Econometric Theory

Econometric Theory

http://en.wikibooks.org/wiki/Econometric_Theory
This Book Is Generated By Wb2PDF
using
RenderX XEP, XML to PDF XSL-FO Formatter

18-04-2011

Table of Contents

Econometric Theory

Econometrics is the branch of economics concerned with the use of mathematics to describe, model, prove, and predict economic theory and systems.

This book can be considered to be three parts.

Chapters 1-4
An introduction and mathematical base needed to perform basic and more advanced econometrics.

Chapters 5, 6
The basics of bivariate and multivariate regression analysis.

Chapters 7-16
Applications of basic econometrics and advanced topics.

Contents

References

Ramanathan, Ramu. "Introductory Econometrics with Applications." South-western Thomson Learning, 2002.

Resources

Wikipedia has related information at ***Econometrics***

- Gretl Gnu Regression, Econometrics and Time Series Library, open source and free software for econometric calculus.

Introduction

The word *econometrics* means "economic measurement". The purpose of econometrics is to quantify and verify predictions from economic theory. It is a mixture of economic theory, mathematical economics, and statistics.

Apart from econometrics, there are two subjects closely related to econometrics. Firstly, mathematical economics is concerned with expressing economic theory in equations. It is algebra of the kind where variables can be subtracted, added, multiplied and divided in an equation. But it is not concerned with measurability or empirical verification. Secondly, we have economic statistics whose concern is to collect, process and present economic data. Many developed countries have government agencies whose job it is to collect and distribute economic data, such as the U.S. Bureau of Labor Statistics, which publishes data on unemployment, inflation, productivity, etc. The Bureau releases this information to the public and it is often used by other levels of government, social scientists, and interested laymen.

Lastly, it is important for econometricians to distinguish between **observational** data and **experimental** data. The use of observational data has three implications: firstly, the econometrician often uses data without knowledge of how the data were sampled; secondly, since the econometrician has no knowledge of how the data were generated, i.e. how people and companies behaved, there is uncertainty as to which economic or statistical model best represents these data; and thirdly, in order to obtain a suitable model that can be used for inference, econometrics is concerned with avoiding misspecification, or else the results of the model may be wrongly interpreted. [1]

Prerequisites

This Wikibook assumes that you have had some experience with the following topics:

- Set Theory
- Real Analysis

and are well-versed in the following:

1. This is the somewhat the view expressed by Judge et al. (1988, p. 5-6) and Gujarati (2003, p.5) Gujarati (2003, p. 23)

- Statistics
- Linear Algebra

1. ↑ This is the somewhat the view expressed by Judge et al. (1988, p. 5-6) and Gujarati (2003, p.5)

Bibliography

- Judge et al. (1988). *Introduction to the Theory and Practice of Econometrics - 2nd ed*. John Wiley & Sons. ISBN 0-471-62414-4.
- Gujarati, D.N. (2003). *Basic Econometrics, International Edition - 4th ed.*. McGraw-Hill Higher Education. ISBN 0-07-112342-3.

Methodology of Econometrics

How does the econometrician go ahead in analysing an economic theory? What is needed is a **methodology**, i.e. a step-by-step procedure. This is similar to other social sciences.

Theory

A theory should have a prediction. In statistics and econometrics, we also speak of **hypothesis**. One example is **the marginal propensity to consume** (MPC) proposed by Keynes. Other examples could be that lower taxes would increase growth, or maybe that it would increase economic inequality, and that introducing a common currency has a positive effect on trade.

Specification of the Mathematical Model

This is where the algebra enters. We need to use mathematical skills to produce an equation. Assume a theory predicting that more schooling increases the wage. In economic terms, we say that the return to schooling is positive. The equation is:

$$Y = \beta_1 + \beta_2 X,$$

where Y is the variable for wage and β_1 is a constant and β_2 is the coefficient of schooling, and X is a measurement of schooling, i.e. the number of years in school. We also call β_1 intercept and β_2 a slope coefficient.

Normally, we would expect both β_1 and β_2 to be positive.

Specification of the Econometric Model

Here, we assume that the mathematical model is correct but we need to account for the fact that it may not be so. We add an **error term**, *u* to the equation above. It is also called a **random (stochastic) variable**. It represents other non-quantifiable or unknown factors that affect Y. It also represents mismeasurements that may have entered the data. The econometric equation is:

$Y = \beta_1 + \beta_2 X + u$.

The error term is assumed to follow some sort of statistical distribution. This will be important later on.

Obtain Data

We need data for the variables above. This can be obtained from government statistics agencies and other sources. A lot of data can also be collected on the Internet in these days. But we need to learn the art of finding appropriate data from the ever increasing huge loads of data

Estimation of the model

Here, we quantify β_1 and β_2, i.e. we obtain numerical estimates. This is done by statistical technique called **regression analysis**.

Hypothesis Testing

Now we go back to the part where we had economic theory. The prediction was that schooling is good for the wage. Does the econometric model support this hypothesis. What we do here is called **statistical inference (hypothesis testing)**. Technically speaking, the β_2 coefficient should be greater than 0.

Forecasting

If the hypothesis testing was positive, i.e. the theory was concluded to be correct, we **forecast** the values of the wage by **predicting** the values of education. For example, how much would someone earn for an additional year of schooling? If the X variable is the years of schooling, the β_2 coefficient gives the answer to the question.

Use for Policy Recommendation

Lastly, if the theory seems to make sense and the econometric model was not refuted on the basis of the hypothesis test, we can go on to use the theory for policy recommendation. If your theory was really good, then maybe you will earn the Nobel Prize of Economics.

Bibliography

- Gujarati, D.N. (2003). *Basic Econometrics, International Edition - 4th ed.*. McGraw-Hill Higher Education. pp. 3-13. ISBN 0-07-112342-3.

Types of Econometrics

There are two branches of econometrics: **theoretical econometrics** and **applied econometrics**.

The former is concerned with methods, both their properties and developing new ones. It is closely related to mathematical statistics, and it states assumptions of a particular method, its properties. It is particularly important to know when they are not fulfilled.

Applied econometrics does what it says: it applies econometric theory to branches within economics, such as unemployment figures, portfolio theory, demand and supply functions, et.c.

- Gujarati, D.N. (2003). *Basic Econometrics, International Edition - 4th ed.*. McGraw-Hill Higher Education. pp. 12. ISBN 0-07-112342-3.

Statistical Packages

Many statistical programs are available for data analysis and statistical computing, including a variety of commercially distributed software: Eviews, SAS, Stata, RATS, TSP, WinBUGS and SPSS, and open source packages, such as Gretl and R.

Wikibooks

- Statistical Analysis: an Introduction using R
- R Programming
- Stata
- SAS

Wikipedia Links

- Econometrics
- Statistics Software

Software

- SHAZAM
- Eviews
- SAS

- Stata
- RATS
- TSP
- WinBugs
- SPSS

External Links

- SHAZAM Software
- R Project
- SAS Software
- SPSS Software
- Stata Software
- Econometric Links
- Free Statistics

Linear Algebra

Linear Algebra

It's assumed that the reader understands basic linear algebra. Throughout this course it is expected that the reader has knowledge of matrix algebra as it will be important knowledge to follow what is called the **matrix approach to the linear regression model**. All that is covered in the link below.

This is the previous content

(I want a book entirely written in matrix algebra so that it will be easier to write formulae and is shorter, despite the complexity. I will try to make some contributions to the book Linear Algebra - Whisky Brewer)

(Is it possible to put the Linear Algebra book at the final chapter of this book on econometric theory?)

It's assumed that the reader understands basic linear algebra. however it's of particular usefulness to know about the properties of a very common class of matrix called a Projection Matrix. These matrices are called projection matrices because they have a geometrical interpretation of projecting data from one plane to another. {Insert more on projection matrices} {Define P and Q} {Show the geometrical implication.}

Important Terms and Concepts of Regression Analysis

In the Introduction chapter, it was mentioned that regression is the tool of econometrics. Here, some important terms and concepts are introduced.

What is Regression?

Historical Meaning of Regression

The *historical meaning* of the term **regression** was coined by statistician Francis Galton. He observed that tall parents tended to have short children and short parents tended to have long children. It seemed as though the height of people was heading to the average of the population across generations. His friend Karl Pearson collected statistics on the heights of individual family members. He found that sons of short fathers tended to be taller than the average height of those fathers and sons of long fathers thus turned out to be shorter than their fathers.

Modern Interpretation of Regression

It can be said that **regression analysis** is used to model relationships between variables and determine the magnitude of those relationships.. Regression analysis is really a more accurate description for regression, but regression is fully comprehensible as an abbreviated term and we will use that one henceforth.

Examples

A number of examples where regression can be used is:[1]

- 1. Galton was interested in finding out why average height was stable over generations. We could also say that we are interested in finding out why the average height of sons would be different from the average height of fathers. A hypothetical distribution can be seen in the *scatter diagram* to the right. In it, you can also see the *regression line* which shows the average height of sons given the average height of the fathers.

1. Examples 1-3 from Gujarati (2003, p.18-21)

{scatter diagram}

- 2. A monopolist wants to know how demand adjusts for changes in price or output. This allows him to maximise profit. This price responsiveness is called *price elasticity*.

- 3. The *Phillips curve* describes the effect of changes in unemployment to the unemployment rate. This could be used to predict a suitable level of inflation that is suitable to the public opinion on unemployment.

- 4. A policy maker may want to see if there is a relationship between expenditure on police and the crime rate in a particular area. It can be useful if we want to know how to tackle crime more efficiently by allocating more money where crime is higher.

Notes

1. ↑ Regression analysis
2. ↑ Examples 1-3 from Gujarati (2003, p.18-21)

Wikipedia links

Regression analysis

Bibliography

- Gujarati, D.N. (2003). *Basic Econometrics, International Edition - 4th ed.*. McGraw-Hill Higher Education. pp. 16-21. 0-07-112342-3.

Regression versus Causation and Correlation

Statistical and deterministic Relationships

A *deterministic relationship* implies that there is an exact mathematical relationship or dependence between variables. An example in physics is Newton's law of gravity:

$$F = k\left(\frac{m_1 m_2}{r^2}\right)$$

, where F, the force, is proportional to a constant, k, the mass of two objects, m_1 and m_2, and inversely to the square of the distance.

A *random* or *stochastic relationship* allows that there is some variation in a relationship. This is where probability distributions will enter later on. The relationship may not be exact due to

- **measurement errors**
- **reporting errors**
- **computing errors**
- **other influence**,

etc. One example is crop yield relative to rain fall. We may not be able to measure the amount of rain accurately (measurement error), we round off to one decimal point (reporting error), there is a bug in the computer software that computes the sum (computing error) and there might be other factors such the quality of fertilisers, the quality of the earth and pollution (other influences).

Regression versus Causation

Regression deals with dependence amongst variables within a model. But it cannot always imply causation. For example, we stated above that rainfall affects crop yield and there is data that support this. However, this is a one-way relationship: crop yield cannot affect rainfall. It means there is no cause and effect reaction on regression if there is no causation.

In short, we conclude that a **statistical relationship does not imply causation**.[1]

Regression versus Correlation

Correlations form a branch of analysis called **correlation analysis**, in which the *degree of linear association* is measured between two variables. If we calculate the correlation between crop yield and rainfall, we might obtain an estimate of, say, 0.69. This is reasonably high to prove that there is a mathematical relationship between them.

There is a distinction in how we regard the relationship between rainfall and crop yield. In statistics, both variables are assumed to be variables with random error in them. Both are treated on an equal footing and there is no distinction between them.

In regression analysis, crop yield is the dependent variable and rainfall is the explanatory variable, according to our theory. The distinction is that the dependent variable has got no random component, all values are fixed from this distribution

This will be important in {section on mismeasurement}.

Notes

{references/}

Wikipedia links

Correlation does not imply causation

1. This is the somewhat the view expressed by Judge et al. (1988, p. 5-6) and Gujarati (2003, p.5) Gujarati (2003, p. 23)

Bibliography

- Gujarati, D.N. (2003). *Basic Econometrics, International Edition - 4th ed.*. McGraw-Hill Higher Education. pp. 22-24. ISBN 0-07-112342-3.

1. ↑ Gujarati (2003, p. 23)

Terminology and Notation

Terminology and Notation

Above we discussed dependent variables and explanatory variables. There are many synonyms of various terms and these are no exception. The dependent variable can also be called:

- Explained variable
- Predictand
- Regressand
- Response
- Endogenous
- Outcome
- Controlled variable

and the explanatory variable:

- Independent variable
- Predictor
- Regressor
- Stimulus
- Exogenous
- Covariate
- Control variable

If we have one explanatory variable. We get what is called **two-variable (simple) regression analysis**. If we include more than one explanatory variable, it is called **multiple regression analysis**.

The **error term** is assumed to be a **random** or **stochastic variable**. It implies that it can take any value within a given probability distribution and there are no other factors determining it. If other influences were known, it would no longer be random.

Bibliography

- Gujarati, D.N. (2003). *Basic Econometrics, International Edition - 4th ed*. McGraw-Hill Higher Education. pp. 24-25. ISBN 0-07-112342-3.

Data

Econometrics in absence of data would not exist. All data can be classified into a category and that can be important as the success of good econometric work depends on the nature, sources and limitations of the data used.

Types of Data

There are three types of data: **time series**, **cross-section**, and a combination of them is called and **pooled data**.

Time series data of a variable have a set of observations on values at different points of time. They are usually collected at fixed intervals, such as daily, weekly, monthly, annually, quarterly, etc. Time series econometrics has applications in macroeconomics, but mainly in financial economics where it is used for price analysis of stocks, derivatives, currencies, etc.

Cross-section data are collected at the same point of time for several individuals. Examples are opinion polls, income distribution, data on GNP per capita in all European countries, etc.

Pooled data is a mixture of time series data and cross-section data. One example is GNP per capita of all European countries over ten years.

Panel, longitudinal or micropanel data is a type that is pooled data of nature. The difference is that we measure over the same cross-sectional unit for individuals, households, firms, etc. This branch of econometrics is called **microeconometrics**.

Sources of Data

There are many sources of data and it can be very time-consuming to find all the data needed. In fact, finding data can take up more of the time than analysis in a project. Some sources are governmental agencies (Eurostat), international agencies (the International Monetary fund (IMF), the World Bank, the World Health Organisation (WHO), etc.), firms, etc.

The Internet has become the newest source of information over the last decade. There are lots of economic and financial to obtain.

Data Accuracy

Because data in the social sciences are seldom generated under controlled conditions, there will always be unknown influences. This makes it difficult to obtain qualitative data for research. This was mentioned in the previous section.

Measurement of Scale Data fall into four categories which are important to know:

- **Ratio scale** refers to quantities such as ratios X_2 / X_1 and distances $X_2 - X_1$. There can be ordering of the data where comparisons are meaningful, such as
  $X_2 \leq X_1$
  .
- **Interval scale** refers to distances as mentioned above
- **Ordinal scale** refers to an order that is not quantitative but qualitative. We can also say that there is a "natural order" of grouping different categories. For example, there are different income classes (high, medium, low), sizes (large, medium, small), etc.
- **Nominal scale** refers to states but there is no ordering amongst them. For instance, genders (male, female), materials (paper, plastics, wood), etc.

Notes

{references/}

Bibliography

- Gujarati, D.N. (2003). *Basic Econometrics, International Edition - 4th ed.*. McGraw-Hill Higher Education. pp. 25-31. ISBN 0-07-112342-3.

Statistical Concepts

This section introduces some statistical concepts that will be useful later in this course. The contents are shown without proof and that can be found in many excellent books on that subject.

Summation and Product Operators

To sum a series of variables *x*, the Greek capital letter sigma *Σ* is used:

$\Sigma_{i=1}^{n} x_i = x_1 + x_2 + \ldots + x_n$
.

There are some properties concerning the summation operator *Σ*:

1.
$\Sigma_{i=1}^{n} k = nk$
, where k is a constant.

2.
$\Sigma_{i=1}^{n} kx_i = k\Sigma_{i=1}^{n} x_i$
, where k is a constant.

3.
$\Sigma_{i=1}^{n} (a + bx_i) = na + b\Sigma_{i=1}^{n} x_i$
, where *a* and *b* are constants. This is a result of rules 1 and 2 above.

4.
$\Sigma_{i=1}^{n} (x_i + y_i) = \Sigma_{i=1}^{n} x_i + \Sigma_{i=1}^{n} y_i$
,

The double summation operator is used to sum up twice for the same variable:

The double summation operator has the following properties:

1.
$\Sigma_{i=1}^{n} \Sigma_{j=1}^{m} x_{ij} = \Sigma_{i=1}^{m} \Sigma_{j=1}^{n} x_{ij}$
. The order of the summation signs is interchangeable.

2.
$\Sigma_{i=1}^{n} \Sigma_{j=1}^{m} x_i y_j = \Sigma_{i=1}^{n} x_i \Sigma_{j=1}^{m} y_j$
.

3.
$\Sigma_{i=1}^{n}\Sigma_{j=1}^{m}(x_i + y_j) = \Sigma_{i=1}^{n} x_i \Sigma_{j=1}^{m} x_i j + \Sigma_{i=1}^{n} x_i \Sigma_{j=1}^{m} y_{ij}$

.

4.
$$\begin{aligned}[\Sigma_{i=1}^{n} x_i]^2 &= \Sigma_{i=1}^{n} {x_i}^2 + 2\Sigma_{i=1}^{n-1}\Sigma_{j=i+1}^{n} x_i x_j \\ &= \Sigma_{i=1}^{n} {x_i}^2 + 2\Sigma_{i<j} x_i x_j\end{aligned}$$

.

Finally, the product operator *Π* is defined as:
$\Pi_{i=1}^{n} x_i = x_1 \cdot x_2 \cdots x_n$

.

Bibliography

- Gujarati, D.N. (2003). *Basic Econometrics, International Edition - 4th ed.*. McGraw-Hill Higher Education. pp. 869-870. ISBN 0-07-112342-3.

Sample Points, Sample Space and Events

The population or the **sample space** refers to the set of all possible outcomes of an experiment that is random. If you toss a coin there are two possible outcomes: head (H) or tail (T). If you toss two coins, there are four outcomes: HH, HT, TH and TT, where the first letter refers to the outcome of the first tossing and the second to that of the second tossing. Each outcome is called a **sample point**.

An **event** is a subset of a sample space. If we want to list all sample points with the occurrence of at least one H, we have HH, HT and TH. Any events are called **mutually exclusive** if the occurrence of one event is independent of the occurrence of another event. For instance, if we got a head on one occasion, we could not simultaneously get a tail.

Events are **exhaustive** if there is no other possible event. For instance, the events two heads (HH), one head and one tail (HT and TH), and two tails (TT) are all possible outcomes if tossing the coin twice. Thus, the list of events is exhaustive.

Bibliography

- Gujarati, D.N. (2003). *Basic Econometrics, International Edition - 4th ed.*. McGraw-Hill Higher Education. pp. 870. ISBN 0-07-112342-3.

Probability Density Function (PDF)

Probability Mass Function of a Discrete Random Variable

A **probability mass function** *f(x)* (PMF) of *X* is a function that determines the probability in terms of the input variable *x*, which is a discrete **random variable** (rv).

A pmf has to satisfy the following properties:

- $$f(x) = \begin{cases} P(X = x_i) & \text{for } i = 1, 2, \cdots, n \\ 0 & \text{for } x \neq x_i \end{cases}$$

- The sum of PMF over all values of *x* is one:

$$\sum_i f(x_i) = 1.$$

Probability Density Function of a Continuous Random Variable

The continuous PDF requires that the input variable *x* is now a continuous rv. The following conditions must be satisfied:

- All values are greater than zero.

- The total area under the PDF is one
- The area under the interval [a, b] is the total probability within this range

Joint Probability Density Functions

Joint **pdf**s are ones that are functions of two or more random variables. The function

is the **continuous joint probability density function**. It gives the joint probability for x and y.

The function

is similarly the **discrete joint probability density function**

Marginal Probability Density Function

The marginal **PDFs** are derived from the joint PDFs. If the joint pdf is integrated over the distribution of the X variable, then one obtains the marginal PDF of y, $f(y)$. The continuous marginal probability distribution functions are:

and the discrete marginal probability distribution functions are

Conditional Probability Density Function

Statistical Independence

- Gujarati, D.N. (2003). *Basic Econometrics, International Edition - 4th ed.*. McGraw-Hill Higher Education. pp. 870-877. ISBN 0-07-112342-3.

Asymptotic Convergence

Wikipedia has related information at ***Convergence of random variables***

Asymptotic Convergence

Modes of Convergence

Convergence in Probability

Convergence in probability is going to be a very useful tool for deriving asymptotic distributions later on in this book. Alongside convergence in distribution it will be the most commonly seen mode of convergence.

Definition

A sequence of random variables

$\{X_n; n = 1, 2, \cdots\}$

converges in probability to *X* if:

$$\forall \epsilon, \delta > 0,$$
$$\exists N \text{ s.t. } \forall n \geq N,$$
$$\Pr\{|X_n - X| > \delta\} < \epsilon$$

an equivalent statement is:

$$\forall \delta > 0,$$
$$\lim_{n \to \infty} \Pr\{|X_n - X| > \delta\} = 0$$

This will be written as either

$X_n \xrightarrow{p} X$

or

$\text{plim} X_n = X$

.

Example

We'll make an intelligent guess that this series converges in probability to the degenerate random variable η. So we have that:

Therefore our definition for convergence in probability in this case is:

$$\forall \epsilon, \delta > 0,$$
$$\exists N \quad \text{s.t.} \forall n \geq N,$$
$$\Pr\{|X_n - \eta| > \delta\} \leq \Pr\{|X_n - \eta| > 0\} = \Pr\{X_n = \theta\} = \tfrac{1}{n} < \epsilon$$

So for any positive values of

$\epsilon \in \mathbb{R}$

we can always find an

$N \in \mathbb{N}$

large enough so that our definition is satisfied. Therefore we have proved that

$X_n \xrightarrow{p} \eta$

.

Convergence Almost Sure

Almost-sure convergence has a marked similarity to convergence in probability, however the conditions for this mode of convergence are stronger; as we will see later, convergence almost surely actually implies that the sequence also converges in probability.

Definition

A sequence of random variables

$\{X_n; n = 1, 2, \cdots\}$

converges almost surely to the random variable *X* if:

$$\forall \delta > 0,$$

$$\lim_{n\to\infty} \Pr\{\bigcup_{m\geq n} |X_m - X| > \delta,\} = 0$$

equivalently

$$\Pr\{\lim_{n\to\infty} X_n = X\} = 1$$

Under these conditions we use the notation

$X_n \xrightarrow{a.s.} X$

or

$\lim_{n\to\infty} X_n = X \,\mathrm{a.s.}$

.

Example

Let's see if our example from the convergence in probability section also converges almost surely. Defining:

we again guess that the convergence is to η. Inspecting the resulting expression we see that:

$$\Pr\{\lim_{n\to\infty} X_n = \eta\} = 1 - \Pr\{\lim_{n\to\infty} X_n \neq \eta\} = 1 - \Pr\{\lim_{n\to\infty} X_n = \theta\} \geq 1 - \lim_{n\to\infty} \tfrac{1}{n} = 1$$

Thereby satisfying our definition of almost-sure convergence.

Convergence in Distribution

Convergence in distribution will appear very frequently in our econometric models through the use of the Central Limit Theorem. So let's define this type of convergence.

Definition

A sequence of random variables

$\{X_n; n = 1, 2, \cdots\}$

asymptotically converges in distribution to the random variable *X* if

$F_{X_n}(\zeta) \to F_X(\zeta)$

for all continuity points.

$F_{X_n}(\zeta)$

and

$F_X(\zeta)$

are the cumulative density functions of X_n and X respectively.

It is the distribution of the random variable that we are concerned with here. Think of a students-T distribution: as the degrees of freedom, n, increases our distribution becomes closer and closer to that of a gaussian distribution. Therefore the random variable

$$Y_n \sim t(n)$$

converges in distribution to the random variable

$$Y \sim N(0,1)$$

(n.b. we say that the random variable

$$Y_n \xrightarrow{d} Y$$

as a notational crutch, what we really should use is

$$f_{Y_n}(\zeta) \xrightarrow{d} f_Y(\zeta)$$

/

Example

Let's consider the distribution X_n whose sample space consists of two points, $1/n$ and 1, with equal probability (1/2). Let X be the binomial distribution with $p = 1/2$. Then X_n converges in distribution to X.

The proof is simple: we ignore 0 and 1 (where the distribution of X is discontinuous) and prove that, for all other points a,

$$\lim F_{X_n}(a) = F_X(a)$$

. Since for $a < 0$ all Fs are 0, and for $a > 1$ all Fs are 1, it remains to prove the convergence for $0 < a < 1$. But

$$F_{X_n}(a) = \frac{1}{2}([a \geq \frac{1}{n}] + [a \geq 1])$$

(using Iverson brackets), so for any a chose $N > 1/a$, and for $n > N$ we have:

$$n > 1/a \rightarrow a > 1/n \rightarrow [a \geq \frac{1}{n}] = 1 \wedge [a \geq 1] = 0 \rightarrow F_{X_n}(a) = \frac{1}{2}$$

So the sequence

$$F_{X_n}(a)$$

converges to

$$F_X(a)$$

for all points where F_X is continuous.

Convergence in R-mean Square

Convergence in R-mean square is not going to be used in this book, however for completeness the definition is provided below.

Definition

A sequence of random variables

$\{X_n; n = 1, 2, \cdots\}$

asymptotically converges in r-th mean (or in the L^r norm) to the random variable *X* if, for any real number $r > 0$ and provided that

$E(|X_n|^r) < \infty$

for all n and

$r \geq 1$

,

Cramer-Wold Device

The Cramer-Wold device will allow us to extend our convergence techniques for random variables from scalars to vectors.

Definition

A random vector

$\mathbf{X}_n \xrightarrow{d} \mathbf{X} \iff \lambda^{\mathrm{T}}\mathbf{X}_n \xrightarrow{d} \lambda^{\mathrm{T}}\mathbf{X} \quad \forall \|\lambda\| \neq 0$

.

Relationships betweeen Modes of Convergence

Law of Large Numbers

Central Limit Theorem

Let

$X_1, X_2, X_3, ...$

be a sequence of random variables which are defined on the same probability space, share the same probability distribution D and are independent. Assume that both the expected value μ and the standard deviation σ of D exist and are finite.

Consider the sum

$S_n = X_1 + ... + X_n$

. Then the expected value of

S_n

is nμ and its standard error is σ n1/2. Furthermore, informally speaking, the distribution of Sn approaches the normal distribution N(nμ,σ2n) as n approaches ∞.

Continuous Mapping Theorem

Slutsky's Theorem

Statistical Inference

Statistical inference is the attempt of making a statement about a population using only sample data that is a subset of that population. This is necessary since in most situations it is either impractical or expensive (or both) to collect data on the whole population.

Statistical inference is broken up into the following parts:

- Identification of the problem.
- Identification of the population.
- Specification of the statistical properties of the parameters in question.

Hypothesis Testing

Basic Concepts

To conduct a successful hypothesis test, the following are required:

- Testable Hypothesis

We need to have a null (H_0) and alternate (H_1) hypothesis.

- Feasible test statistic

A test statistic is a random variable whose value for given sample data determines whether the null is rejected or retained. It is feasible when:

-
    - Its probability distribution is known when the null hypothesis (H_0) is true.
    - Its value can be calculated from the given sample data

- Decision rule

A decision rule clearly delineates the:

-
    - Rejection region - the set of values of the test statistic for which H_0 is to be rejected.
    - Non-rejection region - the set of values of the test statistic for which the H_0 is to be retained.

Procedure for testing a hypothesis

1. Formulate H_0 and H_1.
2. Specify the test statistic and its distribution.
3. Calculate the sample value of the test statistic under H_0 for the given sample data.
4. Select a significance level (α) and determine the corresponding critical values (for the particular distribution).
5. Apply the decision rule and state the conclusion (or inference) implied by the sample value of the test statistic.

Notes

- The null hypothesis will always be the position where there is an equality (either strong or weak), and the alternate hypothesis will have the inequality.

Matrix Algebra

Matrices

A matrix is an array of numbers arranged into rows and columns. Some examples of matrices are,

When describing matrices we indicate the number of rows first, then the number of columns. For example, the matrix *C* with two rows and four columns is said to be a

2×4
matrix.

It is standard notation to name matrices with capital letters and to use lower case letters with subscripts to identify particular entries in a matrix.

For example, to identify the entry in row 1 and column 3 of matrix *A* we would write a_{13}. To indicate that this entry is a six we would write the equation $a_{13} = 6$.

Two matrices are considered to be **equal** only if they are the same size and every pair of corresponding elements are equal.

A **column matrix** is a matrix with only one column. Similarly, a **row matrix** has only one row.

Vectors

A **vector** is an object often defined by a long list of properties. However, for now we will avoid the more complicated definition, and just say that a vector is an ordered list of numbers. Later we will see that vectors can really be much more.

An ordered pair, (x,y), that is used to identify a point in the plane can be considered to be a vector.

Similarly, an ordered triple, (x,y,z) is a vector.

Obviously, row and column matrices can also be considered to be a vector.

It is common to name vectors using variables with arrows above.

For example, we might write

$$\vec{v} = (2, 3, 5, -4), \text{ or } \vec{w} = \begin{bmatrix} 4 \\ 5 \\ 0 \end{bmatrix}$$

.

For the most part, it will convenient to think of vectors as column matrices.

Definitions

Identity Matrix

The identity matrix, with a size of n, is an n-by-n square matrix with ones on the main diagonal and zeros elsewhere. It is commonly denoted as I_n, or simply by I if the size is immaterial or can be easily determined by the context.

$$I_1 = [1] \quad I_2 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \quad I_3 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \quad I_n = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{bmatrix}$$

The most important property of the identity matrix is that, when multiplied by another matrix, A, the result will be A

$$AI_n = A$$

and

$$I_nA = A$$

.

Matrix Differentiation

There are a few simple rules for matrix differentiation. These allow much econometrics to be done in matrix form, which can be simpler and far less cumbersome than using nested summation signs.

Differentiating an inner product with respect to a vector

Let **a** be a given column vector and let **x** be a column choice vector (a vector of values to be chosen). Their transposes can be denoted as **a'** and **x'** . Then the derivative of their inner product, which is a scalar, is a column vector:

$$\partial \mathbf{a'x}/\partial \mathbf{x} = \partial \mathbf{x'a}/\partial \mathbf{x} = \mathbf{a}.$$

Differentiating a quadratic form with respect to a vector

Let **A** be a matrix, either symmetric or non-symmetric, and consider the quadratic form **x'Ax**, which is itself a scalar. The derivative of this quadratic form with respect to the vector **x** is the column vector

$$\partial \mathbf{x'Ax}/\partial \mathbf{x} = (\mathbf{A}+\mathbf{A'})\mathbf{x}.$$

But in econometrics, almost always the matrix in the quadratic form will be symmetric. If **A** is indeed symmetric, the formula can be simplified to

$$\partial \mathbf{x'Ax}/\partial \mathbf{x} = 2\mathbf{Ax}.$$

Application to Ordinary Least Squares

Perhaps the most basic concept in econometrics is ordinary least squares, in which we choose the regression coefficients so as to minimize the sum of squared residuals (mispredictions) of the regression. Suppose the regression model is

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e},$$

where **y** is an n×1 vector of observed values of the dependent variable, **X** is an n×k matrix in which each column is an n×1 vector of observed values of one of the k independent variables ($k<n$), **β** is the k×1 vector of estimated response coefficients to be chosen, and **e** is the n×1 vector of residuals.

The reader may confirm that the sum of squared residuals—the sum of the squares of the elements of **e**—is given by

$$\mathbf{e'e = (y - X\beta)'(y - X\beta) = y'y - y'X\beta - \beta'X'y + \beta'X'X\beta = y'y - 2\beta'(X'y) + \beta'(X'X)\beta}$$

where the last equality holds because $\mathbf{y'X\beta}$ is a scalar and thus is its own transpose. Note that $\mathbf{(X'y)}$ is a $k\times1$ vector and $\mathbf{(X'X)}$ is a $k\times k$ matrix. Using the above rules for differentiation, we have

$$\partial \mathbf{e'e}/\partial\boldsymbol{\beta} = \mathbf{-2X'y + 2(X'X)\beta},$$

where we have used the fact that $\mathbf{X'X}$ is symmetric. When we express the first-order condition by equating this vector derivative to the zero vector, we obtain

$$\mathbf{(X'X)\beta = X'y},$$

which can be solved for the choice vector $\boldsymbol{\beta}$ as

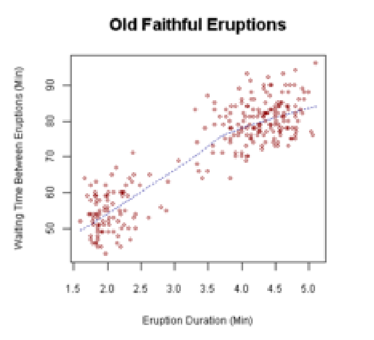
$$\boldsymbol{\beta} = \mathbf{(X'X)^{-1}X'y}.$$

This is the ordinary least squares estimator. Note that so long as $\mathbf{X}$ has full column rank, the expression $\mathbf{(X'X)^{-1}}$ exists and $\mathbf{(X'X)}$ is positive definite, ensuring that the second-order conditions are satisfied.

Classical Normal Linear Regression Model (CNLRM)

Econometrics is all about causality. Economics is full of theory of how one thing causes another: increases in prices cause demand to decrease, better education causes people to become richer, etc. So to be able to test this theory, economists find data (such as price and quantity of a good, or notes on a population's education and wealth levels). However, when this data is placed on a plot, it rarely makes neat lines that are presented in introductory economics text books.

Data always comes out looking like a cloud, and without using proper techniques, it is impossible to determine if this cloud gives any useful information. Econometrics is a tool to establish correlation and hopefully later, causality, using collected data points. We do this by creating an **explanatory function** from the data. The function is linear model and is estimated by *minimizing* the squared distance from the data to the line. The distance is considered an *error term*. This is the process of **linear regression.**



This plot is of waiting time between eruptions of Old Faithful and duration of eruptions, but it might as well be a plot of the supply line for sweater sales

The Line

The point of econometrics is establishing a correlation, and hopefully, causality between two variables. The easiest way to do this is to make a line. The slope of the line will say "if we increase x by so much, then y will increase by this much" and we have an intercept that gives us the value of y when x = 0.

The equation for a line is y = a + b*x (note:a and b take on different written forms, such as alpha and beta, or beta(0) beta(1) but they always mean "intercept" and "slope").

The problem is developing a line that fits our data. Because our data is scattered, an non-linear, it is impossible for this simple line to hit every data point. So we set our line up so that it minimizes the errors (and we need to actually minimize the squared errors). We adjust our first line, or *explanatory function* to have an error term, so that, given x and the error term, we can correctly come up with the right y.

y = a + b*x + error

Basic Example

After blowing our grant money on a vacation, we are pressured by the University to come up with some answers about movements of the sweater industry in Minnesota. We do not have much time, so we only collect data from two different clothing stores in Minnesota on two separate days. Fortunately for us, we get data from one day in the summer and one day in the winter. We ask both stores to tell us how many sweaters they have sold and they tell us the truth. We are looking to see how weather (temperature -- independent variable) affects how many sweaters are sold (dependent variable).

We come up with the following scatter plot.

Now we can add a line (a function) to tell us the relationship of these two variables. We will minimize the sum of errors, and see what we get. The distance to the line from the cold side is +15 and the difference from the hot side to the line is -15. When we add them together, we get 15 - 15 = 0. 0 error, our line must be perfect!

Notice that our line fits the data well. It the differences between the data and the function are evenly distributed (
$\sum \epsilon = 0$
). So there is a relationship between temperature and sweater sales. "Hot weather increases Sweater Sales" will be the title of our famous paper! But it will be wrong, and we'll probably fired from our job at the University.

If we had only minimized distances between the line and the data! Well, here is a plot with a estimated line that does just that. To do this, we minimize the sum of *squared* errors.

Once we minimize the distances between the line and the data, we have a better fit and we can declared that "cold weather increases Sweater Sales" (
$\sum \epsilon^2 = 0$
)

Basic bivariate model

Our basic model is a line that best fits the data.

$Y = \alpha + \beta X_i + \varepsilon_i$

Where
$Y_i \in [1, n] and X_i \in [1, n]$

, α and β are unknown parameters that must be estimated.

$\epsilon_i \in [1, n]$

is the unobserved error term. This term is an iid random variable. **regression coefficient**.

Notes on Notation:

Symbol	**meaning**
Y	Dependant Variable
X	Independent Variable(s)
α,β	Regression Coefficients
ε,*u*	Error or Disturbance term
^	Hat: Estimated

Properties of the error term

The error term, also known as the **disturbance term**, is the unobserved random component which explains the difference between Y_i and $\alpha + \beta X_i$. This term is the combination of four different effects.

1. *Omitted Variables:* in many cases, it is hard to account for every variability in the system. Although cold weather increases sweater sales, but also, the price of heating oil may also have an affect. This was not accounted in our original model, but may be explained in our error term.

2. *Nonlinearities:* The actual relationship may not be linear, but all we have is a linear modeling system. At 30 degrees, 10 people buy sweaters. At 20 degrees 40 people by sweaters. At 10 degrees 80 people buy sweaters. In our model, the error term will account the nonlinearity.

3. *Measurement errors:* Sometimes data was not collected 100% correctly. The store told us that 10 people bought sweaters that day, but after we talked to them, 4 more people bought sweaters. The relationship still exists, but we have some error collected in the error term.

4. *Unpredictable effects:* No matter how well the economic model is specified, there will always be some sort of stochastic that affects it. These effects will be accounted by the error term.

Looking again at our OLS line in our sweater story, we a can have a look at our error terms. The error, is the distance from our data Y and our estimate Ŷ. We get an equation from this:

$\hat{\epsilon}_i = Y_i - \hat{Y}_i = Y_i - \alpha - \beta X_i$

next page>

Assumptions of Classical Linear Regression Model

The estimators that we create through linear regression give us a relationship between the variables. However, performing a regression does not automatically give us a reliable relationship between the variables. In order to create reliable relationships, we must know the properties of the estimators

$\hat{\alpha}, \hat{\beta}$

and show that some basic assumptions about the data are true.

Unbiasedness

Under the following four assumptions, OLS is unbiased. This means that:

$E(\hat{\alpha}) = \alpha$

$E(\hat{\beta}) = \beta$

Linearity

The model must be linear in the parameters.
The parameters are the coefficients on the independent variables, like α and β. These must be linear, so having β^2 or e^{β} would violate this assumption.

Sample Variation

The x_is cannot all have the same value.

Random Sampling

The x_i values must be randomly selected. In other words, there is no correlation between two different x values: $Cov(x_i,x_j) = 0$ for

$i \neq j$

.

Zero Conditional Mean

The mean of the error terms, given a specific value of the independent variable x_i, is zero. $E(\varepsilon_i \mid X_i) = 0$.

Efficiency of OLS

Given the following two assumptions, OLS is the **B**est **L**inear **U**nbiased **E**stimator (BLUE). This means that out of all possible linear unbiased estimators, OLS gives the precise estimates of α and β.

With the third assumption, OLS is the **B**est **U**nbiased **E**stimator (BUE), so it even beats non-linear estimators. Also given this assumption,

$\hat{\alpha}$

is distributed according to the Student's t-distribution about α, and

$\hat{\beta}$

is distributed in such a way about β.

No Heteroskedasticity

The variance of the Error terms are constant. $Var(u_i \mid x_i) = \sigma^2$. This means that the variance of the error term u_i does not depend on the value of x_i. If this is the case, the error terms are called **homoskedastic**. This is not always the case in economic data, for example the variation in a person's wage will vary with their level of education -- someone who is a high-school dropout will not have much variation in their wage, where people with Ph.D.s may see very different wages.

No Serial Correlation

The error terms are independently distributed so that their covariance is 0.
$Cov(u_i, u_j|x_i, x_j) = 0 \forall i \neq j$
.

Normally Distributed Errors

The error terms are *normally distributed*.
$u_i\ N(0, \sigma^2)$

Properties of OLS Estimators

OLS estimators have the following properties

1. Unbiasedness
2. Minimum Variance
3. Efficiency

Unbiasedness

Suppose that the population size is 100 for anything that we are studying. We use samples of size 10 to estimate the α and β of the population. Everytime we use a different sample (a different set of 10 unique parts of the population), we will get a different α and β.

With the OLS method of getting α and β, we get a situation wherein after repeated attempts of trying out different samples of the same size, the mean (average) of all the α and β from the samples will be equal to the actual α and β of the population as a whole.

Basically, this means that if you do the exercise over and over again with different parts of the population, and then you find the mean for all the answers you get, you will have the correct answer (or you will be very close to it).

Minimum variance

This property is what makes the OLS method of estimating α and β the best of all other methods. In the previous section you learned about the principle of unbiasedness. It is possible, however, that two different methods of estimation could be unbiased. How do we decide which one to use in our econometric model? This is where the principle of variance is involved. Variance is a measure of how far the different α and β are from their mean.

An estimator (a function that we use to get estimates), that has a lower variance is one whose individual data points are those that are closer to the mean. This estimator is statistically more likely than others to provide accurate answers. The OLS estimator is one that has a minimum variance.

Efficiency

This property is simply a way to determine which estimator to use.

- An estimator that is unbiased but does not have the minimum variance is not good.
- An estimator that has the minimum variance but is biased is not good
- An estimator that is unbiased and has the minimum variance of all other estimators is the best (efficient).

The OLS estimator is an efficient estimator.

Proofs of properties of β1

Linearity

To be linear,
$\hat{\beta}_1$
must be a linear function of Y_i, as shown below

where k_i is a constant, at any given observation 'i'.

Proof

From the deviation-from-means form of the solution of the OLS Normal Equation for
$\hat{\beta}_1$
, we have

$= \frac{\sum x_i Y_i}{\sum x_i^2}$
, since
$\sum x_i = 0$
.

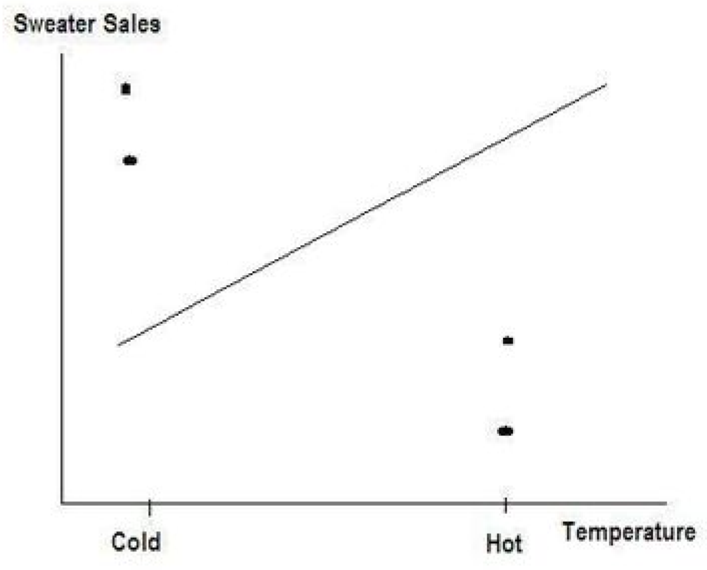
$= \sum k_i Y_i$
, where
$k_i = \frac{x_i}{\sum x_i}$
, which is a constant for any given 'i'-value.

Ordinary Least Squares (OLS)

Ordinary Least Squares or **OLS** is one of the simplest (if you can call it so) methods of linear regression. The goal of OLS is to closely "fit" a function with the data. It does so by minimizing the sum of squared errors from the data.

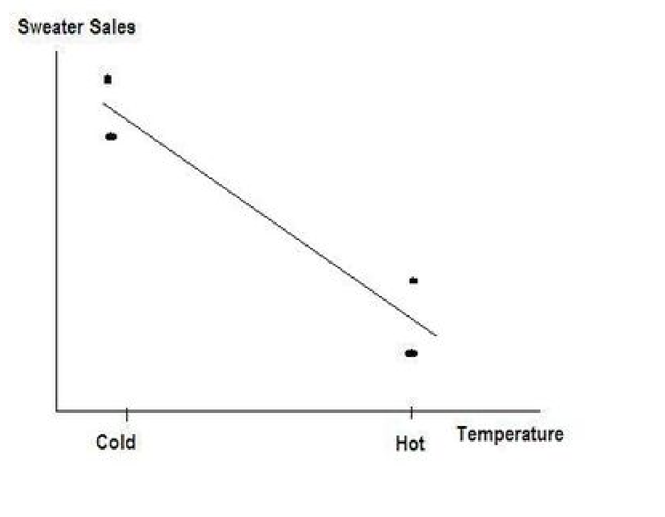
Why we Square Errors before Summing

Notice that we are not just trying to minimize the sum of errors, but rather the sum of *squared* errors. Lets take a brief look at our sweater story again.



Model A

model	data point	error from line
A	1	5
A	2	10
A	3	-5
A	4	-10
B	1	3
B	2	-3
B	3	3
B	4	-3



Model B

Notice that the Sum of Model A is 5 + 10 − 5 − 10 = 0 and that the Sum of Model B is 3 − 3 + 3 − 3 = 0

Both Models sum to 0 and both are great fits! **NO!!**

So to account for the signs, whenever we sum errors, we square the terms first.

The Model

There is a common aspect of these two models. There is an intercept term α, and a slope term β (note, some other textbooks use β_0 instead of α and β_1 instead of β, this is a much better approach once we move to multivariate formulas). We can represent an arbitrary single variable model with the formula: $y_i = \alpha + \beta x_i + u_i$ The y-values are related to the x-values given this formula. We use the subscript i to denote an observation. So y_1 is paired with x_1, y_2 with x_2, etc. The u_t term is the error term, which is the difference between the effect of x_i and the observed value of y_i.

Unfortunately we don't actually know the values of α,β or u_t. We have to approximate them. We can do this by using the ordinary least squares method. The term "least squares" means that we are trying to minimize the sum of squares, or more specifically we are trying to minimize the squared error terms. Since there are two variables that we need to minimize with respect to (α and β), we have two equations:

$$f = \Sigma u_i^2 = \Sigma(y_i - \alpha - \beta x_i)^2$$
$$\frac{\partial f}{\partial \alpha} = -2\Sigma(y_i - \alpha - \beta x_i) = 0$$

$$\frac{\partial f}{\partial \beta} = -2\Sigma(y_i - \alpha - \beta x_i)x_i = 0$$

Call the solutions to these equations

$\hat{\alpha}$

and

$\hat{\beta}$

. Solving we get:

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}$$
$$\hat{\beta} = \frac{\Sigma(x_i - \bar{x})y_i}{\Sigma(x_i - \bar{x})^2}$$

Where

$$\bar{y} = \frac{\Sigma y_i}{n}$$

and

$$\bar{x} = \frac{\Sigma x_i}{n}$$

. Computing these results can be left as an exercise.

It is important to know that

$\hat{\alpha}$

and

$\hat{\beta}$

are not the same as α and β because they are based on a single sample rather than the entire population. If you took a different sample, you would get different values for

$\hat{\alpha}$

and

$\hat{\beta}$

. Let's call

$\hat{\alpha}$

and

$\hat{\beta}$

the OLS estimators of α and β. One of the main goals of econometrics is to analyze the quality of these estimators and see under what conditions these are good estimators and under which conditions they are not.

Once we have

$\hat{\alpha}$

and

$\hat{\beta}$

, we can construct two more variables. The first is the fitted values, or estimates of *y*:

$$\hat{y}_i = \hat{\alpha} + \hat{\beta}$$

The second is the estimates of the error terms, which we will call the **residuals**:

$\hat{u}_i = y_i - \hat{y}_i$

These two variables will be important later on.

Normal Equations Proof

Below is the proof of the **Normal Equations** for OLS.

The goal of OLS is to minimize the sum of squared error terms to find the best fit, also called the Residual Sum of Squares (RSS). This is denoted by

$\sum \hat{\epsilon_i}^2$

.

Defining the RSS

Known:

$\hat{\epsilon_i} = Y_i - \hat{Y_i} = Y_i - \alpha - \beta X_i$

RSS =

$\sum \hat{\epsilon_i}^2$

=

Differentiate the RSS (so that we can then minimise it)

$min_\alpha \sum \hat{\epsilon_i}^2$

=

$min_\beta \sum \hat{\epsilon_i}^2$

=

So we have two equations:

and

$\sum (Y_i - \hat{\alpha} - \hat{\beta} X_i)(-X_i) = 0$

(The two(2) here is divided from both sides)

setting them both equal to
$\sum Y_i$

We get

$\sum Y_i = n\hat{\alpha} + \hat{\beta} \sum X_i$
(This is the first OLS Normal Equation)

and

$\sum Y_i X_i = \hat{\alpha} \sum X_i + \hat{\beta} \sum X_i^2$
(This is the second OLS Normal Equation)

Solve the Normal Equations

Divide the first equation by n

$$\frac{1}{n} \sum Y_i = \hat{\alpha} + \frac{1}{n} \hat{\beta} \sum X_i$$

Leaves us with (
$\sum W_i \frac{1}{n} = \bar{W}$
)

Now we know how to get α(hat), we can work on β(hat)

$$\sum Y_i X_i = \hat{\alpha} \sum X_i + \hat{\beta} \sum X_i^2 = [\bar{Y} - \hat{\beta} \bar{X}] \sum X_i + \hat{\beta} \sum X_i^2 = [\frac{(\sum X_i)(\sum Y_i)}{n}] + \hat{\beta} [\sum X_i^2 - \frac{(\sum X_i)^2}{n}]$$

We can move β(hat) to one side

- n \bar{X} \bar{Y}

And now we have our Normal equations for OLS.

Since we have two equations and two unknowns, we are able to solve for them (
$\hat{\alpha}, \hat{\beta}$
).

Multiple Regression Analysis

Our first regressions (MLE and OLS) were bivariate. Our lines were simple, two variable lines. However, in most economic data, there are a multitude of possible independent things that can effect a dependent variable. So we can expand our explanatory functions to allow multiple independent variables.

Instead of our functions looking like $Y = \alpha + \beta X_i + \varepsilon_i$, our functions look like

$$Y_i = \beta_0 + \beta_1 X_{1,i} + \beta_2 X_{2,i} + \cdots + \beta_n X_{n,i} + \epsilon_i$$

By adding more variables and data to our model, we can hopefully get a better fit and understanding of the dependent variable. However, with the added variables come added problems that will misguide our model.

Goodness of Fit

When we move to the multiple regression case, our goodness of fit looks much like it previously did in the bivariate case. TSS = ESS + RSS. Our R^2 = ESS/TSS) = 1 - (RSS/TSS). We can still use our Coefficient of Determination, R^2 (R^2 = ESS/TSS = 1 - (RSS/TSS)), but there is a problem associated with it. R^2 will never decrease because of an addition of a variable, whether or not it helps us explain our dependent variable. When we add a new variable to the function, the ESS is calculated over a larger set of variables, and ESS will be less than or equal to what we had before. This will cause our R^2 to decrease increase, even if the addition of our new variable hurts our model. There is a tool to fix this problem. R^2 is replaced with *Adjusted R^2* which adjusts it for the added degrees of freedom. Adjusted R^2 is signified by adding a bar above the 'R.'

Inference

Inference is essentially the process of creating a hypothesis of the parameters that describe a population by testing the sample parameters (such as

$\hat{\alpha}$

and

$\hat{\beta}$

) that we already have from a sample of the population.

For example, you have a sample of size N, and you have created a model that can be used to predict changes in the units of the sample. The parameters of this model would be estimates of the actual parameters (such as α and β) for the entire population. Inference is the process used to determine, statistically, what the parameters would be for the whole population.

There are two basic ways of checking estimates for the purposes of statistical inference

- Interval estimation
- Hypothesis testing

To successfully use these methods, one must know two things:

- The probability distribution of the true parameter values (α and β) must be known.
- The formulae used must use data that can be found in the sample data (we can't successfully use unknown parameters in a function).

t-Test

A t-test involves the computation of a t-statistic, which is then compared to the critical values of a t-distribution for a given significance level.

A t-test is essentially the Z-statistic of a variable divided by the square root of an independent chi-square distribution divided by its own degrees-of-freedom. The resulting value is the t-statistic with the same degrees-of-freedom as the chi-squared distribution.

Therefore, the t-statistic of β_1 would be:

- Numerator:

- Denominator:

We know (as an implication of the last assumption of the CLRM) that

$$\frac{(N-2)\hat{\sigma^2}}{\sigma^2} \sim \chi^2[N-2]$$

Therefore,

$$\frac{\hat{\sigma^2}}{\sigma^2} \sim \frac{\chi^2[N-2]}{[N-2]} \Rightarrow \sqrt{\frac{\chi^2[N-2}{[N-2]}} \sim \frac{\hat{\sigma}}{\sigma}$$

Therefore, putting it all together we get,

Notes

- $se(\hat{\beta}_1) = \frac{\sigma}{(\sum X_i^2)^{1/2}}$

- $\hat{se}(\hat{\beta}_1) = \frac{\hat{\sigma}}{(\sum X_i^2)^{1/2}}$

F-Test

An F-test involves the computation of an F-statistic, which is then compared to the critical values of an F-distribution for a given significance and numerator and denominator degrees-of-freedom.

An F-statistic is calculated by dividing a chi-squared distribution divided by its degrees-of-freedom by another (independent) chi-squared distribution by its degrees-of-freedom. The resulting F-statistic has two degrees-of-freedom parameters, one each for the numerator and the denominator.

Therefore, the F-statistic for
$\hat{\beta}_1$
would be:

- Numerator:

We know (somehow) that $[Z(0,1)]^2 = \chi^2[1]$, therefore we set the numerator equal to:

- Denominator:

From the same implication of the last assumption of the CLRM as used by the t-test explanation,

Therefore, putting it all together gives us:

$$F(\hat{\beta}_1) = \frac{(\hat{\beta}_1 - \beta_1)^2(\sum X_i^2)/\sigma^2}{\hat{\sigma^2}/\sigma^2} = \frac{(\hat{\beta}_1 - \beta_1)^2}{\hat{\sigma^2}/\sum X_i^2} = \frac{(\hat{\beta}_1 - \beta_1)^2}{V\hat{a}r(\hat{\beta}_1)} \sim F[1, N-2]$$

Dummy Variables

Health insurance companies often charge differently for different types of people. They know off of their own data, that young adults are in less need of a doctor generally, and so they charge less for insurance. Their data shows that age and health costs are positively correlated and indeed one "causes" another. There are other demographics off of which they organize their prices of coverage. One is whether or not the customer is a smoker, and some even use gender. But how did they come to the conclusion that smokers cost more to cover, or that the cost of covering a man is different from a woman? These are not quantitative pieces of data, so they cannot be regressed, right? Well no, we can make it look like they are in fact quantitative, and not qualitative.

Dummy Variables

Dummy Variables or **Indicator Variables** are these qualitative data points manipulated to be quantitative. In the case of correlating health costs to smoking habits, we can say that a smoker is a *1* and a non-smoker is a *0*. Our dependent variable is health care costs.

Our model will look like this: $Y_i = \alpha + \beta D_i + \varepsilon_i$ where D is our dummy (Smoking) and Y is our Dependant (health care costs). Say that health care costs are \$50 for non-smokers, and \$60 for smokers, our model would then be $Y_i = 50 + 10D + \varepsilon_i$. When the person we are looking at is a non-smoker, D = 0, and when the person we are looking at is a smoker, D = 1.

We can regress with multiple pieces of information (variables) too. We can also mix our normal data with several dummies. $HeathCare_i = \beta_1 age_i + \beta_2 Smoke_i + \beta_3 Gender_i + \varepsilon_i$ (Gender = 1 for male, Gender = 0 for female)

Our estimated model off of the data would be
$HeathCare_i = \hat{\beta_0} age_i + \hat{\beta_1} Smoke_i + \hat{\beta_2} Gender_i + \epsilon_i$

The formula for a 29 year old male who does not smoke would be
$HeathCare_i = \hat{\beta_0} 29 + \hat{\beta_2} + \epsilon$

Our Dummy Variables can be more than just binary. Say health care companies found out that happiness can lead to higher health, and they wanted to use that in their price discrimination scheme. They can ask "how happy are you?, Very Happy, Kind of Happy, Sad." However, they will need to use two dummies for this move. D1 will be 1 and D2 will be 0 if "Very Happy", D1 will be 0 and D2 will be 1 if "Kind of Happy" and D1 will be 0 and D2 will be 0 if "sad."

To add this into our model we will have $HeathCare_i = \hat{\beta}_0 age_i + \hat{\beta}_1 Smoke_i + \hat{\beta}_2 Gender_i + \hat{\beta}_3 VeryHappy_i + \hat{\beta}_4 KindOfHappy_i + \epsilon_i$

Slope vs Intercept shifts

The Dummies can effect the model in two ways. The Dummy can either shift the intercept up or down, or shift the slope shallower or deeper. What has been described above are all intercept shifts. The line stayed neutral for non-smokers, and moved up for smokers. For a slope shift, the dummy is in the same term as the standard variable as in $Y_i = \alpha + \beta_1 X_i + \beta_2 D_i X_i + \varepsilon$ where if D = 1, $Y_i = \alpha + (\beta_1 + \beta_2) X_i + \varepsilon$ and if D = 0 $Y_i = \alpha + \beta_1 X_i + \varepsilon$

Note: The combination of the dummy and the standard in this case is an interaction term. It is often described as one variable as in $D_i X_i = Z_i$

Multicollinearity

While running regressions on multiple explanatory variables, there often is the problem of two variables having the same effects on the dependent variable. For example, say that the reading level of your father and the reading level of your mother can predict your eventual max reading level **(YourLevel = a + b*FatherLevel + c*MotherLevel + u)**. However, there probably is a high probability that your parents have similar reading levels, so that using both may be redundant. This is a problem known as **Multicollinearity.** The problem with this is bias in your estimators. Say we can find yours through your mother's level. **YourLevel = a + c*MotherLevel + u.** If we add on your father's level, we will increase the prediction for your level, although in real life your level will not increase at all.

Detecting Multicollinearity

Detecting multicollinearity can be more difficult than in the above example. But the first step that should be taken is an examination of the theory that is being tested. Is it redundant to have both mother's level and father's level? If this does not yield any results, probably because the theory is more complex, causing multicollinearity to be hidden, several econometric techniques can be used to find problems.

1) Large changes in the estimated regression coefficients when a predictor variable is added or deleted. Running the regression first with 'FatherLevel' and then without it may yield large variation, indicating that there is an error.

2) Non-significant results of simple linear regressions. Obviously if we find that with both FatherLevel and MotherLevel that neither are significant, than again there is something strange happening signaling possible multicollinearity.

3) Estimated regression coefficients have an opposite sign from predicted If in a regression with both FatherLevel and MotherLevel, **b** is positive, but **c** is negative; we know from theory that a higher reading level of the mother does not cause the child to be a worse reader. This is a possible sign of multicollinearity.

4)formal detection-tolerance or the variation inflation factor (VIF)

$$\mathrm{tolerance} = 1 - R^2, \quad \mathrm{VIF} = \frac{1}{\mathrm{tolerance}}.$$

A tolerance of less than 0.1 indicates a multicollinearity problem.

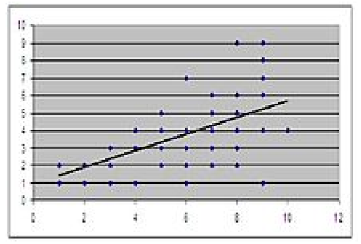
Heteroskedasticity

One of our CLR assumptions for linear regression is that our disturbance terms are **homoscedastic**, meaning they have equal scatter ($Var(\varepsilon) = \sigma^2$). However, there are times that regressions end up with **heteroscedastic** disturbance terms, meaning the scatters are unequal (

$Var(\epsilon) = \sigma_i^2$

).

Causes of Heteroscedasticity


Hetroscedastic, Cone shape

Heteroscedasticity are more common in cross-sectional data than in time series. It is usually due to a scale or size factor.

Example: In basic Keynesian economics, we assume that savings and income are determined by wealth and income. Agents that have more wealth and income are more likely to save, this will produce a hetroscedastic relationship.

Consequences of Hetroscedastity

1) OLS Coefficients are still unbiased for true value.

$E(\hat{\beta}) = \beta$

Unbiased coefficients depend on $E(\varepsilon) = 0, cov(x_i, \varepsilon_i) = 0$

So the regression is safe from hetroscedasity. on this assumption.

2) OLS Coefficients are not efficient. There exists an alternative to the OLS Coefficant that has a smaller variance than the OLS one.

$\exists_{\tilde{\beta}} st.var(\tilde{\beta}) < var(\hat{\beta}^{OLS}) where E(\tilde{\beta}) = \beta$

and therefore, is more efficient.

3) Hypothesis tests, in the presence of hetroscedasity of an OLS coefficient, based on the standard error of the coefficient are invalid.

Bias in estmated variance of OLS Estimators causes ineffiancies.

Recall:
$$\hat{\beta}^{OLS} = \beta + \frac{\sum(x_i + \bar{x})\epsilon_i}{\sum(x_i + \bar{x})^2}$$

When beta-hat is unbiased, the second term goes to zero. However, with hetroscedasicity, the second term is not zero.

Derive
$$var(\hat{\beta}^{OLS})$$
when
$$var(\epsilon) = \sigma_i^2$$
$$var(\hat{\beta}^{OLS}) = var(\frac{\sum(x_i + \bar{x})\epsilon_i}{\sum(x_i + \bar{x})^2})$$
$$= (\frac{1}{\sum(x_i - \bar{x})^2})^2 var(\sum(x_i - \bar{x})\epsilon)$$
$$= (\frac{1}{\sum(x_i - \bar{x})^2})^2 \sum var((x_i - \bar{x})\epsilon)$$
$$= (\frac{1}{\sum(x_i - \bar{x})^2})^2 \sum x_i - \bar{x})^2 var(\epsilon)$$
$$= (\frac{1}{\sum(x_i - \bar{x})^2})^2 \sum x_i - \bar{x})^2 \sigma_i^2$$

We can represent the differences in variances in the error term by the matrix Z

Serial Correlation

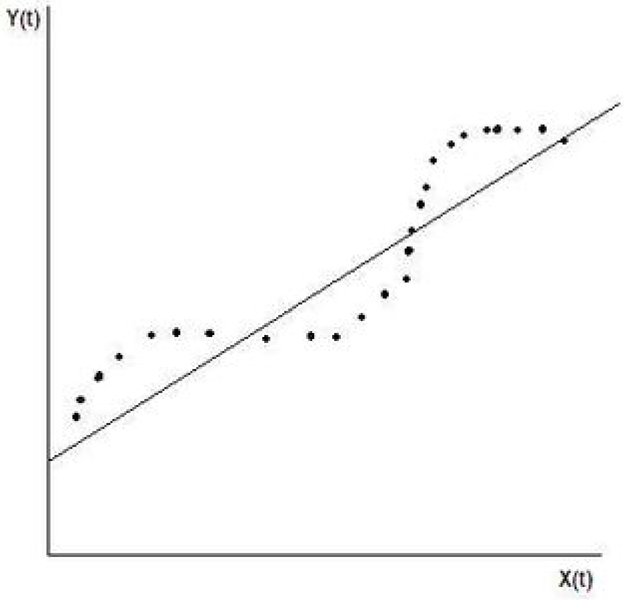
There are times, especially in time-series data, that the CLR assumption of $corr(\varepsilon_t,\varepsilon_{t-1}) = 0$ is broken. This is known in econometrics as **Serial Correlation** or **Autocorrelation**. This means that

$$corr(\epsilon_t, \epsilon_{t-1}) \neq 0$$

and there is a pattern across the error terms. The error terms are then not independently distributed across the observations and are not *strictly* random.

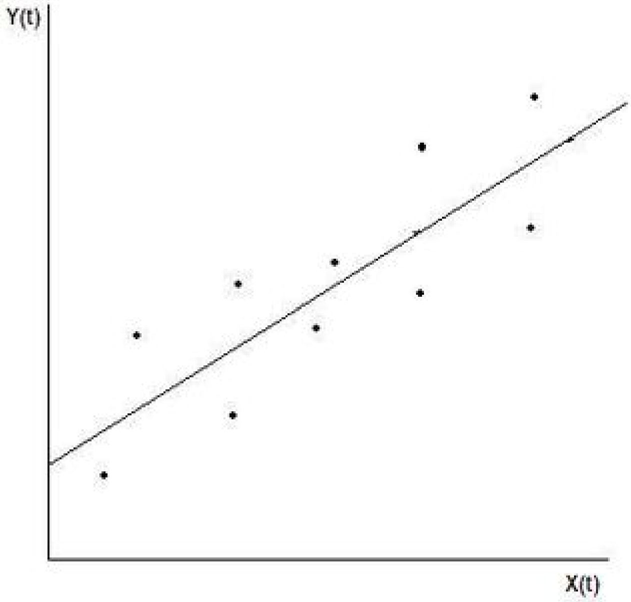
Examples of Autocorrelation

$corr(\varepsilon_t,\varepsilon_{t-1}) > 0$



Positive Autocorrelation

$corr(\varepsilon_t,\varepsilon_{t-1}) < 0$



Negative Autocorrelation

Functional Form

When the error term is related to the previous error term, it can be written in an algebraic equation. $\varepsilon_t = \rho\varepsilon_{t-1} + u_t$ where ρ is the autocorrelation coefficient between the two disturbance terms, and u is the disturbance term for the autocorrelation. This is known as an **Autoregressive Process.** $-1 < / rho = corr(\varepsilon_t, \varepsilon_{t-1}) < 1$The u is needed within the equation because although the error term is less random, it still has a slight random effect.

Serial Correlation of the Nth Order

The Autoregressive model:

- First order Autoregressive Process:$\varepsilon_t = \rho\varepsilon_{t-1} + u_t$
  - This is known as the first order autoregression, due to the error term only depending on the previous error term.
  - It is commonly displayed in textbooks as **AR(1)**

- nth order Autoregressive Process:

  $\epsilon_t = \rho_1\epsilon_{t-1} + \rho_2\epsilon_{t-2} + \cdots + \rho_n\epsilon_{t-n} + u_t$
  - Known as **AR(n)**

Causes of Autocorrelation

1. Spacial Autocorrelation

$corr(\epsilon_t, \epsilon_{t-1}) \neq 0$

Spacial Autocorrelation occurs when the two errors are specially and/or geographically related. In simpler terms, they are "next to each." **Examples:** The city of St. Paul has a spike of crime and so they hire additional police. The following year, they found that the crime rate decreased significantly. Amazingly, the city of Minneapolis, had not adjusted the police force, finds that they have a increase in the crime rate over the same period.

- Note: this type of Autocorrelation occurs over cross-sectional samples.

1. Inertia/Time to Adjust
   1. This often occurs in Macro, time series data. The US interest rate unexpectedly increases and so there is an associated change in exchange rates with other countries. Reaching a new equilibrium could take some time.

2. Prolonged Influences
   1. This is again a Macro, time series issue dealing with economic shocks. It is now expected that the US interest rate will increase. ##The associated exchange rates will slowly adjust up-until the announcement by the Federal Reserve and may overshoot the equilibrium.

3. Data Smoothing/Manipulation
   1. Using functions to smooth data will bring autocorrelation into the disturbance terms

4. Misspecification

1. A regression will often show signs of autocorrelation when there are omitted variables. Because the missing independent variable now exists in the disturbance term, we get a disturbance term that looks like: $\varepsilon_t = \beta_2 X_2 + u_t$ when the correct specification is $Y_t = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + u_t$

Consequences of Autocorrelation

The main problem with Autocorrelation is that it may make a model *look* better than it actually is.

list of consequences

1. Coefficients are still unbiased $E(\varepsilon_t) = 0, cov(X_t, u_t) = 0$
2. True variance of
   $\hat{\beta}$
   is increased by the presence of Autocorrelation.
3. Estimated Variance of
   $\hat{\beta}$
   is smaller due to Autocorrelation (biased downward).
4. A decrease with
   $se(\hat{\beta})$
   and an increase of the t-stats. This results in the estimator looking more accurate than it actually is.
5. R^2 becomes inflated.

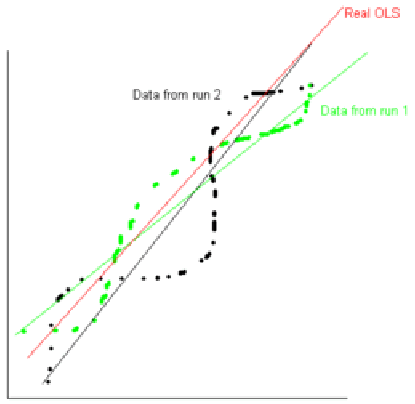
All of these problems result in hypothesis tests becoming invalid.

Testing for AC

1. View a graph of the Dependant variable against the error term (AKA, a residual scatter-plot).

2. Durbin-Watson test.
   1. Assume $\varepsilon_t = \varepsilon_{t-1}\rho + u_t$
   2. test H(0): $\rho = 0$ (no AC) against H(1): $\rho > 0$ (one-tailed test)
   3. Test Stat

$$DW = \frac{\sum(\epsilon_t - \epsilon_{t-1})^2}{\sum \epsilon^2} = 2 - 2\rho$$



Autocorrelation in data. 2 runs, but the real OLS, which we would have never found, is somewhere in the middle.

- Any value under D(L) (in the D-W table) rejects the null hypothesis and AC exists.
- Any value between D(L) and D(W) leaves us with no conclusion of AC.
- Any value larger than D(W) accepts the null hypothesis and AC does not exist.

- Note, this is one tail test, To get the other tail. Use 4 - DW as the test stat instead.

A Moving-Average (MA) Process

A moving average (MA) process is essentially created by lagging the residual term of the time series. For example a time series x with an autoregressive (AR) process of length 1 and a moving average process of length 2 can be expressed as x(t) = x(t-1) + e(t) + e(t-1) + e(t-2). In general, a time series can be represented by an ARMA(p,q) process, where p denotes the length of the autoregressive process, while q denotes the length of the moving average process.

SLR

Use of β_1 or β_0 to denote the Y-intercept is solely discretionary.

$$\widehat{\beta}_2 = \frac{\sum (Y_i - \bar{Y})(X_i - \bar{X})}{\sum (X_i - \bar{X})^2}$$
$$\widehat{\beta}_1 = \bar{Y} - \widehat{\beta}_2 \bar{X}$$
$$Var(\widehat{\beta}_2) = \frac{\sigma^2}{\sum (X_i - \bar{X})^2}$$
$$Var(\widehat{\beta}_1) = \frac{\sum X_i^2 * \sigma^2}{n * \sum (X_i - \bar{X})^2}$$
$$\widehat{\sigma^2} = \frac{\sum \widehat{U_i^2}}{n-2}$$

U and ε have both been used to denote the error term.

$$S_2^2 = \widehat{Var(\widehat{\beta}_2)} = \frac{\widehat{\sigma^2}}{\sum (X_i - \bar{X})^2}$$
$$S_1^2 = \widehat{Var(\widehat{\beta}_1)} = \frac{\sum X_i^2 * \widehat{\sigma^2}}{n * \sum (X_i - \bar{X})^2}$$
$$S.E.(\widehat{\beta}_2) = \sqrt{\widehat{Var(\widehat{\beta}_2)}}$$
$$S.E.(\widehat{\beta}_1) = \sqrt{\widehat{Var(\widehat{\beta}_1)}}$$

S^2 is used to denote a sample variance, S.E. standard error.

$$TSS = \sum (Y_i - \bar{Y})^2$$
$$ESS = \sum (\widehat{Y}_i - \bar{Y})^2$$
$$RSS = \sum \widehat{U_i^2}$$

TSS may also be presented as SST for the Total Sum of Squares, ESS as SSE (error) and RSS as SSR (residuals). Depending on the text, ESS and RSS may become very confusing, as there is great variety in the terminology used.

$$R^2 = \frac{ESS}{TSS} = 1 - \frac{RSS}{TSS}$$

$$\widehat{\log(Y)} = \widehat{\beta_1} + \widehat{\beta_2} \log(X)$$