ECON2206

1 Introduction

1.1 What is econometrics

Use of statistical methods to analyse typically observational/nonexperimental data

- Estimate relationships between economic variables
- Forecasting/evaluating impact of policy changes

1.2 Components of econometric analysis

(1) Economic models (theoretical context)

- Micro/macro models establish relationship between economic variables
- Provides a priori expectations

Example: theory of additional training on productivity

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

(2) Econometric models

- Requires data & decisions on functional form
- Form of data impacts model specification/results
- Non-deterministic relationship with error/disturbance

Example: map theory into econometric model interested in β_3

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

(3) Economic Data

- Cross sectional: observations for units of interest at given point in time
  - Reasonable to assume pure random sampling hence independence
- Time series: observations for a unit of interest collected over time
  - Observations typically serially correlated making order important
  - Typical features include trends, cycles & seasonality

- Pooled cross sections: two or more cross sections combined in one data set
    - Samples drawn independently of each other
    - Often used to evaluate effect of policy changes
- Panel/longitudinal: Same cross-sectional unit followed over time
    - Can account for time-invariant unobservables & lagged responses

1.3 Causality & notion of ceteris paribus

Causal effect of x on y

How does y change if x is changed with all other relevant factors held constant?

Many economic questions are ceteris paribus

- Useful to set up experiment designed to find causal inference
- Harder with observational data as x is often related to other factors

2 Simple Regression Model

2.1 Simple regression model

Variable y explained in terms of variable x

$$y = \beta_0 + \beta_1 x + u$$

y is the dependent (explained, response, endogenous)
x is the explanatory (independent, regressor, exogenous)
u is the disturbance/error term containing unobservables
β_0 is the intercept
β_1 is the slope

For β_1 to be causal interpretation

- All other factors held constant
- Zero conditional mean assumption must hold
    - x does not contain information about mean of u
    - u does not depend on x

$$E(u \mid x) = E(u) = 0$$

$$E(y|x) = E(\beta_0 + \beta_1 x + u|x)$$
$$= \beta_0 + \beta_1 x + E(u|x)$$
$$= \beta_0 + \beta_1 x$$

- Average value of y can be expressed as linear function of x

2.2 Deriving OLS estimates

Let { (xi, yi); i = 1, …n } denote a random sample with equation

$$y_i = \beta_0 + \beta_1 + u_i \text{ for each } i$$

(1) Define regression residuals & fitted values

$$\hat{u}_i = y_i - \hat{y}_i = y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)$$

(2) Solve minimisation problem

$$min \sum_{i=1}^{n} \hat{u}_i^{\,2} \rightarrow \hat{\beta}_0, \hat{\beta}_1$$

(3) Obtain OLS estimates

- β_1 is the sample covariance between x_i & y_i divided by the samples variance of x_i

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}, \quad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

2.3 Algebraic properties of OLS

Several useful properties of OLS estimates

(1) Residuals from fitted regression line sum to zero

$$\sum_{i=1}^{n} \hat{u}_i = 0$$

(2) Covariance between residuals & explanatory variable is zero

$$\sum_{i=1}^{n} x_i \hat{u}_i = 0$$

(3) Sample averages of x,y lie on fitted regression line

$$\bar{y} = \hat{\beta}_0 + \hat{\beta}_1 \bar{x}$$

2.4 Goodness of fit

To measure how well OLS regression line fits the data

- Decompose variation in dependent variable
  - Assume sample average of fitted values equals sample average of y_i

$$(y_i - \bar{y}) = (\hat{y}_i - \bar{y}) + \hat{u}_i$$

Total variation from mean = part explained by x + unexplained deviation

$$SST = SSE + SSR$$

SST = total sum of squares
SSE = explained sum of squares
SSR = residual sum of squares

$$SST = \sum_{i=1}^{n} (y_i - \bar{y})^2;\ SSE = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2;\ SSR = \sum_{i=1}^{n} \hat{u}_i{}^2$$

R-squared measures proportion of sample variation in y explained by x

- High r^2 does not necessarily imply causal relationship

- Tend to be lower in cross-sectional data

$$R^2 = \frac{SSE}{SST} = 1 - \frac{SSR}{SST}$$

2.5 Nonlinear relationships

Possible to incorporate non-linear relationships into linear regression model

- More accurately describes many applied problems

Semi-logarithmic specification

- Avoid logging variables measured in units such as percentage points or can take zero/negative values
- Mitigates outliers & helps secure normality/homoscedasticity
- Slope coefficient invariant to rescaling