Econometrics I

Lecture 4: Inference and Standard Errors

Chris Conlon

Fall 2025

Hypothesis testing

- We are often interested in testing theories, or testing hypotheses about the values of certain parameters
- Simplest example: testing whether mean of a variable $\mu_X \equiv E[X]$ is different from a particular value:

$$\begin{aligned} H_0 : &\qquad \mu_X = a \\ H_1 : &\qquad \mu_X \neq a \end{aligned}$$

- A hypothesis test typically involves a **null hypothesis** and **alternative hypothesis**. The alternative hypothesis could also be about a particular value ($H_1 : \mu_X = b$, or a one-sided rejection of the null ($H_1 : \mu_X > a$).

Review: z test

- If X_i is i.i.d. normal with *known* variance σ^2, then

$$\overline{X} \sim \mathcal{N}\left(\mu_x, \sigma^2/n\right)$$

- In this case, we know the distribution of our estimate $\overline{X}$. We can test

$$H_0 : \mu_x = a \qquad H_1 : \mu_x \neq a$$

  using a **z test**.

- We construct the test statistic, which under the null hypothesis has the standard normal distribution:

$$z = \frac{\overline{X} - a}{n^{-1/2}\sigma} \xrightarrow{d} \mathcal{N}(0, 1)$$

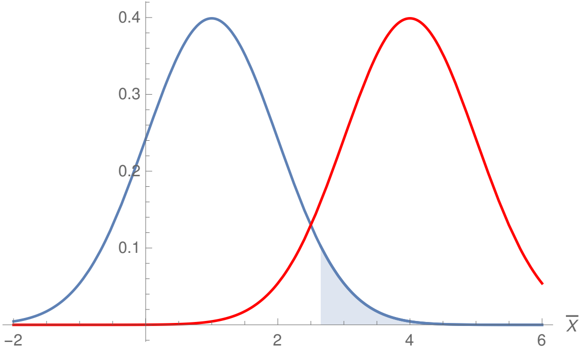
Level and Size of Test

- The **size** (or **level**) of a test is the probability of rejection if the null hypothesis is true. *The size is the rate of false positives or type I errors.*
- When hypothesis testing, we make it hard to reject the null hypothesis. We typically choose the size of the test to be small (most commonly, .01 or .05).

Power of Test

- We typically want to reject only for the outcomes that are least likely under the null hypothesis (or relatively more likely under the alternative hypothesis than the null). For the z test above, we reject only in the tails of the normal distribution. See: Neyman-Pearson Lemma.
- Choosing the rejection region appropriately maximizes the test's **power**, the probability of rejecting the null hypothesis when it is indeed false. Power is often harder to quantify and not something we typically choose. Power is one minus the rate of type II errors (**false negatives**), or failures to reject the null hypothesis when it is false.

Rejection Region and Power

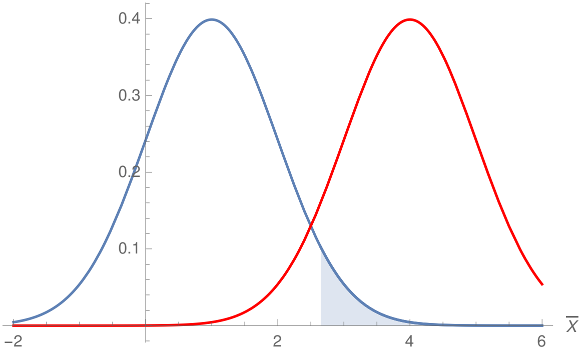


- Suppose $\bar{X}$ is normally distributed with $Var\left(\bar{X}\right) = 1$
- We want to test

$$\begin{aligned} H_0 : \quad & \mu_x = 1 \\ H_1 : \quad & \mu_x = 4 \end{aligned}$$

- The blue and red lines are the PDF of $\bar{x}$ under the null and alternative hypotheses, respectively
- The shaded region is the rejection region with level $\alpha = .05$ that maximizes power. Note that this is for $\bar{X} \geq 2.65$.

Rejection Region and Power



- Note that the rejection region is the region where the PDF of the alternative hypothesis is high relative to the null hypothesis.
- The maximum power test with level .05 is the test that rejects for the 5% of the null-hypothesis PDF in which H_1's likelihood (probability density) is highest relative to H_0's.
- We often take for granted that rejection regions are in the tails of the null-hypothesis PDF; this is why.

Review: t Statistics

- Let’s return to testing the value of a normally distributed random variable’s mean, but now let’s suppose that σ^2 is not known (which is typically the case).
- Our test statistic instead is

$$t = \frac{\bar{X} - a}{n^{-1/2} s}$$

where

$$s = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} \left(X_i - \bar{X} \right)}^{2}$$

- Here, t has a t-distribution with $n - 1$ degrees of freedom.

Testing Paradigm

- We focus on different versions of **Wald tests**, which are based on test statistics that are (approximately) normally distributed.
- Other paradigms:
  - Likelihood Ratio tests and goodness-of-fit-based tests. The idea here is to compare how well different models fit the data.
  - Lagrange multiplier test: for example, testing whether residuals from a restricted model are correlated with excluded variables.

Motivating Small Sample t-Tests

- Last week we learned that if N is large then,

$$\mathbf{b}_{OLS} \overset{a}{\sim} \mathcal{N}\left(\beta, Var(\mathbf{b}_{OLS})\right)$$

  - This hinges on knowing $Var(\mathbf{b}_{OLS})$
  - We rarely know this in practice — we estimate it instead
  - Like testing the mean of a normal random variable, estimating the variance of the test statistic puts us in a t-test situation.

t-Statistics for OLS Parameters (homoscedasticity)

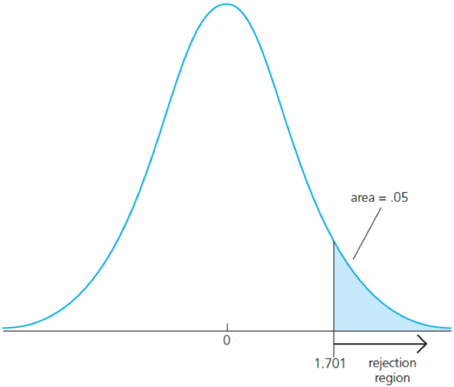
$$\frac{\mathrm{b}_{OLS,k} - \beta_k}{\sqrt{s^2 \left(\mathrm{X}'\mathrm{X}\right)^{-1}_{kk}}} \sim t_{n-K}$$

- where K is the number of parameters, s^2 is the estimator of the variance of ε, and

  $$\left(\mathrm{X}'\mathrm{X}\right)^{-1}_{kk}$$

  refers to the kth diagonal element of $\left(\mathrm{X}'\mathrm{X}\right)^{-1}$.
- Note that the denominator of the above formula is the standard error for the kth estimated parameter $\mathrm{b}_{OLS,k}$.

The t-Distribution



- Similar to the $\mathcal{N}(0,1)$ but parametrized by degrees of freedom
- The tails are fatter but become $\mathcal{N}(0,1)$ as df go to ∞

Example of Reading a t-Table

Example of a table of critical values for t distribution from a textbook:

Degrees of Freedom	.10	.05
1	6.31	12.71
2	2.92	4.30
$\vdots$		
28	1.70	**2.05**
$\vdots$		
∞	1.65	**1.96**

- If N were very large we would use the $\mathcal{N}(0,1)$ approximation which is exactly the case that $df = \infty$
- If $N < \infty$ we can use a table like this, or a computer does it for us
- *Example:* If $N = 30$, $K = 2$, then $df = N - K = 28$ the 5% cutoff value is 2.05

An Example (Bivariate Regression)

Suppose I have the following estimated parameters on 30 observations

$$b_1 = 1.00$$

$$\sum_{i=1}^{N} (X_i - \bar{X})^2 = 14$$

$$\sum_{i=1}^{N} e_i^2 = 100$$

1. First, state the hypothesis:

$$H_0 : \beta_1 = 0$$

$$H_1 : \beta_1 \neq 0$$

Joint hypotheses

- Sometimes we want to test multiple parameters:

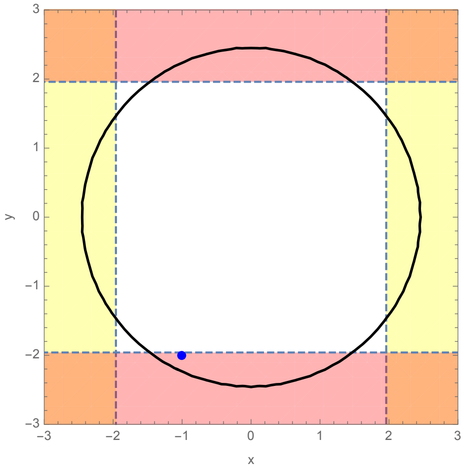
$$H_0: \qquad \beta_{exp} = 0 \quad \text{AND} \quad \beta_{exp2} = 0$$

$$H_1: \qquad \beta_{exp} \neq 0 \quad \text{OR} \quad \beta_{exp2} \neq 0$$

- Note that we do not want to do two separate t-tests for this hypothesis.

Illustration of Two t-Tests Failing

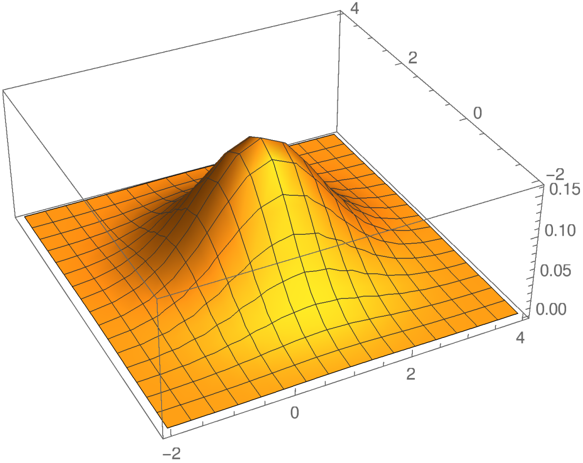
Suppose we have a large sample and t-statistics are -1 and -2. Do we reject null?

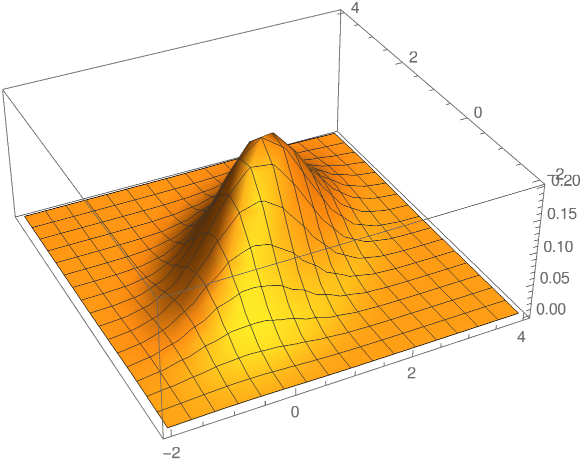


If the t's are independent:

- The circle contains 95% of the probability for two independent t-statistics; the area outside it is the rejection rejoin for the joint t-test.
- The dashed lines are the rejection regions for each of the individual t-tests. (5% level)
- Even though naively would reject, in actuality *not significant*
- What happens with correlated normal RVs?

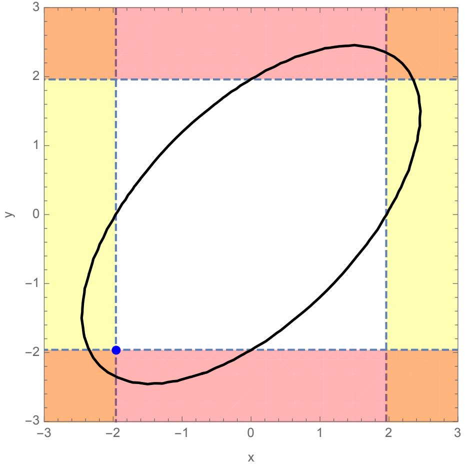
Bivariate normal: independent and correlated





T-Tests with correlation

If the *t*'s are correlated this is the picture:



- Now, the area outside the ellipse is the rejection region for the joint t-test (5% level)
- The dashed lines are the rejection regions for each of the individual t-tests. (5% level)
- Now, notice that even with $t_1 = -2$, $t_2 = -2$, which would be a rejection according to each of the individual tests, is not a rejection of the joint test.

Correcting for Correlation: The F-Test

The issues we have are:

1. Testing a joint hypothesis with independent tests will not give the correct type 1 error
2. Correlated $\hat{\beta}$'s make things very messy

How can we solve this?

- First get a statistic that combines both hypotheses
  - Should be "big" when either t_1 or t_2 or both are big
  - Should include both t's
- Natural candidate:

$$F = t_1^2 + t_2^2$$

  - Always positive and only big when t's are big
  - If t_1 and t_2 are independent normals then $F \sim \chi^2$

Correcting for Correlation: The F-Test, Cont'd

Our candidate test:

$$\frac{1}{2} \times (t_1^2 + t_2^2)$$

- Has a well understood distribution *when* t's are independent
- If not, we can *rotate* the t's so they are
  - Non-matrix formula (for 2 parameters):

$$F = \frac{1}{2} \times \frac{t_1^2 + t_2^2 - 2\rho_{t_1,t_2} t_1 t_2}{1 - \rho_{t_1,t_2}}$$

  - Matrix version (for k parameters):

$$\hat{\beta} - \beta \sim \mathcal{N}\left(0, \Sigma_{\hat{\beta}}\right) \Rightarrow \Sigma_{\hat{\beta}}^{-1/2} \times \left(\hat{\beta} - \beta\right) \sim \mathcal{N}(0, I)$$

This implies,

$$(\hat{\beta} - \beta)' \Sigma^{-1} (\hat{\beta} - \beta)/k \sim \chi_k^2/k = F_k$$

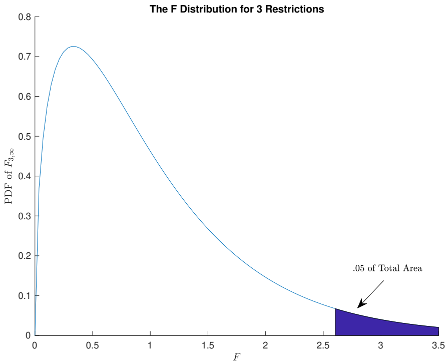
What is the F Distribution?

New test statistic:

$$F = \frac{1}{2} \times \frac{t_1^2 + t_2^2 - 2\rho_{t_1,t_2} t_1 t_2}{1 - \rho_{t_1,t_2}}$$

- Almost always requires a computer
- Ugly formula that follows a simple distribution
- In general, for q restrictions, we will calculate the F statistic and it will be distributed F_q ($F_{q,\infty}$ sometimes)
- Related to take the sum of squared normal random variables
- Critical values will depend on the number of restrictions
- Fun fact: for 1 restriction $F = t^2$

Critical Values of the F



- The distribution looks different than the *t*
- But the testing procedure is the same!
  - Find a critical value so that $P(F > cv) = .05$
  - If *F* is large *given the null* then null is unlikely to be true
  - Critical value depends on number of restrictions, *q*

F-tests: General Definition

- We are interested in testing the following linear restrictions on the parameters:

$$\mathbf{R}\boldsymbol{\beta} = \mathbf{q},$$

  where usually $\mathbf{q} = \mathbf{0}$, but not always.

- What would $\mathbf{R}$ and $\mathbf{q}$ be if we were testing whether two slopes were equal?
- The F statistic (or feasible Wald statistic):

$$F = \frac{(Rb_{OLS} - q)' \left\{R \left[s^2 (X'X)^{-1}\right] R'\right\}^{-1} (Rb_{OLS} - q)}{J},$$

  which has a $F[J, n - K]$ distribution, where J is the number of rows of $\mathbf{R}$ (the number of restrictions).

F-tests: equivalent definitions

$$F = \frac{(Rb_{OLS} - q)' \left\{R\left[s^2 (X'X)^{-1}\right] R'\right\}^{-1} (Rb_{OLS} - q)}{J},$$

- We could also write:

  $$F = \frac{SSE_{CLS} - SSE_{OLS}}{Js^2},$$

  where SSE_{OLS} is the sum of squared residuals for the (unrestricted) OLS estimator, SSE_{CLS} is the sum of squared residuals under the **constrained least squares** estimator with constraints $\mathbf{R}\boldsymbol{\beta} = \mathbf{q}$.
- We are not going to delve into constrained least squares, but the point here is that there are two equivalent ways to think about the F-statistic: (1) as a Wald statistic, which is based on comparing an asymptotically normal parameter estimate to a hypothesized value, and (2) as an assessment of how much better the unconstrained model fits than the constrained model.

Non-Nested Models

- We have considered only nested models thus far. When testing

  $$\mathbf{R}\beta = \mathbf{q},$$

  we are testing a restricted linear model against alternative hypothesis of an unrestricted linear model, *which includes the restricted model as a special case*.

- Sometimes we want to compare non-nested models, which brings us to model selection. The main idea is to balance the model's goodness of fit and number of parameters: see adjusted R^2, Akaike Information Criterion, Bayesian Information Criterion. Machine learning approaches typically try to compare models by directly assessing out-of-sample performance. More on this stuff later and/or next semester.

Confidence Intervals I

- Note that if

$$\mathrm{b} \sim \mathcal{N}\left(\boldsymbol{\beta}, \Sigma)\right),$$

then

$$b_k \sim \mathcal{N}\left(\beta_k, \Sigma_{kk}\right),$$

and

$$Pr\left[b_k - z_{(1-\alpha/2)}\sqrt{\Sigma_{kk}} \leq \beta_k \leq b_k + z_{(1-\alpha/2)}\sqrt{\Sigma_{kk}}\right] = \alpha$$

where $z_{(1-\alpha/2)}$ is the value such that the CDF of the standard normal distribution is $1 - \alpha/2$.

Confidence Intervals II

- Similarly, when

$$\frac{b_k - \beta_k}{\sqrt{\hat{\Sigma}_{kk}}} \sim t_{n-K}$$

because the variance Σ_{kk} has to be estimated, then

$$Pr\left[b_k - t_{(1-\alpha/2),n-K}\sqrt{\hat{\Sigma}_{kk}} \leq \beta_k \leq b_k + t_{(1-\alpha/2),n-K}\sqrt{\hat{\Sigma}_{kk}}\right] = \alpha$$

where $t_{(1-\alpha/2),n-K}$ is the value such that the CDF of the t-distribution with $n-K$ degrees of freedom is $1-\alpha/2$.

Confidence Intervals III

- We define the $1 - \alpha$ **confidence interval** for b_k as

$$\left(b_k - t_{(1-\alpha/2),n-K}\sqrt{\hat{\Sigma}_{kk}}, b_k + t_{(1-\alpha/2),n-K}\sqrt{\hat{\Sigma}_{kk}}\right)$$

- Note that this confidence interval is a function of the data – the end points of the confidence interval are statistics and therefore random variables in their own right.
- Defining the confidence interval in this way, the probability that the confidence interval contains the true parameter is $1 - \alpha$ (if the asymptotic distribution of the estimator is taken as the true distribution). That is, if $\alpha = .05$, this is called a 95% confidence interval, and there is a 95% chance it will contain the true parameter (assuming that the asymptotic approximation holds).

Bootstrap

- Bootstrap takes a different approach.
  - Instead of estimating $\hat{\theta}$ and then assuming a Normal/t-distribution
  - What if we directly tried to construct the sampling distribution of $\hat{\theta}$?
- Our data $(X_1, \ldots, X_n) \sim P$ are drawn from some measure P
  - We can form a nonparametric estimate $\hat{P}$ by just assuming that each X_i has weight $\frac{1}{n}$.
  - We can then simulate a new sample $X^* = (X_1^*, \ldots X_n^*) \sim \hat{P}$.
    - Easy: we take our data and construct n observations by sampling with replacement
  - Compute whatever statistic of X^*, $S(X^*)$ we would like.
    - Could be the OLS coefficients $\beta_1^*, \ldots, \beta_k^*$.
    - Or some function β_1^*/β_2^*.
    - Or something really complicated: estimate parameters of a game $\hat{\theta}^*$ and now find Nash Equilibrium of the game $S(X^*, \hat{\theta}^*)$ changes.
  - Do this B times and calculate at $Var(S_b)$ or $CI(S_1, \ldots, S_b)$.

Bootstrap: Bias Correction

The main idea is that $\hat{\theta}^{1*}, \ldots, \hat{\theta}^{B*}$ approximates the sampling distribution of $\hat{\theta}$. There are lots of things we can do now:

- We already saw how to calculate $Var(\hat{\theta}^{1*}, \ldots, \hat{\theta}^{B*})$.

$$\frac{1}{B-1}\sum_{b=1}^{B}(\hat{\theta}^{*}_{(b)} - \overline{\theta^{*}})^2$$

- Calculate $\mathbb{E}(\hat{\theta}^{*}_{(1)}, \ldots, \hat{\theta}^{*}_{(B)}) = \overline{\theta^{*}} = \frac{1}{B}\sum_{b=1}^{B}\hat{\theta}^{*}_{(b)}$.

Bootstrap: Bias Correction

- We can use the estimated bias to bias correct our estimates

$$\begin{aligned} Bias(\hat{\theta}) &= \mathbb{E}[\hat{\theta}] - \theta \\ Bias_{bs}(\hat{\theta}) &= \overline{\theta^*} - \hat{\theta} \end{aligned}$$

Recall $\theta = \mathbb{E}[\hat{\theta}] - Bias[\hat{\theta}]$:

$$\hat{\theta} - Bias_{bs}(\hat{\theta}) = \hat{\theta} - (\overline{\theta^*} - \hat{\theta}) = 2\hat{\theta} - \overline{\theta^*}$$

- Correcting bias isn't for free - variance tradeoff!
- Linear models are (hopefully) unbiased, but most nonlinear models are consistent but biased.

Bootstrap: Confidence Intervals

There are actually three ways to construct bootstrap CI's:

1. Obvious way: sort $\hat{\theta}^*$ then take $CI : [\hat{\theta}^*_{\alpha/2}, \hat{\theta}^*_{1-\alpha/2}]$.
2. Asymptotic Normal: $CI : \hat{\theta} \pm 1.96\sqrt{V(\hat{\theta}^*)}$. (CLT).
3. Better Way: let $W = \hat{\theta} - \theta$. If we knew the distribution of W then: $Pr(w_{1-\alpha/2} \leq W \leq w_{\alpha/2})$:

$$CI : [\hat{\theta} - w_{1-\alpha/2}, \hat{\theta} - w_{\alpha/2}]$$

We can estimate with $W^* = \hat{\theta}^* - \hat{\theta}$.

$$CI : [\hat{\theta} - w^*_{1-\alpha/2}, \hat{\theta} - w^*_{\alpha/2}] = [2\hat{\theta} - \theta^*_{1-\alpha/2}, 2\hat{\theta} - \theta^*_{\alpha/2}]$$

Why is this preferred? Bias Correction!

Bootstrap: Why do people like it?

- Econometricians like the bootstrap because under certain conditions it is higher order efficient for the confidence interval construction (but not the standard errors).
    - Intuition: because it is non-parametric it is able to deal with more than just the first term in the Taylor Expansion (actually an Edgeworth Expansion).
    - Higher-order asymptotic theory is best left for real econometricians!
- Practitioner's like the bootstrap because it is easy.
    - If you can estimate your model once in a reasonable amount of time, then you can construct confidence intervals for most parameters and model predictions.

Bootstrap: When Does It Fail?

- Bootstrap isn't magic. If you are constructing standard errors for something that isn't asymptotically normal, don't expect it to work!
- The Bootstrap exploits the notion that your sample is IID (by sampling with replacement). If IID does not hold, the bootstrap may fail (but we can sometimes fix it!).
- Bootstrap depends on asymptotic theory. In small samples weird things can happen. We need $\hat{P}$ to be a good approximation to the true P (nothing missing).

Bootstrap: Variants

The bootstrap I have presented is sometimes known as the nonparametric bootstrap and is the most common one.

Parametric Bootstrap ex: if $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$ then we can estimate $(\hat{\beta}_0, \hat{\beta}_1)$ via OLS. Now we can generate a bootstrap sample by drawing an x_i at random with replacement $\hat{\beta}_0 + \hat{\beta}_1$ and then drawing independently from the distribution of estimated residuals $\hat{\epsilon}_i$.

Wild Bootstrap Similar to parametric bootstrap but we rescale ϵ_i to allow for heteroskedasticity

Block Bootstrap For correlated data (e.g.: time series). Blocks can be overlapping or not.

Summary

- Linear regression theory gives us formulas for estimating $Var(\mathbf{b}_{OLS})$
- We can use that variance estimator to test hypotheses about parameters (using t-Tests and f-Tests) as well as construct confidence intervals.
- When the baseline assumptions of the linear regression model are violated (due to correlation or heteroscedasticity), we need to use somewhat more complex formulas to estimate $Var(\mathbf{b}_{OLS})$.