Econometrics I

Sylvia Frühwirth-Schnatter

Department of Finance, Accounting and Statistics
Vienna University of Economics and Business
WS 2012/13

Introductory Econometrics

- Milestone I: Basic Concepts of Econometric Modelling
- Milestone II: The Multiple Regression Model
- Milestone III: Advanced Multiple Regression Models
- Wooldridge, J.: Introductory Econometrics. Thompson South-Western, 2009.
- Hackl, Peter: Einführung in die Ökonometrie. Pearson Verlag, 2005.

Mile stone I

Basic Concepts of Econometric Modelling

- Step 1: What is econometric modelling?
- Step 2: Understanding common data structures
- Step 3: First steps in EViews
- Step 4: The simple regression model

I.1 Econometric Modelling

Econometrics deals with learning about a phenomenon (e.g. status of the economy, influence of product attributes, volatility on financial markets, wage mobility) from data

- **Econometric model**: description of the phenomenon involving quantities that are observable

- **Data** are collected for the observable variables

- **Econometric inference**: draw conclusions from the data about the phenomenon of interest

Econometric Modelling

Example: relationship between price and demand

- Description of the phenomenon involving quantities that are observable
- simplified description of the process behind the data based on a deterministic economic model;
- stochastic model rather than a deterministic model.

Deterministic Economic Model

Exact quantitative relationship between the variables of interest is assumed to be known

Example: Deterministic Relationship between Demand and Price

$$D = f(p),$$

where D is the demand and p is the price.
Linear model:

$$D = \beta_0 + \beta_1 p$$

Non-linear model:

$$D = \beta_0 p^{\beta_1}$$

Econometric Model

Exact quantitative relationship between the variables of interest is NOT known, but disturbed by a (stochastic) error term

Example: Stochastic Relationship between Demand and Price

$$D = f(p, u)$$

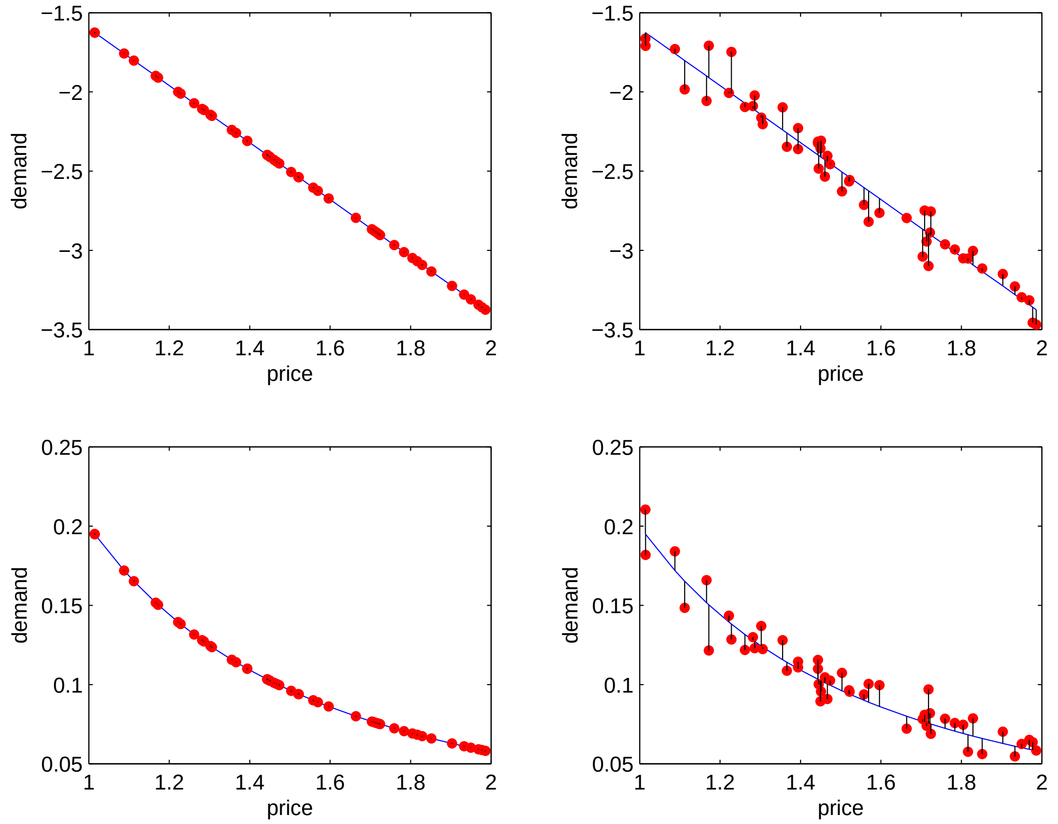
where D is the demand, p is the price, and u is an unobservable error.
Linear model:

$$D = \beta_0 + \beta_1 p + u$$

Non-linear model:

$$D = \beta_0 p^{\beta_1} u$$

Econometric Model



Where does the error come from?

- u aggregates variables, that are not included into the model because
  - their influence is not known apriori
  - these variables are unobservable or difficult to quantify
- u aggregates measurement errors which are caused by quantifying economic variables
- u captures the unpredictable randomness in the left hand side variable of the model

Where does the error come from?

To sum up, an econometric model consists of

- a structural part which describes how the variables are related, if there was no error;
- an error model which describes the properties of the error term.

Econometric Inference

Example: Relationship between Demand and Price
Estimate β_1 and β_2 from the linear model:

$$D = \beta_1 + \beta_2 p + u$$

or from the non-linear model:

$$D = \beta_1 p^{\beta_2} u$$

from data. For the second model β_2 is the price elasticity

Econometric Inference

Econometric inference is, in general, concerned with drawing conclusions from observed data about quantities that are unobserved.

Unobserved quantities:

- quantities that are not directly observable such as parameters that govern the process leading to the observed data, e.g. price elasticities

- potentially observable quantities such as future observations

- hypothesis about the process we observe

Econometric Inference

Due to this impossibility to observe these quantities of interest, any statement about these quantities will be uncertain, even in the light of the data one actually has observed.

- Classical inference: parameter estimation and hypothesis testing to deal with this uncertainty

- Bayesian inference: is based on the concept that the state of knowledge about any unknown quantity is best expressed in terms of a probability distribution.

Practical Econometric Inference

- Model formulation
- Model estimation
- Econometric inference: parameter estimation, hypothesis testing, forecasting
- Model choice
- Model checking

I.2 EViews

- Use a software package for practical econometric inference
- We will use EViews 7
- Detailed instruction on how to use EViews is given in the tutorial

I.3 Data Structure

Experimental data: data obtained through a designed experiment (medicine, travel time to university, ...) - rare in economics (and many other areas without laboratories) to have experimental data.

Non-experimental (observational) data:

- Cross-sectional data
- Time series data
- Panel data

Cross-sectional data

- we are interested in variables (Y, X) (e.g. relationship between demand D and price P) or a set of variables $(Y, X_1, \ldots, X_K)$
- we are observing these variables simultaneously for N subjects drawn randomly from a population (e.g. for various individual, firm, supermarkets, countries) at a point in time

Typically, cross sectional data are indexed as follows:

$$(y_i, x_i), \qquad (y_i, x_{1i}, \ldots, x_{Ki}), \quad i = 1, \ldots, N \tag{1}$$

If the data set is not a random sample, there is a sample-selection problem.

EVIEWS Exercise I.3.1

Demonstrate in EViews how cross-sectional data are organized

- Case Study profit, workfile `profit`;
- Case Study Chicken, workfile `chicken`;
- Case Study Marketing, workfile `marketing`;

Time Series Data

- we are interested in a single variable Y (e.g. the return of a financial asset);
- we are observing this variable over time (e.g. every month)
- data cannot be regarded as random sample; it is important to account for trends and seasonality

Typically, time series data are indexed as follows:

$$y_t, \quad t = 1, \ldots, T \tag{2}$$

EVIEWS Exercise I.3.2

Demonstrate in EViews how time series data are organized

- Case Study Stock Vienna Stocks, workfile `viennastocks`;
- Case Study Stock Returns, workfile `stockreturns`;
- Case Study Yields, workfile `yieldus`;

Panel Data

- Pooled cross-section: Random cross sections can be pooled and treated similar to normal cross section, accounting for differences over time.

- Panel data or longitudinal data: The same (random) individual observations Y_i is followed over time, i.e., we have a time series for each cross-section unit.

Typically, panel data are indexed as follows:

$$y_{it}, \qquad i = 1, \ldots, N, \qquad t = 1, \ldots, T \tag{3}$$

I.4 The Simple Regression Model

- Step 1: Model Formulation and basic assumptions
- Step 2: Ordinary least squares (OLS) estimation
- Step 3: The Log-linear Regression Model
- Step 4: Statistical properties of OLS

Cross-sectional data

- We are interested in a dependent (left-hand side, explained, response) variable Y, which is supposed to depend on an explanatory (right-hand sided, independent, control, predictor) variables X

- Examples: demand is a response variable and price is a predictor variable); wage is a response and years of education is a predictor variable

- Data: we are observing these variables for N subjects drawn randomly from a population (e.g. for various supermarkets, for various individuals): $(y_i, x_i), i = 1, \ldots, N$

I.4.1 Model formulation

The simple linear regression model describes the dependence between the variables X and Y as:

$$Y = \beta_0 + \beta_1 X + u. \tag{4}$$

The parameters β_0 and β_1 need to be estimated:

- β_0 is referred to as the constant or intercept
- β_1 is referred to as slope parameter.

Basic Assumptions

- The average value of the error term u in the population is 0 (not restrictive, we can always use β_0 to normalize $\mathrm{E}(u)$ to 0):

$$\mathrm{E}(u) = 0. \tag{5}$$

- A more crucial assumption is that

$$\mathrm{E}(u|X) = \mathrm{E}(u). \tag{6}$$

Basic Assumptions

- This means that the conditional mean of u is zero, i.e., knowing something about X does not give us any information about u.

- Assumption (6) implies:

$$\mathrm{E}(Y|X) = \beta_0 + \beta_1 X. \tag{7}$$

$\mathrm{E}(Y|X)$ is a linear function of X.

- For a fixed value of $X = x$, the distribution of $Y|X = x$ is centered about its conditional mean $\mathrm{E}(Y|X = x)$.

Understanding the regression model

- Simulate data from a simple regression model with $\beta_0 = 0.2$ and $\beta_1 = -1.8$:

$$Y = 0.2 - 1.8X + u, \tag{8}$$

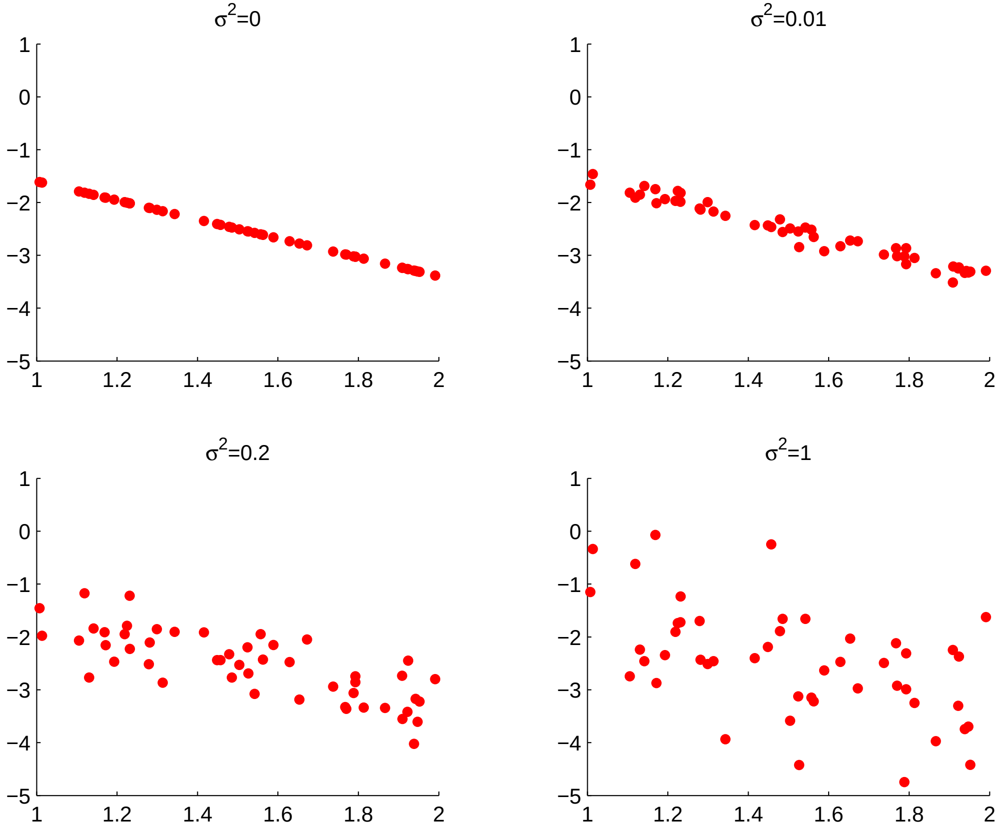
- Specification of the error term:

$$u \sim \mathrm{Normal}\left(0, \sigma^2\right) \tag{9}$$

- Demonstration $\Rightarrow$

MATLAB Code: `regsim.m`

Understanding the regression model



Understanding the parameters

Expected value of Y, given $X = x$:

$$\mathrm{E}(Y|X = x) = \beta_0 + \beta_1 x$$

Expected value of Y, if the predictor X is changed by 1:

$$\mathrm{E}(Y|X = x + 1) = \beta_0 + \beta_1(x + 1).$$

Thus β_1 is the expected absolute change of the response variable Y, if the predictor X is increased by 1:

$$E(\Delta Y|\Delta X = 1) = \mathrm{E}(Y|X = x + 1) - \mathrm{E}(Y|X = x) = \beta_1.$$

Understanding the parameters

- The effect of changing X is independent of the level of X
- The sign shows the direction of the expected change:
  - If $\beta_1 > 0$, then the changes of X and Y go into the same direction.
  - If $\beta_1 < 0$, then the changes of X and Y go into different directions.
  - If $\beta_1 = 0$, then a change in X has no influence on Y.

I.4.2 OLS-Estimation

The population parameters β_0 and β_1 are estimated from a sample. The parameters estimates (coefficients) are typically denoted by a hat: $\hat{\beta}_0$ and $\hat{\beta}_1$.

Let $(y_i, x_i), i = 1, \ldots, N$, denote a random sample of size N from the population. Hence, for each i:

$$y_i = \beta_0 + \beta_1 x_i + u_i. \tag{10}$$

- Estimation problem: how to choose the unknown parameters β_0 and β_1?

OLS-Estimation

- Estimation as Black Box? Very conveniently, the estimation problem is solved by software packages like EViews. It helps, however, to have a deeper understanding of what is going on.

- The commonly used method to estimate the parameters in a simple regression model is ordinary least square (OLS) estimation.

OLS-Estimation

- For each observation x_i, the prediction $\hat{y}_i$ of y_i depends on (β_0, β_1):

$$\hat{y}_i(\beta_0, \beta_1) = \beta_0 + \beta_1 x_i. \tag{11}$$

- For each observation x_i define the regression residuals (prediction error) $u_i(\beta_0, \beta_1)$ as:

$$u_i(\beta_0, \beta_1) = y_i - \hat{y}_i(\beta_0, \beta_1) = y_i - (\beta_0 + \beta_1 x_i). \tag{12}$$

- For each parameter value (β_0, β_1), an overall measure of fit is obtained by aggregating these prediction errors.

OLS-Estimation

- The sum of squared residuals (SSR):

$$\mathrm{SSR} = \sum_{i=1}^{N} u_i(\beta_0, \beta_1)^2 = \sum_{i=1}^{N} (y_i - \beta_0 - \beta_1 x_i)^2. \tag{13}$$

- The OLS-estimator $\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1)$ is the parameter that minimizes the sum of squared residuals.

OLS-Estimation for the Simple Regression Model

Intuitively, OLS is fitting a line through the sample points such that the sum of squared residuals is as small as possible.

Demonstration: $\Rightarrow$

MATLAB Code: `regest.m`

How to compute the OLS Estimator?

Simple regression model:

$$\hat{\beta}_1 = \frac{s_y}{s_x} r_{xy}, \qquad \hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x}, \tag{14}$$

$\overline{x}$ mean of $x_1, \ldots, x_N$, $\overline{y}$ mean of $y_1, \ldots, y_N$

s_x standard deviation of $x_1, \ldots, x_N$, s_y standard deviation of $y_1, \ldots, y_N$

r_{xy} correlation coefficient

The only requirement is that we have sample variation in X ($s_x^2 > 0$).

Proof

The OLS estimator is obtained as solution to the following minimization problem:

$$\min_{\beta_0,\beta_1} \sum_{i=1}^{N} (y_i - \beta_0 - \beta_1 x_i)^2$$

The first-order conditions are:

$$-2 \sum_{i=1}^{N} (y_i - \beta_0 - \beta_1 x_i) = 0, \tag{15}$$

$$-2 \sum_{i=1}^{N} x_i (y_i - \beta_0 - \beta_1 x_i) = 0. \tag{16}$$

Proof

From (15) we have:

$$\overline{y} = \hat{\beta}_0 + \hat{\beta}_1 \overline{x}. \tag{17}$$

Implications (algebraic properties of OLS):

- The regression line passes through the sample midpoint.
- The sum (average) of the OLS residuals $\hat{u}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i$ is equal to zero. Follows from (15):

$$\frac{1}{N}\sum_{i=1}^{N} \hat{u}_i = \frac{1}{N}\sum_{i=1}^{N} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0.$$

Proof

Substituting $\hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x}$ into (16) and solving for $\hat{\beta}_1$ we obtain, provided that $\sum_{i=1}^{N}(x_i - \overline{x})^2 > 0$ (or $s_x^2 > 0$):

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{N}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{N}(x_i - \overline{x})^2} = \frac{s_y}{s_x} r_{xy}. \tag{18}$$

Implications (algebraic properties of OLS):

- The slope estimate is the sample covariance between X and Y, divided by the sample variance of X.
- If X and Y are positively (negatively) correlated, the slope will be positive (negative).

Proof

- The sample covariance between the regressor and the OLS residuals is zero. Follows from (16):

$$\frac{1}{N}\sum_{i=1}^{N} x_i \hat{u}_i = \frac{1}{N}\sum_{i=1}^{N} x_i(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0.$$