Econometrics…
… in a few slides

Topics in Economics

Stefano Usai

1

Overview

- Part 1: Basics of Econometrics
  - Econometrics is…
  - Methodology of Econometrics
  - Elements of Econometrics
- Part 2: Basics of Regression Analysis
  - Regression analysis: some basic ideas, terminology
  - 10 assumptions underlying the linear regression model
  - Goodness of fit
  - Regression modeling selection process
  - What if…
- Part 3: Regression Analysis in praxis
  - Planning, development and maintenance of the model
  - An example: Determinants of atypical work

2

Part 1. Econometrics is…

- … an amalgam of economic theory, mathematical economics, economic statistics, and mathematical statistics

  Gujarati
- … the quantitative analysis of actual economic phenomena based on the concurrent development of theory and observation, related by appropriate methods of inference

  Samuelson, Koopmans & Stone
- … using sample data on observable variables to learn about the functional relationships among economic variables

  Abbott

3

Econometrics consists…

… mainly of:

- ***estimating*** relationships from sample data
- ***testing hypotheses*** about how variables are related
  - the ***existence* of relationships** between variables
  - the ***direction* of the relationships** between the dependent variable and its hypothesized observable determinants
  - the ***magnitude* of the relationships** between a dependent variable and the independent variables thought to determine it

4

Methodology of Econometrics

1. Statement of theory or hypothesis
2. Specification of the mathematical model
3. & of the econometric model of the theory
4. Obtaining the data
5. Estimation of the parameters of the econometric model
6. Hypothesis testing
7. Forecasting or prediction
8. Using the model for control or policy purposes

5

Elements of Econometrics

- **Specification of the *econometric model*** that we think (hope) generated the sample data. It consists of:
  - An ***economic model***: specifies the ***dependent* variable** to be explained and the ***independent* variables** thought to be related to the dependent
    - Suggested or derived from theory
    - Sometimes obtained from informal intuition/ observation
  - A ***statistical model***: specifies the statistical elements of the relationship under investigation, in particular the ***statistical properties* of the *random* variables** in the relationship

6

Elements of Econometrics

- **Collecting and coding the sample data**
  - Most economic data is ***observational*, or *non-experimental***
  - **Sample data** consist of ***observations* on *randomly selected* members of populations** (individual persons, households or families, firms, industries, provinces or states, countries)
- **Estimation** consists of using **the *sample data*** on the ***observable* variables** to compute ***estimates*** of the **numerical values** of all the **unknown parameters** in the model
- **Inference** consists of **using the parameter estimates** computed from sample data **to test hypotheses** about the **numerical values** of the **unknown *population* parameters** that describe the behavior of the population from which the sample was selected

7

Suggestions for further reading

- Damodar N. Gujarati, *Basic Econometrics*, 3d eds, Mc Graw Hill, 1995
- Adrian C. Darnell & J. Lynne Evans, *The limits of Econometrics*, Edward Elgar Publishing Ltd., Hants, England, 1990

8

Part 2. Regression analysis

- Regression models:
  - one variable called the dependent variable is expressed as a linear function of one or more other variables, called the explanatory variables
  - it is assumed implicitly that causal relationships, if any, between dependent and explanatory variables flow in one direction only, namely, from the explanatory variables to the dependent variables

9

Regression Analysis: Some basic ideas

Key idea: the statistical dependence of one variable on one or more other variables

- Objective: to estimate and/or predict the mean or average value of the dependent variable on the basis of the known or fixed values of the explanatory variables

$$Y_i = f(X_{2i}, \ldots, X_{ki}) + u_i = \beta_1 + \beta_2 X_{2i} + \ldots + \beta_k X_{ki} + u_i$$

Population Regression Function (PRF)

The PRF is an idealized concept

- Generally one uses the stochastic sample regression (SRF) function to estimate the PRF
- Ordinary least squares (OLS) is the most used method of constructing the SFR

Sample Data: A random sample of N members of the population for which the observed values of Y and X_{ki}s are measured

- Sample size is critical because it influences the confidence level of conclusions
  - The larger the sample, the higher the associated confidence
  - BUT larger samples require more effort & time

10

Regression Analysis: Terminology

Where

- **Y_i , X_i , u_i variables** of the regression model
  - Y_i ≡ ***dependent*** *or* ***explained*** **variable** or ***regressand***
  - X_i ≡ ***independent*** *or* ***explanatory*** **variable** or ***regressor***
    - **Y_i, X_i** are ***observable*** **variables**
  - **U_i ≡ random error term**
    - it is ***unobservable*** **variable**

      It is called residual for sample observation
- **β_1 and β_2 parameters** of the regression model
  - **β_1 ≡ intercept coefficient**,
  - **β_2 ≡ slope coefficient** of X
    - **they are called regression *coefficients***
    - the **true population values** of β1 and β2 are ***unknown***

      They are called estimators of the regression coefficients for the sample observation

11

10 assumptions underlying the linear regression model

- Linear regression model
  - It is always linear in the coefficients being estimated, not necessarily linear in the variables
  - A scatter-plot is essential to examining the relationship between the two variables
- X values are fixed in repeated sampling
  - More technically X is assumed to be non-stochastic
- Zero mean value of random error term u_i
- Homoscedasticity or equal variance of u_i
  - It means that Y populations corresponding to various X values have the same variance
- No autocorrelation between the random error terms
  - In other words u_i & u_j are uncorrelated

12

10 assumptions underlying the linear regression model

- Zero covariance between u_i and X_i
  - The error term and explanatory variable are uncorrelated
- The number of observations n must be greater than the number of parameters to be estimated
  - Alternatively the number of observations must be greater than the number of explanatory variables
- Variability in X values
  - The X values in a given sample must not be all the same
- The regression model is correctly specified
  - There is no bias or error in the model used in empirical analysis
- There is no perfect multi-collinearity
  - There are no perfect linear relationships among the explanatory variables, the variables should be independent

13

Goodness of fit

- The coefficient of determination r^2 (2 variable case) or R^2 (multiple variable case) is a summary measure that tells how well the regression line fits the data
  - It measure the proportion or percentage of the total variation in Y explained by the regression model
  - It is nonnegative quantity ($0 \leq R^2 \leq 1$)
- The coefficient of correlation r is a measure of the degree of association between two variables
  - Quantity closely related but conceptually very much different from r^2

14

Regression vs Correlation

Correlation analysis

- Primary objective: to measure the degree of linear association between two variables
- Symmetry in the way variables are treated
  - Both variables are assumed to be random
- It does not imply any cause and effect relationship
- The correlation between two random variables is often due only to the fact that both variables are correlated with the same third variable

Regression analysis

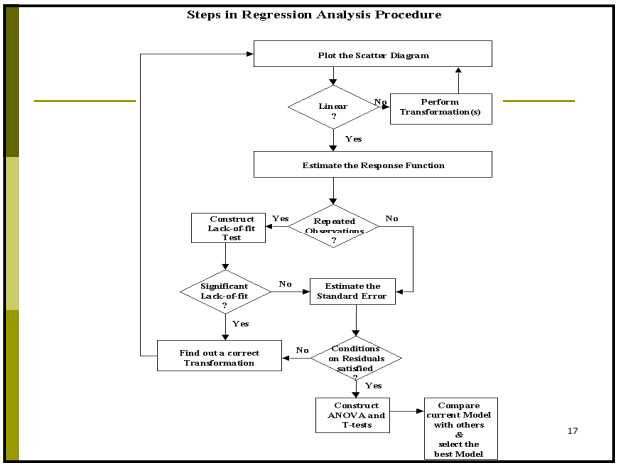
- Primary objective: to estimate or predict the average value of one variable on the basis of fixed values of other variables
- Asymmetry in the way dependent & explanatory variables are treated
  - Dependent variable is random (normal probability distribution)
  - Independent variables are assumed to have fixed values
- It implies causality

15

Regression modeling selection process

- When you have more than one regression equation based on data, to select the "best model", you should compare:
  - R^2 or adjusted R^2, i.e., the percentage of variance in Y accounted for variance in X captured by the model
  - Standard deviation of error terms, i.e., observed y-value/ predicted y-value for each X
  - The T-statistic of individual parameters
  - The values of the parameters and its content to content underpinnings

16



17

What if…we have qualitative explanatory variables?

- Introduce dummy variables D_i (taking values: 1,0)
  - dummies can be used in regression models just as easily as quantitative variables
  - If there is only dummy explanatory variables use ANOVA model (Analysis of Variance)
  - Possibly consider Interactions effects
    - There may be interaction between two dummies so that their effect on mean Y is not simply additive but also multiplicative
- The dummy variable technique must be handled carefully
  - The number of dummy variables must be less than the number of classifications of each qualitative variable
  - The coefficient attached to the dummy variable must be always interpreted in relation to the base, that is the group that gets the value of zero
  - Weight the number of dummies introduced against the number of observations

18

What if… we have dummy dependent variable?

- Use a dummy dependent variable regression model
  - Logistic regression model
  - Probit
  - Logit
  - Tobit

19

What if…

In regression analysis involving time series

…the regression includes not only current but also lagged values of the explanatory variables (Xs)?

- Distributed lag model

…the model includes one or more lagged values of the dependent variable (Yt-1) among its explanatory variables?

- Auto-regressive (dynamic) model

20

What if…

…a regression model has been estimated using the available data sample and an additional data sample become available?

- Use the analysis of covariance (ANCOVA) to test if previous model is still valid or the two separate models are equivalent or not

21

Part 3. Planning the model

Define the problem; select response; suggest variables.
- Are the proposed variables fundamental to the problem, are they variables?
- Are they measurable/countable?
- Can one get a complete set of observations at the same time?
- Is the problem potentially solvable?

- Collect data
- Check the quality of data
  - plot; try models
- Find the basic statistics, correlation matrix and first regression runs
- Establish goal: The final equation should have
  - **adjusted** $R^2 = 0.8$
  - coefficient of variation of less than 0.10
  - appropriate number of predictors
  - estimated coefficients must be significant at m = 0.05
  - no pattern in the residuals
- Are goals and budget acceptable?

22

Development of the model

- Check the regression conditions
  - Remove outliers they may have major impact
  - Examine all the points in the scatter diagram is Y a linear function of X? Consider transformation
  - The distribution of the residual must be normal
  - The residuals should have mean equal to zero, and constant standard deviation
  - The residuals constitute a set of random variables
  - Check for residuals autocorrelation
    - Durbin-Watson (D-W) (values [0,4])
    - No correlation ~2
- Consult experts for criticism
- Plot new variable and examine same fitted model
  - Also transformed predictor variable may be used
- Are goals met?
  - Have you found "the best" model?

23

Validation & maintenance of the model

- Are parameters stable over the sample space?
- Is there a lack of fit?
  - Are the coefficients reasonable?
  - Are any obvious variables missing?
  - Is the equation usable for control or for prediction?
- Maintenance of the Model
  - Need to have control chart to check the model periodically by statistical techniques

24

An example: the consumption function

1. Statement of Theory or Hypothesis

- Keynes states that on average, consumers increase their consumption as their income increases, but not as much as the increase in their income (MPC < 1).

2. Specification of the Mathematical Model of Consumption
(single-equation model)

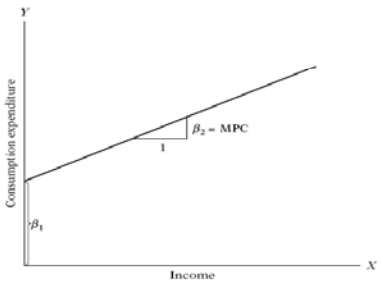
$$Y = \beta_1 + \beta_2 X \qquad 0 < \beta_2 < 1 \qquad (I.3.1)$$

Y = *consumption expenditure and* (dependent variable)
X = income, (independent, or explanatory variable)
β_1 = the intercept
β_2 = the slope coefficient

- The slope coefficient β_2 *measures the MPC.*

Consumption function

- Geometrically,



3. Specification of the Econometric Model of Consumption

- The relationships between economic variables are generally *inexact*. In addition to income, other variables affect consumption expenditure. For example, *size of family, ages of the members in the family, family religion*, etc., are likely to exert some influence on consumption.

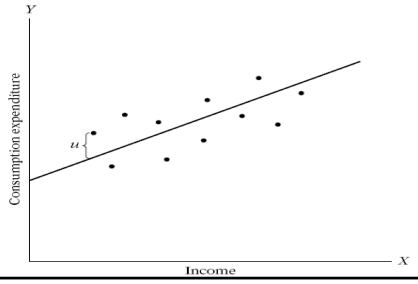
- To allow for the *inexact* relationships between economic variables, (I.3.1) is modified as follows:

- $Y = \beta_1 + \beta_2 X + u$ *(I.3.2)*

- where u, known as *the disturbance, or error, term,* is a random (stochastic) variable that has well-defined *probabilistic properties*. The disturbance term *u may well represent all those factors that affect consumption but are not taken into account* explicitly.

An example of a linear regression model

- (I.3.2) is an example of a linear regression model, i.e., it hypothesizes that *Y is linearly related to X, but that the relationship between the two is not exact*; it is subject to individual variation.

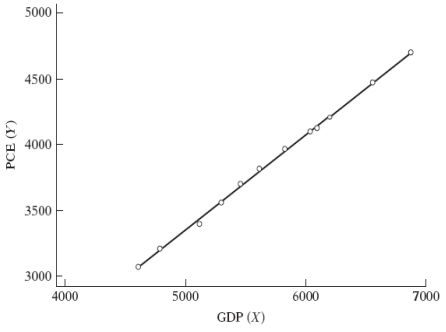


4. Obtaining Data

- To obtain the numerical values of *β1 and β2, we need data.* Look at Table I.1, which relate to the *personal consumption expenditure* (PCE) and the *gross domestic product (GDP)*. The data are in "real" terms.

Year	Y	X
1982	3081.5	4620.3
1983	3240.6	4803.7
1984	3407.6	5140.1
1985	3566.5	5323.5
1986	3708.7	5487.7
1987	3822.3	5649.5
1988	3972.7	5865.2
1989	4064.6	6062.0
1990	4132.2	6136.3
1991	4105.8	6079.4
1992	4219.8	6244.4
1993	4343.6	6389.6
1994	4486.0	6610.7
1995	4595.3	6742.1
1996	4714.1	6928.4

A scatter diagram of PCE on GDP



5. Estimation of the Econometric Model

Regression analysis is the main tool used to obtain the estimates. Using this technique and the data given in Table I.1, we obtain the following estimates of β_1 *and* β_2, *namely,* -184.08 *and* 0.7064. Thus, the estimated consumption function is:

- $\hat{Y} = -184.08 + 0.7064X_i$ *(I.3.3)*
- *The estimated* regression line is shown in Figure I.3. The regression line fits the data quite well. The slope coefficient (i.e., the MPC) was about 0.70, an increase in real income of 1 dollar led, *on average, to an increase of about 70 cents in real consumption*.

6. Hypothesis Testing

- That is to find out whether the estimates obtained in, Eq. (I.3.3) are in accord *with the expectations of the theory that is being tested.* Keynes expected the *MPC to be positive but less than 1*. In our example we found the MPC to be about 0.70. But before we accept this finding as confirmation of Keynesian consumption theory, we must enquire whether this estimate is sufficiently below unity. In other words, is *0.70 statistically less than 1? If it is, it may support* Keynes' theory.
- Such confirmation or refutation of economic theories on the basis of sample evidence is based on a branch of statistical theory known as statistical inference (hypothesis testing).

7. Forecasting or Prediction/a

- To illustrate, suppose we want to predict the mean consumption expenditure for 1997. The GDP value for 1997 was 7269.8 billion dollars consumption would be:

$\hat{Y}1997 = -184.0779 + 0.7064\,(7269.8) = 4951.3$ (I.3.4)

- The *actual value* of the consumption expenditure reported in 1997 was 4913.5 billion dollars. The estimated model (I.3.3) thus over-predicted the actual consumption expenditure by about 37.82 billion dollars. We could say the *forecast error* is about 37.8 billion dollars, which is about 0.76 percent of the actual GDP value for 1997.
- Now suppose the government decides to propose a reduction in the income tax. What will be the effect of such a policy on income and thereby on consumption expenditure and ultimately on employment?

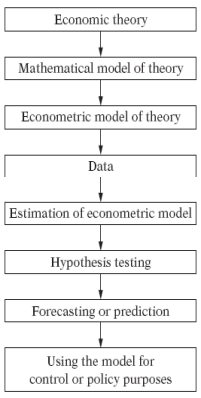
7. Forecasting or Prediction/b

- Suppose that, as a result of the proposed policy change, investment expenditure increases. What will be the effect on the economy? As macroeconomic theory shows, the change in income following, a dollar's worth of change in investment expenditure is given by the income multiplier *M,* which is defined as:
- $M = 1/(1 - \text{MPC})$ (I.3.5)
- The multiplier is about *M = 3.33. That is, an increase (decrease) of a dollar in investment will eventually* lead to more than a threefold increase (decrease) in income; note that it takes time for the multiplier to work.
- The critical value in this computation is *MPC*. Thus, a quantitative estimate *of MPC provides valuable information* for policy purposes. Knowing MPC, one can predict the future course of income, consumption expenditure, and employment following a change in the government's fiscal policies.

8. Use of the Model for Control or Policy Purposes

- Suppose we have the estimated consumption function given in (I.3.3). Suppose further the government believes that consumer expenditure of about 4900 will keep the *unemployment rate at its current level of about 4.2%*. What level of income will guarantee the target amount of consumption expenditure?
- If the regression results given in (I.3.3) seem reasonable, simple arithmetic will show that:
- $4900 = -184.0779 + 0.7064X$ *(I.3.6)*
- which gives *X = 7197, approximately. That is, an income level of about* 7197 (billion) dollars, given an MPC of about 0.70, will produce an expenditure of about 4900 billion dollars. As these calculations suggest, an estimated model may be used for control, or policy, purposes. By appropriate fiscal and monetary policy mix, the government can manipulate the *control variable X to produce the desired level of the target variable Y.*

Anatomy of classical econometric modeling.



9. Choosing among Competing Models

- When a governmental agency (e.g., the U.S. Department of Commerce) collects economic data, such as that shown in Table I.1, it does not *necessarily have any economic theory in mind*. How then does one know that the data really support the Keynesian theory of consumption? Is it because the Keynesian consumption function (i.e., the regression line) shown in Figure I.3 is extremely close to the actual data points? Is it possible that another consumption model (theory) might equally fit the data as well? For example, Milton Friedman has developed a model of consumption, called the *permanent income hypothesis*. *Robert Hall has also developed a model of* consumption, called the *life-cycle permanent income hypothesis*. *Could one* or both of these models also fit the data in Table I.1?
- In short, the question facing a researcher in practice is how to choose among competing hypotheses or models of a given phenomenon, such as the consumption–income relationship.

Some more things

- Multiple models
- Dummy variables
- Interactive variables
- Time dependence…and dynamic models
- Spatial dependence and spatially dynamic models..

conclusions

- The eight-step classical econometric methodology discussed above is neutral in the sense that it can be used to test any of these rival hypotheses. Is it possible to develop a methodology that is comprehensive enough to include competing hypotheses? This is an involved and controversial topic.