Notes on Econometrics

Diego Vilán *

Fall 2011

These are some basic notes on Statistics and Econometrics. The main emphasis will be in the estimation of cross-sectional models. These notes will rely heavily on the references highlighted at the end of the document.

1 Causal Effects and Experiments

It all begins when trying to measure the causal relationship between variables. For example, how could one measure the causal effect on tomato yield (measured in grams) of applying a certain amount of fertilizer per square meter?

One way to address this question is to conduct a controlled experiment. In such an experiment, a researcher plants many plots of tomatoes. Each plot is identical with one exception: some plots get fertilizer while others do not. Moreover, whether a plot gets fertilized or not is determined randomly by a computer algorithm. This ensures that any other differences between the plots are unrelated to whether they receive fertilizer.

The difference between the average yield per square meter of the treated and untreated plots is the effect on tomato production of the fertilizer treatment.

This is an example of a **randomized controlled experiment**. It is **controlled** in the sense that there are both a control group that receives no treatment (no fertilizer) and a treatment group that receives the treatment (fertilizer). Is it **randomized** in the sense that the treatment is assigned randomly. This random assignment eliminates the possibility of a systematic relationship between, for example, how much did each plot receive and whether it received fertilizer, so that the only systematic difference between the treatment and control groups is the treatment itself.

Experiments in econometrics are rare because they are often unethical or prohibitively expensive.The concept of an ideal randomized controlled experiment is useful nonetheless,

[*] **DISCLAIMER**: I wrote these notes as a study aid for myself and for whoever might find them useful. They are work in progress and as such could be incomplete, inaccurate and even somewhat incorrect. Keep that in mind. Comments and suggestions welcomed!

because it provides a theoretical benchmark for an econometric analysis of causal effects using actual data.

1.1 Data Sources:

In econometrics data will generally come from two sources:

1. Experiments
2. Non-experiments

Experimental data comes from experiments designed to evaluate a treatment or policy. Observational data is data obtained by observing actual behavior outside an experiment setting. Observational data are collected using surveys (e.g.: CPS, PSID[1]), administrative records (e.g.: mortgage applications) and transaction histories (e.g.: buy/sell orders, supermarket data).

Observational data pose major challenges to econometric attempts to estimate causal effects because in the an uncontrolled setting (i.e.: the real world) levels of treatments are not assigned at random. Most of the tools in econometrics are aimed at dealing with these challenges.

1.2 Data types:

Whether the data are experimental or observational, there are three main types:

1. Cross-sectional
2. Times Series
3. Panel data

Data on different entities (workers, households, firms, etc.) for a single period of time are called **cross-sectional** data. For example, the data on SAT test scores in the CA school districts for a given period are cross sectional. Those data are for 420 entities (school districts) for a single year (1998). With cross-sectional data we can learn about relationships among variables by studying differences across entities during a single time period. The term pooled cross section, implies a data set constructed by merging cross section data from multiple periods.

Times series data are data for a single entity collected at multiple time periods. For example, the rates of inflation and unemployment in the US is an example of a time series data. By tracking a single entity over time. By studying a single entity over time (in this example,

[1] CPS = Current Population Survey, PSID = Panel Survey of Income Dynamics

the US) we can understand the evolution of variables over time.

Panel data (sometimes called longitudinal data) are data for multiple entities in which each entity is observer at two or more periods.

Note:

1. A panel data keeps track of the same set of entities over time. Pooled cross sections do not.
2. A panel data set is useful for studying the **dynamic behavior** of entities (e.g.: how last year's income of an individual affect this year's expenditure).
3. Pooled cross sections capture how the population **distribution changes** over time (e.g.: how the income distribution changes from a year to another).

2 Probability and Random Variables

2.1 Basic Definitions:

- A process/experiment whose outcome cannot be predicted is called a **random process/experiment**.
  e.g.: Roll a dice once.
- The mutually exclusive potential results of a random process are called **outcomes**.
  eg: $\{1\}, \{2\}, \{3\}, \{4\}, \{5\}\{6\}$
- The set of all possible outcomes is called the **sample space** (Ω).
  e.g.: $\Omega = \{1, 2, 3, 4, 5, 6\}$
- An **event** (A), is a subset of the sample space. That is, an event is a set of one or more outcomes.
  e.g.: Rolling dice and getting an even number. $A = \{2, 4, 6\}$
- A **random variable** is a numerical summary of a random outcome.
  e.g.: X =

  Some random variables are discrete and some continuous. A discrete random variable takes on only a discrete set of values such as -1, 0, 1, 2 ... A continuous random variable may take on a continuous of possible values. Because of these characteristics, random variables are often defined as a function that goes from the random process/experiment's sample space into $\Re$.
- A probability measure P, is a function that assigns a real number to each of the events of a random process/experiment satisfying the following conditions (axioms of probability):

1. $P(A) \geq 0$ for each event A
2. $P(\Omega)$=1
3. For any mutually exclusive events: $A_1, A_2, A_3, ...,$

$$P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i)$$

2.2 Probability distributions of Random Variables

When focusing on the behavior of a random variables X, we might want to treat the values of the random variable as if they were the outcome of the random process/experiment. This creates a new "sample space" and an associated probability measure which will be called the **probability distribution of X**.

eg: page 7, Lect 2
-Def PDF
-Def CDF
- Families of distributions
- Distribution of functions of RV
- Delta Method

2.3 Moments of Random Variables:

-Def Moment
Univariate:
- Def Mean
-Def Variance
Multivariate:
- Joint distribution
- Marginal probability
- Conditional expectations
- Law of iterated expectations
- Conditional Variance
- Conditional distribution
- Independence
- Covariance and Correlation

2.4 Random Sampling:

- Def Random sampling
- Convergence in probability
- Law of large numbers (LLN)
- Central Limit theorem

3 Descriptive Statistics

Theoretical v.s. sample moments.

3.1 Univariate Statistics:

Assume we have n observations on a single random variable X.

Measures of Central tendency:

A measure of central tendency if a statistic that measures the "typical" value. What is typical may have various interpretations.

(a) Mean: $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$

(b) Median: The value in the middle after observations have been ordered from smaller to largest.

(c) Mode: Most frequently occurring value.

Measures of Dispersion:

(a) **Range:** The difference between the largest and the smallest value of the variable amongst the n observations.

Several measures of dispersion are based on deviations/differences of the values of a variable from its mean. For the *ith* observation, the deviation of X from its mean is $d_i = x_i - \bar{x}$.

With n observations there will be n deviations. How do we obtain a measure of the "typical" deviation.

(b) **Mean squared deviation**:

$$\begin{aligned} MSD_x &= \sum_{i=1}^{n} \frac{d_i^2}{n} \\ &= \sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{n} \end{aligned}$$

(c) **Root Mean squared deviation:**

$$\begin{aligned} RMSD_x &= \sqrt{\sum_{i=1}^{n} \frac{d_i^2}{n}} \\ &= \sqrt{\sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{n}} \end{aligned}$$

(d) **Standard Deviation:** A minor alteration of the RMSD yields the standard deviation. The $n-1$ term is called the number of degrees of freedom [2].

$$\begin{aligned} S_x &= \sqrt{\sum_{i=1}^{n} \frac{d_i^2}{n-1}} \\ &= \sqrt{\sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{n-1}} \end{aligned}$$

(e) **Variance:** A minor alteration of the MSD.

$$\begin{aligned} S_x^2 &= \sum_{i=1}^{n} \frac{d_i^2}{n-1} \\ &= \sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{n-1} \end{aligned}$$

[2] The idea is that if we know $n-1$ values of a variable and $\bar{x}$ then the *nth* value can be determined. In this sense the *nth* observation is not free.

Statistic	Theoretical formula	Empirical Definition
Mean squared error	$\sum_{i=1}^{n} \frac{d_i^2}{n} = \sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{x}$	0
Root Mean squared error	n observations	0
Standard Deviation	n observations	0
Variance	$\sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{n-1}$	E[

Table 1: Summary

3.2 Multivariate Statistics:

The multiple regression model:

Classical Assumptions:

Interpretation of the regression coefficients:

PONER TABLA

Hacer de esto un ejemplo

Assume one would want to estimate the following relationship between wage, education and years of experience:

$$wage = \beta_0 + \beta_1 educ + +\beta_2 year + u$$

came down to

$$\hat{w} = -0.90 + 0.54 educ + 0.2 year$$

In this case each coefficient will measure its individual effect on the dependent variable. Naturally, this involves a *ceteris paribus* type of analysis, where we are studying the effects of one of the independent variables on the dependent one, while keeping the rest constant. In the example above we have that each additional year of education is predicted to increase the hourly wage by 0.54 dollars or 54 cents.

However, in the above example, the predicted increase will always be the constant, independent on the number of years of education or experience something which we might find not to be entirely accurate. Instead we could assume that the percentage wage increase will be constant rather than its level. In order to do this we need to modify the model's specification to:

$$\ln(\hat{w}) = \beta_0 + \beta_1 educ + +\beta_2 year + u$$

then the $\%\Delta wage \simeq (100 * \beta_1)\Delta educ$. Notice that you need to multiply $100 * \beta_1$ to get the % change in the wage given one additional year of education. Since the % change in the wage is constant, the change in levels as a function of education is increasing. For example, consider:

$$\ln(\hat{w}) = 0.584 + 0.083 educ + 1.3 year$$

in this case, the subject's wage increases 8.3 percent for every additional year of education. The intercept gives the $\ln(\hat{w})$ when $educ = year = 0$ and is not particularly meaningful.

Other forms of including non-linearities in models are summarized below:

The first model is the level-level model because each variable is in its level form. The third model is called the semi-log form model, where $100 * \beta$ is regarded as the semi-elasticity of

y w.r.t x. The fourth model is the log-form model where the coefficients are regarded as the elasticities of y w.r.t x.

Note that while the mechanics of the simple regression model do not depend on how y and x are defined, the interpretation of the coefficients do.

Geometric interpretation of a regression

Algebra of OLS

Finite sample properties of OLS

Large sample properties of OLS

Other models (GMM, GLS, OGLS)

3.3 IV Estimator

A violation to the orthgonality condition could happen because:

1. Omitted variables
2. Measurement errors
3. Simultaneous causality

When a variable that affects the dependent variable is omitted from the regressor list, and it is correlated with at least one among the regressors of the model, it is likely that the orthogonality condition is violated, so that the OLS estimator becomes inconsistent. Such problem is often referred to as *omitted variable bias*.

In turn, This makes the error term correlated with the regressors. Which makes OLS inconsistent (large samples) or biased (small sample).

References:

- J. H. Stock and M.W. Watson, 2006. Introduction to Econometrics.
- W. H. Greene, 2000. Econometric Analysis.
- F. Hayashi, 2000. Econometrics.
- T. Mirer, 1994. Economic Statistics and Econometrics.