TU-Berlin

Microeconometrics

3^{rd} Tutorial: Binary Models (Introduction), Maximum Likelihood Estimation

Tomas Repasky

WS 2016/17

Technische Universität Berlin
Institut für Volkswirtschaftslehre und Wirtschaftsrecht
FG Ökonometrie und Wirtschaftsstatistik

... Where are we?

- We discussed Pros $\&$ Cons of non-parametric estimation which looks at the data with very few assumptions.
- Data do not speak for themselves, hence, sometimes we need to impose a structure to uncover the relationship.
  - Particularly in cases when the decision depends on multiple characteristics.
- We discussed Linear Probability Model, its interpretation and the consequences of imposing the respective structure on a binary variable.

The (dependent) variable in the `mroz` dataset was `infl` surveyed as follows:

$$Y_i = \begin{cases} 1 & \text{, if the individual } i \text{ was working} \\ 0 & \text{, if it was not} \end{cases}$$

Hence the distribution of Y_i is: Bernoulli Distribution

The respective probabilities are given by:

$$f(y_i = 1) = \pi$$
$$f(y_i = 0) = 1 - \pi$$

We are assuming that the probability π is some function of the observable characteristics

$$f(y_i = 1) = \pi = g(x_i)$$
$$f(y_i = 1) = 1 - \pi = 1 - g(x_i)$$

Interpretation

Expectation of the Benoulli (random) Variable

$$\begin{aligned}\mathbb{E}[Y_i|X=x_i] &= 1 * \mathbb{P}[Y_i = 1|X=x_i] + 0 * \mathbb{P}[Y_i|X=x_i] \\ &= \mathbb{P}[Y_i = 1|X=x_i] = \pi(x_i) = g(x_i)\end{aligned}$$

We sometimes abbreviate $X = x_i$ as x_i e.g. $\mathbb{E}[Y_i|X=x_i]$ as $\mathbb{E}[Y_i|X_i]$.

- Holds only for Bernoulli random Variable.
- We have already used this fact in the interpretation of Linear Probability Model (How?).

Unconditional Probability

How can we interpret the following summary statistics in terms of probabilities?

OUTPUT OF THE COMMAND `table()`

	not working	working	missing (NA)
Nr. of Observations	325	428	0

Conditional Probabilities

PRODUCT OF THE COMMAND **CrossTable** FROM THE **dscr** PACKAGE

	Number of kids at least 6 y. old				
mroz$inlf	0	1	2	3	Total
not working	231	72	19	3	325
	0.381	0.610	0.731	1.000	
working	375	46	7	0	428
	0.619	0.390	0.269	0.000	
Total	606	118	26	3	753

Estimation I

- Conditional probabilities seems to change with changing x_i (look at the Table on the previous page). $=>$ Characteristics x_i potentially influence the decision to work or not to work.
- We could assume LPM. But the downside is that predicted probabilities are not constrained to the $[0-1]$ interval.
- Any solutions?

Estimation II

- We could transform the LPM by some function $G(\mathbf{x}'\boldsymbol{\beta})$, therefore ensuring the resulting probabilities are between 0 and 1.

  More formally, we could assume that:

$$\begin{aligned}\mathbb{E}[Y_i|X = x_i] = \mathbb{P}[Y_i = 1|X = x_i] &= G(\mathbf{x}_i'\boldsymbol{\beta}) \\ &= G(\beta_0 + \beta_1 x_{1i} + ... + \beta_k x_{ki})\end{aligned}$$

Alternative Formulation

Still, there are some questions left

- What are the consequences of the additional assumptions?
- Why do we still assume the same linear-in-parameter function as before inside the $G(.)$ function?

To answer these questions we need an alternative formulation of Probit and Logit, through Latent Variable.

Alternative Formulation II

- The following derivation can be found in Wooldridge (p.576-77)
- Describe all steps of the derivation in your own words.

$$y^* = \beta_0 + \mathbf{x}'\boldsymbol{\beta} + \varepsilon \tag{1}$$

$$y = \begin{cases} 1 & [y^* > 0] \\ 0 & \textit{otherwise} \end{cases} \tag{2}$$

$$\begin{aligned} \mathbb{E}[Y_i|\mathbf{x}] &= \mathbb{P}[Y_i = 1|\mathbf{x}] = \mathbb{P}[y_i^* > 0|\mathbf{x})] = \mathbb{P}[\varepsilon_i > -(\beta_0 + \mathbf{x}_i'\boldsymbol{\beta})|\mathbf{x}] \\ &= 1 - G[-(\beta_0 + \mathbf{x}_i'\boldsymbol{\beta})] = G(\beta_0 + \mathbf{x}_i'\boldsymbol{\beta}) = \Phi(\beta_0 + \mathbf{x}_i'\boldsymbol{\beta}) \end{aligned} \tag{3}$$

- How would the derivation of Logit differ? (only For Your Information)

Latent Variable in Probit

Only for your information = not relevant for the exam (henceafter "FYI")

- Do we have to assume that $\varepsilon|\mathbf{x} \sim N(0,1)$ or can we assume that $\varepsilon|\mathbf{x} \sim N(a,b^2)$, where a,b are unknown constants?

Latent Variable in Probit II - FYI

Whatever we assume, we cannot differentiate between the two cases defined in the previous slide!

$$\begin{aligned}
\mathbb{P}[Y_i = 1|\mathbf{x}] &= \mathbb{P}[y_i^* > 0|\mathbf{x})] = \mathbb{P}[(\beta_0 + \mathbf{x}_i'\boldsymbol{\beta} + \varepsilon_i) > 0|\mathbf{x}] \\
&= \mathbb{P}[\varepsilon_i > -(\beta_0 + \mathbf{x}_i'\boldsymbol{\beta})|\mathbf{x}] \\
&= \mathbb{P}[\frac{\varepsilon_i - \mu_\varepsilon}{\sigma_\varepsilon} > -(\frac{\beta_0 - \mu_\varepsilon}{\sigma_\varepsilon} + \mathbf{x}_i'\frac{\boldsymbol{\beta}}{\sigma_\varepsilon})|\mathbf{x}] \\
&= \mathbb{P}[\xi_i > -(\frac{\beta_0 - \mu_\varepsilon}{\sigma_\varepsilon} + \mathbf{x}_i'\frac{\boldsymbol{\beta}}{\sigma_\varepsilon})|\mathbf{x}] = \Phi(\frac{\beta_0 - \mu_\varepsilon}{\sigma_\varepsilon} + \mathbf{x}_i'\frac{\boldsymbol{\beta}}{\sigma_\varepsilon})
\end{aligned}$$

What does that mean in the context of the estimates?

Latent Variable in Probit III - FYI

- It means we cannot differentiate between the cases ξ & ε
- If we assume $\varepsilon|\mathbf{x} \sim N(0,1)$ then we have identified β_0 & $\boldsymbol{\beta}$.
- If we assume $\varepsilon|\mathbf{x} \sim N(a,b^2)$ then we have identified $\frac{\beta_0 - a}{b}$ & $\frac{\boldsymbol{\beta}}{b}$
- But you can get only one set of numbers out of the estimation! So you cannot know whether you have just estimated $\boldsymbol{\beta}$ or $\frac{\boldsymbol{\beta}}{b}$.
- In sum, we cannot simultaneously identify the magnitude of σ_ε and $\beta_0, \boldsymbol{\beta}$.
- Therefore, we standardised $\varepsilon|\mathbf{x} \sim N(0,1)$

(!) Remember, you don't have to memorize any derivations, you just need to understand the issues addressed by the derivation!

Estimation I

- How can we estimate the parameters given by: $\mathbb{E}[Y_i|\mathbf{x}_i'\boldsymbol{\beta}] = \mathbb{P}[Y_i = 1|\mathbf{x}_i'\boldsymbol{\beta}] = \Phi(\mathbf{x}_i'\boldsymbol{\beta})$?
- We cannot use OLS, since $\Phi(.)$ is a non-linear function.

Maximum Likelihood Estimation I

"The likelihood principle, due to R.A.Fischer(1922), is to choose as estimator of the parameter vector θ_0 *that value* θ *that maximizes the likelihood of observing the actual sample"* [under the assumed distribution].

Cameron & Trivedi, p. 139

Maximum Likelihood Estimation II

From the Wooldridge's Book

$$f(y|\mathbf{x}_i, \boldsymbol{\beta}) = [G(\mathbf{x}_i'\boldsymbol{\beta})]^y * [1 - G(\mathbf{x}_i'\boldsymbol{\beta})]^{1-y}, \quad y = \{0, 1\} \tag{4}$$

$$\ell_i(\boldsymbol{\beta}) = y_i * log[G(\mathbf{x}_i'\boldsymbol{\beta})] + (1 - y_i) * log[1 - G(\mathbf{x}_i'\boldsymbol{\beta})] \tag{5}$$

$$\mathscr{L} = \sum_{i=1}^{N} \ell_i(\boldsymbol{\beta}) \tag{6}$$

State explicitly, which assumptions (if any) were made to formulate equation (4), (5), (6). Be precise!

MLE-Example

MLE of π

1. Your task is to compute the Maximum Likelihood Estimator of $\pi := \mathbb{P}(Y_i = 1)$ of a bernoulli random variable Y_i.
   - Use the information from the previous slides.
2. Explain the difference between unconditional and conditional probability. Come up with a simple example to illustrate the difference.
3. Make sure that you have computed the maximum! (By checking the second order condition (SOC)).

Solution I

The respective Likelihood function looks like:

$$L(\pi, \mathbf{y}) = \prod_{i=1}^{N} f(y_i|\pi) = \prod_{i=1}^{N} \pi^{y_i}(1-\pi)^{1-y_i}$$

The Log-Likelihood:

$$\mathscr{L}(\pi, \mathbf{y}) = \sum_{i=1}^{N} \{y_i * log(\pi) + (1-y_i) * log(1-\pi)\}$$

The respective FOC:

$$\frac{\partial \mathscr{L}(\pi, \mathbf{y})}{\partial \pi} \overset{!}{=} 0$$

FOC - Solution

Left-Hand-Side:

$$\sum_{i=1}^{N} \frac{\partial(y_i log\pi + (1-y_i)log(1-\pi))}{\partial\pi} = \sum_{i=1}^{N} y_i \frac{1}{\pi} + (1-y_i)\frac{1}{1-\pi}(-1)$$

$$= \sum_{i=1}^{N} \frac{y_i(1-\pi) - (1-y_i)\pi}{\pi(1-\pi)} = \sum_{i=1}^{N} \frac{y_i - \pi}{\pi(1-\pi)}$$

putting LHS and RHS together:

$$\frac{\sum_{i=1}^{N} y_i - n\hat{\pi}}{\hat{\pi}(1-\hat{\pi})} = 0$$

$$\frac{1}{N}\sum_{i=1}^{N} y_i = \hat{\pi} = \bar{y}$$

Maximum or Minimum?

We need to check the SOC:

$$\frac{\partial^2 \left(\sum_{i=1}^{N} y_i * log\pi + (1 - y_i * log(1-\pi)) \right)}{\partial^2} =$$

$$\frac{\partial \left(\sum_{i=1}^{N} y_i \frac{1}{\pi} + (1 - y_i) \frac{1}{1-\pi} (-1) \right)}{\partial \pi} =$$

$$\sum_{i=1}^{N} -\frac{y_i}{\hat{\pi}^2} - \frac{(1-y_i)}{(1-\hat{\pi})^2} < 0$$

$\hat{\pi} = \bar{y}$ is a maximum

Main Conclusions from the Discussion in the Tutorials I

1. In the LPM, we have to use heteroscedasticity robust variance-covariance matrix estimation, because the (bernoulli) dependent random variable features heteroscedastic error term.
    - This we know for a fact so we do not have to carry out any formal test for heteroscedasticity.
    - In the LPM, the estimators of the coefficients remain consistent.
2. The situation is very different with the latent variable.
    - If we could observe the latent variable, we could use OLS estimation, since all OLS-Assumptions are fulfilled.
    - We cannot observe it, however, and Probit estimation based on the observed (y) is inconsistent in the presence of heteroscedastic latent error.

Main Conclusions from the Discussion in the Tutorials II

3. The heteroscedasticity of the (observed) bernoulli random variable is no challenge for the standard MLE-Probit estimation, since MLE uses the PDF of the bernoulli distribution and the hereroscedasticity of the error term is just a feature of that distribution.

Please Note

For more details look into the R-Script, where you have the chance to rerun the analysis, look at the graphs again and read the respective commentary.