Estimation of Binary Response Models

(GB: Chapter 13.3–13.6; 15.4-15.7)

Isabel Casas

SYDDANSK UNIVERSITET

Maximum Likelihood Estimation

- The likelihood function
- Maximising the log-likelihood

Motivation

- Likelihood means probability
- Assuming certain probability distribution, what are the parameters of it for a particular sample?

Motivation

- We have coin A and coin B? Would you bid head of tails in either of them?
- What is the value of p for coin A? and for coin B?
- After 10 throws:

```
> set.seed(42)
> A=rbinom(10, 1, 0.7)
> B=rbinom(10, 1, 0.3)
```

  - coin A: 5 heads and 5 tails
  - coin B: 4 heads and 6 tails

- What is the estimated probability of head for each coin?

Motivation

Each thrown follows a Bernoulli distribution:

$$y_j = \begin{cases} 1 & p \\ 0 & (1-p) \end{cases} \qquad f_y(y_j) = p^{y_j}(1-p)^{1-y_j}$$

The joint distribution function of our sample $\{y_1, \ldots, y_{10}\}$, or likelihood function:

$$f_{y_1,\ldots,y_{10}}(y_1, \ldots, y_{10}) = \prod_{j=1}^{10} p^{y_j}(1-p)^{1-y_j} \quad \text{independence}$$

Motivation

The loglikelihood function:

$$\begin{aligned}\ell(p) = f_{y_1,\ldots,y_{10}}(y_1,\ldots,y_{10}) =& \sum_{j=1}^{10}[y_j \log(p) + (1-y_j)\log(1-p)] \\ =& 5\log(p) + 5\log(1-p) \quad \text{coin A} \\ =& 4\log(p) + 6\log(1-p) \quad \text{coin B}\end{aligned}$$

Motivation

What is the value of p that **maximises** the loglikelihood function of my sample?

Take first derivative and equal to zero.

Coin A:

$$\frac{5}{p} - \frac{5}{1-p} = 0 \Rightarrow \hat{p}_A = 0.5$$

Coin B:

$$\frac{4}{p} - \frac{6}{1-p} = 0 \Rightarrow \hat{p}_B = 0.4$$

Motivation

- Let us throw the coin again, after 100 throws:

```
> set.seed(42)
> A=rbinom(100, 1, 0.7)
> B=rbinom(100, 1, 0.3)
```

  - coin A: 66 heads and 34 tails
  - coin B: 37 heads and 63 tails

- What do you think the probability of getting head for each coin is?
- The maximum likelihood estimator of each coin are: $\hat{p}_A = 0.66$ and $\hat{p}_B = 0.37$ which is closer to reality.

The likelihood function

- Assume we have a sample of variables $\mathbf{y}$ and $\mathbf{X}$
- We know the conditional distribution of the $\mathbf{y}$ variable given $\mathbf{X}$
- The particular parameters of this distribution are unknown and depend on $\boldsymbol{\beta}$
- Therefore, estimating the parameters of this distribution will provide estimates for $\boldsymbol{\beta}$
- These estimates are found as the parameters that most likely have generated the observed sample
- Find the values that approximates the sample distribution best

MLE set–up

- We have a sample $\{(y_j, \mathbf{x}_j)\}$ with $j = 1, 2, \ldots, n$
- Assume a conditional density function of y_j given $\mathbf{x}_j$

$$f(y_j|\mathbf{x}_j)$$

- Note that we make assumptions on the shape of the sample distribution
- In OLS we only make assumptions on the shape of the expectation, which is a most less restrictive assumption

MLE set–up

The conditional log–likelihood function for observation j is

$$\ell_j(\boldsymbol{\beta}) = \log f(y_j|\mathbf{x}_j, \boldsymbol{\beta})$$

The conditional log–likelihood function for the whole sample is

$$\ell(\boldsymbol{\beta}) = \sum_{j=1}^{n} \ell_j(\beta) = \sum_{j=1}^{n} \log f(y_j|\mathbf{x}_j, \boldsymbol{\beta})$$

and the maximum likelihood estimator is the value that minimises the function above

$$\hat{\boldsymbol{\beta}}^{ML} = \arg\max_{\boldsymbol{\beta}} \sum_{j=1}^{n} \log f(y_j|\mathbf{x}_j, \boldsymbol{\beta})$$

MLE for Index models

y takes values 0 and 1 (Bernoulli) with probability

$$p_j = P(y_j = 1|\mathbf{X}) = G(\mathbf{x}_j\boldsymbol{\beta})$$

Its conditional probability mass function is:

$$f(y_j|\mathbf{x}_j, \boldsymbol{\beta}) = p_j^{y_j}(1 - p_j)^{1-y_j}, \quad y_j = 0, 1$$

We have made the assumption of the distribution of the sample

The conditional log–likelihood function for observation j is

$$\ell_j(\boldsymbol{\beta}) = \log f(y_j|\mathbf{x}_j, \boldsymbol{\beta}) = y_j \log p_j + (1 - y_j) \log(1 - p_j)$$

MLE for Index models

The conditional log–likelihood function for the whole sample is

$$\ell(\beta) = \sum_{j=1}^{n} \ell_j(\boldsymbol{\beta}) = \sum_{j=1}^{n} [y_j \log G(\mathbf{x}_j\boldsymbol{\beta}) + (1 - y_j) \log(1 - G(\mathbf{x}_j\boldsymbol{\beta}))]$$

The ML estimator is the one that maximises the conditional log–likelihood function.

$$\hat{\boldsymbol{\beta}}^{ML} = \arg\max_{\boldsymbol{\beta}} \ell(\boldsymbol{\beta})$$

How do we maximise a function?

- Find maximum (all first derivatives w.r.t $\beta_i = 0$)
- If second derivative $< 0 \Rightarrow$ maximum

MLE for Index models

$\hat{\boldsymbol{\beta}}^{ML}$ solves

$$\frac{\partial \ell(\beta)}{\partial \beta_i} = \sum_{j=1}^{n} \frac{y_j - G(\mathbf{x}_j\boldsymbol{\beta})}{G(\mathbf{x}_j\boldsymbol{\beta})(1 - G(\mathbf{x}_j\boldsymbol{\beta}))} G'(\mathbf{x}_j\boldsymbol{\beta}) x_{ij} = 0$$

There are $k+1$ of these equations... one for each β_i.

$$s(\beta) = \begin{pmatrix} \frac{\partial \ell(\beta)}{\partial \beta_0} \\ \vdots \\ \frac{\partial \ell(\beta)}{\partial \beta_j} \\ \vdots \\ \frac{\partial \ell(\beta)}{\partial \beta_k} \end{pmatrix} = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ \vdots \\ 0 \end{pmatrix}$$

Computer work!!!!

MLE for Index models

The parameter $\boldsymbol{\beta}$ is a maximum, i.e. it does maximises the log-lik function, if the matrix of second derivatives of the log-lik, the Hessian:

$$H(\beta) = \begin{pmatrix} \frac{\partial^2 \ell(\beta)}{\partial \beta_1^2} & \frac{\partial^2 \ell(\beta)}{\partial \beta_1 \partial \beta_2} & \cdots & \frac{\partial^2 \ell(\beta)}{\partial \beta_1 \partial \beta_k} \\ \frac{\partial^2 \ell(\beta)}{\partial \beta_2 \partial \beta_1} & \frac{\partial^2 \ell(\beta)}{\partial \beta_2^2} & \cdots & \frac{\partial^2 \ell(\beta)}{\partial \beta_2 \partial \beta_k} \\ \vdots & & & \vdots \\ \frac{\partial^2 \ell(\beta)}{\partial \beta_k \partial \beta_1} & \frac{\partial^2 \ell(\beta)}{\partial \beta_k \partial \beta_2} & \cdots & \frac{\partial^2 \ell(\beta)}{\partial \beta_k^2} \end{pmatrix}$$

is negative definite.

MLE for the logit model

$\mathbf{y}$ is binary with probability p_j and we assume we can fit a **logit** model to our data.

$$p_j = \Lambda(\mathbf{x}_j\boldsymbol{\beta}) = \frac{e^{\mathbf{x}_j\boldsymbol{\beta}}}{1 + e^{\mathbf{x}_j\boldsymbol{\beta}}} \quad j = 1, \ldots, n$$

The $\hat{\boldsymbol{\beta}}^{ML}$ estimator solves

$$\sum_{j=1}^{n} \frac{y_j - \Lambda(\mathbf{x}_j\boldsymbol{\beta})}{\Lambda(\mathbf{x}_j\boldsymbol{\beta})(1 - \Lambda(\mathbf{x}_j\beta))} \Lambda'(\mathbf{x}_j\boldsymbol{\beta}) x_{ij} = 0$$

There is not explicit solution, so we would need numerical methods to solve this equation such as the Newton–Raphson iterative procedure.

MLE for the probit model

$\mathbf{y}$ is binary with probability p_j and we assume we can fit a **probit** model to our data.

$$p_j = \Phi(\mathbf{x}_j\boldsymbol{\beta}) \quad j = 1, \ldots, n$$

The ML estimator is the β that solves

$$\sum_{j=1}^{n} \frac{y_j - \Phi(\mathbf{x}_j\boldsymbol{\beta})}{\Phi(\mathbf{x}_j\boldsymbol{\beta})(1 - \Phi(\mathbf{x}_j\beta))}\phi(\mathbf{x}_j\boldsymbol{\beta})x_{ij} = 0$$

There is not explicit solution, so we would need numerical methods to solve this equation such as the Newton–Raphson iterative procedure.

Exercise (5 minutes)

We are using a probit model:

1. Assume that $\mathbf{x}_j\boldsymbol{\beta} = \beta_0 + \beta_1 X_{1j}$ and we have the following sample:
   - $y_1 = 1,\ x_{11} = 1$
   - $y_2 = 0,\ x_{12} = 0.5$
   - $y_3 = 1,\ x_{13} = 2$
2. Write the log–likelihood function of glm model assuming that G is the identity function.

MLE properties

If

- Distribution of $\mathbf{y}$ given $\mathbf{X}$ is correctly specified
- Parameters are identified (not like the variance in the probit)
- The log-lik is smooth (two derivatives)

Then

- Consistent (Theorem 13.1)
- Asymptotically normal (Theorem 13.2)
- Asymptotic efficiency

MLE properties

Two key concepts:

- The Score Vector(dimension $(k+1) \times 1$)
  - The vector of first derivatives of the log-lik with respect to parameters
  - Used for the first order conditions (f.o.c.)
- The Hessian Matrix (dimension $(k+1) \times (k+1)$)
  - The matrix of second derivatives (including cross derivatives) of the log-lik with respect to parameters
  - Used for the second order conditions and for the variance of the estimators.

Score vector

The *score vector* of the log-lik for observation i:

$$\mathbf{s}_j(\beta) = \left(\frac{\partial \ell_j(\beta)}{\partial \beta_0}, \frac{\partial \ell_j(\beta)}{\partial \beta_1}, \ldots, \frac{\partial \ell_j(\beta)}{\partial \beta_k} \right)'$$

The first order condition of the max log-lik is that summing the individual scores $(j = 1, \ldots, n)$ is $=0$:

$$\sum_{j=1}^{n} \mathbf{s}_j(\beta) = \left(\sum_{j=1}^{n} \frac{\partial \ell_j(\beta)}{\partial \beta_0}, \sum_{j=1}^{n} \frac{\partial \ell_j(\beta)}{\partial \beta_1}, \ldots, \sum_{j=1}^{n} \frac{\partial \ell_j(\beta)}{\partial \beta_k} \right) = (0, 0, \ldots, 0)$$

Hessian

The *Hessian matrix* of the log-lik for observation i:

$$\mathbf{H}_j(\boldsymbol{\beta}) = \nabla_\beta \mathbf{s}_j(\beta) \;\Rightarrow\; \mathbf{H} = \sum_{j=1}^{n} \mathbf{H}_j$$

- The Hessian is a $(k+1) \times (k+1)$ matrix
- It is symmetric because

$$\frac{\partial^2 \ell_j(\boldsymbol{\beta})}{\partial \beta_1 \beta_2} = \frac{\partial^2 \ell_j(\boldsymbol{\beta})}{\partial \beta_2 \beta_1}$$

- The Hessian is negative definite
- A variance matrix is positive definite.

Conditional information matrix

It can be shown that,

$$-E[H_j(\boldsymbol{\beta})|\mathbf{x}_j] = Var[s_j(\boldsymbol{\beta})|\mathbf{x}_j] = A(\mathbf{x}_j, \boldsymbol{\beta})$$

When $\boldsymbol{\beta}$ is the true value that minimises the log-lik:

$$-E(H_j(\boldsymbol{\beta}_{true})) = E(s_j(\boldsymbol{\beta}_{true})s_j(\boldsymbol{\beta}_{true})') = A_0$$

because $E[s_j(\boldsymbol{\beta}_{true})|\mathbf{x}_j] = 0$

Consistency (Theorem 13.1)

If

- Distribution of $\mathbf{y}$ given $\mathbf{X}$ is correctly specified
- Parameters are identified (not like the variance in the probit)
- The log-lik is continuous

Then,

$$plim\ \hat{\boldsymbol{\beta}}^{ML} = \boldsymbol{\beta}_{true} \quad \text{as} \quad n \to \infty$$

Asymptotic normality (Theorem 13.2)

If

- Distribution of $\mathbf{y}$ given $\mathbf{X}$ is correctly specified
- Parameters are identified (not like the variance in the probit)
- The log-lik has two derivatives

$$\sqrt{N}(\hat{\boldsymbol{\beta}}^{ML} - \boldsymbol{\beta}_{true}) \rightarrow^{d} N(0, A_0^{-1})$$

where A_0 is a positive definite matrix.

The Hessian is a negative definite matrix, so the negative sign makes it positive definite, as the variance:

$$-E(H_j(\boldsymbol{\beta}_{true})) = E(s_j(\boldsymbol{\beta}_{true})s_j(\boldsymbol{\beta}_{true})')$$

Asymptotic variance of estimator

The asymptotic variance of the estimator is:

$$\mathbf{V}^{ML} = AVar(\hat{\boldsymbol{\beta}}^{ML}) = A_0^{-1}/n = \left[-E(H_j(\boldsymbol{\beta}_{true}))\right]^{-1}/n$$

But, $\boldsymbol{\beta}_{true}$ is unknown... three possible estimators:

$$\hat{\mathbf{V}}^{ML} = \widehat{AVar}(\hat{\boldsymbol{\beta}}^{ML}) = \left[\sum_{j=1}^{n} -H_j(\hat{\boldsymbol{\beta}}^{ML})\right]^{-1}$$

or

$$= \left[\sum_{j=1}^{n} s_j(\hat{\boldsymbol{\beta}}^{ML}) s_j(\hat{\boldsymbol{\beta}}^{ML})'\right]^{-1}$$

or

$$= \left[\sum_{j=1}^{n} -E[H(y_j, \mathbf{x}_j, \hat{\boldsymbol{\beta}}^{ML})|\mathbf{x}_j]\right]^{-1}$$

Asymptotic variance of estimator

Which variance estimator we choose?

- It is up to you.
- The first estimator requires second order derivatives of the log-lik and it is not guaranteed to be positive definite
  - If it is not pos. definite then the se of the estimators will not be defined (sqrt)
- The second estimator is always positive definite. However, it is not very good in moderate sample sizes
- If we know the conditional expectation close form, this is the most attractive estimator.
  - Positive definite (it depends on the informatoin matrix)
  - Behaves well in moderate samples
  - Usually only the first derivative of the log-lik is required

MLE for binary models

- The score vector:

$$s_j(\beta) = \frac{G'(\mathbf{x}_j\boldsymbol{\beta})[y_j - G(\mathbf{x}_j\boldsymbol{\beta})]}{G(\mathbf{x}_j\boldsymbol{\beta})[1 - G(\mathbf{x}_j\boldsymbol{\beta})]}\mathbf{x}_j'$$

- The expected value of the Hessian conditioned on $\mathbf{x}_j$:

$$-E[H_j(\beta)|\mathbf{x}_j] = \frac{[G'(\mathbf{x}_j\boldsymbol{\beta})]^2\mathbf{x}_j'\mathbf{x}_j}{G(\mathbf{x}_j\boldsymbol{\beta})(1 - G(\mathbf{x}_j\boldsymbol{\beta}))} \equiv A(\mathbf{x}_j, \beta)$$

- Hence, we use the third variance matrix estimator in this case.
- What are the standard errors?

MLE for binary models

- For large samples:

$$\hat{\boldsymbol{\beta}}^{ML} \sim N(\beta, \hat{\mathbf{V}}^{ML})$$

where

$$\hat{\mathbf{V}}^{ML} = \left\{ \sum_{j=1}^{n} \frac{[G'(\mathbf{x}_j\hat{\beta})]^2 \mathbf{x}_j' \mathbf{x}_j}{G(\mathbf{x}_j\hat{\beta})(1 - G(\mathbf{x}_j\hat{\beta}))} \right\}^{-1}$$

the standard errors are the square root of the estimated variance.

Reporting result

- The coefficient estimates, standard error and the value of log-likelihood function is given by R
- The marginal effect of the coefficients depends on the function G', so the values $\hat{\beta}$ do not explain the effect of $\mathbf{X}$ on $\mathbf{y}$
- The sign of the coefficients correspond with the sign of the effect

Estimating marginal effects

The marginal effect of variable $\mathbf{x}_j$ continuous:

$$\text{logit}: \quad \frac{\exp(-X\hat{\beta})}{\left[1+\exp(-X\hat{\beta})\right]^2}\hat{\beta}_j$$

$$\text{probit}: \quad \phi(X\hat{\beta})\hat{\beta}_j$$

- They depend on $\hat{\beta}$ and $\mathbf{X}$
- Average of the marginal effects, take the mean of those quantities
- Marginal effects of the sample average of $\mathbf{X}$
- The marginal effects can be compared with the coefficients of the LPM.

Estimating marginal effects

The marginal effect of variable $\mathbf{x}_j$ discrete, from 0 to 1:
Logit:

$$\Lambda(\hat{\beta}_0 + \hat{\beta}_1 \bar{X}_1 + \ldots + \hat{\beta}_{j-1} \bar{X}_{j-1} + \hat{\beta}_j 1 + \hat{\beta}_{j+1} \bar{X}_{j+1} + \ldots + \hat{\beta}_k \bar{X}_k)$$
$$-\Lambda(\hat{\beta}_0 + \hat{\beta}_1 \bar{X}_1 + \ldots + \hat{\beta}_{j-1} \bar{X}_{j-1} + \hat{\beta}_j 0 + \hat{\beta}_{j+1} \bar{X}_{j+1} + \ldots + \hat{\beta}_k \bar{X}_k)$$

Probit:

$$\Phi(\hat{\beta}_0 + \hat{\beta}_1 \bar{X}_1 + \ldots + \hat{\beta}_{j-1} \bar{X}_{j-1} + \hat{\beta}_j 1 + \hat{\beta}_{j+1} \bar{X}_{j+1} + \ldots + \hat{\beta}_k \bar{X}_k)$$
$$-\Phi(\hat{\beta}_0 + \hat{\beta}_1 \bar{X}_1 + \ldots + \hat{\beta}_{j-1} \bar{X}_{j-1} + \hat{\beta}_j 0 + \hat{\beta}_{j+1} \bar{X}_{j+1} + \ldots + \hat{\beta}_k \bar{X}_k)$$

- Average of the variables $\mathbf{X}$

Goodness of fit

- Goodness of fit
- Correctly predicted outcomes
- Deviance
- Pseudo R^2

Test hypothesis on parameters

- For individual effects
  - t-test
- For q linear restrictions
  - F-test not applicable
  - Wald test
  - Likelihood ratio (LR) or Deviance
  - Lagrange multiplier (LM)

Likelihood ratio

The unrestricted model:

$$\mathbf{y} = G(\beta_0 + \beta_1 \mathbf{x}_1 + \ldots + \beta_k \mathbf{x}_k) + \boldsymbol{\epsilon}$$

The restricted model ($H_0 : \beta_1 = \beta_5 = \beta_k = 0$, three restrictions $q = 3$)

$$\mathbf{y} = G(\beta_0 + \beta_2 \mathbf{x}_2 + \ldots + \beta_4 \mathbf{x}_4 + \beta_6 \mathbf{x}_6 + \ldots + \beta_{k-1} \mathbf{x}_{k-1}) + \boldsymbol{\epsilon}$$

$$Deviance = LR \equiv 2[\ell_{unres} - \ell_{res}] \sim \chi^2_q$$

Large LR value $\Rightarrow$ evidence against H_0

In R, deviance test: anova(model.rest, model.unres, test="chisq", lower.test=F)

Wald test

$$\mathbf{y} = \beta_0 + \beta_1 \mathbf{x}_1 + \ldots + \beta_5 \mathbf{x}_5 + \boldsymbol{\epsilon}$$

and we are testing

$$H_0 : \beta_2 = \beta_5 = 0$$

This can be written:

$$\underbrace{\begin{pmatrix} 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}}_{R} \overbrace{\begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \\ \beta_5 \end{pmatrix}}^{\beta} = \underbrace{\begin{pmatrix} 0 \\ 0 \end{pmatrix}}_{r}$$

$$H_0 : R\beta = r$$

Wald test

The test statistics:

$$\mathbf{W} = (\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{R}\boldsymbol{\beta})'(\mathbf{R}\hat{\mathbf{V}}\mathbf{R})^{-1}(\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{R}\boldsymbol{\beta}) \sim^{a} \chi^2_q$$

where $\hat{\mathbf{V}}$ is the estimated asymptotic variance of $\hat{\boldsymbol{\beta}}$.

Large $\mathbf{W}$ value $\Rightarrow$ evidence against H_0

Lagrange multipliers test

- Let us have q restrictions (for example in $H_0 : \beta_2 = \beta_5 = 0$, q=2)
- The estimates of the unrestricted model are $\hat{\beta}$
- The estimates of the restricted model are $\tilde{\beta}$
- Test statistics

$$LM = \left(\sum_{j=1}^{n} \tilde{s}_j \right)' \tilde{V}^{-1} \left(\sum_{j=1}^{n} \tilde{s}_j \right) \sim \chi^2_q$$

- The estimate $\tilde{V}$ can be chosen amongst

$$\sum \tilde{s}_j \tilde{s}_j' \quad \text{or} \quad -\sum \tilde{H}_j \quad \text{or} \quad \sum \tilde{A}_j$$

LM test

There is a special way of getting this test for the Index models.

- We have q restrictions
- The restricted model is $\mathbf{y} = G(\mathbf{V}\boldsymbol{\beta}) + \boldsymbol{\eta}$
- $\mathbf{V}$ contains the $k + 1 - q$ variables of $\mathbf{X}$ after removing the q restrictions
- $\mathbf{Z}$ contains the q variables of $\mathbf{X}$ that we removed for the restricted model
- $\mathbf{X} = (\mathbf{V}, \mathbf{Z})$
- For example: $\mathbf{y} = \beta_0 + \beta_1\mathbf{x}_1 + \beta_2\mathbf{x}_2 + \beta_3\mathbf{x}_3 + \beta_4\mathbf{x}_4 + \beta_5\mathbf{x}_5 + \boldsymbol{\epsilon}$ and we want to test $H_0 : \beta_2 = \beta_4 = 0$ then
  - $V = (1, X_1, X_3, X_4)$ and $Z = (X_2, X_5)$

LM test

1. Estimate the restricted model and get the residuals $\tilde{\eta} = y_j - G(V_j\tilde{\beta})$
2. Construct the standarised residual

$$\tilde{r} = \frac{\tilde{\eta}}{\sqrt{\tilde{G}_j(1-\tilde{G}_j)}} \quad \text{where } \tilde{G}_j = G(V_j\tilde{\beta})$$

3. Do the regressions

$$\tilde{r}_j = \frac{\tilde{G}'_j}{\sqrt{\tilde{G}_j(1-\tilde{G}_j)}} V_j + \frac{\tilde{G}'_j}{\sqrt{\tilde{G}_j(1-\tilde{G}_j)}} Z_j + u_j$$

LM test

- We regress the residuals of the restricted model (without Z) on the standarised V and Z variables.
- If there is no endogeneity, the V cannot explain the residuals
- If H_0 is true, the Z cannot explain the residuals
- We take the R^2 of the last regression and calculate the LM test statistics
- LM statistics is calculated as $n * R^2 \sim \chi^2_q$
- If LM is large then we will reject H_0

Model Misspecification

- Heterogeneity
- Non-normality
- Endogeneity

Omitted variables - neglected heterogeneity

If the omitted variables are exogeneous, we are neglecting heterogeneity.

Let us think of the latent variable model:

$$y^* = \mathbf{X}\beta + \gamma c + \epsilon \quad \epsilon|\mathbf{X}, c \sim N(0,1)$$

where c is an unobserved variable with mean zero and variance τ^2.

The probit model: $P(y = 1|\mathbf{X}, c) = \Phi(\mathbf{X}\beta + \gamma c)$

- Case 1: c and $\mathbf{X}$ are independent
- Case 2: c and $\mathbf{X}$ are dependent

Case 1: c, X independent

- The error term in the latent model is $\gamma c + \epsilon$ with mean zero and variance $\sigma^2 = \gamma^2\tau^2 + 1$
- So when we ommit γc from the model, we are estimating β/σ without knowing
- This is a problem with the probit model, due to its nature
- Similar problem when $E(\epsilon^2) \neq 1$
- Is this a great problem? Not so much, because β is not our partial effects.
- Structural partial effects do not have this problem either.

Case 1: c, X independent

- The partial effect of variable $\mathbf{x}_j$ is:

$$\frac{\partial P(y=1|\mathbf{X},c)}{\partial \mathbf{x}_j} = \phi(X\beta + \gamma c)\beta_j$$

- But because c is unknown, we estimate:

$$\phi\left(X\frac{\beta}{\sigma}\right)\frac{\beta_j}{\sigma}$$

which can be proven to be the APE across the population c,

$$E_c[\phi(X^0\beta + \gamma c)\beta_j] = \phi\left(X^0\frac{\beta}{\sigma}\right)\frac{\beta_j}{\sigma}$$

where X^0 is a fixed value of $\mathbf{X}$.

Case 1: c, X independent

In summary, the omitted variable problem, when this variable is independent of $\mathbf{X}$:

- The partial effects of certain $\mathbf{x}_j$ cannot be estimated consistently with the probit model
- The APE are consistently estimated by **probit** (if c is normally distributed)

Case 2: c, X dependent - endogeneity

- If c is correlated with $\mathbf{X}$ or dependent in any other way: omission of c is a serious problem.
- We cannot get consistent estimates of the APE
- We can find instruments z and run 2SLS on the LPM
- Rivers and Vuong's 2-step approach.