Topic 12 Header Slide

- Topic 12. Trinity of hypothesis testing for nested parametric models in linear regression

  Preliminary discussion: Advanced (i.e., MSc-level) considerations about hypothesis tests

  Main discussion: Wald, Likelihood ratio (LR), and Lagrange multiplier (LM) principles

---

Estimator	Notation	
$1.$	$\hat{\beta}_{OLS}$	
$2.$	$\hat{\beta}_{LAD}$	$= MLE$ with LDE errors
$3.$	$\hat{\beta}_{Lstar}$	needed for optimality results
$4.$	$\hat{\beta}_{GMM}$	$= OLS$
$5a.$	$\hat{\beta}_{IGLS}$	
$5b.$	$\hat{\beta}_{FGLS}$	
$6.$	$\hat{\beta}_{MLE}$	

Key concept 1: null vs. alternative vs. maintained hypotheses

- Null hypothesis, denoted H_0:
  This is the hypothesis that is being tested; the one that is under scrutiny.

- Alternative hypothesis, denoted H_1 (or sometimes H_A):
  This is the "fall-back" position in the event that the data are incompatible with the null.

- Maintained hypothesis
  (Careful: this is a term with which students are typically not familiar!)
  - This refers to the set of features of the underlying data-generating process that are believed to be true and simply just taken for granted.
  - In other words, the maintained hypothesis consists of all those **assumptions** that are not under question; at least not for the purpose of the given hypothesis test.
  - For example, if we were to conduct a standard test (e.g., as reported by default in Stata) for significance of an estimated regression coefficient, say $\hat{\beta}_{1,OLS}$, in the classical NLRM, then $A2linear$ *inter alia* would be part of the maintained hypothesis, whereas $H_0 : \beta_1 = 0$ would be the null, and $H_1 : \beta_1 \neq 0$ the alternative.
  - Notice that the maintained hypothesis allows either the null or the alternative to be true (i.e., it contains them as logical possibilities.)

Examples for key concept 1

- Let us look at a couple of examples:

Example 1: The null consists of a single linear restriction on a single parameter. Moreover, the null is simple (i.e., not composite) and so is the alternative. The maintained hypothesis is not specified explicitly.

$$\begin{aligned} H_0 &: \beta_5^{true} = 7 \\ H_1 &: \beta_5^{true} = 9 \end{aligned}$$

Example 2: The null consists of multiple linear restrictions on multiple parameters. The null is still simple but the alternative is now composite. Again, the maintained hypothesis is not specified explicitly.

$$\begin{aligned} H_0 : 3\beta_2^{true} + 9\beta_3^{true} &= 11 \\ \text{and } \beta_4^{true} &= 9 \end{aligned}$$

$$\begin{aligned} H_1 : 3\beta_2^{true} + 9\beta_3^{true} &> 11 \\ \text{and/or } \beta_4^{true} &= 10 \end{aligned}$$

Key concept 2: restricted ("R") vs. unrestricted ("U") models

- The " U" version of the model is the one corresponding to the maintained hypothesis, while completely disregarding both the null and the alternative (i.e., either of which could be taken to be true by the end of the test).
- The " R" version of the model is the one corresponding to the maintained hypothesis wherein the null is taken for granted (i.e., imposed to be true).
- Let us consider an example:

Example: Suppose the "U" model is $y_s = \beta_0 + \beta_1^{true} X_{s1} + \beta_2^{true} X_{s2} + \varepsilon_s^{true}$ for $s = 1, ..., S$; and suppose $H_0 : \beta_1 = \beta_2$. Then, the "R" model is obtained as

$$\begin{aligned} y_s &= \beta_0^{true} + \beta_1^{true} X_{s1} + \beta_2^{true} X_{s2} + \varepsilon_s^{true} \\ &= \beta_0^{true} + \beta_1^{true} X_{s1} + \beta_1^{true} X_{s2} + \varepsilon_s^{true} \\ &= \beta_0^{true} + \beta_1^{true} (X_{s1} + X_{s2}) + \varepsilon_s^{true}, \end{aligned}$$

which can be estimated via a regression of y_s on a constant and Z_s where $Z_s \equiv X_{s1} + X_{s2}$ for all s.

Key concept 3: nested vs. non-nested hypotheses

- It is worth appreciating the logical distinction between nested and non-nested hypotheses.
- This is because there exist optimality theorems (concerning hypothesis testing procedures) that only apply to nested sets of hypotheses.
- For a setup to be (parametrically) nested, it must be the case that the competing (sets of) hypotheses, H_0 vs. H_1, are special cases of the wider/over-arching maintained hypothesis through restrictions on parameters.
- To understand this, note that:
  - If H_0 is false, it must follow that we should fall back to H_1. That is, logically speaking, the scope of the alternative should be more general than (or at least as wide as) the restricted (or narrow) scope of the null.
  - The maintained hypothesis must be a valid model in its own right; and one that can be shown to logically encompass both possibilities (i.e., whether the true state of the world is H_0 or H_1).

Nested example for key concept 3

- Let us look at a couple of examples.

  Example (nested): Suppose the maintained hypothesis includes the assumption:

  $$A2linear : y = X_A\beta_A^{true} + X_B\beta_B^{true} + \varepsilon^{true}$$

  and the null is

  $$H_0 : \beta_B^{true} = 0.$$

  - This setup is clearly nested because the “R” version of the model is a logical parametric special case of the “U” version of the model.

  - Indeed, the “U” model includes

    $$A2linear : y = X_A\beta_A^{true} + X_B\beta_B^{true} + \varepsilon^{true}$$

    whereby β_A^{true} and β_B^{true} are permitted to take any value in the parameter space.

  - As regards the “R” model, it includes

    $$A2linear.H_0 : y = X_A\beta_A^{true} + \varepsilon^{true}$$

    and β_A^{true} can take any value in the parameter space.

Non-nested example for key concept 3

- In sharp contrast, the following two competing theories (i.e., versions of the model) are not logically ranked. They are instead, from a logical perspective, on a comparable footing.

  Example (non-nested):
  Suppose our competing theories both include the $A2linear$ assumption as per:

  $$A2linear.Theory1 : y = X_A \beta_A^{true} + \varepsilon^{true},$$

  where β_A^{true} can take any value in the parameter space, and

  $$A2linear.Theory2 : y = X_B \beta_B^{true} + \varepsilon^{true},$$

  where β_B^{true} can take any value in the parameter space.

Remark about key concept 3

- There do also exist non-nested testing approaches but these are typically not as philosophically/economically interesting.

- One example of a non-nested approach is to define a so-called "encompassing model" that is a convex combination of the two theories (i.e., the two competing versions of the model).

- That is, to introduce a parameter $\lambda \in [0, 1]$ such that the encompassing model is given by

$$\lambda A2linear.Theory1 + (1 - \lambda)A2linear.Theory2$$

- The intuition is that the parameter λ artificially nests the two competing theories since for $\lambda = 1$, the encompassing model reduces to $A2linear.Theory1$, and for $\lambda = 0$, the encompassing model reduces to $A2linear.Theory2$.

- The main difficulty with this approach is that for $\lambda \in (0, 1)$ (i.e., where the endpoints of the interval are excluded) the implied version of the obtained model may be philosophically/economically invalid.

- That is, economic theory may well be agnostic between the two competing theories; but this does not necessarily allow for "half of $Theory1$" and "half of $Theory2$" to jointly constitute a sensible way to explain the behaviour of economic agents!

Key concept 4: general knowledge from prior studies

This fourth and final key concept slide is ostensibly here to serve as a concept review but, in reality, it is a gentle reminder that you should ensure you are up-to-speed with basic concepts in statistical inference such as:

- Type I and Type II erors
- Significance level of a test
- Power of a test

We will refer to these terms either explicitly or implicitly at various points in the next few weeks. (If you do not know them well, you may struggle to understand, for instance, the next slide.)

Optimality theorems

The main message of this particular slide is that there exist three general procedures for devising hypothesis tests that are optimal and, in fact, equally so. These are the Wald, LR and LM procedures (which we will study in detail next). We refer to them collectively as the trinity of classical hypothesis testing procedures.

- **Theorem 1.** (Exact) Under quite restrictive conditions, the procedures that make up the trinity are equally powerful (for a given significance level) against any given alternative.

- **Theorem 2.** (Asymptotic) Under much less restrictive conditions, asymptotically as $S \to \infty$, the procedures that make up the trinity are equally powerful (for a given significance level) against a sequence of local alternatives (i.e, alternatives that get increasingly close to H_0 at rate $\sqrt{S}$).

  (The intuition behind the phrase "a sequence of local alternatives at rate $\sqrt{S}$" is as follows. Any sensible test statistic will be based on some CUAN estimator that converges to the truth at benchmark rate $\sqrt{S}$, meaning that its variance converges to 0 at rate S. Such a test statistic will trivially reject any null eventually as the sample size grows without bound so long as H_1 differs from H_0 by any fixed, non-zero amount. To overcome this obvious problem for asymptotic analysis, we make instead the alternative H_1 differ from H_0 by an ever-shrinking amount as S grows to infinity. Specifically, we make the fixed, non-zero amount between H_0 and H_1 decay as $\sqrt{S}$.)

Useful analytical framework to study the trinity

- The first message of this particular slide is that there exist three general procedures for devising hypothesis tests that are optimal and, in fact, equally so. These are the Wald, LR and LM procedures (which we will study in detail next). We refer to them collectively as the trinity of optimal hypothesis testing procedures.

- The second message is that a very useful schematic to analyse the trinity is the likelihood function of a parameter, say $\mathcal{L}(\theta)$, plotted on the vertical axis of a graph, against its argument, θ, plotted on the horizontal axis along with the following quantities identified on the plot:

$$\hat{\theta}^R_{best} \text{ and } \hat{\theta}^U_{best}$$

$$\log \mathcal{L}(\hat{\theta}^R_{best}) \text{ and } \log \mathcal{L}(\hat{\theta}^U_{best})$$

$$\left.\frac{\partial \log \mathcal{L}(\theta)}{\partial \theta}\right|_{\theta=\hat{\theta}^R_{best}} \text{ and } \left.\frac{\partial \log \mathcal{L}(\theta)}{\partial \theta}\right|_{\theta=\hat{\theta}^U_{best}}$$

The exercise alluded to above is helpful in making apparent the interpretations for each of the three testing principles.

Trinity I: Wald (pronounced 'Vald')

Key idea: requires best estimation of only the unrestricted model (i.e., $\hat{\theta}^U_{best}$)

- **Step 1.** Estimate efficiently the "U" model and obtain $\hat{\theta}^U_{best}$
- **Step 2.** Consider the restrictions on θ^{true} under H_0 and evaluate in-sample violations of those restrictions at $\hat{\theta}^U_{best}$ with the ultimate aim of checking whether these in-sample violations are (collectively) significantly different from zero.
  - Case A. Suppose all restrictions are linear. Then, we have:

Theoretical:	Estimated:
$H_0 : R\theta^{true} - q = 0_{r\times 1}$	$R\hat{\theta}^U_{best} - q$ to be compared against $0_{r\times 1}$
R is real, known $r \times p$ matrix with rank r	

  - Case B. Suppose at least one restriction is non-linear. Then, we have:

Theoretical:	Estimated:
$H_0 : g(\theta^{true}) = 0_{r\times 1}$	$g(\hat{\theta}^U_{best})$ to be compared against $0_{r\times 1}$
$\nabla g(.)$ is $r \times p$ matrix with rank r	

- **Step 3.** Use the distributional properties of $R\hat{\theta}^U_{best} - q$ for Case A and of $g(\hat{\theta}^U_{best})$ for Case B under H_0 (and the maintained hypothesis) to decide the outcome of the test.

Trinity II: LR

Key idea: requires best estimation of unrestricted and restricted models (i.e., $\hat{\theta}^U_{best}$ and $\hat{\theta}^R_{best}$)

- **Step 1.** Estimate efficiently the "U" model and save the achieved summary metric i.e., maximised $\mathcal{L}(\hat{\theta}^U_{best})$, or maximised $\log \mathcal{L}(\hat{\theta}^U_{best})$, or minimised $RSS(\hat{\theta}^U_{best})$
- **Step 2.** Estimate efficiently the "R" model and save the achieved summary metric i.e., maximised $\mathcal{L}(\hat{\theta}^R_{best})$, or maximised $\log \mathcal{L}(\hat{\theta}^R_{best})$, or minimised $RSS(\hat{\theta}^R_{best})$
- **Step 3.** Use the distributional properties of

  Option 1: $\mathcal{L}(\hat{\theta}^U_{best})/\mathcal{L}(\hat{\theta}^R_{best}) \geq 1$ and/or

  Option 2: $\log \mathcal{L}(\hat{\theta}^U_{best}) - \log \mathcal{L}(\hat{\theta}^R_{best}) \geq 0$ and/or

  Option 3: $RSS(\hat{\theta}^R_{best}) - RSS(\hat{\theta}^U_{best}) \geq 0$

  to decide the outcome of the test.

  Note that the reason these quantities have distributional properties is because they are random variables since they are functions of the (random) estimators.

Trinity III: LM

Key idea: requires efficient estimation of only the restricted model (i.e., $\hat{\theta}^R_{best}$)

- **Step 1.** Estimate efficiently the "R" model. Since, for so-called "regular" parametric problems, the Best CUAN method is MLE, we define $\hat{\theta}^R_{best}$ by the restricted optimisation problem given by

$$\hat{\theta}^R_{best} = \hat{\theta}^R_{restricted\ mle} = \arg\max_{\theta} \log \mathcal{L}(\theta) \text{ s.t. } R\theta = q \text{ when all restrictions are linear;}$$

$$\hat{\theta}^R_{best} = \hat{\theta}^R_{restricted\ mle} = \arg\max_{\theta} \log \mathcal{L}(\theta) \text{ s.t. } g(\theta) = 0 \text{ when } \exists \geq 1 \text{ non-linear restriction}$$

Note 1: The $\log \mathcal{L}(\theta)$ that needs to be maximised is of the unrestricted model. The restricted estimator $\hat{\theta}^R_{best}$ is obtained by maximising subject to the set of constraints in H_0.

Note 2: Under Gaussianity, the problem given by

$$\arg\max_{\theta} \log \mathcal{L}(\theta) \text{ subject to the restrictions in } H_0$$

is equivalent to the problem given by

$$\arg\min_{\theta} RSS(\theta) \text{ subject to the restrictions in } H_0$$

Trinity III: LM

- **Step 2 (LM perspective).** Consider the Lagrangean function that defines the restricted optimisation problem:

$$\begin{aligned} Lagrange(\theta, \lambda) &\equiv \log \mathcal{L}(\theta) + \lambda'(R\theta - q) \text{ when all restrictions are linear} \\ Lagrange(\theta, \lambda) &\equiv \log \mathcal{L}(\theta) + +\lambda' g(\theta) \text{ when there is } \geq 1 \text{ non-linear restriction} \end{aligned}$$

The parameter vector λ is a set of r "shadow values" or "Lagrange multipliers", which characterise how binding is the corresponding restriction.

The restricted optimum is defined by the simultaneous conditions given by:

$$\left(\hat{\theta}^R_{best}, \hat{\lambda}^R_{best} \right)' = \arg \max_\theta, \min_\lambda Lagrange(\theta, \lambda).$$

The further are the elements of λ from 0, the "more binding" are the constraints $R\theta = q$ or $g(\theta) = 0$ at the restricted solution $\hat{\theta}^R_{best}$. Conversely, if the elements of $\hat{\lambda}^R_{best}$ are close to 0 (collectively measured), that would be evidence in support of the restrictions in H_0.

Trinity III: LM

- **Step 2 (score perspective).** The LM test is (synonymously) also referred to as a "score test", where the word "score" refers to the first derivative of the log of the likelihood function. Thus, we now consider an alternative way to understand the LM test.

  Evidence in favour or against the restrictions in H_0 can be assessed according to the score criteria by considering the $p \times 1$ vector of first derivatives as per:

  $$\frac{\partial \log \mathcal{L}(\theta)}{\partial \theta} \quad \text{for the "U" model}$$

  and evaluating it at $\theta = \hat{\theta}^R_{best}$:

  $$\left.\frac{\partial \log \mathcal{L}(\theta)}{\partial \theta}\right|_{\theta = \hat{\theta}^R_{best}}$$

  The further away from 0 are the elements of this vector (collectively measured), the stronger would be the evidence against the restrictions in H_0.

  The intuition here is that we are investigating just how badly the first-order conditions for fully unfettered maximisation are being violated (collectively measured) in the presence of the given constraints.

Trinity III: LM

- **Step 3 (LM perspective).** Use the distributional properties of the vector of Lagrange multipliers (shadow values), $\hat{\lambda}^R_{best}$, to assess evidence for/against H_0. We have:

$$\left(\hat{\lambda}^R_{best}\right)' \left(\mathbb{V}(\hat{\lambda}^R_{best})\right)^{-1} \left(\hat{\lambda}^R_{best}\right) \xrightarrow{d} \chi^2_r \text{ as } S \to \infty.$$

- **Step 3 (score perspective).** Use the distributional properties of the score vector (containing first derivatives of the log of the likelihood function) evaluated at $\hat{\theta}^R_{best}$ to assess evidence for/against H_0. We have:

$$\left(\frac{\partial \log \mathcal{L}(\theta)}{\partial \theta}\Big|_{\theta=\hat{\theta}^R_{best}}\right)' \left(\mathbb{V}\left(\frac{\partial \log \mathcal{L}(\theta)}{\partial \theta}\Big|_{\theta=\hat{\theta}^R_{best}}\right)\right)^{-1} \left(\frac{\partial \log \mathcal{L}(\theta)}{\partial \theta}\Big|_{\theta=\hat{\theta}^R_{best}}\right) \xrightarrow{d} \chi^2_r \text{ as } S \to \infty.$$

- **Remark.** We end with a brief note to say that the LM/score test can sometimes be implemented via means of an auxiliary regression. This implementation is convenient for practical applications. We will delay exploring this “shortcut” method until the next topic (purely for expositional purposes).