Introductory Econometrics

Based on the textbook by Wooldridge: *Introductory Econometrics: A Modern Approach*

Robert M. Kunst
robert.kunst@univie.ac.at

University of Vienna
and
Institute for Advanced Studies Vienna

October 16, 2013

Outline

Introduction

Simple linear regression

Multiple linear regression

Heteroskedasticity

Regressions with time-series observations

Asymptotics of OLS in time-series regression

Serial correlation in time-series regression

Instrumental variables estimation

What is econometrics?

The word ‘econometrics’ may have been created by Pawel Ciompa (1910), an Austro-Hungarian (*jus soli* Ukrainian) economist who used it for a theory of bookkeeping. Ragnar Frisch (1930) extended it to the interface of economics, statistics, and mathematics. Today, econometrics is concerned with statistical methods applied to economic data.

Why is econometrics not to economics what biometrics is to biology? Historical developments, the Econometric Society, the Cowles Commission...

What is so special about economic data?

Like data in astronomy, archeology etc., typical economic data are *non-experimental* (observational). By contrast, the typical backdrop in statistics is experimental data. Typical economic populations are infinite (states, decisions by persons, not persons). Macro-econometrics handles *aggregate data*, micro-econometrics handles *individual data*.

Concepts: economic and econometric model

An economic model reflects a theoretical issue:

$$Y = K^{\alpha} L^{1-\alpha}$$

An econometric model can be estimated from data, variables must be observable:

$$y_t = \alpha k_t + \beta l_t + u_t, \quad u_t = \phi u_{t-1} + \varepsilon_t, \quad \varepsilon_t \sim iid N(0, \sigma_{\varepsilon}^2)$$

Structures in economic data

- **Cross-sectional data** are indexed by individuals, households, firms, cities, sampled at a roughly identical time. Random sampling assumption sometimes appropriate, sometimes not (heterogeneity, heteroskedasticity, sample selection problems);
- **Time-series data** on variables such as prices or economic aggregates are indexed by years, quarters, months, minutes (data frequency, sometimes irregular intervals), sampled for the same country, firm, product. Random sampling assumption usually inappropriate (serial correlation, structural breaks and change);
- **Pooled cross sections** and panel or longitudinal data are indexed in two dimensions.

Causality and *ceteris paribus*

Dependence between two variables is necessary but not sufficient for establishing **causal** effects. In time-series data, causality can be established by the principle of *post hoc ergo propter hoc*. In cross-section data, determining causal directions can be complex.

Some models assume that the effect of one variable x on another variable y can be determined while keeping all other potential influence factors constant: *ceteris paribus*. Often, this is not possible, as x also affects the other factors.

Selected textbooks of econometrics

- BAUM, C.F. (2006). *An Introduction to Modern Econometrics Using Stata*. Stata Press.
- GREENE, W.H. (2008). *Econometric Analysis*. 6th edition, Prentice-Hall.
- HAYASHI, F. (2000). *Econometrics*. Princeton.
- JOHNSTON, J. AND DINARDO, J. (1997). *Econometric Methods*. 4th edition, McGraw-Hill.
- RAMANATHAN, R. (2002). *Introductory Econometrics with Applications*. 5th edition, South-Western.
- STOCK, J.H., AND WATSON, M.W. (2007). *Introduction to Econometrics*. Addison-Wesley.
- VERBEEK, M. (2012). *A Guide to Modern Econometrics*. 4th edition, Wiley.
- WOOLDRIDGE, J.M. (2009). *Introductory Econometrics*. 4th edition, South-Western.

The simple linear regression model

The model

$$y = \beta_0 + \beta_1 x + u$$

is called the **simple linear regression model**, the word ‘simple’ referring to the fact that there is only one regressor variable x. y is called the **dependent variable**, response variable, explained variable, or regressand. x is the **explanatory variable** or regressor (covariate; the term ‘independent variable’ is discouraged). Both x and y are observed.

The unobserved variable u is called the **error term** or disturbance. The (usually unknown) fixed parameters β_0 and β_1 are the **intercept** and the **slope** of the regression.

Errors are zero on average

The classical assumption on the unobserved random variable u is that

$$\mathrm{E}(u) = 0,$$

as otherwise the intercept would not be identified. The current literature prefers the stronger assumption

$$\mathrm{E}(u|x) = 0 \therefore \mathrm{E}(u) = 0,$$

i.e. u is mean independent of x. Mean independence implies $\mathrm{E}(y|x) = \beta_0 + \beta_1 x$. $\mathrm{E}(y|x)$ is also called the **population regression function**.

Deriving OLS as method of moments estimator

In population, $\mathrm{E}(u)=0$ and $\mathrm{cov}(x,u)=0$. The sample counterparts would be

$$n^{-1}\sum_{i=1}^{n} x_i(y_i-\hat\beta_0-\hat\beta_1 x_i)=0, \quad n^{-1}\sum_{i=1}^{n}(y_i-\hat\beta_0-\hat\beta_1 x_i)=0.$$

The latter equation implies that

$$\bar y=\hat\beta_0+\hat\beta_1\bar x,$$

which is inserted into the former equation

$$\sum_{i=1}^{n} x_i(y_i-\bar y+\beta_1\bar x-\beta_1 x_i)=0,$$

and finally

$$\hat\beta_1=\frac{\sum_{i=1}^{n} x_i(y_i-\bar y)}{\sum_{i=1}^{n} x_i(x_i-\bar x)}.$$

The OLS estimator for simple regression

It is easily seen that $\hat{\beta}_1$ can also be written

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}.$$

This is called the **ordinary least squares** (OLS) estimate for the slope β_1, as it is also the solution to the minimization problem

$$\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2 \to \min.$$

The OLS estimate for the intercept follows from evaluation of the equation $\bar{y} = \hat{\beta}_0 + \hat{\beta}_1 \bar{x}$

Fitted values and residuals

The 'systematically explained' portion of y_i is called the **fitted value** $\hat{y}_i$:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i, \quad i = 1, \ldots, n,$$

while the difference between y_i and fitted value is called the **residual**:

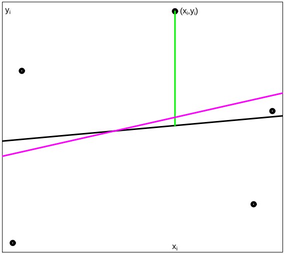
$$\hat{u}_i = y_i - \hat{y}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i, \quad i = 1, \ldots, n.$$

By construction, OLS minimizes the **sum of squared residuals** (SSR)

$$\sum_{i=1}^{n} \hat{u}_i^2 = \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$$

among all possible $\hat{\beta}_0, \hat{\beta}_1$.

OLS residuals



The vertical distance between the value y_i and the height of the point on a fitted OLS line $\hat{\beta}_0 + \hat{\beta}_1 x_i$ is called the *residual* (green) $\hat{u}_i$. The distance to the unknown population regression function (magenta) $\beta_0 + \beta_1 x$ is the *error* term u_i.

Interpretation of intercept and slope

The intercept (β_0 or $\hat{\beta}_0$) is the value for y predicted when $x = 0$ is assumed: $\mathrm{E}(y|x = 0) = \beta_0$. The interpretation does not always make sense.

The slope (β_1 or $\hat{\beta}_1$) is the marginal reaction of y to an infinitesimal change in x:

$$\frac{\partial \mathrm{E}(y|x)}{\partial x} = \beta_1.$$

The slope is usually of central interest in regression analysis.

Low-level properties of OLS I

Some OLS properties do not need any further statistical assumptions. The OLS estimate can always be calculated, as long as the denominator is not 0:

$$\sum_{i=1}^{n}(x_i - \bar{x})^2 \neq 0,$$

i.e. not all observed x_i are the same. A fitted line would be vertical, and the slope would become infinite.

By definition/derivation, it holds that

$$\sum_{i=1}^{n} \hat{u}_i = 0,$$

i.e. the average of the residuals is 0. Note that $\sum_{i=1}^{n} u_i$ is not necessarily 0, whereas $\mathrm{E} u_i = 0$.

Low-level properties of OLS II

By definition/derivation, it holds that

$$\sum_{i=1}^{n} x_i \hat{u}_i = 0,$$

i.e. regressors and residuals are orthogonal. Regressors and residuals have zero sample correlation. Note that population errors and regressors are also uncorrelated.

Interpretation is that if this correlation were non-zero, the residuals would contain some information on y that could be added to the systematic part and could reduce the SSR.

The variance decomposition

OLS decomposes the variation in y into an explained portion and a residual part:

$$\sum_{i=1}^{n}(y_i-\bar{y})^2=\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2+\sum_{i=1}^{n}\hat{u}_i^2,$$

in short

$$SST=SSE+SSR,$$

i.e. the **total sum of squares** equals the **explained sum of squares** plus the **residual sum of squares**.

Proof uses the orthogonality

$$\sum_{i=1}^{n}\hat{u}_i(\hat{y}_i-\bar{y})=0,$$

which follows from the orthogonality of $\hat{u}$ and x.

Goodness of fit

The share of the variation that is explained defines a popular measure for the goodness of fit in a linear regression,

$$R^2 = \frac{SSE}{SST} = 1 - \frac{SSR}{SST}.$$

This $\mathbf{R}^2$, really some approximate square of the sample correlation coefficient of $\hat{y}$ and y, is also called the 'coefficient of determination'. Clearly, it holds that

$$0 \leq R^2 \leq 1.$$

$R^2 = 1$ if all observations (x_i, y_i) are 'on the regression line'. It is not possible to state without context, what value determines a 'good' R^2.

Assumptions for good OLS properties

The really interesting properties of the OLS estimator can be expounded in the context of a statistical model only:

- unbiasedness $\mathrm{E}\hat{\beta} = \beta$;
- consistency $\hat{\beta} \to \beta$ as $n \to \infty$;
- efficiency, i.e. minimal variance among comparable estimators.

Such properties require assumptions on the data-generating model.

Linearity in parameters

The first assumption just defines the considered model class:

SLR.1 In the population, y depends on x according to the scheme

$$y = \beta_0 + \beta_1 x + u.$$

In the following y will be called the dependent variable, x the explanatory variable (regressor), β_0 is the population intercept, β_1 the population slope. β_0 and β_1 are parameters of the model.

This assumption is called ‘linearity in parameters’ or ‘linearity of the regression model’. It is void without further specifications for the error term u.

Random sampling

SLR.2 The available random sample of size n, $\{(x_i, y_i), i = 1, 2, \ldots, n\}$, follows the population model in (SLR.1).

This assumption is violated whenever the sample is selected from certain parts of the population (e.g., only rich households, only females, no unemployed), and when there is dependence among observations (e.g., if observation $i + 1$ is at least as large as observation i, or if there is dependence over time). This assumption implies that u_j and u_i are independent for $i \neq j$.

Do not regress on a constant covariate

SLR.3 The values for the explanatory variable x_i, $1 \leq i \leq n$, are not all identical.

We cannot study the effects of education on wages in a sample of persons with identical education. The case $y_1 = y_2 = \ldots = y_n$ does not make much sense either, but it entails a horizontal regression line $\hat{\beta}_1 = 0$ and is harmless.

Errors are zero on average

SLR.4 The error u has expectation zero given any value of the explanatory variable, in symbols

$$\mathrm{E}(u|x) = 0,$$

which is stronger than simply $\mathrm{E}u = 0$. Traditionally, x has often been assumed as identical across samples (x_i is a fixed 'non-stochastic' value in potential new samples for all i). Then, (SLR.4) is equivalent to $\mathrm{E}u = 0$.

Unbiasedness of OLS

Theorem

Under the assumptions (SLR.1)–(SLR.4), the OLS estimator is unbiased, in symbols

$$\mathrm{E}(\hat{\beta}_0) = \beta_0, \quad \mathrm{E}(\hat{\beta}_1) = \beta_1.$$

Proof: First consider $\hat{\beta}_1$ and substitute for y_i from (SLR.1)

$$\begin{aligned} \mathrm{E}\hat{\beta}_1 &= \mathrm{E}\frac{\sum_{i=1}^n x_i(y_i - \bar{y})}{\sum_{i=1}^n x_i(x_i - \bar{x})} \\ &= \beta_1 + \mathrm{E}\frac{\sum_{i=1}^n x_i(u_i - \bar{u})}{\sum_{i=1}^n x_i(x_i - \bar{x})}. \end{aligned}$$

The latter term is 0 conditional on x, according to (SLR.4), and thus unconditionally 0. (SLR.2) and (SLR.3) are needed implicitly.

Unbiasedness also for the intercept

Now consider $\hat{\beta}_0$, which is defined via

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}.$$

We now know that $\mathrm{E}\hat{\beta}_1 = \beta_1$, thus

$$\begin{aligned} \mathrm{E}(\hat{\beta}_0|x) &= \mathrm{E}(\bar{y}|x) - \mathrm{E}(\hat{\beta}_1|x)\bar{x} \\ &= \beta_0 + \beta_1\bar{x} - \beta_1\bar{x} = \beta_0, \end{aligned}$$

because of assumption (SLR.4).

Homoskedasticity

SLR.5 The error u has a constant and finite variance, conditional on x, in symbols

$$\mathrm{var}(u|x) = \sigma^2.$$

This assumption is stronger than $\mathrm{var}(u) = \sigma^2$. It is called the assumption of (conditional) **homoskedasticity**. If $\mathrm{var}(u|x)$ depends on x, the errors are said to be heteroskedastic. The assumption is often violated in cross-section regressions (richer households have more variation in their consumer preferences). Note that (SLR.5) also implies $\mathrm{var}(y|x) = \sigma^2$.

The variance of the OLS estimator

Theorem

Under assumptions (SLR.1)–(SLR.5), the variance of the OLS slope estimate is

$$\mathrm{var}(\hat{\beta}_1) = \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{\sigma^2}{SST_x},$$

while the variance of the intercept is

$$\mathrm{var}(\hat{\beta}_0) = \frac{\sigma^2 \sum_{i=1}^n x_i^2}{n \sum_{i=1}^n (x_i - \bar{x})^2},$$

in both cases conditional on x.

Here, $SST_x = \sum_{i=1}^n (x_i - \bar{x})^2$, a total sum of squares for x.

Interpretation of the variance formulae

The variance of the slope

$$\frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

increases in σ^2 and decreases in $\mathrm{var}x$. The slope estimate becomes more precise if there is less variation in the errors (points are close to the regression line) and if there is stronger variation in the regressor variable (more information). Also, assuming that the sample variance of x converges, the slope coefficient becomes more precise with increasing n.

The variance of the intercept increases with $\sum_{i=1}^{n} x_i^2$, the uncentered sum of squared regressors. The precision of the intercept estimate decreases when the regressor values are far away from the origin.

Deriving the variance of the OLS slope

Proof of the theorem: note that we consider $\mathrm{var}(\hat{\beta}_1|x)$, not the unconditional variance. Thus, in particular,

$$\begin{aligned}
\mathrm{var}(\hat{\beta}_1|x) &= \mathrm{E}\{\frac{\sum_{i=1}^n (u_i-\bar{u})(x_i-\bar{x})}{\sum_{i=1}^n (x_i-\bar{x})^2}|x\}^2 \\
&= \frac{\mathrm{E}[\{\sum_{i=1}^n (u_i-\bar{u})(x_i-\bar{x})\}^2|x]}{\{\sum_{i=1}^n (x_i-\bar{x})^2\}^2} \\
&= \sigma^2 \frac{\sum_{i=1}^n (x_i-\bar{x})^2}{\{\sum_{i=1}^n (x_i-\bar{x})^2\}^2} \\
&= \sigma^2 \frac{1}{\sum_{i=1}^n (x_i-\bar{x})^2}
\end{aligned}$$

Estimating the error variance

The variance formulae for the coefficient estimates are not operational as σ^2 is an unknown parameter and has to be estimated.

Theorem

Under assumptions SLR.1–SLR.5, the SSR scaled by $n-2$ is an unbiased estimator for the error variance σ^2, in symbols

$$\mathrm{E}\hat{\sigma}^2 = \mathrm{E}\left(\frac{\sum_{i=1}^{n}\hat{u}_i^2}{n-2}\right) = \sigma^2.$$

Proof is technical and omitted. The theorem yields an unbiased estimator for $\mathrm{var}\hat{\beta}_j$, $j = 0, 1$, but not for the standard error σ.

Homogeneous regression

Instead of building on the model $y = \beta_0 + \beta_1 x + u$, one may also start from

$$y = \beta_1 x + u$$

and thus force the regression line through the origin. The corresponding OLS estimator is

$$\tilde{\beta}_1 = \frac{\sum_{i=1}^n x_i y_i}{\sum_{i=1}^n x_i^2}.$$

This **homogeneous regression** is surprisingly rarely used. Statistical entities for this model are grounded in central moments instead of the usual variances and are for this reason not comparable to the inhomogeneous model (e.g., R^2).