Advanced Econometrics

L3 (II) Basic Concept in Hypothesis Testing and Partition Regression

Yonghui Zhang

School of Economics
Renmin University of China

November 2020

Basic Concept in Hypothesis Testing

Hypothesis testing (Why?)

- Hypothesis testing is one of a fundamental problems in Statistics/Econometrics.
- A hypothesis is (usually) an **assertion about the unknown population parameters** such as β_1 in $Y_i = \beta_0 + \beta_1 X_i + U_i$.
- Using the data, the econometrician has to determine whether an assertion is **true or false**.

Example (Phillips curve)

$$Unemployment_t = \beta_0 + \beta_1 Inflation_t + U_t.$$

In this example, we are interested in testing if $\beta_1 = 0$ (no Phillips curve) against $\beta_1 < 0$ (Phillips curve).

Null and alternative hypotheses

- For two competing hypotheses, which of the hypotheses is true *depends on the data (evidence)*.
- **Null hypothesis** $(\mathbb{H}_0)$: A hypothesis that is held to be **true**, **unless** the data provides a sufficient evidence against it.
- **Alternative hypothesis** $(\mathbb{H}_1)$: A hypothesis against which the null is tested. It is held to be true if the null is found false.
- Usually, the econometrician has to carry the "*burden of proof*", and **the case** that he or she is interested in is stated as $\mathbb{H}_1$.
- The econometrician has to prove that his assertion $(\mathbb{H}_1)$ is true by showing that the data rejects $\mathbb{H}_0$.
- The two hypotheses must be **disjoint**: it should be the case that either $\mathbb{H}_0$ is true or $\mathbb{H}_1$ but never together simultaneously.

Decision rule

- The econometrician has to choose between $\mathbb{H}_0$ and $\mathbb{H}_1$.
- The **decision rule** that leads the econometrician to **reject** or **not to reject** $\mathbb{H}_0$ is based on a **test statistic**, T_n, which is a function of the data $\{(Y_i, X_i), i = 1, ..., n\}$.
- **Remark: "not to reject"** $\stackrel{?}{=}$ **"accept"**
- Usually, one **rejects** $\mathbb{H}_0$ if the test statistic falls into a **critical region**.
- A **critical region** is constructed by taking into account the probability of **making a wrong decision**.

Two types of errors in hypothesis testing

- There are two types of errors (Neyman-Pearson):

		Truth	
		$\mathbb{H}_0$	$\mathbb{H}_1$
Decision	$\mathbb{H}_0$	$\surd$	Type II error
	$\mathbb{H}_1$	Type I error	$\surd$

- **Type I error** is the error of rejecting $\mathbb{H}_0$ when $\mathbb{H}_0$ is true.
- The probability of **Type I error** is denoted by α and called **significance level** or **size of a test**:

$$\mathtt{P}(\mathtt{Type\ I\ error}) = \mathtt{P}(\text{reject } \mathbb{H}_0 | \mathbb{H}_0 \text{ is true}) = \alpha.$$

- **Type II error** is the error of not rejecting $\mathbb{H}_0$ when $\mathbb{H}_1$ is true.
- **Power** of a test:

$$1 - \mathtt{P}(\mathtt{Type\ II\ error}) = 1 - \mathtt{P}(\text{not reject } \mathbb{H}_0 | \mathbb{H}_1 \text{ is true}).$$

Errors

- The decision rule depends on a test statistic T_n.
- The real line is split into two regions: **rejection region** (critical region, C_R) and **not rejected region** (C_R^c).
- **Decision:**
  - When $T_n \in C_R^c$, not reject $\mathbb{H}_0$ (and risk making a Type II error).
  - When $T_n \in C_R$, reject $\mathbb{H}_0$ (and risk making a Type I error).

Errors

- **Ideal case**: make two type errors as small as possible.
- Unfortunately, the probabilities of Type I and II errors are inversely related.
  - By decreasing the probability of Type I error a, one makes C_R smaller, which increases the probability of the Type II error ($C_R^c = \mathbb{R} \backslash C_R$ is larger). Thus it is **impossible** to make both errors arbitrary small.
- By convention, we control the Type I error. α is chosen to be a small number, for example, $\alpha = 0.01, 0.05$,or 0.10. (This is in agreement with the econometrician carrying the burden of proof).

Steps or Procedure

- The following are the steps of the hypothesis testing:
  1. Specify $\mathbb{H}_0$ and $\mathbb{H}_1$.
  2. Choose the significance level α.
  3. Define a decision rule (critical region).
  4. Perform the test using the data: given the data compute the test statistic and see if it falls into the critical region.
- The decision depends on the significance level α: larger values of α correspond to bigger critical regions (probability of Type I error is larger).
- It is easier to reject the null for larger values of α.

Test validity and power

- A test is valid if
  1. The test is valid, where the validity of a test means that it has correct size or $\mathtt{P}(\mathtt{Type\ I\ error}) = \alpha$:

$$\mathtt{P}\left(T_n \in C_R | \mathbb{H}_0\right) = \alpha$$

  2. The test has *power:* the test rejects $\mathbb{H}_0$ with probability that exceeds α:

$$\mathtt{P}\left(T_n \in C_R | \mathbb{H}_1\right) > \alpha.$$

  3. The test is *consistent* if $\mathtt{P}\left(T_n \in C_R | \mathbb{H}_1\right) \to 1$ as $n \to \infty$. (Frequentist view. Note that we do not have $\mathtt{P}\left(T_n \in C_R | \mathbb{H}_0\right) \to 0$.)
- In addition, for the choice of C_R, we want $\mathtt{P}\left(T_n \in C_R | \mathbb{H}_1\right)$ to be as large as possible. Note that $\mathtt{P}\left(T_n \in C_R | \mathbb{H}_1\right)$ depends on the true value β_1.

p-value

- **p-value**: Given the data, **the smallest significance level** at which the null can be rejected.
- R. A. Fisher (Calculated only under $\mathbb{H}_0$)
- p-value $\in [0, 1]$
- Depends on the data, p-value is a random variable
- Under $\mathbb{H}_0$, p-value is uniformly distributed on $(0, 1)$.
- p-value also depends on the choice of $C_R(1-\alpha)$.
- Calculate the test statistic T_n, the p-value is given by

$$p\text{-value} = \inf_{\alpha \in (0,1)} \{\alpha : T_n \in C_R(1-\alpha)\}$$

Controversy about p-value

- p-value:
$$P\left(T_n \text{ or "worse" than } T_n | \mathbb{H}_0\right)$$
- The "p-hacking" problem in scientific research: ruin the scientificity
- More appropriate target (?)
$$P\left(\mathbb{H}_0 | T_n\right) \text{ or } P\left(\mathbb{H}_1 | T_n\right)$$
Bayesian approach
- For your interest:
Scientific method: Statistical errors-*Nature*, 2015.
https://www.nature.com/news/scientific-method-statistical-errors-1.14700.
https://www.zhihu.com/question/23680352

Bayesian Approach

- **Bayes factors** (BFs, Kass and Raftery, 1995) can be an effective alternative to p-values for hypothesis testing (Marden, 2000).
- The conventional BF in favor of $\mathbb{H}_0$ against $\mathbb{H}_1$ can be defined as the ratio of two marginal likelihoods, that is,

$$BF_{01} = \frac{p(\mathbf{y}|\mathbb{H}_0)}{p(\mathbf{y}|\mathbb{H}_1)},$$

where $\mathbf{y}$ denotes the observed data, and the marginal likelihood $p(\mathbf{y}|\mathbb{H}_j)$ is obtained by integrating the corresponding likelihood over the parameter space, i.e.,

$$p(\mathbf{y}|\mathbb{H}_j) = \int p(\mathbf{y}|\boldsymbol{\theta}_j, \mathbb{H}_j) p(\boldsymbol{\theta}_j|\mathbb{H}_j) d\boldsymbol{\theta}_j$$

with $p(\boldsymbol{\theta}_j|\mathbb{H}_j)$ being the prior distribution of parameter $\boldsymbol{\theta}_j$ given $\mathbb{H}_j$ for $j = 0, 1$.

Bayesian Approach

- By Bayes' theorem, BFs give a direct answer to $p(\mathbb{H}_0|\mathbf{y})$, the probability that the null hypothesis is true conditional on the observed data, namely,

$$p(\mathbb{H}_0|\mathbf{y}) = \frac{p(\mathbf{y}|\mathbb{H}_0)p(\mathbb{H}_0)}{p(\mathbf{y}|\mathbb{H}_0)p(\mathbb{H}_0) + p(\mathbf{y}|\mathbb{H}_1)p(\mathbb{H}_1)} = \frac{BF_{01}p(\mathbb{H}_0)}{BF_{01}p(\mathbb{H}_0) + p(\mathbb{H}_1)},$$

  Similarly, $p(\mathbb{H}_1|\mathbf{y})$ is given by

$$p(\mathbb{H}_1|\mathbf{y}) = \frac{p(\mathbf{y}|\mathbb{H}_1)p(\mathbb{H}_1)}{p(\mathbf{y}|\mathbb{H}_0)p(\mathbb{H}_0) + p(\mathbf{y}|\mathbb{H}_1)p(\mathbb{H}_1)} = \frac{p(\mathbb{H}_1)}{BF_{01}p(\mathbb{H}_0) + p(\mathbb{H}_1)},$$

- Taking the ratio of these two quantities gives

$$\frac{p(\mathbb{H}_0|\mathbf{y})}{p(\mathbb{H}_1|\mathbf{y})} = \frac{p(\mathbb{H}_0)}{p(\mathbb{H}_1)} \times BF_{01}.$$

  Therefore, BFs transform the *prior odds* to the *posterior odds* of H_0 and H_1.

Bayesian Approach

- When one sets $p(\mathbb{H}_0) = p(\mathbb{H}_1) = 0.5$, the BF is identical to the *posterior odds*.
  - If $BF_{01} > 1$, then $p(\mathbb{H}_0|\mathbf{y}) > p(\mathbb{H}_1|\mathbf{y})$, suggesting an evidence to support $\mathbb{H}_0$ against $\mathbb{H}_1$.
  - However, if $BF_{01} < 1$, there is an evidence to support $\mathbb{H}_1$ against $\mathbb{H}_0$.
- When there are strong prior information about $\mathbb{H}_0$ vs $\mathbb{H}_1$ such as $p(\mathbb{H}_0) = 1 - p(\mathbb{H}_1) = 0.2$, use BF_{01} to calculate $p(\mathbb{H}_0|\mathbf{y})$ and $p(\mathbb{H}_1|\mathbf{y})$.

Bayesian Approach

- **Advantages** of BFs
  1. BFs treat $\mathbb{H}_0$ and $\mathbb{H}_1$ *symmetrically* and no type I and type II error.
  2. The calculation of BF is simple. BFs treat probabilities of alternative model specifications as parameters. Once the priors are specified, the posterior probabilities of competing models (i.e. $p(\mathbb{H}_0|\mathbf{y})$ and $p(\mathbb{H}_1|\mathbf{y})$) can be calculated and then these models can be compared easily.
  3. It is well documented that BFs have the **consistency property,** namely,

$$p(\mathbb{H}_0|\mathbf{y}) \to 1 \text{ under } \mathbb{H}_0, \text{ and } p(\mathbb{H}_1|\mathbf{y}) \to 1 \text{ under } \mathbb{H}_1,$$

  as sample size goes to infinity. *That is why BFs do not suffer from the "p-hacking" problem.*

Bayesian Approach

- **Drawbacks** about BFs
  1. BFs are not well-defined under *improper* priors.
  2. BFs are subject to the *Jeffreys-Lindley-Bartlett's paradox* when proper but vague priors are used. To be specific, BFs always favor H_0 when a very vague prior is used for the parameters under the null hypothesis.
  3. The calculation of BFs requires the evaluation of two marginal likelihood functions, $p(\mathbf{y}|\mathbb{H}_0)$ and $p(\mathbf{y}|\mathbb{H}_1)$. In many cases, marginal likelihood involves *high-dimensional integration* which is usually difficult. Although some interesting approaches have been proposed such as Chib (1995), Chib and Jeliazkov (2001), Friel and Pettitt (2008) and Li et al. (2019), BFs remain challenging in computation, especially in the big data environment.
- For more discussion about BFs, see Li and Yu (2012), Li et al. (2014), Li et al. (2015), Li et al. (2018), Li et al. (2019) and Li et al. (2020).

Partition Regressions

Partitioned regression

- Partitioned regression is useful in many places including the study of multicollinearity.
- Consider the model:

$$\begin{aligned} \mathbf{Y} &= \mathbf{X}\beta + \varepsilon = [\mathbf{X}_1, \mathbf{X}_2] \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix} + \varepsilon \\ &= \mathbf{X}_1\beta_1 + \mathbf{X}_2\beta_2 + \varepsilon \end{aligned}$$

  where $\dim(\mathbf{X}_1) = n \times k_1$ and $\dim(\mathbf{X}_2) = n \times k_2$ with $k_1 + k_2 = k$.
- The normal equation for OLS is

$$\mathbf{X}'\mathbf{X}\hat{\beta} = \mathbf{X}'\mathbf{Y}$$

Partitioned regression

- or equivalently

$$\begin{bmatrix} \mathbf{X}_1'\mathbf{X}_1 & \mathbf{X}_1'\mathbf{X}_2 \\ \mathbf{X}_2'\mathbf{X}_1 & \mathbf{X}_2'\mathbf{X}_2 \end{bmatrix} \begin{pmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{pmatrix} = \begin{pmatrix} \mathbf{X}_1'\mathbf{Y} \\ \mathbf{X}_2'\mathbf{Y} \end{pmatrix}$$

  and

$$\begin{aligned} \mathbf{X}_1'\mathbf{X}_1\hat{\beta}_1 + \mathbf{X}_1'\mathbf{X}_2\hat{\beta}_2 &= \mathbf{X}_1'\mathbf{Y} \quad (1) \\ \mathbf{X}_2'\mathbf{X}_1\hat{\beta}_1 + \mathbf{X}_2'\mathbf{X}_2\hat{\beta}_2 &= \mathbf{X}_2'\mathbf{Y} \quad (2) \end{aligned}$$

- Premultiplying both sides in the above first equation by $\mathbf{X}_2'\mathbf{X}_1 \left(\mathbf{X}_1'\mathbf{X}_1\right)^{-1}$ yields

$$\begin{aligned} \mathbf{X}_2'\mathbf{X}_1\hat{\beta}_1 + \mathbf{X}_2'\mathbf{X}_1 \left(\mathbf{X}_1'\mathbf{X}_1\right)^{-1}\mathbf{X}_1'\mathbf{X}_2\hat{\beta}_2 &= \mathbf{X}_2'\mathbf{X}_1 \left(\mathbf{X}_1'\mathbf{X}_1\right)^{-1}\mathbf{X}_1'\mathbf{Y} \text{ or} \\ \mathbf{X}_2'\mathbf{X}_1\hat{\beta}_1 + \mathbf{X}_2'\mathbf{P}_1\mathbf{X}_2\hat{\beta}_2 &= \mathbf{X}_2'\mathbf{P}_1\mathbf{Y} \quad (3) \end{aligned}$$

  where $\mathbf{P}_1 = \mathbf{X}_1 \left(\mathbf{X}_1'\mathbf{X}_1\right)^{-1}\mathbf{X}_1'$.

Partitioned regression

- Eq(2) minus Eq(3) and gives that

$$\begin{aligned} \mathbf{X}_2'\mathbf{M}_1\mathbf{X}_2\hat{\beta}_2 &= \mathbf{X}_2'\mathbf{M}_1\mathbf{Y} \text{ and} \\ \hat{\beta}_2 &= \left(\mathbf{X}_2'\mathbf{M}_1\mathbf{X}_2\right)^{-1}\mathbf{X}_2'\mathbf{M}_1\mathbf{Y} \end{aligned}$$

  where $\mathbf{M}_1 = I_n - \mathbf{P}_1 = I_n - \mathbf{X}_1\left(\mathbf{X}_1'\mathbf{X}_1\right)^{-1}\mathbf{X}_1'$.

- Similarly, we have

$$\hat{\beta}_1 = \left(\mathbf{X}_1'\mathbf{M}_2\mathbf{X}_1\right)^{-1}\mathbf{X}_1'\mathbf{M}_2\mathbf{Y}$$

  where $\mathbf{M}_2 = I_n - \mathbf{P}_2 = I_n - \mathbf{X}_2\left(\mathbf{X}_2'\mathbf{X}_2\right)^{-1}\mathbf{X}_2'$

PR

- Noting that $\mathbf{M}_1$ and $\mathbf{M}_2$ are projection matrices, we have alternative expressions for $\hat{\beta}_1$ and $\hat{\beta}_2$:

$$\begin{aligned}\hat{\beta}_1 &= \left(\mathbf{X}_1^{*\prime}\mathbf{X}_1^*\right)^{-1}\mathbf{X}_1^{*\prime}\mathbf{Y}_{(2)}^* = \left(\mathbf{X}_1^{*\prime}\mathbf{X}_1^*\right)^{-1}\mathbf{X}_1^{*\prime}\mathbf{Y} \\ \hat{\beta}_2 &= \left(\mathbf{X}_2^{*\prime}\mathbf{X}_2^*\right)^{-1}\mathbf{X}_2^{*\prime}\mathbf{Y}_{(1)}^* = \left(\mathbf{X}_2^{*\prime}\mathbf{X}_2^*\right)^{-1}\mathbf{X}_2^{*\prime}\mathbf{Y}\end{aligned}$$

  where $\mathbf{X}_1^* = \mathbf{M}_2\mathbf{X}_1$, $\mathbf{X}_2^* = \mathbf{M}_1\mathbf{X}_2$, $\mathbf{Y}_{(2)}^* = \mathbf{M}_2\mathbf{Y}$, and $\mathbf{Y}_{(1)}^* = \mathbf{M}_1\mathbf{Y}$.

- Partition regression for $\hat{\beta}_1$: (2 steps or 3 steps OLS regressions)
  - Step 1. Regress $\mathbf{Y}$ on $\mathbf{X}_2$ to get the residual $\mathbf{Y}_{(2)}^*$;
  - Step 2. Regress $\mathbf{X}_1$ on $\mathbf{X}_2$ to get the residual $\mathbf{X}_1^*$;
  - Step 3. Regress $\mathbf{Y}_{(2)}^*$ or $\mathbf{Y}$ on $\mathbf{X}_1^*$ to get $\hat{\beta}_1$

Partitioned regression

- Let $\hat{\beta} = \left(\hat{\beta}_1', \hat{\beta}_2'\right)'$ and $\hat{\boldsymbol{\varepsilon}} = \mathbf{Y} - \mathbf{X}\hat{\beta}$. Then

$$\mathbf{Y} = \mathbf{X}_1\hat{\beta}_1 + \mathbf{X}_2\hat{\beta}_2 + \hat{\boldsymbol{\varepsilon}}$$

- Note that $\mathbf{M}_2\hat{\boldsymbol{\varepsilon}} = \mathbf{M_2}\mathbf{M}\boldsymbol{\varepsilon} = \mathbf{M}\boldsymbol{\varepsilon} = \hat{\boldsymbol{\varepsilon}}$ and $\mathbf{M}_2 X_2 = 0$. We have

$$\mathbf{M}_2\mathbf{Y} = \mathbf{M}_2\mathbf{X}_1\hat{\beta}_1 + \mathbf{M}_2\mathbf{X}_2\hat{\beta}_2 + \mathbf{M}_2\hat{\boldsymbol{\varepsilon}} = \mathbf{M}_2\mathbf{X}_1\hat{\beta}_1 + \hat{\boldsymbol{\varepsilon}}$$

  or

$$\mathbf{Y}^*_{(2)} = \mathbf{X}^*_1\hat{\beta}_1 + \hat{\boldsymbol{\varepsilon}}$$

  which is called the *residual regression* in the literature of econometrics. Note that $\mathbf{X}^{*\prime}_1\hat{\boldsymbol{\varepsilon}} = \mathbf{0}$.

- Since $\hat{\beta}_1$ is precisely the OLS coefficient from a regression of $\mathbf{Y}^*_{(2)}$ on $\mathbf{X}^*_1$ this shows that the residual from this regression is $\hat{\boldsymbol{\varepsilon}}$, numerically the same residual as from the regression of $\mathbf{Y}$ on $\mathbf{X}_1$ and $\mathbf{X}_2$.

Partitioned regression

Example (Regression with detrended data)

Consider the following regression model in the time series framework:

$$Y_t = \alpha_0 + \theta t + \beta' x_t + \varepsilon_t$$

where t is the time trend. One run the above regression when one believes that there is a time trend in Y_t. In most cases, one is interested in the estimation of β but not α_0 and θ. So we can regress Y_t and x_t on constant and t to get the detrended data Y_t^* and x_t^*, respectively, which are defined as above. Then one obtain the estimator for β by regressing Y_t^* on x_t^*.

Partitioned regression

Theorem (Frisch-Waugh-Lovell Theorem)

The OLS estimator of β_1 and the OLS residual $\hat{\varepsilon}$ can be equivalently computed by regressing $\mathbf{Y}^*_{(2)}$ on $\mathbf{X}^*_1$ through the above Steps 1-3.

- In some context, the Frisch-Waugh-Lovell Theorem can be used to speed computation, but in most cases there is little computational advantage to using the residual regression.
- Even so, in practice, partitioned regression is frequently used to analyse some regressors of interests, where the control variables can be left in $\mathbf{X}_2$.
- When $\mathbf{X} = (\iota_n, \mathbf{X}_1)$, we write $\mathbf{M}_0 = \iota_n \iota_n' / n$, and then have

$$\begin{aligned}\hat{\beta}_1 &= \left(\mathbf{X}_1' \mathbf{M}_0 \mathbf{X}_1\right)^{-1} \mathbf{X}_1' \mathbf{M}_0 \mathbf{Y} \\ &= \left[\sum_{i=1}^{n} \left(x_{1i} - \bar{x}_1\right)\left(x_{1i} - \bar{x}_1\right)'\right]^{-1} \sum_{i=1}^{n} \left(x_{1i} - \bar{x}_1\right)\left(y_i - \bar{y}\right)\end{aligned}$$

Partitioned regression

- By the FWL theorem, we have

$$\hat{\beta}_1 = \left(\mathbf{X}_1'\mathbf{M}_2\mathbf{X}_1\right)^{-1}\mathbf{X}_1'\mathbf{M}_2\mathbf{Y}$$

- We can study the finite sample property of $\hat{\beta}_1$ without considering $\hat{\beta}_2$.
  - $E\left(\hat{\beta}_1|\mathbf{X}\right)$
  - $\texttt{Var}\left(\hat{\beta}_1|\mathbf{X}\right)$ and its estimators
  - Distribution of $\hat{\beta}_1$ under normal assumption? Test and CI.

Partitioned regression

- Properties about $\mathbf{P}_1 = \mathbf{X}_1 \left(\mathbf{X}_1'\mathbf{X}_1\right)^{-1} \mathbf{X}_1'$, $\mathbf{P}_2 = \mathbf{X}_2 \left(\mathbf{X}_2'\mathbf{X}_2\right)^{-1} \mathbf{X}_1'$ and $\mathbf{P} = \mathbf{X} \left(\mathbf{X}'\mathbf{X}\right)^{-1} \mathbf{X}'$ with $\mathbf{X} = \left(\mathbf{X}_1, X_2\right)$
  - $\mathbf{P}\mathbf{P}_l = \mathbf{P}_l$ for $l = 1, 2$
  - $\mathbf{M}\mathbf{M}_l = \mathbf{M}$ for $l = 1, 2$
- When $\mathbf{X}_1'\mathbf{X}_2 = 0$, we have

$$\begin{aligned} \mathbf{P}_1\mathbf{P}_2 &= 0 \\ \mathbf{P} &= \mathbf{P}_1 + \mathbf{P}_2 \\ \mathbf{M} &= I_n - \mathbf{P} = I_n - \mathbf{P}_1 - \mathbf{P}_2 = \left(I_n - \mathbf{P}_1\right)\left(I_n - \mathbf{P}_2\right) = \mathbf{M}_1\mathbf{M}_2 \end{aligned}$$