L2: Two-variable regression model

Feng Li
feng.li@cufe.edu.cn

School of Statistics and Mathematics
Central University of Finance and Economics

Revision: September 4, 2014

What we have learned last time...

- Population regression line
- Sample regression line
- The term u_i
- We wished to find $\hat{\beta}_1$ and $\hat{\beta}_2$ so that $\hat{u}_i$ can be minimal.

Today we are going to learn...

1. To find the best β_1 and β_2
2. The properties of ordinary least squares
3. The assumptions for the linear regression model
4. Standard errors of OLS
5. Determination of Goodness of fit

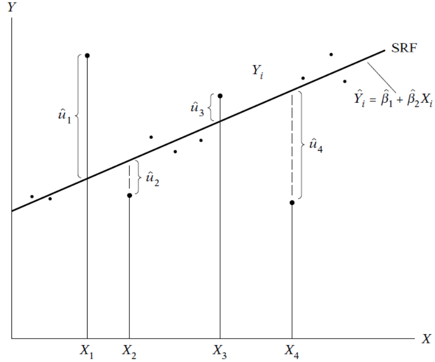
To find the best β_1 and β_2

↪ The problem

- We knew the population regression function is not easy to have.
- Instead we estimate it from the sample regression function, i.e.

$$Y_i = \hat{\beta}_1 + \hat{\beta}_2 X_i + \hat{u}_i$$

- We wish to have small $\hat{u}_i$ for $i = 1, 2, ..., n$
- It's difficult to have a **fair** solution: your regression line resulting some $\hat{u}_i$ are very small, but others are big, which is **unfair**.



To find the best β_1 and β_2
↪ Using the ordinary least squares method

- Recall that the difference between the population mean Y_i and the estimated conditional mean $\hat{Y}_i$

$$\begin{aligned}\hat{u}_i =& Y_i - \hat{Y}_i \\ =& Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_i\end{aligned}$$

- One possible solutions it to let $\sum_{i=1}^{n} \hat{u}_i^2$ to be a minimal so that every observation is considered. **Is this good and why not to minimize** $\sum_{i=1}^{n} u_i^2$**?**
- This yields to minimize

$$\begin{aligned}\sum_{i=1}^{n} \hat{u}_i^2 =& \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 \\ =& \sum_{i=1}^{n} (Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_i)^2\end{aligned}$$

To find the best β_1 and β_2

$\hookrightarrow$ Using the ordinary least squares method

- This is straightforward by applying differential calculations (details in **Appendix 3A**), i.e.

$$\frac{\partial \sum_{i=1}^{n} \hat{u}_i^2}{\partial \hat{\beta}_1} = -2 \sum_{i=1}^{n} (Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_i) = 0$$

$$\frac{\partial \sum_{i=1}^{n} u_i^2}{\partial \hat{\beta}_2} = -2 \sum_{i=1}^{n} (Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_i) X_i = 0$$

- Simplify these equations we have (**how** ?)

$$\sum_{i=1}^{n} Y_i = n\hat{\beta}_1 + \hat{\beta}_2 \sum_{i=1}^{n} X_i$$

$$\sum_{i=1}^{n} Y_i X_i = \hat{\beta}_1 \sum_{i=1}^{n} X_i + \hat{\beta}_2 \sum_{i=1}^{n} X_i^2$$

- Can you obtain $\hat{\beta}_1$ and $\hat{\beta}_2$ now?

To find the best β_1 and β_2

↪ Using the ordinary least squares method

- That is easy, from the first equation, we have

$$\hat{\beta}_1 = \frac{1}{n}\sum_{i=1}^{n} Y_i - \hat{\beta}_2 \frac{1}{n}\sum_{i=1}^{n} X_i = \bar{Y} - \hat{\beta}_2 \bar{X}$$

- Plug this result into the second equation in previous slides

$$\sum_{i=1}^{n} Y_i X_i = (\bar{Y} - \hat{\beta}_2 \bar{X}) \sum_{i=1}^{n} X_i + \hat{\beta}_2 \sum_{i=1}^{n} X_i^2$$

- Solve $\hat{\beta}_2$

$$\hat{\beta}_2 = \frac{\sum_{i=1}^{n} Y_i X_i - \bar{Y} \sum_{i=1}^{n} X_i}{\sum_{i=1}^{n} X_i^2 - \bar{X} \sum_{i=1}^{n} X_i} = \frac{n \sum_{i=1}^{n} Y_i X_i - n\bar{Y} \sum_{i=1}^{n} X_i}{n \sum_{i=1}^{n} X_i^2 - n\bar{X} \sum_{i=1}^{n} X_i} = \frac{n \sum_{i=1}^{n} Y_i X_i - \sum_{i=1}^{n} Y_i \sum_{i=1}^{n} X_i}{n \sum_{i=1}^{n} X_i^2 - (\sum_{i=1}^{n} X_i)^2}$$

$$= \frac{\sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n} (X_i - \bar{X})^2}.$$ **Verify this**!

To find the best β_1 and β_2
$\looparrowright$ Using the ordinary least squares method

- If we let $x_i = X_i - \bar{X}$ and $y_i = Y_i - \bar{Y}$, then the previous result can be written as

$$\hat{\beta}_2 = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2}$$

- Further more (**homework**!),

$$\hat{\beta}_2 = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2} = \frac{\sum_{i=1}^{n} x_i Y_i}{\sum_{i=1}^{n} X_i^2 - n\bar{X}^2} = \frac{\sum_{i=1}^{n} X_i y_i}{\sum_{i=1}^{n} X_i^2 - n\bar{X}^2}$$

and

$$\hat{\beta}_1 = \bar{Y} - \bar{X}\frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2}.$$

- Have you noticed that, the OLS does not depend on the assumption on u_i?

The properties of ordinary least squares (OLS)

- The regression line finally can be expressed as $\hat{Y}_i = \hat{\beta}_1 + \hat{\beta}_2 X_i$ where $\hat{\beta}_1$ and $\hat{\beta}_2$ are determined from previous slides.
- The regression line goes through the sample means of Y and X, i.e., $\bar{Y} = \hat{\beta}_1 + \hat{\beta}_2 \bar{X}$ holds. (**Why**?)
- The mean of our estimated Y, $(\frac{1}{n}\sum \hat{Y}_i)$ is equal to the mean of Y, $(\frac{1}{n}\sum Y_i)$, because (**verify this**!)

$$\begin{aligned}\frac{1}{n}\sum \hat{Y}_i =& \frac{1}{n}\sum(\hat{\beta}_1 + \hat{\beta}_2 X_i) = \frac{1}{n}\sum(\bar{Y} - \hat{\beta}_2\bar{X} + \hat{\beta}_2 X_i)\\ =& \frac{1}{n}\sum \bar{Y} - \hat{\beta}_2 \frac{1}{n}\sum(\bar{X} - X_i) = \frac{1}{n}\sum \bar{Y} = \frac{1}{n}\sum Y_i.\end{aligned}$$

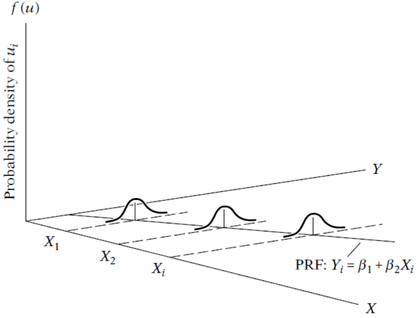
- The mean of the residuals $\hat{u}_i$ is zero which is directly verified by an equation in slide 6. (**which one**?)

The properties of ordinary least squares (OLS)

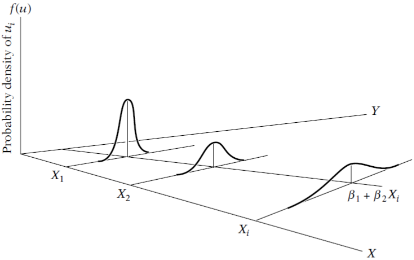
- It is easy to have $y_i = \hat{\beta}_2 x_i + \hat{u}_i$. Think about the equation in the first property and $\bar{Y} = \hat{\beta}_1 + \hat{\beta}_2 \bar{X} + \hat{u}_i$. (**How**?)
- The residuals $\hat{u}_i$ are uncorrelated with the predicted Y_i. Just show that $\sum \hat{u}_i \hat{y}_i = 0$. (**How**?)
- The residuals $\hat{u}_i$ are uncorrelated with X_i. Just show that $\sum \hat{u}_i X_i = 0$. (**How**?)

The assumptions for the linear regression model

1. The **linear** in linear regression model means **linear in the parameters**.
2. The regressor X is fixed (**not random**); X and the error term are independent, i.e., $cov(X_i, u_i) = 0$.
3. Zero mean value of disturbance u_i, i.e., $E(u_i|X_i) = 0$
4. Homoscedasticity (constant variance of u_i), i.e., $var(u_i) = E(u_i - E(u_i|X_i))^2 = E(u_i^2|X_i) = \sigma^2$.



Homoscedasticity.



Heteroscedasticity.

The assumptions for the linear regression model

1. No autocorrelation between the disturbances, i.e., $cov(u_i, u_j | X_i, X_j) = 0$ for $i \neq j$.
2. The number of observations n must be greater than the number of parameters.
3. The X values must not be all the same. (**What will happen if all X_i are the same**?)

Time to think about the assumptions again

1. Are these too realistic?
2. Can our data satisfy all of those assumptions?
3. What will happen if we break some of them?

Standard errors of OLS

1. **Given the Gaussian assumptions**, it is shown (**Appendix 3A**) that

$$var(\hat{\beta}_2) = \frac{\sigma^2}{\sum x_i^2} \Rightarrow se(\hat{\beta}_2) = \frac{\sigma}{\sqrt{\sum x_i^2}}$$

$$var(\hat{\beta}_1) = \frac{\sum X_i^2}{n \sum x_i^2}\sigma^2 \Rightarrow se(\hat{\beta}_1) = \sqrt{\frac{\sum X_i^2}{n \sum x_i^2}}\sigma$$

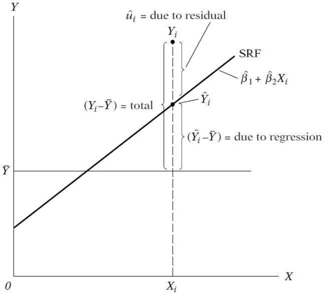
2. The variance of u_i, (σ^2) is estimated by $\hat{\sigma}^2 = \frac{\sum \hat{u}_i^2}{n-2}$, where $n-2$ is known as the **degrees of freedom**, and $\sum \hat{u}_i^2$ is called the **residual sum of squares (RSS)**. Further more $\hat{\sigma} = \sqrt{\frac{\sum \hat{u}_i^2}{n-2}}$ is called the **standard error (se)** of the regression.

3. The parameters $\hat{\beta}_1$ and $\hat{\beta}_2$ are dependent on each other, that is (**Section 3A.4**)

$$cov(\hat{\beta}_1, \hat{\beta}_2) = -\bar{X} \cdot var(\hat{\beta}_2) = -\bar{X}\frac{\sigma^2}{\sum x_i^2}$$

Determination of Goodness of fit

↛ The idea



Breakdown of the variation of Y_i into two components.

1. The **total sum of squares (TSS)** is the variation of Y about there sample mean, i.e.,

$$\begin{aligned} \sum y_i^2 &= \sum \hat{y}_i^2 + \sum \hat{u}_i^2 \quad (\textbf{verify this!}) \\ \sum (Y_i - \bar{Y})^2 &= \sum (\hat{Y}_i - \bar{Y})^2 + \sum \hat{u}_i^2 \\ \text{TSS} &= \text{ESS} + \text{RSS} \end{aligned}$$

2. A good model should be ESS → TSS, RSS → 0 (but this is not the sufficient condition).

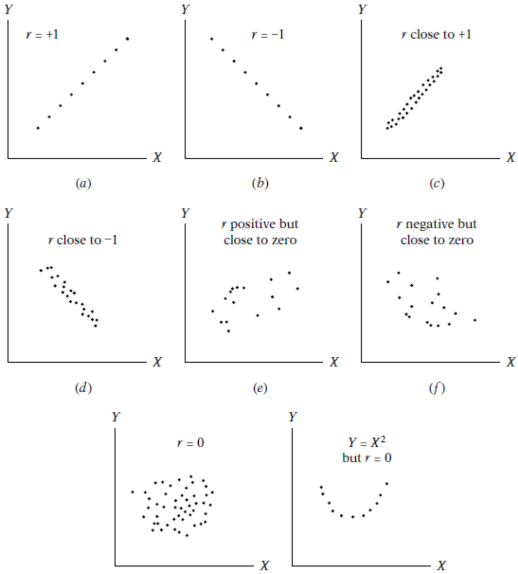
Determination of Goodness of fit
↛ The goodness of fit coefficient, r^2

1. Define the coefficient of determination of goodness of fit r^2 $(0 \leqslant r^2 \leqslant 1)$ as

$$r^2 = \frac{ESS}{TSS} = 1 - \frac{RSS}{TSS}$$

2. Properties of r^2
   1. r^2 can be linked with $\hat{\beta}_2$: $r^2 = \hat{\beta}_2^2 \frac{\sum x_i^2}{\sum y_i^2}$
   2. r^2 can be linked with sample variance of X and Y: $r^2 = \hat{\beta}_2^2 \frac{S_x^2}{S_y^2}$
3. The **coefficient of correlation** for X and Y is actually $r = \pm\sqrt{r^2}$
   1. Its traditional formula is $r = \frac{\sum x_i y_i}{\sqrt{\sum x_i^2 \sum y_i^2}}$
   2. correlation can be positive and negative, $-1 \leqslant r \leqslant 1$
   3. $r_{xy} = r_{yx}$.
   4. Correlation coefficients can only determine linear correlation.

The correlation coefficient, r
↛ A visual example



Take home questions

1. Verify the properties in **slides 9 and 10**.
2. Do the numerical example in the end of **Chapter 3** with Excel or a calculator.
3. Exercises (S1): **2.7, 2.13, 3.1, 3.6, 3.7, 3.14, 3.16, 3.19**
4. How do you appliy **maximum likelihood method** to find the coefficients?