Applied Health Economics (for B.Sc.)

Helmut Farbmacher

Department of Economics
University of Mannheim

Autumn Semester 2017

Outlook

1 Linear models (OLS, Omitted variables, 2SLS)

2 Limited and qualitative dependent variables

3 Organizational issues (Timeline, Topics, etc.)

1. Linear models

1.1 Main Concepts in Econometric Analysis

1.2 The OLS Estimator

1.3 The 2SLS Estimator

1.1 Main Concepts in Econometric Analysis

Introduction

- In economics we are often interested in causal effects. For instance, what is the causal effect of school education x on later income y.

- Ideally, we would have an experiment, where the researcher could manipulate x and then measure y. Then the causal effect can be easily estimated as

$$E(y|x = \text{good educ}) - E(y|x = \text{bad educ}) \quad \text{if } x \text{ is binary/discrete}$$

$$\frac{\partial\, E(y|x)}{\partial\, x} \quad \text{if } x \text{ is continuous}$$

1.1 Main Concepts in Econometric Analysis

- In economics we often do not have experimental data.
- Estimating causal effects from observational data is harder. We could, for instance,
  - control for **sufficiently** many other factors (ceteris paribus analysis)
  - use instrumental variables
  - use panel data estimators
  - ...

1.1 Main Concepts in Econometric Analysis

The Population Model

- We assume that our data come from a random sample of the relevant population.

- Suppose the observations ($i = 1, \ldots, n$) in our dataset have been generated by the following population model ("data generating process" - DGP)

$$y_i = f(x_i, u_i, \beta) = \beta_0 + \beta_1 \, x_{1i} + \beta_2 \, x_{2i} + \ldots + \beta_p \, x_{pi} + u_i = x_i' \beta + u_i$$

- Stacking the observations, we can also write $y = X\beta + u$. Here, y and u are $n \times 1$ vectors and X is a $n \times p$ matrix.

1.1 Main Concepts in Econometric Analysis

- Consider our n realizations of the population model

$$y_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_k x_{ik} + u_i$$

which can be written as

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & \ldots & x_{1k} \\ 1 & x_{21} & \ldots & x_{2k} \\ \vdots & \vdots & & \vdots \\ 1 & x_{n1} & \ldots & x_{nk} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{pmatrix} + \begin{pmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{pmatrix}$$

or simply

$$y = X\beta + u$$

1.1 Main Concepts in Econometric Analysis

- Remarks:
  - β is a $p \times 1$ vector of unknown parameters, which we want to estimate.
  - Because $f(x_i, u_i, \beta) = x_i'\beta + u_i$ we know that the true functional form is linear. In practice, we don't know the true DGP and it could be far more complicated. Giving rise to semi- or non-parametric estimation techniques.
  - The population model does not necessarily have to be the model we are going to estimate (e.g., because of functional form misspecification, one of the x's is not in our data or measured with error,...).

1.1 Main Concepts in Econometric Analysis

Conditional Expectations

- In regression analysis we estimate conditional expectations.
- In the previous model, the conditional expectation is

$$E(y|X) = E(X\beta + u|X) = E(X|X)\beta + E(u|X) = X\beta + E(u|X)$$

- Remarks:
  - The second equality follows as the expectation operator is linear and β is a constant population parameter.
  - The third equality follows as $E(X|X) = X$.
  - Now, assuming $E(u|X) = 0$, we get $E(y|X) = X\beta$.

1.2 The OLS Estimator

Asymptotic properties

- We use asymptotic theory to derive
  - what can be learned from data when the sample is unlimited ("identification")
  - what can be concluded from a sample of finite size ("statistical inference")

- Assumption OLS.1: $E(X'u) = \underset{p\times 1}{0}$. This means that u is uncorrelated with each regressor.

1.2 The OLS Estimator

- A regressor x_j is called **endogenous** (in the econometric sense) if it is correlated with the error term (u), where $cov(x_j, u) \neq 0$ implies $E(x_j' u) \neq 0$. Otherwise, it is called **exogenous** and Assumption OLS.1 is fulfilled for x_j.

- Potential threats of identification
  - Omitted variables
  - Measurement error in observed variables
  - Reverse causality: Not only x affects y but y simultaneously affects x

1.2 The OLS Estimator

- Assumption OLS.2: rank$(X'X) = p$. Reason: $X'X$ is invertible only if it has full rank, i.e., there is no exact linear relationship among the regressors. Assumption OLS.2 is purely technical ruling out perfect multicollinearity. It is, for instance, not fulfilled if you
  - include k mutually exclusive dummy variables in the model (here we can only identify $k - 1$ parameters which are relative to kth category)
  - include both $log(age)$ and $log(age^2)$

1.2 The OLS Estimator

Omitted Variables

- Let's go back to our example: What is the effect of education on income?
- Suppose the **population** model is

$$y = \beta x + \gamma q + u\,.$$

  We assume that our dataset has been generated by this process. However, while x is observed, q is not. We are also willing to assume that $E(x'u) = 0$.

- Suppose we are interested in the parameter β.

1.2 The OLS Estimator

- As data on q is missing, the best we can do is to **estimate** the following model

$$y = \delta x + w$$

  where $w = \gamma q + u$ is a composite error term.

- What is the probability limit of the OLS estimator for δ?

- Can we say something about the potential direction of the bias?

1.2 The OLS Estimator

- We have to include q in our regression to avoid falsely attributing the explanatory power of q to x. Example: The parameter of education is biased if we do not control for ability.

- The estimator of β will be consistent **only** when either $Cov(x, q) = 0$ or $\gamma = 0$.

- Otherwise, it will suffer from omitted-variables bias.

1.3 The 2SLS Estimator

- We are still interested in the same population model

$$y = \beta x + \gamma q + u \text{ with } w = \gamma q + u$$

- We know OLS is inconsistent because of the correlation between x and q. Without additional information, we cannot estimate β consistently.

1.3 The 2SLS Estimator

- Let's assume there is an additional variable Z which satisfies the following conditions
  - $cov(w, Z) = 0$ (exogeneity assumption)
  - $cov(x, Z) \neq 0$ (relevance assumption)
- We call such a variable Z an instrumental variable (or just instrument).

1.3 The 2SLS Estimator

- Denote the number of available instruments m and the number of endogenous variables k, then
  - a model is called "just-identified" if $m = k$
  - a model is called "over-identified" if $m > k$

- More formally the assumptions to make 2SLS consistent are
  - Assumption 2SLS.1: $E(Z'w) = \underset{m \times 1}{0}$
  - Assumption 2SLS.2a: $\text{rank}(Z'x) = k$
  - Assumption 2SLS.2b: $\text{rank}(Z'Z) = m$

1.3 The 2SLS Estimator

- The two-stage least squares (2SLS) estimator can be derived in two stages.
  - First regress the endogenous variable x on Z:

$$x = Z\widehat{\pi} + \widehat{v}$$

    $\widehat{\pi}$ is such that $cov(Z, \widehat{v}) = 0$.
  - Now, predict $\widehat{x} = Z\widehat{\pi}$ and regress $\widehat{x}$ on y

$$y = [Z\widehat{\pi} + \widehat{v}]\beta + \gamma q + u = Z\widehat{\pi}\beta + \widehat{v}\beta + \gamma q + u$$

1.3 The 2SLS Estimator

- Remarks: We only used the OLS estimation technique to derive the 2SLS estimator! Although the point estimates are identical using two stages, the SEs are wrong. Why?

- To test the relevance assumption, we can test whether all elements in π are jointly zero using an F-test.

- If the model is over-identified, we can also (partly) test the exogeneity assumption.

1.3 The 2SLS Estimator

Final remark: Proxy variables

- Proxy variables can also provide a solution to the omitted-variable problem (e.g., IQ test results as proxy for ability)

- However, there is an important difference between instruments and proxy variables:
  - While a proxy should be highly correlated with the omitted variable to account/"control" for this omitted variable,
  - the instrument has to be uncorrelated with the omitted variable. This is because the omitted variable is part of the error term and we want the instrument to be uncorrelated with the error term.

2. Limited and qualitative dependent variables

2.1 Introduction

2.2 Binary dependent variables

2.3 Ordered dependent variables

2.1 Introduction

Introduction

- Conceptually, setting up and estimating models with limited and qualitative (discrete) dependent variables involves four steps:
  1. Specify an **observation rule** for the dependent variable, y. Typically, this rule is based on a latent (unobserved) continuous outcome, y^*.
  2. Specify how the explanatory variables enter the model (typically, via a **linear index**, $X\beta$).s

2.1 Introduction

3. Specify how an **unobserved variable** enters and its distribution. Typically, the unobserved variable enters additively and is normally distributed (conditional on the explanatory variables).
4. Derive the **likelihood function** to estimate the parameters by maximum likelihood.

- In the remainder of this chapter, we will discuss two examples of this approach, which are often used in health economics.

2.2 Binary dependent variables

- Let $y \in \{0, 1\}$ be the dependent variable and $X = (x_1, x_2, \ldots, x_k)$ the k explanatory variables.

- One object of interest is the probability of observing $y = 1$ which is a (yet to be specified) function of the explanatory variables:

$$Pr(y = 1|X) = E(y|X)$$

2.2 Binary dependent variables

- Even more interesting is how this probability changes with the explanatory variables. The marginal effects for continuous explanatory variables x_j are

$$\frac{\partial Pr(y = 1|X)}{\partial x_j}$$

- If x_j is discrete, one can interpret the change in the probability as its partial effect:

$$Pr(y = 1|x_j = 1, X_{-j}) - Pr(y = 0|x_j = 1, X_{-j})$$

2.2 Binary dependent variables

The linear probability model (LPM)

- Ignore the binary nature of the dependent variable and estimate the model by OLS as if y were continuous. Thus, we write

$$E(y|X) = Pr(y = 1|X) = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k$$

- Problem: We can find values of X for which $Pr(y = 1|X) \notin [0, 1]$. Unless we estimate a saturated model.

- The linear probability model should best be considered as an approximation to the data.

2.2 Binary dependent variables

Index models for binary dependent variables

- Use a nonlinear function to model $Pr(y = 1|X)$, where the explanatory variables enter via a linear index, $X\beta$. Thus, we write

$$Pr(y = 1|X) = G(X\beta)$$

- The linear probability model has $G(X\beta) = X\beta \notin [0, 1]$.
- We are looking for functions G that satisfy $G(X\beta) \in [0, 1]$.
- Obvious candidates are cumulative distribution functions since they are limited to $[0, 1]$ by definition.

2.2 Binary dependent variables

- Index models can be made consistent with utility maximization via a latent variable, y^*, which we can interprete as a utility difference:

$$y^* = X\beta + e$$

- The observation rule is consistent with utility maximization if y^* is interpreted as the utility difference between options 0 and 1:

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

2.2 Binary dependent variables

- If we assume that the CDF of e is G, we get the density from the probabilities of the two outcomes:

$$f(y|X,\beta) = [G(X\beta)]^y[1 - G(X\beta)]^{1-y}$$

- Next, we derive the log-likelihood contribution of observation i:

$$\ell_i(\beta) = y_i \, ln \, G(\mathbf{x}_i'\beta) + (1 - y_i) \, ln(1 - G(\mathbf{x}_i'\beta))$$

2.2 Binary dependent variables

- The common choice for the CDF in health economics is the standard normal distribution (Probit model), $G(z) = \Phi(z) = \int_{-\infty}^{z} \phi(v)\mathrm{d}v$

- The log-likelihoood function is

$$\mathcal{L} = \sum_{i=1}^{N} \{y_i \, ln \, G(\mathbf{x}_i'\beta) + (1 - y_i) \, ln(1 - G(\mathbf{x}_i'\beta))\}$$

2.2 Binary dependent variables

- In practice, an important issue is how the value of the parameters (β's) should be interpreted.
- It is now common to report marginal effects rather than the parameters themselves.
- The marginal effect of a continuous regressor x_j on $Pr(y = 1|X)$ is

$$\begin{aligned} \frac{\partial Pr(y=1|X)}{\partial x_j} &= \beta_j && \text{(LPM)} \\ \frac{\partial Pr(y=1|X)}{\partial x_j} &= \phi(\mathbf{x}_i'\beta)\beta_j && \text{(Probit)} \end{aligned}$$

- The marginal effects of Probit needs to be evaluated at some specific value of $\mathbf{x}$

2.2 Binary dependent variables

- Usually, marginal effects are evaluated for each observation i and then the average is reported (average marginal effects)
- Alternatively, marginal effects are evaluated at the sample means of the elements of $\mathbf{x}$ (marginal effects at the average).
- `margins` is a special Stata command which computes marginal effects after non-linear estimation.

2.2 Binary dependent variables

Generalizations of binary dependent variables

- Models for binary choice can be generalized to situations in which the dependent variable takes **more than two values**. There are two cases:
  - **Multinomial dependent variables** that have no natural odering. The leading example is product choice (cars, transportation modes, etc.)
  - **Ordered dependent variables** that have a natural ordering, such as credit ratings, self-assessed health status, etc. They have no cardinal interpretation, however.

3. Organizational Issues

- You have to write a term paper (approx. 10 pages without figures and tables) based on assigned research articles.
- And you have to present (approx. 20 minutes) in the seminar:
  - Friday (10.11.) 1pm-7pm, room 003, L9, 1-2
  - Saturday (11.11.) 8am-11am, room P044, L7, 3-5
- The term paper has to be written between 2 October and 29 October (4 weeks).
- Grading: term paper (50%), presentation and active participation in the seminar (50%).

3. Organizational Issues

- You have to hand in the term paper **in pdf format** on 29 October 2017, via email.

- You have to hand in your presentation slides (preferably in pdf format) on 9 November 2017, 6pm (at the latest), by email to farbmacher@uni-mannheim.de.

3. Organizational Issues

Topics

- Mendelian randomization (in economic research)
- The income-health relationship
- Health and wages
- The demand for health and health care (count data models!)
- Evaluation of policy reforms in health economics

Send me an email with your first and second preference until Friday, **22 September**.