Introductory Econometrics

Lecture 4: Simple Linear Regression and OLS

Instructor: Ma, Jun

Renmin University of China

March 7, 2023

Introduction

- The simple linear regression model is used to study the relationship between two variables.
- It has many limitations, but nevertheless there examples in the literature where the simple linear regression is applied (e.g. stock returns predictability).
- It is also a good starting point to learning the regression technique.

Definitions

TABLE 2.1

Terminology for Simple Regression

y	x
Dependent variable	Independent variable
Explained variable	Explanatory variable
Response variable	Control variable
Predicted variable	Predictor variable
Regressand	Regressor

Sample and population

- The econometrician observes random data:

observation	explained variable	explanatory variable
1	Y_1	X_1
2	Y_2	X_2
$\vdots$	$\vdots$	$\vdots$
n	Y_n	X_n

- A pair X_i, Y_i is called an observation.
- Sample: $\{(X_i, Y_i) : i = 1, \ldots, n\}$.
- The population is the joint distribution of the sample.

The model

- We model the relationship between Y and X using the conditional expectation:

$$\mathrm{E}\left[Y_i \mid X_i\right] = \alpha + \beta X_i.$$

- Intercept: $\alpha = \mathrm{E}\left[Y_i \mid X_i = 0\right]$.
- Slope: β measures the effect of a unit change in X on Y :

$$\begin{aligned} \beta &= \mathrm{E}\left[Y_i \mid X_i = x+1\right] - \mathrm{E}\left[Y_i \mid X_i = x\right] \\ &= \left[\alpha + \beta(x+1)\right] - \left[\alpha + \beta x\right]. \end{aligned}$$

- $$\beta = \frac{d\mathrm{E}\left[Y_i \mid X_i\right]}{dX_i}.$$

- The effect is the same for all x!

The model

- α and β in $\mathrm{E}[Y_i \mid X_i] = \alpha + \beta X_i$ are unknown!
- Residual (error):

$$U_i = Y_i - \mathrm{E}[Y_i \mid X_i] = Y_i - (\alpha + \beta X_i).$$

  U_i's are unobservable!
- The model:

$$\begin{aligned} Y_i &= \alpha + \beta X_i + U_i, \\ \mathrm{E}[U_i \mid X_i] &= 0. \end{aligned}$$

Functional form

- We consider a model that is linear in the coefficients $\alpha, \beta : Y_i = \alpha + \beta X_i + U_i$.
- The dependent variable and the regressor can be nonlinear functions of some other variables.
- The most popular function is $\log$.

Functional form: the log-linear model

- Consider the following model:

$$\log Y_i = \alpha + \beta X_i + U_i.$$

- In this case

$$\begin{aligned} \beta &= \frac{d\left(\log Y_i\right)}{dX_i} \\ &= \frac{dY_i/Y_i}{dX_i} = \frac{dY_i/dX_i}{Y_i}. \end{aligned}$$

- β measures percentage change in Y as a response to a unit change in X.
- In this model, it is assumed that the percentage change in Y is the same for all values of X (constant).
- In $\log\left(\text{Wage}_i\right) = \alpha + \beta \times \text{Education}_i + U_i$, β measures the return to education.

Functional form: the log-log model

- Consider the following model:

$$\log Y_i = \alpha + \beta \log X_i + U_i.$$

- In this model,

$$\beta = \frac{d \log Y_i}{d \log X_i} = \frac{dY_i/Y_i}{dX_i/X_i} = \frac{dY_i}{dX_i}\frac{X_i}{Y_i}.$$

- β measures the elasticity: the percentage change in Y as a response to 1% change in X. Here, the elasticity is assumed to be the same for all values of X.
- Example: Cobb-Douglas production function: $Y = \alpha K^{\beta_1} L^{\beta_2} \implies \log Y = \log \alpha + \beta_1 \log K + \beta_2 \log L$ (two regressors log of capital and log of labour).

Orthogonality of residuals

The model

$$Y_i = \alpha + \beta X_i + U_i.$$

We assume that $\mathrm{E}[U_i \mid X_i] = 0$.

- $\mathrm{E}[U_i] = 0$.

$$\mathrm{E}[U_i] \overset{\text{Law of Iterated Expectation}}{=} = \mathrm{E}[\mathrm{E}[U_i \mid X_i]] = \mathrm{E}[0] = 0.$$

- $\mathrm{Cov}[X_i, U_i] = \mathrm{E}[X_i U_i] = 0$.

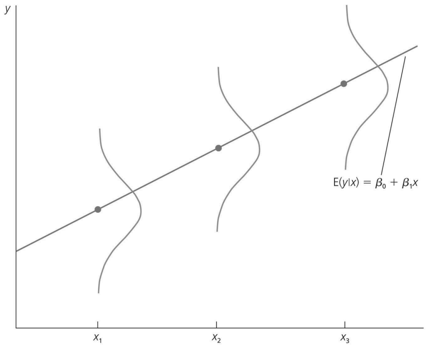
$$\begin{aligned} \mathrm{E}[X_i U_i] &\overset{\text{Law of Iterated Expectation}}{=} \mathrm{E}[\mathrm{E}[X_i U_i \mid X_i]] \\ &= \mathrm{E}[X_i \mathrm{E}[U_i \mid X_i]] = \mathrm{E}[X_i 0] = 0. \end{aligned}$$

The model

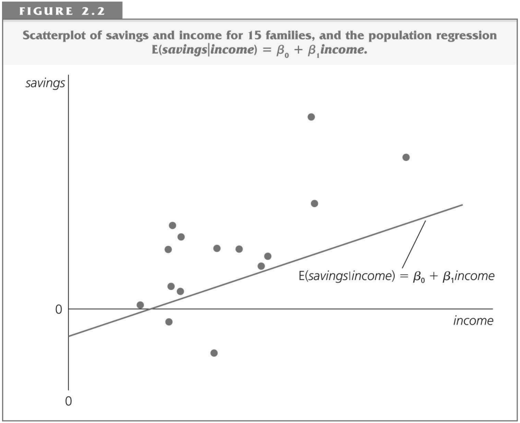
$$Y_i = \underbrace{\alpha + \beta X_i}_{\text{Predicted by } X} + \underbrace{U_i}_{\text{Orthogonal to } X}$$

FIGURE 2.1

E(y|x) as a linear function of x.



Estimation problem



Problem: estimate the unknown parameters α and β using the data (n observations) on Y and X.

Method of moments

- We assume that

$$\begin{aligned} \mathrm{E}\left[U_i\right] &= \mathrm{E}\left[Y_i - \alpha - \beta X_i\right] = 0. \\ \mathrm{E}\left[X_i U_i\right] &= \mathrm{E}\left[X_i\left(Y_i - \alpha - \beta X_i\right)\right] = 0. \end{aligned}$$

- An estimator is a function of the observable data, it can depend only on observable X and Y. Let $\hat{\alpha}$ and $\hat{\beta}$ denote the estimators of α and β.
- Method of moments: Replace expectations with averages. Normal equations:

$$\begin{aligned} \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \hat{\alpha} - \hat{\beta}X_i\right) &= 0. \\ \frac{1}{n}\sum_{i=1}^{n}X_i\left(Y_i - \hat{\alpha} - \hat{\beta}X_i\right) &= 0. \end{aligned}$$

Solution

- Let $\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$ and $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ (averages).
- $\frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \hat{\alpha} - \hat{\beta}X_i\right) = 0$ implies

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n} Y_i - \frac{1}{n}\sum_{i=1}^{n}\hat{\alpha} - \hat{\beta}\frac{1}{n}\sum_{i=1}^{n} X_i &= 0 \text{ or} \\
\bar{Y} - \hat{\alpha} - \hat{\beta}\bar{X} &= 0.
\end{aligned}$$

- The fitted regression line goes through the averages:

$$\hat{\alpha} = \bar{Y} - \hat{\beta}\bar{X}.$$

Solution

$\hat{\alpha} = \bar{Y} - \hat{\beta}\bar{X}$ and therefore

$$\begin{aligned} 0 &= \frac{1}{n}\sum_{i=1}^{n} X_i\left(Y_i - \hat{\alpha} - \hat{\beta}X_i\right) \\ &= \sum_{i=1}^{n} X_i\left(Y_i - \left(\bar{Y} - \hat{\beta}\bar{X}\right) - \hat{\beta}X_i\right) \\ &= \sum_{i=1}^{n} X_i\left[\left(Y_i - \bar{Y}\right) - \hat{\beta}\left(X_i - \bar{X}\right)\right] \\ &= \sum_{i=1}^{n} X_i\left(Y_i - \bar{Y}\right) - \hat{\beta}\sum_{i=1}^{n} X_i\left(X_i - \bar{X}\right). \end{aligned}$$

Solution

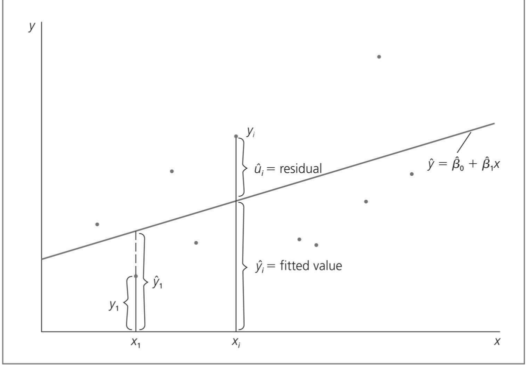
- $$0 = \sum_{i=1}^{n} X_i \left(Y_i - \bar{Y}\right) - \hat{\beta} \sum_{i=1}^{n} X_i \left(X_i - \bar{X}\right) \text{ or}$$
  $$\hat{\beta} = \frac{\sum_{i=1}^{n} X_i \left(Y_i - \bar{Y}\right)}{\sum_{i=1}^{n} X_i \left(X_i - \bar{X}\right)}.$$
- Since
  $$\begin{aligned} \sum_{i=1}^{n} X_i \left(Y_i - \bar{Y}\right) &= \sum_{i=1}^{n} \left(X_i - \bar{X}\right)\left(Y_i - \bar{Y}\right) = \sum_{i=1}^{n} \left(X_i - \bar{X}\right) Y_i \text{ and} \\ \sum_{i=1}^{n} X_i \left(X_i - \bar{X}\right) &= \sum_{i=1}^{n} \left(X_i - \bar{X}\right)\left(X_i - \bar{X}\right) = \sum_{i=1}^{n} \left(X_i - \bar{X}\right)^2 \end{aligned}$$
  we can also write
  $$\hat{\beta} = \frac{\sum_{i=1}^{n} \left(X_i - \bar{X}\right) Y_i}{\sum_{i=1}^{n} \left(X_i - \bar{X}\right)^2}.$$

Fitted line

- Fitted values: $\hat{Y}_i = \hat{\alpha} + \hat{\beta} X_i$.
- Fitted residuals: $\hat{U}_i = Y_i - \hat{\alpha} - \hat{\beta} X_i$.

FIGURE 2.4

Fitted values and residuals.

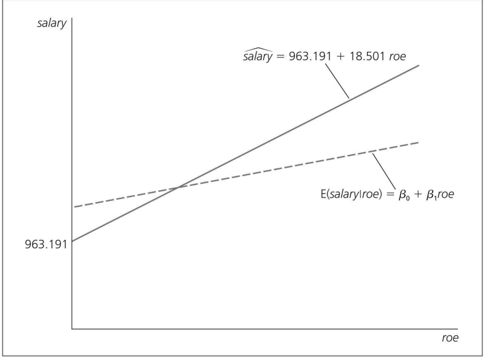


True line and fitted line

- True: $Y_i = \alpha + \beta X_i + U_i$, $\mathrm{E}[U_i] = \mathrm{E}[X_i U_i] = 0$.
- Fitted: $Y_i = \hat{\alpha} + \hat{\beta} X_i + \hat{U}_i$, $\sum_{i=1}^{n} \hat{U}_i = \sum_{i=1}^{n} X_i \hat{U}_i = 0$.

FIGURE 2.5

The OLS regression line $\widehat{salary} = 963.191 + 18.501\, roe$ and the (unknown) population regression function.



Ordinary Least Squares (OLS)

- Minimize $Q(a, b) = \sum_{i=1}^{n} (Y_i - a - bX_i)^2$ w.r.t. a and b.
- Derivatives:

$$\frac{dQ(a,b)}{da} = -2\sum_{i=1}^{n}(Y_i - a - bX_i).$$

$$\frac{dQ(a,b)}{db} = -2\sum_{i=1}^{n}(Y_i - a - bX_i)X_i.$$

- First-order conditions:

$$0 = \sum_{i=1}^{n}\left(Y_i - \hat{\alpha} - \hat{\beta}X_i\right) = \sum_{i=1}^{n}\hat{U}_i.$$

$$0 = \sum_{i=1}^{n}\left(Y_i - \hat{\alpha} - \hat{\beta}X_i\right)X_i = \sum_{i=1}^{n}\hat{U}_iX_i.$$

- Method of moments = OLS

$$\hat{\beta} = \frac{\sum_{i=1}^{n}\left(X_i - \bar{X}\right)Y_i}{\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2} \text{ and } \hat{\alpha} = \bar{Y} - \hat{\beta}\bar{X}.$$