ECON 5350 Class Notes
The Linear Regression Model

1 Introduction

The **multiple linear regression model** can be written as

$$y_i = \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_k x_{ik} + \epsilon_i \quad i = 1, ..., n \tag{1}$$

where y_i is the dependent variable, x_{ij} is the j^{th} explanatory variable and ϵ_i is the error term.

- There are k explanatory variables.
- There are n observations.
- x_{i1} is often set equal to one, $i = 1, ..., n$, so β_1 is an intercept.
- The βs are coefficients or parameters to be estimated.

Using matrices, model (1) can be written more compactly as

$$Y = X\beta + \epsilon \tag{2}$$

where Y is an $n \times 1$ column vector of dependent variables, X is an $n \times k$ matrix of explanatory variables, β is a $k \times 1$ column vector of parameters and ϵ is an $n \times 1$ column vector of errors.

For example:

$$Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \qquad X = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1k} \\ x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nk} \end{bmatrix} \qquad \beta = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{bmatrix}.$$

2 Data-Generating Assumptions

There are six data-generating assumptions associated with model (1) or (2).

2.1 Linearity

The model must take the form of (1) so that it is linear in the parameters (β) and the error term (ϵ).

- The model need not be linear in the Xs or Ys. However, it must be transformable into a form such as (1).

- For example, after taking natural logs, $y_i = \exp(\beta_1 x_i + \epsilon_i)$ can be transformed into $\ln(y_i) = \beta_1 x_i + \epsilon_i$, which is in the form of (1) with an appropriate redefining of y.

2.2 Full Rank

The columns of X need to be linearly independent and there must be at least k observations.

- In other words, $Rank(X) = k$.

2.3 Mean-Zero Errors

Conditional on X, the error terms are mean zero.

- In other words, $E[\epsilon|X] = 0$.
- This implies that $E[Y|X] = X\beta$, (i.e., the regression of Y on X is the conditional mean $X\beta$).
- Including a constant term will guarantee this assumption holds. Assume $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$ has the property $E[\epsilon_i] = \mu \neq 0$. By redefining the constant term, $\epsilon_i^* = \epsilon_i - \mu$, and intercept $\beta_0^* = \beta_0 + \mu$, the model can be written with mean-zero errors, $y_i = \beta_0^* + \beta_1 x_i + \epsilon_i^*$.

2.4 Spherical Disturbances

The error terms should display **homoscedasticity** (i.e., error variances are constant across observations) and no **autocorrelation** (i.e., errors are uncorrelated across observations).

- Homoscedasticity: $Var[\epsilon_i] = \sigma^2$ for all $i = 1, ..., n$.
- No autocorrelation: $Cov[\epsilon_i, \epsilon_j] = 0$ for all $i \neq j$.
- Matrix representation: $Var[\epsilon] = E[\epsilon\epsilon'] = \begin{bmatrix} Var[\epsilon_1] & Cov[\epsilon_2, \epsilon_1] & \cdots & Cov[\epsilon_n, \epsilon_1] \\ Cov[\epsilon_1, \epsilon_2] & Var[\epsilon_2] & \cdots & Cov[\epsilon_n, \epsilon_2] \\ \vdots & \vdots & & \vdots \\ Cov[\epsilon_1, \epsilon_n] & Cov[\epsilon_2, \epsilon_n] & \cdots & Var[\epsilon_n] \end{bmatrix} = \sigma^2 I_n.$

2.5 Nonstochastic Regressors

The explanatory variables are fixed in repeated sampling.

- This is often true in scientific experiments.
- This is generally not true in the social sciences.
- We can relax the assumption, so long as $corr(X, \epsilon) = 0$.

2.6 Normality

The error terms will follow a normal distribution. This is supported by the Central Limit Theorem and is necessary for inference. It is not necessary, however, to show the optimal properties of least squares estimators.