The Consequences of Unobserved Heterogeneity in a Sequential Logit Model

Maarten L. Buis

Institut für Soziologie
Eberhard Karls Universität Tübingen
maarten.buis@ifsoz.uni-tuebingen.de

The aim of this talk is

The aim of this talk is

- introduce the Mare model, and the problem of unobserved heterogeneity,

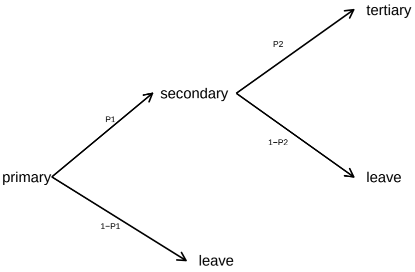
The aim of this talk is

- introduce the Mare model, and the problem of unobserved heterogeneity,
- assess how sensitive the conclusions from this model are to unobserved heterogeneity, and

The aim of this talk is

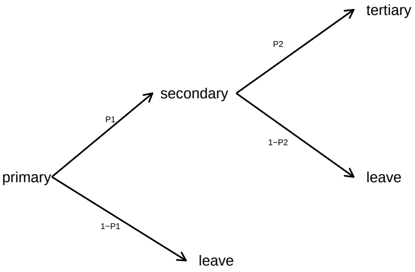
- introduce the Mare model, and the problem of unobserved heterogeneity,
- assess how sensitive the conclusions from this model are to unobserved heterogeneity, and
- introduce Stata

The sequential logit model



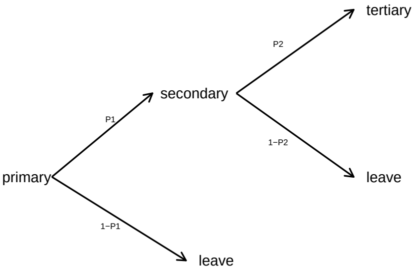
- A sequential logit model models a sequence of transitions.

The sequential logit model



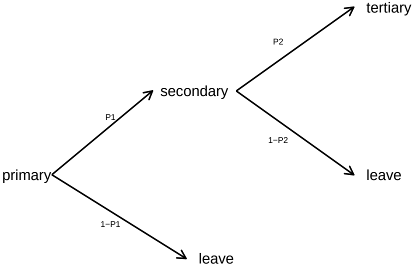
- A sequential logit model models a sequence of transitions.
- Each transition is modeled as a (multinomial) logistic regression using the sample which is 'at risk'.

The sequential logit model



- A sequential logit model models a sequence of transitions.
- Each transition is modeled as a (multinomial) logistic regression using the sample which is 'at risk'.
- $p_1 = \Lambda(\beta_{01} + \beta_{11} x + \beta_{21} z)$
- $p_2 = \Lambda(\beta_{02} + \beta_{12} x + \beta_{22} z)$

The sequential logit model



- A sequential logit model models a sequence of transitions.
- Each transition is modeled as a (multinomial) logistic regression using the sample which is 'at risk'.
- $p_1 = \Lambda(\beta_{01} + \beta_{11}x + \beta_{21}z)$
- $p_2 = \Lambda(\beta_{02} + \beta_{12}x + \beta_{22}z)$
- $\Lambda(u) = \frac{\exp(u)}{1+\exp(u)}$

Estimation

- Relatively simple

Estimation

- Relatively simple
- Create an indicator variable indicating whether or not someone has more than primary education.

Estimation

- Relatively simple
- Create an indicator variable indicating whether or not someone has more than primary education.
- Create an indicator variable indicating whether or not someone has more than secondary education

Estimation

- Relatively simple
- Create an indicator variable indicating whether or not someone has more than primary education.
- Create an indicator variable indicating whether or not someone has more than secondary education , but has a missing value when someone has only primary education (not at risk).

Estimation

- Relatively simple
- Create an indicator variable indicating whether or not someone has more than primary education.
- Create an indicator variable indicating whether or not someone has more than secondary education , but has a missing value when someone has only primary education (not at risk).
- Use standard logistic regression with these indicator variables as dependent variable.

Outline

The problem with unobserved variables

- The unobserved variable(s) could be confounding variables.

The problem with unobserved variables

- The unobserved variable(s) could be confounding variables.
- Even if the unobserved variable(s) are not confounding variables, they will still influence the results through 2 mechanisms:

The problem with unobserved variables

- The unobserved variable(s) could be confounding variables.
- Even if the unobserved variable(s) are not confounding variables, they will still influence the results through 2 mechanisms:
  - The Averaging Mechanism (scale identification problem)
  - The Selection Mechanism (dynamic selection bias)

The Averaging Mechanism

- leaving a variable out means averaging the probability over this variable

The Averaging Mechanism

- leaving a variable out means averaging the probability over this variable
- So, if we think that the true model is:

$$Pr(pass) = \Lambda(\beta_0 + \beta_1 x + \beta_2 z)$$

The Averaging Mechanism

- leaving a variable out means averaging the probability over this variable
- So, if we think that the true model is:

$$Pr(pass) = \Lambda(\beta_0 + \beta_1 x + \beta_2 z)$$

- and we cannot observe z, then we should use:

$$E_z[Pr(pass)] = E_z[\Lambda(\beta_0 + \beta_1 x + \beta_2 z)]$$

The Averaging Mechanism

- leaving a variable out means averaging the probability over this variable
- So, if we think that the true model is:

$$Pr(pass) = \Lambda(\beta_0 + \beta_1 x + \beta_2 z)$$

- and we cannot observe z, then we should use:

$$E_z[Pr(pass)] = E_z[\Lambda(\beta_0 + \beta_1 x + \beta_2 z)]$$

- Because $\Lambda()$ is a non-linear transformation, this is not the same as a simple logistic regression excluding z:

$$E_z[Pr(pass)] \neq \Lambda(\beta_0 + \beta_1 x + E_z[\beta_2 z])$$

The Averaging Mechanism

- leaving a variable out means averaging the probability over this variable
- So, if we think that the true model is:

$$Pr(pass) = \Lambda(\beta_0 + \beta_1 x + \beta_2 z)$$

- and we cannot observe z, then we should use:

$$E_z[Pr(pass)] = E_z[\Lambda(\beta_0 + \beta_1 x + \beta_2 z)]$$

- Because $\Lambda()$ is a non-linear transformation, this is not the same as a simple logistic regression excluding z:

$$E_z[Pr(pass)] \neq \Lambda(\underbrace{\beta_0^*}_{=\beta_0 + E_z[\beta_2 z]} + \beta_1 x)$$

The Selection Mechanism

- At higher transitions the sample at risk is a selected sample

The Selection Mechanism

- At higher transitions the sample at risk is a selected sample
- This selection is likely to produce a negative correlation between the observed and unobserved variables

The Selection Mechanism

- At higher transitions the sample at risk is a selected sample
- This selection is likely to produce a negative correlation between the observed and unobserved variables
- This means that the unobserved variable is likely to become a confounding variable at higher transitions, even if it was not one at the first transition

Outline

The problem

A solution

An application

Sensitivity analysis

- The aim of a sensitivity analysis is to show what would happen to the estimates of our model under a range of scenarios concerning unobserved heterogeneity.

Sensitivity analysis

- The aim of a sensitivity analysis is to show what would happen to the estimates of our model under a range of scenarios concerning unobserved heterogeneity.
- These scenarios are not intended to be true, but together they are meant to show what unobserved heterogeneity could do to the estimates.

Sensitivity analysis

- The aim of a sensitivity analysis is to show what would happen to the estimates of our model under a range of scenarios concerning unobserved heterogeneity.
- These scenarios are not intended to be true, but together they are meant to show what unobserved heterogeneity could do to the estimates.
- A sensitivity analysis is intended to show which conclusions are robust and which are not.

Scenarios concerning unobserved heterogeneity

- There is not one unobserved variable but many.

Scenarios concerning unobserved heterogeneity

- There is not one unobserved variable but many.
- So, if we have one observed variable x, the true model is for transition k is:
  $p_k = \Lambda(\beta_{0k} + \beta_{1k}x + \underbrace{\gamma_{1k}z_1 + \gamma_{2k}z_2 + \cdots + \gamma_{lk}z_l}_{\varepsilon})$

Scenarios concerning unobserved heterogeneity

- There is not one unobserved variable but many.
- So, if we have one observed variable x, the true model is for transition k is:
$p_k = \Lambda(\beta_{0k} + \beta_{1k}x + \underbrace{\gamma_{1k}z_1 + \gamma_{2k}z_2 + \cdots + \gamma_{lk}z_l}_{\varepsilon})$
- Basic scenario:
  - ε is normally distributed at the first transition.

Scenarios concerning unobserved heterogeneity

- There is not one unobserved variable but many.
- So, if we have one observed variable x, the true model is for transition k is:
  $p_k = \Lambda(\beta_{0k} + \beta_{1k}x + \underbrace{\gamma_{1k}z_1 + \gamma_{2k}z_2 + \cdots + \gamma_{lk}z_l}_{\varepsilon})$
- Basic scenario:
  - ε is normally distributed at the first transition.
  - The scenario is represented by an initial standard deviation of ε and an initial correlation between x and ε

Scenarios concerning unobserved heterogeneity

- There is not one unobserved variable but many.
- So, if we have one observed variable x, the true model is for transition k is:
  $p_k = \Lambda(\beta_{0k} + \beta_{1k}x + \underbrace{\gamma_{1k}z_1 + \gamma_{2k}z_2 + \cdots + \gamma_{lk}z_l}_{\varepsilon})$
- Basic scenario:
  - ε is normally distributed at the first transition.
  - The scenario is represented by an initial standard deviation of ε and an initial correlation between x and ε
  - This standard deviation can be seen as an constant effect of a standardized version of ε.

Scenarios concerning unobserved heterogeneity

- There is not one unobserved variable but many.
- So, if we have one observed variable x, the true model is for transition k is:
  $p_k = \Lambda(\beta_{0k} + \beta_{1k}x + \underbrace{\gamma_{1k}z_1 + \gamma_{2k}z_2 + \cdots + \gamma_{lk}z_l}_{\varepsilon})$
- Basic scenario:
  - ε is normally distributed at the first transition.
  - The scenario is represented by an initial standard deviation of ε and an initial correlation between x and ε
  - This standard deviation can be seen as an constant effect of a standardized version of ε.
- Extensions:

Scenarios concerning unobserved heterogeneity

- There is not one unobserved variable but many.
- So, if we have one observed variable x, the true model is for transition k is:

$$p_k = \Lambda(\beta_{0k} + \beta_{1k}x + \underbrace{\gamma_{1k}z_1 + \gamma_{2k}z_2 + \cdots + \gamma_{lk}z_l}_{\varepsilon})$$

- Basic scenario:
  - ε is normally distributed at the first transition.
  - The scenario is represented by an initial standard deviation of ε and an initial correlation between x and ε
  - This standard deviation can be seen as an constant effect of a standardized version of ε.
- Extensions:
  - These "effects" can change over transitions. (Hauser and Andrew 2006)

Scenarios concerning unobserved heterogeneity

- There is not one unobserved variable but many.
- So, if we have one observed variable x, the true model is for transition k is:

  $$p_k = \Lambda(\beta_{0k} + \beta_{1k}x + \underbrace{\gamma_{1k}z_1 + \gamma_{2k}z_2 + \cdots + \gamma_{lk}z_l}_{\varepsilon})$$

- Basic scenario:
  - ε is normally distributed at the first transition.
  - The scenario is represented by an initial standard deviation of ε and an initial correlation between x and ε
  - This standard deviation can be seen as an constant effect of a standardized version of ε.
- Extensions:
  - These "effects" can change over transitions. (Hauser and Andrew 2006)
  - ε may be non-normally distributed

Outline

The problem

A solution

An application