QUALITATIVE RESPONSE REGRESSION MODELS

- Regression models involving nominal scale dependent variables are among a broader class of models known as qualitative response regression models.
- Today we will consider the simplest of such models, the binary or dichotomous or dummy dependent variable regression models.

Dichotomous Dependent Variables

- So far, having described the OLS, we have assumed a statistical experiment that draws from a normal distribution.
- There are various problems associated with estimating a dichotomous dependent variable under assumptions of a statistical experiment that draws from a normal distribution, i.e., using regression.
- Obviously the statistical experiment is *not* draws from a normal distribution, but from something called a Bernoulli distribution. Thus, estimation with OLS is likely to be inefficient. It is also *theoretically* inconsistent with the nature of the statistical experiment.

Recall of a Bernoulli distribution

- Before introducing a Binomial random distribution let us introduce the Bernoulli distribution.
- A random variable X is called Bernoulli $X \sim B(1; p)$ if assumes only two values: 0 with probability 1-p and 1 with probability p, where 0≤p≤1, or equivalently

x_i	p_i
0	$1-p$
1	p
	1

- This random variable can be generated by drawing an outcome by a population whose units can ssume only two carachters, such as: YES-NO; RIGHT-WRONG; etc.

Recall of a Bernoulli distribution

- From a Bernoulli it is straightforward to calculate the rth moment

$$\mu_r = \sum_{i=1}^{k} x_i^r \, p_i = 0^r\,(1\text{-}p) + 1^r p = p.$$

- In particular:

$$\mu = p; \qquad \sigma^2 = p\,q;$$

A **Binomial** is a generalization of a Bernoulli (draws with ripetition/remission)

$$X \sim B(N; p).$$

Bernoulli indipendent random variables

$$X \sim B(N; p) = \sum_{i=1}^{N} B_i(1; p)$$

Binary response models

- Model for mutually exclusive binary outcomes focus on the determinants of the probability p of the occurrence of one outcome rather than an alternative outcome that occurs with a probability of 1-p (in regression analysis we want to measure how the probability p varies across individuals as function of regressors).
- An alternative is predicting the propensity score p, the conditional probability of participation of an individual in a treatment program

  - Linear Probability Model
  - Probit and Logit Regression
  - Estimation, Marginal Effect and Inference

Basic models

- Suppose the outcome *y* takes one of the two values:

$$y = \begin{cases} 1 & \text{with probability } p \\ 0 & \text{with probability } 1-p \end{cases}$$

- Given our interest in modeling p as a function of regressors x, there is no loss of generality in setting the outcome values to 1 and 0. The probability density function for the observed outcome *y* is $p^y(1\text{-}p)^{1\text{-}y}$, with E(*y*) = p and Var(*y*)=p(1-p).
- A regression model is formed by parametrizising the probability *p* on a regressor vector **x** and a Kx1 parameter vector beta. The commonly used models are of single-index form with conditional probability given by $p_i \equiv \Pr[y_i = 1|\mathbf{x}] = F(\mathbf{x}_i'\beta)$,

$$f(y_i|\mathbf{x}_i) = p_i^{y_i}(1-p_i)^{1-y_i}, \quad y_i = 0, 1,$$

This yields probabilities p_i and $(1-p_i)$ since $p^1(1-p)^0 = p$ and $f(0) = p^0(1-p)^1 = 1-p$.

Basic models

$$p_i \equiv \Pr[y_i = 1|\mathbf{x}] = F(\mathbf{x}_i'\boldsymbol{\beta}),$$

- Where F(.) is a specified parametric function. To ensure that 0 ≤ p ≤ 1 it is natural to specify F(.) to be a cumulative distribution function (CDF).
- Note that if F(.) is a CDF, then this cdf is only being used to model the parameter p and does not denote the cdf of *y* itself, except for the LPM that does not use cdf.

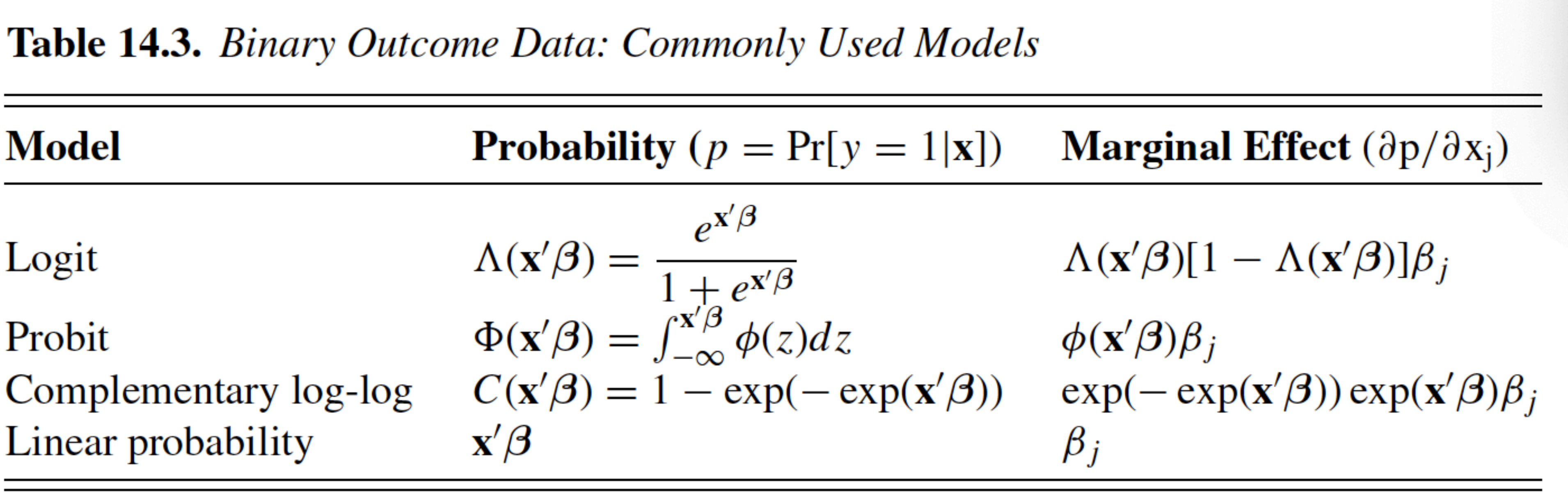


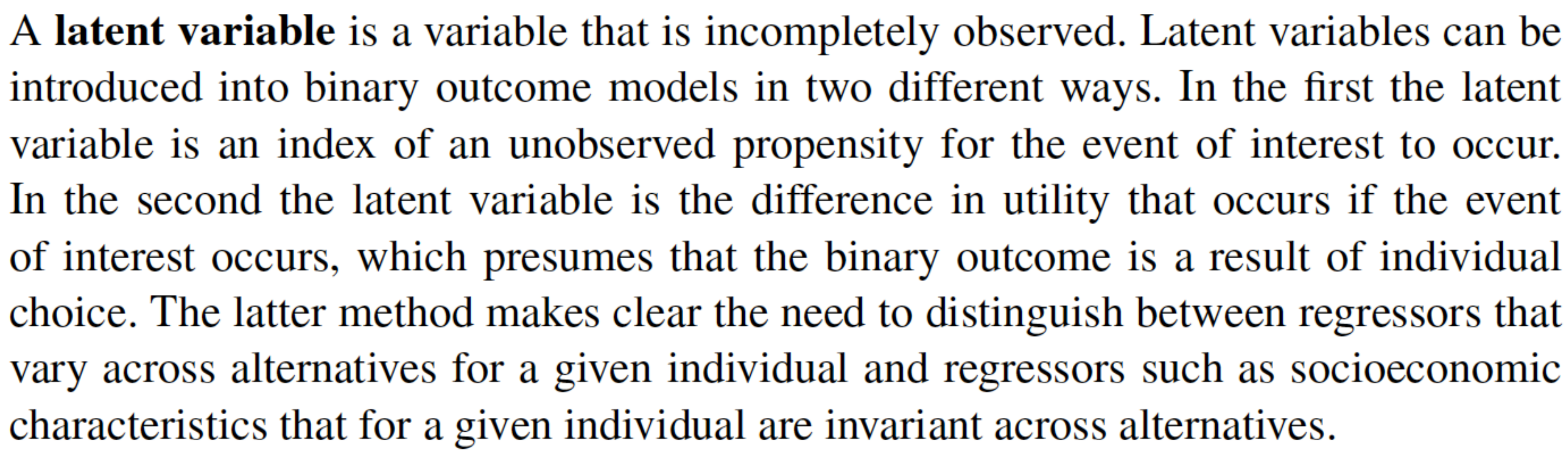
Table 14.3. *Binary Outcome Data: Commonly Used Models*

Model	**Probability** ($p = \Pr[y = 1\|\mathbf{x}]$)	**Marginal Effect** ($\partial p/\partial x_j$)
Logit	$\Lambda(\mathbf{x}'\boldsymbol{\beta}) = \dfrac{e^{\mathbf{x}'\boldsymbol{\beta}}}{1+e^{\mathbf{x}'\boldsymbol{\beta}}}$	$\Lambda(\mathbf{x}'\boldsymbol{\beta})[1-\Lambda(\mathbf{x}'\boldsymbol{\beta})]\beta_j$
Probit	$\Phi(\mathbf{x}'\boldsymbol{\beta}) = \int_{-\infty}^{\mathbf{x}'\boldsymbol{\beta}} \phi(z)dz$	$\phi(\mathbf{x}'\boldsymbol{\beta})\beta_j$
Complementary log-log	$C(\mathbf{x}'\boldsymbol{\beta}) = 1 - \exp(-\exp(\mathbf{x}'\boldsymbol{\beta}))$	$\exp(-\exp(\mathbf{x}'\boldsymbol{\beta}))\exp(\mathbf{x}'\boldsymbol{\beta})\beta_j$
Linear probability	$\mathbf{x}'\boldsymbol{\beta}$	β_j

Latent variable interpretation and identification

- Binary outcome models can be given a latent-variable interpretation. This provides a link with the linear regression model, explains more deeply the difference between logit and probit and provides the basis for extention to some multinomial models.

From Cameron and Trivedi (2005)



A **latent variable** is a variable that is incompletely observed. Latent variables can be introduced into binary outcome models in two different ways. In the first the latent variable is an index of an unobserved propensity for the event of interest to occur. In the second the latent variable is the difference in utility that occurs if the event of interest occurs, which presumes that the binary outcome is a result of individual choice. The latter method makes clear the need to distinguish between regressors that vary across alternatives for a given individual and regressors such as socioeconomic characteristics that for a given individual are invariant across alternatives.

Latent variable interpretation and identification

- We distinguish between the observed binary outcome, y, and an underlying continuous unobservable (or latent) variable, y^*, that satisfies the single- index model

$$y^* = \mathbf{x}'\boldsymbol{\beta} + u.$$

Altough y^* is not observed, we do observe

$$y = \begin{cases} 1 \text{ if } y^* > 0, \\ 0 \text{ if } y^* \leq 0, \end{cases}$$

Where the zero threshold is a normalization that is of no consequence if x includes an intercept. Given the latent-variable models just described, we have

$$\begin{aligned} \Pr[y = 1|\mathbf{x}] &= \Pr[y^* > 0] \\ &= \Pr[\mathbf{x}'\boldsymbol{\beta} + u > 0] \\ &= \Pr[-u < \mathbf{x}'\boldsymbol{\beta}] \\ &= F(\mathbf{x}'\boldsymbol{\beta}), \end{aligned}$$

F(.) is the CDF of $-u$. If u is standard normally distributed PROBIT; if logistically distributed LOGIT

Latent variable interpretation and identification

- We distinguish between the observed binary outcome, y, and an underlying continuous unobservable (or latent) variable, y^*, that satisfies the single- index model

$$y^* = \mathbf{x}'\beta + u.$$

Altough y^* is not observed, we do observe

$$y = \begin{cases} 1 \text{ if } y^* > 0, \\ 0 \text{ if } y^* \leq 0, \end{cases}$$

Where the zero threshold is a normalization that is of no consequence if x includes an intercept. Given the latent-variable models just described, we have

$$\begin{aligned} \Pr[y = 1|\mathbf{x}] &= \Pr[y^* > 0] \\ &= \Pr[\mathbf{x}'\beta + u > 0] \\ &= \Pr[-u < \mathbf{x}'\beta] \\ &= F(\mathbf{x}'\beta), \end{aligned}$$

F(.) is the CDF of *-u*. If *u* is standard normally distributed PROBIT; if logistically distributed LOGIT

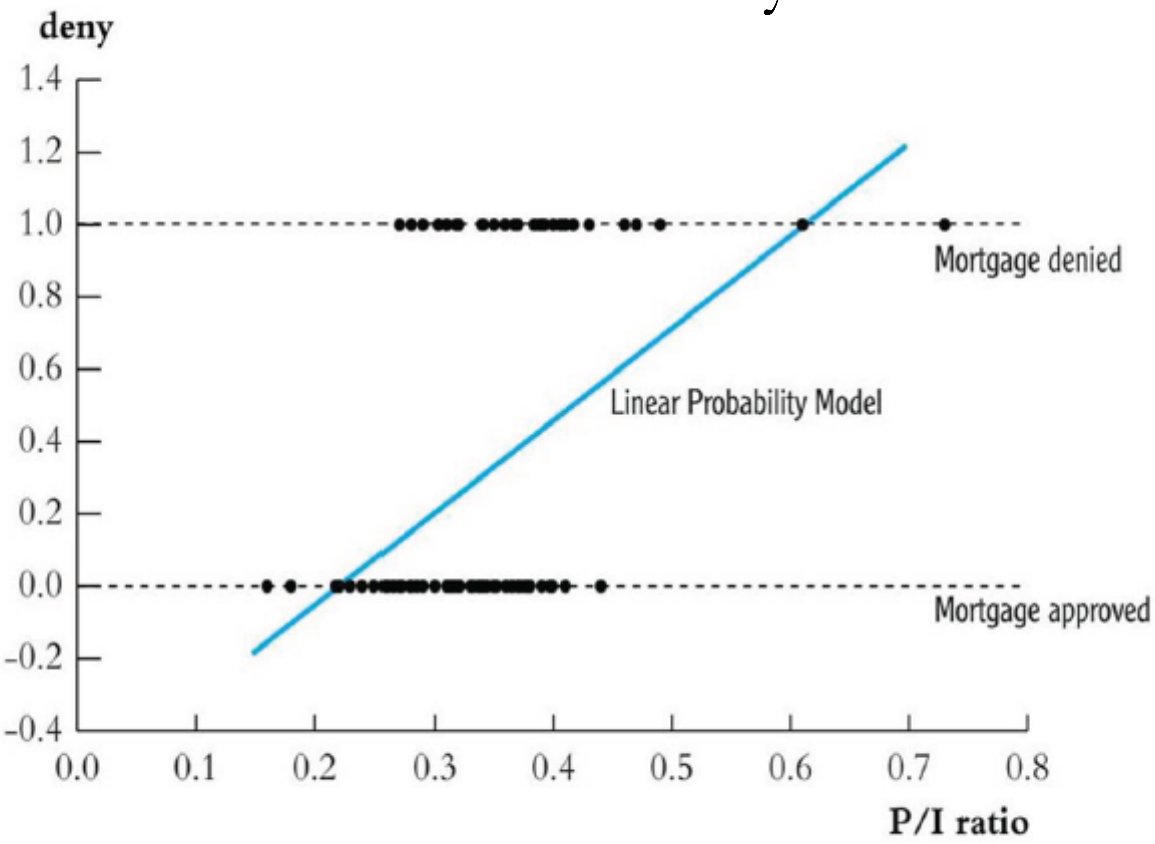
Some more information

- For binary models other than the LPM, estimation is by ML. The density for a single observation can be compatly written as is $p_i^{yi}(1-p_i)^{1-yi}$, where $p_i \equiv \Pr[y_i = 1|\mathbf{x}] = F(\mathbf{x}_i'\boldsymbol{\beta})$,
- For a sample of N independent observations, the MLE maximizes the associated log-likelihood function

$$\mathcal{L}_N(\boldsymbol{\beta}) = \sum_{i=1}^{N} \left\{ y_i \ln F(\mathbf{x}_i'\boldsymbol{\beta}) + (1 - y_i) \ln(1 - F(\mathbf{x}_i'\boldsymbol{\beta})) \right\}.$$

- The MLE is obtained by iterative methods and is asymptotically normally distributed. More iterations might mean high degree of multicollinearity.

Linear Probability model



- Note that a linear regression line through the actual data cuts through the data at the point of greatest concentration on each end.
- The residuals from this regression line will only be close to the regression line if the X variable is also Bernoulli distributed. This means that measures of fit or hypothesis tests involving the squared errors will be silly. The regression line will seldom lie near the data.

PROBLEMS WITH LINEAR PROBABILITY MODEL

- The linear probability model (LPM) uses the OLS method to determine the probability of an outcome.
- *Problems*:
  - 1. The LPM assumes that the probability of the outcome moves linearly with the value of the explanatory variable, no matter how small or large that value is.
  - 2. The probability value must lie between 0 and 1, yet there is no guarantee that the estimated probability values from the LPM will lie within these limits.
  - 3. The usual assumption that the error term is normally distributed cannot hold when the dependent variable takes only values of 0 and 1, since it follows the binomial distribution.
  - 4. The error term in the LPM is heteroscedastic, making the traditional significance tests suspect.

- Relatedly, this feature also means that the residuals from the linear model will be dichotomous and heteroskedastic, rather than normal, raising questions about hypothesis tests.

  When y=1, the residual will depend on X and be:

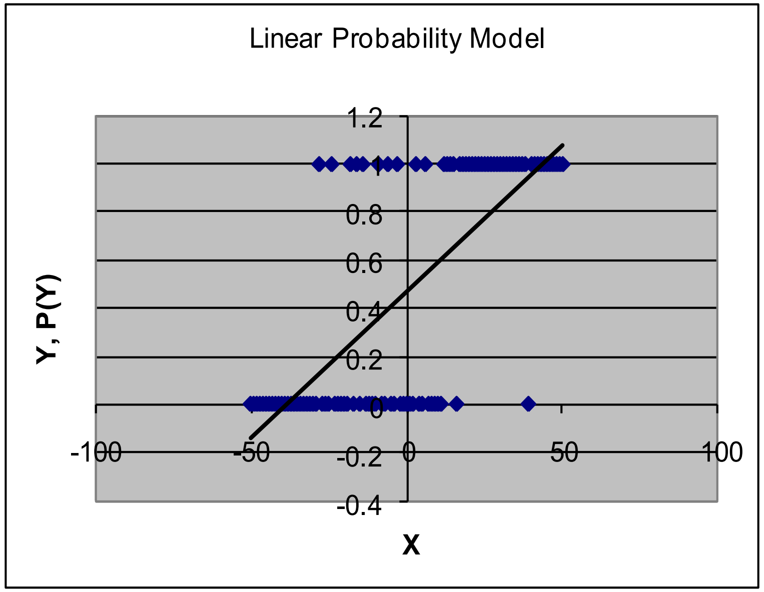
$$y = \hat{\beta}X + \varepsilon$$

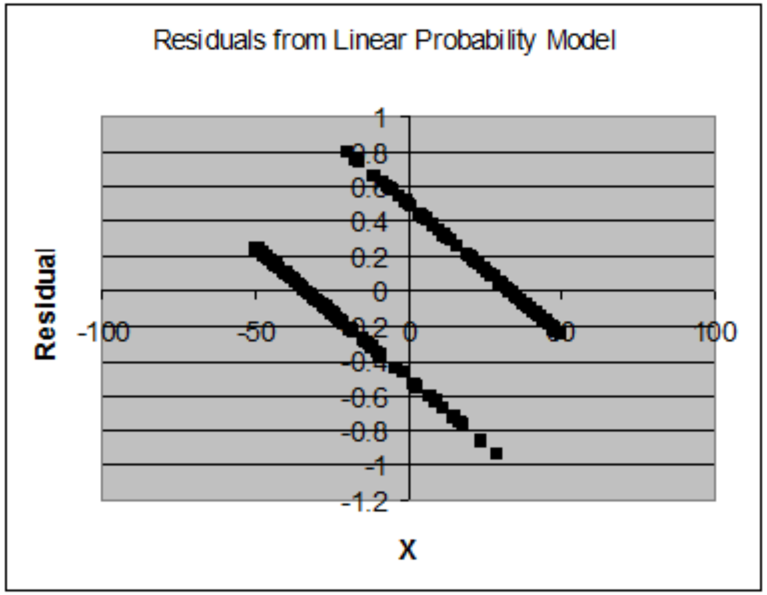
When y=0, the residual will depend on X and be: $\varepsilon = 1 - \hat{\beta}X$

$$y = \hat{\beta}X + \varepsilon$$

$$\varepsilon = -\hat{\beta}X$$

- This means that the residuals from the linear probability model will be heteroskedastic and have a dichotomous character.

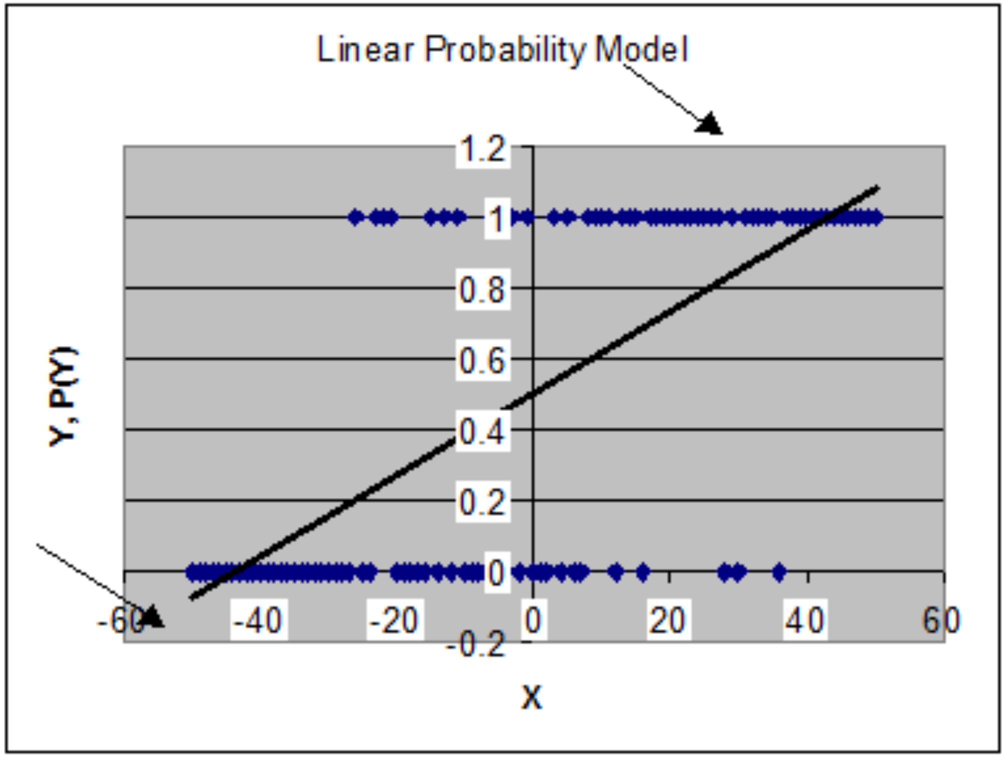




- **Note that the residuals change systematically with the values of X. This implies what it termed endogeneity.They are also not distributed normally.**

  We could "fix" this problem by estimating the linear probability model using weighted least squares.

  However, the problem with this model runs deeper. We must be able to interpret results from this model as expected values of probabilities. However, the graph below suggests further problems.



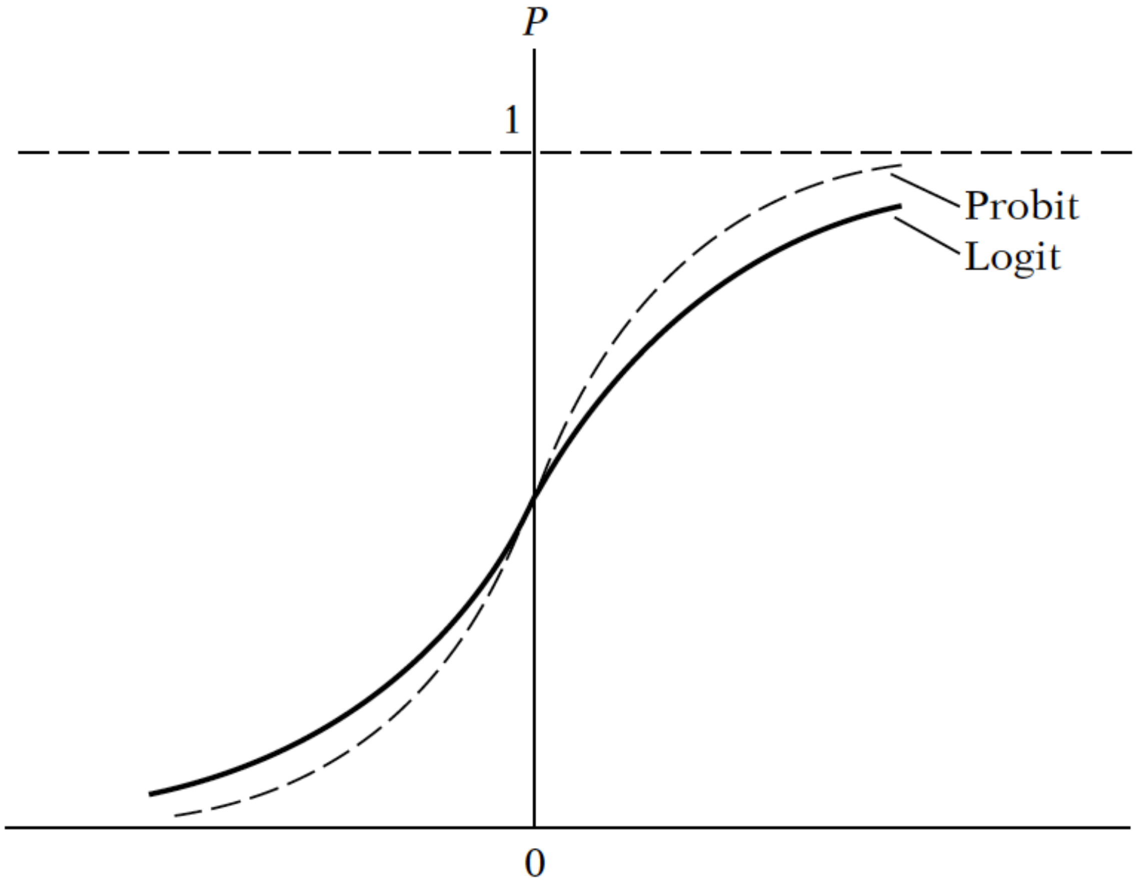
- **Observe that some of the probabilities lie above 1 and below zero. This is not consistent with the rules of probability. We could truncate the model at 0 and 1 to "fix" this problem.**

  However, note that probability, according to this model, is alleged to change in linear fashion with changes in X. Yet, this may not be consistent with reality in many real world situations. For example, consider the probability of home ownership as a function of income.

  Suppose we have prospective buyers with income around 10k per year. If we change their income by 1k, how much does the probability that they will buy a home change? Suppose we have prospective buyers with income around 30k. If we change their income by 1k, how much does the probability that they will own a home change? Suppose we have prospective buyers with income around 80k. If we change their income by 1k, how much does the probability that they will own a home change?

- **In practice, there are many situations where the probability of a yes outcome follows an S shaped distribution, rather than the linear distribution alleged by the linear probability model.**

Non-Linear Probability Models



CHARACTERISTICS OF LOGIT/PROBIT MODEL

- 1. As P_i goes from 0 to 1, L_i goes from -∞ to ∞.
- 2. Although L_i is linear in X_i, the probabilities themselves are not.
- 3. If L_i is positive, when the value of the explanatory variable(s) increases, the odds of the outcome increase. If it negative, the odds of the outcome decrease.
- 4. The interpretation of the logit model is as follows: Each slope coefficient shows how the log of the odds in favor of the outcome changes as the value of the *X* variable changes by a unit.
- 5. Once the coefficients of the logit model are estimated, we can easily compute the probabilities of the outcome.
- 6. In the LPM the slope coefficient measures the marginal effect of a unit change in the explanatory variable on the probability of the outcome, holding other variables constant. In the logit model, the marginal effect of a unit change in the explanatory variable not only depends on the coefficient of that variable but also on the level of probability from which the change is measured. The latter depends on the values of all the explanatory variables in the model.

LOGIT vs PROBIT

- Logit and probit models generally give similar results.
- The main difference between the two models is that the logistic distribution has slightly fatter tails.
  - The conditional probability P_i approaches 0 or 1 at a slower rate in logit than in probit.
- In practice there is no compelling reason to choose one over the other.
- Many researchers choose the logit over the probit because of its comparative mathematical simplicity.

$$\begin{aligned}\widehat{\beta}_{\text{Logit}} &\simeq 4\widehat{\beta}_{\text{OLS}},\\ \widehat{\beta}_{\text{Probit}} &\simeq 2.5\widehat{\beta}_{\text{OLS}},\\ \widehat{\beta}_{\text{Logit}} &\simeq 1.6\widehat{\beta}_{\text{Probit}}.\end{aligned}$$

STATA

Measures of fit

The R^2 and R^2 corrected don't make sense here. So, two other specialized measures are used:

- The **fraction correctly predicted = fraction of** Y 's for which predicted probability is >50% (if $Y_i = 1$) or is <50% (if $Y_i = 0$).
- The **pseudo-R^2** measure the fit using the likelihood function: measures the improvement in the value of the log likelihood, relative to having no X's. This simplifies to the R^2 in the linear model with normally distributed errors.

Interpretation

- **Interpreting Dichotomous Logit and Probit**
  - **Coefficients- The actual coefficients in a logit or probit analysis are limited in their immediate interpretability.**
  - **The signs are meaningful, but the magnitudes may not be, particularly when the variables are in different metrics.**
  - **Above all, note that you cannot interpret the coefficients directly in terms of units of change in y for a unit change in x, as in regression analysis.**
- **There are various approaches to imparting substantive meaning into logit and probit results, including:**
  - **Probability Calculations**
  - **Graphical methods**
  - **Odd Ratio**
  - **First Partial derivatives.**

Marginal Effect

However, what we really care is not $\hat{\beta}_1$ itself. We want to know how the change of X will affect the probability that Y = 1. For the probit model,

$$\begin{aligned} \Pr(Y = 1|X) &= \Phi(\hat{\beta}_0 + \hat{\beta}_1 X) \\ \frac{\partial \Pr(Y = 1|X)}{\partial X} &= \phi(\hat{\beta}_0 + \hat{\beta}_1 X)\hat{\beta}_1 \end{aligned}$$

wher $\phi(\cdot)$ is pdf of the standard normal distribution.

The effect of the change in X on Pr(Y = 1|X) depends on the value of X. In practice, we usually evalute the *marginal effect* at the sample average $\bar{X}$. i.e. The marginal effect is

$$\phi(\hat{\beta}_0 + \hat{\beta}_1 \bar{X})\hat{\beta}_1$$

When X is binary, it is not clear what does the sample average mean.

The marginal effect then measures the probability difference between X = 1 and X = 0.

$$
\begin{aligned}
& \Pr(Y = 1|X = 1) - \Pr(Y = 1|X = 0) \\
= \; & \Phi(\hat{\beta}_0 + \hat{\beta}_1) - \Phi(\hat{\beta}_0)
\end{aligned}
$$