Erasmus
School of
Economics

MOOC Econometrics

Lecture 1.4 on Simple Regression: Evaluation

Philip Hans Franses

Erasmus University Rotterdam

Prediction interval

- Least squares: data $(x_i, y_i), i = 1, 2, \ldots, n \quad \rightarrow \quad a$ and b
- Regression line: $y = a + bx$

  Residuals: $e_i = y_i - a - bx_i$

  Residual standard deviation: $s = \sqrt{s^2} = \sqrt{\frac{1}{n-2}\sum_{i=1}^{n} e_i^2}$
- Question: Predict outcome of y_0 for new value of x_0.
- Actual value: $y_0 = \alpha + \beta x_0 + \varepsilon_0$

  Point prediction: $\hat{y}_0 = a + bx_0$

  Interval for ε_0: $(-ks, ks)$.
- Prediction interval for $y_0 : (\hat{y}_0 - ks, \hat{y}_0 + ks)$

Test question

- Prediction interval for $y_0 : (\hat{y}_0 - ks, \hat{y}_0 + ks)$.

Test

Which prediction interval has the highest confidence to contain y_0: for $k = 1$ or $k = 2$?

- Answer: $k = 2$, as the interval is wider.

Assumptions

- A1: Data Generating Process is $y_i = \alpha + \beta x_i + \varepsilon_i$.
- A2: The n observations of x_i are fixed numbers.
- A3: The n error terms ε_i are random, with $E(\varepsilon_i) = 0$.
- A4: The variance of the n errors is fixed, $E(\varepsilon_i^2) = \sigma^2$.
- A5: The errors are uncorrelated, $E(\varepsilon_i \varepsilon_j) = 0$ for all $i \neq j$.
- A6: α and β are unknown, but fixed for all n observations.
- A7: $\varepsilon_1, \ldots, \varepsilon_n$ are jointly normally distributed; with A3, A4, A5: $\varepsilon_i \sim NID(0, \sigma^2)$.

Statistical properties of b: preliminaries

- Least squares slope estimator: $b = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}$
- Derive properties of b from those of error terms ε_i (see A1-A7).
- We will show that $b = \beta + \sum_{i=1}^n c_i \varepsilon_i$, where $\boxed{c_i = \frac{x_i - \bar{x}}{\sum_{i=1}^n (x_i - \bar{x})^2}}$ are fixed numbers (see A2).
- Next slide shows steps needed for this result.

Derivation of the constants c_i

- $b = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}$ (A)
- A1: $y_i - \bar{y} = (\alpha + \beta x_i + \varepsilon_i) - (\alpha + \beta \bar{x} + \bar{\varepsilon})$
  $= \beta(x_i - \bar{x}) + (\varepsilon_i - \bar{\varepsilon})$ (B)
- $b \overset{(A,B)}{=} \beta + \frac{\sum_{i=1}^n (x_i - \bar{x})(\varepsilon_i - \bar{\varepsilon})}{\sum_{i=1}^n (x_i - \bar{x})^2}$ (C)
- $\sum_{i=1}^n (x_i - \bar{x})\bar{\varepsilon} = \bar{\varepsilon} \sum_{i=1}^n (x_i - \bar{x}) = \bar{\varepsilon}(\sum_{i=1}^n x_i - n\bar{x})$
  $= \bar{\varepsilon}(\sum_{i=1}^n x_i - \sum_{i=1}^n x_i) = 0$ (D)
- $b \overset{(C,D)}{=} \beta + \frac{\sum_{i=1}^n (x_i - \bar{x})\varepsilon_i}{\sum_{i=1}^n (x_i - \bar{x})^2} = \beta + \sum_{i=1}^n c_i \varepsilon_i$ with $c_i = \frac{x_i - \bar{x}}{\sum_{i=1}^n (x_i - \bar{x})^2}$

The mean of b: unbiased

- $b = \beta + \sum_{i=1}^n c_i \varepsilon_i$ with c_i fixed (due to A2).
- $E(b) = E(\beta) + \sum_{i=1}^n E(c_i \varepsilon_i)$.
- A6: β fixed, hence $E(\beta) = \beta$.
- c_i fixed, so $E(c_i \varepsilon_i) = c_i E(\varepsilon_i) = 0$ (due to A3).
- Hence: $E(b) = \beta + \sum_{i=1}^n 0 = \beta$.
- So b is unbiased estimator of slope parameter β.

Formula for $\sigma_b^2 = \text{var}(b)$

- $b = \beta + \sum_{i=1}^n c_i \varepsilon_i$ with $c_i = \frac{x_i - \bar{x}}{\sum_{i=1}^n (x_i - \bar{x})^2}$
- $$\begin{aligned} \sigma_b^2 &= E((b - E(b))^2) = E((b - \beta)^2) \\ &= E((\sum_{i=1}^n c_i \varepsilon_i)^2) = E(\sum_{i=1}^n \sum_{j=1}^n c_i c_j \varepsilon_i \varepsilon_j) \\ &\overset{(A2)}{=} \sum_{i=1}^n \sum_{j=1}^n c_i c_j E(\varepsilon_i \varepsilon_j) \\ &\overset{(A5)}{=} \sum_{i=1}^n c_i^2 E(\varepsilon_i^2) \\ &\overset{(A4)}{=} \sigma^2 \sum_{i=1}^n c_i^2 \\ &\overset{(*)}{=} \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \end{aligned}$$
- Notes: Step A5: $E(\varepsilon_i \varepsilon_j) = 0$ for all $i \neq j$

  Step (*): $\sum_{i=1}^n c_i^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{(\sum_{i=1}^n (x_i - \bar{x})^2)^2} = \frac{1}{\sum_{i=1}^n (x_i - \bar{x})^2}$

Test on slope parameter β

- Seen before: $b = \beta + \sum_{i=1}^n c_i \varepsilon_i$, $E(b) = \beta$, and $\sigma_b^2 = \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2}$
- A7: ε_i normal, then b normal (see Building Blocks)
- So: $b \sim N(\beta, \sigma_b^2)$ and $Z = \frac{b-\beta}{\sigma_b} \sim N(0,1)$
- Replace unknown σ^2 by $s^2 = \frac{1}{n-2} \sum_{i=1}^n e_i^2$, then $s_b^2 = \frac{s^2}{\sum_{i=1}^n (x_i - \bar{x})^2}$
- $t_b = \frac{b-\beta}{s_b} \sim t(n-2)$ (compare Building Blocks)
- t-test on H_0: $\beta = 0$ based on $t_b = \frac{b}{s_b}$

  Rule-of-thumb for large n: reject H_0 if $t_b < -2$ or $t_b > 2$.

Confidence and prediction intervals

- Approximate 95% confidence interval for b:

  $\frac{b-\beta}{s_b} \approx N(0,1)$ with interval $(-2, 2)$
- $-2 \leq \frac{b-\beta}{s_b} \leq 2 \quad \rightarrow \quad b - 2s_b \leq \beta \leq b + 2s_b$
- Approximate 95% prediction interval for y (see before):

  $a + bx - 2s \leq y \leq a + bx + 2s$
- Note: $-2s \leq \varepsilon \leq 2s$ is approximate 95% confidence interval for ε, uncertainty in a and b is neglected here.

Test question

Test

Let measurement scale of the dependent variable y be fixed, and compare two scales for the explanatory factor x: first x is measured in 10 units (recorded value of 5 corresponds to 50 units), and later in 100 units (recorded value of 5 corresponds to 500 units).
Which case gives the widest confidence interval for b?

- Answer: For U units, the value of x changes from $\frac{U}{10}$ to $\frac{U}{100}$, so x becomes 10 times as small.
- As $b = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}$, b is multiplied by $\frac{\frac{1}{10}}{\frac{1}{100}} = 10$.
- So b becomes 10 times as large, and same for confidence interval.

Seven assumptions

Assumption	Violation	Lecture
A1: $y_i = \alpha + \beta x_i + \varepsilon_i$	More than one x var.	2
	Choose among x-var.	3
	Binary y_i (0 or 1)	5
A2: x_i fixed	Random x_i	4
A3/A6: $E(\varepsilon_i) = 0$, α, β fixed	Parameter breaks	3
A4: Homoskedastic	Heteroskedastic errors	6
A5: Uncorrelated	Correlated errors	6
A7: ε normal	Often not needed	2-6

TRAINING EXERCISE 1.4

- 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).