Class 4: Asymptotics & Inference

MFE 402

Dan Yavorsky

Last Class

1. Explored the properties of the
   - the unconditional CEF Error Variance $\mathbb{E}[e^2] = \sigma^2$, a scalar
   - the conditional CEF Error Variance $\mathbb{E}[e^2|X] = \sigma^2(X)$, a scalar function of X
2. Derived the OLS Estimator Variance, $\mathbf{V}_{\hat{\beta}} = \text{Var}(\hat{\beta}|\mathbf{X})$,
   - it is a $k \times k$ matrix function of $\mathbf{X}$,
   - it is also, in general, a function of the conditional CEF Error Variance $\sigma^2(X)$, and
   - under homoskedasticity, it is a function of the unconditional CEF Error Variance σ^2
3. Introduced fitted values, residuals, and projection matrices P and M
4. Constructed feasible estimators
   - s^2
   - $\hat{\mathbf{V}}_{\hat{\beta}} = s^2(\mathbf{X}'\mathbf{X})^{-1}$
   - $\hat{\mathbf{V}}_{\hat{\beta}}^{\text{HC0}} = (\mathbf{X}'\mathbf{X})^{-1} \left(\sum_{i=1}^{n} X_i X_i' \hat{e}_i^2\right) (\mathbf{X}'\mathbf{X})^{-1}$
   - $\hat{\mathbf{V}}_{\hat{\beta}}^{\text{HC1}} = n/(n-k) \times \hat{\mathbf{V}}_{\hat{\beta}}^{\text{HC0}}$
5. Introduced Coefficient of Determination (R^2) and the "Adjusted" $\bar{R}^2$

Topics for Today

- Tools for Asymptotics
- Consistency of $\hat{\beta}$
- **Asymptotic Distribution of** $\hat{\beta}$
- Confidence Intervals
- Hypothesis Tests
- Assume Errors are Normally Distributed ($+$ Computation)
- Other linear hypothesis tests ($+$ Computation)

Tools for Asymptotics

Limits and Convergence generally

Definition:

A sequence a_n has the limit a if
for all $\delta > 0$ there is some $n_\delta < \infty$ such that for all $n \geq n_\delta$, we have $|a_n - a| \leq \delta$

Translation:

a_n has the limit a if the sequence gets closer and closer to a as n gets larger.

Notation:

- $a_n \to a$ as $n \to \infty$
- or $\lim_{n\to\infty} a_n = a$
- or a_n converges to a as n diverges (ie, increases without bound)

Convergence in Probability

We have one definition of convergence with a sequence of numbers. We have several types of convergence for random variables. The most common is convergence in probability.

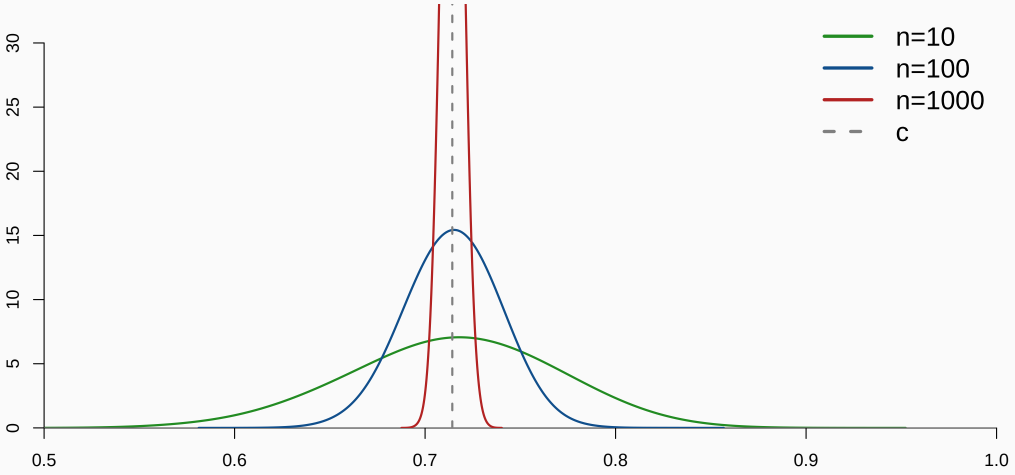
Definition:

A sequence of random variables $Z_n \in \mathbb{R}$ converges in probability to c as $n \to \infty$ if for all $\varepsilon > 0$ we have $\lim_{n\to\infty} \mathbb{P}(|Z_n - c| \leq \varepsilon) = 1$

Translation:

If a sequence of random variables Z_n has probability limit (or "plim") c, this means that the distribution of Z_n is concentrating around c and in the limit (if there were such a thing) the distribution would be a point-mass on c.

Example: Convergence in Probability



WLLN: Weak Law of Large Numbers

Definition:

The Weak Law of Large Numbers (LLN) states that the distribution of the sample average converges in probability to the expectation.

Suppose X_i are independent and with the same distribution F_X with finite mean ($\mathbb{E}[X] < \infty$). Then:

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow{p} \mathbb{E}[X]$$

Translation:

The distribution of the sample average concentrates on the expectation.

Continuity

Recall the **definition** of a continuous function:

A function $h(x)$ is continuous at $x = c$ if for all $\varepsilon > 0$ there is some $\delta > 0$ such that $|x - c| \leq \delta$ implies $|h(x) - h(c)| \leq \varepsilon$

Translation:

Small changes ($< \delta$) in the input x result in small changes ($< \varepsilon$) in the output $h(x)$

CMT

The Continuous Mapping Theorem (CMT) states that continuous functions are limit-preserving.

Definition:

If $Z_n \xrightarrow{p} c$ as $n \to \infty$ and $h(\cdot)$ is continuous at c, then $h(Z_n) \xrightarrow{p} h(c)$ as $n \to \infty$

Translation:

Convergence in probability is preserved by continuous mappings:
When a continuous function $h(\cdot)$ is applied to a random variable which converges in probability to c, the result is a new random variable which converges in probability to $h(c)$.

Consistency

Consistency of $\hat{\beta}$

(1) The OLS estimator $\hat{\beta}$ can be written as a continuous function of sample moments

$$\hat{\beta} = \left(\frac{1}{n}\sum_{i=1}^{n} X_i X_i'\right)^{-1} \left(\frac{1}{n}\sum_{i=1}^{n} X_i Y_i\right) = \hat{\mathbf{Q}}_{XX}^{-1}\hat{\mathbf{Q}}_{XY}$$

(2) The WLLN shows that sample moments converge in probability to "population" moments

$$\hat{\mathbf{Q}}_{XX} = n^{-1}\sum_{i=1}^{n} X_i X_i' \xrightarrow{p} \mathbb{E}[XX'] = \mathbf{Q}_{XX}$$

$$\hat{\mathbf{Q}}_{XY} = n^{-1}\sum_{i=1}^{n} X_i Y_i' \xrightarrow{p} \mathbb{E}[XY] = \mathbf{Q}_{XY}$$

(3) The CMT shows that continuous functions preserve convergence in probability

- Define $g(A, b) = A^{-1}b$ and take $A = \hat{\mathbf{Q}}_{XX}$ and $b = \hat{\mathbf{Q}}_{XY}$.
- Then $\hat{\beta} = g(\hat{\mathbf{Q}}_{XX}, \hat{\mathbf{Q}}_{XY}) = \hat{\mathbf{Q}}_{XX}^{-1}\hat{\mathbf{Q}}_{XY} \xrightarrow{p} \mathbf{Q}_{XX}^{-1}\mathbf{Q}_{XY} = \beta$

Comments on Consistency

1. Consistency is **different** from unbiasedness
   - Consistency is about the value on which a sampling distribution concentrates as $n \to \infty$
   - Unbiasedness is about the center of a sampling distribution for any given n
2. Consistency is a **good** (and basic) property for an estimator to possess
   - For almost any data distribution, there is a sufficiently-large sample size such that the estimator $\hat{\theta}$ will be arbitrarily close to the true value θ with high probability
   - Conversely: how much would you trust an inconsistent estimator that concentrates on the "wrong" answer as your sample size increases indefinitely?
3. However, there is **no practical guidance** on how large n has to be in order to believe our finite-sample results are approximately equal to their asymptotic counterparts
   - One option is to rely on simulation to assess the finite-sample properties of estimators

Consistency of Error Variance Estimators $\hat{\sigma}^2$ and s^2

Express the residuals as $\hat{e}_i = Y_i - X_i'\hat{\beta} = e_i - X_i'(\hat{\beta} - \beta)$. Then

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}\hat{e}_i^2 = \underbrace{\frac{1}{n}\sum_{i=1}^{n}e_i^2}_{\xrightarrow{p}\sigma^2} - 2\left(\frac{1}{n}\sum_{i=1}^{n}e_iX_i'\right)\underbrace{(\hat{\beta} - \beta)}_{\xrightarrow{p}0} + (\hat{\beta} - \beta)'\left(\frac{1}{n}\sum_{i=1}^{n}X_iX_i'\right)\underbrace{(\hat{\beta} - \beta)}_{\xrightarrow{p}0}$$

Because

1. $\hat{\beta} \xrightarrow{p} \beta$ as $n \to \infty$ and
2. by the WLLN $n^{-1}\sum_{i=1}^{n} e_i^2 \xrightarrow{p} \mathbb{E}[e_i^2] = \sigma^2$

It follows that $s^2 = \frac{n}{n-k}\hat{\sigma}^2 \xrightarrow{p} 1 \times \sigma^2 = \sigma^2$.

Thus we have that $\hat{\sigma}^2$ and s^2 are each consistent for σ^2.

Consistency of the Homoskedastic Covariance Matrix Estimator

Recall the formulas for $\text{var}(\hat{\beta})$ and its estimator under homoskedasticity:

$$\mathbf{V}^0_{\hat{\beta}} = \sigma^2(\mathbf{X}'\mathbf{X})^{-1} \quad \text{and} \quad \hat{\mathbf{V}}^0_{\hat{\beta}} = s^2(\mathbf{X}'\mathbf{X})^{-1}$$

We've already seen that $(\mathbf{X}'\mathbf{X}) = \hat{\mathbf{Q}}_{XX} \xrightarrow{p} \mathbf{Q}_{XX}$ and $s^2 \xrightarrow{p} \sigma^2$.

So by the CMT $\hat{\mathbf{V}}^0_{\hat{\beta}} \xrightarrow{p} \mathbf{V}^0_{\hat{\beta}}$ and thus the least squares covariance matrix estimator is consistent

Consistency of the Heteroskedastic Covariance Matrix Estimator

Recall the formulas for $\text{var}(\hat{\beta})$ and n times its estimator under heteroskedasticity:

$$\mathbf{V}_{\hat{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbb{E}[\mathbf{e}\mathbf{e}']\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \quad \text{and} \quad n\hat{\mathbf{V}}_{\hat{\beta}}^{\text{HC0}} = \underbrace{\left(\frac{1}{n}\sum_{i=1}^{n} X_i X_i'\right)^{-1}}_{\hat{\mathbf{Q}}_{XX}^{-1}} \underbrace{\left(\frac{1}{n}\sum_{i=1}^{n} X_i X_i' \hat{e}_i^2\right)}_{\hat{\boldsymbol{\Omega}}} \underbrace{\left(\frac{1}{n}\sum_{i=1}^{n} X_i X_i'\right)^{-1}}_{\hat{\mathbf{Q}}_{XX}^{-1}}$$

We've seen $\hat{\mathbf{Q}}_{XX}^{-1} \xrightarrow{p} \mathbf{Q}_{XX}$. Need to show $\hat{\boldsymbol{\Omega}} \xrightarrow{p} \boldsymbol{\Omega} = \mathbb{E}[XX'e^2]$.

Add and subtract $X_i X_i' e_i^2$ from each term in the middle sum to yield:

$$\hat{\boldsymbol{\Omega}} = \frac{1}{n}\sum_{i=1}^{n} X_i X_i' \hat{e}^2 = \underbrace{\frac{1}{n}\sum_{i=1}^{n} X_i X_i' e_i^2}_{\xrightarrow{p} \mathbb{E}[XX'e^2]} + \underbrace{\frac{1}{n}\sum_{i=1}^{n} X_i X_i' (\hat{e}_i^2 - e_i^2)}_{\xrightarrow{p} 0}$$

The first term converges by the WLLN. The second term converges to zero because

$$\hat{e}_i^2 - e_i^2 = \left(e_i - X_i'(\hat{\beta} - \beta)\right)^2 - e_i^2 = \underbrace{e_i^2 - e_i^2}_{=0} - 2e_i X_i' \underbrace{(\hat{\beta} - \beta)}_{\xrightarrow{p} 0} + \underbrace{\left(X_i'(\hat{\beta} - \beta)\right)^2}_{\xrightarrow{p} 0}$$

Asymptotic Distribution of $\hat{\beta}$

Convergence in Distribution

Another form of convergence for a random variable is Convergence in Distribution.

Definition:

Let Z_n be a sequence of random variables (or vectors) with distribution $G_n(u) = \mathbb{P}(Z_n \leq u)$.

Z_n converges in distribution to Z as $n \to \infty$ if for all u at which $G(u) = \mathbb{P}(Z \leq u)$ is continuous, $G_n(u) \to G(u)$ as $n \to \infty$.

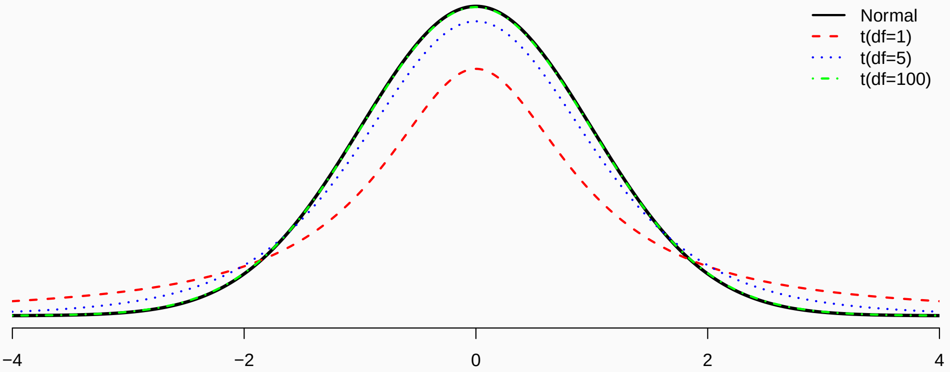
Translation:

Pointwise, the CDFs curves of Z_n converge to the CDF curve of Z

Notation:

We say Z_n converges in distribution to Z or that $Z_n \xrightarrow{d} Z$.

Example: Convergence in Distribution



CMT(d)

There is a Continuous Mapping Theorem for convergence in distribution

Definition:

Let Z_n be a sequence of random variables (or random vectors)

If $Z_n \xrightarrow{d} Z$ as $n \to \infty$ and $g(\cdot)$ is any* continuous function, then $g(Z_n) \xrightarrow{d} g(Z)$ as $n \to \infty$

Translation:

Convergence in distribution is preserved by continuous mappings.

Slutsky's Theorem

Common applications of the two CMTs get their own name: Slutsky's Theorem

If $Z_n \xrightarrow{d} Z$ and $c_n \xrightarrow{p} c$ as $n \to \infty$, then

$$Z_n + c_n \xrightarrow{d} Z + c$$

$$c_n Z_n \xrightarrow{d} cZ$$

$$Z_n / c_n \xrightarrow{d} Z/c \text{ if } c \neq 0$$

CLT: Central Limit Theorem

For scalar random variable Z:

If $Z_i \in \mathbb{R}$ are independent with the same distribution F_Z and $\bar{Z} = \frac{1}{n}\sum_{i=1}^{n} Z_i$ then as $n \to \infty$

$$\sqrt{n}(\bar{Z} - \mu) \xrightarrow{d} N(0, \sigma^2)$$

where $\mu = \mathbb{E}[Z]$ and $\sigma^2 = \mathbb{E}[(Z - \mu)^2]$

For random vector Z:

If $Z_i \in \mathbb{R}^k$ are independent with the same joint distribution F_Z and $\bar{Z} = \frac{1}{n}\sum_{i=1}^{n} Z_i$ then as $n \to \infty$

$$\sqrt{n}(\bar{Z} - \mu) \xrightarrow{d} N(0, \mathbf{V})$$

where $\mu = \mathbb{E}[Z]$ and $\mathbf{V} = \mathbb{E}[(Z - \mu)(Z - \mu)']$

For proofs of the CLT, see BHP Ch. 8

Asymptotic Normality of $X'e/\sqrt{n}$

Recall the conversion from matrix to summation notation for $n \times k$ matrix $\mathbf{X}$ and $n \times 1$ vector $\mathbf{e}$:

$$\frac{1}{n}\mathbf{X}'\mathbf{e} = \frac{1}{n}\sum_{i=1}^{n} X_i' e_i$$

Each row of the resulting $k \times 1$ vector is an average.

And for random vector X and random variable e, we know:

- $\mathbb{E}[Xe] = \mathbf{Q}_{Xe} = 0$
- $\text{var}(Xe) = \mathbb{E}[(Xe)(Xe)'] = \mathbb{E}[XX'e^2] = \mathbf{\Omega}$

Then by the CLT:

$$\sqrt{n}\left(\frac{1}{n}\mathbf{X}'\mathbf{e} - 0\right) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} X_i' e_i \xrightarrow{d} N(0, \mathbf{\Omega})$$

Asymptotic Normality of $\hat{\beta}$

Re-write $\hat{\beta}$ as $\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{X}\beta + \mathbf{e}) = \beta + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{e}$

$$\Rightarrow (\hat{\beta} - \beta) = \left(\frac{1}{n}\sum_{i=1}^{n} X_i X_i'\right)^{-1} \left(\frac{1}{n}\sum_{i=1}^{n} X_i' e_i\right) = \hat{\mathbf{Q}}_{XX}^{-1}\hat{\mathbf{Q}}_{Xe}$$

By the CLT, and CMT(d), and linearity of Normal Distributions

$$\sqrt{n}(\hat{\beta} - \beta) \xrightarrow{d} \mathbf{Q}_{XX}^{-1}\, N(0, \boldsymbol{\Omega}) = N(0, \mathbf{V}_{\beta})$$

where $\mathbf{V}_{\beta} = \mathbf{Q}_{XX}^{-1}\,\boldsymbol{\Omega}\,\mathbf{Q}_{XX}^{-1}$ and $\boldsymbol{\Omega} = \mathrm{Var}(Xe) = \mathbb{E}[XX'e^2]$

In words: we know the asymptotic distribution of $\hat{\beta}$! It is Normal with mean β and variance $\mathbf{V}_{\beta}$.

A Comment on Variance-Covariance Matrices

Don't confuse these:

- $\mathbf{V}_\beta = \mathbf{Q}_{XX}^{-1}\,\boldsymbol{\Omega}\,\mathbf{Q}_{XX}^{-1}$ is the variance of the asymptotic distribution of $\sqrt{n}(\hat\beta - \beta)$ and therefore is often referred to as the **asymptotic covariance matrix**. It is useful for asymptotic theory.
- $\mathbf{V}_{\hat\beta} = \text{var}(\hat\beta|\mathbf{X}) = (\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\mathbf{D}\mathbf{X})(\mathbf{X}'\mathbf{X})^{-1}$ is the **exact conditional variance** of $\hat\beta$. An estimate of it is useful for practical inference (ie, calculating standard errors).

They are, unsuprisingly, related:

$$n\mathbf{V}_{\hat\beta} = \left(\frac{1}{n}\mathbf{X}'\mathbf{X}\right)^{-1}\left(\frac{1}{n}\mathbf{X}'\mathbf{D}\mathbf{X}\right)\left(\frac{1}{n}\mathbf{X}'\mathbf{X}\right)^{-1} = \left(\frac{1}{n}\sum_{i=1}^{n} X_i X_i'\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n} X_i X_i'\mathbb{E}[e_i^2|X]\right)\left(\frac{1}{n}\sum_{i=1}^{n} X_i X_i'\right)^{-1}$$

$$\xrightarrow{p} (\mathbb{E}[XX'])^{-1}\,\mathbb{E}[XX'e^2]\,(\mathbb{E}[XX'])^{-1} = \mathbf{Q}_{XX}^{-1}\,\boldsymbol{\Omega}\,\mathbf{Q}_{XX}^{-1} = \mathbf{V}_\beta$$

Asymptotic Covariance Matrix under Homoskedasticity

Under homoskedasticity:

$$\boldsymbol{\Omega} = \mathbb{E}[XX'e^2] = \mathbb{E}[XX']\mathbb{E}[e^2] = \sigma^2 \mathbf{Q}_{XX}$$

And so the asymptotic covariance matrix becomes:

$$\mathbf{V}_\beta = \mathbf{Q}_{XX}^{-1}\, \boldsymbol{\Omega}\, \mathbf{Q}_{XX}^{-1} = \sigma^2 \mathbf{Q}_{XX}^{-1}$$

Confidence Intervals

Estimation Error and Pivotal Quantities

- When we estimate a parameter, we essentially guess it's value. This should be a "well-educated" guess (ie, based on the data). However, statistical estimation is made with error.
- Presenting just the estimate – no matter how good it is – **is not enough**. We should give an idea about the estimation error; that is, we should **quantify the uncertainty in the estimate**.
- We cannot directly estimate the estimation error, because if we could, we would use it to improve our estimate! But we can estimate its "order of magnitude."
- This is usually done by finding a pivot, which is a function of the data and the parameters, and whose distribution is **known** and does **not** depend on the parameters.

t-Statistic

Let $s(\hat{\beta}) = \sqrt{\left[\hat{\mathbf{V}}_{\hat{\beta}}\right]_{jj}}$ denote the standard error of one parameter β_j

Let $T(\beta_j^0)$ be the pivotal test statistic $T(\beta_j^0) = (\hat{\beta}_j - \beta_j^0)/s(\hat{\beta}_j)$. Then:

$$T(\beta_j^0) = \frac{\hat{\beta}_j - \beta_j^0}{\sqrt{\left[\hat{\mathbf{V}}_{\hat{\beta}}\right]_{jj}}} = \frac{\sqrt{n}(\hat{\beta}_j - \beta_j^0)}{\sqrt{n\left[\hat{\mathbf{V}}_{\hat{\beta}}\right]_{jj}}} \xrightarrow{d} \frac{N\left(0, \left[\mathbf{V}_\beta\right]_{jj}\right)}{\sqrt{\left[\mathbf{V}_\beta\right]_{jj}}} = N\left(0,\ \left[\mathbf{V}_\beta\right]_{jj}^{-1/2} \left[\mathbf{V}_\beta\right]_{jj} \left[\mathbf{V}_\beta\right]_{jj}^{-1/2}\right) = N(0,1)$$

The result follows from:

- The CLT: $\sqrt{n}(\hat{\beta} - \beta_j^0) \xrightarrow{d} N(0, \mathbf{V}_\beta)$
- The LLN: $n\hat{\mathbf{V}}_{\hat{\beta}} \xrightarrow{p} \mathbf{V}_\beta$ because $\hat{\mathbf{V}}_{\hat{\beta}} \xrightarrow{p} \mathbf{V}_{\hat{\beta}}$ and $n\mathbf{V}_{\hat{\beta}} \xrightarrow{p} \mathbf{V}_\beta$
- And Slutsky's Theorem

Note: BEH calls $T(\beta_0)$ the *t-statistic*, regardless of the asymptotic distribution of $T(\beta)$

Confidence Intervals for $\hat{\beta}$

We use $T(\beta_j)$ to construct a $(1 - \alpha) \times 100\%$ **confidence interval** for β_j.

The set of β_j values such that $T(\beta_j)$ is smaller (in absolute value) than c is:

$$\hat{C} = \{\beta_j : |T(\beta_j)| \leq c\} = \left\{\beta_j : -c \leq \frac{\hat{\beta}_j - \beta_j}{s(\hat{\beta}_j)} \leq c\right\} = \left[\hat{\beta}_j - c \times s(\hat{\beta}_j),\ \hat{\beta}_j + c \times s(\hat{\beta}_j)\right]$$

The coverage probability of this confidence interval is:

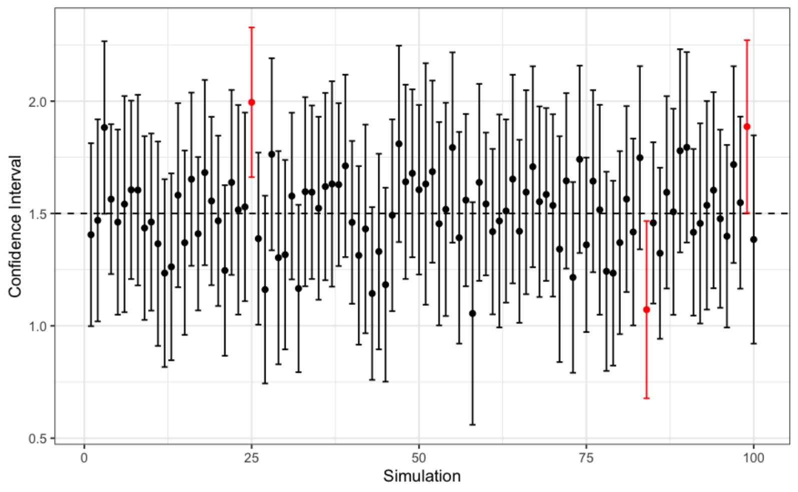
$$\Pr\left(\beta_j \in \hat{C}\right) = \Pr\left(|T(\beta_j)| \leq c\right) = \Pr\left(|Z| \leq c\right) = 1 - \alpha$$

The "standard" coverage probability for confidence intervals is 95%, leading to the choice of $c = 1.96$ because $\Pr(|Z| \leq 1.96) = 0.95$, which is often rounded to $c = 2$ resulting in the most commonly used confidence interval in econometric practice:

$$\hat{C} = \left[\hat{\beta} - 2s(\hat{\beta}),\ \hat{\beta} + 2s(\hat{\beta})\right]$$

Example Plot of CIs from Repeated Samples

A confidence interval is a function of the data and hence is random.



Hypothesis Tests

Null and Alternative Hypotheses, and Test Statistics

In many situations, the research questions yields one of two possible answers:

- The Null Hypothesis is the hypothesis to be tested: $H_0 : \theta = \theta_0$
- The Alternative Hypothesis is the set $\{\theta \in \Theta : \theta \neq \theta_0\}$

The statistician is required to choose the "correct" one, based on the data $\{\mathbf{y}, \mathbf{X}\}$

- We need to find a rule (a **test statistic** $T(\mathbf{y}, \mathbf{X})$) that maps the data to a decision
- Typically, $T(\mathbf{y}, \mathbf{X})$ is chosen such that it tends to be **small** when the Null Hypothesis is **true**; the **larger** it is, the stronger is the evidence **against** H_0 in favor of H_1

The Null Hypothesis is **rejected** if $T(\mathbf{y}, \mathbf{X})$ is larger than some critical value c

- When a test rejects H_0, it is common to say that the parameter estimate is **statistically significant**
- "Reject H_0" has has some **strength**: the evidence is inconsistent with H_0
- "Accept H_0" by comparison is **not** a strong statement – there is insufficient evidence to reject H_0. This does not mean H_0 is true! It might just mean we don't (yet) have enough evidence to reject it.

t-Test for $\hat{\beta}$

To test $H_0 : \beta = \beta_0$ against $H_1 : \beta \neq \beta_0$ we use the t-statistic $T(\beta_0) = \left(\hat{\beta} - \beta_0\right) / s(\hat{\beta})$

Note the default is statistical software is $H_0 : \beta = 0$

Since we know $T(\beta_0) \xrightarrow{d} Z = N(0, 1)$, we have that $\Pr(|T(\beta_0)| > c|H_0) \to \alpha$

Thus, to test the hypothesis using the test statistic, you:

- specify α
- find c satisfying $2(1 - \Phi(c)) = \alpha$ where Φ is the standard Normal CDF ($c = 1.96$ for $\alpha = 0.05$)
- reject H_0 if $|T(\beta_0)| > c$

p-Values for $\hat{\beta}$

Hypothesis tests dichotemize the outcomes: accept or reject H_0.

However, the **magnitude** of the test statistic T suggests a "degree of evidence" against H_0.

- Let $G(T)$ be the CDF of T, then the p-value is $p = 2\big(1 - G(|T|)\big)$
- There is a correspondence between the critical value c and the p-value p: instead of rejecting H_0 at the significance level α if $T > c$, we can equivalently reject H_0 if $p < \alpha$.

Thus, to test the hypothesis using the p-value, you:

- calculate $|T(\beta_0)| = |\hat{\beta} - \beta_0|/s(\hat{\beta})$
- find the p-value by calculating $p = 2 \times \Pr\big(T > |T(\beta_0)| \,|H_0\big)$
- reject H_0 if $p < \alpha$

Caution and Advice on Statistical Reporting

Always report the parameter estimates and their standard errors.

- Confidence intervals and p-values are good too.

Do **not** simply report something is (or is not) “statistically significant”.

Do **not** include asterisks without also reporting the p-values.

Scientific beliefs/conclusions are the result of a **body of evidence**

Caution and Advice on Interpretation

Hypothesis tests are **binary** statements about **precision**.

- "Significant" is just a term to abbreviate the exact meaning, which is: using the statistic T, the hypothesis H_0 can be rejected at the (asymptotic) $(1 - \alpha) \times 100\%$ level because we observed a value $(\hat{\beta} - \beta)$ from data of size n with error variability σ^2 that was unlikely to occur if H_0 were true.
- Do not confuse statistical significance with **economic/practical significance**. Many companies have data such that n is quite large, so that even small (practically irrelevant) estimates are statistically significantly different from zero.

A p-value is **not** the probability the Null or Alternative Hypotheses is true, **not** conclusive of any hypothesis, and **not** indicative of the size of the effect.

- It's the probability of receiving a test statistic as large (or larger) than the one calculated, under the assumption that the Null Hypothesis is true.

What if $e \sim N(0, \sigma^2)$

The Normal Regression Model

Suppose instead of the Linear CEF Model, we have the Normal ("Classical") CEF Model with an independent Normal error:

$$Y = X'\beta + e \quad \text{with} \quad e|X \sim N(0, \sigma^2)$$

Notice that homoskedasticity is "built in" to this model

The main estimators remain:

- Estimator $\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$ for estimand β
- Estimator $s^2 = \hat{\mathbf{e}}'\hat{\mathbf{e}}/(n - k)$ for estimand σ^2

Distribution of $\hat{\beta}$

Distribution of the error vector:

- The normality assumption $e|X \sim N(0, \sigma^2)$ combined with the independence of observations has a distributional implication: the error vector is distributed multivariate normal

$$\mathbf{e}|\mathbf{X} \sim N(0, I_n\sigma^2)$$

Distribution of the OLS estimator:

- The OLS estimator satisfies $\hat{\beta} - \beta = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{e}$ which is a linear (affine) function of $\mathbf{e}$
- Since affine functions of Normals are also Normal, we have:

$$\begin{aligned}\hat{\beta} - \beta|\mathbf{X} &\sim (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\, N(0, I_n\sigma^2) \\ &\sim N(0, \sigma^2(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}) \\ &= N(0, \sigma^2(\mathbf{X}'\mathbf{X})^{-1})\end{aligned}$$

- Or that $\hat{\beta} \sim N(\beta, \sigma^2(\mathbf{X}'\mathbf{X})^{-1})$. Notice that this is **exact**. No asymptotics needed.

Distribution of Residuals

Recall that $\hat{\mathbf{e}} = \mathbf{Me}$ where $\mathbf{M} = I_n - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$.

- Thus $\hat{\mathbf{e}}$ is linear in $\mathbf{e}$
- And so $\hat{\mathbf{e}} = \mathbf{Me}|\mathbf{X} \sim N(0, \sigma^2\mathbf{MM}) = N(0, \sigma^2\mathbf{M})$

Additionally, since $\hat{\mathbf{e}} = \mathbf{Me}$ and $\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$, we have that $\hat{\mathbf{e}}$ and $\hat{\beta}$ are independent because:

- They are orthogonal: $\hat{\mathbf{e}}'\hat{\beta} = \mathbf{e}'\mathbf{MX}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = \mathbf{0}$ because $\mathbf{MX} = \mathbf{0}$
- Normal random variables are independent IFF they are uncorrelated (i.e., orthogonal)

Thus, $(n - k)s^2/\sigma^2 = n\hat{\mathbf{e}}'\hat{\mathbf{e}}/\sigma^2 \sim \chi^2_{n-k}$

t-Statistic

Consider the t-statistic for β_j. It has all the ingredients for an **exact** t-distribution.

$$
\begin{aligned}
T &= \frac{\hat{\beta}_j - \beta_j}{s(\hat{\beta}_j)} \\
&= \frac{\hat{\beta}_j - \beta_j}{\sqrt{s^2[(\mathbf{X}'\mathbf{X})^{-1}]_{jj}}} \times \frac{\sqrt{\sigma^2(n-k)}}{\sqrt{\sigma^2(n-k)}} \\
&= \frac{\hat{\beta}_j - \beta_j}{\sqrt{\sigma^2[(\mathbf{X}'\mathbf{X})^{-1}]_{jj}}} \Bigg/ \sqrt{\frac{(n-k)s^2}{\sigma^2}/(n-k)} \\
&\sim \frac{N(0,1)}{\sqrt{\chi^2_{n-k}/(n-k)}} \\
&\sim t_{n-k}
\end{aligned}
$$

Confidence Interval and Hypothesis Tests

Confidence intervals and hypothesis tests both leverage the test statistic T.

- In the Linear CEF Model, T had an **asymptotically normal** distribution
- In the Normal ("Classical") CEF Model, T has an **exact t** distribution

With either statistic:

- Confidence intervals are $\hat{C} = [\hat{\beta} - c \times s(\hat{\beta}),\ \hat{\beta} + c \times s(\hat{\beta})]$
- Tests for $H_0 : \beta = \beta_0$ use $P(|T| > c|H_0) = 2(1 - F(c))$ where
  - F is the CDF of the t_{n-k} distribution (for the Normal CEF Model) and
  - F is the (asymptotic) CDF of standard normal distribution (for the Linear CEF Model)
- Select c so that this probability equals a pre-selected value α: $F(c) = 1 - \alpha/2$

In practice: many researchers do not assume that errors are normally distributed, but nevertheless use the t distribution for inference, since it leads to slightly more conservative (ie, larger) confidence intervals and p-values

Computation

Computation in R: `lm()`

```r
dat <- read.table("support/cps09mar.txt")
exper <- dat[,1] - dat[,4] - 6
lwage <- log( dat[,5]/(dat[,6]*dat[,7]) )
sam <- dat[,11]==4 & dat[,12]==7 & dat[,2]==0
```

```r
out <- lm(lwage[sam] ~ exper[sam])
summary(out)
```

```
Call:
lm(formula = lwage[sam] ~ exper[sam])

Residuals:
    Min      1Q  Median      3Q     Max
-2.3583 -0.4215  0.0042  0.4718  2.3569

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 2.876515   0.067631  42.532   <2e-16 ***
exper[sam]  0.004776   0.004335   1.102    0.272
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.7122 on 266 degrees of freedom
Multiple R-squared:  0.004542,  Adjusted R-squared:  0.0007998
F-statistic: 1.214 on 1 and 266 DF,  p-value: 0.2716
```

Computation in R: $\hat{y}$ and $\hat{e}$

```r
y <- matrix(lwage[sam], ncol=1)
x <- cbind(1, exper[sam])

xxi <- solve(crossprod(x))
xy <- crossprod(x,y)
betahat <- xxi %*% xy

yhat <- x %*% betahat # fitted values
ehat <- y - yhat # residuals
```

```r
n <- nrow(y)
k <- ncol(x)

# residual standard error
sig2hat <- sum(ehat^2) / n
s2      <- sum(ehat^2) / (n-k)

# std err (homosk)
V0 <- s2*xxi
s_beta <- sqrt(diag(V0))
```

Computation in t-Statistics and p-Values

```
# t-stats
tstats <- (betahat - 0) / s_beta
tstats
```

```
          [,1]
[1,] 42.532241
[2,]  1.101689
```

```
# p-values
p_vals <- 2 * (1 - pt(tstats, df=n-k))
round(p_vals,5)
```

```
        [,1]
[1,] 0.00000
[2,] 0.27159
```

```
# Asymptotic T
zstats <- (betahat - 0) / s_beta
zstats
```

```
          [,1]
[1,] 42.532241
[2,]  1.101689
```

```
# p-vals
p_vals_asymp <- 2 * (1 - pnorm(zstats))
round(p_vals_asymp,5)
```

```
       [,1]
[1,] 0.0000
[2,] 0.2706
```

Other Linear Hypotheses

Single Linear Restriction

Suppose your Null Hypothesis is **one** linear restriction on **multiple** coefficients for some scalar value w and length-k vector r:

$$H_0 : r_1\beta_1 + r_2\beta_2 + \ldots + r_k\beta_k = w$$

We can show that:

- $r'\hat{\beta}$ estimates $r'\beta$ (and is the BLUE under homoskedasticity)
- with variance $V_{r'\hat{\beta}} = r'V_{\hat{\beta}}r$
- and variance estimator $\hat{V}_{r'\hat{\beta}} = r'\hat{V}_{\hat{\beta}}r$

Construct the test statistic $T(r'\hat{\beta})$ to test the hypothesis or inform construction of confidence intervals:

$$T(r'\hat{\beta}) = \frac{r'\hat{\beta} - r'\beta}{se(r'\hat{\beta})} = \frac{r'\hat{\beta} - w}{\sqrt{r'\hat{V}_{\hat{\beta}}r}}$$

Then $T \sim t_{n-k}$ if you assume $e \sim N(0, \sigma^2)$, otherwise $T \xrightarrow{d} N(0, 1)$

Multiple Linear Restrictions

Suppose your Null Hypothesis is a **set** of linear restrictions, written with $q \times k$ matrix R and length-q vector w:

$$R'\beta = w$$

Calculate the Wald statistic W, which is a generalization of the t-statistic to multiple restrictions:

$$W = \left(R'\hat{\beta} - w\right)' \left(R'\hat{V}_{\hat{\beta}}R\right)^{-1} \left(R'\hat{\beta} - w\right) \xrightarrow{d} \chi^2_q$$

For hypothesis tests, choose cutoff c satisfying

- $1 - G_q(c) = \alpha$ (where G_q is the χ^2_q CDF)
- $\Pr(W > c | H_0) \to \alpha$

Thus we reject H_0 if $W > c$

F-Tests

If you assume $e \sim N(0, \sigma^2)$, then $W/q \sim F_{q,n-k}$:

$$F = W/q = \left(R'\hat{\beta} - w\right)' \left(R'(X'X)^{-1}R\right)^{-1} \left(R'\hat{\beta} - w\right) / qs^2 \sim \frac{\chi^2_q/q}{\chi^2_{n-k}/(n-k)} = F_{q,n-k}$$

Let tilde denote estimators from a restricted regression (ie, where some parameters are hypothesized to be zero). Then, with quite a bit of algebra, we can show

$$F = \frac{[\text{SSE}(\tilde{\beta}) - \text{SSE}(\hat{\beta})]/q}{\text{SSE}(\hat{\beta})/(n-k)} = \frac{(\tilde{\sigma}^2 - \hat{\sigma}^2)/q}{\hat{\sigma}^2/(n-k)} = \frac{(R^2 - \tilde{R}^2)/q}{(1-R^2)/(n-k)}$$

And when the Null Hypothesis is that **all** slope coefficients equal zero (ie, $H_0 : \beta_j = 0$ for all $j > 0$) – which is the F-statistic printed by the `summary()` function in R – the formula simplifies further:

$$F = \frac{R^2/q}{(1-R^2)/(n-k)} = \frac{\text{SSR}/q}{\text{SSE}/(n-k)}$$

Comments on t, χ^2 and F tests

t & F: Suppose you test the joint hypothesis $H_0 : \beta_4 = 0$ & $\beta_5 = 0$.

- It is possible that the F-test is statistically significant but the two separate t-tests are not
- It is possible that the two t-tests are statistically significant but the F-test is not

F & χ^2:

- $F_{q,n\text{-}k} : \chi^2$ relationship is analogous to $t_{n\text{-}k} : N(0,1)$
- A small sample exact result (under normal errors) compared to the asymptotic result

The default F-stat from `summary()`:

- Tests whether all slope (but not the intercept) coefficients are 0
- Useful for small n to answer the question "is there *any* explanatory power in the regression?"
- A statistically insignificant result indicates that the model is not significantly better than $\bar{y}$
- However, a statistically significant result does **not** enable you to conclude that the model is good, perfect, valid, or correct

Computation in R: t and F

```r
# t-stat
tstat <- (betahat[2] - 0)/s_beta[2]
tstat
```

```
[1] 1.101689
```

```r
p_t <- 2 * (1 - pt(tstat, df=n-k))
p_t
```

```
[1] 0.2715926
```

```r
tstat^2 # same as F!
```

```
[1] 1.213719
```

```r
# F-test
sigYtilde <- sum((y - mean(y))^2) / n
rsq <- 1 - sig2hat/sigYtilde

Fstat <- (rsq/1) / ((1-rsq)/(n-k))
Fstat
```

```
[1] 1.213719
```

```r
p_F <- 1 - pf(Fstat, 1, n-k)
p_F
```

```
[1] 0.2715926
```

Next Time

Practical tools for developing regression models:

- Categorical X Variables
- Log-Linear an Log-Log models
- Multicollinearity
- Errors in Variables
- Omitted Variables
- Leverage and Outliers
- Forecasting