Introductory Econometrics

Based on the textbook by RAMANATHAN: *Introductory Econometrics*

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

University of Vienna
and
Institute for Advanced Studies Vienna

September 23, 2011

Outline

Empirical economic research and econometrics

Empirical economic research is the internal wording for introductory econometrics. Econometrics focuses on the interface of economic theory and the actual economic world.

The information on economics comes in the shape of *data*.

This data (*quantitative* rather than qualitative) is the subject of the analysis.

In current usage, methods for the statistical analysis of the data are called 'econometrics', not for the gathering or compilation of the data.

Econometrics

Word appears for the first time around 1900: analogy construction to biometrics etc.

Founding of the *Econometric Society* and its journal *Econometrica* (1930, RAGNAR FRISCH and others): *mathematical and statistical* methods in economics.

Today, mathematical methods in economics (mathematical economics) are no more regarded as 'econometrics', while they continue to dominate the Econometric Society and also Econometrica.

Central issues of econometrics

In the early days, the focus is on the collection of data (*national accounts*).

Cowles Commission: genuine autonomous research on procedures for the estimation of linear dynamic models for aggregate variables with simultaneous feedback (private consumption $\rightarrow$ GDP $\rightarrow$ income $\rightarrow$ private consumption): issues that are otherwise unusual in the field of statistics.

In recent decades, a shift of attention toward microeconometric analysis, decreasing dominance of the macroeconometric model: microeconometrics is closer to the typical approach of statistics used in other disciplines.

Selected textbooks of econometrics

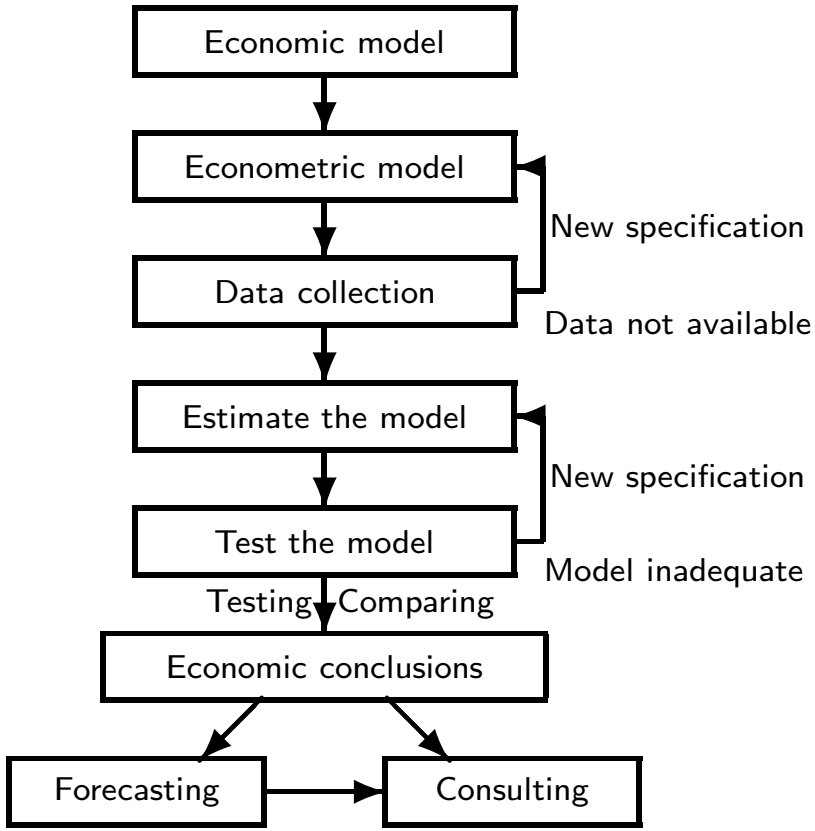
- Ramanathan, R. (2002). *Introductory Econometrics with Applications*. 5th edition, South-Western.
- Berndt, E.R. (1991). *The Practice of Econometrics*. Addison-Wesley.
- Greene, W.H. (2008). *Econometric Analysis*. 6th edition, Prentice-Hall.
- Hayashi, F. (2000). *Econometrics*. Princeton.
- Johnston, J. and DiNardo, J. (1997). *Econometric Methods*. 4th edition, McGraw-Hill.
- Stock, J.H., and Watson, M.W. (2007). *Introduction to Econometrics*. Addison-Wesley.
- Wooldridge, J.M. (2009). *Introductory Econometrics*. 4th edition, South-Western.

Aims of econometric analysis

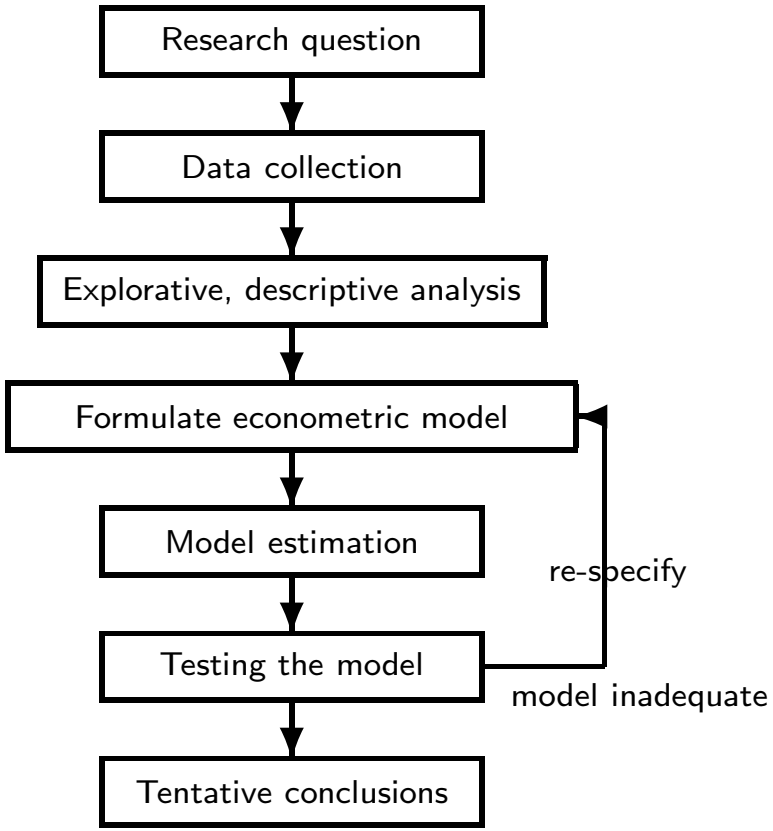
- Testing economic hypotheses;
- Quantifying economic parameters;
- Forecasting;
- Establishing new facts from statistical evidence (e.g. empirical relationships among variables).

theory-driven or *data-driven*

Flow chart following Ramanathan



Modified flow chart



Classification of data according to index subscript

1. **Cross-section data**: mostly in microeconometrics. Index i denotes a person, a firm etc.;
2. **Time-series data**: mostly in macroeconometrics and in finance. Index t denotes a time point (year, quarter, month, day, ...);
3. **Panel data**: two-dimensional, one index denotes the 'individual' i, the other one the time point t;
4. multi-dimensional data, e.g. with spatial dimension.

In the following, all observations will be indexed by t, even when this t does not have a temporal interpretation: t denotes the *t*ypical observation.

Time series of flows and stocks

1. Most macroeconomic 'accounts' variables are *flows*: GDP for the year 2005;
2. Quantities and prices are *stocks*: e.g. unemployed persons in January 2009; these can be measurements at a specific time or averages over time intervals.

Differences between the two types concern temporal aggregation: with flows, annual data evolve from monthly data by sums (accumulation); with stocks, one may use averages over monthly observations, though sometimes other conventions are used.

Classification of data according to their genesis

1. *Experimental* data are rare outside experimental economics;
2. *Samples*: e.g. 100 persons out of 3 Mio. randomly selected: microeconomic situations;
3. *Observational data*: GDP 2005 exists only once and is provided in the Statistik Austria database: typical case in macroeconomics.

Observational data cannot be augmented, their population is an abstract concept.

Typical macroeconomic data: aggregate consumption and disposable household income

year	consumption C	income YD
1976	71.5	84.2
1977	75.9	86.6
1978	74.8	88.9
...	...	...
2007	133.5	154.0
2008	134.5	156.7

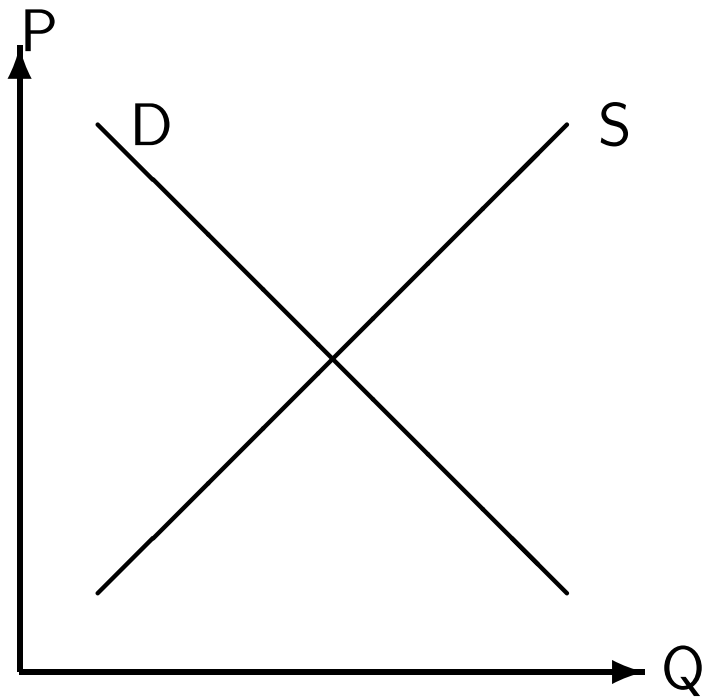
What is a model?

Every science has its specific concept of a model.

- An economic model contains concepts regarding cause-effect relationships among variables, variables need not be observed;
- An econometric model contains assumptions on statistical distributions of (potential) data-generating mechanisms for observed variables.

The translation between the two model concepts is a typical weak point of empirical projects.

Example: economic but not econometric model



Supply and demand: variables are not observed, model statistically inadequate (not identified).

Example: econometric model with economic problem of interpretation

$$\begin{aligned} y_t &= \alpha + \beta x_t + \varepsilon_t \\ \varepsilon_t &\sim N(0, \zeta_0 + \zeta_1 \varepsilon_{t-1}^2) \end{aligned}$$

Regression model with ARCH errors: what is ζ_0? A lower bound for the variance of shocks in very calm episodes?

The sample

Notation: There are n observations for the variable X:

$$X_1, X_2, \ldots, X_n,$$

or

$$X_t, t = 1, \ldots, n.$$

The *sample size* is n (the number of observations).

In statistical analysis, the sample is seen as the *realization of a random variable*. The typical statistical notation, with capitals denoting random variables and lower-case letters denoting realizations ($X = x$), is not adhered to in econometrics.

'Sample' with observational data

Example: Observations on private consumption C and (disposable) income YD 1976–2008. There exists the backdrop economic model of a 'consumption function'

$$C_t \approx \alpha + \beta YD_t.$$

Are these 33 realizations of C and YD or *one* realization of $(C_1, \ldots, C_{33})$ and $(YD_1, \ldots, YD_{33})$?

Usefulness or honesty?

- 33 realizations of *one* random variable: C and YD are increasing over time, time dependence (business cycle, inertia in the reaction of households), changes in economic behavior: implausible. Parameters of the consumption function can be estimated with reasonable precision.
- One realization: Interpretation of unobserved realizations becomes problematic (parallel universes?). Parameters of the consumption function cannot be estimated from one observation.

Way out: not the variables, but the unobserved error term is the random variable: 33 observations, 'shocks stationary', consumption function again estimable.

Definition of a parameter

A parameter ('beyond the measurable') is an unknown constant. It describes the statistical properties of the variables, for which the observations are realizations.

Example: In the regression model

$$C_t = \alpha + \beta YD_t + u_t, \quad u_t \sim N(0, \sigma^2),$$

α, β, σ^2 are parameters. Not every parameter is economically interesting.

Definition of an estimator

An *estimator* is a specific function of the data that is designed to approximate a parameter. Its realization for given data is called *estimate*.

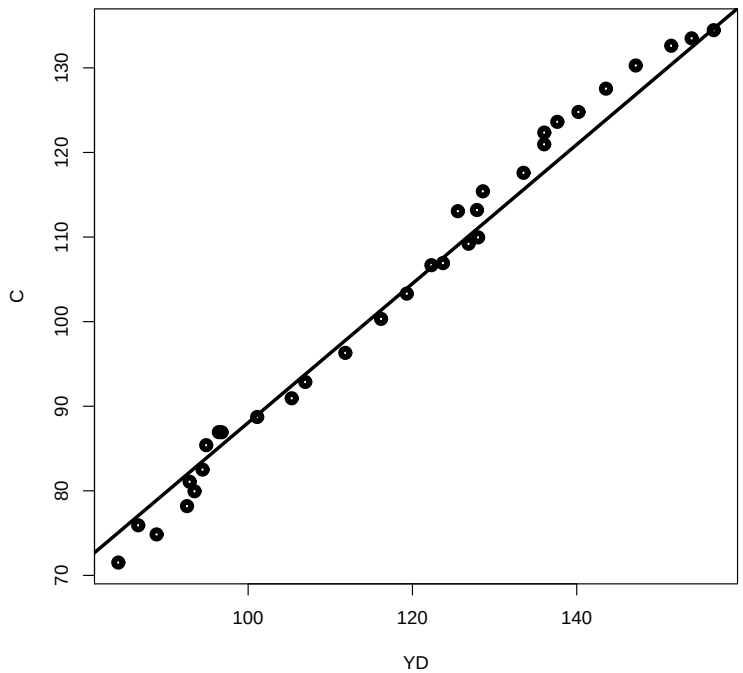
The word 'estimator' is used for the random variable that evolves from calculating the function of a virtual sample—a statistic—and for the functional form proper. Statistics calculated from observed data are observed. If data are assumed to be realizations of random variables, all statistics calculated from them and also the estimators are random variables, while estimates are realizations of these random variables. Parameters are unobserved.

Notation for estimates and parameters

Parameters are usually denoted by Greek letters: α,β,θ,...

The corresponding estimates or estimators are denoted by a hat: $\hat{\alpha}, \hat{\beta}, \hat{\theta}, \ldots$ Latin letters (*a* estimates α etc.) are also used by some authors.

Examples of estimators



Consumption C depending on income YD can be represented as a consumption function (line) and as data points in a scatter diagram: the intercept $\hat{\alpha}$ estimates the autonomous consumption, the slope $\hat{\beta}$ estimates the marginal propensity to consume of the households.

Definition of the hypothesis test

A *test* is a statistical decision rule. The aim is a decision whether the data suggest that a specified hypothesis is incorrect ('to reject') or correct ('not to reject', 'to accept', 'fail to reject').

It is customary first to calculate a *test statistic* from the data, a real number that is a function of the data. If this statistic is in the *rejection region* (critical region), the test is said to reject.

Tests: common mistakes

- **A test is not a test statistic**. The test statistic is a real number, whereas the test itself can only indicate rejection or acceptance ('non-rejection'), a decision that may be coded by 0/1. A test cannot be 1.5.
- **A test is not a null hypothesis**. A test cannot be rejected. The tested hypothesis can be rejected. The test rejects.

Do economic theories deserve a presumption of innocence?

The *null hypothesis* (or short *null*) is a (typically sharp) statement on a parameter. Example: $\beta = 1$. These statements often correspond to an economic theory.

The *alternative hypothesis* (or short *alternative*) is often the negation of the null within the framework of a general hypothesis (*maintained hypothesis*). Example: $\beta \in (0, 1) \cup (1, \infty)$. This statement often corresponds to the invalidity of an economic theory.

Notation: H_0 for the null hypothesis and e.g. H_A for the alternative. $H_0 \cup H_A$ is the general or maintained hypothesis. Example: the maintained hypothesis $\beta > 0$ is not tested but assumed ('window on the world').

Hypotheses: common mistakes

- **A hypothesis cannot be a statement on an estimate**. "$H_0 : \hat{\beta} = 0$" is meaningless, as $\hat{\beta}$ can be determined exactly from data.
- **Alternatives cannot be rejected**. The classical hypotheses test is asymmetrical. Only H_0 can be rejected or 'accepted'. (Bayes tests are non-classical and follow different rules)
- **Hypotheses do not have probabilities**. H_0 is never correct with a probability of 90% after testing. (Bayes tests allot probabilities to hypotheses)

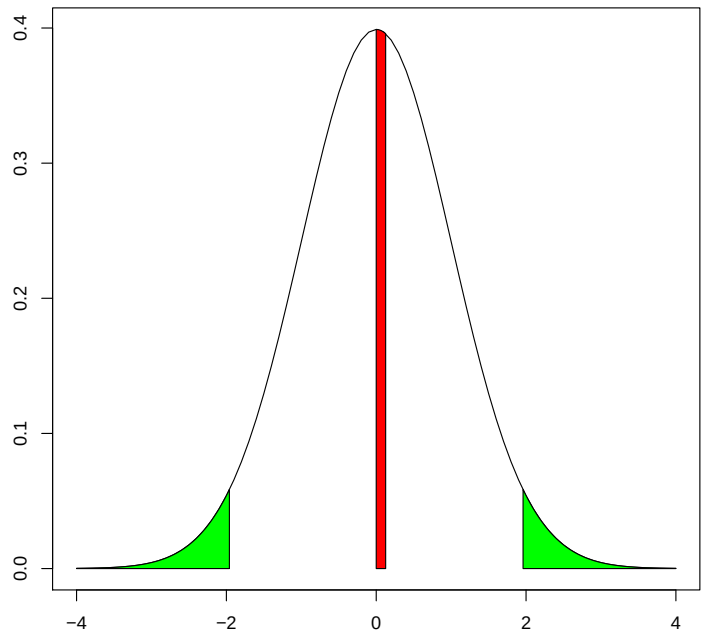
Errors of type I and of type II

The parameters are unobserved and can only be determined approximately from the sample. For this reason, decisions are often incorrect:

- If the test rejects, although the null hypothesis is correct, this is a **type I error**;
- If the test accepts the null, although it is incorrect, this is a **type II error**.

A basic construction principle of classical hypothesis tests is to prescribe a bound on the probability of type I errors (e.g. by 1%, 5%) and, given this bound, to minimize the probability of a type II error.

Two tests on a 5% level



Density of a test statistic under H_0. 2 tests use the same test statistic. The one with red rejection region is viewed as a worse test than the one with green rejection region, as it implies more type II errors.

More vocabulary on tests

The often pre-specified probability of a type I error is called the **size** of the test or the significance level. For simple null hypotheses, it will be approximately constant under H_0, otherwise it is an upper bound.

The probability *not* to make a type II error is called the **power** of the test. It is defined on the alternative, and it depends on the 'distance' from H_0. Close to H_0, power will be low (close to size); at locations far from H_0, it approaches 1.

If the test attains a power of 1 on the entire alternative, as the sample size grows toward ∞, the test will be called **consistent**.

Best tests

A concept analogous to efficient estimators, a test that attains maximal power at given size, exists for some simple problems only. For most test problems, however, the *likelihood-ratio test* (LR test, the test statistic is a ratio of the maxima of the likelihood under H_0 and H_A) has good power properties.

Often, the *Wald test* and the *LM test* (Lagrange multiplier) represent 'cheaper' approximations to LR and still they have identical properties for large n .

Attention: LR, LM and Wald test are *construction principles* for tests, not specific tests. It does not make sense to say that 'the Wald test rejects', unless it is also indicated what is being tested.

Can a presumption of innocence apply to the invalidity of a theory?

Economic statements need not correspond to the null of a hypothesis test. Sometimes, the null hypothesis is the invalidity of a theory. For example, theory may tell that the export ratio depends on the exchange rate.

Even then, the null of the test corresponds to the more restrictive statement. The economic statement becomes the alternative. The invalidity of the economic theory can be rejected. For example, H_0 says that the coefficient of the export ratio that depends on the exchange rate is 0, and this is rejected.

How are significance points found?

Tests are usually defined by a test statistic and by critical values. Critical values (significance points) are usually quantiles of the distribution of the test statistic under H_0.

Determination of critical values:

1. Tables: Quantiles of important distributions have been tabulated by simulation or by analytical calculation.
2. Calculation: Computer inverts the distribution function.
3. p–values: Computer calculates marginal significance levels.
4. Bootstrap: Computer simulates distribution function under H_0 and takes characteristics of the sample into account.

Definition of the p–value

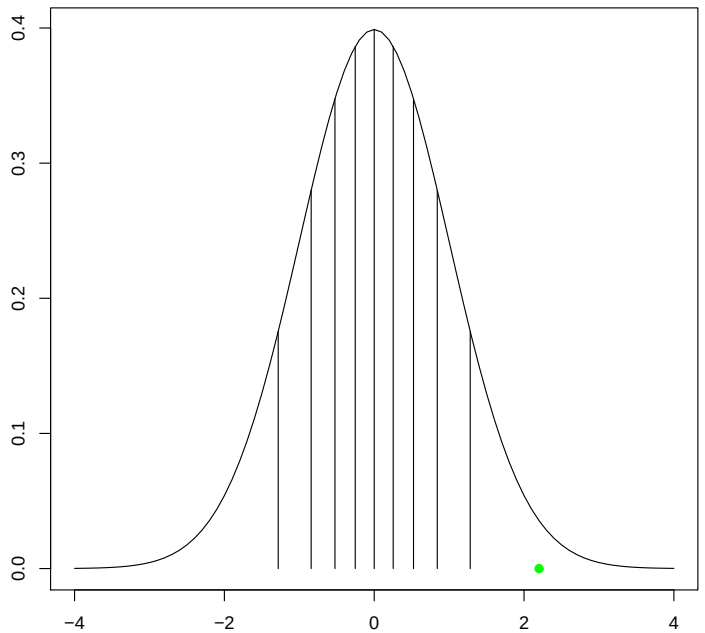
Correct definitions:

- The p–value is the significance level, at which the test becomes indifferent between rejection and acceptance for the sample at hand (the calculated value of the test statistic);
- The p–value is the probability of generating values for the test statistic that are, under the null hypothesis, even more unusual (less typical, often 'larger') than the one calculated from the sample.

Incorrect definition:

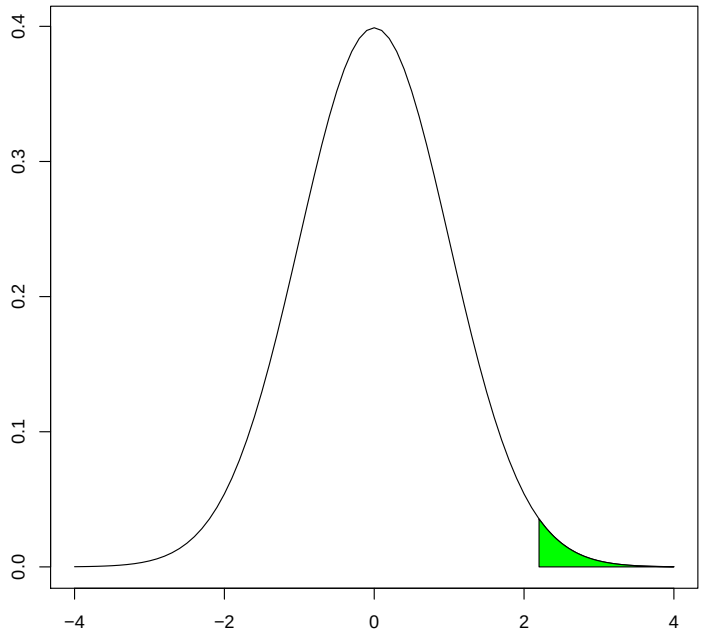
- The p–value is the probability of the null hypothesis for this sample.

Test based on quantiles



10% to 90% quantiles of the normal distribution. The observed value of 2.2 for the test statistic that is, under H_0, normally distributed is significant at 10% for the one-sided test.

Test based on p–values



The area under the density curve to the right of the observed value of 2.2 is 0.014, which yields the p–value. The one-sided test rejects on the levels of 10%, 5%, but not 1%.

Definition of expectation

The *expectation* of a random variable X with density $f(x)$ is defined by

$$\mathrm{E}X = \int_{-\infty}^{\infty} xf(x)dx.$$

For discrete random variables with probabilities $P(X = x_j)$, the corresponding definition is

$$\mathrm{E}(X) = \sum_{j=1}^{n} x_j P(X = x_j).$$

Linearity of expectation

For two (even dependent) random variables X and Y and arbitrary $\alpha \in \mathbb{R}$,

$$\mathrm{E}(X + \alpha Y) = \mathrm{E}X + \alpha \mathrm{E}Y,$$

holds, and the expectation of a constant 'random variable' a is a.

Beware, however, that in general $\mathrm{E}(XY) \neq \mathrm{E}(X)\mathrm{E}(Y)$, and $\mathrm{E}g(X) \neq g(\mathrm{E}X)$ for a general function $g(.)$, particularly $\mathrm{E}(1/X) \neq 1/\mathrm{E}X$.

Estimation of the expectation

The most usual estimator for the expectation of a random variable is the *sample mean*

$$\bar{X} = \sum_{t=1}^{n} X_t / n \quad .$$

Pair of concepts:

- Expectation $\mathrm{E}X$: *population mean*, parameter, unobserved.
- Mean $\bar{X}$: *sample mean*, statistic, observed.

Definition of the variance

The *variance* of a random variable X with density $f(x)$ is defined by

$$\mathrm{var}X = \mathrm{E}(X - \mathrm{E}X)^2 = \mathrm{E}X^2 - (\mathrm{E}X)^2.$$

It measures the dispersion of a random variable around its expectation (centered 2. Moment). A customary notation is $\sigma_X^2 = \mathrm{var}X$.

The square root of the variance is called the *standard deviation* (preferred for observed X) or *standard error* (preferred for statistics) and is often denoted by σ_X.

Properties of the variance

1. $\mathrm{var}X \geq 0$;
2. $X \equiv \alpha \in \mathbb{R} \therefore \mathrm{var}X = 0$;
3. $\mathrm{var}(X + Y) = \mathrm{var}X + \mathrm{var}Y + \mathrm{cov}(X, Y)$ with $\mathrm{cov}(X, Y) = \mathrm{E}\{(X - \mathrm{E}X)(Y - \mathrm{E}Y)\}$;
4. $\alpha \in \mathbb{R} \therefore \mathrm{var}(\alpha X) = \alpha^2 \mathrm{var}X$.

The variance operator is not linear, $\mathrm{var}X$ can also be $+\infty$.

Estimation of the variance

The most customary estimator for the variance of a random variable is the *empirical variance*

$$\widehat{\mathrm{var}}(X) = \frac{1}{n-1}\sum_{t=1}^{n}(X_t - \bar{X})^2 = \frac{1}{n-1}\sum_{t=1}^{n}X_t^2 - \frac{n}{n-1}\bar{X}^2.$$

Pair of concepts:

- Variance $\mathrm{var}X$: *population variance*, parameter, unobserved.
- Sample variance $\widehat{\mathrm{var}}X$: statistic, observed.

Bias of estimators

Suppose $\hat{\theta}$ is an estimator for the parameter θ. The value

$$\mathrm{E}\hat{\theta} - \theta$$

is called the *bias* of the estimator. If the bias is 0, i.e.

$$\mathrm{E}\hat{\theta} = \theta,$$

then the estimator is called *unbiased*. This is certainly a desirable property for estimators: the distribution of the estimates is centered around the true value.

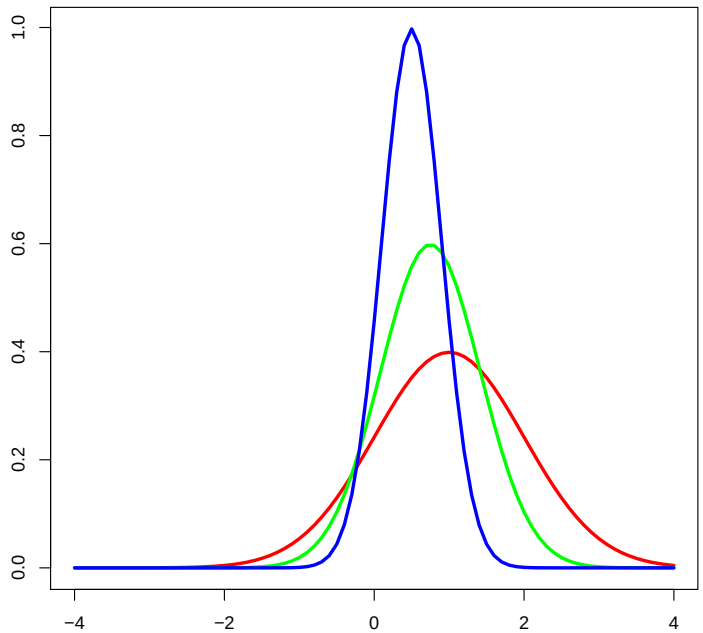
Consistency of estimators

Assume that the sample size diverges toward ∞. Denote the estimate for the sample size n by $\hat{\theta}_n$. If, in any reasonable definition of convergence

$$\hat{\theta}_n \to \theta, n \to \infty,$$

then the estimator is called *consistent*. This is another desirable property of estimators: the distribution of the estimates shrinks toward the true value, as the sample size grows.

Consistency is more important than unbiasedness



Consistent estimator with bias for the true value of 0. Small sample red, intermediate green, largest blue.

Consistency and unbiasedness do not imply each other

Example: Let X_t be independently drawn from N(a,1). Someone believes *a priori* b to be a plausible value for the expectation. $\hat{a}_1 = \bar{X} + b/n$ is consistent, usually biased, and a reasonable estimator.

Example: Same assumptions, $\hat{a}_2 = (X_1 + X_n)/2$ is unbiased, but inconsistent and an implausible estimator.

These examples are constructed but typical: inconsistent estimators should be discarded, biased estimators can be useful or inevitable.

Efficiency of estimators

An unbiased estimator $\hat{\theta}$ is called more efficient than another unbiased estimator $\tilde{\theta}$, if

$$\mathrm{var}\hat{\theta} < \mathrm{var}\tilde{\theta}.$$

If $\hat{\theta}$ has minimal variance among all unbiased estimators, it is called *efficient*.

Remark: Excluding biased estimators is awkward, but we will work with this definition for the time being.

Simple linear regression

The model explains (describes) a variable Y by a variable X, with n observations available for each of the two:

$$y_t = \alpha + \beta x_t + u_t, \quad t = 1, \ldots, n,$$

where α and β are unknown parameters (coefficients). The regression is called

- *simple*, as only one variable is used to explain Y;
- *linear*, as the dependence of Y on X is modelled by a linear (affine) function.

Terminology

In words, Y is said to be *regressed on* X.

Y is called the *dependent variable* of the regression, also the regressand, the explained variable, the response;

X is called the *explanatory variable* of the regression, also the *regressor*, the design variable, the covariate.

Usage of the wording 'independent variable' for X is strongly discouraged. Only in true experiments will X be set independently. In the important multiple regression model, typically all regressors are mutually dependent.

α and β

α is the *intercept* of the regression or the regression constant. It represents the value for $x = 0$ on the theoretical regression line $y = \alpha + \beta x$, i.e. the location where the regression line intersects the y–axis. Example: in the consumption function, α is autonomous consumption.

β is the *regression coefficient* of X and indicates the marginal reaction of Y on changes in X an. It is the slope of the theoretical regression line. Example: in the consumption function, β is the marginal propensity to consume.

What is u?

In the descriptive regression model, u is simply the unexplained remainder.

In the statistical regression model, the true u_t is an unobserved random variable, the *error* or disturbance term of the regression.

The word 'residuals' describes the observed(!) errors that evolve after estimating coefficients. Some authors, however, call the unobserved errors u_t 'true residuals'.

Is simple regression an important model?

The simple linear regression model offers hardly anything new relative to correlation analysis for two variables, and it is rather uninteresting.

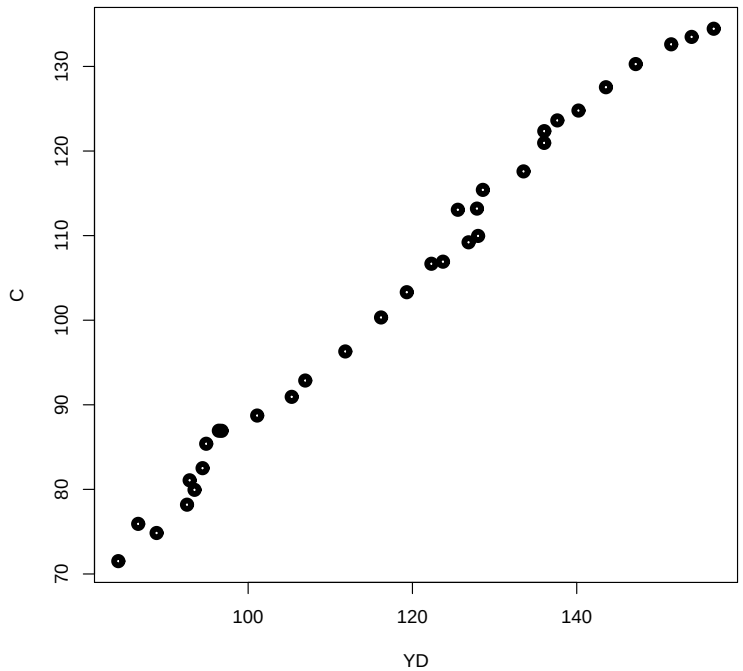
It is a didactic tool that permits to introduce new concepts that continue to exist in the *multiple regression model*. This multiple model is the most important model of empirical economics.

Descriptive regression

The descriptive problem does not make any assumptions on the statistical properties of variables and errors. It does not admit any statistical conclusions.

The descriptive linear regression problem is to fit a straight line through a scatter of points, such that it fits as well as possible.

Example for scatter diagram: aggregate consumption

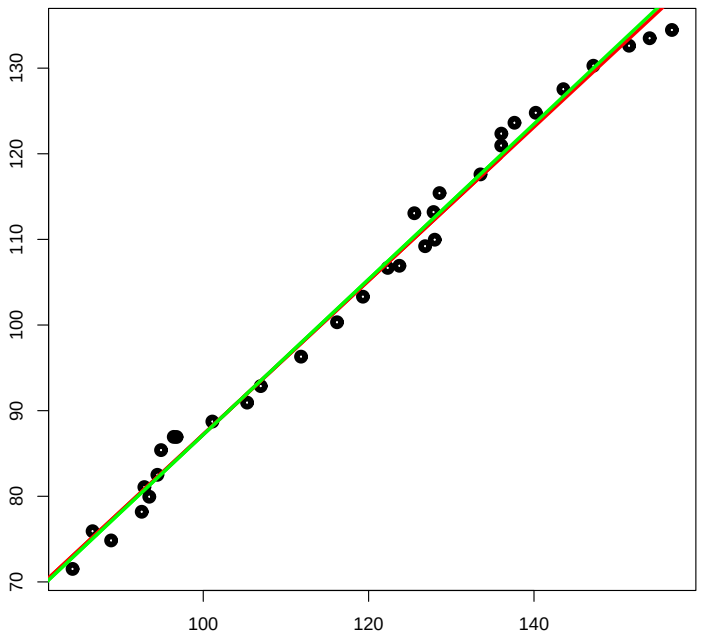


Consumption C and income YD for Austria as data points in a scatter diagram.

Suggestions how to fit a line through a scatter

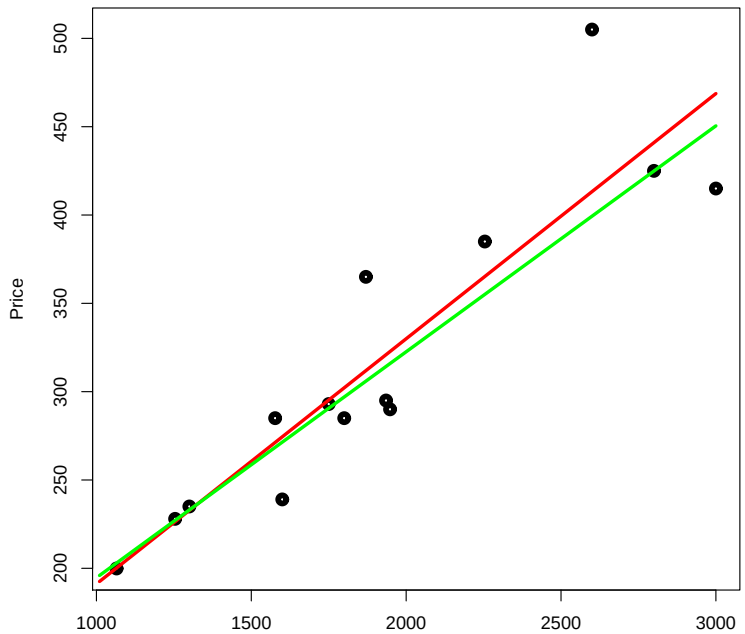
1. The line should hit as many points as possible exactly: usually, the number of points that can be forced to lie on the line is just two;
2. 'Free-hand method': subjective;
3. Minimize distances (of line to points) in the plane: orthogonal regression, important but rarely used method. Variables are treated symmetrically, like in correlation analysis;
4. Minimize the sum of vertical absolute distances: *least absolute deviations* (LAD). Robust procedure;
5. Minimize the sum of vertical squared distances: *ordinary least squares* (OLS). Most commonly used method.

Results of regression methods are similar in good samples



LAD regression (green) and OLS regression (red) applied to consumption data.

Sizeable discrepancies among methods in samples of moderate quality



LAD (green) and OLS (red) applied to RAMANATHAN's sample of 14 houses in San Diego, their sizes and their prices.

The OLS regression problem is easily solved

We minimize

$$Q(\alpha, \beta) = \sum_{t=1}^{n} (y_t - \alpha - \beta x_t)^2$$

in α and β, the minima are denoted by $\hat{\alpha}$ and $\hat{\beta}$. Taking derivatives with respect to α and equating to zero

$$\frac{\partial}{\partial \alpha} Q(\alpha, \beta) = 0,$$

which yields

$$-2 \sum_{t=1}^{n} (y_t - \hat{\alpha} - \hat{\beta} x_t) = 0 \therefore \bar{y} = \hat{\alpha} + \hat{\beta} \bar{x}.$$

This is the so called first normal equation: empirical means are exactly on the regression line.

The second normal equation

Taking derivatives with respect to β and equating to zero

$$\frac{\partial}{\partial \beta} Q(\alpha, \beta) = 0$$

yields

$$-2 \sum_{t=1}^{n} (y_t - \hat{\alpha} - \hat{\beta} x_t) x_t = 0 \therefore \frac{1}{n} \sum_{t=1}^{n} y_t x_t = \hat{\alpha} \bar{x} + \hat{\beta} \frac{1}{n} \sum_{t=1}^{n} x_t^2.$$

This is the second normal equation. 2 equations in 2 variables $\hat{\alpha}$ and $\hat{\beta}$ can be solved easily.

The OLS estimator

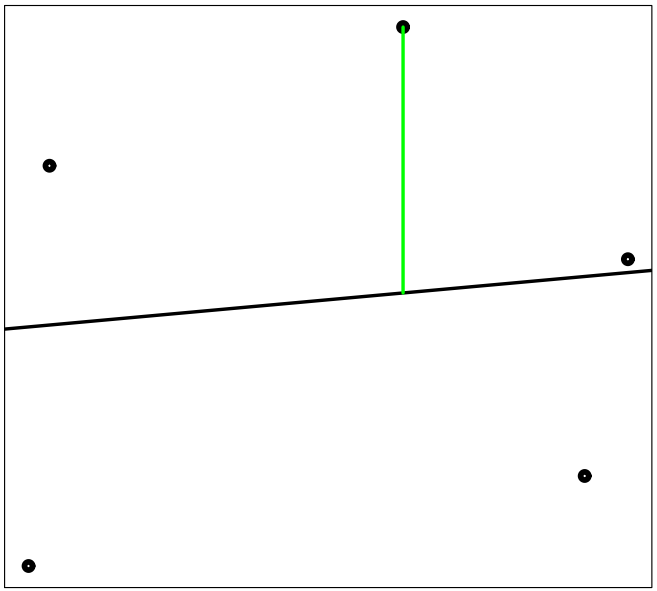
Solving the system of both normal equations yields

$$\hat{\beta} = \frac{\frac{1}{n}\sum_{t=1}^{n}(x_t - \bar{x})(y_t - \bar{y})}{\frac{1}{n}\sum_{t=1}^{n}(x_t - \bar{x})^2}.$$

The regression coefficient is an empirical covariance of X and Y divided by an empirical variance of X. The solution for $\hat{\alpha}$ is less easy to interpret:

$$\hat{\alpha} = \frac{\bar{y}\frac{1}{n}\sum_{t=1}^{n} x_t^2 - \bar{x}\frac{1}{n}\sum_{t=1}^{n} x_t y_t}{\frac{1}{n}\sum_{t=1}^{n} x_t^2 - \bar{x}^2}$$

OLS residuals

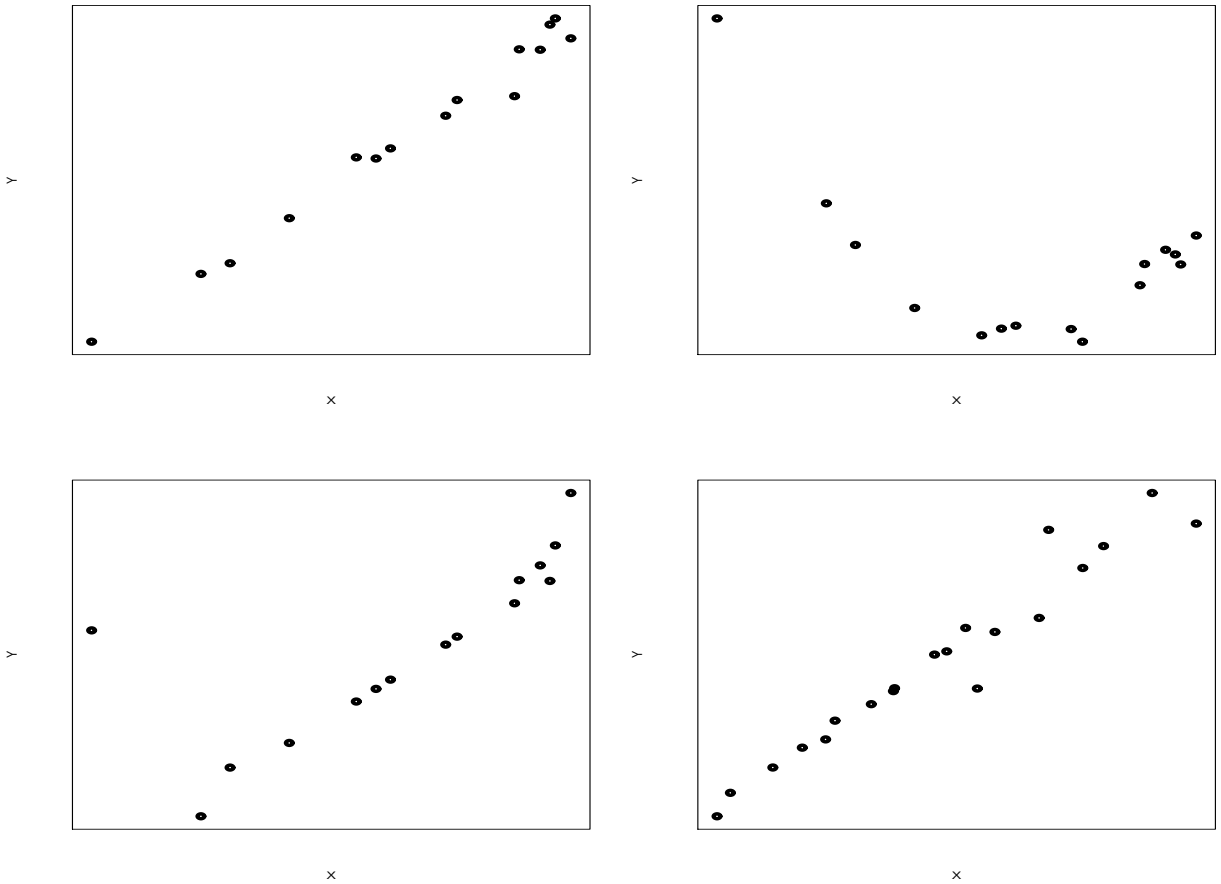


The vertical distance between the value y_t and the height of the point on a fitted OLS line $\hat{\alpha} + \hat{\beta} x_t$ (below or above) is called the *residual*.

Scatter diagrams

Scatter plots) with dependent variable on the y–axis and regressor on the x–axis are an important graphic tool in simple regression.

Because a *linear* regression function is fitted to the data, a *linear* relationship should be recognizable.



From left to right and from top to bottom: nice diagram, suspicion of nonlinearity, outlier, suspicion of heteroskedasticity

Stochastic linear regression: the idea

In the classical regression model

$$y_t = \alpha + \beta x_t + u_t,$$

a non-random input X impacts on an output Y and is disturbed by the random variable U. Thus, also Y becomes a random variable.

Assumptions on the 'design' X and the statistical properties of U admit statistical statements on estimators and tests.

Stochastic linear regression: critique

X as well as Y are typically observational data, there is no experimental situation. Should not both X and Y be modelled as random variables and assumptions be imposed on these two?

In principle, this critique is justified. It is, however, didactically simpler to start with this simple model, and to generalize and modify it later. So called *fully stochastic* models often use restrictive assumptions on the independence of observations.

Assumption 1: Linearity of the model

Assumption 1. The observations correspond to the linear model

$$y_t = \alpha + \beta x_t + u_t, \quad t = 1, \ldots, n.$$

Conceptually, this assumption is void without further assumptions on the errors u_t.

Empirical consequence: nonlinear relationships often become linear after transformations of X and/or Y.

Assumption 2: Reasonable design

Assumption 2. Not all observed values for X_t are identical.

Experiments on the effect of X on Y are only meaningful if X varies. One observation is insufficient. Vertical regression lines are undesirable. Convenient: this assumption can be checked immediately by a simple inspection of the data (without any hypothesis tests and exactly).

Assumption 3: Stochastic errors

Assumption 3. The errors u_t are realizations of random variables U_t, and it should also hold that

$$\mathrm{E}U_t = 0 \quad \forall t.$$

Actually, these are 2 assumptions: U is an unobserved random variable, and its expectation is 0. Admitting an expectation unequal 0 would be meaningless, as then the regression constant would not be defined (not identified).

Assumption 4: Natural experiment

Assumption 4. The explanatory variable X_t is not stochastic. The values $x_t, t = 1, \ldots, n$, are given data.

If the experiment is repeated, it will shows the same values for X_t, whereas the values for u_t and, in consequence, for y_t will change. Technical auxiliary assumption.

If the regressor conceptually depends on Y or if it is a lagged Y, this assumption must be abandoned and replaced by different assumptions.

The meaning of assumptions 1–4

With assumptions 1–4, the regression model is a meaningfully defined model.

With **assumptions 1–4**, the OLS estimators $\hat{\alpha}$, $\hat{\beta}$ for the coefficients α and β is **unbiased**;

Together with some technical auxiliary assumptions, assumptions 1–4 suffice to guarantee the **consistency** of the OLS estimator.

Unbiasedness of $\hat{\beta}$

Remember that

$$\hat{\beta} = \frac{\frac{1}{n}\sum_{t=1}^{n}(x_t - \bar{x})(y_t - \bar{y})}{\frac{1}{n}\sum_{t=1}^{n}(x_t - \bar{x})^2}.$$

Application of the expectation operators

$$\mathrm{E}\hat{\beta} = \frac{\frac{1}{n}\sum_{t=1}^{n}(x_t - \bar{x})\mathrm{E}(y_t - \bar{y})}{\frac{1}{n}\sum_{t=1}^{n}(x_t - \bar{x})^2},$$

as both the numerator and all x–terms are non-stochastic (assumption 4). Assumption 2 guarantees that $\hat{\beta}$ can be evaluated (denominator is not 0).

Evaluation of the expectation term by substitution from assumption 1:

$$\mathrm{E}(y_t - \bar{y}) = \mathrm{E}(\alpha + \beta x_t + u_t - \alpha - \beta\bar{x} - \bar{u}),$$

because of

$$\begin{aligned} \frac{1}{n}\sum_{t=1}^{n} y_t &= \frac{1}{n}\sum_{t=1}^{n}(\alpha + \beta x_t + u_t) \\ &= \alpha + \beta\bar{x} + \bar{u}, \end{aligned}$$

with the usual notation for arithmetic means. Due to assumption 3, the expectation of the u terms is 0, which yields

$$\mathrm{E}(y_t - \bar{y}) = \beta(x_t - \bar{x}),$$

not stochastic because of assumption 4.

Substitution yields immediately

$$\mathrm{E}\hat{\beta} = \frac{\frac{1}{n}\sum_{t=1}^{n}(x_t - \bar{x})\beta(x_t - \bar{x})}{\frac{1}{n}\sum_{t=1}^{n}(x_t - \bar{x})^2} = \beta.$$

Can this be proved even without assumption 4?

- If X is stochastic, but 'exogenous', expectations can be evaluated conditional on X, using $\mathrm{E}\{\mathrm{E}(\hat{\beta}|X)\} = \mathrm{E}\hat{\beta}$: OLS is again unbiased;
- If X is a lagged Y, this trick does not work, OLS will typically be biased.

Unbiasedness of $\hat{\alpha}$

Because of the first normal equation

$$\bar{y} = \hat{\alpha} + \hat{\beta}\bar{x},$$

it follows that

$$\begin{aligned} \mathrm{E}\hat{\alpha} &= \mathrm{E}(\bar{y}) - \mathrm{E}(\hat{\beta})\bar{x} \\ &= \alpha + \beta\bar{x} - \beta\bar{x} = \alpha, \end{aligned}$$

and hence the OLS estimator for the regression constant is also unbiased.

Consistency of $\hat{\beta}$

Re-arranging yields

$$\hat{\beta} = \beta + \frac{\frac{1}{n}\sum_{t=1}^{n}(x_t - \bar{x})(u_t - \bar{u})}{\frac{1}{n}\sum_{t=1}^{n}(x_t - \bar{x})^2}.$$

For $n \to \infty$, the denominator converges to a sort of variance of X, and the numerator to 0, assuming the X are set independently of U. For the formal proof, some complex auxiliary assumptions are required on how to 'set' the non-stochastic X.

The proof of consistency does not use assumption 4, as long as X and U are independent. Even when X is a lagged Y, OLS is usually consistent (under assumption 6, see below).

Assumption 5: homoskedastic errors

Assumption 5. All errors U_t are identically distributed with finite variance σ_u^2.

Here, U and not only U_t is a random variable, its variance is finite. With cross-section data, assumption 5 is often violated.

homoskedasticity=identical variance, *heteroskedasticity*=varying variance.

Assumption 6: independent errors

Assumption 6. The errors U_t are statistically independent.

If the variance of U_t is finite, it follows that the U_t are also uncorrelated (*no autocorrelation*). With time-series data, assumption 6 is often violated. *Autocorrelation* = correlation of a variable over time 'with itself' (*auto*).

Common mistake: Not the residuals $\hat{u}_t$ are independent, but the errors u_t. OLS residuals are usually (slightly) autocorrelated.

OLS becomes linear efficient

Assumptions 1–6 imply that the OLS estimators $\hat{\alpha}$ and $\hat{\beta}$ are the linear unbiased estimators with the smallest variance (o.c.s.).
BLUE=*best linear unbiased estimator*.

A *linear* estimator can be represented as a linear combination of y_t, and this also holds for OLS. OLS is not necessarily the best estimator (efficient).

This BLUE property (not shown here) under assumptions 1–6 is also called the *Gauss-Markov Theorem*.

Assumption 7: sample sufficiently large

Assumption 7. The sample must be larger than the number of estimated regression coefficients.

The simple model has 2 coefficients, thus the sample should have at least 3 observations. Fitting a straight line through 2 points exactly does not admit any inference on errors. Technically convenient assumption: can be checked without any hypothesis tests, just by counting.

Assumption 8: normal regression

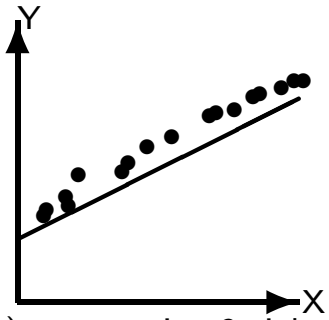
Assumption 8. All errors U_t are normally distributed.

It is convenient if the world is normally distributed, but is it realistic? In small samples, it is impossible to test this assumption. In large samples, there is often evidence of it being violated. Regression under assumption 8 is called *normal regression*.

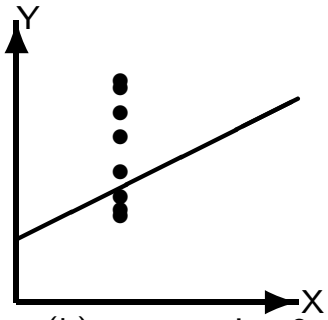
Implications of assumption 8

- Under **Assumptions 1–8**, OLS becomes the **efficient** estimator, there is no other unbiased estimator with smaller variance. Usually, however, a search for more efficient nonlinear estimators is fruitless for small or moderate n;
- Under assumption 8, hypothesis tests have exact distributions (t, F, DW). Without assumption 8, all distributions are approximative only.

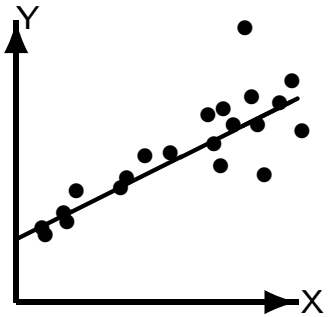
Violation of specific assumptions in scatter diagrams



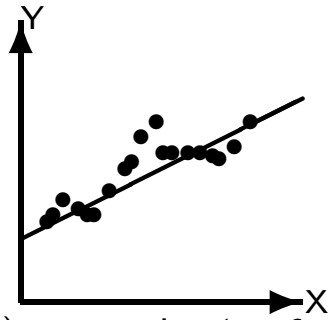
(a): assumption 3 violated



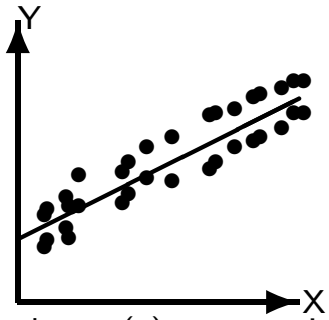
(b): assumption 2 violated



(c): assumption 5 verletzt



(d): assumption 1 or 6 violated



(e): assumption 8 violated

The variance von $\hat{\beta}$

Under assumptions 1–6, OLS is the BLUE estimator with smallest variance. For $\hat{\beta}$, this variance is given by

$$\mathrm{var}(\hat{\beta}) = \frac{\sigma_u^2}{n\widehat{\mathrm{var}}^*(X)},$$

using the n–weighted empirical variance of X

$$\widehat{\mathrm{var}}^*(X) = \frac{1}{n}\sum_{t=1}^{n}(x_t - \bar{x})^2.$$

Proof by direct evaluation of expectation operators. Assumptions 5 and 6 are used in the proof.

Interpretation of the variance formula for $\hat{\beta}$

$$\frac{\sigma_u^2}{n\widehat{\mathrm{var}}^*(X)}$$

depends on three factors:

1. σ_u^2: smaller variation in the errors implies that the data points are closer to the regression line and the estimation becomes more precise;
2. $\widehat{\mathrm{var}}^*(X)$: larger variation in the regressor variable implies a a good and informative design and the estimation becomes more precise;
3. n: larger sample, better estimate (consistency).

The variance of $\hat{\alpha}$

Under assumptions 1–6, the variance of the OLS estimate for the regression constant is given by

$$\operatorname{var}(\hat{\alpha}) = \frac{\sigma_u^2}{n} \frac{\frac{1}{n}\sum_{t=1}^{n} x_t^2}{\widehat{\operatorname{var}}^*(X)}.$$

If $\bar{x}$ is close to 0, i.e. if X is centered in 0, then the ratio is close to 1 and the intercept estimate is quite reliable. If the distributional center of X is far away from 0, the estimate becomes less precise.

The covariance of the coefficient estimates

Under assumption 1–6, the covariance of the estimates $\hat{\alpha}$ and $\hat{\beta}$ is given by

$$\operatorname{cov}(\hat{\alpha}, \hat{\beta}) = -\frac{\sigma_u^2}{n} \frac{\bar{x}}{\widehat{\operatorname{var}}^*(X)}.$$

For positive $\bar{x}$, this is negative: a larger intercept balances out a smaller slope. For $\bar{x} = 0$, both estimates are uncorrelated.

Estimation of the error variance

The formulae for variance and covariance of the OLS estimator are not directly applicable, as the error variance σ_u^2 is unknown. Can we use the observed OLS residuals $\hat{u}_t^2$ to estimate it?

The residuals have a lower variance than the errors, naive estimates such as

$$\frac{1}{n}\sum_{t=1}^{n}\hat{u}_t^2, \quad \frac{1}{n-1}\sum_{t=1}^{n}\hat{u}_t^2,$$

systematically underestimate σ_u^2.

Unbiased variance estimator

Under the assumptions 1–7, the estimator

$$\hat{\sigma}_u^2 = \frac{1}{n-2} \sum_{t=1}^{n} \hat{u}_t^2$$

will be unbiased for σ_u^2. (Assumption 7 is used, otherwise the denominator is 0.)

$n-2$ corresponds to the concept of 'degrees of freedom': n observations, 2 parameters are used up in the OLS regression.

$\hat{\sigma}_u^2$ is a *biased* estimator for the variance of the *residuals*.

The root $\hat{\sigma}_u$ is often called the (*standard error of regression*).

Standard errors of regression coefficients

The feasible variance estimate of $\hat{\beta}$ is

$$\hat{\sigma}_{\beta}^2 = \frac{\hat{\sigma}_u^2}{n\widehat{\mathrm{var}}^*(X)},$$

which, under assumptions 1–7, provides an unbiased estimate for $\mathrm{var}\hat{\beta} = \sigma_{\beta}^2$. An analogous argument applies to σ_{α}^2 and to the covariance.

The standard error or the estimated standard deviation is the square root of this value. This estimator is not unbiased for σ_{β}.

Homogeneous linear regression

Sometimes, it may look attractive (for empirical or theoretical reasons) to constrain the *inhomogeneous* regression

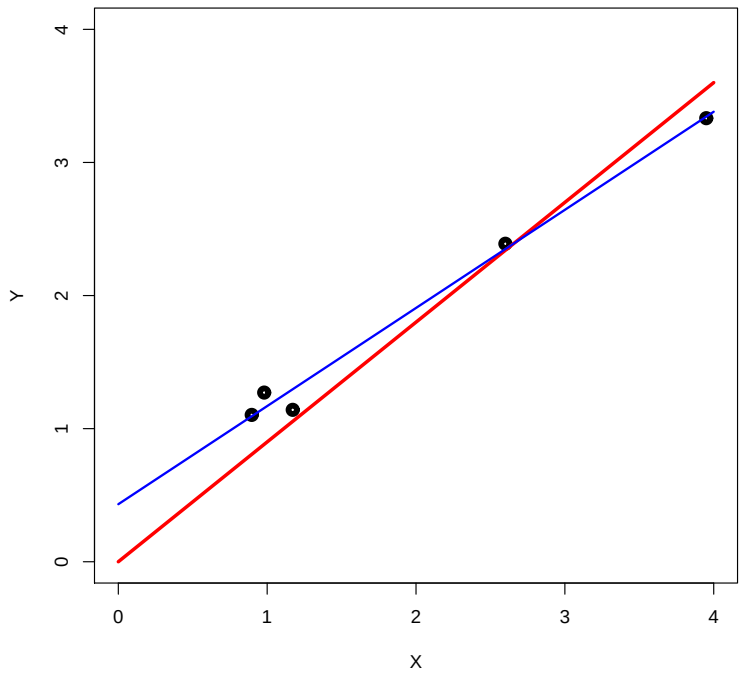
$$y_t = \alpha + \beta x_t + u_t$$

and to replace it by the *homogeneous* regression,

$$y_t = \beta x_t + u_t$$

which forces the regression line through the origin.

Homogeneous regression: Example



Blue line = OLS regression line $\hat{\alpha} + \hat{\beta}X$; red line = result of homogeneous OLS regression $\check{\beta}X$.

The homogeneous OLS estimator

The task to minimize the sum of squared vertical distance under the constraint that the regression line passes through the origin, can be written in symbols

$$\min Q_0(\beta) = \sum_{t=1}^{n} (y_t - \beta x_t)^2,$$

which yields the solution

$$\check{\beta} = \frac{\sum_{t=1}^{n} x_t y_t}{\sum_{t=1}^{n} x_t^2}.$$

Inhomogeneous OLS in the homogeneous model

Suppose assumptions 2–6 hold, and also assumption 1H

$$y_t = \beta x_t + u_t$$

instead of assumption 1. Then, the usual inhomogeneous OLS estimator $\hat{\beta}$ is still unbiased, but it is not BLUE, as it does not fulfill the condition $\alpha = 0$. All variance estimators remain valid, as (1H) is a special case of assumption 1.

Homogeneous OLS in the inhomogeneous model

Assuming $\alpha \neq 0$, if we estimate using $\check{\beta}$, then this estimator is generally biased and inconsistent because of

$$\begin{aligned} \mathrm{E}\check{\beta} &= \frac{\sum_{t=1}^{n} \mathrm{E}\{x_t(\alpha + \beta x_t + u_t)\}}{\sum_{t=1}^{n} x_t^2} \\ &= \beta + \alpha \frac{\bar{x}}{\sum_{t=1}^{n} x_t^2}. \end{aligned}$$

If $\bar{x} = 0$, i.e. if X vary around 0, then the 'incorrect' estimator is unbiased.

A game against nature

		true model	
		intercept	no intercept
estimated model	with intercept	very good	not efficient
	without intercept	biased, inconsistent	very good

In any case of doubt, the risk of a certain inefficiency will weight less, and the inhomogeneous regression will be preferred. This is recommended anyway, for reasons of compatibility and comparability with other specifications.

The t–statistic

In order to determine the significance of regression coefficients, the value of $\hat{\beta}$ can be compared to its standard error $\hat{\sigma}_\beta$. If **assumptions 1–8** hold, o.c.s. that the t–statistic

$$t_\beta = \frac{\hat{\beta}}{\hat{\sigma}_\beta},$$

under the null hypothesis '$H_0 : \beta = 0$', will be t–distributed with $n - 2$ degrees of freedom.

Similarly, one may also test '$H_0 : \alpha = 0$' using the corresponding quotient t_α, which is again, under H_0, distributed t_{n-2}.

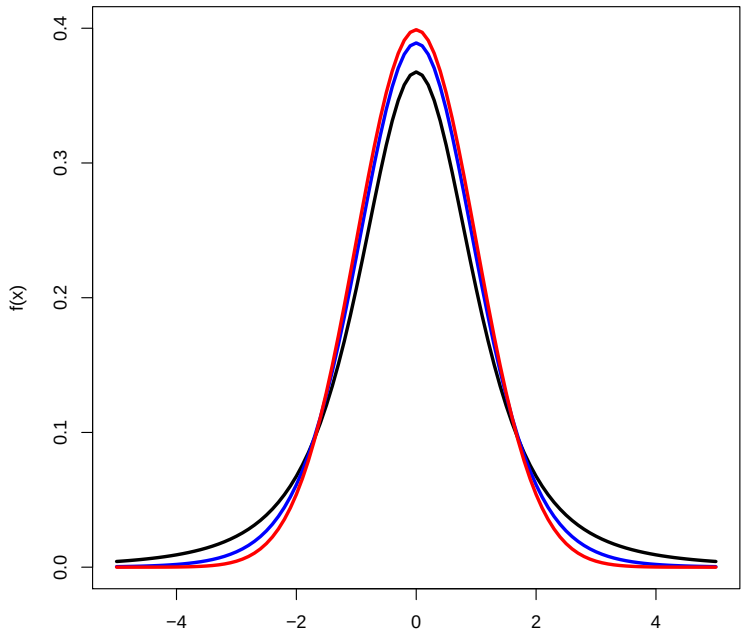
Common mistakes: '$H_0 : \hat{\beta} = 0$' and '$H_0 : \beta$ insignificant' are no useful null hypotheses.

Why is t_β t–distributed?

A proof proceeds along the following steps:

1. The coefficient estimate $\hat{\beta}$ is normally distributed $N(\beta, \sigma^2_\beta)$;
2. the standardized variance estimate $(n-2)\hat{\sigma}^2_\beta/\sigma^2_\beta$ is χ^2–distributed with $n-2$ degrees of freedom;
3. numerator and denominator are independent; if A is normal N(0,1) distributed and if B is χ^2_m–distributed and if A and B are independent, then $A/\sqrt{B/m}$ is distributed t_m.

t–distribution and normal distribution



Density of the t_3–distribution (black), of the t_{10}–distribution (blue), and of the N(0,1) normal distribution (red). For $n > 30$, the significance points of N(0,1) are used in the t–test almost exclusively.

t–Test if assumption 8 is invalid

If assumption 8 does not hold and n is 'large', then the t–statistic is approximately normal N(0,1) distributed. Usually, $n > 30$ is seen as a criterion for large samples here.

Because the two-sided 5% points of the N(0,1) are around ± 1.96, many empirical researchers use the t–statistic for a thumb-rule evaluation only, in the sense that it is significant when it is greater two in absolute value.

If assumption 8 does not hold and the sample is small, then the t–test is unreliable.

t–test on '$H_0 : \beta = \beta_0$'

One may be interested in testing whether β corresponds to a specific pre-specified value (Example: if the propensity to consume equals 1). The null hypothesis '$H_0 : \beta = \beta_0$' is tested using the t–statistic

$$\frac{\hat{\beta} - \beta_0}{\hat{\sigma}_\beta},$$

which again, under H_0, is distributed t_{n-2}.

This statistic is not supplied automatically, but it is easy to determine from the formula $t_\beta - \beta_0/\hat{\sigma}_\beta$.

Statistical regression programs and the t–test

- Often, $\hat{\beta}$, $\hat{\sigma}_\beta$, and the ratio t_β are supplied, increasingly even the p–value of the t–test;
- The software assumes assumptions 1–8 to hold and routinely uses the t–distribution, even when assumption 8 is invalid: take care;
- The p–value refers to the two-sided test against '$H_A : \beta \neq 0$': for the one-sided test (for example '$H_A : \beta > 0$') you have to adjust.

Coefficient of determination: idea

One may be interested in a descriptive statistic that quantifies to what degree Y is described linearly by X. The empirical coefficient of correlation

$$\widehat{\text{corr}}(X,Y) = \frac{\frac{1}{n}\sum_{t=1}^{n}(x_t - \bar{x})(y_t - \bar{y})}{\sqrt{\frac{1}{n}\sum_{t=1}^{n}(x_t - \bar{x})^2 \frac{1}{n}\sum_{t=1}^{n}(y_t - \bar{y})^2}}$$

does just that. For random variables(!) X and Y, it estimates the correlation

$$\text{corr}(X,Y) = \frac{\text{cov}(X,Y)}{\sqrt{\text{var}(X)\text{var}(Y)}},$$

which measures linear dependence. In regression analysis, it is customary to use the square of this statistic and to call it R^2.

OLS decomposes the dependent variable

OLS estimation decomposes y_t into an explained and an unexplained component:

$$y_t = \hat{\alpha} + \hat{\beta} x_t + \hat{u}_t = \hat{y}_t + \hat{u}_t.$$

$\hat{u}_t$ is called the residual. $\hat{y}_t$ is the systematic, explained part of y_t, and it is often called the *predictor*.

Variance decomposition

The second normal equation implies

$$\sum_{t=1}^{n} x_t \hat{u}_t = 0,$$

and, because of the first normal equation, $\sum_{t=1}^{n} \hat{u}_t = 0$. It follows that

$$\widehat{\mathrm{cov}}(\hat{U}, X) = \frac{1}{n} \sum_{t=1}^{n} (\hat{u}_t - \bar{\hat{u}})(x_t - \bar{x}) = 0.$$

However, $\hat{y}_t$ is only a linear transformation of x_t, hence it is uncorrelated with the residuals. It follows that

$$\sum_{t=1}^{n} (y_t - \bar{y})^2 = \sum_{t=1}^{n} (\hat{y}_t - \bar{y})^2 + \sum_{t=1}^{n} \hat{u}_t^2.$$

Notation for the variance components

Instead of

$$\sum_{t=1}^{n}(y_t - \bar{y})^2 = \sum_{t=1}^{n}(\hat{y}_t - \bar{y})^2 + \sum_{t=1}^{n}\hat{u}_t^2$$

one may write in short

$$TSS = ESS + RSS,$$

i.e. *total sum of squares* equals *explained sum of squares* plus *residual sum of squares*. [**Attention**: these abbreviations have not been standardized, many authors use instead of RSS acronyms such as SSR, ESS, SSE, and similarly for our ESS.] The total variance has been decomposed into an explained and an unexplained part.

Definition of R^2

The coefficient of determination R^2 is defined as the ratio

$$R^2 = \frac{ESS}{TSS} = \frac{\sum_{t=1}^{n}(\hat{y}_t - \bar{y})^2}{\sum_{t=1}^{n}(y_t - \bar{y})^2} = 1 - \frac{RSS}{TSS}.$$

Obvious properties:

- All sums of squares are non-negative, hence $0 \leq R^2 \leq 1$;
- $R^2 = 0$ is very 'bad', nothing is explained;
- $R^2 = 1$ is very 'good', the residuals are 0 and all points are on the regression line.

Properties of R^2

- The so defined R^2 is exactly the square of the empirical correlation of Y and X (easy to show);
- R^2 is the product of $\hat{\beta}$ and the OLS–coefficient in the inverse regression of X on Y;
- R^2 is meant to be a *descriptive statistic*, not an estimator for the correlation of X and Y, which does not exist because of assumption 4;
- there is no universally accepted rule, which value should be the minimum R^2 for the regression to be 'acceptable' (in cross-section data, often $R^2 \approx 0.3$; in time-series data, often $R^2 \approx 0.9$).

The adjusted R^2

If, in spite of its descriptive intention, R^2 is interpreted as an estimator for the correlation of random variables X and Y, then it is heavily upward biased. The statistic (after Wherry and Theil)

$$\bar{R}^2 = 1 - \frac{n-1}{n-2}(1 - R^2)$$

is also biased, but its bias is smaller.

- $\bar{R}^2$ is a tool for model selection in the multiple model with k regressors, where the denominator is replaced by $n - k$;
- $\bar{R}^2$ can become negative: indicator for bad regression.

R^2 as a test statistic

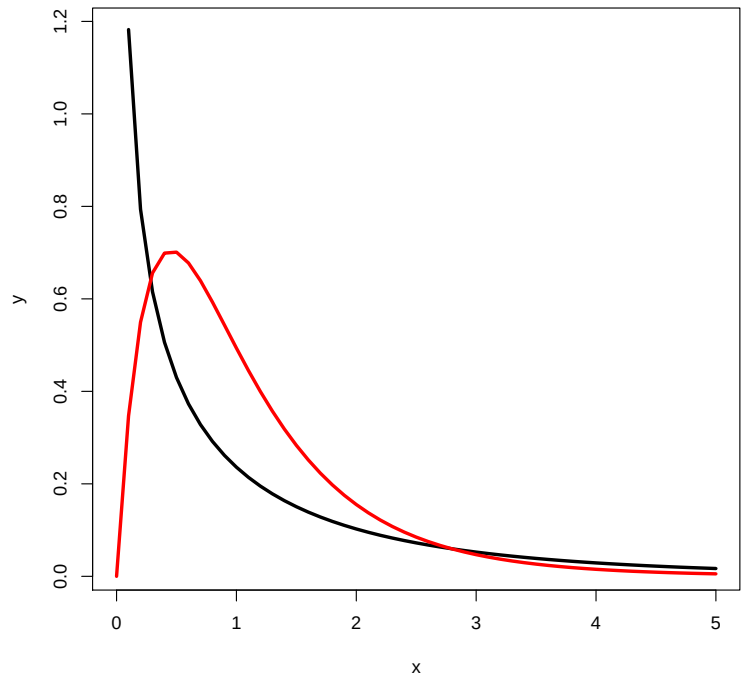
As a function of the data y_t, which are realizations of random variables according to assumptions 1–8, R^2 is a *statistic*, and hence a random variable with a probability distribution. Even under '$H_0 : \beta = 0$', however, R^2 does not have a useful or accessible distribution.

The simple transformation

$$F_c = \frac{R^2 (n-2)}{1-R^2} = \frac{ESS}{RSS} (n-2),$$

however, under this H_0 and assumptions 1–8, is F–distributed with $(1, n-2)$ degrees of freedom.

The F–distribution



Density of the distributions $F(1, 20)$ (black) and $F(4, 20)$ (red). In the simple regression model, only $F(1, .)$ distributions occur for the statistic F_c.

Why is F_c $F(1, n-2)$–distributed?

- The $F(m, n)$–distribution after SNEDECOR (the 'F' honors FISHER) is the distribution of a random variable $(A/m)/(B/n)$, if A and B are independent χ^2–distributed with m respectively n degree of freedom: therefore, numerator degrees of freedom m and denominator degrees of freedom n;
- the explained and the unexplained shares ESS and RSS are independent (because uncorrelated and assumption 8); under H_0, ESS is $\chi^2(1)$–distributed, RSS is distributed $\chi^2(n-2)$;
- all F–distributions describe non-negative random variables, rejection of H_0 at large values (always one-sided test).

F_c in the simple regression model

- F_c is provided by all regression programs, often with p–value;
- the F-total–test tests the same H_0 as the t–test for β, and even $F_c = t_\beta^2$, thus F_c has no additional information in the simple model;
- F_c becomes more important in the multiple model.