*_{iJ}\}$$

  so that y_i takes on a value in $\{j = 0, 1, .., J\}$.

- As shown by McFadden (1974), if the a_{ij}, $j = 0, 1, .., J$ are independently distributed with cdf $F(a) = \exp(-\exp(-a))$, the type I extreme value distribution, then

  $$p_{ji}(\mathbf{x}_i) = P(y_i = j \,|\, \mathbf{x}_i) = \frac{\exp(\mathbf{x}'_{ij}\boldsymbol{\beta})}{\sum_{h=0}^{J} \exp(\mathbf{x}'_{ih}\boldsymbol{\beta})}$$

  These response probabilities constitute what is usually called the **conditional logit (CL) model**.

Probabilistic choice models

- Dropping the subscript i and differentiating shows that the marginal effects are given by

$$\frac{\partial p_j(\mathbf{x})}{\partial x_{jl}} = p_j(\mathbf{x})(1 - p_j(\mathbf{x}))\beta_l, \quad j = 0, 1, .., J \text{ and } l = 1, 2, .., k$$

$$\frac{\partial p_j(\mathbf{x})}{\partial x_{hl}} = -p_j(\mathbf{x})p_h(\mathbf{x})\beta_l, \quad h \neq j, \quad j = 0, 1, .., J, \text{ and } l = 1, 2, .., k$$

  where β_l is the l-th element of vector $\boldsymbol{\beta}$.

- As usual, if the $\mathbf{x}_j$ contain nonlinear functions of underlying explanatory variables, this fact will be reflected in the partial derivatives.
- The conditional logit and multinomial logit models have similar response probabilities, but they differ in some important respects.

Probabilistic choice models

- In the MNL model:
  - The conditioning variables do not change across alternatives: for each i, $\mathbf{x}_i$ contains variables specific to the individual but not to the alternatives.
  - The MNL model allows the individual characteristics to have different effects on the relative probabilities between any two choices.
  - This model is appropriate for problems where characteristics of the alternatives are unimportant or are not of interest, or where the data are simply not available.
  - For example, in a model of occupational choice, we do not usually know how much someone could make in every occupation. What we can usually collect data on are things that affect individual productivity and tastes, such as education and past experience.

Probabilistic choice models

- The CL model:
  - This model is intended specifically for problems where consumer or firm choices are at least partly made based on observable attributes of each alternative.
  - The utility level of each choice is assumed to be a linear function in choice attributes, $\mathbf{x}_{ij}$, with common parameter vector $\boldsymbol{\beta}$.

Probabilistic choice models

- The CL model actually contains the MNL model as a special case by appropriately choosing $\mathbf{x}_{ij}$.
    - Suppose $\mathbf{w}_i$ is a vector of individual characteristics and $P(y_i = j \mid \mathbf{w}_i)$ follows the MNL in equation (7) with parameters $\boldsymbol{\delta}_j$, $j = 1, 2, .., J$.
    - We can cast this model as the conditional logit model by defining $\mathbf{x}_{ij} = (d1_j\mathbf{w}_i, d2_j\mathbf{w}_i, .., dJ_J\mathbf{w}_i)$, where dh_j is a dummy variable equal to unity when $j = h$, and $\boldsymbol{\beta} = (\boldsymbol{\delta}_1', \boldsymbol{\delta}_2', .., \boldsymbol{\delta}_J')'$.
    - Consequently, some authors refer to the conditional logit model as the multinomial logit model, with the understanding that alternative specific characteristics are allowed in the response probability.
- Empirical applications of the conditional logit model often include individual specific variables by allowing them to have separate effects on the latent utilities. A general model is

$$y_{ij}^* = \mathbf{z}_{ij}'\gamma + \mathbf{w}_i'\boldsymbol{\delta}_j + a_{ij}, \quad j = 0, 1, .., J$$

with $\delta_0 = 0$ as a normalization, where $\mathbf{z}_{ij}$ varies across j and possibly i. This model is called the **mixed logit model**.

Independence from irrelevant alternatives

- The conditional logit model is very convenient for modeling probabilistic choice, but it has some limitations.
- An important restriction is

$$\frac{p_j(\mathbf{x})}{p_h(\mathbf{x})} = \frac{\exp(\mathbf{x}_j'\boldsymbol{\beta})}{\exp(\mathbf{x}_h'\boldsymbol{\beta})} = \exp\left((\mathbf{x}_j - \mathbf{x}_h)'\boldsymbol{\beta}\right) \tag{9}$$

  so that relative probabilities for any two alternatives do not depend on the attributes of the other alternatives.
- This is called the **independence from irrelevant alternatives (IIA) assumption** because it implies that adding another alternative or changing the characteristics of a third alternative does not affect the relative odds between alternatives *j* and *h*.

Independence from irrelevant alternatives

- This implication is implausible when alternatives are similar.
- A well-known example due to McFadden (1974) is the following:
  - Consider commuters initially choosing between two modes of transportation, car and red bus, with equal probability, 0.5, so that the ratio in equation (9) is unity.
  - Now suppose a third mode, blue bus, is added. If bus commuters do not care about the color of the bus, consumers will choose between these with equal probability, so that $P_{\text{Red bus}}/P_{\text{Blue bus}} = 1$.
  - The IIA implies that $P_{\text{Red bus}}/P_{\text{Car}}$ does not change when the blue bus is introduced.
  - Then, since $P_{\text{Red bus}}/P_{\text{Blue bus}} = 1$ and $P_{\text{Red bus}}/P_{\text{Car}} = 1$, the probability of each mode is $1/3$, and therefore, the fraction of commuters taking a car would fall from $1/2$ to $1/3$ when the blue bus is introduced, a result that is not very realistic.
- This example is admittedly extreme—in practice, we would put the blue and red buses into the same category, provided there are no other differences—but it indicates that the IIA property can impose unwanted restrictions in the CL model.

Independence from irrelevant alternatives

- Some models that relax the IIA assumption have been suggested.
- In the context of the random utility model the IIA assumption comes about because the $\{a_{ij},\ j = 0, 1, .., J\}$ are assumed to be independent and follow a type I extreme value distribution.
- A more flexible assumption is that $\mathbf{a}_i = (a_{i0}, a_{i1}, a_{i2}, .., a_{iJ})$ has a multivariate normal distribution with arbitrary correlations between a_{ij} and a_{ih}, all $j \neq h$. The resulting model is called the **multinomial probit model**.

Ordered response models

- Another kind of multinomial response is an ordered response. As the name suggests, if y is an ordered response, then the values we assign to each outcome are no longer arbitrary.
- For example, y might be a credit rating on a scale from zero to six, with $y = 6$ representing the highest rating and $y = 0$ the lowest rating.
    - The fact that six is a better rating than five conveys useful information, even though the credit rating itself only has ordinal meaning.
    - For example, we cannot say that the difference between four and two is somehow twice as important as the difference between one and zero.

Ordered probit and ordered logit

- Let y be an ordered response taking on the values $\{0, 1, 2, .., J\}$ for some known integer J.
- The **ordered probit** model for y (conditional on explanatory variables $\mathbf{x}$) can be derived from a latent variable model.
- Assume that a latent variable y^* is determined by

$$y^* = \mathbf{x}'\boldsymbol{\beta} + \varepsilon, \quad \varepsilon \,|\, \mathbf{x} \sim N(0, 1)$$

  where $\mathbf{x}$ does not contain a constant
- Let $\alpha_1 < \alpha_2 < .. < \alpha_J$ be unknown cut points and define

$$\begin{aligned} &y = 0 \text{ if } y^* \leq \alpha_1 \\ &y = 1 \text{ if } \alpha_1 < y^* \leq \alpha_2 \\ &\vdots \\ &y = J \text{ if } y^* > \alpha_J \end{aligned}$$

Ordered probit and ordered logit

- Given the standard normal assumption for ε, it is straightforward to derive the conditional distribution of y given $\mathbf{x}$; we simply compute each response probability:

$$
\begin{array}{l}
P(y = 0 \,|\, \mathbf{x}) = P(y^* \leq \alpha_1 \,|\, \mathbf{x}) = P(\varepsilon \leq \alpha_1 - \mathbf{x}'\boldsymbol{\beta} \,|\, \mathbf{x}) = \Phi(\alpha_1 - \mathbf{x}'\boldsymbol{\beta}) \\
P(y = 1 \,|\, \mathbf{x}) = P(\alpha_1 < y^* \leq \alpha_2 \,|\, \mathbf{x}) = \Phi(\alpha_2 - \mathbf{x}'\boldsymbol{\beta}) - \Phi(\alpha_1 - \mathbf{x}'\boldsymbol{\beta}) \\
\vdots \\
P(y = J - 1 \,|\, \mathbf{x}) = P(\alpha_{J-1} < y^* \leq \alpha_J \,|\, \mathbf{x}) = \Phi(\alpha_J - \mathbf{x}'\boldsymbol{\beta}) - \Phi(\alpha_{J-1}(- \\
P(y = J \,|\, \mathbf{x}) = P(y)^* > \alpha_J \,|\, \mathbf{x}) = 1 - \Phi(\alpha_J - \mathbf{x}'\boldsymbol{\beta})
\end{array}
$$

- When $J = 1$ we get the binary probit model: $P(y = 1 \,|\, \mathbf{x}) = 1 - P(y = 0 \,|\, \mathbf{x}) = 1 - \Phi(\alpha_1 - \mathbf{x}'\boldsymbol{\beta}) = \Phi(\mathbf{x}'\boldsymbol{\beta} - \alpha_1)$, and so $-\alpha_1$ is the intercept inside Φ.
  - When there are only two outcomes, zero and one, we set the single cut point to zero and estimate the intercept; this approach leads to the standard probit model.

Ordered probit and ordered logit

- The parameters $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ can be estimated by maximum likelihood. The log-likelihood function is:

$$\log L(\boldsymbol{\beta}, \boldsymbol{\alpha}; y_1, ..y_N \,|\, \mathbf{x}_1, .., \mathbf{x}_N) = \sum_{i=1}^{N} \{1(y_i = 0) \log \Phi(\alpha_1 - \mathbf{x}_i' \boldsymbol{\beta})$$
$$+1(y_i = 1) \log (\Phi(\alpha_2 - \mathbf{x}' \boldsymbol{\beta}) - \Phi(\alpha_1 - \mathbf{x}' \boldsymbol{\beta}))$$
$$+ ... + 1(y_i = J) \log (1 - \Phi(\alpha_J - \mathbf{x}' \boldsymbol{\beta})) \}$$

- Other distribution functions can be used in place of Φ. Replacing Φ with the standard logistic distribution, Λ, gives the **ordered logit model**.
- Both in ordered logit and ordered probit $\boldsymbol{\beta}$, by itself, is of limited interest. In most cases, we are interested in the response probabilities $P(y = j \,|\, \mathbf{x})$, just as in the previous models.

Ordered probit and ordered logit

- For the ordered probit model

$$
\begin{aligned}
\frac{\partial P(y=0\,|\,\mathbf{x})}{\partial x_l} &= -\phi(\alpha_1 - \mathbf{x}'\boldsymbol{\beta})\beta_l \\
\frac{\partial P(y=j\,|\,\mathbf{x})}{\partial x_l} &= \left(\phi(\alpha_j - \mathbf{x}'\boldsymbol{\beta}) - \phi(\alpha_{j+1} - x'\boldsymbol{\beta})\right)\beta_l \quad ,j = 1,..J-1 \\
\frac{\partial P(y=J\,|\,\mathbf{x})}{\partial x_l} &= \phi(\alpha_J - \mathbf{x}'\boldsymbol{\beta})\beta_l
\end{aligned}
$$

  and the formulas for the ordered logit model are similar.

- In making comparisons across different models—in particular, comparing ordered probit and ordered logit— we must remember to compare estimated response probabilities at various values of **x**, such as $\bar{\mathbf{x}}$; the $\boldsymbol{\beta}$ are not directly comparable.

Ordered probit and ordered logit

- While the direction of the effect of x_l on the probabilities $P(y = 0 \,|\, \mathbf{x})$ and $P(y = J \,|\, \mathbf{x})$ is unambiguously determined by the sign of β_l, the sign of β_l does not always determine the direction of the effect for the intermediate outcomes, $1, 2, .., J - 1$.
- To see this point, suppose there are three possible outcomes, 0, 1, and 2, and that $\beta_l > 0$. Then $\frac{\partial P(y=0 \,|\, \mathbf{x})}{\partial x_l} < 0$ and $\frac{\partial P(y=2 \,|\, \mathbf{x})}{\partial x_l} > 0$, but $\frac{\partial P(y=1 \,|\, \mathbf{x})}{\partial x_l}$ could be either sign. If $|\alpha_1 - \mathbf{x}'\boldsymbol{\beta}| < |\alpha_2 - \mathbf{x}'\boldsymbol{\beta}|$, the scale factor, $\phi(\alpha_j - \mathbf{x}'\boldsymbol{\beta}) - \phi(\alpha_{j+1} - \mathbf{x}'\boldsymbol{\beta})$, is positive; otherwise it is negative. (This conclusion follows because the standard normal pdf is symmetric about zero, reaches its maximum at zero, and declines monotonically as its argument increases in absolute value.)
- As with multinomial logit, for ordered responses we can compute the percent correctly predicted, for each outcome as well as overall: our prediction for y is simply the outcome with the highest probability.

Example 3: Attitude towards gays and lesbians

- We are interested in studying the effect of some characteristics of the people on their attitude towards gays and lesbians.
- The European Social Survey includes the following question: Do you agree with the following sentence "gays and lesbians should be free to live life as they wish"?
- The answers have been categorized in four:
  - $freehomosex = 0$ if the person "disagree"
  - $freehomosex = 1$ if the person "neither agree nor disagree"
  - $freehomosex = 2$ if the person "agree"
  - $freehomosex = 3$ if the person "strongly agree".
- The data set ESS2010.DTA has information for a subsample of people from the European Social Survey 2010.

Example 3: Attitude towards gays and lesbians

Order probit estimates

	Dependent variable: *freehomosex*
eduyrs	0.0256 (0.0057)
age	0.0010 (0.0075)
agesq	−000168 (0.000076)
female	0.3381 (0.0542)
somerelig	−0.3793 (0.0666)
quiterelig	−0.5476 (0.0849)
veryrelig	−0.9734 (0.1212)
/*cut*1	−1.8566 (0.1988)
/*cut*2	−1.2122 (0.1967)
/*cut*3	0.0418 (0.1953)
Observations	1810

Example 3: Attitude towards gays and lesbians

- For a 45 years old male, with 12 years of education, who is very religious, an increase of one year in age increases the probability of "disagree" by 0.38 percentage points and the probability of "neither agree nor disagree" by 0.17 percentage points and decreases the probability of "agree" by 0.20 percentage points and the probability of "strongly agree" by 0.34 percentage points.
- On average, the probability of "disagree" is 4.0 percentage points lower for females than for males, the probability of "neither agree nor disagree" is 3.8 percentage points lower for females than for males, the probability of "agree" is 3.9 percentage points lower for females than for males, and the probability of "strongly agree" is 11.8 percentage points larger for females than for males.