Empirical Methods

TA Session 1

Foundations of the CLRM, notation and more

Important links

R Studio download:
https://www.rstudio.com/products/rstudio/download/#download
Group formation form: (▸ google sheet)

Random Variables I

- Let's recap from slide 10 of top01a and spend little more time on these topics

- Consider our dependent variable, y_i, and our explanatory variables, x_i

- In econometrics, we treat both y_i and x_i as *random variables*

  - $\equiv$ a variable whose possible values are numerical outcomes of (what we take to be) a random phenomenon

  - "(what we take to be)" is actually quite important:

    - Individual i's behavior may be perfectly deterministic (i.e. non-random)

    - BUT we as econometricians may not know all the factors that make it deterministic

    - $\implies$ from our perspective it *is* random

Random Variables II

- To define a random variable, you need two things:

    1 The values (outcomes) it takes

    2 A probability distribution that assigns probabilities to each of these values

- Types of RVs:

    ► When the probability distribution concerns only one RV, we call it a *univariate distribution*

        ★ And the RV a "univariate random variable"

    ► When the probability distribution concerns several RVs, we call it a *multivariate distribution*

        ★ And the RVs "multivariate random variables"

- Let's start with univariate RVs

Univariate Random Variables I

- RVs can be discrete or continuous:
    - Discrete $\equiv$ taking a limited number of distinct values (e.g. i's final grade)
    - Continuous $\equiv$ taking an infinite number of values over a continuum (e.g. i' hours working on econometrics)
- For a discrete RV, the probability distribution is defined by the probability of each of the $r = 1, ..., R$ discrete events,
    - $P(y_i = y_r) = p_r$
        - ⋆ see example on slide 12, top01a

Univariate Random Variables II

- For a continuous RV, the probability distribution is defined by its "probability density function" (PDF)

  - PDF $\equiv$ is a function whose value at any point in the set of values the random variable can take provides the *relative likelihood* that the value of the random variable equals that point

  - Most common use of the PDFs of a random variable, x_i, $f(x_i)$:

    - The area under the PDF between two points, a and b, equals the probability of observing values of the random variable between a and b:

$$P(a \leq x_i \leq b) = \int_a^b f(x_i)dx_i$$

Mean, Variance, and Standard Deviation I

For a univariate RV, in econometric we rely most on two properties:

1. Its expected value or mean

    ▶ Synonymous measures of the "center" of the distribution

    ▶ Denoted $E(x_i)$ or Ex or μ_x

    ▶ Defined as:

        ★ Discrete: $Ex = \sum_{s=1}^{R} p_s x_s$

        ★ Continuous: $Ex = \int_a^b x f(x) dx$
           (where a and $b \equiv$ the lowest and highest possible values for x)

Mean, Variance, and Standard Deviation II

2. Its variance or standard deviation

   ▶ Two related measures of the "spread" of the distribution around its mean

   ▶ Variance is denoted $V(x_i)$ or Vx or σ_x^2 and is defined as

   $$V(x_i) = E[(x_i - Ex)^2], \text{ or}$$

   $$V(x_i) = E(x_i^2) - (Ex)^2$$

Mean, Variance, and Standard Deviation III

2. Its variance or standard deviation, cont:

   ▶ Standard deviation is denoted $SD(x_i)$ or σ_x and is just the square root of the variance:

   ★ $SD(x_i) = \sqrt{V(x_i)}$

   ★ Equivalently, $\sqrt{\sigma_x^2} = \sigma_x$

   ▶ Remember slide 14 (topic1a): If x_i is measured in hours, then σ_x^2 is measured in *squared hours*

   ★ By contrast, σ_x is measured in hours

Mean, Variance, and Standard Deviation IV
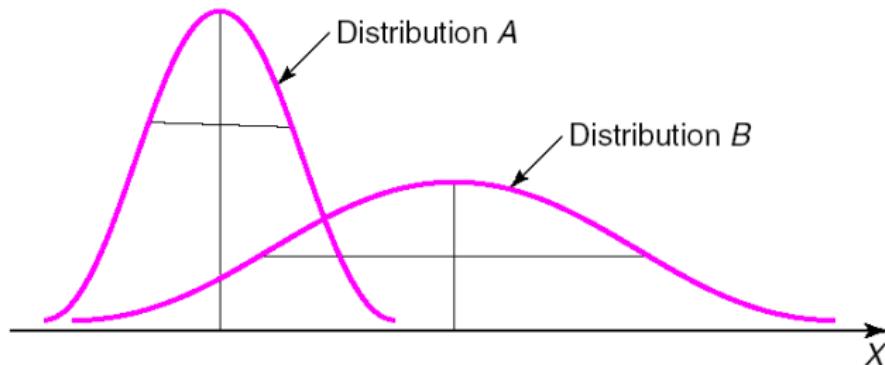
Figure 4.1

- Comparing these two distributions
  - A is "centered" to the left of B, i.e. $\mu_A < \mu_B$
  - A is "less spread out" than B, i.e. $\sigma_A^2 < \sigma_B^2$
    - ★ Equivalently, $\sigma_A < \sigma_B$

  Do you remember the linear rules? See slide 15 topic01a

Common Random Variables: Normal I

- As shown in the lecture (slide 17-24 topic01a) we rely on several common RVs in econometrics

- The most common is a RV distributed as a *Normal* distribution:
  - For x_i normally distributed with mean μ_x and variance σ_x^2
    - ★ We denote this $x_i \sim N(\mu_x, \sigma_x^2)$
 - Its PDF is given by:
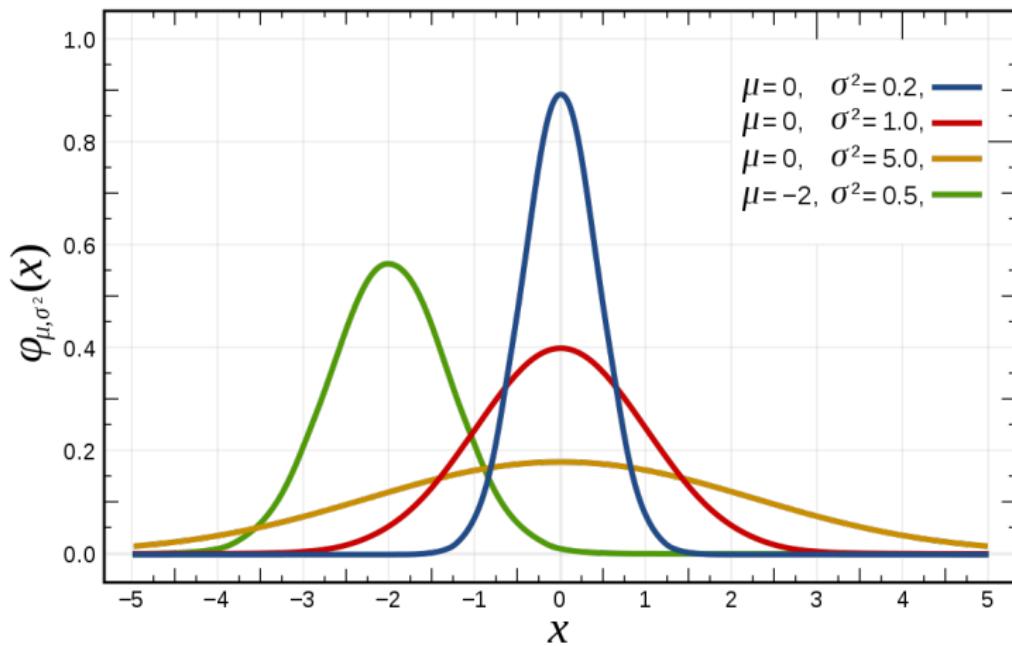
$$f(x_i) = \frac{1}{\sqrt{2\pi}\sigma_x} e^{-\frac{(x_i - \mu_x)^2}{2\sigma_x^2}}$$

Common Random Variables: Normal II

Properties of a Normal distribution:

- Is a "bell-shaped" distribution Figure

- Has about 67% of its area within one standard deviation of the mean

- Has about 95% of its area within two standard deviations of the mean

- "Closed under linear transformation"

  - ▸ ≡ $\boxed{\text{If } x_i \sim N(\mu_x, \sigma_x^2)}$ and y_i is a linear function of x_i

    - ★ $\boxed{\text{And } y_i = a + bx_i}$...

  - ▸ $\boxed{\text{Then } y_i \sim N(*, *)}$

  - ▸ Can use this property to "standardize" any normal distribution

Normal distribution

Common Random Variables: Normal III

- A "standard normal" RV, denoted z
  - is distributed with mean 0 and stdev 1, i.e. $z \sim N(0,1)$

- Its PDF is denoted $\phi(z)$:

$$\phi(z) = \frac{1}{\sqrt{2\pi}}e^{-z^2/2}$$

- If $x_i \sim N(\mu_x, \sigma_x^2)$, then we can "build it" from a standard normal by writing $x_i = \mu_x + \sigma_x z$ (verify it!)

- Equivalently, if $x_i \sim N(\mu_x, \sigma_x^2)$, then can "standardize" it:

$$z = \frac{x - \mu_x}{\sigma_x} \sim N(0,1)$$
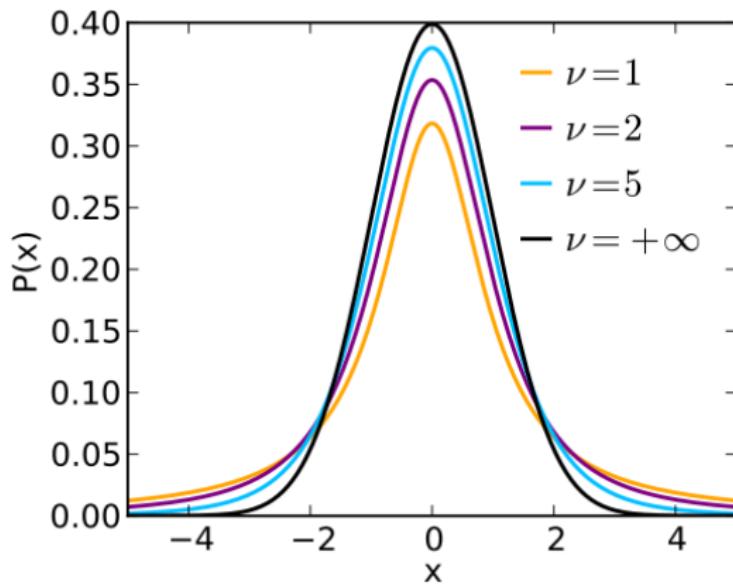
  - Remember: this is the form of many hypothesis tests in econometrics

Common Random Variables: t-distribution

Other common RVs in econometrics:

2. The ("Student's") t-distribution with v degrees of freedom

   ▸ Denoted $x_i \sim t_\nu$

   ▸ Similar to the Normal distribution but with "fatter tails"

      ★ Incorporates the additional uncertainty from estimating σ_x^2 when doing hypothesis tests

   ▸ If have N observations and K covariates in the regression, then have $v = N - K$ degrees of freedom

      ★ Differs from a normal only when v is small Figure

Student's t-distribution
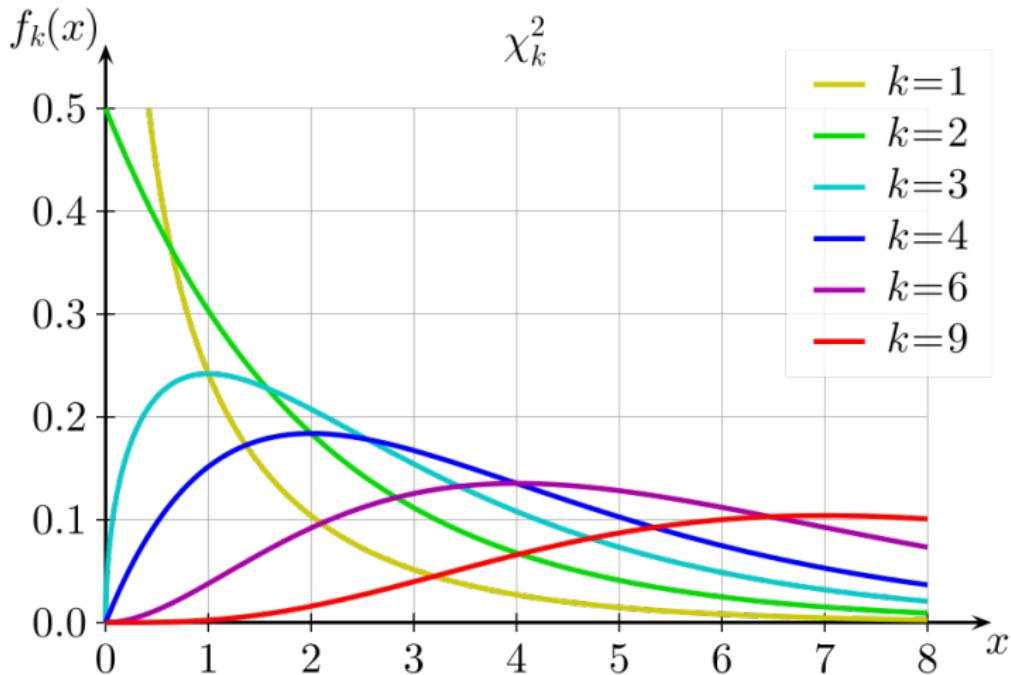
Common Random Variables: χ^2 distribution

Other common RVs in econometrics, cont:

3. The χ^2 distribution (with k degrees of freedom)

   ▶ Denoted $x_i \sim \chi^2_k$

   ▶ Is the distribution of the sum of the squares of k independent standard normal RVs Figure

   ▶ In econometrics, k usually corresponds to the number of restrictions in the hypothesis we're testing

      ★ e.g., $H_0 : \beta_1 = 0$ **and** $\beta_2 = 0$

      ★ $\Rightarrow k = 2$

χ^2 distribution

Common Random Variables: F distribution

Other common RVs in econometrics, cont:

4. F-distribution

   - With v_1 (numerator) d.o.f. and v_2 (denominator) d.o.f

   - Denoted $x_i \sim F(v_1, v_2)$

   - Is the small-sample analog of the χ^2 distribution

     ★ Where v_1 corresponds to k, the number of squared independent normal RVs in the χ^2

     ★ Where v_2 corresponds to $N - 1$, the number of observations

Common Random Variables: Wrapup

As you have seen in the lecture notes:

- In econometrics, "F is to χ^2 as t is to Normal"

- In other words:

  ▶ We (mostly) rely on two kinds of tests in econometrics

    1. Tests of single hypotheses, e.g. $H_0 : \beta_1 = 0$

    2. Tests of multiple hypotheses, e.g. $H_0 : \beta_1 = 0, \beta_2 = 0, ..., \beta_k = 0$

- For each kind of test, there are two distributions one *could* use:

  ▶ For (1), we use the Normal or t distribution

    ⋆ Depending on whether you know σ^2 or have to estimate it

  ▶ For (2), we use the χ^2 or F distribution

Multivariate Random Variables
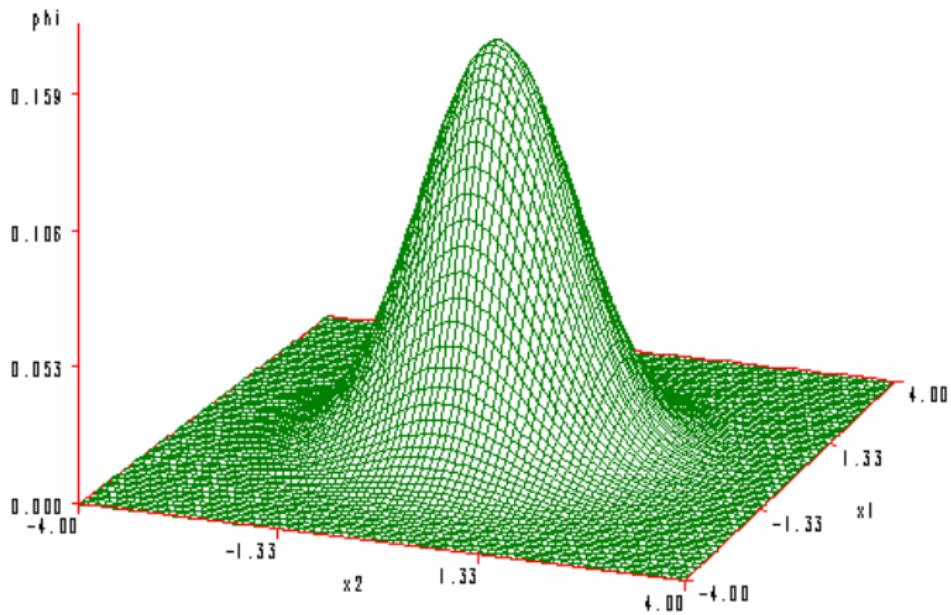
For a multivariate discrete RV, its probability distribution is defined by the probability of each possible combination of values of the K random variables. Recap from the lectures:

- So for a *bivariate* discrete RV,

  - $P(x_i = x_s, y_i = y_r) = p_{sr}$,

- For a multivariate continuous RV, its probability distribution is defined by a joint probability density function (joint PDF), $f(x_i, y_i)$

- For which the probability that x_i and y_i lie within particular ranges is given by the area under the PDF:

$$P(a \leq x_i \leq b, c \leq y_i \leq d) = \int_c^d \int_a^b f(x_i, y_i) dx_i dy_i$$

Example: Bivariate normal distribution I

Bivariate Normal Density — r= 0.0

Multivariate Random Variables II

- Recap from lecture: in econometrics, we rely on several concepts and properties of multivariate RVs:

  1 Marginal distributions

  2 Conditional distributions

  3 Covariance and correlation

  4 Properties of linear combinations of RVs

Marginal Probability Distributions I

- Given a joint probability distribution of K random variables, one can always get a distribution of one of its components by adding up over all of the other components
  - This is its "marginal distribution"

- So for a discrete bivariate RV, for each value r, we can get:

$$P(y_i = y_r) = \sum_s P(x_i = x_s, y_i = y_r)$$
$$= \sum_s p_{sr}$$

- See slide 29 (topic01a) for an example

Marginal Probability Distributions II

And for a continuous random variable

$$f(x_i) = \int_{y_{i,min}}^{y_{i,max}} f(x_i, y_i) dy_i$$

where

- $y_{i,min}$ is the smallest possible value for y_i and
- $y_{i,max}$ is the largest possible value for y_i
  - could be $-\infty$ and ∞

Conditional Probability Distributions

- Slide 30-33 (topic01a)

- With a multivariate RVs, we can now calculate the probability that one random variable (y_i) takes a certain value *given the others have a particular value* ($x_i = x_s$).

  ▶ This is Conditional Probability Distribution

  ▶ (Note we're fixing the value of all K of the x's in x_i)

- For a discrete distribution,

$$P(y_i = y_r | x_i = x_s) = \frac{P(x_i = x_s, y_i = y_r)}{P(x_i = x_s)}$$

- But only if ...?

Means and variances of multivariate RVs

- As for any univariate random variable, we can calculate the mean and variance of
  - Any one of a set of multivariate random variables, (x_i, y_i),
    - $E(x_i)$ and/or $E(y_i)$
    - e.g. $E(x_i) = ?$
  - Any one of a set of conditional random variables, $(y_i | x_i = x_s)$
    - We denote this $E(y_i | x_i)$
    - This is an expectation over the random variable y_i for given values of the random variable x_i

Covariance and Correlation I

- With multivariate random variables, we can now also introduce covariance and correlation

- Covariance measures the degree of (linear) association between two random variables, denoted $Cov(x_i, y_i)$ or σ_{xy}

$$Cov(x_i, y_i) = \sigma_{xy} = E[(x_i - Ex)(y_i - Ey)], \text{ or}$$

$$\boxed{Cov(x_i, y_i) = E(x_i y_i) - ExEy}$$

Covariance and Correlation II

- Remember from the lecture (slide 29, topic01a): the same "units issue" arises for covariance as for variance

  - If x_i is measured in hours and y_i is measured in grade points, then σ_{xy} is measured in hours-grade points.

    - And thus the value of the covariance varies with the units of x_i or y_i

- Correlation, denoted ρ_{xy} or just ρ, solves this problem by providing a *unitless* measure of the relationship between x_i and y_i:

$$\rho_{xy} = \frac{Cov(x_i, y_i)}{SD(x_i)SD(y_i)}, \text{ or}$$

$$\boxed{\rho_{xy} = \frac{\sigma_{xy}}{\sigma_x \sigma_y}}$$

  - Important: $\rho \in [-1, 1]$

Properties of multivariate random variables I

- Properties of multivariate random variables we often use in econometrics:

  ▶ Two random variables are *independent* if and only if their joint probability distribution can be written as the product of respective marginal distributions:

    ★ i.e., if $P(x_i = x_s, y_i = y_r) = P(x_i = x_s)P(y_i = y_r)$

  ▶ Linear rules for combinations of random variables:

    $$E(x_i + y_i) = E(x_i) + E(y_i)$$

    $$V(x_i + y_i) = V(x_i) + V(y_i) + 2Cov(x_i, y_i)$$

    $$V(x_i - y_i) = V(x_i) + V(y_i) - 2Cov(x_i, y_i)$$

Properties of multivariate random variables II

- Putting the two together we get an important result:
  - ▶ If two random variables are independent then they are uncorrelated:
    - ★ i.e. $P(x_i = x_s, y_i = y_r) = P(x_i = x_s)P(y_i = y_r) \implies Cov(x_i, y_i) = 0$
  - ▶ So if x_i and y_i are independent we get:

$$\boxed{\begin{array}{c} V(x_i + y_i) = V(x_i) + V(y_i) \\[2mm] V(x_i - y_i) = V(x_i) + V(y_i) \end{array}}$$

Notation

Notation

- It's now time for you to see what notation we'll use throughout the course ...or equivalently, it's time for you to dislike your TA

- Let's get our hands dirty

Matrix Notation in General

- Consider a $N \times K$ matrix, A:

$$
A = \begin{bmatrix} a_{11} & \cdots & a_{1K} \\ \vdots & \ddots & \vdots \\ a_{N1} & \cdots & a_{NK} \end{bmatrix}
$$

  - a_{ij} is the element in the ith row and j^{th} column of A

Matrix Notation in General, cont.

- Two convenient ways to write A are

  1 As a vector with each element being one of the K columns of A:

  $$A = \begin{bmatrix} a_1 & \cdots & a_k & \cdots & a_K \end{bmatrix}$$

  where a_k is a $N \times 1$ column vector with elements $a_{1k}, a_{2k}, ..., a_{Nk}$

  2 As a column vector with each element being one of the N rows of A:

  $$A = \begin{bmatrix} a_1' \\ \vdots \\ a_i' \\ \vdots \\ a_N' \end{bmatrix}$$

  where a_i' is a $1 \times K$ row vector with elements $a_{i1}, a_{i2}, ..., a_{iK}$

Matrix Notation in General, cont.

- **Note**:

  ▶ Matrices are written in upper case

  ▶ Vectors are written in lower case

    ★ The default vector is a column vector, e.g. a_k, is $N \times 1$

    ★ Row vectors are written with a transpose, e.g. a_i', is $1 \times K$

The CLRM in Matrix Notation

The CLRM in Matrix Notation I

- We write the CLRM as

$$
\begin{aligned}
y_i &= \beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3} + \ldots + \beta_K x_{iK} + \epsilon_i, \qquad i = 1, \ldots, N \\
&= \sum_{k=1}^{K} \beta_k x_{ik} + \epsilon_i \qquad\qquad\qquad\qquad \text{where } x_{i1} = 1, \forall i \\
&= x_i' \beta + \epsilon_i
\end{aligned}
$$

$$
\text{where} \quad x_i = \begin{bmatrix} 1 \\ x_{i2} \\ \vdots \\ x_{iK} \end{bmatrix} \quad \text{and} \quad \beta = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_K \end{bmatrix}
$$

- In sum: $\boxed{y_i = x_i'\beta + \epsilon_i, \qquad i = 1, \ldots, N}$

The CLRM in Matrix Notation II

- We can write this using matrix notation as:

$$y = X\beta + \epsilon$$

$$\text{where} \quad y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix} \quad , \quad \epsilon = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_N \end{bmatrix}$$

$$\text{and} \quad X = \begin{bmatrix} 1 & x_{12} & ... & x_{1K} \\ 1 & x_{22} & ... & x_{2K} \\ \vdots & & & \vdots \\ 1 & x_{N2} & ... & x_{NK} \end{bmatrix} = \begin{bmatrix} x_1' \\ x_2' \\ \vdots \\ x_N' \end{bmatrix}$$

- In sum: $\boxed{y = X\beta + \epsilon}$

The CLRM in Matrix Notation III

- It is important to make sure your matrices are *conformable*, which means:

  ▶ for matrix addition: the matrices must have the identical dimensions

  ▶ for matrix multiplication: the number of columns in the left matrix must match the number of rows in the right

- **Always check conformability!**

The CLRM in Matrix Notation IV

- For individual i:

$$\underbrace{y_i}_{(1\times 1)} = \underbrace{x_i'}_{(1\times K)} \underbrace{\beta}_{(K\times 1)} + \underbrace{\epsilon_i}_{(1\times 1)}$$

- And for all the observations in the dataset:

$$\underbrace{y}_{(N\times 1)} = \underbrace{X}_{(N\times K)} \underbrace{\beta}_{(K\times 1)} + \underbrace{\epsilon}_{(N\times 1)}$$

  ▶ Ready for some derivations?

Derivation of the OLS Formula

OLS I

- We estimate the CLRM using OLS

- OLS selects $\hat{\beta}$ to minimize the sum of squared residuals (SSR)

- Let $S(\hat{\beta})$ denote the SSR: *

OLS II

$$S(\hat{\beta}) = y'y - 2\hat{\beta}'X'y + \hat{\beta}'X'X\hat{\beta}$$

- Differentiating with respect to $\hat{\beta}$ and setting to zero give the *normal equations* (or FOCs) in matrix notation:

$$\frac{\partial S}{\partial \hat{\beta}} = *$$

$$\Rightarrow \boxed{\hat{\beta} = *}$$

Notation for OLS Matrices I

So far we have

$$y = X\beta + \epsilon \qquad \hat{\beta} = (X'X)^{-1}X'y$$

- Let's see another common type of notation:

  - How can we rewrite X'?

  - How can we rewrite X?

  - How can we rewrite $X'X$?

  - How can we rewrite $X'y$?

  - So how does the OLS become?

Notation for OLS Matrices IV

- We have thus two equivalent representations:

$$\hat{\beta} = (X'X)^{-1}X'y \quad \Leftrightarrow \quad \hat{\beta} = (\sum x_i x_i')^{-1} \sum x_i y_i$$

  ▶ No panic, these are merely two different ways to write the same thing

  ▶ You are going to see more of this in class

More algebra of the OLS

Assumptions 3 and 4 of the CLRM

- The CLRM relies on many assumptions. You will see all of them in class. Here I only need to state 3 and 4 to show how they translate in matrix notation You're so gonna love the next two slides...

  ▶ Assumption 3: Homoskedasticity

  ▶ Assumption 4: No Correlation

- The mean of a random vector, ϵ, is:

$$E(\epsilon) = E \left[\begin{array}{c} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_N \end{array} \right] = \left[\begin{array}{c} E(\epsilon_1) \\ E(\epsilon_2) \\ \vdots \\ E(\epsilon_N) \end{array} \right]$$

Assumptions 3 and 4 in Matrix Notation I

- The variance-covariance *matrix* of a random vector, ϵ, is:

$$
\begin{aligned}
V(\epsilon) &\equiv E(\epsilon\epsilon') - E(\epsilon)E(\epsilon') \\
&= E\left(\left[\begin{array}{c} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_N \end{array}\right] \left[\begin{array}{cccc} \epsilon_1 & \epsilon_2 & \cdots & \epsilon_N \end{array}\right]\right) - \left[\begin{array}{c} E(\epsilon_1) \\ E(\epsilon_2) \\ \vdots \\ E(\epsilon_N) \end{array}\right] \left[\begin{array}{cccc} E(\epsilon_1) & E(\epsilon_2) & \cdots & E(\epsilon_N) \end{array}\right] \\
&= \left[\begin{array}{cccc} E(\epsilon_1^2) - [E(\epsilon_1)^2] & E(\epsilon_1\epsilon_2) - [E(\epsilon_1)E(\epsilon_2)] & \cdots & E(\epsilon_1\epsilon_N) - [E(\epsilon_1)E(\epsilon_N)] \\ E(\epsilon_2\epsilon_1) - [E(\epsilon_2)E(\epsilon_1)] & E(\epsilon_2^2) - [E(\epsilon_2)^2] & \cdots & E(\epsilon_2\epsilon_N) - [E(\epsilon_2)E(\epsilon_N)] \\ \vdots & \vdots & \ddots & \vdots \\ E(\epsilon_N\epsilon_1) - [E(\epsilon_N)E(\epsilon_1)] & E(\epsilon_N\epsilon_2) - [E(\epsilon_N)E(\epsilon_2)] & \cdots & E(\epsilon_N^2) - [E(\epsilon_N)^2] \end{array}\right]
\end{aligned}
$$

Assumptions 3 and 4 in Matrix Notation II

- For the CLRM, this equals:

$$
V(\epsilon) = \begin{bmatrix}
E(\epsilon_1^2) - [E(\epsilon_1)^2] & E(\epsilon_1\epsilon_2) - [E(\epsilon_1)E(\epsilon_2)] & \cdots & E(\epsilon_1\epsilon_N) - [E(\epsilon_1)E(\epsilon_N)] \\
E(\epsilon_2\epsilon_1) - [E(\epsilon_2)E(\epsilon_1)] & E(\epsilon_2^2) - [E(\epsilon_2)^2] & \cdots & E(\epsilon_2\epsilon_N) - [E(\epsilon_2)E(\epsilon_N)] \\
\vdots & \vdots & \ddots & \vdots \\
E(\epsilon_N\epsilon_1) - [E(\epsilon_N)E(\epsilon_1)] & E(\epsilon_N\epsilon_2) - [E(\epsilon_N)E(\epsilon_2)] & \cdots & E(\epsilon_N^2) - [E(\epsilon_N)^2]
\end{bmatrix}
$$

$$
= \begin{bmatrix}
E(\epsilon_1^2) & E(\epsilon_1\epsilon_2) & \cdots & E(\epsilon_1\epsilon_N) \\
E(\epsilon_2\epsilon_1) & E(\epsilon_2^2) & \cdots & E(\epsilon_2\epsilon_N) \\
\vdots & \vdots & \ddots & \vdots \\
E(\epsilon_N\epsilon_1) & E(\epsilon_N\epsilon_2) & \cdots & E(\epsilon_N^2)
\end{bmatrix}
$$

$$
= \begin{bmatrix}
V(\epsilon_1) & Cov(\epsilon_1, \epsilon_2) & \cdots & Cov(\epsilon_1, \epsilon_N) \\
Cov(\epsilon_2, \epsilon_1) & V(\epsilon_2) & \cdots & Cov(\epsilon_2, \epsilon_N) \\
\vdots & \vdots & \ddots & \vdots \\
Cov(\epsilon_N, \epsilon_1) & Cov(\epsilon_N, \epsilon_2) & \cdots & V(\epsilon_N)
\end{bmatrix} \quad \text{by definition of V/Cov when } E(\epsilon_i) = 0
$$

$$
= \begin{bmatrix}
\sigma^2 & 0 & \cdots & 0 \\
0 & \sigma^2 & \cdots & 0 \\
\vdots & \vdots & \ddots & \vdots \\
0 & 0 & \cdots & \sigma^2
\end{bmatrix} \quad \text{under (A3, Homoskedasticity) and (A4, No Correlation)}
$$

$$
= \sigma^2 I_N \quad \text{where } I_N \text{ is an } N \times N \text{ identity matrix}
$$

- In sum: $\boxed{V(\epsilon) = \sigma^2 I_N}$

In synthesis

- These assumptions (plus assumption 2) can be stated as:

$$\epsilon_i | x_i \sim (0, \sigma^2) , \quad Cov(\epsilon_i, \epsilon_j) = 0, \quad i, j = 1, ..., N, \quad i \neq j$$

- or, using matrix notation:

$$\epsilon | X \sim (0, \sigma^2 I_N)$$

Interval Estimation

The Basics of Interval Estimation I

- An interval estimate of a population parameter, β_k, consists of two bounds within which we expect β_k to lie, i.e.
  - $LB \leq \beta_k \leq UB$
    - ⋆ Where LB and UB are the lower and upper bounds

- The probability that β_k lies within the provided interval estimate is called the confidence coefficient
  - It is denoted $1 - \alpha$
    - ⋆ Where α is the *significance level* of the confidence interval
  - The significance level (α) will come back when we do hypothesis testing.

- For a specified α, the interval $LB \leq \beta_k \leq UB$ is referred to as a $100(1 - \alpha)\%$ confidence interval.

The Basics of Interval Estimation II

Example:
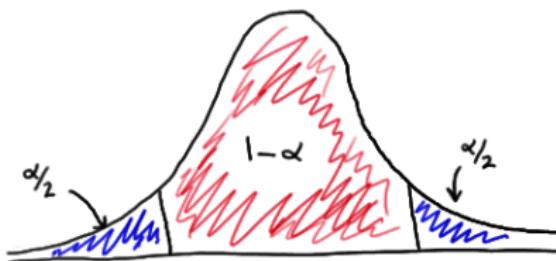
- Suppose $\alpha = .05$ or 5%.

    ▸ Then $1 - \alpha = .95$ or 95%

- Under our CLRM assumptions, it is possible to derive an interval within which 95 times out of 100 the true β_k will reside,

    ▸ $P(LB \leq \beta_k \leq UB) = 1 - \alpha = 95\%$

The Basics of Interval Estimation III
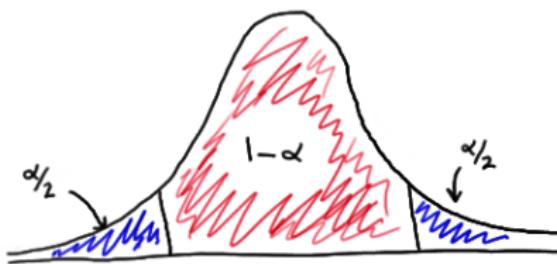
- How?   Three steps:

  1. Find the lower and upper bounds for a t distribution with $N - K$ degrees of freedom (denoted "dof" or "df") for a given α

     * We will call these lower/upper bounds "critical values"

     * (And use them again in hypothesis testing)

  2. "Standardize" the distribution of the population parameter

     * (Recall slide 14 above)

  3. Plug (2) into (1) to obtain a confidence interval for β_k

The Basics of Interval Estimation IV

- Step 1: Calculate the bounds for a t distribution with $N - K$ df
  - Let's let $N - K = 100$ for simplicity

- To do so, define $\bar{t}_{1-\alpha/2}$ be the "critical value" of a t_{100} random variable for which $Pr(t_{100} < \bar{t}_{1-\alpha/2}) \equiv F_{t_{100}}(\bar{t}_{1-\alpha/2}) = 1 - \alpha/2$
  - Where $F_{t_{100}}(\cdot)$ is the CDF of a t distribution with 100 df

- Thus, $\bar{t}_{1-\alpha/2}$ is the value of a t_{100} with $1 - \alpha/2$ probability to its left.

The Basics of Interval Estimation V

- Symmetry of the t distribution implies symmetry of critical values:
  - $\bar{t}_{\alpha/2} = -\bar{t}_{1-\alpha/2}$

- Thus if $\alpha = 0.05$, then $1 - \alpha/2 = .975$ and $\alpha/2 = .025$.
  - Since $F_{t_{100}}(1.984) = 0.975$, the upper critical value, $\bar{t}_{1-\alpha/2}$, is 1.984
    - ★ (See "Inverse t calculator" or a standard t-table to get 1.984)

  - By symmetry, the lower critical value $\bar{t}_{\alpha/2} = -\bar{t}_{1-\alpha/2} = -1.984$

- **Bottom Line (Step 1):** $P(-\bar{t}_{1-\alpha/2} \leq t_{100} \leq \bar{t}_{1-\alpha/2}) = 1 - \alpha$
  - For our example: $P(-1.984 \leq t_{100} \leq 1.984) = 1 - 0.05 = 0.95$

The Basics of Interval Estimation VI

- Step 2: Standardize the distribution of $\hat{\beta}_k$

$$\hat{\beta}_k \sim N(\beta_k, \sigma^2 (X'X)_{kk}^{-1})$$

$$\Rightarrow \quad \frac{\hat{\beta}_k - \beta_k}{\sigma \sqrt{(X'X)_{kk}^{-1}}} \sim N(0,1)$$

$$\Rightarrow \quad \frac{\hat{\beta}_k - \beta_k}{s \sqrt{(X'X)_{kk}^{-1}}} \sim t_{N-K}$$

$$\Rightarrow \quad \frac{\hat{\beta}_k - \beta_k}{stderr(\hat{\beta}_k)} \sim t_{N-K}$$

⋆ where $stderr(\hat{\beta}_k) = s \sqrt{(X'X)_{kk}^{-1}}$ is our estimate of the standard deviation of the distribution of $\hat{\beta}_k$

The Basics of Interval Estimation VII

- Step 3: Combine the results of Steps 1 and 2 to form an interval

$$
\begin{aligned}
P(-\bar{t}_{1-\alpha/2} \leq t_{100} \leq \bar{t}_{1-\alpha/2}) &= 1-\alpha \\
P(-\bar{t}_{1-\alpha/2} \leq \frac{\hat{\beta}_k - \beta_k}{s\sqrt{(X'X)_{kk}^{-1}}} \leq \bar{t}_{1-\alpha/2}) &= 1-\alpha \\
P(-\bar{t}_{1-\alpha/2}s\sqrt{(X'X)_{kk}^{-1}} \leq \hat{\beta}_k - \beta_k \leq \bar{t}_{1-\alpha/2}s\sqrt{(X'X)_{kk}^{-1}}) &= 1-\alpha \\
P(-\hat{\beta}_k - \bar{t}_{1-\alpha/2}s\sqrt{(X'X)_{kk}^{-1}} \leq -\beta_k \leq -\hat{\beta}_k + \bar{t}_{1-\alpha/2}s\sqrt{(X'X)_{kk}^{-1}}) &= 1-\alpha \\
P(+\hat{\beta}_k + \bar{t}_{1-\alpha/2}s\sqrt{(X'X)_{kk}^{-1}} \geq +\beta_k \geq +\hat{\beta}_k - \bar{t}_{1-\alpha/2}s\sqrt{(X'X)_{kk}^{-1}}) &= 1-\alpha \\
P(\hat{\beta}_k - \bar{t}_{1-\alpha/2}s\sqrt{(X'X)_{kk}^{-1}} \leq \beta_k \leq \hat{\beta}_k + \bar{t}_{1-\alpha/2}s\sqrt{(X'X)_{kk}^{-1}}) &= 1-\alpha
\end{aligned}
$$

- For $\alpha = 0.05$ and $N - K = 100$:

$$
P(\hat{\beta}_k - 1.984s\sqrt{(X'X)_{kk}^{-1}} \leq \beta_k \leq \hat{\beta}_k + 1.984s\sqrt{(X'X)_{kk}^{-1}}) = 95\%
$$

Basics of Hypothesis Testing

Hypothesis Testing: Intro

- Remember a few of our side points from earlier in the slides:
    - We'll use the t distribution instead of the Normal for our tests
        - As we'll have estimated s^2 for σ^2
    - We'll focus initially on tests of single hypotheses about single parameters
        - Then we'll move to other types of tests

The Basics of Hypothesis Testing I

There are four steps (in two flavors) to constructing a hypothesis test

1. Explicitly define two opposing hypotheses

   - The *Null Hypothesis* and the *Alternative Hypothesis*

   - Denoted H_0 and H_1

2. Construct a test statistic under the null hypothesis

   - Usually the standardized distribution of our estimate of a population parameter

   - Denote our test statistic $\tilde{t}$

- The next steps depend on which of two paths we take

The Basics of Hypothesis Testing II

First path:

3. Select a level of significance, α, and calculate its critical value

   ▶ as we did for confidence intervals a few slides ago

4. Compare (the absolute value of) our test statistic to the critical value and:

   ▶ either (a) reject or (b) "fail to reject" H_0

   ▶ We use the absolute value as either large positive or large negative values of the test statistic can cause you to reject

The Basics of Hypothesis Testing III

If we take what we call the "p-value" path:

3. Calculate the probability of getting a value of the test statistic as big as it is (in absolute value) if the null hypothesis is true

   ▶ This is the hypothesis test's *p-value*

4. Compare this probability (p-value) to a desired level of significance

   ▶ and either (a) reject or (b) fail to reject H_0

The Basics of Hypothesis Testing IV

Let's illustrate both approaches using a sample exam question:

- What is the best way to reduce traffic fatalities? In 2009, the UK government announced plans to cut the number of deaths on UK roads by a third by reducing speed limits to 20 mph (from 30) on residential roads and to 50 mph (from 60) on single-lane highways.

  Similar concerns worry policy-makers the world over. To address this question in the U.S., a researcher obtained information on vehicle fatality rates for each of the 48 mainland U.S. states between 1982 and 1988 ($N = 336$) and estimated the following relationship:

  $$vfr_{it} = \alpha + \beta_1 beertax_{it} + \beta_2 unemp_{it} + \beta_3 yngdrv_{it} + \beta_4 inc_{it} + \beta_5 vmiles_{it} + d_t + a_i + u_{it}$$

  where d_t is a set of 7 dummy variables (one for each of 1983-1988) with (e.g.) the dummy variable for 1983 equal to 1 for each of the observations in 1983, vfr_{it} is the vehicle fatality rate (per 10,000 population), $beertax_{it}$ is the tax on a case (24 12-oz. cans) of beer, $unemp_{it}$ is the unemployment rate, $yngdrv_{it}$ is the share of young drivers, inc_{it} is the per-capita income, and $vmiles_{it}$ is the average number of vehicle miles driven in state i in year t.

  - ▶ Note: this data is actually "panel data" as we have variation both across states (i) and years (t). This was done on purpose as we'll come back to this example in the slides later in the course.

The Basics of Hypothesis Testing V

A pooled OLS regression with standard errors clustered by state yielded the following results (year dummies omitted):

$$vfr_{it} = \underset{(0.949)}{1.771} + \underset{(0.116)}{0.094 beertax_{it}} + \underset{(0.025)}{0.016 unemp_{it}} + \underset{(2.060)}{2.150 yngdrv_{it}}$$

$$- \underset{(0.029)}{0.099 inc_{it}} + \underset{(0.086)}{0.133 vmiles_{it}} + \hat{v}_{it}$$

where the standard error is listed in parentheses below each coefficient estimate

1 Assume the CLRM assumptions hold. Interpret the coefficient on $beertax_{it}$ in your results.

2 Test the hypothesis that $\beta_1 = 0$ at a 5% significance level.

The Basics of Hypothesis Testing VI

- Step 1: Explicitly define two opposing hypotheses:

  - The *Null Hypothesis*, denoted H_0, specifies a specific value for a population parameter,

    - ★ For our exam question, $H_0 : \beta_1 = 0$

  - The *Alternative Hypothesis*, denoted H_1, specifies the set of alternatives the econometrician is willing to consider

    - ★ For our exam question, $H_1 : \beta_1 \neq 0$

The Basics of Hypothesis Testing VII

- Step 2: Construct a test statistic under the null hypothesis
- We said earlier that we could standardize the distribution of one of our OLS estimates, $\hat{\beta}_k$, as:

$$\hat{\beta}_k \sim N(\beta_k, \sigma^2 (X'X)^{-1}_{kk})$$

$$\Rightarrow \quad \frac{\hat{\beta}_k - \beta_k}{\sigma \sqrt{(X'X)^{-1}_{kk}}} \sim N(0, 1)$$

$$\Rightarrow \quad \frac{\hat{\beta}_k - \beta_k}{s \sqrt{(X'X)^{-1}_{kk}}} \sim t_{N-K}$$

The Basics of Hypothesis Testing VIII

- For our exam question, $N - K = 336 - 6 = 330$, and we can rewrite the last line as:

$$
\begin{aligned}
\frac{\hat{\beta}_k - \beta_k}{s\sqrt{(X'X)^{-1}_{kk}}} &\sim t_{N-K} \\
\Rightarrow \quad \frac{\hat{\beta}_1 - \beta_1}{stderr(\hat{\beta}_1)} &\sim t_{330}
\end{aligned}
$$

where $stderr(\hat{\beta}_1) = \sqrt{\hat{V}(\hat{\beta}_1)} = s\sqrt{(X'X)^{-1}_{11}}$

- If H_0 is true, then $\beta_1 = 0$ and we can calculate our test statistic, $\tilde{t}$:

$$
\tilde{t} = \frac{0.094 - 0}{0.116}
$$

$$
= 0.810
$$

The Basics of Hypothesis Testing IX

Let's first finish up on the "critical value" path

- Step 3a: Select a level of significance, α, and calculate its critical value

    ▸ For our exam question, we are given that $\alpha = 5\%$ and we can go to a t-table or inverse t calculator

        ★ Solving for the $\bar{t}_{0.975}$ such that $P(t_{330} <= \bar{t}_{0.975}) = 0.975)$ yields $\bar{t}_{0.975} = 1.967$

- Step 4a: Compare our test statistic to the critical value and...

    ▸ Fail to reject as 0.810 is not greater than 1.967!

The Basics of Hypothesis Testing X

Alternatively we could have taken the "p-value" path:
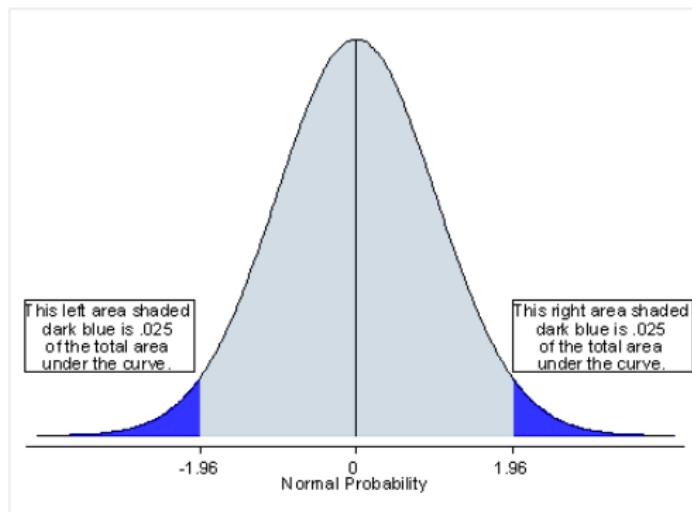
- Step 3b: Calculate the probability of getting a value of our test statistic as big as it is under the assumption H_0 is true

    - For which we need to calculate the probability that a t_{330} random variable could have the value as high or higher than 0.810 in absolute value

        * i.e. as high or higher than 0.810 or as low or lower than -0.810

    - We do so by evaluating $P(|t_{330}| \geq 0.810)$

        * Going again to a t-table or t calculator, we obtain $2 * P(|t_{330}| \geq 0.810) = 0.4185$

        * This is our p-value

- Step 4b: If our p-value is *less than* our selected level of significance, α, we reject H_0, else we "fail to reject" H_0.

    - Thus we again fail to reject as 0.4185 is greater than 0.05. Thus it isn't surprising to get a value of our test statistic of 0.810 if indeed the true $\beta_1 = 0$

One-sided hypothesis tests II

- We sometimes concern ourselves with one-sided alternatives when testing hypotheses.
- As a practical matter, doing a one-sided test changes both
  - ► Calculating the critical value of a test
    - ★ (in the "critical value" path)
  - ► Calculating the probability of seeing the test statistic that you do
   - ★ (in the "p-value" path)
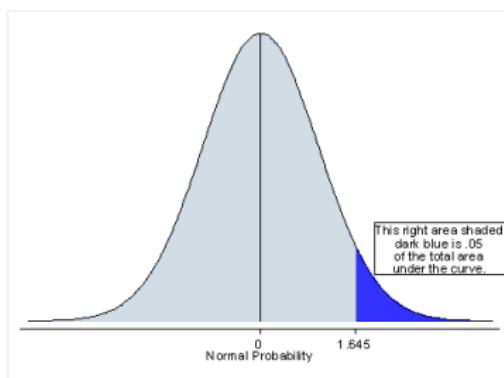
One- versus two-sided hypothesis tests I

- In particular, with a two-sided hypothesis test (e.g. $H_0 : \beta_k = 0$)

  ▶ Either large positive *or* large negative values of $\hat{\beta}_k$ - and thus large positive or negative values of $\tilde{t}$ - would cause you to reject H_0

One- versus two-sided hypothesis tests II

- Whereas with a one-sided hypothesis test (e.g. $H_0 : \beta_k < 0$)
    - Only large *positive* values of $\hat{\beta}_k$ - and thus large positive values of $\tilde{t}$ - would cause you to reject H_0:

    - Note the lower critical value for a one-sided test
        - (As *all 5%* is now in the upper tail)
        - (Whereas previously only 2.5% was up there)

One-sided hypothesis tests III

- Indeed for our exam question, it's hard to imagine that increasing beer taxes would cause *higher* vehicle fatality rates

  - Thus a better test would have been $H_0 : \beta_1 > 0$ versus $H_1 : \beta_1 < 0$

    - ⋆ Note the "game" in hypothesis testing: you actually *want* to reject the hypothesis

    - ⋆ Thus H_0 is the thing you "don't want"

  - In which case only significant *negative* values of our test statistic would cause us to reject

One-sided hypothesis tests IV

- Note the value of our test statistic *doesn't* change
    - $\tilde{t}$ is still equal to 0.810
- What changes is the probabilities we would consider
    - i.e. only big negative values of $\tilde{t}$ would cause us to reject our test
    - And the p-value for this one-sided test should have been $P(t_{330} < 0.810) = 0.791$
    - We *definitely* wouldn't reject H_0!