Topic 8 Header Slide

- Topic 8. Three fundamental aims of statistical inference in linear/non-linear regression (all of which require distributions of $R\hat{\beta}_{method}$ (and also of $g(\hat{\beta}_{method})$)
  - Aim #1. Confidence interval estimation
  - Aim #2. Out-of-sample prediction
  - Aim #3. Hypothesis testing

Let us briefly recap our list of estimators so far:

Estimator	Notation
1.	$\hat{\beta}_{OLS}$
2.	$\hat{\beta}_{LAD}$
3.	$\hat{\beta}_{Lstar}$
4.	$\hat{\beta}_{GMM}$
5a.	$\hat{\beta}_{IGLS}$
5b.	$\hat{\beta}_{FGLS}$
6.	$\hat{\beta}_{MLE}$

Aim 1: interval estimation

- The construction of confidence intervals around true parameters is one of the key aims of statistical inference. It is called interval estimation (as opposed to point estimation).
- Consider a parameter θ^{true} and a "good" estimator $\hat{\theta}_{method}$ for it, obtained from a sample (y, X) of size S. (Note: recall what "good" estimator may connote.)
- Suppose that the distribution of $\hat{\theta}_{method}$ is given by:

$$\hat{\theta}_{method} \sim D(\theta^{true}, V_{method}).$$

- Using the distribution D, for some given $\alpha \in (0, 1)$, we can define the statement:

$$\Pr(\text{LeftCutoff} < \theta^{true} < \text{Rightcutoff}) = 1 - \alpha$$

by "pivoting" around the θ^{true} and $\hat{\theta}_{method}$ values so as to place θ^{true} in the middle of the left- and right-inequalities. For example, we can choose $\alpha = 0.05$, or $\alpha = 0.10$, so that the statement

$$[\text{LeftCutoff} < \theta^{true} < \text{Rightcutoff}]$$

will be true with 95% and 90% probability respectively.

A note on interpretation of confidence intervals

- Before the actual data (y, X) are plugged into the estimator $\hat{\theta}_{method}$, *LeftCutoff* and *RightCutoff* will be random/stochastic; so the probability statement on the previous slide makes sense.
- Once the left and right cutoffs are calculated based on the data, however, then *CalculatedLeftCutoff* and *CalculatedRightCutoff* will be fixed/non-stochastic numbers. This means that the probability of the statement of inequality will be either 1 or 0, depending on whether

$$[\text{CalculatedLeftCutoff} < \theta^{true} < \text{CalculatedRightcutoff}]$$

  happens to be true or false.
- For this reason, we then write:

$$CI(\text{CalculatedLeftCutoff} < \theta^{true} < \text{CalculatedRightcutoff}) = 1 - \alpha$$

  to mean that the constructed confidence interval has a confidence level of $1 - \alpha$, so that over hypothetically repeated experiments, the percentage of the constructed intervals encompassing θ^{true} will be $100(1 - \alpha)\%$, and the percentage of the constructed intervals not encompassing θ^{true} will be $100\alpha\%$.

Four examples

We will illustrate these definitions and ideas with four examples.

- We consider the NLRM with A4GM(iid) for which the BUE for $(k > 5)$-dimensional parameter vector β^{true} is $\hat{\beta}_{OLS}$; and for which $\hat{s}^2_{OLS} \equiv \frac{RSS_{OLS}}{S-k}$ is an unbiased estimator for the true σ^2_ε.
- The four examples are:

  Example 1:

  $$CI \text{ for } \beta_5^{true}$$

  Example 2:

  $$CI \text{ for } \beta_2^{true} - 3\beta_3^{true}$$

  Example 3:

  $$CI \text{ for } \sigma^2_\varepsilon$$

  Example 4:

  $$CI \text{ for } \beta_2^{true} \cdot \beta_4^{true} - \beta_5^{true}$$

Statistical inference in the NLRM

If we look at the NLRM with A4GM(iid), and we drop any results that are not relevant for our four examples, we see the following:

CASE I:	**NLRM with A4GM(iid)**
Result 1	[result suppressed]
Result 2	Conditionally on X, $\frac{(S-K)\hat{s}^2_{OLS}}{\sigma^2_\varepsilon} \sim \chi^2(S-k)$
Result 3	[result suppressed]
Result 4A	[result suppressed]
Result 4B	[result suppressed]
Result 5	Conditionally on X, $\tau \equiv \frac{\hat{\beta}_{j,OLS} - \beta_j^{true}}{\sqrt{\left[\hat{s}^2_{OLS}(X'X)^{-1}\right]_{jj}}} \sim t(S-k)$
Result 6	[result suppressed]
Result 7	For any continuous function $g(.) : R^p \to R^r$ with continuous first-derivative matrix $\left[\frac{\partial g(.)}{\partial \theta}\right]$, it follows (delta method) that: $g(\hat{\theta}_{method}) \overset{approx.}{\sim} N\left(g(\theta^{true}), \left[\frac{\partial g(.)}{\partial \theta}\right] V_{GM}(\hat{\theta}_{method}) \left[\frac{\partial g(.)}{\partial \theta}\right]'\right)$ for very large S.

Example 1:

$$CI \text{ for } \beta_5^{true}$$

- Result 5 states that: $\tau \equiv \frac{\hat{\beta}_{j,OLS} - \beta_j^{true}}{\sqrt{\left[\hat{s}^2_{OLS}(X'X)^{-1}\right]_{jj}}} \sim t(S-k)$. Hence:

$$\tau \equiv \frac{\hat{\beta}_{5,OLS} - \beta_5^{true}}{\sqrt{\left[\hat{s}^2_{OLS}(X'X)^{-1}\right]_{55}}} = \frac{\hat{\beta}_{5,OLS} - \beta_5^{true}}{SE(\hat{\beta}_{5,OLS})} \sim t(S-k)$$

- Denote by c_{Left} the lower cutoff value of the $t(S-k)$ distribution such that left tail probability is $\alpha/2$, and by c_{Right} the upper cutoff value of $t(S-k)$ distribution such that right tail probability is $\alpha/2$. (Note: Given the symmetry of the $t(.)$ distribution around 0, it follows $c_{Left} = -c_{Right}$.) The probability statement is then:

$$\begin{aligned} 1-\alpha &= \Pr\left(c_{Left} < \frac{\hat{\beta}_{5,OLS} - \beta_5^{true}}{SE(\hat{\beta}_{5,OLS})} < c_{Right}\right) \\ &= \Pr\left(\hat{\beta}_{5,OLS} - c_{Right} \cdot SE(\hat{\beta}_{5,OLS}) < \beta_5^{true} < \hat{\beta}_{5,OLS} + c_{Right} \cdot SE(\hat{\beta}_{5,OLS})\right) \end{aligned}$$

- Once we calculate $\hat{\beta}_{5,OLS}$ and $SE(\hat{\beta}_{5,OLS})$ based on the actual data, we then obtain the $100(1-\alpha)\%$ CI for β_5^{true}.

Example 2

Example 2:

$$CI \text{ for } \beta_2^{true} - 3\beta_3^{true}$$

- Define $\gamma^{true} \equiv \beta_2^{true} - 3\beta_3^{true} = r'\beta^{true}$, where $r' = (0, 1, -3, 0, \cdots, 0)$.
- We then follow the exact same steps as in Example 1 to obtain:

$$\begin{aligned} 1-\alpha &= \Pr\left(c_{Left} < \frac{\hat{\gamma}_{OLS} - \gamma^{true}}{SE(\hat{\gamma}_{OLS})} < c_{Right}\right) \\ &= \Pr\left(\hat{\gamma}_{OLS} - c_{Right} \cdot SE(\hat{\gamma}_{OLS}) < \gamma^{true} < \hat{\gamma}_{OLS} + c_{Right} \cdot SE(\hat{\gamma}_{OLS})\right) \end{aligned}$$

  where the BUE for $\gamma^{true} \equiv \beta_2^{true} - 3\beta_3^{true}$ is $\hat{\gamma}_{OLS} \equiv \hat{\beta}_{2OLS} - 3\hat{\beta}_{3OLS} = r'\hat{\beta}_{OLS}$ with estimated standard error

$$SE(\hat{\gamma}_{OLS}) = \sqrt{r'\hat{\beta}_{OLS}} = \sqrt{r'(\hat{s}^2_{OLS}(X'X)^{-1})r} = \sqrt{\hat{v}_{22} + 9\hat{v}_{33} - 6\hat{v}_{23}}$$

- Once we calculate $\hat{\gamma}_{OLS}$ and $SE(\hat{\gamma}_{OLS})$ based on the actual data, we then obtain the $100(1-\alpha)\%$ CI for $\gamma^{true} \equiv \beta_2^{true} - 3\beta_3^{true}$.

Example 3:

CI for σ_ε^2

- Result 2 states that: $\frac{(S-K)\hat{s}^2_{OLS}}{\sigma^2_\varepsilon} \sim \chi^2(S-k)$. Let c_{Left} be the $\alpha/2$ left tail probability of the $\chi^2(S-k)$ distribution and c_{Right} to be the $\alpha/2$ right probability of the $\chi^2(S-k)$ distribution. (Note: since the $\chi^2(\cdot)$ distribution has positive support and is not symmetric, $c_{Left} < c_{Right}$ and both will be positive.)

- We can then state that:

$$\begin{aligned} 1-\alpha &= \Pr\left(c_{Left} < \frac{(S-k)\hat{s}^2_{OLS}}{\sigma^2_\varepsilon} < c_{Right}\right) \\ &= \Pr\left(\frac{(S-k)\hat{s}^2_{OLS}}{c_{Right}} < \sigma^2_\varepsilon < \frac{(S-k)\hat{s}^2_{OLS}}{c_{Left}}\right) \end{aligned}$$

  and the $100(1-\alpha)\%$ CI follows once we plug in the calculated $\hat{s}^2_{OLS}$ based on actual data.

Example 4

Example 4:

$$CI \text{ for } \beta_2^{true} \cdot \beta_4^{true} - \beta_5^{true}$$

- Assuming very large S, the asymptotic/approximate Result 7 will be used, which states that:

$$g(\hat{\theta}_{method}) \overset{approx.}{\sim} N\left(g(\theta^{true}), \left[\frac{\partial g(.)}{\partial \theta}\right] V_{GM}(\hat{\theta}_{method}) \left[\frac{\partial g(.)}{\partial \theta}\right]'\right) \text{ for very large } S.$$

- The continuous function in this case is $g(\theta^{true}) = \beta_2^{true} \cdot \beta_4^{true} - \beta_5^{true}$ and, correspondingly, we have $\hat{g} = \hat{\beta}_{2,OLS} \cdot \hat{\beta}_{4,OLS} - \hat{\beta}_{5,OLS}$.
- Finally, the gradient is given by:

$$\frac{\partial g(\beta^{true})}{\partial \beta} = \left[\begin{array}{ccccccc} 0 & \beta_4^{true} & 0 & \beta_2^{true} & -1 & 0, \cdots, 0 \end{array}\right]$$

and the OLS-estimated gradient:

$$\frac{\partial g(\hat{\beta}_{OLS})}{\partial \beta} = \left[\begin{array}{ccccccc} 0 & \hat{\beta}_{4,OLS} & 0 & \hat{\beta}_{2,OLS} & -1 & 0, \cdots, 0 \end{array}\right]$$

- The left- and right-tail probabilities of the $N(0,1)$ distribution will allow us to define the (asymptotically valid/approximate) $1-\alpha$ probability statement.
- Plugging in the OLS estimates will give us the required $100(1-\alpha)\%$ CI.

Final Remarks (1 of 2)

- Note that for CIs for all linear relations, $R\beta^{true}$, we can rely on basic NLRM results; whereas when there exists even a single non-linear (continuous) relation among the true parameters, we need the delta method.
- Note that CIs will be more compact/narrow the "better" (i.e., the lower the variance of) the method of estimation employed.

Final Remarks (2 of 2)

- It is NOT very useful to attempt to consider multi-dimensional statements involving true parameter vectors. For example, suppose one considered two-dimensional functions of the underlying β^{true} vector:

$$\text{Linear} \ : \ R\beta^{true} = \begin{bmatrix} \beta_2^{true} - 3\beta_3^{true} \\ \beta_7^{true} \end{bmatrix} \text{ or}$$

$$\text{Non-linear} \ : \ g(\beta^{true}) = \begin{bmatrix} \beta_2^{true} - 3\beta_3^{true} \\ \beta_4^{true} \cdot \beta_5^{true} - \beta_6^{true} \end{bmatrix}$$

  Assuming the NLRM-A4GM(iid) setup, the BUE for $R\beta^{true}$ will be $R\hat{\beta}_{OLS}$ and the Best CUAN for $g(\beta^{true})$ will be $g(\hat{\beta}_{OLS})$. But now to define $1-\alpha$ probability statements for $R\beta^{true}$ and $R\hat{\beta}_{OLS}$ and for $g(\beta^{true})$ and $g(\hat{\beta}_{OLS})$ respectively, we will have to construct multi-dimensional regions (as opposed to intervals). It will be practically impossible to interpret such regions, and to somehow "pivot" around the true parameters in the centre of the inequalities, whatever that might mean.

Aim 2: out-of-sample (OofS) prediction

Using the basic available data $\{(y_s, x_s') : s = 1, ..., S\}$, we postulate the exact NLRM: $A1, A2linear, \geq A3Rmi, A4GM(= iid), A5Gaussian.$

Now suppose the regressor data is expanded by x_f' (single additional observation, $f = S+1$) or by X_f matrix of n_f new observations $S+1, \cdots, S+1+n_f$. In the first case, we would want to predict the single outcome y_f or its expectation; while in the second case, we would want to predict the n_f outcomes y_f or their expectation.

Some points for further elaboration by DrVH during the lecture:

- **NB1.** We will need an additional assumption of "model stability/no structural break". (Cases to elaborate: same β^{true} over OofS period; same σ_ε^2 over OofS period; A4GM(=iid) applies across both the original and OofS periods; and similarly $A5Gaussian.$)
- **NB2.** To explain why the prediction for ε_f^{true} is (the mean, median and mode) 0.
- **NB3.** To explain why second prediction will have a larger total forecast error, even though the same predictive functions — the two sub-aim predictions (see next slides) are identical.
- **NB4.** To explain that depending on the assumed underlying model, the "best" (LUE, CUAN, or whatever) estimation method may not be OLS.

OofS prediction of $\mathbb{E}(y_f|X, x_f)$ versus of y_f

Sub-aim 1:

- single new observation:
  construct prediction interval for scalar $\mathbb{E}(y_f|X, x_f) = x_f^{'}\beta^{true} = R\beta^{true}$ *using*

$$\widetilde{\mathbb{E}(y_f|X, x_f)} = x_f^{'}\hat{\beta}_{OLS} = R\hat{\beta}_{OLS}$$

  (or other preferred estimators as the case may be)

- n_f new observations:
  construct prediction intervals for vector $\mathbb{E}(y_f|X, X_f) = X_f\beta^{true} = R\beta^{true}$ *using*

$$\widetilde{\mathbb{E}(y_f|X, X_f)} = X_f\hat{\beta}_{OLS} = R\hat{\beta}_{OLS}$$

  (or other preferred estimators as the case may be)

OofS prediction of $\mathbb{E}(y_f|X, x_f)$ versus of y_f

Sub-aim 2:

- single new observation:
  construct prediction interval for scalar $y_f = x_f'\beta^{true} + \varepsilon_f^{true} = R\beta^{true} + \varepsilon_f^{true}$ *using*

$$\widetilde{y_f}|X, x_f = \widetilde{\mathbb{E}(y_f|X, x_f)} + \widetilde{\mathbb{E}(\varepsilon_f^{true}|X, x_f)} = x_f'\hat{\beta}_{OLS} + 0 = R\hat{\beta}_{OLS}$$

  (or other preferred estimators as the case may be)

- n_f new observations:
  construct prediction intervals for vector $y_f = X_f\beta^{true} + \varepsilon_f^{true} = R\beta^{true} + \varepsilon_f^{true}$ *using*

$$\widetilde{y_f}|X, X_f = \widetilde{E(y_f|X, X_f)} + \widetilde{E(\varepsilon_f^{true}|X, X_f)} = X_f\hat{\beta}_{OLS} + 0 = R\hat{\beta}_{OLS}$$

  (or other preferred estimators as the case may be)

Aim 3: hypothesis testing of true unknown parameters

The final aim of statistical inference is to construct critical rejection regions for hypotheses involving linear and non-linear restrictions on the true parameters, $R\beta^{true}$ and $g(\beta^{true})$ (and possibly also unknown parameters other than β^{true}) using similar functions of $\hat{\beta}_{OLS}$ (or other preferred estimators as the case may be).

Aim 3: hypothesis testing of true unknown parameters

- Note 1: To be as general as possible, we let the $p \times 1$ vector θ^{true} denote the vector of true unknown parameters. Depending on the context, we can have, for example,

$$\theta^{true} = \begin{pmatrix} \beta^{true} \\ \sigma_\varepsilon^2 \end{pmatrix}, \text{ or}$$

$$\theta^{true} = \begin{pmatrix} \beta^{true} \\ c^2 \\ \text{upper triangle of } \Omega \end{pmatrix}, \text{ or}$$

$$\theta^{true} = \begin{pmatrix} \beta^{true} \\ c^2 \\ \lambda \end{pmatrix} \text{ as in } \Omega(\lambda), \text{ or}$$

$$\theta^{true} = \begin{pmatrix} \beta^{true} \\ c^2 \\ \lambda_1 \\ \lambda_2 \end{pmatrix}, \text{ as in } \Omega(\lambda_1),$$

and λ_2 denotes parameters for non-Gaussian distributions, e.g., skewness, kurtosis, etc.

Aim 3: hypothesis testing of true unknown parameters

- Note 2: Greek words transliterated:
  $\Upsilon\pi o\theta\acute{\varepsilon}\sigma\varepsilon\iota\varsigma$ = hypotheses (plural) versus $\Upsilon\pi\acute{o}\theta\varepsilon\sigma\iota\varsigma$ = hypothesis (singular). So the singular iota "ι" is transliterated to "i", while the plural dipthong $\varepsilon\iota$ is transliterated to "e".

Topic 9 Header Slide

- Topic 9. Two fundamental concepts for statistical inference in linear/non-linear regression
  - Concept #1. "True" data-generating process (DGP)
  - Concept #2. Assumed minimal set of conditions for feasibility of estimation method

---

- There is a distinction between the true DGP and the assumed, possibly wrong, set of minimal assumptions necessary to define an estimator for practical use. We refer to this Minimum set/specification of Assumptions for Feasibility of Estimation using the acronym "MAFE".
- The true DGP (or just "DGP") comprises a specific choice under all five assumptions. While we are not always talking about experiments in the laboratory or Monte Carlo sense, we could think of the DGP as comprising a set of assumptions that would allow the researcher to notionally generate the given data from a computer simulation (at least in principle).
- Please note that the DGP need not coincide with the MAFE, and we would refer to such a discrepancy as a "model mis-specification". We will study this extensively in future weeks.
- We consider some examples on the next slide.

DGP versus MAFE – an illustration

Suppose the true DGP is given by $A1, A2linear, \geq A3Rmi, A4GMiid$, and $A5LDE$, where $A5LDE$ represents the Laplace Double Exponential distributional assumption.

Let us consider alternative minimal specifications defining particular estimators:

- Minimal assumptions for the OLS estimator,

$$\hat{\beta}_{OLS} = (X'X)^{-1}X'y,$$

  would be MAFE-OLS: $A1$.

- Minimal assumptions for the IGLS estimator assuming, say $\Omega_{AR(1)}$,

$$\hat{\beta}_{IGLS.AR(1)} = (X'\Omega_{AR(1)}^{-1}X)^{-1}X'\Omega_{AR(1)}^{-1}y$$

  would be MAFE-IGLS.$AR(1)$: $A1, A2linear, \geq A3Rmi, A4\Omega_{AR(1)}$.

- Let us also consider examples of minimal assumptions for the ML estimator (see next slide). Note that MLE is the most demanding estimation method on our list in terms of the intensity of underlying assumptions needed for it to be made operational in practice. Recall that MLE requires all five assumptions for it to even just be defined.

MAFE for MLE – examples

Continue to suppose that DGP: $A1, A2linear, \geq A3Rmi, A4GMiid$, and $A5LDE$.

- Consider MLE under the Laplace Double Exponential distribution. We have:

$$\hat{\beta}_{MLE.iid.LDE} = \hat{\beta}_{LAD},$$

  whereby the minimal set of assumptions would be
  MAFE-LAD: $A1, A2linear, \geq A3Rmi, A4GMiid$ and $A5LDE$.

- Consider MLE assuming the Logistic distribution. We have:
  MAFE-MLE.iid.Logistic: $A1, A2linear, \geq A3Rmi, A4GMiid$ and $A5Logistic$.

- Consider MLE assuming the Logistic distribution where we assume independence but allow for non-identical distributions across error terms. We have:
  MAFE-MLE.inid.Logistic: $A1, A2linear, \geq A3Rmi$, $A4\Omega diagonal$ and $A5Logistic$.

MAFE versus MAFE+

There is a distinction between being able to define an estimator and being able not only to define an estimator but also its SEV. We refer to the minimum set of conditions for the former as MAFE and for the latter as "MAFE+".

As an example, consider the OLS estimator. We have $\hat{\beta}_{OLS}$ defined as

$$\hat{\beta}_{OLS} = (X'X)^{-1}X'y$$

under MAFE: $A1$, but the sampling error defined, given $y = X\beta + \varepsilon$, as

$$SEV(\hat{\beta}_{OLS}) = (X'X)^{-1}X'\varepsilon$$

under MAFE+: $A1, A2linear$.

Sampling Error Vector (SEV)

Note that the SEV can be defined under false assumptions (i.e., under model mis-specification); but properties of the SEV must be analysed under the true DGP. Consider the example below.

- Suppose the true DGP is $A1,\ A2correct, \geq\ A3Rmi$, some $A4$ and $A5specific$, where $A2correct$ imposes $y = X\beta + Z\gamma + \varepsilon$.
- Say we mis-specify via MAFE+:$A1,\ A2wrong$, where $A2wrong$ assumes $y = X\beta + u$. Under mis-specification, we can define both the estimator,

$$\hat{\beta}_{OLS.wrong} = (X'X)^{-1}X'y$$

and its sampling error,

$$SEV(\hat{\beta}_{OLS.wrong}) = (X'X)^{-1}X'u.$$

- However, we must analyse its properties under $A2correct$ whereby $u = Z\gamma + \varepsilon$, so

$$\mathbb{E}(SEV(\hat{\beta}_{OLS.wrong})|X, Z) = (X'X)^{-1}X'Z\gamma,$$

under $A3Rmi$ which states that $\mathbb{E}(\varepsilon|X, Z) = 0$. (Note that $\mathbb{E}(u|X) \neq 0$.)

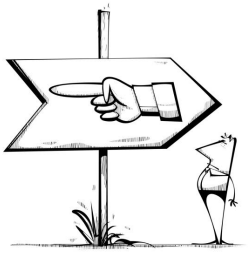
Review Quiz for Topics 8 and 9

Question 1. (What follows below is effectively a list of background vocabulary in the topic of statistical inference. These terms are likely to be familiar from your previous studies, but it would be good for you to refresh your knowledge of the precise meanings of each of these concepts.)

What is a hypothesis test? What is the power of a test? What is a pivotal function? What is a confidence interval? What is a prediction interval? What is the difference between an in-sample and an out-of-sample prediction? What is a forecast? (It's a good idea to also formulate examples of your own to understand these concepts. You can use the model: $y_i = \beta x_i + \varepsilon_i$ for $i = 1, ..., n$, where $A3F$, $A4GMiid$ and $A5Gaussian$ hold, if you wish.)

Question 2. What is the meaning of the acronyms DGP, MAFE and MAFE+? Explain with examples.

Question 3. This is not really a question but more a prompt for you to check that you understand Examples 1–4 (i.e., for confidence intervals) in the lecture slides above. (In this regard, you may need to review sampling distributions such as the t, χ^2 and F; and the delta method.)



Signposts 8 and 9

For statistical inference, you can see *Statistical Inference* by Casella & Berger.