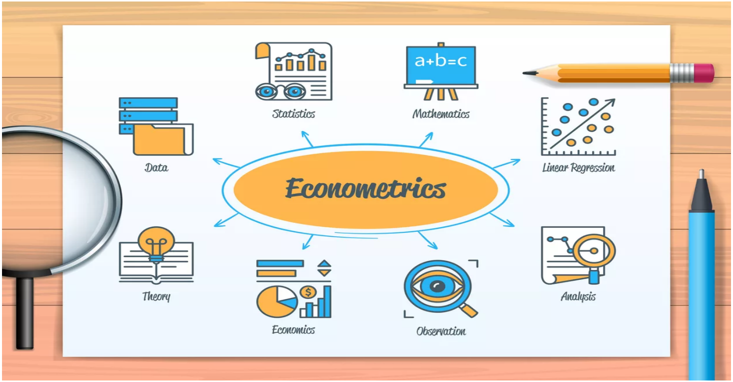
Fundamentals of Econometrics

Lecture 1: Nature of Econometrics & Economic Data



Section 1

Introduction

What is econometrics?

The study and application of statistical methods to the analysis of economic phenomena. Tintner (1953), Econometrica

Econo + metrics → Economic Measurement

- In practice, it is a set of tools from statistics that are combined with results from economic theory. (Leads to a later point on possible differences between the two. . .)
- Purpose: Econometrics gives **empirical content** to a priori reasoning in economics.

Section 2

Econometric Analysis

Steps in Econometric Analysis

1. Economic Models: Formulate a model of the economic relationship.
   - May be micro- or macroeconomic in nature.
   - Often use optimizing behavior of economic agents, equilibrium modeling, . . .
   - Establish relationship between economic variables
     - Examples: demand equations, pricing equations, production functions, etc.

2. Econometric Model: Specify the variables, functional form, and assumptions, etc.

Kinds of Application

- Economics suggests important relationships, often with policy implications, but virtually never suggests quantitative magnitudes of causal effects.
  - What is the *quantitative* effect of reducing class size on student achievement?
  - How does another year of education change earnings?
  - What is the price elasticity of cigarettes?
  - What is the effect on output growth of a 1 percentage point increase in interest rates by the Fed?
  - What is the effect on housing prices of environmental improvements?

Major Areas of Concern

- Estimation of economic relationships between economic variables.
- Testing of economic theories and hypotheses.
- Forecasting future values of economic variables.
- Policy analysis and evaluation.

Causal Effect

To estimate the causal effect of one variable on another.

$$\Downarrow$$

The effect of one variable on another, holding all other relevant factors constant (ceteris paribus).

Section 3

Examples

Size of the Police Force and Crime Rates

Question: What is the effect of the size of the police force on crime?

$$\#\text{Crimes} = \alpha + \beta \times \text{Size of Police Force} + u$$

Usually, cities with a lot of criminal activity have a bigger police force. Simple correlations can **spuriously** indicate that the size of the police force has a positive effect on the crime rates.

Always remember that correlation does not imply causation!

Firm Minimization Problem

Consider the Cobb-Douglas production function:

$$q = f(x, y) = Ax^{\alpha}y^{\beta}$$

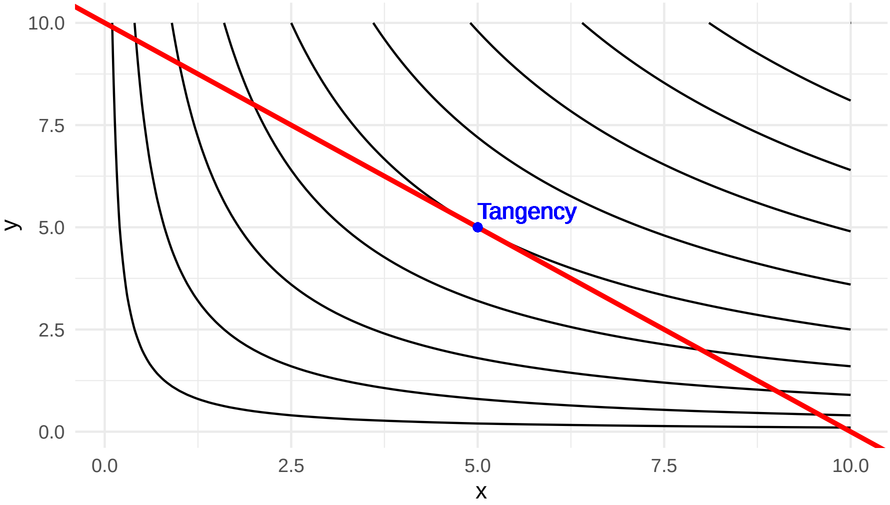
where q is output, x and y are inputs. A (a constant), α, and β are parameters.

Question: What is the cost minimizing factor demands for x and y?

- We must first assume prices for x and y, (p_x and p_y).
- The firm then has to minimize cost subject to some desired production level.

Firm's Problem

$Ax^{\alpha}y^{\beta} = 2x^{0.5}y^{0.5}$



Cobb-Douglas Production Function

The firm's problem is to minimize cost:

$$\text{(Cost Function):}\ \min_{x,y} p_x x + p_y y$$

$$\text{s.t}$$

$$\text{(Isoquant):}\ Ax^{\alpha}y^{1-\alpha} = \bar{q},$$

Lagrangian:

$$\mathcal{L}(x, y, \lambda) = p_x x + p_y y + \lambda(\bar{q} - Ax^{\alpha}y^{1-\alpha})$$

Cobb-Douglas Production Function

With some algebra, we can estimate the conditional factor demands for x^c and y^c via regression analyses using the first order conditions (FOCs) from the Lagrangian.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 \tag{1}$$

where $y = \ln(x^c)$, $x_1 = \ln(\bar{q})$, $x_2 = \ln(p_y)$, $x_3 = \ln(p_x)$, and $\beta_0 = \tilde{\gamma}$.

Theory tells us that $\beta_1 = 1$ and $\beta_2 = -\beta_3$ since $\beta_2 = 1 - \alpha$ and $\beta_3 = \alpha - 1$.

With data, we can test these propositions/hypotheses!!

Deterministic vs. Stochastic Models

- Our model in (1) is deterministic.
- This is fine for economic theory, but wrong if we want to:
  - use statistical methods to estimate the parameters $\beta_0, \beta_1, \beta_2,$ and β_3.
  - do inference– test hypotheses regarding the values of and/or relationships among the parameters.
- We need to introduce a stochastic error term to account for the fact that the conditional factor demand for x is not deterministic.

Nature of Econometrics

- We must make the model one that is stochastic!
    - Append a random error term, u, to the linear regression model
    - Note that, assuming all of the independent or right-hand-side (RHS) variables are deterministic, that is, fixed in repeated draws (which they rarely will be), then the model obtains all of its random properties from the error term.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \underbrace{\boxed{u}}_{???}$$

Economic model of crime

Becker (1968) posited that the decision to commit a crime is based on a cost-benefit analysis.

$$y = f(x_1, x_2, x_3, x_4, x_5, x_6, x_7)$$

where

- y is the number of hours spent in criminal activity,
- x_1 is the wage from criminal activities,
- x_2 is the wage from legal activities,
- x_3 is other income,
- x_4 is the probability of getting caught,
- x_5 is the probability of conviction if caught,
- x_6 is the expected sentence, and
- x_7 is the age of the individual.
- Unspecified functional form of relationship... linear?

Here equation is postulated without economic modeling.

Nature of Econometrics

Econometric model of crime activity

- The functional form has to be specified.
- Variables may have to be approximated by other quantities.

$$crime = \beta_0 + \beta_1 wagelegal + \beta_2 othinc + \beta_3 freqarr + \beta_4 freqconv + \beta_5 avgsen + \beta_6 age + u$$

where

- wagelegal is the wage from legal employment,
- othinc is other income, freqarr is the frequency of prior arrests,
- freqconv is the frequency of convictions,
- avgsen is the average sentence length after conviction, and
- age is the age of the individual.
- u is the error term and represents the unobserved determinants of criminal activity. E.g. wage from criminal activities, family background, etc.

Nature of Econometrics

- Model of job training and worker productivity
  - What is the effect of additional job training on worker productivity?
  - Formal theory not really needed to derive the equation:

$$wage = f(educ, exper, training)$$

where

- wage is the hourly wage,
- educ is the years of formal education,
- exper is the years of workforce experience, and
- training is the number of weeks spent in job training.
- Other factors may be relevant, but these may be most important (??)

Nature of Econometrics

Econometric model of job training and worker productivity

$$wage = \beta_0 + \beta_1 educ + \beta_2 exper + \beta_3 training + u$$

where

- *wage* is the hourly wage,
- *educ* is the years of formal education,
- *exper* is the years of workforce experience, and
- *training* is the number of weeks spent in job training.
- u is the error term and represents the unobserved determinants of worker productivity/wages. E.g. innate ability, family background, quality of education, etc.

- Most of econometrics deals with the specification of the error, u.
- Econometric models may be used for hypothesis testing
    - For example, the parameter β_3 represents effect of training on wage.
    - How large is this effect? Is it **statistically** different from zero?

Nature of Econometrics

Are futures markets efficient predictors of (future) cash market/spot prices for that item/commodity?

$$\text{Spot Price}_t = \beta_0 + \beta_1 \text{Futures Price}_{t-1} + u_t$$

where

- Spot Price_t is the price of the commodity at time t (say today),
- Futures Price_{t-1} is the price of the futures contract for the commodity at time $t-1$ (yesterday), and
- u_t is the error term which captures the unobserved determinants of the spot price. E.g. unexpected changes in demand, weather shocks, etc.

Economic theory suggests that futures prices should be unbiased predictors of future spot prices. Therefore, if the market is truly “strong-form efficient,” then $\beta_1 = 1$ and $\beta_0 = 0$.

What is this course about?

- Ideally, we would like an experiment
  - What would an experiment to estimate the effect of class size on standardized test scores?
- However, almost always, we have only **observational data**.
  - Returns on education
  - Household income
  - Unemployment rates
- Most of the course addresses difficulties arising from using obesrvational data to estimate causal effects.
  - Confounding Effects (ommitted variables)
  - Measurement Error
  - Simultaneity (reverse causality)

Economic Data

- Econometric analysis requires data.
- Different kinds of economic data are available:
  1. Cross-sectional data
  2. Time Series data
  3. Pooled cross-sections
  4. Panel data
- Econometric methods depend on the nature of the data used.
  - Use of inappropriate methods can lead to incorrect results and conclusions.

Nature of Economic Data

Cross-sectional data sets

- Sample of individuals, households, firms, cities, states, countries, or other units of interest **at a given point of time/in a given period**.
  - For example, data on individuals in the 2010 Census, data on firms in the 2010 Annual Survey of Manufactures, etc.
- Cross-sectional observations are typically independent
  - For example, my income spent on food is likely independent of your income spent on food
- Pure random sampling from a population often assumed. Also, ordering doesn't matter.
- Sometimes pure random sampling is violated, e.g. units refuse to respond in surveys
- Cross-sectional data typically encountered in applied microeconomics.

Nature of Economic Data

Sample Cross-sectional data set on wages and other characteristics

Individual	**Wage**	**educ**	**exper**	**female**	**married**
1	5.50	12	5	0	1
2	6.00	16	7	1	0
3	5.75	14	9	1	1
4	6.25	18	12	0	0
5	5.50	12	5	0	1
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.
525	6.75	16	10	1	1
526	7.00	18	12	0	0

Here $female$ and $married$ are dummy/indicator variables where (1 = yes, 0 = no).

Nature of Economic Data

Time Series Data

- Observations of one or more variables **over time**
  - For example, stock prices, money supply, consumer price index, gross domestic product, annual homicide rates, unemployment rates, ...
- Time series observations are typically **serially correlated**
  - Producer price index this month is highly correlated with that from last month
- Ordering of observations conveys important information. Time series data are therefore presented in chronological order.
- Data frequency: daily, weekly, monthly, quarterly, annually, ...
- Typical features of time series: **trends** and **seasonality**
- Typical applications: applied macroeconomics and finance

Nature of Economic Data

Sample Time Series data set on GDP, wages, inflation, and unemployment

obsno	**Year**	**GDP**	**Wage**	**Inflation**	**Unemployment**
1	1990	1000	5.50	0.02	0.05
2	1991	1050	6.00	0.03	0.04
3	1992	1100	5.75	0.02	0.03
4	1993	1150	6.25	0.04	0.03
5	1994	1200	5.50	0.03	0.04
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.
33	2022	2000	7.00	0.02	0.05
34	2023	2050	7.25	0.03	0.04

Nature of Economic Data

Pooled Cross-sections

- Combination of (independently drawn) cross-sectional data sets over time
  - For example, data on individuals in the 2010 Census, 2011 Census, 2012 Census, etc.
- Pooled cross-sections are typically used to study changes in economic relationships over time or evaluate the impact of policy changes over time.
  - For example, the impact of raising the federal minimum wage on employment.
  - Impact of raising the raising the legal drinking age in a state.
  - Compare effects before and after the introduction of a new technology.

Nature of Economic Data

Sample Pooled Cross-section data set on wages and other characteristics

Individual	**Year**	**Wage**	**educ**	**exper**	**female**	**married**
1	2010	5.50	12	5	0	1
2	2010	6.00	16	7	1	0
3	2010	5.75	14	9	1	1
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.
300	2011	6.75	16	10	1	1
301	2011	7.00	18	12	0	0
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.
670	2011	7.50	20	15	1	1

Nature of Economic Data

Panel or longitudinal data

- Combination of cross-sectional and time series data. The **same** cross-sectional units are followed **over time**.
- Panel data have a **cross-sectional and a time series dimension**
- Panel data can be used to account for time-invariant unobservables
    - What econometricians call unobserved heterogeneity
- Panel data can be used to model lagged responses
- Example:
    - City crime statistics; each city is observed in two years
    - Time-invariant unobserved city characteristics may be modeled
    - Effect of police on crime rates may exhibit time lag

Nature of Economic Data

Sample Panel data set on city crime rates

obsno	City	Year	murders	population	unemp	police
1	1	1985	10	100000	5	150
2	1	1990	12	105000	6	155
3	2	1985	15	200000	7	200
4	2	1990	20	210000	8	210
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.
497	149	1985	20	480000	8	280
498	149	1990	25	490000	9	290
499	150	1985	30	500000	10	300
500	150	1990	35	510000	11	310

Nature of Econometrics & Economic Data

Causality and the notion of ceteris paribus

- Econometric models are used to estimate causal relationships between economic variables.
    - For example, the effect of education on wages, the effect of inflation on unemployment, etc.

Formally, we define a causal effect of variable X on variable Y as the change in Y that results from a change in X, **while holding all other relevant factors constant**.

- Most economic questions are *ceteris paribus* questions
- It is important to define which causal effect one is interested in.

Nature of Econometrics & Economic Data

Causality effect of fertilizer on crop yield

- By how much will the production of maize increase if one increases the amount of fertilizer applied
- **Implicit assumption:** all other factors that influence crop yield such as quality of land, rainfall, presence of pests etc. are held fixed

Measuring the return to education

- If a person is chosen from the population and given another year of education, by how much will his/her wage increase?
- **Implicit assumption:** all other factors that influence wages such as experience, family background, intelligence etc. are held fixed.

Section 4

Appendix

Cobb-Douglas Production Function (In detail)

Our first order conditions (FOCs) are:

$$\frac{\partial \mathcal{L}}{\partial x} = p_x - \lambda \alpha A x^{(\alpha-1)} y^{1-\alpha} = 0 \tag{2}$$

$$\frac{\partial \mathcal{L}}{\partial y} = p_y - \lambda (1-\alpha) A x^{\alpha} y^{(-\alpha)} = 0 \tag{3}$$

$$\frac{\partial \mathcal{L}}{\partial \lambda} = \bar{q} - A x^{\alpha} y^{1-\alpha} = 0 \tag{4}$$

Cobb-Douglas Production Function

Combine (1) and (2) we obtain:

$$\frac{p_x}{p_y} = \frac{\lambda \alpha A x^{(\alpha-1)} y^{1-\alpha}}{\lambda(1-\alpha) A x^{\alpha} y^{(-\alpha)}} = \underbrace{\boxed{\frac{\alpha}{1-\alpha}\frac{y}{x}}}_{\text{Tangency condition in graph!}}$$

or

$$y = \frac{1-\alpha}{\alpha}\frac{p_x}{p_y}x \tag{5}$$

Substitute (4) into (3) we obtain:

$$\bar{q} = A x^{\alpha} \underbrace{\left(\frac{1-\alpha}{\alpha}\frac{p_x}{p_y}x\right)^{1-\alpha}}_{y}$$

Cobb-Douglas Production Function

or solving for x:

$$\underbrace{\boxed{x = \frac{\bar{q}}{A}\left(\frac{\alpha}{1-\alpha}\frac{p_y}{p_x}\right)^{1-\alpha}}}_{\text{Conditional factor demand for x}} = x^c(p_x, p_y, \bar{q})$$

And substituting $x^c(p_x, p_y, \bar{q})$ back into (4) we obtain:

$$\underbrace{\boxed{y = \frac{\bar{q}}{A}\left(\frac{1-\alpha}{\alpha}\frac{p_x}{p_y}\right)^{\alpha}}}_{\text{Conditional factor demand for y}} = y^c(p_x, p_y, \bar{q})$$

- This appears to be an algebraic mess, but it possesses important economic and empirical implications.

Cobb-Douglas Production Function

- Now define

$$\gamma_1 = \frac{1}{A}\left(\frac{\alpha}{1-\alpha}\right)^{1-\alpha} \quad \text{and} \quad \gamma_2 = \frac{1}{A}\left(\frac{1-\alpha}{\alpha}\right)^{\alpha}$$

so that we can write the conditional factor demands as:

$$x^c(p_x, p_y, \bar{q}) = \gamma_1 \bar{q}\left(\frac{p_y}{p_x}\right)^{1-\alpha} \quad \text{and} \quad y^c(p_x, p_y, \bar{q}) = \gamma_2 \left(\frac{p_x}{p_y}\right)^{\alpha} \bar{q}$$

Cobb-Douglas Production Function

Also, recall the properties of natural logarithms:

$$\begin{aligned}
\ln(x/y) &= \ln(x) - \ln(y) \\
\ln(xy) &= \ln(x) + \ln(y) \\
\ln(x^\alpha) &= \alpha \ln(x) \\
\ln(x^\alpha y^\beta) &= \alpha \ln(x) + \beta \ln(y) \\
etc.
\end{aligned}$$

Cobb-Douglas Production Function

- We can now take the natural logarithm of the conditional factor demands to obtain:

$$\ln(x^c) = \underbrace{\tilde{\gamma}}_{\ln(\gamma_1)} + \ln(\bar{q}) + (1-\alpha)\ln(p_y) + (\alpha - 1)\ln(p_x)$$

$$\ln(y^c) = \underbrace{\tilde{\gamma}}_{\ln(\gamma_2)} + \ln(\bar{q}) - \alpha\ln(p_y) + \alpha\ln(p_x)$$

Focusing on the conditional factor demand for x, we can write in the form of a linear regression model:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 \tag{6}$$

where $y = \ln(x^c)$, $x_1 = \ln(\bar{q})$, $x_2 = \ln(p_y)$, $x_3 = \ln(p_x)$, and $\beta_0 = \tilde{\gamma}$.

Theory tells us that $\beta_1 = 1$ and $\beta_2 = -\beta_3$ since $\beta_2 = 1 - \alpha$ and $\beta_3 = \alpha - 1$.

With data, we can test these propositions/hypotheses!!