Limited Dependent Variable Models and Sample Selection Corrections.

- **Limited dependent variables (LDV)**
  - LDV are substantively restricted
    - Binary vavariables take only two values, e.g. employed/not employed
    - Nonnegative variables, e.g. wages, prices, interest rates
    - Nonnegative variables with excess zeros, e.g. labor supply
    - Count variables, e.g. the number of arrests in a year

- **Sample selection models**
  - The sample used to infer population relationships is endogenously selected, e.g. wage offer regression but data only about working women.

- **Logit and Probit models for binary response**

- Disadvantages of the LPM for binary dependent variables
  - Predictions sometimes lie outside the unit interval
  - Partial effects of explanatory variables are constant

- Nonlinear models for binary response
  - Response probability is a nonlinear function of explanat. variables

$$P(y=1|\mathbf{x}) = G(\beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k) = G(\mathbf{x}\beta)$$

Probability of a "success" given explanatory variables

A cumulative distribution function $0 < G(z) < 1$. The response probability is thus a function of the explanatory variables x.

Shorthand vector notation: the vector of explanatory variables x also contains the constant of the model.

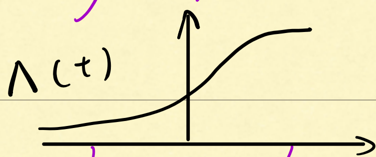
$$P(y=1|x) = G(\underbrace{\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k}_{z}) = G(x\beta).$$

$$0 \le G(z) \le 1.$$

Candidates for G:

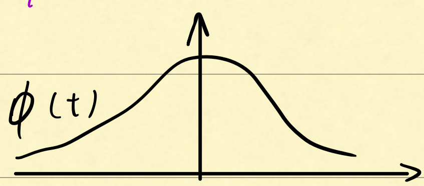
① logit model: $G(z) = \dfrac{e^{z}}{1+e^{z}} = \Lambda(z)$

c.d.f of standard logistic distribution



② probit model: $G(z) = \Phi(z) = \int_{-\infty}^{z} \phi(z)\,dz$

c.d.f. of standard normal distribution.



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

代码: glm (y ~ $x_1 + x_2 + \cdots + x_k$, family = binomial (link = probit / logit))

Latent variable formulation of the logit and probit models.

$y^* = x\beta + e$ and $y = 1\cdot[y^* > 0]$.

e: { standard logistic distribution $L(0,1)$; standard normal distribution $N(0,1)$ } ; indep of x

$$\Longrightarrow P(y=1 \mid x) = P(y^* > 0 \mid x)$$
$$= P(x\beta + e > 0 \mid x)$$
$$= P(e > -x\beta \mid x)$$
$$= 1 - \underset{\text{or } \Phi}{\Lambda}(-x\beta \mid x)$$
$$= \underset{\text{or } \Phi}{\Lambda}(x\beta \mid x)$$

- **Interpretation of coefficients in Logit and Probit models**

Continuous explanatory variables:

$$\frac{\partial P(y=1|\mathbf{x})}{\partial x_j} = g(\mathbf{x}\boldsymbol{\beta})\beta_j \quad \text{where} \quad g(z) \equiv \partial G(z)/\partial z > 0$$

How does the probability for y = 1 change if explanatory variable x_j changes by one unit?

Discrete explanatory variables:

$$G[\beta_0 + \beta_1 x_1 + \ldots + \beta_k(c_k+1)] - G[\beta_0 + \beta_1 x_1 + \ldots + \beta_k c_k]$$

For example, explanatory variable x_k increases by one unit.

- Partial effects are nonlinear and depend on the level of x.

- **Maximum likelihood estimation of Logit and Probit models**

$$f(y_i|\mathbf{x}_i;\boldsymbol{\beta}) = [G(\mathbf{x}_i\boldsymbol{\beta})]^{y_i}[1-G(\mathbf{x}_i\boldsymbol{\beta})]^{1-y_i}$$

The probability that individual i's outcome is y_i given that his/her characteristics are x_i

$$\log L(\boldsymbol{\beta}) = \log\left(\prod_{i=1}^{n} f(y_i|\mathbf{x}_i;\boldsymbol{\beta})\right) = \sum_{i=1}^{n} \log f(y_i|\mathbf{x}_i;\boldsymbol{\beta})$$

Under random sampling

$$\max \ \log L(\boldsymbol{\beta}) \quad \rightarrow \quad \widehat{\beta}_0, \widehat{\beta}_1, \ldots, \widehat{\beta}_k$$

Maximum likelihood estimates

- Properties of maximum likelihood estimators
  - Maximum likelihood estimators are consistent, asymptotically normal, and asymptotically efficient if the distributional assumptions hold.
- **Hypothesis testing after maximum likelihood estimation**
  - The usual t-tests and confidence intervals can be used.
  - There are three alternatives to test multiple hypotheses:
    - Lagrange multiplier or score test (not discussed here)
    - Wald test (requires only estimation of unrestricted model)
    - Likelihood ratio test (restricted and unrestricted models needed)

$$LR = 2(\log L_{ur} - \log L_r) \sim \chi^2_q$$

Chi-square distribution with q degrees of freedom

The null hypothesis that the q hypotheses hold is rejected if the growth in maximized likelihood is too large when going from the restricted to the unrestricted model

- **Goodness-of-fit measures for Logit and Probit models**
  - Percent correctly predicted $= \frac{\# \text{correctly predicted } \tilde{y}}{\# \ \tilde{y}}$.

$$\tilde{y}_i = \begin{cases} 1 & \text{if } G(\mathbf{x}_i\hat{\boldsymbol{\beta}}) \geq .5 \\ 0 & \text{otherwise} \end{cases}$$

predicted.

Individual i's outcome is predicted as one if the probability for this event is larger than .5, then percentage of correctly predicted y = 1 and y = 0 is counted

- Pseudo R-squared

$$\tilde{R}^2 = 1 - \log L_{ur} / \log L_0$$

Compare maximized log-likelihood of the model with that of a model that only contains a constant (and no explanatory variables)

- Correlation based measures

$$Corr(y_i, \tilde{y}_i), \ Corr(y_i, G(\mathbf{x}_i\hat{\boldsymbol{\beta}}))$$

Look at correlation (or squared correlation) between predictions or predicted prob. and true values

- **Reporting partial effects of explanatory variables**
  - The difficulty is that partial effects are not constant but depend on.
  - Partial effects at the average:

$$\widehat{PEA}_j = g(\bar{\boldsymbol{x}}\hat{\boldsymbol{\beta}})\hat{\beta}_j$$

The partial effect of explanatory variable x_j is considered for an "average" individual (this is problematic in the case of explanatory variables such as gender)

  - Average partial effects:

$$\widehat{APE}_j = n^{-1} \sum_{i=1}^{n} g(\mathbf{x}_i\hat{\boldsymbol{\beta}})\hat{\beta}_j$$

The partial effect of explanatory variable x_j is computed for each individual in the sample and then averaged across all sample members (makes more sense)

  - Analogous formulas hold for discrete explanatory variables.

```
glm ( y ~ x1 + x2 + ... + xk , family = binomial (link = probit))
                                                         logit
summary (A)
LogLik (A)
```

Likelihold rate: $LR = 2(\mathrm{LogLik}(ur) - \mathrm{LogLik}(r))$

$$\sim \chi^2_q$$

$$P(y=1 \mid x) = \int_{-\infty}^{z} \lambda(z)\, dz = \int_{-\infty}^{z} \frac{e^v}{(1+e^v)^2}\, dv.$$

$$= \frac{e^z}{(1+e^z)^2}$$

$$= \frac{e^{\beta_0+\beta_1 x_1+\cdots \beta_k x_k}}{1+e^{\beta_0+\beta_1 x_1+\cdots+\beta_k x_k}}$$

$$\widehat{P(y=1 \mid x)} = \frac{e^{\hat\beta_0+\hat\beta_1 x_1+\cdots \hat\beta_k x_k}}{1+e^{\hat\beta_0+\hat\beta_1 x_1+\cdots+\hat\beta_k x_k}}$$