Erasmus School of Economics

MOOC Econometrics

Lecture 1.3 on Simple Regression: Estimation

Philip Hans Franses

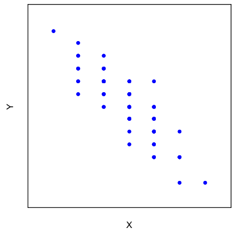
Erasmus University Rotterdam

Model and data

- Simple regression model: $y_i = \alpha + \beta x_i + \varepsilon_i$
- In econometrics, we do not know α and β (and ε_i)
  we do have observations on x_i and y_i
- Use observed data on x_i and y_i to find "optimal" values of a and b so that $y_i \approx a + bx_i$.
- The line $y = a + bx$ is called the regression line.

Data and regression line

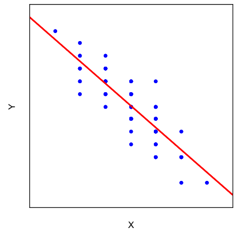
- Data: n pairs of observations (x_i, y_i) for $i = 1, 2, \ldots, n$



- Fitted line: $y_i = a + bx_i$
  - a: intercept
  - b: slope
- Residuals: $e_i = y_i - a - bx_i$
- Choose fitted line such that e_i are small.

Data and regression line

- Data: n pairs of observations (x_i, y_i) for $i = 1, 2, \ldots, n$



- Fitted line: $y_i = a + bx_i$
  - a: intercept
  - b: slope
- Residuals: $e_i = y_i - a - bx_i$
- Choose fitted line such that e_i are small.

Least squares

- Least squares criterion: find a and b by minimizing

  $S(a,b) = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - a - bx_i)^2$

- Get a and b by solving $\frac{\partial S}{\partial a} = 0$ and $\frac{\partial S}{\partial b} = 0$.

- We first analyze $\frac{\partial S}{\partial a} = 0$ (and later we consider $\frac{\partial S}{\partial b} = 0$):

  $0 = \frac{\partial S}{\partial a} = -2\sum_{i=1}^{n} (y_i - a - bx_i) = -2\sum_{i=1}^{n} e_i$

- Note: One residual follows from the other n-1 residuals:

  $e_n = -(e_1 + e_2 + ... + e_{n-1})$

Solving $\frac{\partial S}{\partial a} = 0$

- $0 = \frac{\partial S}{\partial a} = -2\sum_{i=1}^{n} (y_i - a - bx_i) = -2\sum_{i=1}^{n} y_i + 2na + 2b\sum_{i=1}^{n} x_i$

- Denote sample means by $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ and $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$,

  then above equation gives: $-2\bar{y} + 2a + 2b\bar{x} = 0$

- So: $a = \bar{y} - b\bar{x}$

Test question

Test

Suppose we apply least squares on de-meaned data, with dependent variable $y_i^* = y_i - \bar{y}$ and explanatory factor $x_i^* = x_i - \bar{x}$.
Which values do a and/or b take in this special case?

- Answer:

  Check that $\bar{y}^* = \frac{1}{n}\sum_{i=1}^{n} y_i - \frac{1}{n}\sum_{i=1}^{n} \bar{y} = \bar{y} - \bar{y} = 0$,

  and likewise $\bar{x}^* = 0$.

- So: $a = \bar{y}^* - b\bar{x}^* = 0 - b \times 0 = 0$.

- Later we will see that b is not affected by de-meaning.

Solving $\frac{\partial S}{\partial b} = 0$

- $S(a,b) = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - a - bx_i)^2$

- $0 = \frac{\partial S}{\partial b} = -2\sum_{i=1}^{n} x_i(y_i - a - bx_i) = -2\sum_{i=1}^{n} x_i e_i$

- Note: if $x_1 \neq 0$, then $e_1 = -(x_2e_2 + x_3e_3 + ... + x_ne_n)/x_1$

- $$\begin{aligned} 0 &= \sum_{i=1}^{n} x_i(y_i - a - bx_i) \\ &= \sum_{i=1}^{n} x_iy_i - a\sum_{i=1}^{n} x_i - b\sum_{i=1}^{n} x_i^2 \\ &= \sum_{i=1}^{n} x_iy_i - (\bar{y} - b\bar{x})\sum_{i=1}^{n} x_i - b\sum_{i=1}^{n} x_i^2 \\ &= \sum_{i=1}^{n} x_iy_i - \sum_{i=1}^{n} x_i\bar{y} + b\sum_{i=1}^{n} x_i\bar{x} - b\sum_{i=1}^{n} x_i^2 \\ &= \sum_{i=1}^{n} x_i(y_i - \bar{y}) - b\sum_{i=1}^{n} x_i(x_i - \bar{x}) \end{aligned}$$

- So: $b = \frac{\sum_{i=1}^{n} x_i(y_i - \bar{y})}{\sum_{i=1}^{n} x_i(x_i - \bar{x})}$

Solving $\frac{\partial S}{\partial b} = 0$

- $b = \frac{\sum_{i=1}^n x_i(y_i - \bar{y})}{\sum_{i=1}^n x_i(x_i - \bar{x})}$
- Use that $\sum_{i=1}^n (y_i - \bar{y}) = \sum_{i=1}^n y_i - n\bar{y} = 0$, and similarly
  $\sum_{i=1}^n (x_i - \bar{x}) = \sum_{i=1}^n x_i - n\bar{x} = 0$, hence
  $\bar{x}\sum_{i=1}^n (y_i - \bar{y}) = 0$ and $\bar{x}\sum_{i=1}^n (x_i - \bar{x}) = 0$
- We get:
  $$b = \frac{\sum_{i=1}^n x_i(y_i - \bar{y})}{\sum_{i=1}^n x_i(x_i - \bar{x})} = \boxed{\frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}}$$

Test

What value does b take if all observations of y_i are equal to 93?

- Answer: $b = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}$ with $y_i - \bar{y} = 93 - 93 = 0$.
  So: $b = 0$.

R-squared

- Data (x_i, y_i) give numerical values of a and b, with $a = \bar{y} - b\bar{x}$.
- Then $y_i = a + bx_i + e_i = \bar{y} - b\bar{x} + bx_i + e_i$, so
  $$y_i - \bar{y} = b(x_i - \bar{x}) + e_i \qquad (*)$$
- Deviation $y_i - \bar{y}$ partly explained by $x_i - \bar{x}$ (e_i is unexplained).
- Seen before: $\sum_{i=1}^n e_i = 0$ and $\sum_{i=1}^n x_i e_i = 0$, hence
  $\sum_{i=1}^n (x_i - \bar{x})e_i = \sum_{i=1}^n x_i e_i - \bar{x}\sum_{i=1}^n e_i = 0$.
- Squaring and Summing (SS) both sides of $(*)$ therefore gives:
  $$\sum_{i=1}^n (y_i - \bar{y})^2 = b^2 \sum_{i=1}^n (x_i - \bar{x})^2 + \sum_{i=1}^n e_i^2$$
  SSTotal = SSExplained + SSResidual
- $R^2 = \frac{\text{SSExplained}}{\text{SSTotal}} = \boxed{1 - \frac{\sum_{i=1}^n e_i^2}{\sum_{i=1}^n (y_i - \bar{y})^2}}$

Estimate of error variance

- $y_i = \alpha + \beta x_i + \varepsilon_i$ with $\varepsilon_i \sim NID(0, \sigma^2)$
- Unknown σ^2 is estimated from residuals $e_i = y_i - a - bx_i$.
- Residuals $e_i, i = 1, 2, \ldots, n$, have $n - 2$ free values (seen before).
- $s^2 = \frac{1}{n-2} \sum_{i=1}^n (e_i - \bar{e})^2$.
- Seen before: $\sum_{i=1}^n e_i = 0$, so $\bar{e} = 0$.
- Therefore: $\boxed{s^2 = \frac{1}{n-2} \sum_{i=1}^n e_i^2}$
- (see Building Blocks for case $n - 1$)

TRAINING EXERCISE 1.3

- Train yourself by making the training exercise (see the website).
- After making this exercise, check your answers by studying the webcast solution (also available on the website).