Introductory Econometrics

Simple Linear Regression Model (III)

Yaohan Chen

School of Big Data and Statistics, Anhui University

Spring, 2025

Hypothesis Testing

- Hypothesis testing is frequently needed when we conduct statistical inference in the regression framework. It can be used to evaluate the validity of economic theory, to detect absence of structure, among many other things.
- **Example (Production Function)**
  Given the Cobb-Douglas production function $Y = AK^{\beta_2}L^{\beta_3}$, we want to test

$$\begin{aligned} H_0 &: \beta_2 + \beta_3 = 1 \quad \text{(constant return to scale)} \\ \text{versus } H_1 &: \beta_2 + \beta_3 < 1 \quad \text{(decreasing return to scale)} \end{aligned}$$

  To conduct the test, we can consider the following linear regression model

$$\ln(Y) = \beta_1 + \beta_2 \ln(K) + \beta_3 \ln(L) + \varepsilon$$

Hypothesis Testing

- **Example (Structural Change)**
  Let GDP_t stands for the gross domestic product of China at year t. We are interested in checking whether there is a structural change around year 1979. For this purpose, define a dummy variable $D_t = 1(t \geqslant 1979)$, and consider the following regression model

$$\ln(GDP_t) = (\beta_1 + \beta_3 D_t) + (\beta_2 + \beta_4 D_t)\, t + \varepsilon_t$$

  The null of interest is $H_0 : \beta_3 = \beta_4 = 0$ (no structural change) versus $H_1 : \beta_3 \neq 0$ or $\beta_4 \neq 0$ (having structural change)

Single Linear Restriction: t-test

- Under the classical assumptions, i.e. **Assumption 1 to Assumption 5**,

$$\hat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma^2}{\sum x_i^2}\right)$$

- We use the unbiased estimator of $\hat{\sigma}^2 = \frac{\sum e_i^2}{n-2}$ as the proxy for σ^2 and obtain the statistics as follows

$$t = \frac{\hat{\beta}_1 - \beta_1}{\sqrt{\hat{\sigma}^2 / \sum x_i^2}} = \frac{\hat{\beta}_1 - \beta_1}{S_{\hat{\beta}_1}} \sim t(n-2)$$

Single Linear Restriction: t-test

- For large sample, **Assumption 5** can be relaxed, and asymptotically

$$t = \frac{\hat{\beta}_1 - \beta_1}{\sqrt{\hat{\sigma}^2 / \sum x_i^2}} = \frac{\hat{\beta}_1 - \beta_1}{S_{\hat{\beta}_1}} \overset{a}{\sim} t(n-2).$$

- For β_0

$$t = \frac{\hat{\beta}_0 - \beta_0}{\sqrt{\hat{\sigma}^2 \sum X_i^2 / n \sum x_i^2}} = \frac{\hat{\beta}_0}{S_{\hat{\beta}_0}} \sim t(n-2).$$

- Theoretically, β_1 can take any reasonable values. Usually we are interested in checking whether $\beta_1 = 0$, i.e.

$$H_0 : \beta_1 = 0 \text{ versus } H_1 : \beta_1 \neq 0$$

Basic Steps for Hypothesis Testing

(1) Proposing hypothesis

$$H_0 : \beta_1 = 0 \text{ versus } H_1 : \beta_1 \neq 0.$$

(2) Constructing t-statistics under H_0 and evaluating t-statistics corresponding to the sample

$$t = \frac{\hat{\beta}_1}{S_{\hat{\beta}_1}}.$$

(3) For specific significance level α, comparing t with the threshold $t_{\alpha/2}(n-2)$, and

- Reject H_0, accept H_1, if $|t| > t_{\alpha/2}(n-2)$
- Reject H_1, accept H_0, if $|t| \leq t_{\alpha/2}(n-2)$

Hypothesis Testing: A Example

- For the example 2.2.1 in textbook, we calculate sample estimation of σ^2 as

$$\hat{\sigma}^2 = \frac{\sum e_i^2}{n-2} = \frac{\sum y_i^2 - \hat{\beta}_1^2 \sum x_i^2}{n-2} = \frac{3354955 - 0.67^2 \times 425000}{10-2} = 2734$$

and sample variance for $\hat{\beta}_1$ and $\hat{\beta}_0$

$$\begin{aligned} S_{\hat{\beta}_1} &= \sqrt{\hat{\sigma}^2 / \sum x_i^2} = \sqrt{2734/7425000} = \sqrt{0.0004} = 0.019 \\ S_{\hat{\beta}_0} &= \sqrt{\hat{\sigma}^2 \sum X_i^2 / n \sum x_i^2} \\ &= \sqrt{2734 \times 53650000/10 \times 7425000} = 44.45 \end{aligned}$$

Hypothesis Testing: An Example

- Calculating t-statistics corrresponding to $\hat{\beta}_0$ and $\hat{\beta}_1$ respectively

$$\begin{aligned} t_1 &= \hat{\beta}_1 / S_{\hat{\beta}_1} = 0.67/0.019 = 35.26 \\ t_0 &= \hat{\beta}_0 / S_{\hat{\beta}_0} = 142.40/44.45 = 3.20 \end{aligned}$$

- For a given $\alpha = 0.05$, we have $t_{\alpha/2}(10-2) = 2.306$. Since $|t_1| > t_{\alpha/2}(10-2)$ and $|t_0| > t_{\alpha/2}(10-2)$, $\hat{\beta}_1$ and $\hat{\beta}_0$ are both significant for the given significance level $\alpha = 5\%$.

Confidence Interval

- Hypothesis testing to some extent can help claim whether the parameters of interest take specific values in statistical sense. **BUT** no information about how well the estimator serves as a proxy for the parameter of interest.
- To check the "wellness", we need to construct an interval using sample information and a given probability that the constructed interval contains the underlying parameter of interest.
- The interval we construct is referred to as **confidence interval**.

Confidence Interval

- For a given probability that confidence interval cover the underlying parameter of interest, denoted by $1-\alpha$. Confidence interval as **random interval** with "radius" δ such that

$$P(\hat{\beta}-\delta \leq \beta \leq \hat{\beta}+\delta) = 1-\alpha.$$

where $1-\alpha$ is referred to as the **confidence coefficient**, α as the **level of significance**, and $\hat{\beta}-\delta$ and $\hat{\beta}+\delta$ as the **confidence limit**.

Confidence Interval for OLS Estimator in Simple Linear Regression

- For simple linear regression,

$$t = \frac{\hat{\beta}_i - \beta_i}{S_{\hat{\beta}_i}} \sim t(n-2)$$

  where $i = 0, 1$.

- For a given confidence coefficient $1 - \alpha$

$$P\left(-t_{\frac{\alpha}{2}} < t < t_{\frac{\alpha}{2}}\right) = 1 - \alpha$$

and accordingly

$$P\left(-t_{\frac{\alpha}{2}} < \frac{\hat{\beta}_i - \beta_i}{S_{\hat{\beta}_i}} < t_{\frac{\alpha}{2}}\right) = 1 - \alpha$$

$$P\left(\boxed{\hat{\beta}_i - t_{\frac{\alpha}{2}} \times S_{\hat{\beta}_i}} < \beta_i < \boxed{\hat{\beta}_i + t_{\frac{\alpha}{2}} \times S_{\hat{\beta}_i}}\right) = 1 - \alpha$$

Prediction

- Evaluation of sample regression function (SRF) at $X = X_0$ as the approximation of population regression function (PRF) evaluated at $X = X_0$, i.e. $\mathrm{E}(Y \mid X = X_0)$. To some extent, this can be regarded as "prediction" in generic sense.
- In simple linear regression framework,"prediction" can be expressed as

$$\hat{Y}_0 = \hat{\beta}_0 + \hat{\beta}_1 X_0.$$

- When model is correctly specified,

$$Y = \beta_0 + \beta_1 X + u,$$

and

$$\mathrm{E}(Y \mid X = X_0) = \beta_0 + \beta_1 X_0.$$

Prediction

- When model is correctly specified

$$\mathrm{E}\left(\hat{Y}_0\right) = \mathrm{E}\left(Y \mid X = X_0\right).$$

  That is, $\hat{Y}_0$ is unbiased estimator of $\mathrm{E}\left(Y \mid X = X_0\right)$.
- $\hat{Y}_0$ as the random variable, what is the corresponding distributions ?
- Conditional on sample $X_1, \ldots, X_n$ and given X_0

$$\hat{Y}_0 \sim N\left(\beta_0 + \beta_1 X_0, \sigma^2\left(\frac{1}{n} + \frac{\left(X_0 - \bar{X}\right)^2}{\sum x_i^2}\right)\right).$$

Prediction

- By replacing σ^2 with $\hat{\sigma}^2$, we can similarly construct the t-statics

$$t = \frac{\hat{Y}_0 - (\beta_0 + \beta_1 X_0)}{S_{\hat{Y}_0}} \sim t(n-2)$$

where

$$S_{\hat{Y}_0} = \sqrt{\hat{\sigma}^2 \left(\frac{1}{n} + \frac{\left(X_0 - \bar{X}\right)^2}{\sum x_i^2} \right)}$$

- Confidence interval for $\mathrm{E}\left(Y \mid X = X_0\right)$ is

$$\hat{Y}_0 - t_{\frac{\alpha}{2}} \times S_{\hat{Y}_0} < E\left(Y \mid X_0\right) < \hat{Y}_0 + t_{\frac{\alpha}{2}} \times S_{\hat{Y}_0}.$$

Prediction

- Since $Y_0 = \beta_0 + \beta_1 X_0 + u$,

$$Y_0 \sim N\left(\beta_0 + \beta_1 X_0, \sigma^2\right)$$

and accordingly

$$\hat{Y}_0 - Y_0 \sim N\left(0, \sigma^2\left(1 + \frac{1}{n} + \frac{\left(X_0 - \bar{X}\right)^2}{\sum x_i^2}\right)\right).$$

- By replacing σ^2 with $\hat{\sigma}^2$, we can similarly construct t-statistics

$$t = \frac{\hat{Y}_0 - Y_0}{S_{\hat{Y}_0 - Y_0}} \sim t(n-2)$$

Prediction

where

$$S_{\hat{Y}_0-Y_0} = \sqrt{\hat{\sigma}^2\left(1+\frac{1}{n}+\frac{\left(X_0-\bar{X}\right)^2}{\sum x_i^2}\right)}$$

and accordingly confidence interval for Y_0 is

$$\hat{Y}_0 - t_{\frac{\alpha}{2}} \times S_{\hat{Y}_0-Y_0} < Y_0 < \hat{Y}_0 + t_{\frac{\alpha}{2}} \times S_{\hat{Y}_0-Y_0}.$$

- Note that $S_{\hat{Y}_0-Y_0} > S_{\hat{Y}_0}$, hence for a given a given confidence coefficient, $\left[\hat{Y}_0 - t_{\frac{\alpha}{2}} \times S_{\hat{Y}_0-Y_0}, \hat{Y}_0 + t_{\frac{\alpha}{2}} \times S_{\hat{Y}_0-Y_0}\right]$ is wider than $\left[\hat{Y}_0 - t_{\frac{\alpha}{2}} \times S_{\hat{Y}_0}, \hat{Y}_0 + t_{\frac{\alpha}{2}} \times S_{\hat{Y}_0}\right]$.

Prediction: A Example

- OLS estimation corresponding to the income-expenditure example in textbook.

$$\begin{array}{rcccc} \hat{Y}_i & = & 2372.62 & + & 0.6232 X_i \\ & & (546.03) & & (0.018) \end{array}$$

- Based on the OLS estimation, we can predict the expenditure level for a given income level $X_0 = 20000$ CNY,

$$\hat{Y}_0 = 2372.62 + 0.6232 \times 20000 = 14836.6 \text{ CNY}$$

- Calculating $\bar{X} = 28166.1$ ans $\sum x_i^2 = 3943671436$.

Prediction: A Example

Confidence interval for $\mathrm{E}(Y \mid X = X_0)$:

$$14836.6 \pm 2.045 \times \sqrt{\frac{37041610}{31-2} \times \left(\frac{1}{31} + \frac{(20000 - 28166.1)^2}{3943671436}\right)}$$
$$= 14836.6 \pm 512.5$$

Confidence interval for Y_0:

$$14836.6 \pm 2.045 \times \sqrt{\frac{37041610}{31-2} \times \left(1 + \frac{1}{31} + \frac{(20000 - 28166.1)^2}{3943671436}\right)}$$
$$= 14836.6 \pm 2367.3$$

Summary

- Basic logic of hypothesis testing.
- Hypothesis testing for simple linear regression.
- Prediction and best linear predictor.
- Homework for Chapter 2: Textbook Page 52 - Page 53, 8,9,10.