Chapter 1: Linear regression Models

Tuan A. Luong

De Montfort University - VCREME - IREEDS

June 2018

Table of contents

Specification

- The regression model

$$y_t = \mathbf{x}_\mathbf{t}'\boldsymbol{\beta} + u_t \tag{1}$$

- y_t is the dependent variable. Example: student exam grade
- $\mathbf{x_t}$ is a vector ($kx1$) of independent variables. Example: student characteristics (attendance, entrance score)
- u_t is the error term or the disturbance term. Example: luck.

The goals

- The observed sample: $(y_1, y_2, ..., y_T)$ and $\begin{bmatrix} x_{11}, x_{12}, \cdots, x_{1T} \\ x_{21}, x_{22}, \cdots, x_{2T} \\ \vdots \\ x_{k1}, x_{k2}, \cdots, x_{kT} \end{bmatrix}$
- u_t is not observed.
- Our objectives are:
  - Suppose the first characteristic is attendance. Can we say that more attendance leads to higher grade?
  - Can we predict a student score based on his(her) characteristics?

Linear algebra-OLS

- To estimate the parameter β, we need to minimize the residual sum of squares (RSS):

$$RSS = \sum_{t=1}^{T} (y_t - \mathbf{x}_t' \boldsymbol{\beta})^2 \tag{2}$$

- The solution is given by

$$\hat{\boldsymbol{\beta}} = \left[\sum_{t=1}^{T} (\mathbf{x}_t \mathbf{x}_t') \right]^{-1} \left[\sum_{t=1}^{T} (\mathbf{x}_t y_t) \right] \tag{3}$$

Conditions

- Consider the regression model

$$y_t = \beta_1 x_t + \beta_2 z_t + u_t \tag{4}$$

- Let $X_t = [x_t; z_t]$.
- We can only employ the OLS estimator when $X_t X' t$ can be invertible.
- If x_t is a linear function of z_t, e.g. $x_t = \alpha z_t$ then $X_t X' t$ is not invertible.

Examples

- Let $y_t = 5 * x_t$.
- Let $z_t = 2 * x_t$.
- If we have $y_t = \alpha x_t + \beta z_t$.
- What is the value of α and β?

Case 1: Simple regression without constant

- When $k = 1$ we have:

$$y_t = \beta x_t + u_t \tag{5}$$

- Apply Equation 3 we have:

$$\hat{\beta} = \frac{\sum_{t=1}^{T} y_t x_t}{\sum_{t=1}^{T} x_t^2} \tag{6}$$

Case 2: Multivariate regression without constant

- When $k = 2$ we have:

$$y_t = \beta_1 x_t + \beta_2 z_t + u_t \tag{7}$$

- Apply Equation 3 we have:

$$\begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{bmatrix} = \begin{bmatrix} \sum_{t=1}^{T} x_t^2 & \sum_{t=1}^{T} x_t z_t \\ \sum_{t=1}^{T} x_t z_t & \sum_{t=1}^{T} z_t^2 \end{bmatrix}^{-1} \begin{bmatrix} \sum_{t=1}^{T} y_t x_t \\ \sum_{t=1}^{T} y_t z_t \end{bmatrix} \tag{8}$$

Case 2: Multivariate regression without constant (cont')

- Note that

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix}^{-1} = \frac{1}{ad - bc} \begin{bmatrix} d & -b \\ -c & a \end{bmatrix} \tag{9}$$

- We then have:

$$\hat{\beta}_1 = \frac{\sum_{t=1}^{T} y_t x_t \sum_{t=1}^{T} z_t^2 - \sum_{t=1}^{T} y_t z_t \sum_{t=1}^{T} x_t z_t}{\sum_{t=1}^{T} x_t^2 \sum_{t=1}^{T} z_t^2 - (\sum_{t=1}^{T} x_t z_t)^2} \tag{10}$$

- What is the estimate of β_2?

Case 3: Single regression with constant

- Regression model:

$$y_t = \beta_0 + \beta_1 x_t + u_t \tag{11}$$

- What is the estimate of β_1?

Case 4: Multivariate regression with constant

- Regression model:

$$y_t = \beta_0 + \beta_1 x_t + \beta_2 z_t + u_t \tag{12}$$

- if we denote $\bar{x} = \frac{\sum_{t=1}^{T} x_t}{T}$, $\bar{z} = \frac{\sum_{t=1}^{T} z_t}{T}$, $\bar{y} = \frac{\sum_{t=1}^{T} y_t}{T}$
- then we have

$$\hat{\beta}_1 = \frac{\sum_{t=1}^{T}(y_t - \bar{y})(x_t - \bar{x})}{\sum_{t=1}^{T}(x_t - \bar{x})^2} \tag{13}$$

- and

$$\hat{\beta}_2 = \frac{\sum_{t=1}^{T}(y_t - \bar{y})(z_t - \bar{z})}{\sum_{t=1}^{T}(z_t - \bar{z})^2} \tag{14}$$

Assumptions

- Assumption 1: the vector $\mathbf{x_t}$ is deterministic.
- Assumption 2: the error term u_t is random vector following an identical and independent distribution with zero mean and variance σ^2.
- Assumption 3: u_t is Gaussian.

Unbiasedness

- Under Assumptions 1 and 2, the OLS estimators are unbiased.
- In other words, they are the random variables with mean equal the true value:

$$E(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta}$$

Estimator variance

- Under Assumptions 1 and 2, the OLS estimators are random variables with variance.

$$Var(\hat{\beta}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1} \tag{15}$$

OLS estimators distribution

- Under Assumptions 1,2 and 3, the OLS estimators follow the normal distribution.

$$\hat{\boldsymbol{\beta}} \sim N(\boldsymbol{\beta}, \sigma^2(\mathbf{X}'\mathbf{X})^{-1}) \tag{16}$$

- Under Assumptions 1,2 and 3, no unbiased estimators are more efficient than the OLS estimators: they all have bigger variance.

Example: Single regression model without constant

- Recall the regression model:

$$y_t = \beta x_t + u_t$$

- Under Assumption 1,2 and 3, the OLS estimator $\hat{\beta}$ follow the normal distribution.

$$\hat{\beta} \sim N(\beta, \frac{\sigma^2}{\sum_{t=1}^{T} x_t^2})$$

- the larger the sample size, the more accurate the estimator.

R-squared

- Based on the estimators, we can predict

$$\hat{y}_t = \mathbf{x}_\mathbf{t}' \hat{\boldsymbol{\beta}} \tag{17}$$

- How well this predicted value fit the sample data?
- When the model has the constant term, the centered R^2 is given by:

$$R^2 = \frac{\sum_{t=1}^{T} (\hat{y}_t - \bar{y})^2}{\sum_{t=1}^{T} (y_t - \bar{y})^2} \tag{18}$$

- This goodness of fit varies from 0 to 1: the higher R^2, the better fit the model.

Residuals variance

- Practically, we might not know the variance σ^2 of the error terms u_t.
- In this case we can estimate the variance as:

$$\hat{\sigma}^2 = \frac{\sum_{t=1}^{T}(y_t - \hat{y}_t)^2}{T - k} \tag{19}$$

- Note that $\hat{u}_t = y_t - \hat{y}_t$ is the estimated error term.
- We can show that

$$E(\hat{\sigma}^2) = \sigma^2$$