Q.1

A

20-PBD-002

Shraddha P Jain

3951: Introduction to Econometrics and Finance

CIA-I Assignment

Q.1 What do you mean by emperical analysis? Explain the steps to be kept in mind while undertaking econometric analysis

Ans: An analysis that uses data to test a theory or to estimate a relationship is called emperical analysis. In emperical analysis, you test a theory or an relationship statistically using data, and draw conclusions from it.

The steps to be kept in mind while undertaking ~~economic~~ econometric analysis are

1. Carefully pose a question: You should clearly define # your objective of the analysis, including the theory that you wish to test. Based on your question, your hypothesis will be formulated.

2 Specify an economic or conceptual model: An economic model consists of a functional setup ~~mathematical equations~~ that describe various relationships between the variables under study

It is usually of the form:

variable under study
= f (various other factors that we have)

Q. 2

A

and

2 Turn the above economic model into econometric model:

In an economic model, you do not have any parameters that allow you to explicitly measure the effect of various factors on the target variable.

By converting an economic model to an econometric model, you are building a setup that allows you to measure the ~~most~~ impact of each of the factors on the variable under study.

An econometric model is a mathematical formulation of the problem.

An example of econometric model is:

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \mu.$$

where Y is the variable under study

x_1, x_2 are the factors on which Y depends

- Collect the data on variables, and use statistical methods to estimate parameters, construct confidence intervals, and test hypotheses.

Q. 2. (i) Define cross-sectional data and time series data.

Ans

A cross-sectional data set consists of a random sample of individuals, households, firms or a variety of other units, taken at a given point in time.

Sometimes, the data on all units do not correspond to precisely the same time period, but we can ignore minor timing differences in collecting the data.

A key feature of cross-sectional data is that the ordering of data does not matter in econometric analysis

A cross-sectional data does not have time dimension attached to it

A time series data set consists of observations on a variable or several variables over time, for eg example stock prices, GDP, etc.

The chronological ordering of observations in time series conveys potentially important information, and the observations can rarely be assumed to be independent across time.

Q 2 (ii) Define Panel data and explain the ingredients of Panel Data regression model.

Ans A Panel data also called longitudinal data set consists of time series for each cross-sectional member in dataset.

Eg:, we have wage, education and employment history of a set of 1000 individuals followed over a 10 year period.

In Panel Data regression model; the unobservable characteristics (like a person's beauty, ability, etc) are by necessity excluded from the set of explanatory variables, and included in random error term.

⇒ Panel Data Regression Model, we have the following components (ingredients):

1) The variables that do not change over time (have no time subscript). are called time invariant, and are represented as:
$w_{i1}, w_{i2}, \ldots w_{im} : \omega_i$.

2). There are also unobserved, omitted factors in each time period, for each individual that will compose the error terms of regression

(a) Unobserved, time invariant individual characteristics, called unobserved heterogeneity are represented as $u_{1i}, u_{2i}, \ldots, u_{ki} : u_i$

These u_i's summarise the unobserved factors leading to individual differences

(b) The unobserved, individual time varying characteristics, called ~~idiosyntatic~~ idiosyncratic errors are represented as:

$$e_{1it}, e_{2it}, \ldots e_{kit} : e_{it}$$

(c) The unobserved, time specific error that varies ~~not~~ over time but not individuals is represented as

$$m_{1t}, m_{2t} \ldots m_{Tt} : m_t$$

(3) $x_{1it} = 1$ is the intercept, with x_{2it}, x_{kit} being observations on $k-1$ factors that vary across individual and time!

(4) The y_{it} values represent the outcome variable, dependent on both individual, and time.

Hence, the regression model for panel data is represented as:

$$y_{it} = \beta_1 + \beta_2 x_{2it} + \beta_3 x_{3it} \ldots + a_1 w_{it} + \ldots + (u_i + e_{it}).$$

Q3. Elaborate the main advantages of Panel Data models.

Ans The advantages of Panel Data model are:

1. They are realistic and flexible:
A pure cross-sectional data model does not take into account change of parameters over time while a pure time series data does not take into account change of parameters accross individuals.
A panel data model takes into account of both of those things.

2. They are more efficient:
Panel data models provide estimates with least variance, as there are more number of observations, and hence, higher degrees of freedom, leading to more efficient parameters.

- Panel data models enable identification of certain parameters, without the need to make restrictive assumptions.

- Panel data models can identify dynamic processes

5. They can easily correct time-invariant unobservable effects, correlated with observable regressors

6. Panel data models are in general more informative, lesser variance of parameters, more efficient, more flexible and realistic.

Q. 4. Discuss error component in a Panel Data Regression model.

Ans In Panel Data regression model, the unobservable, omitted factors in each time period, for each individual compose the regression's random error term.

In Panel Data models, we can identify several type of unobserved effects.

1. First are the unobserved / unmeasureable, time invariant individual characteristics, denoted as $u_{1i}, u_{2i}, \ldots u_{ki}$.

But these cannot be observed, so ~~the~~ their combined effect is represented as u_i.

u_i represents unobserved heterogeneity, and summarises the unobserved factors leading to individual differences.

2) Second are the time- as well as individual-varying unobservable/unmeasurable factors denoted as $e_{1t}, e_{2t}, \ldots$, and their combined effect is e_{it}. These errors correspond to the usual type of errors of regression, and are called idiosyncratic errors.

3) Third are the time ~~specific~~ variant, individual invarying error components. represented as $m_{1t}, m_{2t}, \ldots$ and their combined effect represented as m_t.

Q 5. Discuss fixed effects estimator

Ans. The fixed effects model is simply a linear regression model in which the intercept terms vary over individual units i, that is.

$$y_{it} = \alpha_i + x_{it}'\beta + \mu_{it} \qquad \mu_{it} \sim IID(0, \sigma_u^2)$$

where it is usually assumed that all x_{it} are independent of all μ_{it}.

The fixed effects models all follow the strict exogeneity assumption, that states that:

Given the values of explanatory variable, in all time periods, time invariant and unobserved heterogeneity, the best prediction of idosyncratic error is 0,

i.e., $E(e_{it} | X_{2it}, \omega_i, u_i) = 0$.

i.e., there is no information in these factors about the idosyncratic error.

Note that this assumption does not require the unobserved heterogeneity (ω) to be uncorrelated with the values of explanatory variables.

The other assumptions of fixed effect ~~estimators~~ models is:

1. model has a parameter and individual specific effect.
2. The samples should have been taken randomly.
3. There should be no perfect collinearity between the explanatory variables.
4. There should be no homoskedasticity.
5. There should be no serial correlation among the ~~idosyncle~~ idosyncratic errors.
6. $e_{it} \sim N(0, \sigma^2)$.

Under fixed effects estimators, we have the following estimators

1. difference estimator
2. within estimator
3. Least Square Dummy Variable estimator

1. Difference estimator →

This method of estimation best works when you have panel data with as few as $T=2$ observations per individual.

The two observations can be written as:

$$y_{i1} = \beta_1 + \beta_2 x_{2i1} + \alpha_1 \omega_{1i} + \mu_i + e_{i1} \quad -①$$

$$y_{i2} = \beta_1 + \beta_2 x_{2i2} + \alpha_1 \omega_{1i} + \mu_i + e_{i2} \quad -②$$

Subtracting eq" ② from ①, we have.

$$(y_{i2} - y_{i1}) = \beta_2 (x_{2i2} - x_{2i1}) + (e_{i2} - e_{i1}) \quad -③.$$

OR $\Delta y_i = \beta_2 \Delta x_{i2} + \Delta e_i$

The OLS estimator of eq" ③ is called the difference estimator.

The time-invariant terms $\beta_1, \alpha_1 \omega_{1i}, u_i$ have been removed by subtraction

The difference estimator is consistent if:

① Δe_i has zero mean

② Δe_i is uncorrelated with Δx_{i2}

③ Δx_{i2} takes more than two values.

2. Within Estimator →

Within estimator is similar to difference estimator, but with the advantage that it can be generalized nicely to situations the data having $T > 2$.

First we eliminate the individual effects a_i by transforming the data, and finding time average of the equations.

We get the following equation.

$$\bar{y}_{i.} = \beta_1 + \beta_2 \bar{x}_{2i.} + a_1 w_{i.} + u_i + \bar{e}_{i.} \quad - (1)$$

Note that the averaging does not affect the model parameters or time invariant terms β_1, $w_{i.}$, and u_i

Now, within transformation subtracts eq (1) from original observations to obtain

$$y_{it} - \bar{y}_{i.} = \beta_2 (x_{2it} - \bar{x}_{2i.}) + (e_{it} - \bar{e}_{i.}) \quad - (2).$$

OR $$\tilde{y}_{it} = \beta_2 \tilde{x}_{2it} + \tilde{e}_{it} \quad - (3).$$

The OLS estimator of β_2 in this case is called the within estimator.

Note that here, instead of the first differences we have differences from variable means

The within estimator is consistent if-

1) $E(\tilde{e}_{it}) = 0$
2) $Cov(\tilde{e}_{it}, \tilde{x}_{jit}) = 0$
3) $\tilde{x}_{jit}$ takes more than two values.

3 Least Square Dummy Variable (LSDV) estimator

Consider the general regression equation:

$$y_{it} = \beta_1 + \beta_1 x_{2it} + \ldots + \beta_k x_{kit} + \alpha_1 \omega_{1i} + \ldots + \alpha_m \omega_{mi} + (u_i + e_{it}). \quad \text{—(1)}$$

In the above equations, $x_{1it} = 1$ & $(k-1) = K_s$ variable vary ~~across~~ across individual and time, also M variable that are time invariant.

We now introduce dummy variable D_{ji} for each of the unit i in model to control unobserved heterogeneity. as.

$$D_{1i} = \begin{cases} 1 & i=1 \\ 0 & \text{otherwise} \end{cases} \quad \ldots \quad D_{Ni} = \begin{cases} 1 & i=N \\ 0 & \text{otherwise} \end{cases}$$

So, the regression equation now becomes:

$$y_{it} = \beta_{11} D_{1i} + \beta_{12} D_{2i} + \ldots + \beta_{1N} D_{Ni} + \beta_1 + \beta_2 x_{2it} + \ldots + \beta_k x_{kit} + \alpha_i \omega_{it} + \ldots + \alpha_m \omega_{mi} + (u_i + e_{it}) \quad \text{—(2)}$$

But in the above equation we have exact collinearity as

$$D_{1i} + D_{2i} + \dots + D_{Ni} = 1$$

To deal with this, we drop the constant term $x_{1it} = 1$, time-invariant variables $w_{1i}, w_{2i}, \dots, w_{mi}$, and the unobserved heterogeneity u_i.

∴ We now have

$$y_{it} = \beta_1 D_{1i} + \beta_2 D_{2i} + \dots + \beta_N D_{Ni} + \beta_2 x_{2it} + \dots + \beta_k x_{kit} + e_{it} \quad \text{---(3)}$$

Eq (3) is called LSDV ~~estimat~~ model, and the OLS estimators of eq (3) parameters are called LSDV estimators.

This estimator is not used in practice unless N is small.

The LSDV estimator is consistent if

1) $E(e_{it}) = 0$.
2) $Cov(e_{it}, x_{2it}) = 0$.
3) x_{2it} takes more than 2 values.