What is econometrics?

-> Econometrics is the quantitative application of statistical and mathematical models using data to develop theories or test existing hypotheses

in economics and to forecast future trends from historical data. It subjects real-world data to statistical trials and then compares and contrasts

the results against the theory or theories being tested.

-> Econometrics is concerned with the empirical determination of economic laws.

METHODOLOGY OF ECONOMETRICS

Broadly speaking, traditional econometric methodology proceeds along

the following steps:

1. Statement of theory or hypothesis.
2. Specification of the mathematical model of the theory
3. Specification of the statistical, or econometric, model
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

Difference between correlation and regression

Basis	Correlation	Regression

Meaning	Correlation is a statistical measure that determines the association or co-relationship between two variables.	Regression describes how to numerically relate an independent variable to the dependent variable.
Usage	To represent a linear relationship between two variables.	To fit the best line and to estimate one variable based on another.
Indicates	Correlation coefficient indicates the extent to which two variables move together.	Regression indicates the impact of a change of unit on the estimated variable (y) in the known variable (x).
Objective	To find a numerical value expressing the relationship between variables.	To estimate values of random variables on the basis of the values of fixed variables.
Variables	We treat any (two) variables symmetrically; there is no distinction between the dependent and explanatory variables.	In regression analysis there is an asymmetry in the way the dependent and explanatory variables are treated.
Math	No scope for further mathematical treatment	Can be used for further mathematical treatment
Origin and scale	Independent of origin and scale	Independent of origin but not of scale
Measure	Relative measure	Absolute measure
Symmetrical / Asymmetric	Symmetrical	Asymmetrical

Difference between All the Data Structures

	Cross Sectional Data	Time Series Data	Panel Data	Pooled Data

Definition	Cross-section data are data on one or more variables collected at the single point in time	A time series is a set of observations on the values that a variable takes at different times.	A panel or longitudinal data set consists of two or more sets observations on the same sample of cross-sectional units at two or more points in time.	A panel or longitudinal data set consists of two or more sets observations on the different sample of cross-sectional units at two or more points in time.
Ordering	With cross-sectional data the ordering of the data does not matter	Unlike cross-sectional data, the ordering of the data is important in time-series data.	The order within a cross section of a dataset panel does not matter, but the order in the time dimension is relevant.	Order does not matter within time and cross section.
Application	Researchers generally use cross-sectional data to make comparisons between subgroups	Time Series Analysis is used for many applications such as Economic Forecasting, Sales Forecasting, Budgetary Analysis, Stock Market Analysis	professionals often use it for statistical, financial and economic research.	One example is GNP per capita of all European countries over ten years

Multicollinearity

Assumption : No correlation b/w two independent variables.

Problem : Multicollinearity

Illness : Uncertainity : Inefficiency of coefficient

Measure : Variance inflation factor (VIF)

VIF = 1/(1-r^2) (independent calculation for each variable)

Income pocketmoney example.

Solution :

1. Leave model alone if VIF is small
2. Exclude variable if VIF is greater than threshold (cannot remove focus variable; can remove control variable)
3. Change the measure
4. Increase in sample size

Multicollinearity exists whenever an independent variable is highly correlated with one or more of the other independent variables in a multiple regression equation.

Multicollinearity is a problem because it undermines the statistical significance of an independent variable.

Multicollinearity can only be defined with linear relationships among X variables.

⇨ Detection of Multicollinearity (R2 P E A E T)

1) High **R2** but few significant t ratios: If R2 is high, say, in excess of 0.8, the F test in most cases will reject the hypothesis that the partial slope coefficients are simultaneously equal to zero, but the individual t tests will show that none or very few of the partial slope coefficients are statistically different from zero
2) High **pair-wise** correlations among regressors: High zero-order correlations are a sufficient but not a necessary condition for the existence of multicollinearity because it can exist even though the zero-order or simple correlations are comparatively low.

   In models involving more than two explanatory variables, the simple or zero-order correlation will not provide an infallible guide to the presence of multicollinearity. Of course, if there are only two explanatory variables, the zero-order correlations will suffice.
3) **E**xamination of partial correlations
4) **A**uxiliary regressions
5) **E**igenvalues and condition index
6) **T**olerance and variance inflation factor

⇨ Remedial Measures of Multicollinearity

Rule of thumb procedures

1) A **Priori Information**: $Yi = \beta1 + \beta2X2i + \beta3X3i + ui$ where Y = consumption, X2 = income, and X3 = wealth. As noted before, income and wealth variables tend to be highly collinear. But suppose a priori we believe that $\beta3 = 0.10\beta2$; that is, the rate of change of consumption with respect to wealth is one-tenth the corresponding rate with respect to income.

   $Yi = \beta1 + \beta2X2i + 0.10\beta2X3i + ui$

   $= \beta1 + \beta2Xi + ui$

   where $Xi = X2i + 0.1X3i$. Once we obtain $\beta2$, we can estimate $\beta3$ from the postulated relationship between $\beta2$ and $\beta3$.

   We can obtain a priori information with the help of empirical work.

2) **Combining** cross-sectional and time series data

   Combination of Cross-sectional and time series data results in a technique named pooling of data. Although it might be appealing technique but it might create problems of interpretation because we are assuming that estimated value obtained from cross sectional data is same as time series data.

3) **Dropping** a variable(s) and specification bias

   A straightforward method of correcting multicollinearity is removing one or more variables showing a high correlation. This assists in reducing the multicollinearity linking correlated features.

   Can drop control variables not focus variables.

4) **Transformation** the Variables

   It is also helpful for abetting the reduction in multicollinearity. One of the methods for transformation is Difference Form of the equation. Second method is ratio transformation. Such transformations will reduce collinearity in original variables.

5) **Additional** or new data

   Statistically, a regression model with more data is likely to suffer less variance due to a larger sample size. This will reduce the impact of multicollinearity.

6) **Reducing** collinearity **in polynomial** regressions

Also there are other methods which can reduce the multicollinearity such as factor analysis , Principal Components.

Heteroscedasticity

Assumption : Error term has a constant variance

Holds assumption : Homoscedasticity

Violates : Heteroscedasticity (problem of cross sectional data)

(> Autocorrelation was the problem of time series data.)

Proposional vector : term for which variance is not remaining constant.

Illness : Issues in significance

Significant variable seems insignificant and vice versa.

Measures :

1. Park test

   Take log

2. White test

   H0 : no heteroscedasticity

   P>0.05 Accept

   P<0.05 Reject

Removals :

Weighted least square method

Hac test

As we have seen in the 4th assumption of CLRM equal variance or homoscedasticity, in heteroscedasticity we have unequal variance

$E(u_i^2) = \sigma_i^2$ Here the subscript denotes that variance is not equal, it is variable.

⇨ Sources of heteroscedasticity (L DC O A S)

1) In the models to learn anything to them in time their **Learning** capability will increase and hence variance decreases and hence heteroscedasticity comes into picture.
2) As **Data Collecting** techniques improve, σ^2_i is likely to decrease.
3) Heteroscedasticity is mainly due to the presence of **Outlier** in the data. Outlier in Heteroscedasticity means that the observations that are either small or large with respect to the other observations are present in the sample.
4) Another source of heteroscedasticity arises from violating **Assumption** 9 of CLRM, namely, that the regression model is correctly specified.
5) Another source of heteroscedasticity is **Skewness** in the distribution of one or more regressors included in the model.

⇨ Consequences of Heteroscedasticity

1) The OLS estimators and regression predictions based on them remains unbiased and consistent.
2) The OLS estimators are no longer the BLUE (Best Linear Unbiased Estimators) because they are no longer efficient, so the regression predictions will be inefficient too.
3) Because of the inconsistency of the covariance matrix of the estimated regression coefficients, the tests of hypotheses, (t-test, F-test) are no longer valid.

⇨ Detection of Heteroscedasticity

Informal methods

1) Nature of the Problem
2) Graphical method

Formal Methods

3) Park Test
4) Glejser Test
5) Spearman's Rank Correlation Test
6) Goldfeld-Quandt Test
7) Breusch–Pagan–Godfrey Test

8) White's General Heteroscedasticity Test

⇨ Remedial Measures

- When σ2 i Is **Known**: The Method of Weighted Least Squares

  The most straightforward method of correcting heteroscedasticity is by means of **weighted least squares**, for the estimators thus obtained are BLUE.

- When σ2 i Is **Not Known**

  White's Heteroscedasticity-Consistent Variances and Standard Errors.

  The most common way to correct heteroskedasticity is to calculate **robust standard errors (RSE)** (also called **White**-corrected standard errors or heteroskedasticity-consistent standard errors). These robust standard errors are then used to recalculate the f-statistics using the original regression coefficients.

Autocorrelation

Assumption : error term observations are not correlated with each other.

GDP, investment, consumption, income are autocorrelated (with itself).

Error should be random.

If error term is autocorrelated than there is a problem. (There should be not consistency)

Error predicts same sign : positive correlation

Error predicts opposite sign : negative correlation

Illness : No reliable of significance

Measure : Derwan Watson statistics

0-2 +ve

2- no autocorrelation

2-4 -ve

Serial correlation test : H0 no autocorrelation

p>0.05 : accept

P<0.05 : reject

Removals :

1. Addition of relavant variables.
2. GLS (Generilzed Least Square) or cockrain ocrad.
3. AR1
4. For large size test : HAC test

Autocorrelation refers to the degree of correlation between the values of the same variables across different observations in the data.

Symbolically, $E(u_i u_j) = 0$, i != j

⇨ Source of Serial Correlation

1) Specification Bias: **Excluded Variables** Case

$$Y_t = \beta_1 + \beta_2 X_{2t} + \beta_3 X_{3t} + \beta_4 X_{4t} + u_t$$

tity of beef demanded, X_2 = price of beef, X
ice of pork, and t = time.[4] However, for some
gression:

$$Y_t = \beta_1 + \beta_2 X_{2t} + \beta_3 X_{3t} + v_t$$

In regression modeling, it is not possible to include all the variables in the model. There can be various reasons for this, e.g., some variable may be qualitative, sometimes direct observations may not be available on the variable etc. The joint effect of such deleted variables gives rise to autocorrelation in the data.

2) Specification Bias: **Incorrect Functional Form**

$$\text{Marginal cost}_i = \beta_1 + \beta_2 \text{ output}_i + \beta_3 \text{ output}_i^2 + u_i$$

he following model:

$$\text{Marginal cost}_i = \alpha_1 + \alpha_2 \text{ output}_i + v_i$$

It is assumed that the form of relationship between study and explanatory variables is linear. If there are log or exponential terms present in the model so that the linearity of the model is questionable, then this also gives rise to autocorrelation in the data.

3) **Cobweb** Phenomenon. The supply of many agricultural commodities reflects the so-called cobweb phenomenon, where supply reacts to price with a lag of one time period because supply decisions take time to implement (the gestation period).

$$\text{Supply}_t = \beta_1 + \beta_2 P_{t-1} + u_t$$

4) **Lags**:

$$\text{Consumption}_t = \beta_1 + \beta_2 \text{ income}_t + \beta_3 \text{ consumption}_{t-1} + u_t \qquad \textbf{(12.1.7)}$$

A regression such as (12.1.7) is known as autoregression because one of the explanatory variables is the lagged value of the dependent variable. This might lead to autocorrelation.

5) "**Manipulation**" of Data.

Often manipulation of data results into this problem. Manipulation happens in many ways, firstly if adding any value to current data then it might lead to autocorrelation. Secondly using the techniques such as interpolation or extrapolation can cause a severe effect on the data.

6) **Data Transformation**.

7) **Stationary**

$$Y_t = \beta_1 + \beta_2 X_t + u_t \qquad \textbf{(12.1.8)}$$

In a regression model such as (12.1.8), it is quite possible that both Y and X are nonstationary and therefore the error u is also nonstationary. In that case, the error term will exhibit autocorrelation

⇨ Remedies

1) If the autocorrelation is pure autocorrelation, then one can use appropriate transformation of the original model so that in the transformed model we do not have the problem of (pure) autocorrelation.

2) In large samples, we can use the Newey-West method to obtain standard errors of OLS estimators that are corrected for autocorrelation (extension of White's heteroscedasticity-consistent standard errors method.

Difference between correlation and regression

Correlation | Regression

1. Nature of Relationship :

- Tests the closeness of variables

- Measures the extent of change in dependent variable y due to change in independent variable X.

2. Relationship

- Only closeness of variables is studied. it does not study cause and effect relationship between variables.

- Cause and effect relationship in variables moving in same or opposite direction is studied

3. Mutual dependence of variables

- Studies mutual dependence of variables

- In regression analysis the functional relationship showing dependence of one variable upon other is analyzed

4. Spurious correlation

- chances of spurious correlation b/w two variables having no practical importance may be observed.

- There is no chance of existence of such type of relation in regression analysis.

5. Mathematical treatment

- no scope for further mathematical treatment

- can be used for further mathematical treatment

6. Origin and scale

- is independent of change of origin as well as change of scale.

- is independent of change of origin but not of change of scale.

7. relative and absolute measures

- it is a relative measure of linear relationship between x and y and is independent of measurement. it is a number which lies from +1 to -1.

- it is an absolute measure showing the change in the value of y or x for unit change in value of x or y.

8. Applicability

- It has very limited scope of application. it is limited to linear relationship between two variable.

- scope of applicability is very wide. It can be covered under linear as well as non linear relationship b/w variables.

9. Differentiation in variables

- Both variables are considered at per study purpose

- Variables are differentiated as dependent and independent variables.

10. Symmetrical or asymmetrical formation

- correlation is symmetrical in formation

    - rxy = ryx

    - r is a both way relationship of x on y or y on x.

- the approach is treating one variable as dependent and other as independent thus making analysis asymmetrical i.e.

    - byx != bxy

Measurement scale

Ratio

Interval - ratio doesn't matter

Ordinal - natural ordering, distance doesn't matter

Nominal - nothing matters. they are categories.

Signifiance of stochastic disturbance term

1. Vagueness of theory

2. Unavailability of data

3. Core variables vs peripheral variables

4. Instrinstic randomness in human behaviour

5. Poory proxy variables

6. Principal of parcimony

7. Wrong functional form

CLRM Assumptions

1. Linear in parameters

2. X values fixed in repeated sampling

3. zero mean value of ui

4. Homoscedasticity

5. No autocorrelation b/w disturbances

6. Zero covariance b/w ui and x

7. number of observations > numbers of parameters

8. Variability in X values

9. Regression model is correctly specified

10. No perfect multicolinearity

Properties of R

properties of r

1. Positive, Negative :

sample cov. of 2 variables

2. $-1 < r < 1$

3. symmetrical :

$r_{xy} = r_{yx}$

4. independent of origin and scale

5. zero correlation !=> independence

6. measure of lienar dependence; no meaning for nonlinear relations.

7. not imply any cause-and-effect relationship.

Normality assumtions

1. CLT

normality of u

(large ind. distributed random vars.)

2. number of var. not very large then also

their sum still be normally distributed

3. b1, b2 - linear functions of ui

=> ui normally distributed.

>linear function of normally distributed vars

is itself normally distributed

4. simple distribution involves mean,

variance. seems to follow normal distrib.

5. enables us to use t,F,X^2 tests along

with helping us to derive the exact

prob. distr. of OLS.

Properties of OLS estimators under normality assumptions:

1. unbiased

2. minimum variance

3. consistency

4. b1 normally distributed

5. b2 normally distributed

6. chi square distribution

7. b1,b2 distributed independently of variance.

8. b1, b2 have min variance in entire class of unbiased estimators (BUE)