Lecture notes III 03.09.2007

Linear regression with one regressor. (SW ch. 4)

The linear regression model

$$Y_i = \beta_0 + \beta_1 X_i + u_i$$

where

i is observation unit: $i = 1, \dots, n$

Y_i is dependent variable

X_i is the regressor (independent var.)

β_0 is intercept (parameter)

β_1 is slope coeff. (parameter)

$\beta_0 + \beta_1 X_i$ is regression line

u_i is the error term (unobserved)

Main identifying assumption (A.1)

A.1: $E(u_i \mid X_i) = 0$

$$\Rightarrow E(Y_i \mid X_i) = \beta_0 + \beta_1 X_i \quad \text{(regression line)}$$

$$\beta_1 = \frac{\partial E(Y_i \mid X_i)}{\partial X_i} \equiv \text{marginal effect}$$

Example: - Y_i is score / grade of a student
- X_i is the schools # teachers per student

Note that A.1 $\Rightarrow Cov(u_i, X_i) = 0$

since a) $E(u_i) = E(E(u_i \mid X_i)) = 0$

rule of double / iterated expectation

b) $Cov(u_i, X_i) = E(u_i X_i) - \overbrace{E(u_i)}^{=0} E(X_i)$

$= E(u_i X_i) = E(\underbrace{E(u_i \mid X_i)}_{=0} X_i) = 0$

Estimation of β_0 and β_1

The Ordinary Least Squares (OLS) estimator

minimizes $\underbrace{\sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)^2}_{= SSR \text{ (sum of squares of residuals)}}$ with respect

to β_0 and β_1

1. order condition:

$$\frac{\partial SSR}{\partial \beta_0} = 2 \sum_{i=1}^{n} (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i) = 0 \qquad (1)$$

$$\frac{\partial SSR}{\partial \beta_1} = 2 \sum_{i=1}^{n} (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i) X_i = 0 \qquad (2)$$

(3): From (1) $\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$, $\bar{X} = \frac{1}{n} \sum_i X_i$, $\bar{Y} = \frac{1}{n} \sum_i Y_i$

Then, from (2)

$$\sum_{i=1}^{n} (Y_i - \underbrace{(\bar{Y} - \hat{\beta}_1 \bar{X})}_{= \hat{\beta}_0} - \hat{\beta}_1 X_i) X_i = 0$$

$$\Leftrightarrow \sum_{i=1}^{n} (Y_i - \bar{Y}) X_i = \hat{\beta}_1 \sum_{i=1}^{n} (X_i - \bar{X}) X_i$$

$$\Leftrightarrow \hat{\beta}_1 = \frac{\sum_{i=1}^{n} (X_i - \bar{X}) Y_i}{\sum_{i=1}^{n} (X_i - \bar{X}) X_i}$$

Note that

$$\sum_{i=1}^{n} (X_i - \bar{X}) X_i = \sum_{i=1}^{n} (X_i - \bar{X})(X_i - \bar{X}) \quad \text{and}$$

$$\sum_{i=1}^{n} (X_i - \bar{X}) Y_i = \sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y})$$

Hence

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n} (X_i - \bar{X})^2} = \frac{S_{XY}}{S_{XX}}$$

where for <u>any</u> variables X and Z

$$S_{XZ} \equiv \frac{1}{n} \sum_{i=1}^{n} (X_i - \bar{X})(Z_i - \bar{Z})$$ and where

$\bar{X}$ and $\bar{Z}$ is average of X_i and Z_i, respectively

That is S_{XZ} is the <u>empirical</u> covariance between X and Z.

Why is this a meaningful estimator?

$$\hat{\beta}_1 = \frac{S_{XY}}{S_{XX}} \to \frac{Cov(X,Y)}{Var(X)}$$ by ~~central limit thm.~~ law of large numbers

Furthermore, the true β_1 satisfies

$$Y_i - E(Y_i) = \beta_1 (X_i - E(X_i)) + u_i$$

$$\Downarrow$$

$$(Y_i - \mu_Y) X_i = \beta_1 (X_i - \mu_X) X_i + u_i X_i$$

Taking expectations of both sides gives:

$$Cov(Y_i, X_i) = \beta_1 Var(X_i) + \underbrace{Cov(u_i, X_i)}_{= 0}$$

$$\Downarrow$$

$$\beta_1 = \frac{Cov(Y_i, X_i)}{Var(X_i)}$$

Hence $\hat{\beta}_1$ is a moment estimator: theoretical moments are replaced by empirical ones!

The least squares assumption:

Under certain key assumptions the OLS estimator is:

i) unbiased: $E(\hat{\beta}_1) = \beta_1$, $E(\hat{\beta}_0) = \beta_0$

ii) consistent: $\hat{\beta}_1 \xrightarrow{p} \beta_1$ and $\hat{\beta}_0 \xrightarrow{p} \beta_0$

iii) $\dfrac{\hat{\beta}_1 - \beta_1}{SE(\hat{\beta}_1)} \sim N(0,1)$ when n is large, i.e. asymptotic normality

$$\frac{\hat{\beta}_0 - \beta_0}{SE(\hat{\beta}_0)} \sim N(0,1)$$

Key OLS assumptions:

A.1 $E(u_i \mid X_i) = 0$ $(\Rightarrow Cov(u_i, X_i) = 0)$

A.2 (X_i, Y_i) for $i = 1, \ldots, n$ are i.i.d vectors of random variables
[Alternatively: (X_i, u_i) are i.i.d]

A.3 $E(X_i^4) < \infty$ and $E(u_i^4) < \infty$

$\Rightarrow$ the probability of ~~the~~ extreme data is negligible