Hypothesis Testing

AECN 396/896-002

Before we start

Learning objectives

Learn the theory of statistical hypothesis testing and learn how to conduct various tests.

Table of contents

Review on Statistical Hypothesis Testing

Hypothesis Testing: General Steps

Here is the general step of any hypothesis testing:

- Step 1: specify the null (H_0) and alternative (H_1) hypotheses
- Step 2: find the distribution of the test statistic if the null hypothesis is true
- Step 3: calculate the test statistic based on the data and regression results
- Step 4: define the significance level
- Step 5: check how unlikely that you get the actual test statistic (found at Step 3) if indeed the null hypothesis is true

Hypothesis testing: An Example

Goal

Suppose you want to test if the expected value of a normally distributed random variable (x) is 1 or not.

State of Knowledge

We do know x follows a normal distribution and its variance is 4 for some reason.

Your Estimator

Your estimator is the sample mean: $\theta = \sum_{i=1}^{J} x_i / J$

Math Asides

Math Aside 1

$Var(ax) = a^2 Var(X)$

So, we know that $\theta \sim N(\alpha, 4/J)$ (of course α is not known).

Math Aside 2

If $x \sim N(a, b)$, then,

$x - a \sim N(0, b)$ (shift)

Further,

$\frac{x-a}{\sqrt{b}} \sim N(0, 1)$ (see Math Aside 1)

Math Aside 1

$Var(ax) = a^2 Var(X)$

Math Aside 2

If $x \sim N(a, b)$, then, $\frac{x-a}{\sqrt{b}} \sim N(0, 1)$ (see Math Aside 1)

Therefore,

Since $\theta = \sum_{i=1}^{J} x_i / J$ and $x_i \sim N(\alpha, 4)$,

$Var(\theta) = J \times \frac{1}{J^2} Var(x) = 4/J$

$\frac{\sqrt{J}}{2} \cdot (\theta - \alpha) \sim N(0, 1).$

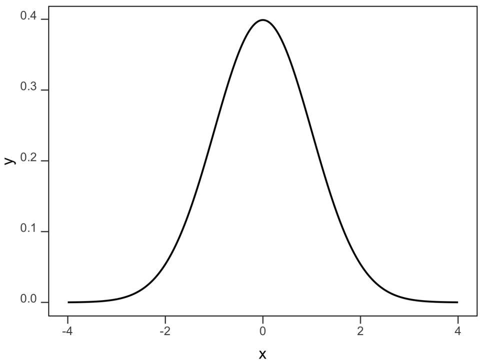
Test statistic and its distribution under the null hypothesis

We established that $\frac{\sqrt{J}}{2} \cdot (\theta - \alpha) \sim N(0,1)$.

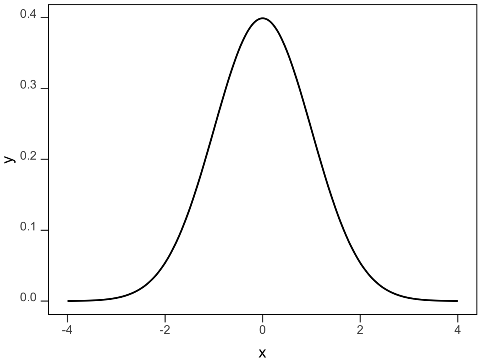
The null hypothesis is $\alpha = 1$.

If $\alpha = 1$ is indeed true, then $\sqrt{J} \times (\theta - 1)/2 \sim N(0,1)$.

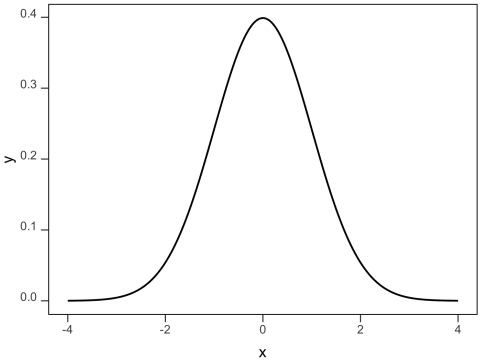
In other words, if you multiply the sample mean by the square root of the number of observations and divide it by 2, then it follows the standard normal distribution like below.



Hypothesis testing: An Example (Cont.)



Hypothesis testing: An Example (Cont.)



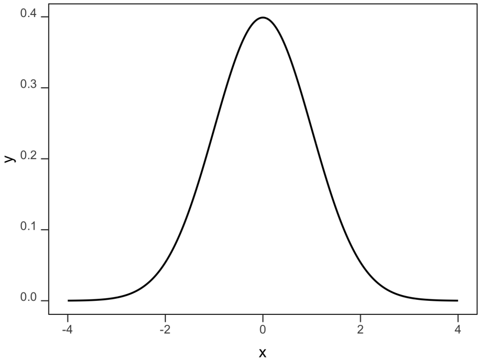
Case 1

Suppose you have obtained 100 samples $(J = 100)$ and calculated θ (sample mean), which turned out to be 2.

Then, your test statistic is $\sqrt{100} \times (2 - 1)/2 = 5$.

How unlikely is it to get the number you got (5) if the null hypothesis is indeed true?

Hypothesis testing: An Example (Cont.)



Case 1

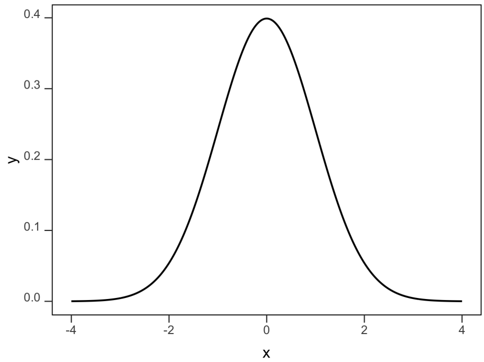
Suppose you have obtained 100 samples $(J = 100)$ and calculated θ (sample mean), which turned out to be 2.

Then, your test statistic is $\sqrt{100} \times (2 - 1)/2 = 5$.

How unlikely is it to get the number you got (5) if the null hypothesis is indeed true?

Very unlikely! So, the null is very much likely wrong!

Hypothesis testing: An Example (Cont.)



Case 1

Suppose you have obtained 100 samples $(J = 100)$ and calculated θ (sample mean), which turned out to be 2.

Then, your test statistic is $\sqrt{100} \times (2 - 1)/2 = 5.$

How unlikely is it to get the number you got (5) if the null hypothesis is indeed true?

Very unlikely! So, the null is very much likely wrong!

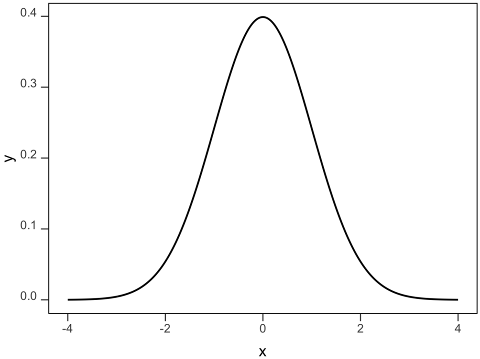
Case 2

Suppose you have obtained 400 samples $(J = 400)$ and calculated θ (sample mean), which turned out to be 1.02.

Then, your test statistic is $\sqrt{400} \times (1.02 - 1)/2 = 0.2.$

How unlikely is it to get the number you got (0.2) if the null hypothesis is indeed true?

Hypothesis testing: An Example (Cont.)



Case 1

Suppose you have obtained 100 samples $(J = 100)$ and calculated θ (sample mean), which turned out to be 2.

Then, your test statistic is $\sqrt{100} \times (2 - 1)/2 = 5$.

How unlikely is it to get the number you got (5) if the null hypothesis is indeed true?

Very unlikely! So, the null is very much likely wrong!

Case 2

Suppose you have obtained 400 samples $(J = 400)$ and calculated θ (sample mean), which turned out to be 1.02.

Then, your test statistic is $\sqrt{400} \times (1.02 - 1)/2 = 0.2$.

How unlikely is it to get the number you got (0.2) if the null hypothesis is indeed true?

Very much possible! So, we cannot say confidently that the null is wrong.

Hypothesis testing: Side Note

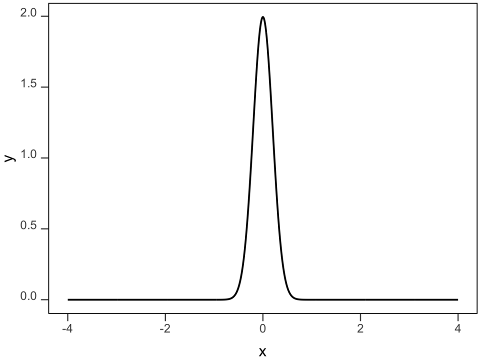
Note that you do not really need to use $\sqrt{J} \times (\theta - \alpha)/2$ as your test statistic.

Hypothesis testing: Side Note

Note that you do not really need to use $\sqrt{J} \times (\theta - \alpha)/2$ as your test statistic.

You could alternatively use $\theta - \alpha$. But, in that case, you need to be looking at $N(0, 4/J)$ instead of $N(0, 1)$.

For example, when the number of observations is 100 $(J = 100)$,

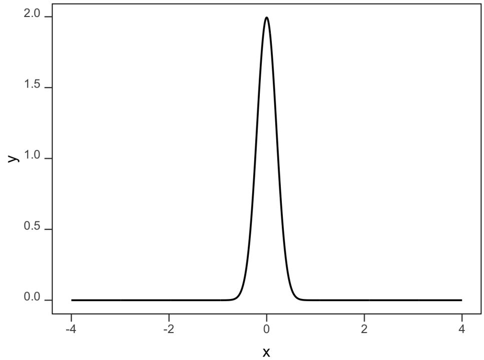


Hypothesis testing: Side Note

Note that you do not really need to use $\sqrt{J} \times (\theta - \alpha)/2$ as your test statistic.

You could alternatively use $\theta - \alpha$. But, in that case, you need to be looking at $N(0, 4/J)$ instead of $N(0, 1)$.

For example, when the number of observations is 100 $(J = 100)$,



Reconsider the case 1

Suppose you have obtained 100 samples $(J = 100)$ and calculated θ (sample mean), which turned out to be 2.

Then, your test statistic is $2 - 1 = 1$.

Is it unlikely for you to get 1 if the null hypothesis is true?

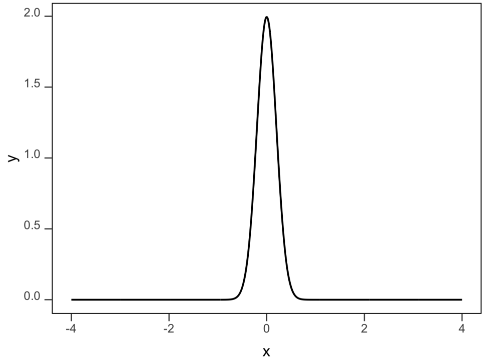
The conclusion would be exactly the same as using $\sqrt{J} \times (\theta - \alpha)/2$ because the distribution under the null is adjusted according to the test statistic you use.

Hypothesis testing: Side Note

Note that you do not really need to use $\sqrt{J} \times (\theta - \alpha)/2$ as your test statistic.

You could alternatively use $\theta - \alpha$. But, in that case, you need to be looking at $N(0, 4/J)$ instead of $N(0, 1)$.

For example, when the number of observations is 100 $(J = 100)$,



Reconsider the case 1

Suppose you have obtained 100 samples $(J = 100)$ and calculated θ (sample mean), which turned out to be 2.

Then, your test statistic is $2 - 1 = 1$.

Is it unlikely for you to get 1 if the null hypothesis is true?

The conclusion would be exactly the same as using $\sqrt{J} \times (\theta - \alpha)/2$ because the distribution under the null is adjusted according to the test statistic you use.

Note

We always use normalize test statistic so that we can always look up the same distribution.

Summary

What do we need?

- test-statistic of which we know the distribution (e.g., t-distribution, Normal distribution) assuming the null hypothesis

What do we (often) do?

- transform (most of the time) a raw random variable (e.g., sample mean in the example above) into a test statistic of which we know the distribution assuming that the null hypothesis is true
    - e.g., we transformed the sample mean so that it follows the standard Normal distribution.
- check if the actual number you got from the test statistic is likely to happen or not (formal criteria has not been discussed yet)

Exercise

You have collected data on annual salary for those who graduated from University A and B. You are interested in testing whether the difference in annual salary between the universities (call it x) is not equal 10 on average. You know (for unknown reasons) know that the difference is distributed as $N(\theta, 16)$.

1. What is the null hypothesis?
2. Under the null hypothesis, what is the distribution of the difference (test-statistic).
3. Normalize the test statistic so that the transformed version follows $N(0, 1)$.
4. The actual difference you observed is 30. What is the probability that you observe a number greater than 30 if the null hypothesis is true? Use `prnom()`.

Exercise

You have collected data on annual salary for those who graduated from University A and B. You are interested in testing whether the difference in annual salary between the universities (call it x) is not equal 10 on average. You know (for unknown reasons) know that the difference is distributed as $N(\theta, 16)$.

1. What is the null hypothesis?
2. Under the null hypothesis, what is the distribution of the difference (test-statistic).
3. Normalize the test statistic so that the transformed version follows $N(0, 1)$.
4. The actual difference you observed is 30. What is the probability that you observe a number greater than 30 if the null hypothesis is true? Use `prnom()`.

Note:

In reality,

- we need to find out what the distribution of the test statistic is
- we need to formerly define when we accept or not accept the null hypothesis

Hypothesis Testing of the Coefficients of a Linear Model

Hypothesis Testing of the Coefficients

Hypotheses examples:

Consider the following model,

$$wage = \beta_0 + \beta_1 educ + \beta_2 exper + u$$

- Hypothesis 1: education has no impact on wage $(\beta_1 = 0)$
- Hypothesis 2: experience has a positive impact on wage $(\beta_2 > 0)$

Hypothesis Testing of the Coefficients

Hypotheses examples:

Consider the following model,

$$wage = \beta_0 + \beta_1 educ + \beta_2 exper + u$$

- Hypothesis 1: education has no impact on wage $(\beta_1 = 0)$
- Hypothesis 2: experience has a positive impact on wage $(\beta_2 > 0)$

Statistical hypothesis testing

- If $\hat{\beta}_1$ is non-random, but just a scalar, all you have to do is just check if $\hat{\beta}_1 = 0$ or not

Hypothesis Testing of the Coefficients

Hypotheses examples:

Consider the following model,

$$wage = \beta_0 + \beta_1 educ + \beta_2 exper + u$$

- Hypothesis 1: education has no impact on wage $(\beta_1 = 0)$
- Hypothesis 2: experience has a positive impact on wage $(\beta_2 > 0)$

Statistical hypothesis testing

- If $\hat{\beta}_1$ is non-random, but just a scalar, all you have to do is just check if $\hat{\beta}_1 = 0$ or not
- But, the estimate you get is just one realization of the range of values $\hat{\beta}_1$ could take because it is a random varaible

Hypothesis Testing of the Coefficients

Hypotheses examples:

Consider the following model,

$$wage = \beta_0 + \beta_1 educ + \beta_2 exper + u$$

- Hypothesis 1: education has no impact on wage $(\beta_1 = 0)$
- Hypothesis 2: experience has a positive impact on wage $(\beta_2 > 0)$

Statistical hypothesis testing

- If $\hat{\beta}_1$ is non-random, but just a scalar, all you have to do is just check if $\hat{\beta}_1 = 0$ or not
- But, the estimate you get is just one realization of the range of values $\hat{\beta}_1$ could take because it is a random varaible
- This means that even if $\beta_1 = 0$ in the population, it is possible to get an estimate that is very far from 0

Hypothesis Testing of the Coefficients

Hypotheses examples:

Consider the following model,

$$wage = \beta_0 + \beta_1 educ + \beta_2 exper + u$$

- Hypothesis 1: education has no impact on wage $(\beta_1 = 0)$
- Hypothesis 2: experience has a positive impact on wage $(\beta_2 > 0)$

Statistical hypothesis testing

- If $\hat{\beta}_1$ is non-random, but just a scalar, all you have to do is just check if $\hat{\beta}_1 = 0$ or not
- But, the estimate you get is just one realization of the range of values $\hat{\beta}_1$ could take because it is a random varaible
- This means that even if $\beta_1 = 0$ in the population, it is possible to get an estimate that is very far from 0

Hypothesis Testing in General

You have gotten an estimate of β $(\hat{\beta})$ and are wondering if the true value of β (which you will never know) is α (a specific constant).

Here is the underlying concept of hypothesis testing.

- What would be the distribution of $\hat{\beta}$ (the estimator) if the true value of β is indeed α?
- If so, how likely that you would have gotten the value you have gotten for $\hat{\beta}$

Hypothesis Testing of the Coefficients

Hypotheses examples:

Consider the following model,

$$wage = \beta_0 + \beta_1 educ + \beta_2 exper + u$$

- Hypothesis 1: education has no impact on wage $(\beta_1 = 0)$
- Hypothesis 2: experience has a positive impact on wage $(\beta_2 > 0)$

Statistical hypothesis testing

- If $\hat{\beta}_1$ is non-random, but just a scalar, all you have to do is just check if $\hat{\beta}_1 = 0$ or not
- But, the estimate you get is just one realization of the range of values $\hat{\beta}_1$ could take because it is a random varaible
- This means that even if $\beta_1 = 0$ in the population, it is possible to get an estimate that is very far from 0

Hypothesis Testing in General

You have gotten an estimate of β $(\hat{\beta})$ and are wondering if the true value of β (which you will never know) is α (a specific constant).

Here is the underlying concept of hypothesis testing.

- What would be the distribution of $\hat{\beta}$ (the estimator) if the true value of β is indeed α?
- If so, how likely that you would have gotten the value you have gotten for $\hat{\beta}$

So, let's discuss the distribution of β_j now.

Hypothesis Testing: Additional Assumption

So far, we learned that:

- Expected value of the OLS estimators under MLR.1 ~ MLR.4
- Variance of the OLS estimators under MLR.1 ~ MLR.5

Hypothesis Testing: Additional Assumption

So far, we learned that:

- Expected value of the OLS estimators under MLR.1 ~ MLR.4
- Variance of the OLS estimators under MLR.1 ~ MLR.5

We have NOT made any assumptions about the distribution of the error term!!

Hypothesis Testing: Additional Assumption

So far, we learned that:

- Expected value of the OLS estimators under MLR.1 ~ MLR.4
- Variance of the OLS estimators under MLR.1 ~ MLR.5

We have NOT made any assumptions about the distribution of the error term!!

In order to perform hypothesis testing, we need to make assumptions about the distribution of error term (this is not strictly true, but more on this later)

Hypothesis Testing: Additional Assumption

So far, we learned that:

- Expected value of the OLS estimators under MLR.1 ~ MLR.4
- Variance of the OLS estimators under MLR.1 ~ MLR.5

We have NOT made any assumptions about the distribution of the error term!!

In order to perform hypothesis testing, we need to make assumptions about the distribution of error term (this is not strictly true, but more on this later)

Normality Assumption

The population error u is independent of the explanatory variables $x_1, \ldots, x_k$ and is normally distributed with zero mean and variance σ^2:

$$u \sim N(0, \sigma^2)$$

Hypothesis Testing: Additional Assumption

So far, we learned that:

- Expected value of the OLS estimators under MLR.1 ~ MLR.4
- Variance of the OLS estimators under MLR.1 ~ MLR.5

We have NOT made any assumptions about the distribution of the error term!!

In order to perform hypothesis testing, we need to make assumptions about the distribution of error term (this is not strictly true, but more on this later)

Normality Assumption

The population error u is independent of the explanatory variables $x_1, \ldots, x_k$ and is normally distributed with zero mean and variance σ^2:

$$u \sim N(0, \sigma^2)$$

Note

The normality assumption is much more than error term being distributed as Normal.

Independence of the error term implies

- $E[u|x] = 0$
- $Var[u|x] = \sigma^2$

So, we are necessarily assuming MLR.4 and MLR.5 hold by the independence assumption.

The Implications of All the Assumptions

distribution of the dependent variable

The distribution of y conditional on x is a Normal distribution

$y|x \sim N(\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k, \sigma^2)$

- $E[y|x]$ is $\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k$
- $u|x$ is $N(0, \sigma^2)$

The Implications of All the Assumptions

distribution of the dependent variable

The distribution of y conditional on x is a Normal distribution

$y|x \sim N(\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k, \sigma^2)$

- $E[y|x]$ is $\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k$
- $u|x$ is $N(0, \sigma^2)$

distribution of the OLS estimator

If the MLR.1 through MLR.6 are satisfied, OLS estimators are also Normally distributed!

$\hat{\beta}_j \sim N(\beta_j, Var(\hat{\beta}_j))$

which means,

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

The Implications of All the Assumptions

distribution of the dependent variable

The distribution of y conditional on x is a Normal distribution

$$y|x \sim N(\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k, \sigma^2)$$

- $E[y|x]$ is $\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k$
- $u|x$ is $N(0, \sigma^2)$

distribution of the OLS estimator

If the MLR.1 through MLR.6 are satisfied, OLS estimators are also Normally distributed!

$$\hat{\beta}_j \sim N(\beta_j, Var(\hat{\beta}_j))$$

which means,

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

Okay, so are we going to use this for testing involving β_j?

The Implications of All the Assumptions

distribution of the dependent variable

The distribution of y conditional on x is a Normal distribution

$$y|x \sim N(\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k, \sigma^2)$$

- $E[y|x]$ is $\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k$
- $u|x$ is $N(0, \sigma^2)$

distribution of the OLS estimator

If the MLR.1 through MLR.6 are satisfied, OLS estimators are also Normally distributed!

$$\hat{\beta}_j \sim N(\beta_j, Var(\hat{\beta}_j))$$

which means,

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

Okay, so are we going to use this for testing involving β_j?

No.

The Implications of All the Assumptions

distribution of the dependent variable

The distribution of y conditional on x is a Normal distribution

$$y|x \sim N(\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k, \sigma^2)$$

- $E[y|x]$ is $\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k$
- $u|x$ is $N(0, \sigma^2)$

distribution of the OLS estimator

If the MLR.1 through MLR.6 are satisfied, OLS estimators are also Normally distributed!

$$\hat{\beta}_j \sim N(\beta_j, Var(\hat{\beta}_j))$$

which means,

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

Okay, so are we going to use this for testing involving β_j?

No.

t-statistic and t-distribution

But, in practice, we need to estimate $sd(\beta_j)$. If we use $\widehat{se(\hat{\beta}_j)}$ instead of $se(\hat{\beta}_j)$, then,

$$\frac{\hat{\beta}_j - \beta_j}{\widehat{se(\hat{\beta}_j)}} \sim t_{n-k-1}$$

where $n - k - 1$ is the degree of freedom of residual.

(Note: $\widehat{se(\hat{\beta}_j)} = \hat{\sigma}^2 / \left[SST_X \cdot (1 - R_j^2)\right]$)

Recap on notations

- k: the number of explanatory variables included except the intercept
- σ^2: the true variance of the error term
- $\hat{\sigma}^2$: the estimator (estimate) of the variance of the error term
  - $\hat{\sigma}^2 = \frac{\sum \hat{u}_i^2}{n-k-1}$
- $SST_X = \sum(x_i - \bar{x})^2$
- $\hat{\beta}_j$: OLS estimator (estimate) on explanatory variable x_j
  - $\hat{\beta} = \frac{\sum(x_i-\bar{x})(y_i-\bar{y})}{SST_X}$ (for simple univariate regression)
- $var(\hat{\beta}_j)$: the true variance of $\hat{\beta}_j$
- R_j^2: R^2 when you regres x_j on all the other covariates (mathematical expression omitted)
- $\widehat{var}(\hat{\beta}_j)$: Estimator (estimate) of $var(\hat{\beta}_j)$
  - $\widehat{var}(\hat{\beta}_j) = \frac{\hat{\sigma}^2}{SST_X \cdot (1-R_j^2)}$
- $se(\hat{\beta}_j)$: square root of $var(\hat{\beta}_j)$
- $\widehat{se}(\hat{\beta}_j)$: square root of $\widehat{var}(\hat{\beta}_j)$

For a univariate regression case,

Normalized version of $\hat{\beta}$

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

So, $[\frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{SST_X} - \beta_j] \cdot \frac{SST_X}{\sigma^2} \sim N(0,1)$

Estimated version

$$\frac{\hat{\beta}_j - \beta_j}{\widehat{se(\hat{\beta}_j)}} \sim t_{n-k-1}$$

So, $[\frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{SST_X} - \beta_j] \cdot SST_X \cdot \frac{n-k-1}{\sum \hat{u}_i^2} \sim t_{n-k-1}$

Null and alternative hypotheses

Statistical hypothesis testing involves two hypotheses: Null and Alternative hypotheses.

Pretend that you are an attorney who indicted a defendent who you think commited a crime.

Null Hypothesis

Hypothesis that you would like to reject (defendent is not guilty)

Alternative Hypothesis

Hypothesis you are in support of (defendent is guilty)

One-sided and Two-sided Alternatives Hypotheses

one-sided alternative:

$H_0 : \beta_j = 0 \; H_1 : \beta_j > 0$

One-sided and Two-sided Alternatives Hypotheses

one-sided alternative:

$H_0 : \beta_j = 0 \; H_1 : \beta_j > 0$

You look at the positive end of the t-distribution to see if the t-statistic you obtained is more extreme than the level of error you accept (significance level).

One-sided and Two-sided Alternatives Hypotheses

one-sided alternative:

$H_0 : \beta_j = 0 \; H_1 : \beta_j > 0$

You look at the positive end of the t-distribution to see if the t-statistic you obtained is more extreme than the level of error you accept (significance level).

two-sided alternative:

$H_0 : \beta_j = 0 \; H_1 : \beta_j \neq 0$

You look at the both ends of the t-distribution to see if the t-statistic you obtained is more extreme than the level of error you accept (significance level).

Significance level

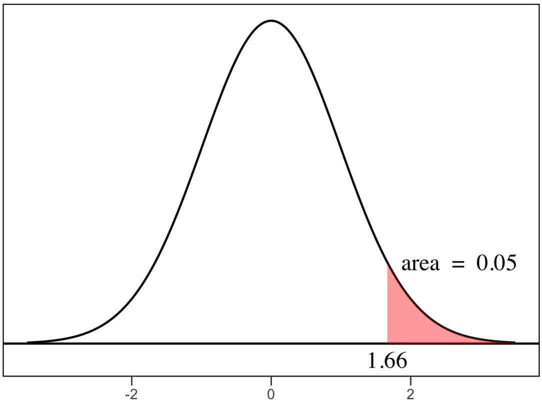
Definition:

The probability of rejecting the null when the null is actually true (The probability that you wrongly claim that the null hypothesis is wrong even though it's true in reality: Type I error)

The lower the significance level, you are more sure that the null is indeed wrong when you reject the null hypothesis

One-sided test: 5% significance level

This is the distribution of $\frac{\hat{\beta}_j - \beta_j}{\widehat{se(\hat{\beta}_j)}}$ if $\beta_j = 0$ (the null hypothesis is true).



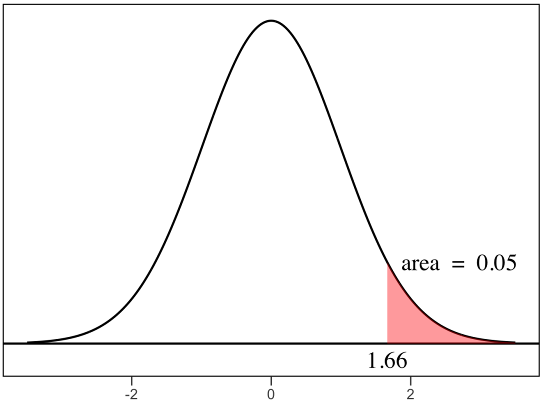
The probability that you get a value larger than 1.66 is 5% (0.05 in area).

Decision rule

Reject the null hypothesis if the t-statistic is greater than 1.66 (95% quantile of the t-distribution)

One-sided test: 5% significance level

This is the distribution of $\frac{\hat{\beta}_j - \beta_j}{\widehat{se(\hat{\beta}_j)}}$ if $\beta_j = 0$ (the null hypothesis is true).



The probability that you get a value larger than 1.66 is 5% (0.05 in area).

Decision rule

Reject the null hypothesis if the t-statistic is greater than 1.66 (95% quantile of the t-distribution)

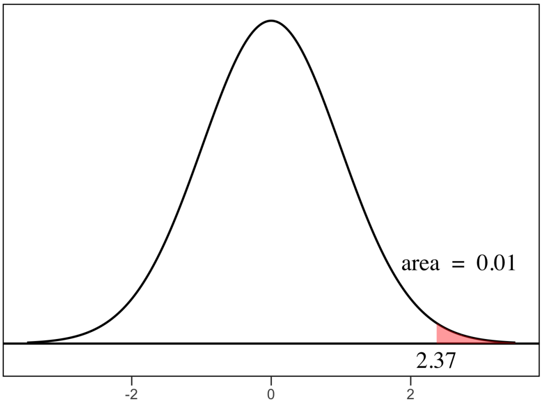
If you follow the decision rule, then you have a 5% chance that you are wrong in rejecting the null hypothesis of $\beta_j = 0$.

Here,

- 5% is the significance level
- 1.66 is the critical value above which you will reject the null

One-sided test: 1% significance level

This is the distribution of $\frac{\hat{\beta}_j - \beta_j}{\widehat{se(\hat{\beta}_j)}}$ if $\beta_j = 0$ (the null hypothesis is true).



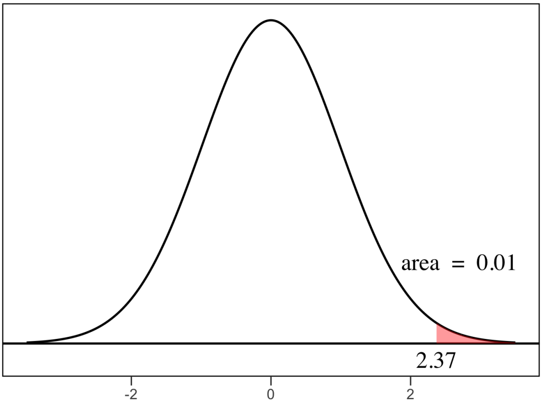
The probability that you get a value larger than 2.37 is 1% (0.01 in area).

Decision rule

Reject the null hypothesis if the t-statistic is greater than 2.37 (99% quantile of the t-distribution)

One-sided test: 1% significance level

This is the distribution of $\frac{\hat{\beta}_j - \beta_j}{\widehat{se(\hat{\beta}_j)}}$ if $\beta_j = 0$ (the null hypothesis is true).



The probability that you get a value larger than 2.37 is 1% (0.01 in area).

Decision rule

Reject the null hypothesis if the t-statistic is greater than 2.37 (99% quantile of the t-distribution)

If you follow the decision rule, then you have a 1% chance that you are wrong in rejecting the null hypothesis of $\beta_j = 0$.

Here,

- 1% is the significance level
- 2.37 is the critical value above which you will reject the null

One-sided test: an example

Estimated Model

The impact of experience on wage:

- $\widehat{log(wage)} = 0.284 + 0.092 \times educ + 0.0041 \times exper + 0.022 \times tenure$
- $se(\widehat{\hat{\beta}_{exper}}) = 0.0017$
- $n = 526$

One-sided test: an example

Estimated Model

The impact of experience on wage:

- $log(wage) = 0.284 + 0.092 \times educ + 0.0041 \times exper + 0.022 \times tenure$
- $se(\widehat{\hat{\beta}_{exper}}) = 0.0017$
- $n = 526$

Hypothesis

- H_0: $\beta_{exper} = 0$
- H_1: $\beta_{exper} > 0$

One-sided test: an example

Estimated Model

The impact of experience on wage:

- $log(wage) = 0.284 + 0.092 \times educ + 0.0041 \times exper + 0.022 \times tenure$
- $\widehat{se(\hat{\beta}_{exper})} = 0.0017$
- $n = 526$

Hypothesis

- H_0: $\beta_{exper} = 0$
- H_1: $\beta_{exper} > 0$

Test

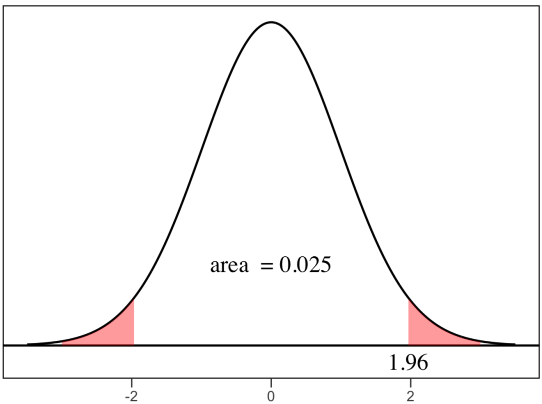
t-statistic $= 0.0041/0.0017 = 2.41$

The critical value is the 99% quantile of $t_{526-3-1}$, which is 2.33 (it can be obtained by `qt(0.95, 522)`)

Since $2.41 > 2.33$, we reject the null in favor of the alternative hypothesis at the 1% level.

Two-sided test: 5% significance level

This is the distribution of $\frac{\hat{\beta}_j - \beta_j}{\widehat{se(\hat{\beta}_j)}}$ if $\beta_j = 0$ (the null hypothesis is true).



The probability that you get a value more extreme than 1.96 or -1.96 is 5% (0.05 in area cobining the two area at the edges).

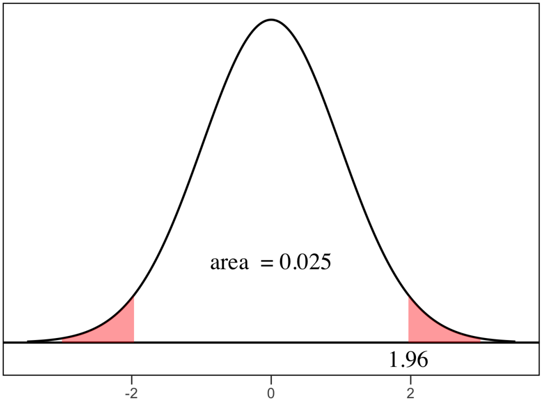
(Note: irrespective of the type of tests, the distribution of t-statistics is the same.)

Decision rule

Reject the null hypothesis if the absolute value of the t-statistic is greater than 1.96.

Two-sided test: 5% significance level

This is the distribution of $\frac{\hat{\beta}_j - \beta_j}{\widehat{se(\hat{\beta}_j)}}$ if $\beta_j = 0$ (the null hypothesis is true).



The probability that you get a value more extreme than 1.96 or -1.96 is 5% (0.05 in area cobining the two area at the edges).

(Note: irrespective of the type of tests, the distribution of t-statistics is the same.)

Decision rule

Reject the null hypothesis if the absolute value of the t-statistic is greater than 1.96.

If you follow the decision rule, then you have a 5% chance that you are wrong in rejecting the null hypothesis of $\beta_j = 0$.

Here,

- 5% is the significance level
- 1.96 is the critical value above which you will reject the null

p-value

Definition

The smallest significance level at which the null hypothesis would be rejected (the probability of observing a test statistic at least as extreme as we did if the null hypothesis is true)

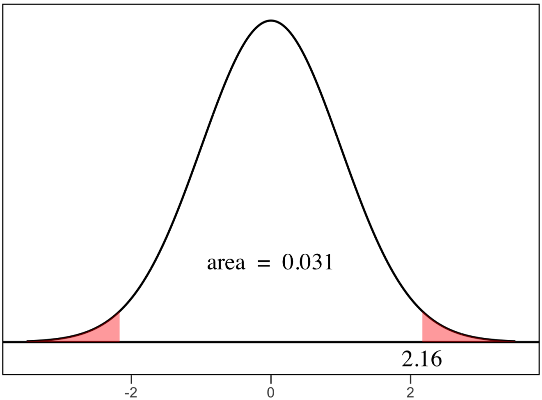
Suppose the t-statistic you got is 2.16. Then, there's a 3.1% chance you reject the null when it is actually true, if you use it as the critical value.

p-value

Definition

The smallest significance level at which the null hypothesis would be rejected (the probability of observing a test statistic at least as extreme as we did if the null hypothesis is true)

Suppose the t-statistic you got is 2.16. Then, there's a 3.1% chance you reject the null when it is actually true, if you use it as the critical value.



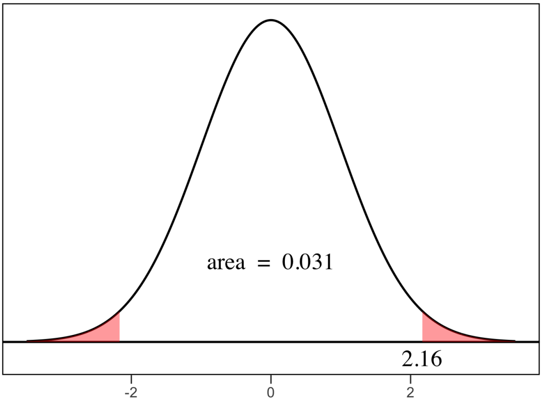
So, the lower significance level the null hypothesis is rejected is 3.1%, which is the definition of p-value.

p-value

Definition

The smallest significance level at which the null hypothesis would be rejected (the probability of observing a test statistic at least as extreme as we did if the null hypothesis is true)

Suppose the t-statistic you got is 2.16. Then, there's a 3.1% chance you reject the null when it is actually true, if you use it as the critical value.



So, the lower significance level the null hypothesis is rejected is 3.1%, which is the definition of p-value.

Decision rule

If the p-value is lower than your choice of significance level, then you reject the null.

This decision rule of course results in the same test results as the one we saw that uses a t-value and critical value.

R implementation

Get the data and run a regression

```
#--- get the data ---#
wage <- read_csv("wage1.csv")

#--- run a regression ---#
reg_wage <- feols(wage ~ educ + exper + tenure, data = wage)
```

R implementation

Get the data and run a regression

```r
#--- get the data ---#
wage <- read_csv("wage1.csv")

#--- run a regression ---#
reg_wage <- feols(wage ~ educ + exper + tenure, data = wage)
```

Obtain t-statistics

You can apply the `tidy()` function from the `broom` package to access the regression results (`reg_wage` here) as a `tibble` (`data.frame`).

```r
library(broom)

#* tidy up the regression results
reg_results_table <- tidy(reg_wage)

#* display
reg_results_table
```

```
## # A tibble: 4 × 5
##   term        estimate std.error statistic  p.value
##   <chr>          <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept)  -2.87      0.729      -3.94 9.22e- 5
## 2 educ          0.599     0.0513     11.7  3.68e-28
## 3 exper         0.0223    0.0121      1.85 6.45e- 2
## 4 tenure        0.169     0.0216      7.82 2.93e-14
```

- `estimate`: coefficient estimates $(\hat{\beta}_j)$
- `std.error`: $se(\hat{\beta}_j)$
- `statistic`: t-statistic for the null of $\beta_j = 0$
- `p.value`: p-value (for the two sided test with the null of $\beta_j = 0$)

So, for the t-test of of $\beta_j = 0$ is already there. You do not need to do anything further.

For the null hypothesis other than $\beta_j = 0$, you need further work.

R implementation (by hand)

Suppose you are testing the null hypothesis of $\beta_{educ} = 1$ against the alternative hypothesis of $\beta_{educ} \neq 1$ (so, this is a two-sided test).

The t-value for this test is not available from the summary.

```
#--- coefficient estimate on educ ---#
beta_educ <-
  reg_results_table %>%
  filter(term == "educ") %>%
  pull(estimate)

#--- se of the coefficient on educ ---#
se_beta_educ <-
  reg_results_table %>%
  filter(term == "educ") %>%
  pull(std.error)

#--- get the t-value ---#
(
  t_value <- (beta_educ - 1) / se_beta_educ
)
```

```
## [1] -7.819953
## attr(,"type")
## [1] "IID"
```

R implementation (by hand)

Suppose you are testing the null hypothesis of $\beta_{educ} = 1$ against the alternative hypothesis of $\beta_{educ} \neq 1$ (so, this is a two-sided test).

The t-value for this test is not available from the summary.

```
#--- coefficient estimate on educ ---#
beta_educ <-
  reg_results_table %>%
  filter(term == "educ") %>%
  pull(estimate)

#--- se of the coefficient on educ ---#
se_beta_educ <-
  reg_results_table %>%
  filter(term == "educ") %>%
  pull(std.error)

#--- get the t-value ---#
(
  t_value <- (beta_educ - 1) / se_beta_educ
)
```

```
## [1] -7.819953
## attr(,"type")
## [1] "IID"
```

The degree of freedom $(n - k - 1)$ of the t-distribution can be obtained by applying `degrees_freedom(reg_wage, "resid")`:

```
(
  df_t <- degrees_freedom(reg_wage, "resid")
)
```

```
## [1] 522
```

R implementation (by hand)

Suppose you are testing the null hypothesis of $\beta_{educ} = 1$ against the alternative hypothesis of $\beta_{educ} \neq 1$ (so, this is a two-sided test).

The t-value for this test is not available from the summary.

```
#--- coefficient estimate on educ ---#
beta_educ <-
  reg_results_table %>%
  filter(term == "educ") %>%
  pull(estimate)

#--- se of the coefficient on educ ---#
se_beta_educ <-
  reg_results_table %>%
  filter(term == "educ") %>%
  pull(std.error)

#--- get the t-value ---#
(
  t_value <- (beta_educ - 1) / se_beta_educ
)
```

```
## [1] -7.819953
## attr(,"type")
## [1] "IID"
```

The degree of freedom $(n - k - 1)$ of the t-distribution can be obtained by applying `degrees_freedom(reg_wage, "resid")`:

```
(
  df_t <- degrees_freedom(reg_wage, "resid")
)
```

```
## [1] 522
```

You can get the 97.5% quantile of the t_{522} using the `qt()` function:

```
(
  critical_value <- qt(0.975, df = df_t)
)
```

```
## [1] 1.964519
```

R implementation (by hand)

Suppose you are testing the null hypothesis of $\beta_{educ} = 1$ against the alternative hypothesis of $\beta_{educ} \neq 1$ (so, this is a two-sided test).

The t-value for this test is not available from the summary.

```
#--- coefficient estimate on educ ---#
beta_educ <-
  reg_results_table %>%
  filter(term == "educ") %>%
  pull(estimate)

#--- se of the coefficient on educ ---#
se_beta_educ <-
  reg_results_table %>%
  filter(term == "educ") %>%
  pull(std.error)

#--- get the t-value ---#
(
  t_value <- (beta_educ - 1) / se_beta_educ
)
```

```
## [1] -7.819953
## attr(,"type")
## [1] "IID"
```

The degree of freedom $(n-k-1)$ of the t-distribution can be obtained by applying `degrees_freedom(reg_wage, "resid")`:

```
(
  df_t <- degrees_freedom(reg_wage, "resid")
)
```

```
## [1] 522
```

You can get the 97.5% quantile of the t_{522} using the `qt()` function:

```
(
  critical_value <- qt(0.975, df = df_t)
)
```

```
## [1] 1.964519
```

Since the absolute value of the t-value (-7.8199528) is greater than the critical value (1.9645189), you reject the null.

Confidence Interval (CI)

Confidence Interval (CI): Definition

Definition

If you calculate 95% CI on multiple different samples, 95% of the time, the calculated CI includes the true parameter

What confidence interval is not

The probability that a realized CI calculated from specific sample data includes the true parameter

How to get CI (in general)

General Procedure:

For the **assumed** distribution of statistic x, the $A\%$ confidence interval of x is the range with

- lower bound: 100 − A/2 percent quantile of x
- upper bound: 100-(100 − A)/2 percent quantile of x

How to get CI (in general)

General Procedure:

For the **assumed** distribution of statistic x, the $A\%$ confidence interval of x is the range with

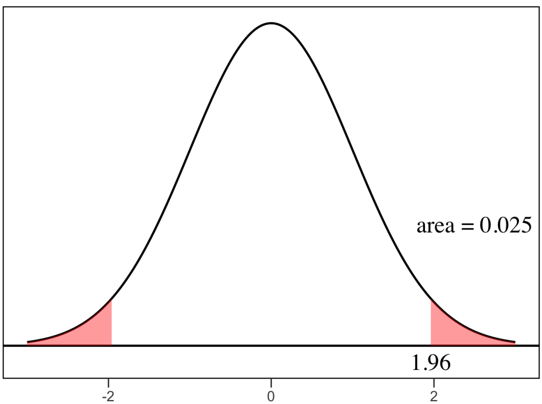
- lower bound: 100 − A/2 percent quantile of x
- upper bound: 100-(100 − A)/2 percent quantile of x

Example:

For the 95% CI (A = 95),

- lower bound: 2.5 (100 − 95/2) percent quantile of x
- upper bound: 97.5 (100-(100 − 95)/2) percent quantile of x

If x follos the standard normal distribution $(x \sim N(0,1))$, then,



The 2.5% and 97.5% quantiles are -1.96 and 1.96, respectively.

So, the 95% CI of x is [-1.96, 1.96].

How to get the CI of coefficients

Under the assumption of MLR.1 through MLR.6 (which includes the normality assumption of the error), we learned that

$$\frac{\hat{\beta}_j - \beta_j}{\widehat{se(\hat{\beta}_j)}} \sim t_{n-k-1}$$

How to get the CI of coefficients

Under the assumption of MLR.1 through MLR.6 (which includes the normality assumption of the error), we learned that

$$\frac{\hat{\beta}_j - \beta_j}{\widehat{se(\hat{\beta}_j)}} \sim t_{n-k-1}$$

So, following the general procedure we discussed in the previous slide, the A% confidence interval of $\frac{\hat{\beta}_j - \beta_j}{\widehat{se(\hat{\beta}_j)}}$ is

- lower bound: $(100 - A)/2$% quantile of the t_{n-k-1} distribution (let's call this Q_l)
- upper bound: $100 - (100 - A)/2$% quantile of the t_{n-k-1} distribution (let's call this Q_h)

How to get the CI of coefficients

Under the assumption of MLR.1 through MLR.6 (which includes the normality assumption of the error), we learned that

$\frac{\hat{\beta}_j - \beta_j}{\widehat{se(\hat{\beta}_j)}} \sim t_{n-k-1}$

So, following the general procedure we discussed in the previous slide, the A% confidence interval of $\frac{\hat{\beta}_j - \beta_j}{\widehat{se(\hat{\beta}_j)}}$ is

- lower bound: $(100 - A)/2$% quantile of the t_{n-k-1} distribution (let's call this Q_l)
- upper bound: $100 - (100 - A)/2$% quantile of the t_{n-k-1} distribution (let's call this Q_h)

But, we want the A% CI of β_j, not $\frac{\hat{\beta}_j - \beta_j}{se(\hat{\beta}_j)}$. Solving for β_j,

$\beta_j = t_{n-k-1} \times se(\hat{\beta}_j) + \hat{\beta}_j$

How to get the CI of coefficients

Under the assumption of MLR.1 through MLR.6 (which includes the normality assumption of the error), we learned that

$\frac{\hat{\beta}_j - \beta_j}{\widehat{se(\hat{\beta}_j)}} \sim t_{n-k-1}$

So, following the general procedure we discussed in the previous slide, the A% confidence interval of $\frac{\hat{\beta}_j - \beta_j}{\widehat{se(\hat{\beta}_j)}}$ is

- lower bound: $(100 - A)/2$% quantile of the t_{n-k-1} distribution (let's call this Q_l)
- upper bound: $100 - (100 - A)/2$% quantile of the t_{n-k-1} distribution (let's call this Q_h)

But, we want the A% CI of β_j, not $\frac{\hat{\beta}_j - \beta_j}{se(\hat{\beta}_j)}$. Solving for β_j,

$\beta_j = t_{n-k-1} \times se(\hat{\beta}_j) + \hat{\beta}_j$

So, to get the A% CI of β_j, we scale the CI of $\frac{\hat{\beta}_j - \beta_j}{se(\hat{\beta}_j)}$ by $se(\hat{\beta}_j)$ and then shift by $\hat{\beta}_j$.

- lower bound: $Q_l \times se(\hat{\beta}_j) + \hat{\beta}_j$
- lower bound: $Q_h \times se(\hat{\beta}_j) + \hat{\beta}_j$

CI: An Example

Get data

```
wage <- read_csv("wage1.csv") %>%
  select(wage, educ, exper, tenure)
```

Take a look at the data:

```
head(wage)
```

```
## # A tibble: 6 × 4
##    wage  educ exper tenure
##   <dbl> <dbl> <dbl>  <dbl>
## 1  3.10    11     2      0
## 2  3.24    12    22      2
## 3  3       11     2      0
## 4  6        8    44     28
## 5  5.30    12     7      2
## 6  8.75    16     9      8
```

Run OLS and extract necessary information

```
wage_reg <- feols(wage ~ educ + exper + tenure, data = wage)
```

Applying `tidy()` to `wage_reg`,

```
tidy(wage_reg)
```

```
## # A tibble: 4 × 5
##   term        estimate std.error statistic  p.value
##   <chr>          <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept)  -2.87      0.729      -3.94 9.22e- 5
## 2 educ          0.599     0.0513     11.7  3.68e-28
## 3 exper         0.0223    0.0121      1.85 6.45e- 2
## 4 tenure        0.169     0.0216      7.82 2.93e-14
```

The degrees of freedom of residual (n-k-1) is

```
df <- degrees_freedom(reg_wage, "resid")
```

CI: An Example (Cont.)

We are interested in getting the 90% confidence interval of the coefficient on `educ` (β_{educ}).

Under all the assumptions (MLR.1 through MLR.6), we know that in general,

$$\frac{\hat{\beta}_{educ} - \beta_{educ}}{\widehat{se(\hat{\beta}_{educ})}} \sim t_{n-k-1}$$

CI: An Example (Cont.)

We are interested in getting the 90% confidence interval of the coefficient on `educ` (β_{educ}).

Under all the assumptions (MLR.1 through MLR.6), we know that in general,

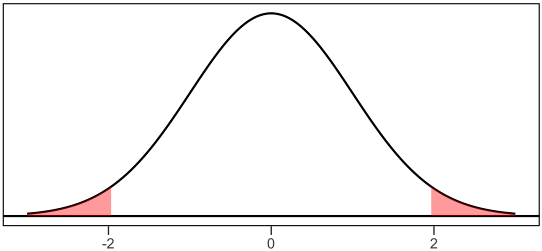
$\frac{\hat{\beta}_{educ} - \beta_{educ}}{\widehat{se(\hat{\beta}_{educ})}} \sim t_{n-k-1}$

Specifically for this regression,

- $\hat{\beta}_{educ}$ = 0.5989651
- $\widehat{se(\hat{\beta}_{educ})}$ = 0.0512835
- $n - k - 1 = 522$

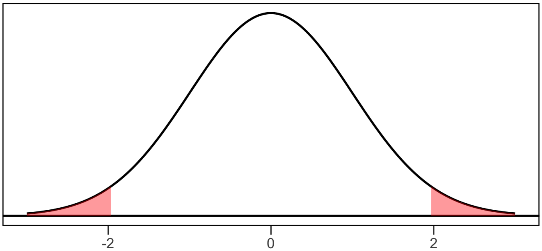
CI: An Example (Cont.)

Here is the distribution of t_{522}:



CI: An Example (Cont.)

Here is the distribution of t_{522}:



Now, we need to find the 5% ((100-90)/2) and 95% (100-(100-90)/2) quantile of t_{522}.

```
qt(0.05, df = 522)
```

```
## [1] -1.647778
```

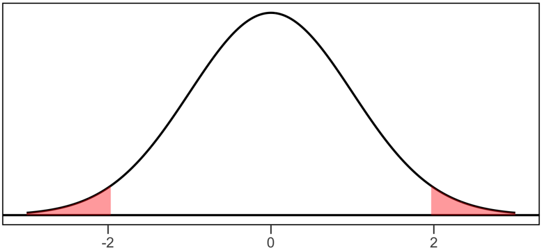
```
qt(0.95, df = 522)
```

```
## [1] 1.647778
```

Yes, we could have just gotten one of them and multiply it by -1 to get the other since t distribution is symmetric around 0.

CI: An Example (Cont.)

Here is the distribution of t_{522}:



Now, we need to find the 5% ((100-90)/2) and 95% (100-(100-90)/2) quantile of t_{522}.

```
qt(0.05, df = 522)
```

```
## [1] -1.647778
```

```
qt(0.95, df = 522)
```

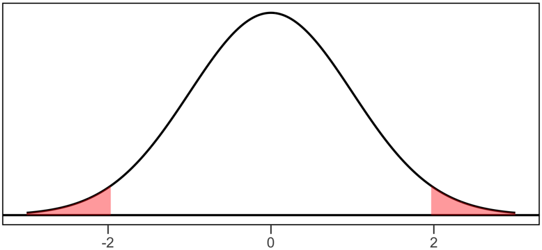
```
## [1] 1.647778
```

Yes, we could have just gotten one of them and multiply it by -1 to get the other since t distribution is symmetric around 0.

So, the 90% CI of $\frac{0.599-\beta_{educ}}{0.051} \sim t_{522}$ is [-1.6477779, 1.6477779]

CI: An Example (Cont.)

Here is the distribution of t_{522}:



Now, we need to find the 5% ((100-90)/2) and 95% (100-(100-90)/2) quantile of t_{522}.

```
qt(0.05, df = 522)
```

```
## [1] -1.647778
```

```
qt(0.95, df = 522)
```

```
## [1] 1.647778
```

Yes, we could have just gotten one of them and multiply it by -1 to get the other since t distribution is symmetric around 0.

So, the 90% CI of $\frac{0.599-\beta_{educ}}{0.051} \sim t_{522}$ is [-1.6477779, 1.6477779]

By scaling and shifting, the lower and upper bounds of the 90% CI of β_{educ} are:

- lower bound: 0.599 + 0.051 $\times$ -1.6477779 = 0.5149633
- upper bound: 0.599 + 0.051 $\times$ 1.6477779 = 0.6830367

In practice ...

You can just use `tidy()` with `conf.int = TRUE`, `conf.level = confidence level` like below:

```
tidy(wage_reg, conf.int = TRUE, conf.level = 0.9) %>%
  relocate(term, conf.low, conf.high)
```

```
## # A tibble: 4 × 7
##   term        conf.low conf.high estimate std.error statistic  p.value
##   <chr>          <dbl>     <dbl>    <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept) -4.07      -1.67    -2.87      0.729      -3.94 9.22e- 5
## 2 educ         0.514      0.683    0.599     0.0513     11.7  3.68e-28
## 3 exper        0.00247    0.0422   0.0223    0.0121      1.85 6.45e- 2
## 4 tenure       0.134      0.205    0.169     0.0216      7.82 2.93e-14
```

Linear Combination of Multiple Coefficients

Linear Combination of Multiple Coefficients

Example

$$log(wage) = \beta_0 + \beta_1 jc + \beta_2 univ + \beta_3 exper + u$$

- jc: 1 if you attended 2-year college, 0 otherwise
- $univ$: 1 if you attended 4-year college, 0 otherwise

Linear Combination of Multiple Coefficients

Example

$$log(wage) = \beta_0 + \beta_1 jc + \beta_2 univ + \beta_3 exper + u$$

- jc: 1 if you attended 2-year college, 0 otherwise
- $univ$: 1 if you attended 4-year college, 0 otherwise

Does the impact of education on wage is greater if you attend a 4-year college than 2-year college?

Linear Combination of Multiple Coefficients

Example

$$log(wage) = \beta_0 + \beta_1 jc + \beta_2 univ + \beta_3 exper + u$$

- jc: 1 if you attended 2-year college, 0 otherwise
- $univ$: 1 if you attended 4-year college, 0 otherwise

Does the impact of education on wage is greater if you attend a 4-year college than 2-year college?

The null and alternative hypotheses would be:

- $H_1 : \beta_1 < \beta_2$
- $H_0 : \beta_1 = \beta_2$

Linear Combination of Multiple Coefficients

Example

$$log(wage) = \beta_0 + \beta_1 jc + \beta_2 univ + \beta_3 exper + u$$

- jc: 1 if you attended 2-year college, 0 otherwise
- $univ$: 1 if you attended 4-year college, 0 otherwise

Does the impact of education on wage is greater if you attend a 4-year college than 2-year college?

The null and alternative hypotheses would be:

- $H_1 : \beta_1 < \beta_2$
- $H_0 : \beta_1 = \beta_2$

Rewriting them,

- $H_1 : \beta_1 - \beta_2 < 0$
- $H_0 : \beta_1 - \beta_2 = 0$

Linear Combination of Multiple Coefficients

Example

$$log(wage) = \beta_0 + \beta_1 jc + \beta_2 univ + \beta_3 exper + u$$

- jc: 1 if you attended 2-year college, 0 otherwise
- $univ$: 1 if you attended 4-year college, 0 otherwise

Does the impact of education on wage is greater if you attend a 4-year college than 2-year college?

The null and alternative hypotheses would be:

- $H_1 : \beta_1 < \beta_2$
- $H_0 : \beta_1 = \beta_2$

Rewriting them,

- $H_1 : \beta_1 - \beta_2 < 0$
- $H_0 : \beta_1 - \beta_2 = 0$

Or,

- $H_1 : \alpha < 0$
- $H_0 : \alpha = 0$

where $\alpha = \beta_1 - \beta_2$

Note that α is a linear combination of β_1 and β_2.

Linear Combination of Multiple Coefficients

Important fact

For any linear combination of the OLS coefficients, denoted as $\hat{\alpha}$, the following holds:

$$\frac{\hat{\alpha} - \alpha}{se(\hat{\alpha})} \sim t_{n-k-1}$$

Where α is the true value (it is $\beta_1 - \beta_2$ in the example in the previous slide).

Linear Combination of Multiple Coefficients

Important fact

For any linear combination of the OLS coefficients, denoted as $\hat{\alpha}$, the following holds:

$$\frac{\hat{\alpha} - \alpha}{se(\hat{\alpha})} \sim t_{n-k-1}$$

Where α is the true value (it is $\beta_1 - \beta_2$ in the example in the previous slide).

So, using the example, this means that

$$\frac{\hat{\alpha} - \alpha}{se(\hat{\alpha})} = \frac{\hat{\beta}_1 - \hat{\beta}_2 - (\beta_1 - \beta_2)}{se(\hat{\beta}_1 - \hat{\beta}_2)} \sim t_{n-k-1}$$

Linear Combination of Multiple Coefficients

Important fact

For any linear combination of the OLS coefficients, denoted as $\hat{\alpha}$, the following holds:

$$\frac{\hat{\alpha}-\alpha}{se(\hat{\alpha})} \sim t_{n-k-1}$$

Where α is the true value (it is $\beta_1 - \beta_2$ in the example in the previous slide).

So, using the example, this means that

$$\frac{\hat{\alpha}-\alpha}{se(\hat{\alpha})} = \frac{\hat{\beta}_1 - \hat{\beta}_2 - (\beta_1 - \beta_2)}{se(\hat{\beta}_1 - \hat{\beta}_2)} \sim t_{n-k-1}$$

This is great because we know how to do t-test!

Linear Combination of Multiple Coefficients

Important fact

For any linear combination of the OLS coefficients, denoted as $\hat{\alpha}$, the following holds:

$$\frac{\hat{\alpha} - \alpha}{se(\hat{\alpha})} \sim t_{n-k-1}$$

Where α is the true value (it is $\beta_1 - \beta_2$ in the example in the previous slide).

So, using the example, this means that

$$\frac{\hat{\alpha} - \alpha}{se(\hat{\alpha})} = \frac{\hat{\beta}_1 - \hat{\beta}_2 - (\beta_1 - \beta_2)}{se(\hat{\beta}_1 - \hat{\beta}_2)} \sim t_{n-k-1}$$

This is great because we know how to do t-test!

Going back to the example

Our null hypothesis is $\alpha = 0$ (or $\beta_1 - \beta_2 = 0$).

So, If indeed the null hypothesis is true, then

$$\frac{\hat{\alpha} - 0}{se(\hat{\alpha})} = \frac{\hat{\beta}_1 - \hat{\beta}_2 - 0}{se(\hat{\beta}_1 - \hat{\beta}_2)} \sim t_{n-k-1}$$

Linear Combination of Multiple Coefficients

Important fact

For any linear combination of the OLS coefficients, denoted as $\hat{\alpha}$, the following holds:

$$\frac{\hat{\alpha} - \alpha}{se(\hat{\alpha})} \sim t_{n-k-1}$$

Where α is the true value (it is $\beta_1 - \beta_2$ in the example in the previous slide).

So, using the example, this means that

$$\frac{\hat{\alpha} - \alpha}{se(\hat{\alpha})} = \frac{\hat{\beta}_1 - \hat{\beta}_2 - (\beta_1 - \beta_2)}{se(\hat{\beta}_1 - \hat{\beta}_2)} \sim t_{n-k-1}$$

This is great because we know how to do t-test!

Going back to the example

Our null hypothesis is $\alpha = 0$ (or $\beta_1 - \beta_2 = 0$).

So, If indeed the null hypothesis is true, then

$$\frac{\hat{\alpha} - 0}{se(\hat{\alpha})} = \frac{\hat{\beta}_1 - \hat{\beta}_2 - 0}{se(\hat{\beta}_1 - \hat{\beta}_2)} \sim t_{n-k-1}$$

So, all you need to do is to substitute $\hat{\beta}_1$, $\hat{\beta}_2$, $se(\hat{\beta}_1 - \hat{\beta}_2)$ into the formula and see if the value is beyond the critical value for your chosen level of statistical significance.

Linear Combination of Multiple Coefficients

But,

$$se(\hat{\beta}_1 - \hat{\beta}_2) = \sqrt{Var(\hat{\beta}_1 - \hat{\beta}_2)} \neq \sqrt{Var(\hat{\beta}_1) + Var(\hat{\beta}_2)}$$

Linear Combination of Multiple Coefficients

But,

$$se(\hat{\beta}_1 - \hat{\beta}_2) = \sqrt{Var(\hat{\beta}_1 - \hat{\beta}_2)} \neq \sqrt{Var(\hat{\beta}_1) + Var(\hat{\beta}_2)}$$

If the following was true,

$$se(\hat{\beta}_1 - \hat{\beta}_2) = \sqrt{Var(\hat{\beta}_1) + Var(\hat{\beta}_2)}$$

then, we could have just extracted $Var(\hat{\beta}_1)$ and $Var(\hat{\beta}_2)$ individually from the regression object on R, sum them up, and take a square root of it.

Linear Combination of Multiple Coefficients

But,

$$se(\hat{\beta}_1 - \hat{\beta}_2) = \sqrt{Var(\hat{\beta}_1 - \hat{\beta}_2)} \neq \sqrt{Var(\hat{\beta}_1) + Var(\hat{\beta}_2)}$$

If the following was true,

$$se(\hat{\beta}_1 - \hat{\beta}_2) = \sqrt{Var(\hat{\beta}_1) + Var(\hat{\beta}_2)}$$

then, we could have just extracted $Var(\hat{\beta}_1)$ and $Var(\hat{\beta}_2)$ individually from the regression object on R, sum them up, and take a square root of it.

Math Aside

$$Var(ax + by) = a^2Var(x) + 2abCov(x, y) + b^2Var(y)$$

Linear Combination of Multiple Coefficients

But,

$$se(\hat{\beta}_1 - \hat{\beta}_2) = \sqrt{Var(\hat{\beta}_1 - \hat{\beta}_2)} \neq \sqrt{Var(\hat{\beta}_1) + Var(\hat{\beta}_2)}$$

If the following was true,

$$se(\hat{\beta}_1 - \hat{\beta}_2) = \sqrt{Var(\hat{\beta}_1) + Var(\hat{\beta}_2)}$$

then, we could have just extracted $Var(\hat{\beta}_1)$ and $Var(\hat{\beta}_2)$ individually from the regression object on R, sum them up, and take a square root of it.

Math Aside

$$Var(ax + by) = a^2Var(x) + 2abCov(x, y) + b^2Var(y)$$

So,

$$se(\hat{\beta}_1 - \hat{\beta}_2) = \sqrt{Var(\hat{\beta}_1 - \hat{\beta}_2)} = \sqrt{Var(\hat{\beta}_1) - 2Cov(\hat{\beta}_1, \hat{\beta}_2) + Var(\hat{\beta}_2)}$$

Demonstration using R

Regression

```r
twoyear <- readRDS("twoyear.rds") # import data

reg_sc <- feols(lwage ~ jc + univ + exper, data = twoyear) # OLS
```

Demonstration using R

Regression

```r
twoyear <- readRDS("twoyear.rds") # import data

reg_sc <- feols(lwage ~ jc + univ + exper, data = twoyear) # OLS
```

Variance covariance matrix

Variance covariance matrix is a matrix where

- $VCOV_{i,i}$: the variance of ith variable's coefficient estimator
- $VCOV_{i,j}$: the covariance between ith and jth variables' estimators

You can get is by applying `vcov()` to regression results:

```r
(
  vcov_sc <- vcov(reg_sc) # variance covariance matrix
)
```

```
##               (Intercept)            jc          univ         exper
## (Intercept)  4.435337e-04 -1.741432e-05 -1.573472e-05 -3.104756e-06
## jc          -1.741432e-05  4.663243e-05  1.927929e-06 -1.718296e-08
## univ        -1.573472e-05  1.927929e-06  5.330230e-06  3.933491e-08
## exper       -3.104756e-06 -1.718296e-08  3.933491e-08  2.479792e-08
```

For example, $vcov_{sc}[2,2]$ is the variance of $\hat{\beta}_{jc}$, and $vcov_{sc}[2,3]$ is the covariance between $\hat{\beta}_{jc}$ and $\hat{\beta}_{univ}$.

Demonstration using R

Calculate the t-statistic

```
vcov_sc
```

```
##               (Intercept)            jc          univ         exper
## (Intercept)  4.435337e-04 -1.741432e-05 -1.573472e-05 -3.104756e-06
## jc          -1.741432e-05  4.663243e-05  1.927929e-06 -1.718296e-08
## univ        -1.573472e-05  1.927929e-06  5.330230e-06  3.933491e-08
## exper       -3.104756e-06 -1.718296e-08  3.933491e-08  2.479792e-08
```

```
reg_sc_td <- tidy(reg_sc)

beta_jc <-
  reg_sc_td %>%
  filter(term == "jc") %>%
  pull(estimate)

beta_univ <-
  reg_sc_td %>%
  filter(term == "univ") %>%
  pull(estimate)

numerator <- beta_jc - beta_univ
denominator <-
  sqrt(
    vcov_sc["jc", "jc"] - 2 * vcov_sc["jc", "univ"] + vcov_sc["univ", "univ"]
  )
t_stat <- numerator / denominator
t_stat
```

```
## [1] -1.467657
## attr(,"type")
## [1] "IID"
```

Multiple Linear Restrictions: F-test

Multiple Linear Restrictions: F-test

Example

$$log(salary) = \beta_0 + \beta_1 years + \beta_2 gamesyr + \beta_3 bavg + \beta_4 hrunsyr + \beta_5 rbisyr + u$$

- $salary$: salary in 1993
- $years$: years in the league
- $gamesyr$: average games played per year
- $bavg$: career batting average
- $hrunsyr$: home runs per year
- $rbisyr$: runs batted in per year

Multiple Linear Restrictions: F-test

Example

$$log(salary) = \beta_0 + \beta_1 years + \beta_2 gamesyr + \beta_3 bavg + \beta_4 hrunsyr + \beta_5 rbisyr + u$$

- $salary$: salary in 1993
- $years$: years in the league
- $gamesyr$: average games played per year
- $bavg$: career batting average
- $hrunsyr$: home runs per year
- $rbisyr$: runs batted in per year

Hypothesis

Once years in the league and games per year have been controlled for, the statistics measuring performance ($bavg$, $hrunsyr$, $rbisyr$) have no effect on salary collectively.

H_0: $\beta_3 = 0$, $\beta_4 = 0$, and $\beta_5 = 0$

H_1: H_0 is not true

Multiple Linear Restrictions: F-test

Example

$$log(salary) = \beta_0 + \beta_1 years + \beta_2 gamesyr + \beta_3 bavg + \beta_4 hrunsyr + \beta_5 rbisyr + u$$

- $salary$: salary in 1993
- $years$: years in the league
- $gamesyr$: average games played per year
- $bavg$: career batting average
- $hrunsyr$: home runs per year
- $rbisyr$: runs batted in per year

Hypothesis

Once years in the league and games per year have been controlled for, the statistics measuring performance ($bavg$, $hrunsyr$, $rbisyr$) have no effect on salary collectively.

H_0: $\beta_3 = 0$, $\beta_4 = 0$, and $\beta_5 = 0$

H_1: H_0 is not true

Questions

How do we test this?

- H_0 holds if all of β_3, β_4, or β_5 are zero.
- Conduct t-test for each coefficient individually?

Individual t-test on the coefficients

Running a regression

```r
library(readstata13)

#--- red the data ---#
mlb_data <- read.dta13("MLB1.dta")

#--- run a rergession ---#
mlb_reg <- feols(log(salary) ~ years + gamesyr + bavg
  + hrunsyr + rbisyr, data = mlb_data)

#--- take a look at the results ---#
tidy(mlb_reg)
```

```
## # A tibble: 6 × 5
##   term        estimate std.error statistic   p.value
##   <chr>          <dbl>     <dbl>     <dbl>     <dbl>
## 1 (Intercept) 11.2        0.289      38.8   4.19e-128
## 2 years        0.0689     0.0121      5.68  2.79e-  8
## 3 gamesyr      0.0126     0.00265     4.74  3.09e-  6
## 4 bavg         0.000979   0.00110     0.887 3.76e-  1
## 5 hrunsyr      0.0144     0.0161      0.899 3.69e-  1
## 6 rbisyr       0.0108     0.00717     1.50  1.34e-  1
```

Individual t-test on the coefficients

Running a regression

```
library(readstata13)

#--- red the data ---#
mlb_data <- read.dta13("MLB1.dta")

#--- run a rergession ---#
mlb_reg <- feols(log(salary) ~ years + gamesyr + bavg
  + hrunsyr + rbisyr, data = mlb_data)

#--- take a look at the results ---#
tidy(mlb_reg)
```

```
## # A tibble: 6 × 5
##   term        estimate std.error statistic   p.value
##   <chr>          <dbl>     <dbl>     <dbl>     <dbl>
## 1 (Intercept) 11.2        0.289      38.8   4.19e-128
## 2 years        0.0689     0.0121      5.68  2.79e-  8
## 3 gamesyr      0.0126     0.00265     4.74  3.09e-  6
## 4 bavg         0.000979   0.00110     0.887 3.76e-  1
## 5 hrunsyr      0.0144     0.0161      0.899 3.69e-  1
## 6 rbisyr       0.0108     0.00717     1.50  1.34e-  1
```

Question

What do you find?

Individual t-test on the coefficients

Running a regression

```r
library(readstata13)

#--- red the data ---#
mlb_data <- read.dta13("MLB1.dta")

#--- run a rergession ---#
mlb_reg <- feols(log(salary) ~ years + gamesyr + bavg
  + hrunsyr + rbisyr, data = mlb_data)

#--- take a look at the results ---#
tidy(mlb_reg)
```

```
## # A tibble: 6 × 5
##   term        estimate std.error statistic   p.value
##   <chr>          <dbl>     <dbl>     <dbl>     <dbl>
## 1 (Intercept) 11.2        0.289      38.8   4.19e-128
## 2 years        0.0689     0.0121      5.68  2.79e-  8
## 3 gamesyr      0.0126     0.00265     4.74  3.09e-  6
## 4 bavg         0.000979   0.00110     0.887 3.76e-  1
## 5 hrunsyr      0.0144     0.0161      0.899 3.69e-  1
## 6 rbisyr       0.0108     0.00717     1.50  1.34e-  1
```

Question

What do you find?

None of the coefficients on `bavg`, `hrunsyr`, and `rbisyr` is statistically significantly different from 0 even at the 10% level!!

So, does this mean that they collectively have no impact on the salary of MLB players?

Individual t-test on the coefficients

Running a regression

```
library(readstata13)

#--- red the data ---#
mlb_data <- read.dta13("MLB1.dta")

#--- run a rergession ---#
mlb_reg <- feols(log(salary) ~ years + gamesyr + bavg
  + hrunsyr + rbisyr, data = mlb_data)

#--- take a look at the results ---#
tidy(mlb_reg)
```

```
## # A tibble: 6 × 5
##   term        estimate std.error statistic   p.value
##   <chr>          <dbl>     <dbl>     <dbl>     <dbl>
## 1 (Intercept) 11.2        0.289      38.8   4.19e-128
## 2 years        0.0689     0.0121      5.68  2.79e-  8
## 3 gamesyr      0.0126     0.00265     4.74  3.09e-  6
## 4 bavg         0.000979   0.00110     0.887 3.76e-  1
## 5 hrunsyr      0.0144     0.0161      0.899 3.69e-  1
## 6 rbisyr       0.0108     0.00717     1.50  1.34e-  1
```

Question

What do you find?

None of the coefficients on `bavg`, `hrunsyr`, and `rbisyr` is statistically significantly different from 0 even at the 10% level!!

So, does this mean that they collectively have no impact on the salary of MLB players?

If you were to conclude that they do not have statistically significant impact jointly, you would turn out to be wrong!!

SSR (or R^2) turns out to be useful for testing their impacts jointly.

F-test

In doing an F-test of the null hypothesis, we compare sum of squared residuals (SSR) of two models:

Unrestricted Model

$$log(salary) = \beta_0 + \beta_1 years + \beta_2 gamesyr + \beta_3 bavg + \beta_4 hrunsyr + \beta_5 rbisyr + u$$

Restricted Model

$$log(salary) = \beta_0 + \beta_1 years + \beta_2 gamesyr + u$$

The coefficients on $bavg$, $hrunsyr$, and $rbisyr$ are restricted to be 0 following the null hypothesis.

F-test

In doing an F-test of the null hypothesis, we compare sum of squared residuals (SSR) of two models:

Unrestricted Model

$$log(salary) = \beta_0 + \beta_1 years + \beta_2 gamesyr + \beta_3 bavg + \beta_4 hrunsyr + \beta_5 rbisyr + u$$

Restricted Model

$$log(salary) = \beta_0 + \beta_1 years + \beta_2 gamesyr + u$$

The coefficients on $bavg$, $hrunsyr$, and $rbisyr$ are restricted to be 0 following the null hypothesis.

Question

If the null hypothesis is indeed true, then what do you think is going to happen if you compare the SSR of the two models?

Sum of Squared Residuals (SSR) for F-test

SSR of the unrestricted model: SSR_u.

```
#--- run OLS ---#
res_u <- feols(log(salary) ~ years + gamesyr + bavg
  + hrunsyr + rbisyr, data = mlb_data)

#--- SSR ---#
sum(res_u$residuals^2)
```

```
## [1] 183.1863
```

SSR of the restricted model: SSR_r.

```
#--- run OLS ---#
res_r <- feols(log(salary) ~ years + gamesyr, data = mlb_data)

#--- SSR ---#
sum(res_r$residuals^2)
```

```
## [1] 198.3115
```

Sum of Squared Residuals (SSR) for F-test

SSR of the unrestricted model: SSR_u.

```
#--- run OLS ---#
res_u <- feols(log(salary) ~ years + gamesyr + bavg
  + hrunsyr + rbisyr, data = mlb_data)

#--- SSR ---#
sum(res_u$residuals^2)
```

```
## [1] 183.1863
```

SSR of the restricted model: SSR_r.

```
#--- run OLS ---#
res_r <- feols(log(salary) ~ years + gamesyr, data = mlb_data)

#--- SSR ---#
sum(res_r$residuals^2)
```

```
## [1] 198.3115
```

Questions

- Which SSR is larger? Does that make sense?

Sum of Squared Residuals (SSR) for F-test

SSR of the unrestricted model: SSR_u.

```
#--- run OLS ---#
res_u <- feols(log(salary) ~ years + gamesyr + bavg
  + hrunsyr + rbisyr, data = mlb_data)

#--- SSR ---#
sum(res_u$residuals^2)
```

```
## [1] 183.1863
```

SSR of the restricted model: SSR_r.

```
#--- run OLS ---#
res_r <- feols(log(salary) ~ years + gamesyr, data = mlb_data)

#--- SSR ---#
sum(res_r$residuals^2)
```

```
## [1] 198.3115
```

Questions

- Which SSR is larger? Does that make sense?

SSR_r should be large because the restricted model has a smaller explanatory power than the unrestricted model.

Sum of Squared Residuals (SSR) for F-test

SSR of the unrestricted model: SSR_u.

```
#--- run OLS ---#
res_u <- feols(log(salary) ~ years + gamesyr + bavg
  + hrunsyr + rbisyr, data = mlb_data)

#--- SSR ---#
sum(res_u$residuals^2)
```

```
## [1] 183.1863
```

SSR of the restricted model: SSR_r.

```
#--- run OLS ---#
res_r <- feols(log(salary) ~ years + gamesyr, data = mlb_data)

#--- SSR ---#
sum(res_r$residuals^2)
```

```
## [1] 198.3115
```

Questions

- Which SSR is larger? Does that make sense?

SSR_r should be large because the restricted model has a smaller explanatory power than the unrestricted model.

- What does $SSR_r - SSR_u$ measure?

Sum of Squared Residuals (SSR) for F-test

SSR of the unrestricted model: SSR_u.

```
#--- run OLS ---#
res_u <- feols(log(salary) ~ years + gamesyr + bavg
  + hrunsyr + rbisyr, data = mlb_data)

#--- SSR ---#
sum(res_u$residuals^2)
```

```
## [1] 183.1863
```

SSR of the restricted model: SSR_r.

```
#--- run OLS ---#
res_r <- feols(log(salary) ~ years + gamesyr, data = mlb_data)

#--- SSR ---#
sum(res_r$residuals^2)
```

```
## [1] 198.3115
```

Questions

- Which SSR is larger? Does that make sense?

SSR_r should be large because the restricted model has a smaller explanatory power than the unrestricted model.

- What does $SSR_r - SSR_u$ measure?

The contribution from the three excluded variables in explaining the dependent variable.

Sum of Squared Residuals (SSR) for F-test

SSR of the unrestricted model: SSR_u.

```
#--- run OLS ---#
res_u <- feols(log(salary) ~ years + gamesyr + bavg
  + hrunsyr + rbisyr, data = mlb_data)

#--- SSR ---#
sum(res_u$residuals^2)
```

```
## [1] 183.1863
```

SSR of the restricted model: SSR_r.

```
#--- run OLS ---#
res_r <- feols(log(salary) ~ years + gamesyr, data = mlb_data)

#--- SSR ---#
sum(res_r$residuals^2)
```

```
## [1] 198.3115
```

Questions

- Which SSR is larger? Does that make sense?

SSR_r should be large because the restricted model has a smaller explanatory power than the unrestricted model.

- What does $SSR_r - SSR_u$ measure?

The contribution from the three excluded variables in explaining the dependent variable.

- Is the contribution large enough to say that the excluded variables are important?

Sum of Squared Residuals (SSR) for F-test

SSR of the unrestricted model: SSR_u.

```
#--- run OLS ---#
res_u <- feols(log(salary) ~ years + gamesyr + bavg
  + hrunsyr + rbisyr, data = mlb_data)

#--- SSR ---#
sum(res_u$residuals^2)
```

```
## [1] 183.1863
```

SSR of the restricted model: SSR_r.

```
#--- run OLS ---#
res_r <- feols(log(salary) ~ years + gamesyr, data = mlb_data)

#--- SSR ---#
sum(res_r$residuals^2)
```

```
## [1] 198.3115
```

Questions

- Which SSR is larger? Does that make sense?

SSR_r should be large because the restricted model has a smaller explanatory power than the unrestricted model.

- What does $SSR_r - SSR_u$ measure?

The contribution from the three excluded variables in explaining the dependent variable.

- Is the contribution large enough to say that the excluded variables are important?

Cannot tell at this point.

F-test in general

Setup

Consider a following general model:

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + u$$

Suppose we have q restrictions to test: that is, the null hypothesis states that q of the variables have zero coefficients.

$$H_0 : \beta_{k-q+1} = 0, \beta_{k-q+2} = 0, \ldots, \beta_k = 0$$

When we impose the restrictions under H_0, the restricted model is the following:

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_{k-q} x_{k-q} + u$$

F-statistic

If the null hypothesis is true, then,

$$F = \frac{(SSR_r - SSR_u)/q}{SSR_u/(n-k-1)} \sim F_{q,n-k-1}$$

- q: the number of restrictions
- $n - k - 1$: degrees of freedom of residuals

Questions

- Is the above F-statistic always positive?

F-test in general

Setup

Consider a following general model:

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + u$$

Suppose we have q restrictions to test: that is, the null hypothesis states that q of the variables have zero coefficients.

$$H_0 : \beta_{k-q+1} = 0, \beta_{k-q+2} = 0, \ldots, \beta_k = 0$$

When we impose the restrictions under H_0, the restricted model is the following:

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_{k-q} x_{k-q} + u$$

F-statistic

If the null hypothesis is true, then,

$$F = \frac{(SSR_r - SSR_u)/q}{SSR_u/(n-k-1)} \sim F_{q,n-k-1}$$

- q: the number of restrictions
- $n-k-1$: degrees of freedom of residuals

Questions

- Is the above F-statistic always positive?

Yes, because $SSR_r - SSR_u$ is always positive.

F-test in general

Setup

Consider a following general model:

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + u$$

Suppose we have q restrictions to test: that is, the null hypothesis states that q of the variables have zero coefficients.

$$H_0 : \beta_{k-q+1} = 0, \beta_{k-q+2} = 0, \ldots, \beta_k = 0$$

When we impose the restrictions under H_0, the restricted model is the following:

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_{k-q} x_{k-q} + u$$

F-statistic

If the null hypothesis is true, then,

$$F = \frac{(SSR_r - SSR_u)/q}{SSR_u/(n-k-1)} \sim F_{q,n-k-1}$$

- q: the number of restrictions
- $n-k-1$: degrees of freedom of residuals

Questions

- Is the above F-statistic always positive?

Yes, because $SSR_r - SSR_u$ is always positive.

- The greater the joint contribution of the q variables, the (greater or smaller) the F-statistic?

F-test in general

Setup

Consider a following general model:

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + u$$

Suppose we have q restrictions to test: that is, the null hypothesis states that q of the variables have zero coefficients.

$$H_0 : \beta_{k-q+1} = 0, \beta_{k-q+2} = 0, \ldots, \beta_k = 0$$

When we impose the restrictions under H_0, the restricted model is the following:

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_{k-q} x_{k-q} + u$$

F-statistic

If the null hypothesis is true, then,

$$F = \frac{(SSR_r - SSR_u)/q}{SSR_u/(n-k-1)} \sim F_{q,n-k-1}$$

- q: the number of restrictions
- $n-k-1$: degrees of freedom of residuals

Questions

- Is the above F-statistic always positive?

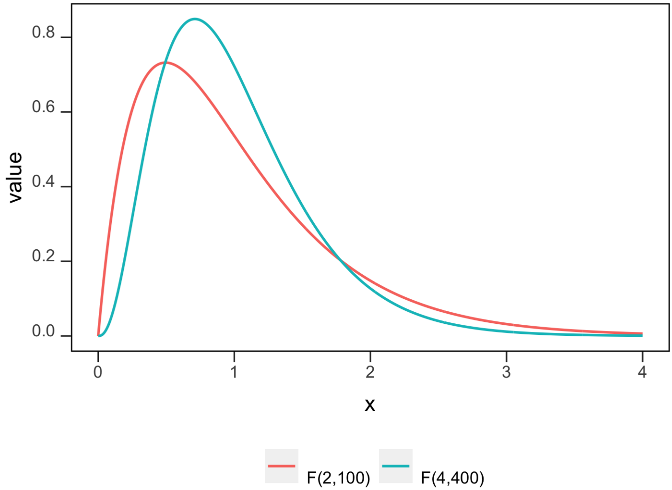
Yes, because $SSR_r - SSR_u$ is always positive.

- The greater the joint contribution of the q variables, the (greater or smaller) the F-statistic?

Greater.

F-distribution and F-test steps

F-distribution



F-test steps

- Define the null hypothesis
- Estimate the unrestricted and restricted models to obtains their SSR
- Calculate F-statistic
- Define the significance level and corresponding critical value according to the F distribution with appropriate degrees of freedoms
- Reject if your F-statistic is greater than the critical value, otherwise do not reject

F-test by hand

Step 1: estimate the unrestricted and restricted models

```r
#--- unrestricted model ---#
reg_u <- feols(log(salary) ~ years + gamesyr +
  bavg + hrunsyr + rbisyr, data = mlb_data)

SSR_u <- sum(reg_u$residuals^2)

#--- restricted model ---#
reg_r <- feols(log(salary) ~ years + gamesyr, data = mlb_data)

SSR_r <- sum(reg_r$residuals^2)
```

F-test by hand

Step 1: estimate the unrestricted and restricted models

```
#--- unrestricted model ---#
reg_u <- feols(log(salary) ~ years + gamesyr +
  bavg + hrunsyr + rbisyr, data = mlb_data)

SSR_u <- sum(reg_u$residuals^2)

#--- restricted model ---#
reg_r <- feols(log(salary) ~ years + gamesyr, data = mlb_data)

SSR_r <- sum(reg_r$residuals^2)
```

Step 2: calculate F-stat

```
df_q <- 3 # the number of restrictions
df_ur <- degrees_freedom(reg_u, "resid") # degrees of freedom for the unrestricted mo
F_stat_num <- (SSR_r - SSR_u) / df_q # numerator of F-stat
F_stat_denom <- SSR_u / df_ur # denominator of F-stat
F_sta <- F_stat_num / F_stat_denom # F-stat
F_sta
```

```
## [1] 9.550254
```

F-test by hand

Step 1: estimate the unrestricted and restricted models

```
#--- unrestricted model ---#
reg_u <- feols(log(salary) ~ years + gamesyr +
  bavg + hrunsyr + rbisyr, data = mlb_data)

SSR_u <- sum(reg_u$residuals^2)

#--- restricted model ---#
reg_r <- feols(log(salary) ~ years + gamesyr, data = mlb_data)

SSR_r <- sum(reg_r$residuals^2)
```

Step 2: calculate F-stat

```
df_q <- 3 # the number of restrictions
df_ur <- degrees_freedom(reg_u, "resid") # degrees of freedom for the unrestricted mo
F_stat_num <- (SSR_r - SSR_u) / df_q # numerator of F-stat
F_stat_denom <- SSR_u / df_ur # denominator of F-stat
F_sta <- F_stat_num / F_stat_denom # F-stat
F_sta
```

```
## [1] 9.550254
```

Step 3: find the critical value

```
alpha <- 0.05 # 5%  significance level
c_value <- qf(1 - alpha, df1 = df_q, df2 = df_ur)
c_value
```

```
## [1] 2.630641
```

F-test by hand

Step 1: estimate the unrestricted and restricted models

```
#--- unrestricted model ---#
reg_u <- feols(log(salary) ~ years + gamesyr +
  bavg + hrunsyr + rbisyr, data = mlb_data)

SSR_u <- sum(reg_u$residuals^2)

#--- restricted model ---#
reg_r <- feols(log(salary) ~ years + gamesyr, data = mlb_data)

SSR_r <- sum(reg_r$residuals^2)
```

Step 2: calculate F-stat

```
df_q <- 3 # the number of restrictions
df_ur <- degrees_freedom(reg_u, "resid") # degrees of freedom for the unrestricted mo
F_stat_num <- (SSR_r - SSR_u) / df_q # numerator of F-stat
F_stat_denom <- SSR_u / df_ur # denominator of F-stat
F_sta <- F_stat_num / F_stat_denom # F-stat
F_sta
```

```
## [1] 9.550254
```

Step 3: find the critical value

```
alpha <- 0.05 # 5%  significance level
c_value <- qf(1 - alpha, df1 = df_q, df2 = df_ur)
c_value
```

```
## [1] 2.630641
```

Step 4: F-stat > critical value?

```
F_sta > c_value
```

```
## [1] TRUE
```

F-test by hand

Step 1: estimate the unrestricted and restricted models

```
#--- unrestricted model ---#
reg_u <- feols(log(salary) ~ years + gamesyr +
  bavg + hrunsyr + rbisyr, data = mlb_data)

SSR_u <- sum(reg_u$residuals^2)

#--- restricted model ---#
reg_r <- feols(log(salary) ~ years + gamesyr, data = mlb_data)

SSR_r <- sum(reg_r$residuals^2)
```

Step 2: calculate F-stat

```
df_q <- 3 # the number of restrictions
df_ur <- degrees_freedom(reg_u, "resid") # degrees of freedom for the unrestricted mo
F_stat_num <- (SSR_r - SSR_u) / df_q # numerator of F-stat
F_stat_denom <- SSR_u / df_ur # denominator of F-stat
F_sta <- F_stat_num / F_stat_denom # F-stat
F_sta
```

```
## [1] 9.550254
```

Step 3: find the critical value

```
alpha <- 0.05 # 5%  significance level
c_value <- qf(1 - alpha, df1 = df_q, df2 = df_ur)
c_value
```

```
## [1] 2.630641
```

Step 4: F-stat > critical value?

```
F_sta > c_value
```

```
## [1] TRUE
```

So, the performance variables have statistically significant impacts on salary jointly!!

F-test by hand

Step 1: estimate the unrestricted and restricted models

```
#--- unrestricted model ---#
reg_u <- feols(log(salary) ~ years + gamesyr +
  bavg + hrunsyr + rbisyr, data = mlb_data)

SSR_u <- sum(reg_u$residuals^2)

#--- restricted model ---#
reg_r <- feols(log(salary) ~ years + gamesyr, data = mlb_data)

SSR_r <- sum(reg_r$residuals^2)
```

Step 2: calculate F-stat

```
df_q <- 3 # the number of restrictions
df_ur <- degrees_freedom(reg_u, "resid") # degrees of freedom for the unrestricted mo
F_stat_num <- (SSR_r - SSR_u) / df_q # numerator of F-stat
F_stat_denom <- SSR_u / df_ur # denominator of F-stat
F_sta <- F_stat_num / F_stat_denom # F-stat
F_sta
```

```
## [1] 9.550254
```

Step 3: find the critical value

```
alpha <- 0.05 # 5%  significance level
c_value <- qf(1 - alpha, df1 = df_q, df2 = df_ur)
c_value
```

```
## [1] 2.630641
```

Step 4: F-stat > critical value?

```
F_sta > c_value
```

```
## [1] TRUE
```

So, the performance variables have statistically significant impacts on salary jointly!!

What happened?

Simulation (multicollinearity, t-test, and F-test)

Data generation

```
N <- 300 # num observations
mu <- runif(N) # term shared by indep vars 1 and 2
x1 <- 0.1 * runif(N) + 2 * mu # indep 1
x2 <- 0.1 * runif(N) + 2 * mu # indep 2
x3 <- runif(N) # indep 3
u <- rnorm(N) # error
y <- 1 + x1 + x2 + x3 + u # generate y
data <- data.table(y = y, x1 = x1, x2 = x2, x3 = x3) # combine into a data.table
```

`x1` and `x2` are highly correlated with each other:

```
cor(x1, x2) # correlation between x1 and x2
```

```
## [1] 0.9974609
```

Regression

```
reg_u <- feols(y ~ x1 + x2 + x3, data = data) # OLS
tidy(reg_u) # results
```

```
## # A tibble: 4 × 5
##   term        estimate std.error statistic       p.value
##   <chr>          <dbl>     <dbl>     <dbl>         <dbl>
## 1 (Intercept)    0.830     0.145     5.74  0.0000000237
## 2 x1             1.89      1.31      1.44  0.150
## 3 x2             0.182     1.31      0.139 0.889
## 4 x3             1.18      0.188     6.27  0.00000000131
```

Both `x1` and `x2` are statistically insignificant individually.

Simulation (multicollinearity, t-test, and F-test)

F-test

```
#--- unrestricted ---#
SSR_u <- sum(reg_u$residuals^2)

#--- restricted ---#
reg_r <- feols(y ~ x3, data = data)
SSR_r <- sum(reg_r$residuals^2)

#--- degrees of freedom of residuals ---#
df_resid <- degrees_freedom(reg_u, "resid")

#--- F ---#
F_stat <- ((SSR_r - SSR_u) / 2) / (SSR_u / df_resid)

#--- critical value ---#
alpha <- 0.05
(
  c_value <- qf(1 - alpha, df1 = 2, df2 = df_resid)
)
```

```
## [1] 3.026257
```

The F-statistic for the hypothesis testing is:

```
#--- F > critical value? ---#
F_stat
```

```
## [1] 245.9419
```

```
F_stat > c_value
```

```
## [1] TRUE
```

The F-statistic is very high, meaning they collectively affect the dependent variable significantly.

Simulation (multicollinearity, t-test, and F-test)

F-test

```
#--- unrestricted ---#
SSR_u <- sum(reg_u$residuals^2)

#--- restricted ---#
reg_r <- feols(y ~ x3, data = data)
SSR_r <- sum(reg_r$residuals^2)

#--- degrees of freedom of residuals ---#
df_resid <- degrees_freedom(reg_u, "resid")

#--- F ---#
F_stat <- ((SSR_r - SSR_u) / 2) / (SSR_u / df_resid)

#--- critical value ---#
alpha <- 0.05
(
  c_value <- qf(1 - alpha, df1 = 2, df2 = df_resid)
)
```

```
## [1] 3.026257
```

The F-statistic for the hypothesis testing is:

```
#--- F > critical value? ---#
F_stat
```

```
## [1] 245.9419
```

```
F_stat > c_value
```

```
## [1] TRUE
```

The F-statistic is very high, meaning they collectively affect the dependent variable significantly.

Important

Standard error of the coefficients on `x1` and `x2` are very high because they are so highly correlated that your estimation of the model had such a difficult time to distinguish the their individual impacts.

But, collectively, they have large impacts. F-test was able to detect the statistical significance of their impacts collectively.

MLB example

Here is the correlation coefficients between the three variables:

```
select(mlb_data, bavg, hrunsyr, rbisyr) %>% cor()
```

```
##              bavg   hrunsyr    rbisyr
## bavg    1.0000000 0.1905958 0.3291454
## hrunsyr 0.1905958 1.0000000 0.8907428
## rbisyr  0.3291454 0.8907428 1.0000000
```

MLB example

Here is the correlation coefficients between the three variables:

```
select(mlb_data, bavg, hrunsyr, rbisyr) %>% cor()
```

```
##              bavg   hrunsyr    rbisyr
## bavg    1.0000000 0.1905958 0.3291454
## hrunsyr 0.1905958 1.0000000 0.8907428
## rbisyr  0.3291454 0.8907428 1.0000000
```

As you can see, `brunsyr` and `hrunsyr` are highly correlated with each other.

They are not so highly correlated with `bavg`.

F-test (easier way)

You can use the `linearHypothesis()` function from the `car` package.

Syntax

```
linearHypothesis(regression, hypothesis)
```

`regression` is the name of the regression results of the unrestricted model

`hypothesis` is a text of null hypothesis:

example

`c("x1 = 0", "x2 = 1")` means the coefficients on $x1$ and $x2$ are 0 and 1, respectively

Demonstration

```
#--- load the car package ---#
library(car)

#--- unrestricted regression ---#
reg_u <- feols(y ~ x1 + x2 + x3, data = data)

#--- F-test ---#
linearHypothesis(reg_u, c("x1=0", "x2=0"))
```

```
## Linear hypothesis test
##
## Hypothesis:
## x1 = 0
## x2 = 0
##
## Model 1: restricted model
## Model 2: y ~ x1 + x2 + x3
##
##   Df  Chisq Pr(>Chisq)
## 1
## 2  2 491.88  < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Multiple coefficients again

- The test of a linear combination of the parameters we looked at earlier is a special case where the number of restriction is 1.
- It can be shown that square root of $F_{1,df}$ follows the t_{df} distribution.
- So, we can actually use F-test for this type of hypothesis testing because $F_{1,t-n-k} \sim t^2_{t-n-k}$.

```
#--- OLS with lm() ---#
reg_sc <- feols(lwage ~ jc + univ + exper, data = twoyear) # OLS

#--- F-test ---#
F_test <- linearHypothesis(reg_sc, c("jc-univ=0"))

#--- t-statistic ---#
sqrt(F_test$Chisq[2])
```

```
## [1] 1.467657
```

Check with the previous slide of the same t-test and confirm that the t-statistics we got there and here are the same.