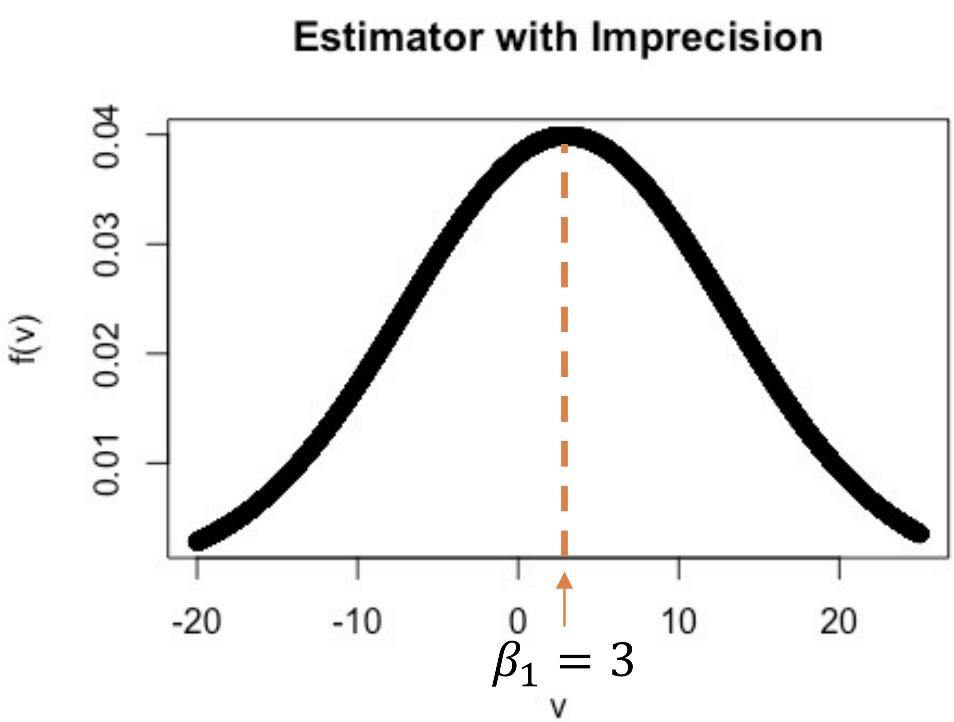
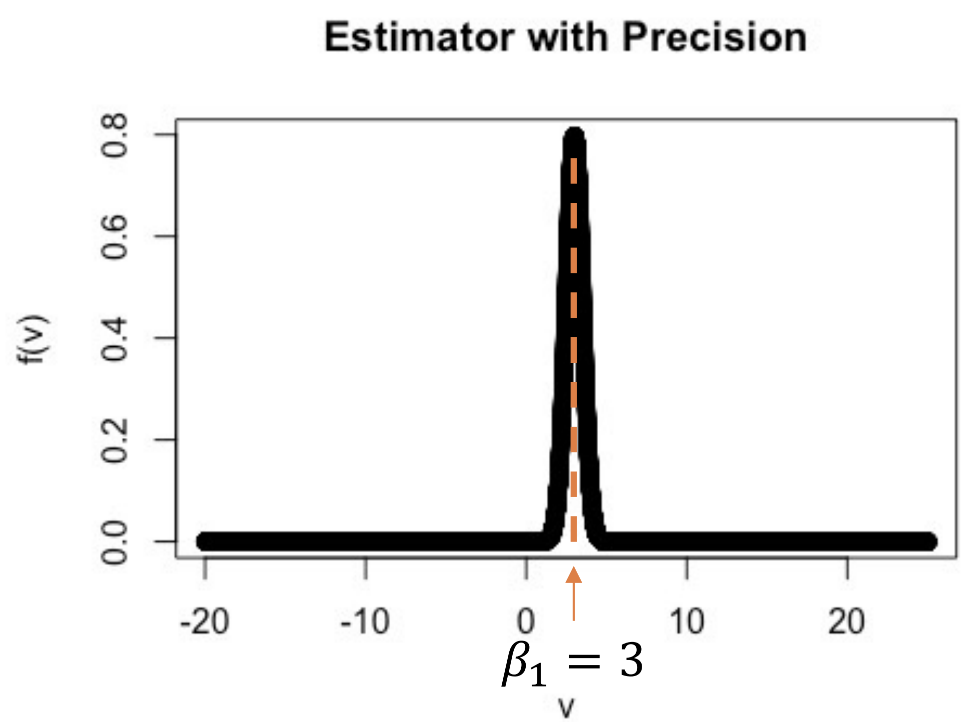
ECON B253: INTRODUCTION TO ECONOMETRICS

Sebastian Anti, Assistant Professor of Economics
Bryn Mawr College

Lecture 5: Interval Estimation and Hypothesis Testing

Visualizing Precision Again

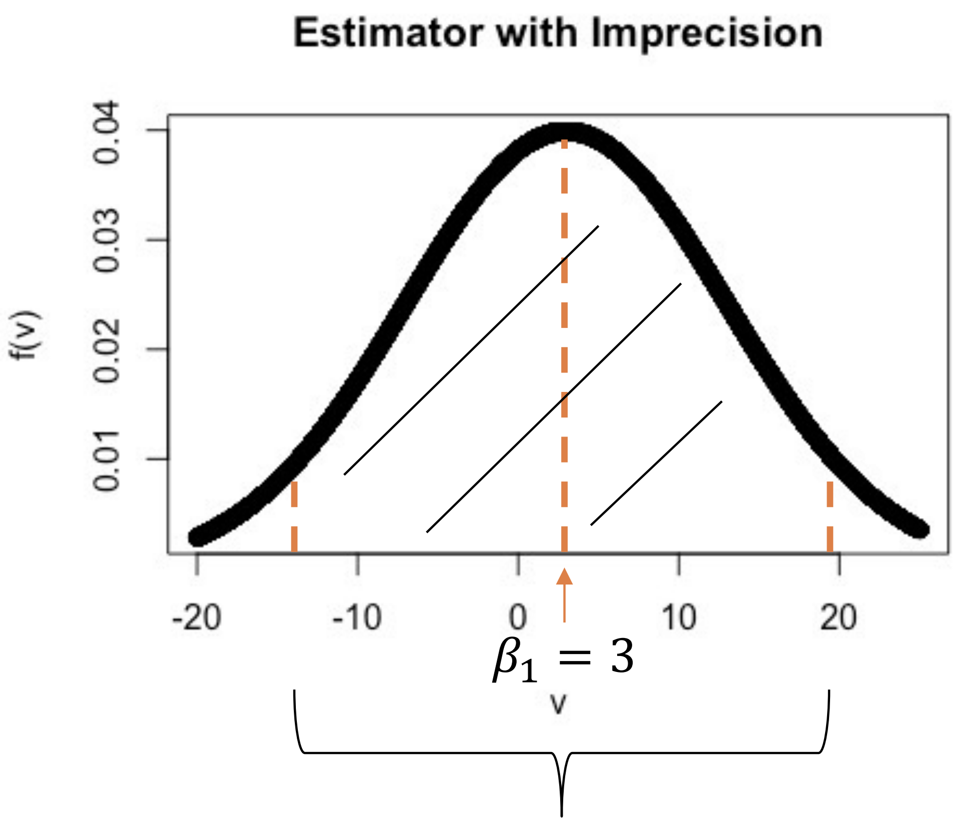




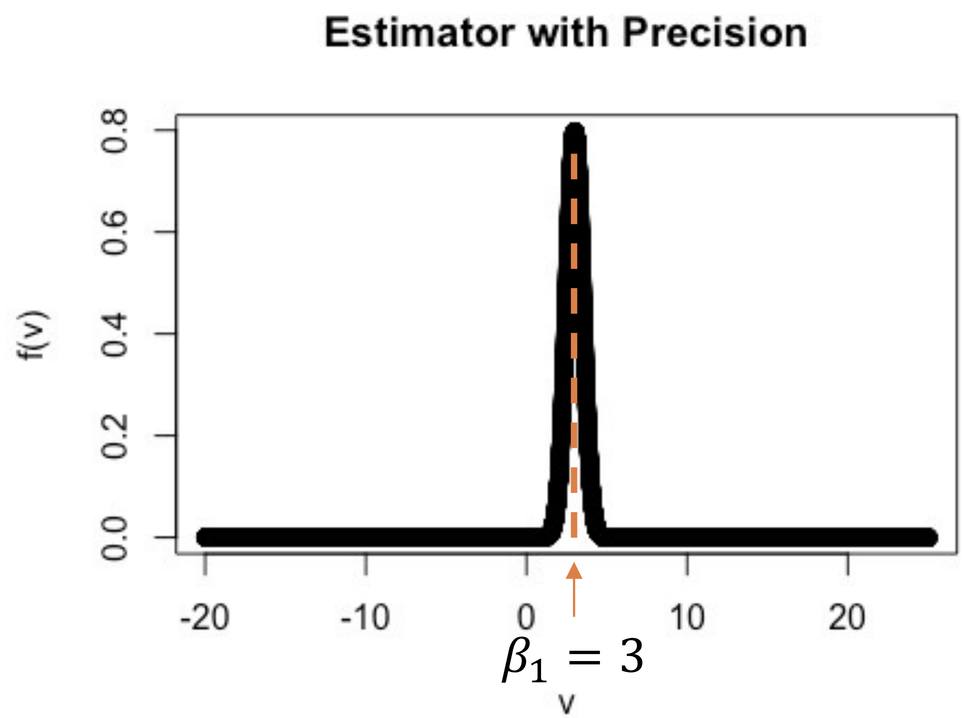
Imprecision means that it's very likely your actual estimate will be negative or positive, or deceptively close to zero

Precision means that you're much more likely to estimate something close to the true value. Almost all possible estimates you can get here are positive and bunched around 3

Intuition of Interval Estimation



Normal distribution: 95% of observations within 2 SDs of expected value. Here it is very likely for the estimator to take on negative values or something close to zero. Low likelihood that our estimate is close to the true value of 3. Low confidence



Normal distribution: 95% of observations within 2 SDs of expected value. Here it is very unlikely for the estimator to take on negative values or something close to zero. High level of confidence that our estimate is close to the true value of 3. High confidence

Motivation for Interval Estimation

- **To do this, we'll do the following:**
  - We'll look at our estimate of the variance around the point estimate, $\hat{\beta}_1$
  - We'll use this to construct bounds for what $\hat{\beta}_1$ could be, given certain levels of confidence
    - This provides an upper value and a lower value within which we have some level of confidence our true value falls within
    - This is called a "confidence interval"
  - **We'll use this to conduct precise statements over whether our estimate is "statistically significant" from zero or not**
    - This is a deceptively powerful statement

Difference Between Point Estimate and Interval Estimate

Point Estimate	Interval Estimate
The point estimate conveys our best estimate of the relationship between X_j and Y, BUT it doesn't convey any information about whether the <u>data are powerful enough to give us confidence that this estimate is correct, or close to correct.</u>	The interval estimate conveys information both on the range of likely relationships between X_j and Y but also the extent to which we should be willing to be confident in these estimates.

Some Conditions for this Discussion

- **We will focus on matters of *precision* here**
  - We'll assume that our estimates are unbiased and estimation is efficient to keep the discussion narrowly focused
  - In other words we'll take Assumptions 1-4 to be true
- **We'll spend a lot of the rest of the class dealing with bias and inefficiency from violations of Assumptions 1-4**
  - In reality researchers and analysts will need to contend with precision issues as well as bias and inefficiency when they're doing work

What Assumptions Are Needed?

- **We need Assumptions 1 through 4 plus…**
- **It turns out we need to make one more assumption:**
  - Assumption 5: The error term, u_i, is normally distributed

- **All of these assumptions together are called the "classical linear model"**
  - Note that Wooldridge includes a 6th assumption

More on Assumption 5

- **Why is Assumption 5 reasonable?**
  - You can think of u as being an average of all the variables unaccounted for in the model that are randomly distributed across the population.
    - This average of randomly distributed variables allows us to invoke the Central Limit Theorem (CLT)
- **The CLT roughly states:**
  - If you average a large number of independent random variables, the resulting random variable is approximately normally distributed no matter what the distribution of the variables you averaged

Quick Review and Extension

- Assumptions 1 through 4 got us the following:

$$E\left(\hat{\beta}_j | X\right) = \beta_j \tag{9}$$

$$Var\left(\hat{\beta}_j\right) = \frac{\sigma^2}{\sum_{i=1}^{N}\left(X_{ji} - \bar{X}\right)^2\left(1 - R_j^2\right)} \tag{10}$$

- If we can make Assumption 5, we can go further and say:

$$\hat{\beta}_j \sim \mathrm{N}\left(\beta_j, \sigma_{\hat{\beta}_j}^2\right), where\ \sigma_{\hat{\beta}_j}^2 = \frac{\sigma^2}{\sum_{i=1}^{N}\left(X_{ji} - \bar{X}\right)^2\left(1 - R_j^2\right)}. \tag{11}$$

Assumption 5 $\Rightarrow \widehat{\beta}_j \sim \mathrm{N}\left(\beta_j, \sigma^2_{\widehat{\beta}_j}\right)$

$$\hat{\beta}_j \sim \mathrm{N}\left(\beta_j, \sigma^2_{\widehat{\beta}_j}\right), where\ \sigma^2_{\widehat{\beta}_j} = \frac{\sigma^2}{\sum_{i=1}^{N}\left(X_{ji} - \bar{X}\right)^2 \left(1 - R_j^2\right)}. \tag{12}$$

- **What this means is:**
  - In addition to knowing the mean and variance of the OLS estimator, we know that its sampling distribution is ***normal***! But why?!

Working Through Why Assumption 5 $\Longrightarrow$ $\widehat{\beta}_j \sim N\left(\beta_j, \sigma^2_{\widehat{\beta}_j}\right)$

- The CLT tells us that a linear function of normal random variables is itself a normal random variable
- We can see that the OLS estimator is a linear function of the us, which is a normal random variable by Assumption 5
- You can see this as follows on the next slide

Proving OLS Estimator is a Linear Function of u

- **Recall,**

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{N}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{N}(X_i - \bar{X})^2} = \frac{\sum_{i=1}^{N}(X_i - \bar{X})Y_i}{\sum_{i=1}^{N}(X_i - \bar{X})X_i} \qquad (13)$$

- **Substitute in:** $\boldsymbol{Y_i = \beta_1 X_i + u_i}$

$$\hat{\beta}_1 = \beta_1 + \frac{\sum_{i=1}^{N}(X_i - \bar{X})u_i}{\sum_{i=1}^{N}(X_i - \bar{X})^2} \qquad (14)$$

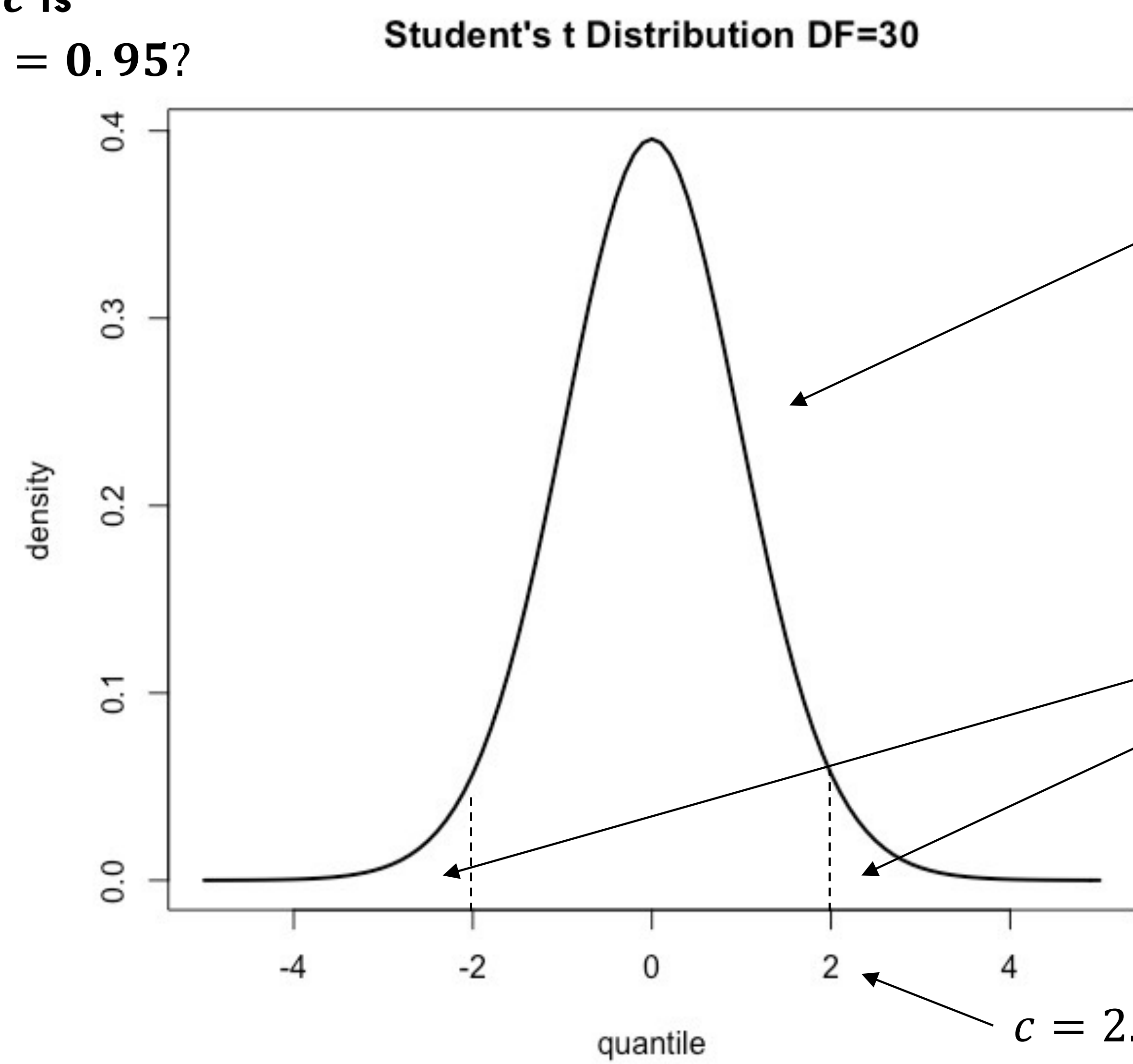
$$\Longrightarrow \hat{\beta}_1 = \beta_1 + \sum_{i=1}^{N} k_i u_i \qquad (15)$$

Using Distributions in Econometrics

- **We have a statistic ($\widehat{\beta}_1$) that has a distribution**
  - By the CLT, it has a normal distribution, which can then be converted into a Student's t distribution
    - <u>Thus, we'll work with it having a Student's t distribution</u>
- **We want to use this information to answer the following types of questions:**
  - For what value of c is $Pr\left(-c < \hat{\beta}_1 < c\right) = 0.95$?
  - For what value of c is $Pr\left(\hat{\beta}_1 > c\right) = 0.1$?
  - For what value of c is $Pr\left(\hat{\beta}_1 < -c \; or \; \hat{\beta}_1 > c\right) = 0.05$?

What Does this Mean?

For what value of c is
$Pr(-c < \hat{\beta}_1 < c) = 0.95$?



PDF of sampling distribution of $\hat{\beta}_1$

Area = 0.05

$c = 2.042$

Back to Interval Estimation

- **Okay, having gotten that detour into statistics out of the way, let's get back to what our goal is: <u>Interval Estimation</u>**
  - Recall, interval estimation is a way of making a specific statement about the *precision* of our estimation

- **To do this, we are trying to calculate a range around $\hat{\beta}_1$ for which we can make a statement like:**
  - "This interval is calculated in such a way that in 95 percent of the samples we could draw from the true model, such an interval would contain β_1"

More Definitions

- **Hypothesis Testing:**
  - A statistical test of the null hypothesis against an alternative hypothesis

- **Critical Value:**
  - In hypothesis testing, the value against which a test statistic is compared to determine whether or not the null hypothesis is rejected

Part 1: Confidence Intervals

t Table

cum. prob	$t_{.50}$	$t_{.75}$	$t_{.80}$	$t_{.85}$	$t_{.90}$	$t_{.95}$	$t_{.975}$	$t_{.99}$	$t_{.995}$	$t_{.999}$	$t_{.9995}$
one-tail	**0.50**	**0.25**	**0.20**	**0.15**	**0.10**	**0.05**	**0.025**	**0.01**	**0.005**	**0.001**	**0.0005**
two-tails	**1.00**	**0.50**	**0.40**	**0.30**	**0.20**	**0.10**	**0.05**	**0.02**	**0.01**	**0.002**	**0.001**
df											
1	0.000	1.000	1.376	1.963	3.078	6.314	12.71	31.82	63.66	318.31	636.62
2	0.000	0.816	1.061	1.386	1.886	2.920	4.303	6.965	9.925	22.327	31.599
3	0.000	0.765	0.978	1.250	1.638	2.353	3.182	4.541	5.841	10.215	12.924
4	0.000	0.741	0.941	1.190	1.533	2.132	2.776	3.747	4.604	7.173	8.610
5	0.000	0.727	0.920	1.156	1.476	2.015	2.571	3.365	4.032	5.893	6.869
6	0.000	0.718	0.906	1.134	1.440	1.943	2.447	3.143	3.707	5.208	5.959
7	0.000	0.711	0.896	1.119	1.415	1.895	2.365	2.998	3.499	4.785	5.408
8	0.000	0.706	0.889	1.108	1.397	1.860	2.306	2.896	3.355	4.501	5.041
9	0.000	0.703	0.883	1.100	1.383	1.833	2.262	2.821	3.250	4.297	4.781
10	0.000	0.700	0.879	1.093	1.372	1.812	2.228	2.764	3.169	4.144	4.587
11	0.000	0.697	0.876	1.088	1.363	1.796	2.201	2.718	3.106	4.025	4.437
12	0.000	0.695	0.873	1.083	1.356	1.782	2.179	2.681	3.055	3.930	4.318
13	0.000	0.694	0.870	1.079	1.350	1.771	2.160	2.650	3.012	3.852	4.221
14	0.000	0.692	0.868	1.076	1.345	1.761	2.145	2.624	2.977	3.787	4.140
15	0.000	0.691	0.866	1.074	1.341	1.753	2.131	2.602	2.947	3.733	4.073
16	0.000	0.690	0.865	1.071	1.337	1.746	2.120	2.583	2.921	3.686	4.015
17	0.000	0.689	0.863	1.069	1.333	1.740	2.110	2.567	2.898	3.646	3.965
18	0.000	0.688	0.862	1.067	1.330	1.734	2.101	2.552	2.878	3.610	3.922
19	0.000	0.688	0.861	1.066	1.328	1.729	2.093	2.539	2.861	3.579	3.883
20	0.000	0.687	0.860	1.064	1.325	1.725	2.086	2.528	2.845	3.552	3.850
21	0.000	0.686	0.859	1.063	1.323	1.721	2.080	2.518	2.831	3.527	3.819
22	0.000	0.686	0.858	1.061	1.321	1.717	2.074	2.508	2.819	3.505	3.792
23	0.000	0.685	0.858	1.060	1.319	1.714	2.069	2.500	2.807	3.485	3.768
24	0.000	0.685	0.857	1.059	1.318	1.711	2.064	2.492	2.797	3.467	3.745
25	0.000	0.684	0.856	1.058	1.316	1.708	2.060	2.485	2.787	3.450	3.725
26	0.000	0.684	0.856	1.058	1.315	1.706	2.056	2.479	2.779	3.435	3.707
27	0.000	0.684	0.855	1.057	1.314	1.703	2.052	2.473	2.771	3.421	3.690
28	0.000	0.683	0.855	1.056	1.313	1.701	2.048	2.467	2.763	3.408	3.674
29	0.000	0.683	0.854	1.055	1.311	1.699	2.045	2.462	2.756	3.396	3.659
30	0.000	0.683	0.854	1.055	1.310	1.697	2.042	2.457	2.750	3.385	3.646
40	0.000	0.681	0.851	1.050	1.303	1.684	2.021	2.423	2.704	3.307	3.551
60	0.000	0.679	0.848	1.045	1.296	1.671	2.000	2.390	2.660	3.232	3.460
80	0.000	0.678	0.846	1.043	1.292	1.664	1.990	2.374	2.639	3.195	3.416
100	0.000	0.677	0.845	1.042	1.290	1.660	1.984	2.364	2.626	3.174	3.390
1000	0.000	0.675	0.842	1.037	1.282	1.646	1.962	2.330	2.581	3.098	3.300
z	0.000	0.674	0.842	1.036	1.282	1.645	1.960	2.326	2.576	3.090	3.291
	0%	50%	60%	70%	80%	90%	95%	98%	99%	99.8%	99.9%
	Confidence Level										

Derivation of the Confidence Interval

- **Knowing the OLS estimator is normally distributed, we can standardize it and say:**

$$\frac{\hat{\beta}_j - \beta_j}{\sigma_{\hat{\beta}_j}} \sim N(0,1) \qquad (16)$$

So with the standardization, $\hat{\beta}_j$ now has a standard normal distribution

- **Now, if we knew the true standard deviation of the sampling distribution ($\sigma_{\hat{\beta}_j}$) we'd be done and we could just use Table G.1 in Wooldridge to make precise statements about $\hat{\beta}_j$ taking certain values**
  - <u>But we don't… So we have to do a little more work</u>

Derivation of the Confidence Interval

- **With Assumptions 1 through 5 one can derive the following:**

$$\frac{\hat{\beta}_j - \beta_j}{\hat{\sigma}_{\hat{\beta}_j}} \sim t_{n-k-1} \qquad (17)$$

$$\hat{\sigma}_{\hat{\beta}_j} = \sqrt{\frac{\hat{\sigma}^2}{\sum_{i=1}^{N}\left(X_{ji} - \bar{X}\right)^2\left(1 - R_j^2\right)}} \qquad (18)$$

$$\hat{\sigma}^2 = \frac{1}{n-k-1} \times \sum_{i=1}^{N} (\hat{u}_i)^2 \qquad (19)$$

We have replaced the <u>true standard deviation</u> with the <u>estimated standard deviation</u> and conclude that the standardized OLS estimate is very close to normally distributed

Derivation of the Confidence Interval

$$\frac{\hat{\beta}_j - \beta_j}{\hat{\sigma}_{\hat{\beta}_j}} \sim t_{n-k-1} \quad (20)$$

Mean of $\hat{\beta}_j$ if OLS is unbiased

Estimated standard deviation of $\hat{\beta}_j$

- **This statistic is just the distance between our OLS estimate and the true value divided by our *estimated* standard deviation.**

- **This standardized distance has a specific <u>Student's t</u> distribution!**
  - Proof is available multiple places online

- **This is what we need to (finally!) construct confidence intervals**

Derivation of the Confidence Interval

- **We can use the tables of critical values for the Student's t distribution to figure out the combinations of confidence levels (α) and c that satisfy,**

$$Pr\left(-c \leq \frac{\hat{\beta}_j - \beta_j}{\hat{\sigma}_{\hat{\beta}_j}} \leq c\right) = 1 - \alpha \qquad (21)$$

  - Where $-c$ and c are the critical values and α is the confidence level (0.1, 0.05, or 0.01)

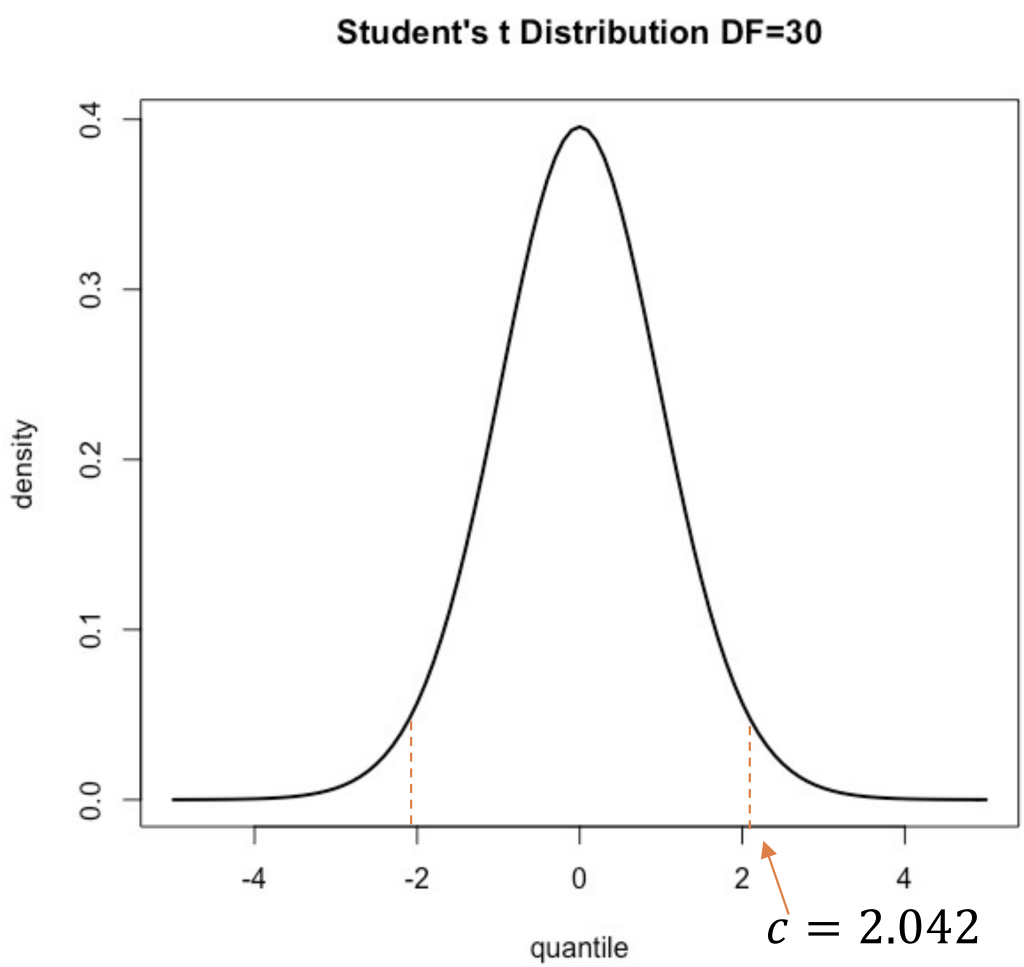
Derivation of the Confidence Interval

$$Pr\left(-c \leq \frac{\hat{\beta}_j - \beta_j}{\hat{\sigma}_{\hat{\beta}_j}} \leq c\right) = 1 - \alpha \qquad (21)$$

- **What is this saying in words?**
  - Suppose $\alpha = .05$. It's saying that there is a 95% probability that the standardized level of $\hat{\beta}_j$ exists betweeen $-c$ and c in the t distribution

Confidence Interval Graphical Representation

- Graphically, we can use t distribution tables to find the value of c such that an area under the distribution is equal to $\frac{\alpha}{2}$ to the right of c and to the left of $-c$



Final Derivation of the Confidence Interval

- **Start with,**

$$Pr\left(-c \leq \frac{\hat{\beta}_j - \beta_j}{\hat{\sigma}_{\hat{\beta}_j}} \leq c\right) = 1 - \alpha \qquad (22)$$

- **Rearrange as follows,**

$$\Rightarrow Pr\left(-c\hat{\sigma}_{\hat{\beta}_j} \leq \hat{\beta}_j - \beta_j \leq c\hat{\sigma}_{\hat{\beta}_j}\right) = 1 - \alpha \qquad (23)$$

$$\Rightarrow Pr\left(-c\hat{\sigma}_{\hat{\beta}_j} - \hat{\beta}_j \leq -\beta_j \leq c\hat{\sigma}_{\hat{\beta}_j} - \hat{\beta}_j\right) = 1 - \alpha \qquad (24)$$

$$\boxed{\Rightarrow Pr\left(\hat{\beta}_j - c\hat{\sigma}_{\hat{\beta}_j} \leq \beta_j \leq c\hat{\sigma}_{\hat{\beta}_j} + \hat{\beta}_j\right) = 1 - \alpha \qquad (25)}$$

What Does This Mean?

$$Pr\left(\underbrace{\hat{\beta}_j - c\hat{\sigma}_{\hat{\beta}_j}}_{\text{Lower Bound}} \leq \beta_j \leq \underbrace{c\hat{\sigma}_{\hat{\beta}_j} + \hat{\beta}_j}_{\text{Upper Bound}}\right) = 1 - \alpha \qquad (26)$$

- **This is saying that there is a $1-\alpha$ probability that β_j exists between $\hat{\beta}_j - c\hat{\sigma}_{\hat{\beta}_j}$ and $c\hat{\sigma}_{\hat{\beta}_j} + \hat{\beta}_j$**
- **Again, the value of α is up to the researcher, but 0.01, 0.05, and 0.1 are common**

What Does This Mean?

$$Pr\left(\underbrace{\hat{\beta}_j - c\hat{\sigma}_{\hat{\beta}_j}}_{\text{Lower Bound}} \leq \beta_j \leq \underbrace{c\hat{\sigma}_{\hat{\beta}_j} + \hat{\beta}_j}_{\text{Upper Bound}}\right) = 1 - \alpha \qquad (26)$$

- This is the formula for the confidence interval equal to $(1-\alpha)\times 100$

- If we choose the appropriate critical value from Student's t table, then an interval constructed in this way will contain the true value β_j in 95 percent of the samples we could have drawn (for $\alpha = 0.05$)

What Does This Mean?

$$Pr\left(\hat{\beta}_j - c\hat{\sigma}_{\hat{\beta}_j} \leq \beta_j \leq c\hat{\sigma}_{\hat{\beta}_j} + \hat{\beta}_j\right) = 1 - \alpha \qquad (27)$$

- **Well, it means that we have pretty high (though not complete) confidence that our true value is in this range (again assuming our OLS estimate is unbiased)**

- **If the interval is narrow, then we are more confident that our estimate is close to the true value**

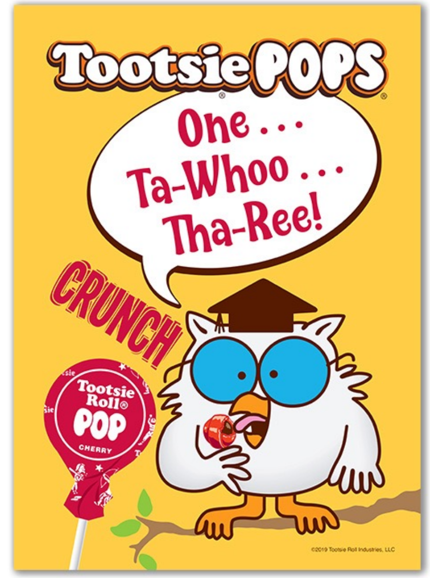
- **If the interval is wide, then we are less confident**
  - If the interval bounds zero, then we are much less confident because the relationship could be positive, negative, or zero! Yikes!

- **Recall that for this discussion we have assumed we have no bias in our estimation. If we don't then we have problems with this approach**

The Recipe for Calculating Confidence Intervals

Three steps!

1. **Choose α (common choices: 0.05 and 0.01)**
2. **Find c by going to Student's t table**
   - Choose the row relevant to the degrees of freedom ($DF = n - k - 1$), and find the c such that $Pr(t > c)$
3. **Construct the interval by plugging all this into the formula**



Let's Do It! Step 1

- **Choose α:**
  - The smaller the α the larger your CI
  - If $\alpha = 0.1$ our confidence that our CI contains our β_j is equal to $(1 - 0.1)\times 100 =$ **90 percent**
  - If $\boxed{\alpha = 0.05}$ our confidence that our CI contains our β_j is equal to $(1 - 0.05)\times 100 =$ **95 percent**
  - If $\alpha = 0.01$ our confidence that our CI contains our β_j is equal to $(1 - 0.01)\times 100 =$ **99 percent**

Let's Do It! Step 2

- **Find c by going to Student's t table**
  - Choose the row relevant to the degrees of freedom ($DF = n - k - 1$), and find the c such that $Pr(t > c)$
  - Be careful to read the labels of the table carefully
  - A good rule of thumb is that for a 95 percent confidence level and a large sample, where DF>50, c will be around 2
- **We'll say for this example that $\boxed{DF = 25}$**
  - We're interested in a "two-tailed" test

t Table

cum. prob	$t_{.50}$	$t_{.75}$	$t_{.80}$	$t_{.85}$	$t_{.90}$	$t_{.95}$	$t_{.975}$	$t_{.99}$	$t_{.995}$	$t_{.999}$	$t_{.9995}$
one-tail	0.50	0.25	0.20	0.15	0.10	0.05	0.025	0.01	0.005	0.001	0.0005
two-tails	1.00	0.50	0.40	0.30	0.20	0.10	0.05	0.02	0.01	0.002	0.001
df											
1	0.000	1.000	1.376	1.963	3.078	6.314	12.71	31.82	63.66	318.31	636.62
2	0.000	0.816	1.061	1.386	1.886	2.920	4.303	6.965	9.925	22.327	31.599
3	0.000	0.765	0.978	1.250	1.638	2.353	3.182	4.541	5.841	10.215	12.924
4	0.000	0.741	0.941	1.190	1.533	2.132	2.776	3.747	4.604	7.173	8.610
5	0.000	0.727	0.920	1.156	1.476	2.015	2.571	3.365	4.032	5.893	6.869
6	0.000	0.718	0.906	1.134	1.440	1.943	2.447	3.143	3.707	5.208	5.959
7	0.000	0.711	0.896	1.119	1.415	1.895	2.365	2.998	3.499	4.785	5.408
8	0.000	0.706	0.889	1.108	1.397	1.860	2.306	2.896	3.355	4.501	5.041
9	0.000	0.703	0.883	1.100	1.383	1.833	2.262	2.821	3.250	4.297	4.781
10	0.000	0.700	0.879	1.093	1.372	1.812	2.228	2.764	3.169	4.144	4.587
11	0.000	0.697	0.876	1.088	1.363	1.796	2.201	2.718	3.106	4.025	4.437
12	0.000	0.695	0.873	1.083	1.356	1.782	2.179	2.681	3.055	3.930	4.318
13	0.000	0.694	0.870	1.079	1.350	1.771	2.160	2.650	3.012	3.852	4.221
14	0.000	0.692	0.868	1.076	1.345	1.761	2.145	2.624	2.977	3.787	4.140
15	0.000	0.691	0.866	1.074	1.341	1.753	2.131	2.602	2.947	3.733	4.073
16	0.000	0.690	0.865	1.071	1.337	1.746	2.120	2.583	2.921	3.686	4.015
17	0.000	0.689	0.863	1.069	1.333	1.740	2.110	2.567	2.898	3.646	3.965
18	0.000	0.688	0.862	1.067	1.330	1.734	2.101	2.552	2.878	3.610	3.922
19	0.000	0.688	0.861	1.066	1.328	1.729	2.093	2.539	2.861	3.579	3.883
20	0.000	0.687	0.860	1.064	1.325	1.725	2.086	2.528	2.845	3.552	3.850
21	0.000	0.686	0.859	1.063	1.323	1.721	2.080	2.518	2.831	3.527	3.819
22	0.000	0.686	0.858	1.061	1.321	1.717	2.074	2.508	2.819	3.505	3.792
23	0.000	0.685	0.858	1.060	1.319	1.714	2.069	2.500	2.807	3.485	3.768
24	0.000	0.685	0.857	1.059	1.318	1.711	2.064	2.492	2.797	3.467	3.745
25	0.000	0.684	0.856	1.058	1.316	1.708	2.060	2.485	2.787	3.450	3.725
26	0.000	0.684	0.856	1.058	1.315	1.706	2.056	2.479	2.779	3.435	3.707
27	0.000	0.684	0.855	1.057	1.314	1.703	2.052	2.473	2.771	3.421	3.690
28	0.000	0.683	0.855	1.056	1.313	1.701	2.048	2.467	2.763	3.408	3.674
29	0.000	0.683	0.854	1.055	1.311	1.699	2.045	2.462	2.756	3.396	3.659
30	0.000	0.683	0.854	1.055	1.310	1.697	2.042	2.457	2.750	3.385	3.646
40	0.000	0.681	0.851	1.050	1.303	1.684	2.021	2.423	2.704	3.307	3.551
60	0.000	0.679	0.848	1.045	1.296	1.671	2.000	2.390	2.660	3.232	3.460
80	0.000	0.678	0.846	1.043	1.292	1.664	1.990	2.374	2.639	3.195	3.416
100	0.000	0.677	0.845	1.042	1.290	1.660	1.984	2.364	2.626	3.174	3.390
1000	0.000	0.675	0.842	1.037	1.282	1.646	1.962	2.330	2.581	3.098	3.300
z	0.000	0.674	0.842	1.036	1.282	1.645	1.960	2.326	2.576	3.090	3.291
	0%	50%	60%	70%	80%	90%	95%	98%	99%	99.8%	99.9%
	Confidence Level										

Let's Do It! Step 3

- **Construct the interval by plugging all this into the formula,**

$$Pr\left(\hat{\beta}_j - c\hat{\sigma}_{\hat{\beta}_j} \leq \beta_j \leq c\hat{\sigma}_{\hat{\beta}_j} + \hat{\beta}_j\right) = 1 - \alpha \tag{28}$$

- **So if $\boldsymbol{\widehat{\beta}_j = 1}$, $\boldsymbol{\widehat{\sigma}_{\widehat{\beta}_j} = 2}$, what is the confidence interval for $\boldsymbol{\beta_j}$?**

$$Pr\left(1 - c2 \leq \beta_j \leq c2 + 1\right) = 95\ Percent\ CI \tag{29}$$

$$Pr\left(1 - (2.06)2 \leq \beta_j \leq (2.06)2 + 1\right) = 95\ Percent\ CI \tag{30}$$

$$[-3.12, 5.12] = 95\ Percent\ CI \tag{31}$$

Recall Our Bivariate Wage Regression


```
. reg earn_wk educ

      Source |       SS           df       MS      Number of obs   =       741
-------------+----------------------------------   F(1, 739)       =     13.00
       Model |   818614972         1   818614972   Prob > F        =    0.0003
    Residual |  4.6527e+10       739    62959464   R-squared       =    0.0173
-------------+----------------------------------   Adj R-squared   =    0.0160
       Total |  4.7346e+10       740  63980620.1   Root MSE        =    7934.7

------------------------------------------------------------------------------
     earn_wk |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |   364.1439   100.9866     3.61   0.000     165.8891    562.3987
       _cons |   2.125245   768.8046     0.00   0.998    -1507.176    1511.426
------------------------------------------------------------------------------
```

- $\hat{\beta}_1 = 364.144, \hat{\sigma}_{\hat{\beta}_1} = 100.987, \boxed{\alpha = 0.05}, DF = 741 - 2 = 739$

$\Longrightarrow c = 1.96$

$CI = [364.144 - 1.96\times 100.987, 364.144 + 1.96\times 100.987] = \underbrace{[165.889, 562.399]}$

Interpretation of a Confidence Interval

- **Consider this 95 percent confidence interval,**

$$CI = [165.889, 562.399]$$

- **What does the upper bound suggest?**
  - Suggests that one more year of education is associated with an increase *as big as* 562.4 USD per week in earnings, *within our range of confidence*
- **What does the lower bound suggest?**
  - Suggests that one more year of education could be associated with an increase *as little as* 165.89 USD per week in earnings, *within our range of confidence*

Interpretation of a Confidence Interval

- **So we are 95 percent confident that the true relationship lies in the interval $[\mathbf{165.889, 562.399}]$**
  - This type of statement is very powerful
- **But what if the interval were $[\mathbf{-165.889, 562.399}]$?**
  - Why does this lessen the power of our statement about or estimate of β_1?

Part 2:
Hypothesis Testing

Hypothesis Testing Motivation

- **Hypothesis testing is a highly related subject to CI estimation**
- **Consider you've estimated the following model:**

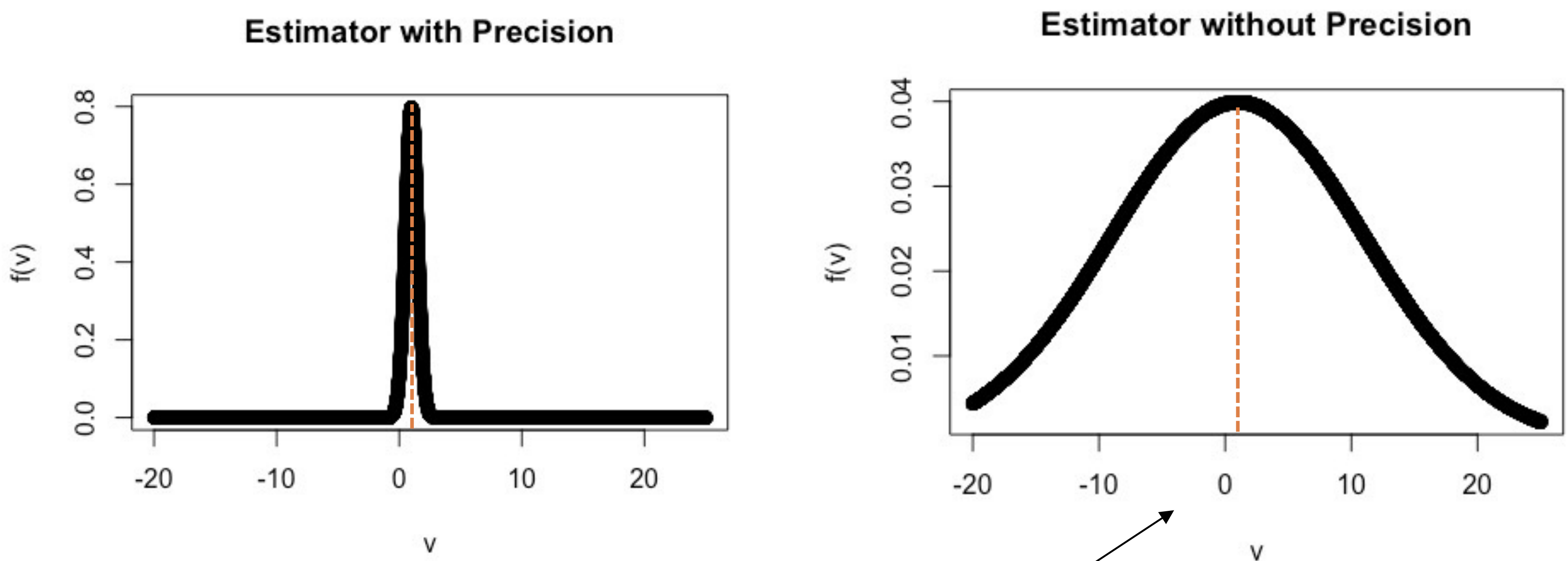
$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots \beta_k X_{ki} + u_i \qquad (32)$$

- **Suppose, $\widehat{\boldsymbol{\beta}}_1 = \mathbf{0.8}$ and $\widehat{\boldsymbol{\sigma}}^2_{\widehat{\boldsymbol{\beta}}_1} = \mathbf{0.35}$**
- **Based on this, we want to test the hypothesis that $\boldsymbol{\beta}_1 = \mathbf{1}$**
  - How can we go about assessing whether this hypothesis is correct?

Hypothesis Testing Motivation

- **Does $\widehat{\boldsymbol{\beta}}_{\mathbf{1}} = \mathbf{0.8} \implies \boldsymbol{\beta}_{\mathbf{1}} \neq \mathbf{1}$?**
  - No! Even if $\beta_1 = 1$, an unbiased and efficient estimate is unlikely to to equal 1. The more sampling variance the more likely this is. Why?

Example Illustration



If you're drawing from a distribution of betas with wider variance, it's more likely you'll measure beta that's not close to the true relationship

Hypothesis Testing Motivation

- **Does $\widehat{\boldsymbol{\beta}}_{\mathbf{1}} = \mathbf{0.8} \implies \boldsymbol{\beta}_{\mathbf{1}} \neq \mathbf{1}$?**
  - No! Even if $\beta_1 = 1$, an unbiased and efficient estimate is unlikely to to equal 1. The more sampling variance the more likely this is. Why?

- **Let's say in the context of the study that $\mathbf{0.8} \approx \mathbf{1}$. Does this mean the hypothesis is true?**
  - No! Even if the true value is substantially different from 1, our unbiased and efficient estimate has some chance of providing us an estimate close to 1

Hypothesis Testing Motivation

- **Hypothesis Testing provides a systematic and data-driven approach to answering these types of questions**
- **It allows us to conclude whether the data are <u>consistent with the hypothesis</u> (in this last example the hypothesis is that in fact $\beta_1 = 1$) or lead us to <u>reject it</u>**
  - Note that "consistent with the hypothesis" is a much weaker statement than saying the hypothesis is true
    - Note the asymmetry here. Statements rejecting the hypothesis are much stronger than the data being consistent

Intuition of Hypothesis Testing

- Okay, we use OLS to get $\widehat{\beta}_j$, and we still don't know β_j
- Given the $\widehat{\beta}_j$ we got, and our understanding of its distribution, can we reject that notion that $\beta_j = \beta_j^*$

What We Need for Hypothesis Testing

- **Just like CI estimation, we need the following,**
  - We need to know the distribution of $\hat{\beta}_j$
  - We need Assumptions 1 through 5 to hold

- **Recall that all this allows us to say the following,**

$$\frac{\hat{\beta}_j - \beta_j}{\hat{\sigma}_{\hat{\beta}_j}} \sim t_{n-k-1}, where \tag{33}$$

$$\hat{\sigma}_{\hat{\beta}_j} = \sqrt{\frac{\hat{\sigma}^2}{\sum_{i=1}^{N}\left(X_{ji} - \bar{X}\right)^2\left(1 - R_j^2\right)}} \tag{34}$$

$$\hat{\sigma}^2 = \frac{1}{n-k-1} \times \sum_{i=1}^{N}(\hat{u}_i)^2 \tag{35}$$

Derivation of Hypothesis Test

- **We'll take equation (33),**

$$\frac{\hat{\beta}_j - \beta_j}{\hat{\sigma}_{\hat{\beta}_j}} \sim t_{n-k-1} \tag{33}$$

- **This result tells us that our OLS estimator has a distribution that is very close to a normal distribution with a mean equal to the true value, variance equal to the estimated variance, but…**
- **We don't know what the true value is. So this tells us how tall and skinny the sampling distribution is, but not where it is located**
- **Now, we'll just take $\boldsymbol{\beta_j}$ and replace it with our hypothesized value $\boldsymbol{\beta_j^*}$. If Assumptions 1-5 hold then the resulting "t-statistic" has the following distribution,**

$$\frac{\hat{\beta}_j - \beta_j^*}{\hat{\sigma}_{\hat{\beta}_j}} \sim t_{n-k-1} \tag{36}$$

Stating the Null and Alternative Hypothesis

- **Here we can state the general forms of the Null Hypothesis $(\boldsymbol{H_0})$ and Alternative Hypothesis $(\boldsymbol{H_A})$**
  - $H_0: \beta_j = \beta_j^*, and$
  - $H_A: \beta_j \neq \beta_j^*$

Derivation of Hypothesis Test

$$\text{“}t\text{”} = \frac{\hat{\beta}_j - \beta_j^*}{\hat{\sigma}_{\hat{\beta}_j}} \sim t_{n-k-1} \quad (36)$$

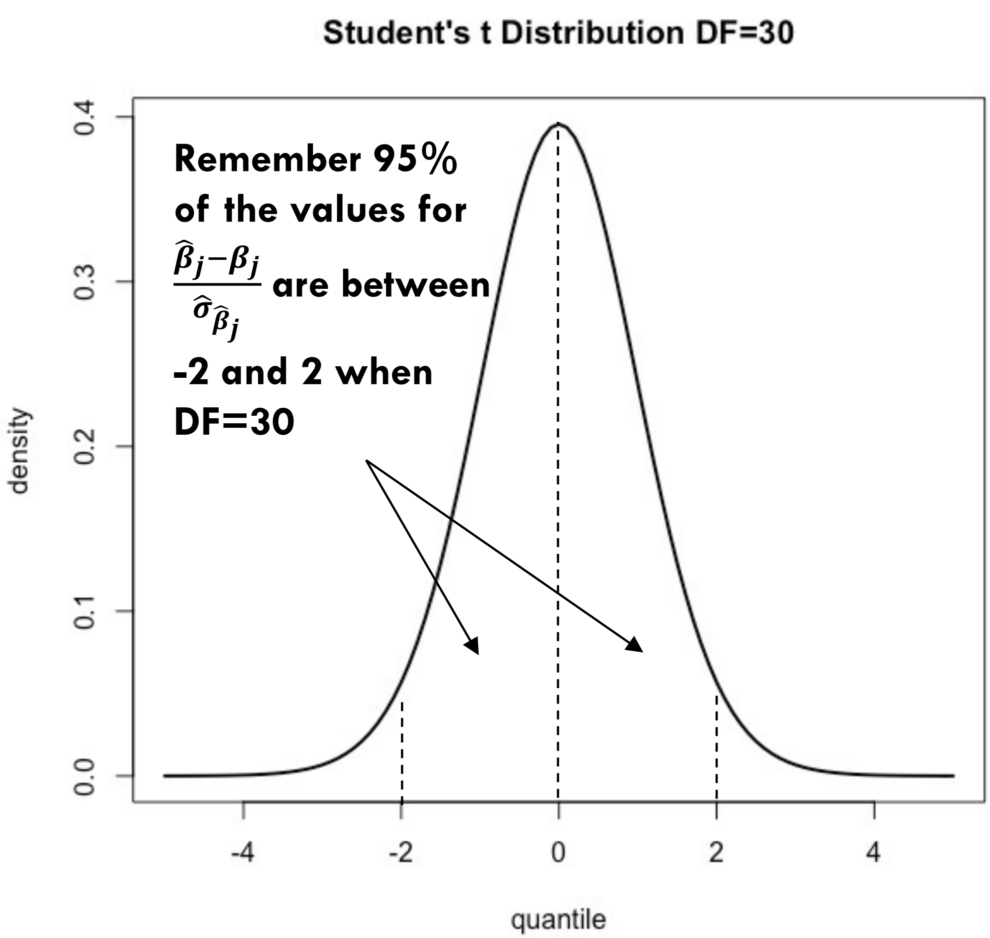
- **This is a statistic we can calculate after running an OLS regression**
- **The numerator is the distance between our OLS estimate and the hypothesized value**
- **The denominator is the estimated standard error of our OLS estimates**
- **This result tells us that if the null hypothesis (H_0: $\beta_j = \beta_j^*$) is true, then we not only know how tall and skinny the sampling distribution is but we also know where its peak should be (over the hypothesized value β_j^*)**

Derivation of Hypothesis Test

$$\text{“}t\text{”} = \frac{\hat{\beta}_j - \beta_j^*}{\hat{\sigma}_{\hat{\beta}_j}} \sim t_{n-k-1} \tag{36}$$

- **If $\boldsymbol{\widehat{\beta}_j = \beta_j^*}$ then the numerator will be 0 and the t-statistic = = 0**
- **This means that**
  - If our estimate is close to the hypothesized value, the t-statistic should just be 0
  - If our estimate is far away from the hypothesized value, the t-statistic should be large (in magnitude) and far away and in the tails on the t-distribution

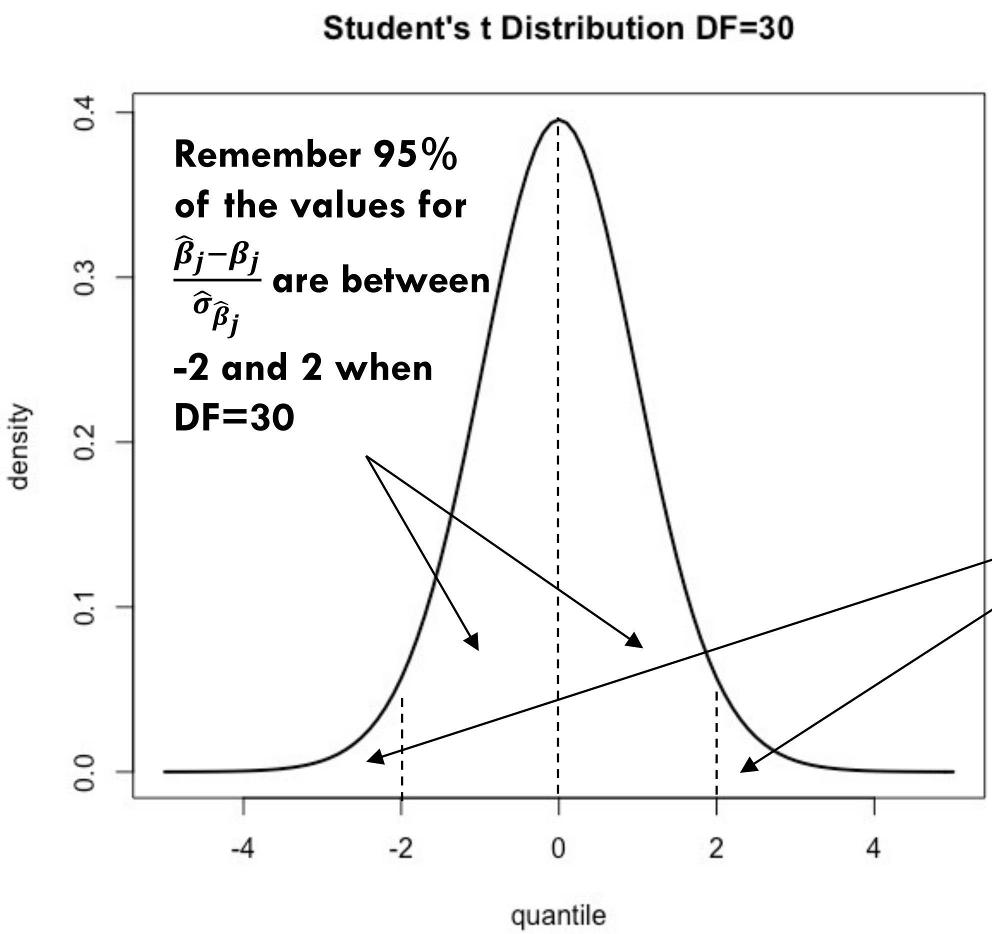
How to Use this Result



$$“t” = \frac{\hat{\beta}_j-\beta_j^*}{\hat{\sigma}_{\hat{\beta}_j}} \sim t_{n-k-1} \qquad (36)$$

- **If our hypothesis is correct it should have a Student's t distribution**
- **It is highly unlikely that we will get a t-statistic greater than 2 or less than -2 if the hypothesis is true with a reasonably large DF**

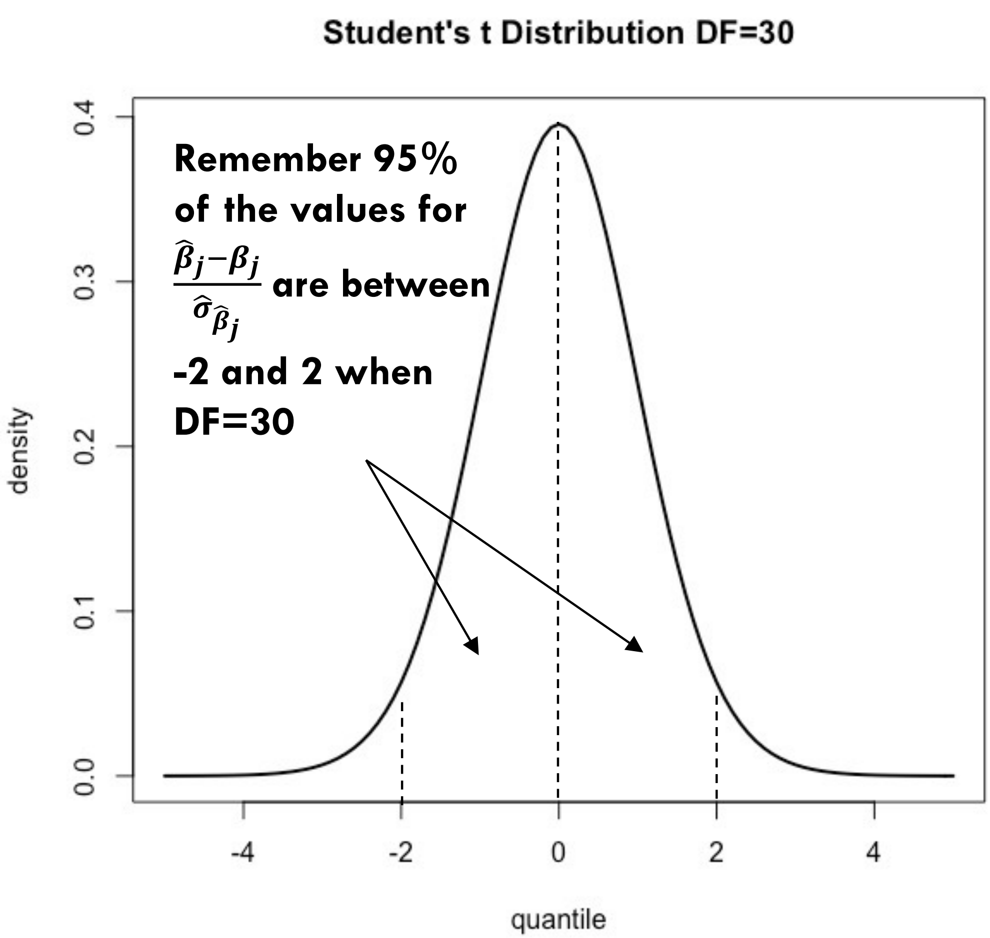
How to Use this Result



$$\text{“}t\text{”} = \frac{\hat{\beta}_j - \beta_j^*}{\hat{\sigma}_{\hat{\beta}_j}} \sim t_{n-k-1} \qquad (36)$$

- **So if $\boldsymbol{\beta_j^*}$ is really different from $\boldsymbol{\hat{\beta}_j}$, then $\boldsymbol{t}$ is very different from zero. If absolute value is greater than 2**
  - $\Longrightarrow$ It's very unlikely to be this big if $\beta_j^* = \hat{\beta}_j$. So we reject the hypothesis that $\beta_j^* = \hat{\beta}_j$ (with 95% certainty)

How to Use this Result



$$\text{“}t\text{”} = \frac{\hat{\beta}_j - \beta_j^*}{\hat{\sigma}_{\hat{\beta}_j}} \sim t_{n-k-1} \qquad (36)$$

- **So if $\boldsymbol{\beta_j^*}$ is relatively close to $\boldsymbol{\hat{\beta}_j}$, then $\boldsymbol{t}$ is relatively close to 0**
  - $\Longrightarrow$ Conclude that we don't have sufficient evidence to reject the idea that $\beta_j^* \neq \hat{\beta}_j$

How to Use this Result

$$“t” = \frac{\hat{\beta}_j - \beta_j^*}{\hat{\sigma}_{\hat{\beta}_j}} \sim t_{n-k-1} \qquad (36)$$

- **More specifically, there is a 95 percent chance that a t random variable (with a high degrees of freedom) will take values between -2 and 2 (-1.96 and 1.96 with DF>1000)**
- **This means that if the hypothesis is true (remember that $\boldsymbol{\beta_j = \beta_j^*}$), in 95 percent of the samples we could have drawn from the true model, the statistic we get by taking the difference between the OLS estimate and the hypothesized value and dividing it by the standard error will be between -2 and 2**
- **The takeaway: Absolute values of t greater than 2 or less than 2 ($|t| > 2$) are highly unlikely if the hypothesis is true**
  - Recall “2” is only an approximation. You need t table to to get the actual threshold in your case.

How to Use this Result

$$\text{“}t\text{”} = \frac{\hat{\beta}_j - \beta_j^*}{\hat{\sigma}_{\hat{\beta}_j}} \sim t_{n-k-1} \tag{36}$$

- **If we do get a large absolute value for the t-statistic we can conclude:**

“Well… if the hypothesis really were true, this t-statistic value would be highly unlikely. Since I got this number, I’m going to conclude with reasonable confidence (but not absolute certainty) that the hypothesis must be wrong”

How to Use this Result

$$“t” = \frac{\hat{\beta}_j - \beta_j^*}{\hat{\sigma}_{\hat{\beta}_j}} \sim t_{n-k-1} \quad (36)$$

- **Does it make sense that this would be useful in making concise statements about our findings and our confidence in them? Yes.**
  - The numerator is the difference between our finding from OLS and our hypothesized value. The bigger it is, the bigger the t-statistic and the more confident we are that $\beta_j \neq \beta_j^*$
  - The denominator is the estimated standard error of our regression coefficient. The larger this is the less precise our estimator and the larger the variance in our sampling distribution and the wider the range of possible values $\hat{\beta}_j$ can take if the hypothesis is true
    - Recall that less noise in the data and more data will reduce the standard error and make it more likely that you'll be able to reject the hypothesis

Steps to Hypothesis Testing (The Recipe)

- **The "Classical" Approach to hypothesis testing has <u>6 steps</u>:**
  1. State the null hypothesis
  2. State the alternative hypothesis
  3. Choose a confidence level (significance level)
  4. Use the t table to find critical values
  5. Calculate the t statistic
  6. Apply the rules for conclusions

Steps to Hypothesis Testing

- **Step 1: State the null hypothesis**
  - This is a statement about the true value of our parameter of interest in our model. For example,

$$ln(Wage_i) = \beta_0 + \beta_1 Educ_i + \beta_2 Exp_i + u_i \qquad (37)$$

$$H_0: \beta_1 = 3 \qquad (38)$$

- **Our objective: Figure out whether or not the data are consistent with the hypothesis that $\boldsymbol{\beta_1 = 3}$**

Steps to Hypothesis Testing

- **Step 2: Clarify the nature of our interest in this hypothesis by specifying an alternative hypothesis**
  - This should embody the ways in which the null hypothesis would be false
- **Here we can state this several possible ways:**
  1. $H_A\colon \beta_1 \neq 3$ — Leads to a selection of a "2-tailed" test (most common)
  2. $H_A\colon \beta_1 > 3$
  3. $H_A\colon \beta_1 < 3$

  Either one leads to a selection of a "1-tailed" test

Steps to Hypothesis Testing

- **Step 3: Choose a confidence/significance level, (denoted usually as α in textbooks)**
  - α indicates how "unlikely" a value for our t-statistic must be in order for us to conclude that our data are inconsistent with the null hypothesis
  - Common choices are 0.1, 0.05, and 0.01
  - $\alpha = 0.05$ means that we will reject the null hypothesis if we get a value of our t-statistic that would happen in only 5 percent or fewer of the samples we could have drawn from the true model if this null hypothesis were correct
    - Means a 95% confidence level

Steps to Hypothesis Testing

- **Step 3: Choose a confidence/significance level, (denoted usually as α in textbooks)**
  - As our choice of α gets smaller and smaller we are employing a more and more stringent criteria for what "unlikely" is
  - As α gets smaller it will be harder and harder to reject the null hypothesis. <u>But when we do reject the null hypothesis we'll do it with more confidence</u>
  - If we can reject the null hypothesis even with a tiny α ($\alpha \leq 0.01$) then our data provide very strong evidence against the null hypothesis

Steps to Hypothesis Testing

- **Step 4: Use a table of Student's t distribution to find critical values**
  - The critical values are the values that we will compare our t-statistic to in deciding whether or not our t-statistic is highly unlikely and the null should be rejected
  - All this requires is knowing our α, the degrees of freedom in our regression, and whether we are interested in a 1-tailed (upper or lower) or a 2-tailed test
    - For a "2-tailed" test, we find c so that: $Pr(t_{n-k-1} > c) = \frac{\alpha}{2}$
    - For a "upper-tailed" test, we find c so that: $Pr(t_{n-k-1} > c) = \alpha$
    - For a "lower-tailed" test, we find c so that: $Pr(t_{n-k-1} > c) = \alpha$
      - But for the lower-tailed test, we use $-c$ as our critical value

t Table

cum. prob	$t_{.50}$	$t_{.75}$	$t_{.80}$	$t_{.85}$	$t_{.90}$	$t_{.95}$	$t_{.975}$	$t_{.99}$	$t_{.995}$	$t_{.999}$	$t_{.9995}$
one-tail	**0.50**	**0.25**	**0.20**	**0.15**	**0.10**	**0.05**	**0.025**	**0.01**	**0.005**	**0.001**	**0.0005**
two-tails	**1.00**	**0.50**	**0.40**	**0.30**	**0.20**	**0.10**	**0.05**	**0.02**	**0.01**	**0.002**	**0.001**
df											
1	0.000	1.000	1.376	1.963	3.078	6.314	12.71	31.82	63.66	318.31	636.62
2	0.000	0.816	1.061	1.386	1.886	2.920	4.303	6.965	9.925	22.327	31.599
3	0.000	0.765	0.978	1.250	1.638	2.353	3.182	4.541	5.841	10.215	12.924
4	0.000	0.741	0.941	1.190	1.533	2.132	2.776	3.747	4.604	7.173	8.610
5	0.000	0.727	0.920	1.156	1.476	2.015	2.571	3.365	4.032	5.893	6.869
6	0.000	0.718	0.906	1.134	1.440	1.943	2.447	3.143	3.707	5.208	5.959
7	0.000	0.711	0.896	1.119	1.415	1.895	2.365	2.998	3.499	4.785	5.408
8	0.000	0.706	0.889	1.108	1.397	1.860	2.306	2.896	3.355	4.501	5.041
9	0.000	0.703	0.883	1.100	1.383	1.833	2.262	2.821	3.250	4.297	4.781
10	0.000	0.700	0.879	1.093	1.372	1.812	2.228	2.764	3.169	4.144	4.587
11	0.000	0.697	0.876	1.088	1.363	1.796	2.201	2.718	3.106	4.025	4.437
12	0.000	0.695	0.873	1.083	1.356	1.782	2.179	2.681	3.055	3.930	4.318
13	0.000	0.694	0.870	1.079	1.350	1.771	2.160	2.650	3.012	3.852	4.221
14	0.000	0.692	0.868	1.076	1.345	1.761	2.145	2.624	2.977	3.787	4.140
15	0.000	0.691	0.866	1.074	1.341	1.753	2.131	2.602	2.947	3.733	4.073
16	0.000	0.690	0.865	1.071	1.337	1.746	2.120	2.583	2.921	3.686	4.015
17	0.000	0.689	0.863	1.069	1.333	1.740	2.110	2.567	2.898	3.646	3.965
18	0.000	0.688	0.862	1.067	1.330	1.734	2.101	2.552	2.878	3.610	3.922
19	0.000	0.688	0.861	1.066	1.328	1.729	2.093	2.539	2.861	3.579	3.883
20	0.000	0.687	0.860	1.064	1.325	1.725	2.086	2.528	2.845	3.552	3.850
21	0.000	0.686	0.859	1.063	1.323	1.721	2.080	2.518	2.831	3.527	3.819
22	0.000	0.686	0.858	1.061	1.321	1.717	2.074	2.508	2.819	3.505	3.792
23	0.000	0.685	0.858	1.060	1.319	1.714	2.069	2.500	2.807	3.485	3.768
24	0.000	0.685	0.857	1.059	1.318	1.711	2.064	2.492	2.797	3.467	3.745
25	0.000	0.684	0.856	1.058	1.316	1.708	2.060	2.485	2.787	3.450	3.725
26	0.000	0.684	0.856	1.058	1.315	1.706	2.056	2.479	2.779	3.435	3.707
27	0.000	0.684	0.855	1.057	1.314	1.703	2.052	2.473	2.771	3.421	3.690
28	0.000	0.683	0.855	1.056	1.313	1.701	2.048	2.467	2.763	3.408	3.674
29	0.000	0.683	0.854	1.055	1.311	1.699	2.045	2.462	2.756	3.396	3.659
30	0.000	0.683	0.854	1.055	1.310	1.697	2.042	2.457	2.750	3.385	3.646
40	0.000	0.681	0.851	1.050	1.303	1.684	2.021	2.423	2.704	3.307	3.551
60	0.000	0.679	0.848	1.045	1.296	1.671	2.000	2.390	2.660	3.232	3.460
80	0.000	0.678	0.846	1.043	1.292	1.664	1.990	2.374	2.639	3.195	3.416
100	0.000	0.677	0.845	1.042	1.290	1.660	1.984	2.364	2.626	3.174	3.390
1000	0.000	0.675	0.842	1.037	1.282	1.646	1.962	2.330	2.581	3.098	3.300
z	0.000	0.674	0.842	1.036	1.282	1.645	1.960	2.326	2.576	3.090	3.291
	0%	50%	60%	70%	80%	90%	95%	98%	99%	99.8%	99.9%
	Confidence Level										

Steps to Hypothesis Testing

- **Step 5: Calculate the t-statistic:**

$$\text{“}t\text{”} = \frac{\hat{\beta}_j - \beta_j^*}{\hat{\sigma}_{\hat{\beta}_j}} \sim t_{n-k-1} \tag{39}$$

- **This requires two pieces of information from the Stata printout. Which ones?**

```
      Source |       SS           df       MS      Number of obs   =       741
-------------+----------------------------------   F(1, 739)       =     13.00
       Model |   818614972         1   818614972   Prob > F        =    0.0003
    Residual |  4.6527e+10       739    62959464   R-squared       =    0.0173
-------------+----------------------------------   Adj R-squared   =    0.0160
       Total |  4.7346e+10       740  63980620.1   Root MSE        =    7934.7

------------------------------------------------------------------------------
     earn_wk |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |   364.1439   100.9866     3.61   0.000     165.8891    562.3987
       _cons |   2.125245   768.8046     0.00   0.998    -1507.176    1511.426
------------------------------------------------------------------------------
```

Steps to Hypothesis Testing

- **Step 6: Apply the appropriate one of the following rules to reach a conclusion:**
  - If $H_0: \beta_1 = 3$, then
    1. $H_A: \beta_1 \neq 3$, we reject H_0 if $t < -c$ or $t > c$
    2. $H_A: \beta_1 > 3$, we reject H_0 if $t > c$
    3. $H_A: \beta_1 < 3$, we reject H_0 if $t < c$
  - If we do not reject, the conclusion is we **"fail to reject."** In other words, we conclude that our data are consistent with the null hypothesis, but **we can't claim to have proven that the null hypothesis is true**
    - Doing this is a classic novice mistake

That's it!

How to Formulate Your Hypothesis

- **Once you've estimated your regression and you're confident that you have minimized bias, you can formulate your Null and Alternative hypotheses:**
  - **Null hypotheses** are always simple. They hypothesize that your parameter of interest is some specific value
    - So they always involve an equality rather than an inequality
  - **Alternative hypotheses** are always composites. They allow the parameter to take some range
    - So they always involve inequalities

The "Zero Null Hypothesis"

- **The most common hypothesis test regarding OLS coefficients is the following,**

$$H_0: \beta_j = 0 \tag{40}$$
$$H_A: \beta_j \neq 0 \tag{41}$$

- **Why is this an interesting test in many circumstances?**
  - If we reject the null, then we can say our estimate is evidence that the effect of X_j on Y is not zero
    - This is a deceptively powerful statement

The Zero Null Hypothesis

- **How do we perform this test? Easy:**

$$t = \frac{\hat{\beta}_j - \beta_j^*}{\hat{\sigma}_{\hat{\beta}_j}} = \frac{\hat{\beta}_j - 0}{\hat{\sigma}_{\hat{\beta}_j}} = \frac{\hat{\beta}_j}{\hat{\sigma}_{\hat{\beta}_j}} \tag{42}$$

- $\frac{\hat{\beta}_j}{\hat{\sigma}_{\hat{\beta}_j}}$ **is often called the "t-ratio." It is the default t-statistic reported by Stata**

Steps to Hypothesis Testing

- **Step 5: Calculate the t-statistic:**

$$\text{“}t\text{”} = \frac{\hat{\beta}_j - \beta_j^*}{\hat{\sigma}_{\hat{\beta}_j}} \sim t_{n-k-1} \qquad (43)$$

- **This requires two pieces of information from the Stata printout. Which ones?**

```
      Source |       SS           df       MS      Number of obs   =       741
-------------+----------------------------------   F(1, 739)       =     13.00
       Model |   818614972         1   818614972   Prob > F        =    0.0003
    Residual |  4.6527e+10       739    62959464   R-squared       =    0.0173
-------------+----------------------------------   Adj R-squared   =    0.0160
       Total |  4.7346e+10       740  63980620.1   Root MSE        =    7934.7

------------------------------------------------------------------------------
     earn_wk |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |   364.1439   100.9866     3.61   0.000     165.8891    562.3987
       _cons |   2.125245   768.8046     0.00   0.998    -1507.176    1511.426
------------------------------------------------------------------------------
```

The Zero Null Hypothesis

- **If the value of the t-ratio is high or low enough that we would reject the null hypothesis, this coefficient is said to be "<u>statistically significant</u>"**
  - This is a term that many people use without really understanding it
  - A more accurate phrasing would be "statistically significantly different from zero"
  - With a large number of degrees of freedom (> 1000), if $|t - ratio| > 1.96$, then the finding is "statistically significant"

A Word of Caution

- **If you "fail to reject" $H_0\!: \beta_j = 0$, <u>it does not mean that the true value is zero!</u>**
  - As I mentioned, mixing this up is a classic novice mistake
  - It just means that your data don't allow you to state with high confidence that the relationship isn't zero. This is a very different conclusion
  - As said before, "failing to reject" the null is a much less conclusive statement than rejecting the null
    - A good and accepted phrasing is that your estimate is not "measurably different from zero"

Other Null Hypotheses

- **Often theory either assumes or predicts a specific non-zero relationship between variables**
  - For example, if a theory predicts that two things Y and X should always adjust to be equal to each other then you could run a regression of Y on X and test the hypothesis that the coefficient is 1

- **Sometimes you may have a question about how much or whether related phenomena are the same or different in how they affect Y**
  - For example, you may be interested in the effect a mine opening on employment of rural people in rural sub-Saharan Africa. You may want to compare that effect of a mine opened by a foreign company to the known effect of a mine opened by a domestic mining company and see if they are statistically different from one another

Other Null Hypotheses

- **Sometimes you'll look at previous studies that conclude some relationship between X and Y equal to $\widehat{\gamma}_j$. Perhaps you see flaws in their studies that you think leads to bias in this consensus and can be rectified with new data you have**
  - In this case you can estimate your $\hat{\beta}_j$ and test whether it is statistically different from the value advanced by this consensus
  - What would be the null and alternative hypotheses in this case?
    - $H_0: \beta_j = \hat{\gamma}_j, H_A: \beta_j \neq \hat{\gamma}_j$

Economic Versus Statistical Significance

- **A common weakness in research is to focus too narrowly on statistical significance (of differences from zero)**
  - An estimate can be statistically significant in that the researcher is confident that the effect of X on Y is not zero, but still so small that the effect is economically unimportant
  - Similarly, an estimate can be statistically insignificant, but the estimated effect could be very large. In this case, perhaps the data are just too weak or noisy to provide a ton of confidence in this estimate, but point to a large effect!
- **Good empirical research will always discuss the magnitude of the estimate and what it means as well as statistical significance**
  - This is why you worked so hard early on to learn how to interpret coefficients

Part 3:
Joint Hypothesis Testing

Joint Hypothesis Testing

- Often we are interested in testing hypotheses that involve multiple coefficients
- These tests are called "Joint Hypothesis Tests"
- This involves a standard multiple regression model with many RHS variables and an idea about more than one coefficient
- Examples (Next Slide)

Joint Hypothesis Testing

- **There are four common types of Joint Hypothesis Tests:**
  1. Multiple zeros
  2. All slopes equal to zero
  3. Equality of coefficients
  4. Coefficients sum to a specific value

- **I'll explain each one**

Multiple Zeros

- **Suppose you are estimating the basic multivariate regression model,**

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_k X_{ki} + u_i. \tag{47}$$

- **The multiple zeros hypothesis would be,**

$$H_0: \beta_1 = 0 \; and \; \beta_2 = 0 \tag{48}$$

  - This is different from checking separately whether $\beta_1 = 0$ and then $\beta_2 = 0$
    - Why? Sometimes data will not be able to tell precisely about the individual values of the betas but will be able to tell us that jointly the two effects are important (or not)

Multiple Zeros Example

- In a study of wage outcomes, we might want to test the hypothesis that the slope coefficients are the same for men and women
- Consider the regression,

$$Wage_i$$

$$= \beta_0 + \beta_1 Female_i + \beta_2 Educ_i Female_i + \beta_3 Educ_i$$

$$+ \beta_4 Exp_i Female_i + \beta_5 Exp_i + u_i \qquad (49)$$

- What would be the multiple zeros null hypothesis be here?

All Slopes Equal to Zero

- **Going back to our model,**

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_k X_{Ki} + u_i. \qquad (50)$$

- **The "all slopes equal to zero" hypothesis would be**

$$H_0 \colon \beta_1 = 0, \beta_1 = 0 \ldots \beta_k = 0 \qquad (51)$$

- **This is testing the hypothesis that none of the RHS variables matter in determining Y**
  - This is called a test of "overall significance"
- **If the null is not rejected then knowing values of Xs doesn't help predict Y at all**

Equality of Coefficients

- **Here the null hypothesis is,**

$$H_0: \beta_1 = \beta_2 \qquad (52)$$

- **We are testing for whether the coefficients for two (or more) variables are the same**
- **Why might we be interested in this?**
  - Well we might have different but similar things we are measuring. Example: The effects of a factory coming to your town might differ by domestic versus foreign ownership

Coefficients Sum to a Specific Value

- **Here the null hypothesis is,**

$$H_0: \beta_1 + \beta_2 = W \tag{53}$$

Where W is some constant

- **Can you think of an example of where we might use this example?**
  - One example: Production function estimation w/Cobb-Douglas Functional Form
  - Coefficients should sum to 1

Alternative Hypothesis in Joint Significance Tests

- **So far I've only outlined the null hypotheses for the four joint significance test we'll look at**
  - But to proceed with the tests, we need to outline the alternative hypotheses as well
- **In general, the alternative hypotheses for these four tests is of the form, "<u>at least one thing in the null hypothesis is not correct</u>"**

The Alternative Hypotheses

Type of Hypothesis	Joint Null Hypothesis	Alternative Hypothesis
Multiple Zeros	$H_0: \beta_1 = 0\ and\ \beta_2 = 0$	$H_A:\ \beta_1 \neq or\ \beta_2 \neq 0$
All Slopes Equal to Zero	$H_0: \beta_1 = 0, \beta_1 = 0 \ldots\ \beta_k = 0$	$H_A:\ \beta_1 \neq or\ \beta_2 \neq 0\ \ldots or \beta_k \neq 0$ (or some combination)
Equality of Coefficients	$H_0: \beta_1 = \beta_2$	$H_A: \beta_1 \neq \beta_2$
Coefficients Sum to a Specific Value	$H_0: \beta_1 + \beta_2 = W$	$H_A: \beta_1 + \beta_2 \neq W$

Implementing the Tests

- **To construct Joint Hypothesis Tests, you need to run two regressions:**
  1. The **restricted regression:** the regression you would run if you knew for sure the null hypothesis is true
  2. The **unrestricted regression**: The more general regression that allows for the possibility that either the null or the alternative hypothesis might be true
     - This is usually the regression that you start with and you have questions about
- **<u>You examine whether the unrestricted regression fits the data better than the restricted regression. If it does, then you reject the null hypothesis</u>**

The Unrestricted Regression

- **Let's take the general model we used before,**

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_k X_{ki} + u_i \qquad (54)$$

- **This is the unrestricted regression we must run if we aren't sure that the null hypothesis is true**
- **To get the restricted regression from this, you:**
  1. Start with the unrestricted regression
  2. Plug in the null hypothesis
  3. Re-arrange the resulting equation to get something that looks like a regression we can run

Example with the Multiple Zeros Hypothesis

- **The unrestricted regression is,**

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i} + u_i \quad (55)$$

- $\boldsymbol{H_0: \beta_1 = 0 \; and \; \beta_2 = 0}$
- **Plugging in $\boldsymbol{H_0}$,**

$$Y_i = \beta_0 + (0)X_{1i} + (0)X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i} + u_i \qquad (56)$$

$$\Longrightarrow \boxed{Y_i = \beta_0 + \beta_3 X_{3i} + \beta_4 X_{4i} + u_i} \qquad (57)$$

Example with the All Zeros Hypothesis

- **The unrestricted regression is,**

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i} + u_i \quad (58)$$

- $\boldsymbol{H_0: \beta_1 = 0, \beta_2 = 0, \beta_3 = 0, \beta_4 = 0}$
- **Plugging in** $\boldsymbol{H_0}$**,**

$$Y_i = \beta_0 + (0)X_{1i} + (0)X_{2i} + (0)X_{3i} + (0)X_{4i} + u_i \quad (59)$$

$$\Longrightarrow \boxed{Y_i = \beta_0 + u_i} \quad (60)$$

Example with the Equality of Coefficients Hypothesis

- **The unrestricted regression is,**

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i} + u_i \quad (61)$$

- $\boldsymbol{H_0}: \boldsymbol{\beta_1} = \boldsymbol{\beta_2}$
- **Plugging in $\boldsymbol{H_0}$,**

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_1 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i} + u_i \quad (62)$$

$$\Longrightarrow \boxed{Y_i = \beta_0 + \beta_1 (X_{1i} + X_{2i}) + \beta_3 X_{3i} + \beta_4 X_{4i} + u_i.} \quad (63)$$

Example with the Sum to Specific Value Hypothesis

- **The unrestricted regression is,**

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i} + u_i \quad (64)$$

- $\boldsymbol{H_0}$: $H_0: \beta_1 + \beta_2 = W \Longrightarrow \beta_2 = W - \beta_1$
- **Plugging in** $\boldsymbol{H_0}$**,**

$$Y_i = \beta_0 + \beta_1 X_{1i} + (W - \beta_1) X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i} + u_i \qquad (65)$$

$$\Longrightarrow \boxed{Y_i = \beta_0 + \beta_1 (X_{1i} - X_{2i}) + W X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i} + u_i.} \qquad (66)$$

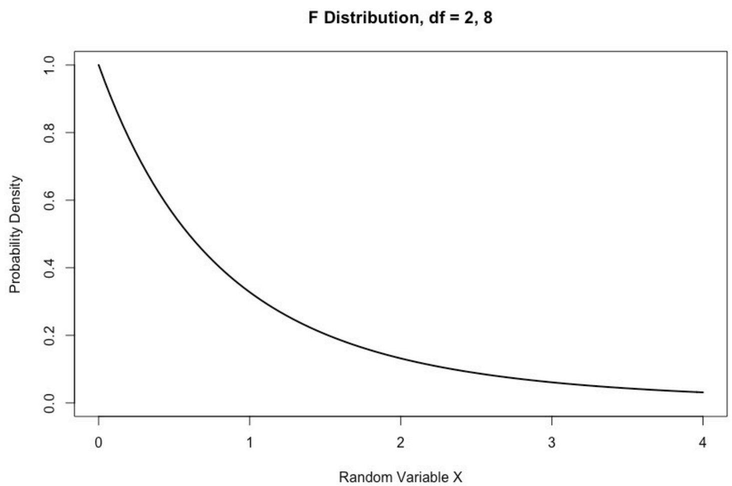
Statistical Foundation of Joint Hypothesis Testing

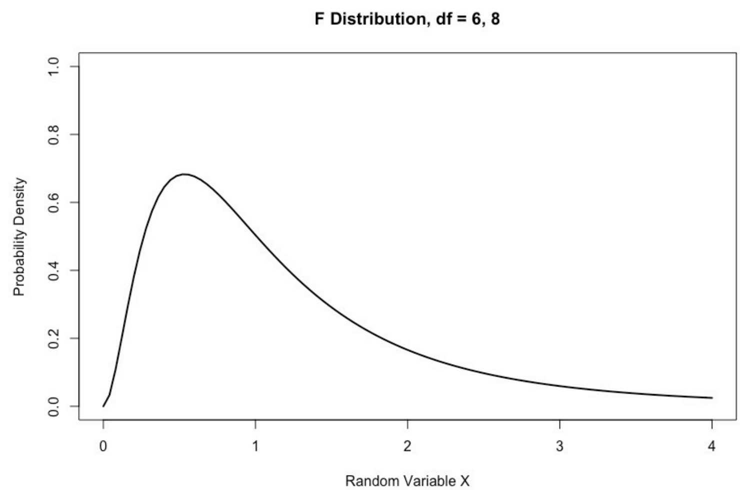
- **Under assumptions 1 through 4 and the assumption that the error term is normally distributed, one can show that**

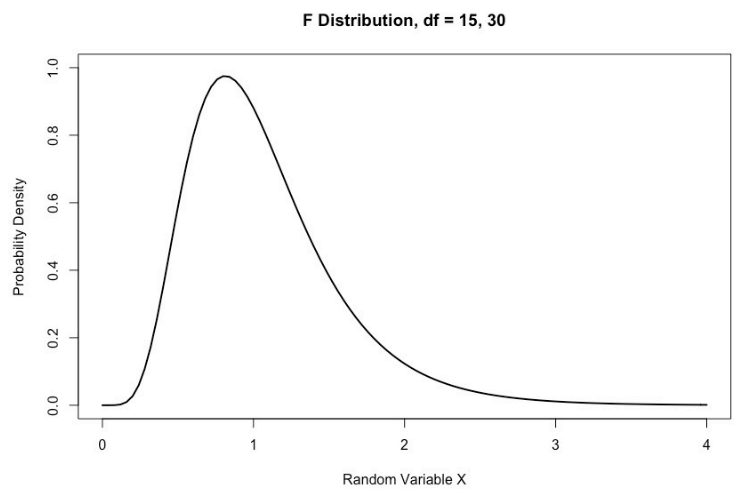
$$\frac{(RSS_R - RSS_U)/_{m}}{RSS_U/_{(n-k-1)}} \sim \underbrace{F(m, n-k-1)}_{\textbf{The } F \textbf{ distribution}} \tag{67}$$

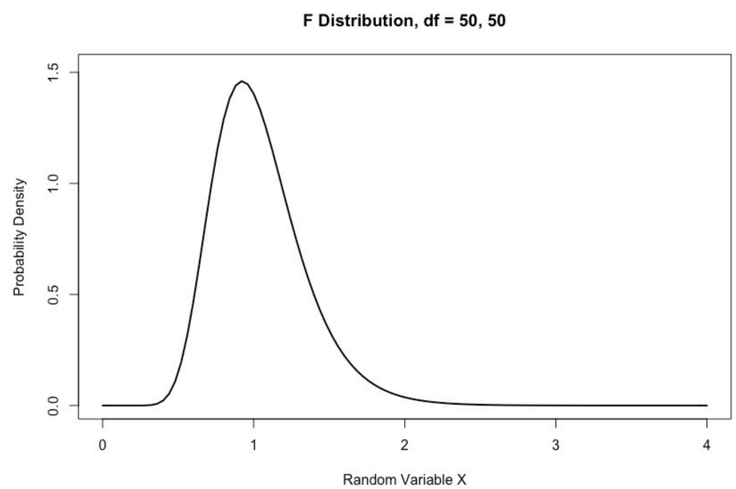
  - RSS_R is the sum of squared residuals from the restricted regression
  - RSS_U is the sum of squared residuals from the unrestricted regression
  - m is the number of restrictions (the difference between the number of parameters in the two models)
  - k is the number of RHS variables in the unrestricted regression

F Distribution Shapes









Statistical Foundation of Joint Hypothesis Testing

$$\frac{(RSS_R - RSS_U)/_{m}}{RSS_U/_{(n-k-1)}} \sim F(m, n-k-1) \qquad (71)$$

- **Remember the residual sum of squares is just,**

$$RSS = \sum_{i=1}^{N} \hat{u}_i^2 \qquad (72)$$

Why Care about $\frac{(RSS_R - RSS_U)/m}{RSS_U/(n-k-1)}$?

- **Why is $\frac{(RSS_R - RSS_U)/m}{RSS_U/(n-k-1)}$ even relevant?**
  - **Remember: we are examining whether the unrestricted regression fits the data better than the restricted regression. If it does, then you reject the null hypothesis**
  - If the **null hypothesis is true** then the unrestricted regression would show us that the null hypothesis is true by giving us coefficients that come very close to satisfying the null hypothesis
    - And in that case, the unrestricted regression will not fit the data much better than the restricted regression ($\Longrightarrow RSS_R \approx RSS_U$)
      - The numerator will be close to zero
  - If the **null hypothesis is false**, then our unrestricted regression will be very different from the restricted regression and because it (the restricted regression) has more parameters should fit better
    - The numerator will be large if the null hypothesis is false

Why Care about $\frac{(RSS_R - RSS_U)/m}{RSS_U/(n-k-1)}$?

- **Why is $\frac{(RSS_R - RSS_U)/m}{RSS_U/(n-k-1)}$ even relevant?**
  - So, if this statistic is basically a measure of how much better the unrestricted regression fits the data
    - If the measure is large, then we will be more likely to reject the null hypothesis
    - If the measure is small then we will be more likely to fail to reject the null hypothesis
    - To determine whether we in fact reject or fail to reject the null hypothesis we apply criteria
      - Just like in hypothesis testing of a single parameter, we need to compare the resulting value to a critical value in the distribution that this statistic would take (the F distribution in this case)

Testing Joint Hypotheses: The 9-Step Recipe

- **Step 1:** State clearly the null hypothesis
- **Step 2:** State clearly the alternative hypothesis
- **Step 3:** Choose a significance value
- **Step 4:** Estimate the unrestricted regression and find the sum of squared residuals from the Stata output (RSS_U)
- **Step 5:** Estimate the restricted regression and find the sum of squared residuals from the Stata output (RSS_R)

Testing Joint Hypotheses: The Recipe

- **Step 6:** Count the number of restrictions, which equals the difference between the number of parameters in the unrestricted and restricted regressions. This is m
- **Step 7:** Find the critical value for your significance level for an F random variable with m degrees of freedom in the numerator and with $n-k-1$ degrees of freedom in the denominator
- **Step 8:** Calculate the "F Statistic"

$$F = \frac{(RSS_R - RSS_U)\big/ m}{RSS_U \big/ (n-k-1)} \tag{73}$$

Testing Joint Hypotheses: The Recipe

- **Step 9 (The Last Step):** Compare the critical value from the F distribution to your F Statistic
  - If the statistic exceeds the critical value we reject the null. Otherwise we fail to reject
  - You can find the critical value in the table in the back of your econometrics book for the F distribution
  - You can also call this table in Stata using the command `invFtail`

F Test in Stata

- **This sounds like a lot of steps, but as usual it is very easy to implement this in Stata**
  - The is no default joint hypothesis test, so you need to ask Stata to do it after running your regression that you want to test
- **Relevant Code:**
  - **Multiple Zeros:** `test var1 var2`
  - **All slopes Equal Zero:** `test var1 var2 … vark`
  - **Equality of Coefficients:** `test var1 = var2`
  - **Coefficients Sum to a Specific Value (W):** `test var1 + var2 = W`

F Test in Stata

- **The output will look unimpressive, but it'll give you a p-value, which is all you need to reject or fail to reject the null, recall:**
  - Remember if $p<0.05$ then statistically significant at 95% confidence or above
  - if $p<0.01$ then statistically significant at 99% confidence or above
  - if $p<0.1$ then statistically significant at 90% confidence
- **See the Stata manual for more details: https://www.stata.com/manuals/rtest.pdf**