Statistical Testing

Taro Mieno

AECN 896-003: Applied Econometrics

Hypothesis Testing: Lecture Outline

1. examples of hypotheses
2. additional assumption we need to make to perform statistical performance
3. the distribution of the OLS estimators in the population
4. t-distribution and t-statistic
5. hypothesis testing (single parameter)
   - two-sided
   - one-sided
6. hypothesis testing (multiple parameter)

Hypothesis Testing: Examples

Consider the following model,

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

Hypotheses examples:

Hypothesis 1 : education has no impact on wage ($\beta_1 = 0$)

Hypothesis 2 : experience has a positive impact on wage ($\beta_2 > 0$)

Education has no impact on wage ($\beta_1 = 0$)

- 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

Education has no impact on wage ($\beta_1 = 0$)

- 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

Education has no impact on wage ($\beta_1 = 0$)

- 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 $\hat{\beta}$

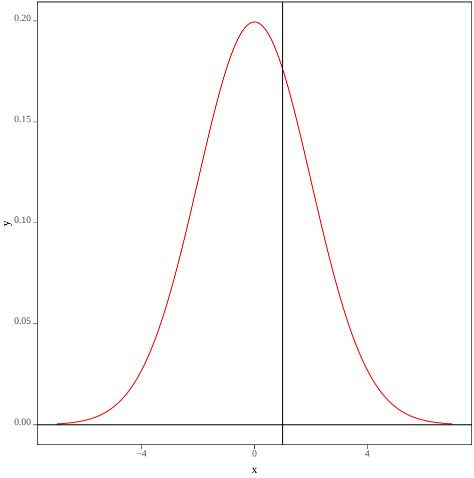
Hypothesis Testing

Education example

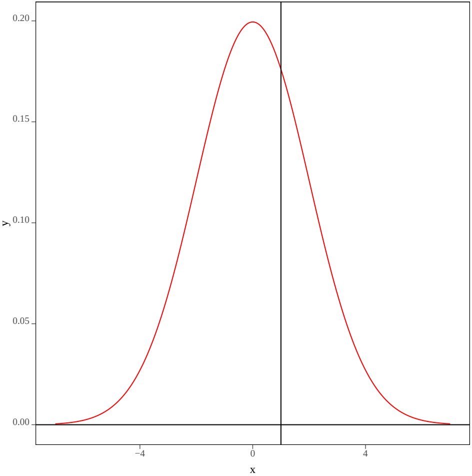
You have gotten an estimate of the impact of education on income ($\hat{\beta}_1$) and are wondering if the true value of β (which you will never know) is 0.

- What would be the distribution of $\hat{\beta}$ (the estimator) if the true value of β is indeed 0?
- If so, how likely that you would have gotten the value you have gotten for $\hat{\beta}$

Distribution of $\hat{\beta}_1$ if $\beta_1 = 0$



Distribution of $\hat{\beta}_1$ if $\beta_1 = 0$



Question

Would you say $\hat{\beta}_1$ is different from 0? Let's formalize this process in a statistical manner.

Hypothesis Testing

So far,

we learned how to find, under certain conditions:

- Expected value of the OLS estimators ($MLR.1 \sim MLR.4$)
- Variance of the OLS estimators ($MLR.1 \sim MLR.5$)

Important

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

Now,

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

A popular (the) choice of distribution is (mostly out of convenience),

MLR.6: Normality

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:

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

Normality Assumption

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

Independence implies,

$$
\begin{aligned}
E[u|x] =& 0 \\
Var[u|x] =& \sigma^2
\end{aligned}
$$

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

Normality Assumption

Does the normality of error term hold in practice?

- It almost always does NOT hold
- It may be a good approximation of the true distribution of the error term
- It is an empirical matter (you may or may not depending on what problem you are working on).
- It seems more realistic than some other distributions like uniform distribution

Classical Linear Model (CLM) assumption

- Assumptions $MLR.1$ through $MLR.6$ are called collectively the classical linear model (CLM) assumption.
- Under this assumption, OLS can be shown to be the minimum variance unbiased estimators (including not only linear, but non-linear estimators)

Note:

This theorem has almost no practical relevance. You can forget about this theorem.

Under the Classical Linear Model (CLM) assumption

The distribution of y conditional on x is a Normal distribution

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

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

Under the CLM assumption

If the CLM assumption is satisfied, OLS estimator also has a Normal distribution!

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

which means,

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

t-distribution and t-statistic

In the population,

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

But, in practice, we need to estimate $sd(\beta_j)$. If we use $se(\hat{\beta}_j)$ (an estimator of $sd(\hat{\beta}_j)$) instead of $sd(\hat{\beta}_j)$ (, which you will never know), then,

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

where $n - k - 1$ is the degree of freedom when $sd(\hat{\beta}_j)$ is estimated.

Hypothesis Testing: General Steps

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

Hypothesis testing about a single parameter

One-sided Alternative

H_1: $\beta_j > 0$

Hypothesis testing about a single parameter

One-sided Alternative

H_1: $\beta_j > 0 \Rightarrow H_0$: $\beta_j \leq 0$

Hypothesis testing about a single parameter

One-sided Alternative

H_1: $\beta_j > 0 \Rightarrow H_0$: $\beta_j \leq 0$

But,

The null value that is hardest to reject in favor of the H_1 is $\beta_j = 0$. That is, if we reject $\beta_j = 0$ in favor of the H_1, you will automatically reject other values of $\beta_j < 0$. This means, it is sufficient to test $H_0 : \beta_j = 0$ against $H_1 : \beta_j > 0$.

Hypothesis testing about a single parameter

One-sided Alternative

H_1: $\beta_j > 0 \Rightarrow H_0$: $\beta_j \leq 0$

But,

The null value that is hardest to reject in favor of the H_1 is $\beta_j = 0$. That is, if we reject $\beta_j = 0$ in favor of the H_1, you will automatically reject other values of $\beta_j < 0$. This means, it is sufficient to test $H_0 : \beta_j = 0$ against $H_1 : \beta_j > 0$.

So,

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

One-sided Alternative

We learned that,

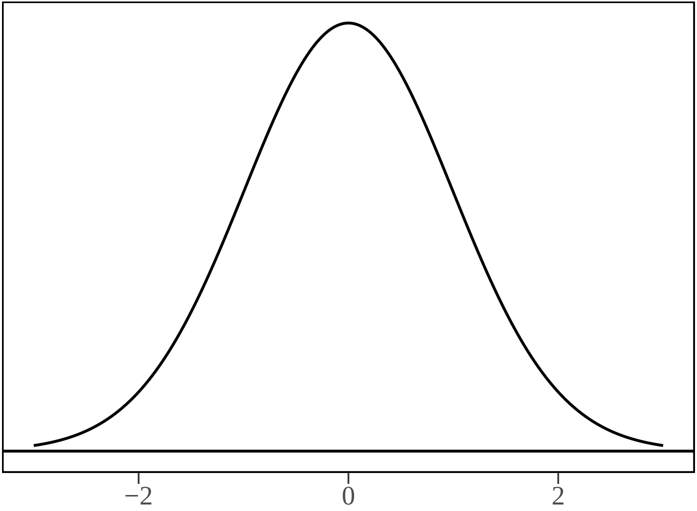
$$\frac{\hat{\beta}_j - \beta_j}{se(\hat{\beta}_j)} \sim t_{n-k-1}$$

Under the null ($\beta_j = 0$),

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

Visualization of the distribution of $\hat{\beta}_j / se(\hat{\beta}_j)$

Figure: The distribution of t_{90} ($N = 100$ and $k = 9$)



Define the significance level: Step 3

Significance Level

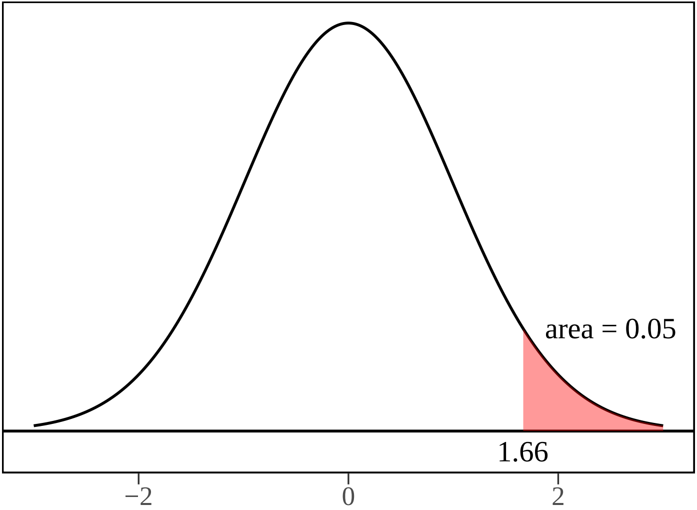
The probability of rejecting the null when the null is actually true (The probability that you wrongly cliam that the null hypothesis is wrong even though it's true in reality: Type I error)

So,

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

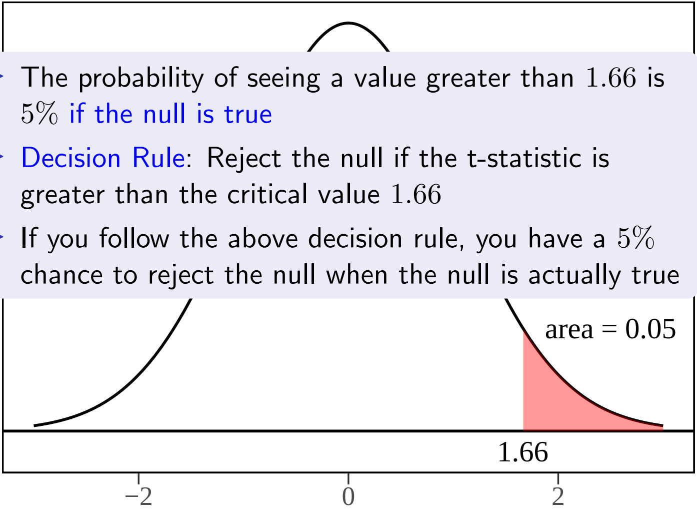
Significance Level

Figure: $\alpha = 0.05$



Significance Level

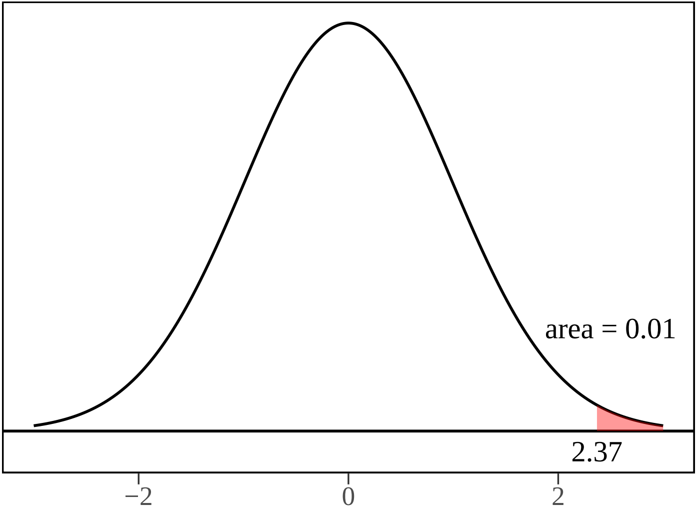
Figure: $\alpha = 0.05$



- The probability of seeing a value greater than 1.66 is 5% if the null is true
- Decision Rule: Reject the null if the t-statistic is greater than the critical value 1.66
- If you follow the above decision rule, you have a 5% chance to reject the null when the null is actually true

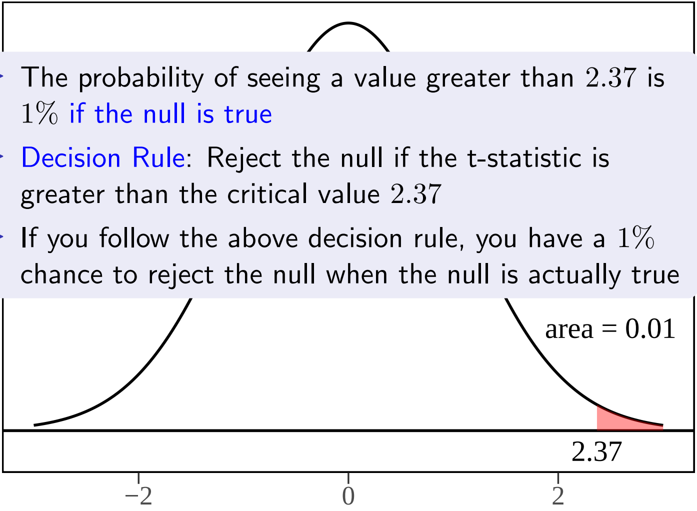
Significance Level

Figure: $\alpha = 0.01$



Significance Level

Figure: $\alpha = 0.01$



- The probability of seeing a value greater than 2.37 is 1% if the null is true
- Decision Rule: Reject the null if the t-statistic is greater than the critical value 2.37
- If you follow the above decision rule, you have a 1% chance to reject the null when the null is actually true

Steps 4 and 5

Step 4 : Plug in actual numbers into $\frac{\hat{\beta}_j}{se(\hat{\beta}_j)}$ to obtain the t-statistic.

Step 5 : Follow the decision rule you specified to determine whether you should reject or not reject the null

An Example

The impact of experience on wage

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

Hypothesis

H_0: $\beta_{exper} = 0$

H_1: $\beta_{exper} > 0$

Test

- $t = 0.0041/0.0017 = 2.41$
- $2.41 > c_{0.01} = 2.33$?

An Example

Test Results and Conclusion

Test Results :$2.41 > c_{0.01} = 2.33$

Conclusion : $\hat{\beta}_{exper}$ is statistically greater than zero at the 1% significance level.:

Two-sided Alternatives

Two-sided Alternative

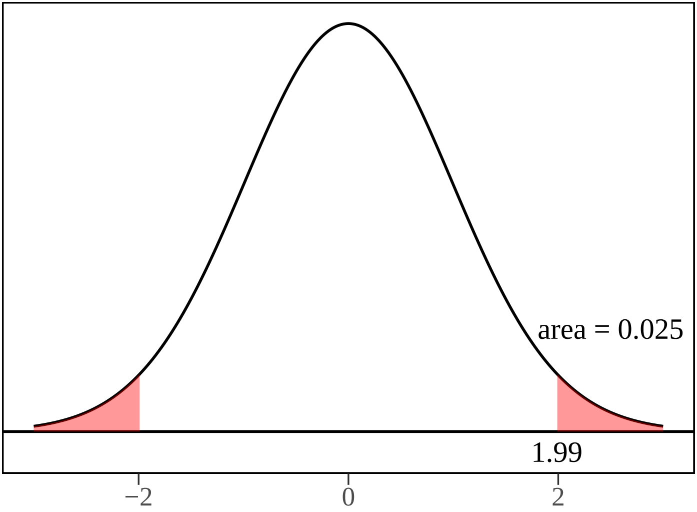
H_1: $\beta_j \neq 0$ (positive or negative not specified)

Null

H_0: $\beta_j = 0$

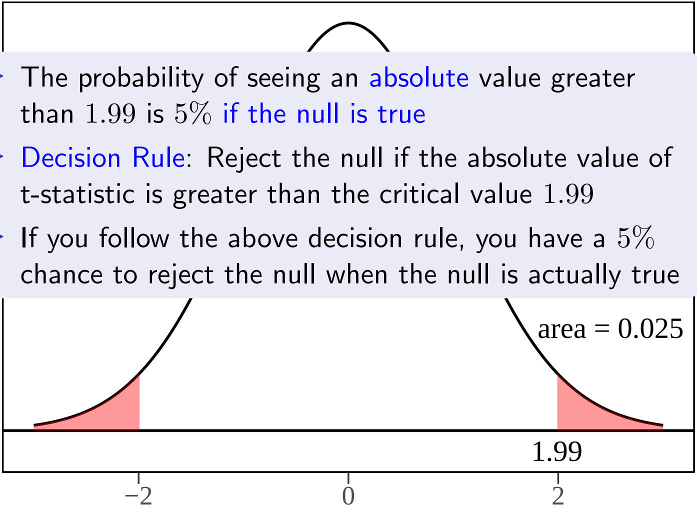
Two-sided Alternatives

Figure: $\alpha = 0.05$



Two-sided Alternatives

Figure: $\alpha = 0.05$



- The probability of seeing an absolute value greater than 1.99 is 5% if the null is true
- Decision Rule: Reject the null if the absolute value of t-statistic is greater than the critical value 1.99
- If you follow the above decision rule, you have a 5% chance to reject the null when the null is actually true

R Implementation

Model

$$
\begin{aligned}
wage =& \beta_0 + \beta_1 \times educ + \beta_2 \times exper \\
& + \beta_3 \times tenure + u
\end{aligned}
$$

Hypothesis

$$
\begin{aligned}
H_0 &: \beta_1 = 0 \\
H_1 &: \beta_1 \neq 0
\end{aligned}
$$

R Implementation

R code: Import wage data

```
wage <- readRDS('wage1.rds') %>% data.table()
wage[,.(wage,educ,exper,female,married)]
```

```
      wage educ exper female married
  1:  3.10   11     2      1       0
  2:  3.24   12    22      1       1
  3:  3.00   11     2      0       0
  4:  6.00    8    44      0       1
  5:  5.30   12     7      0       1
 ---
522: 15.00   16    14      1       1
523:  2.27   10     2      1       0
524:  4.67   15    13      0       1
525: 11.56   16     5      0       1
526:  3.50   14     5      1       0
```

R Implementation

R code: Hypothesis Testing

```
  reg_wage <- summary(lm(wage~educ+exper+tenure,data=wage))
  reg_wage$coef
```

```
              Estimate Std. Error   t value     Pr(>|t|)
(Intercept) -2.87273482 0.72896429 -3.940844 9.224742e-05
educ         0.59896507 0.05128355 11.679478 3.681353e-28
exper        0.02233952 0.01205685  1.852849 6.446818e-02
tenure       0.16926865 0.02164461  7.820361 2.934527e-14
```

```
  #--- calculate t-statistic ---#
  beta_educ <- reg_wage$coef[2,1] # coefficient estimate on educ
  se_beta_educ <- reg_wage$coef[2,2] # se of the coefficient on educ
  t <- beta_educ/se_beta_educ # t-statistic
  t
```

```
[1] 11.67948
```

R Implementation

R code: Hypothesis Testing

```
  #--- degree of freedom for t-distribution ---#
  df <- reg_wage$df[2]
  df
```

```
[1] 522
```

```
  #--- specify significance level ---#
  alpha <- 0.05

  #--- find the critical value ---#
  c_value <- qt(alpha/2,df) %>% abs()
  c_value
```

```
[1] 1.964519
```

```
  #--- follow the decision rule ---#
  t>c_value
```

```
[1] TRUE
```

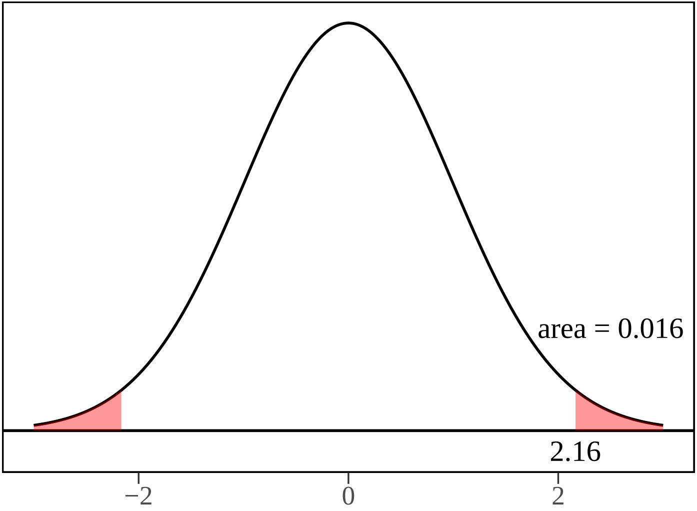
$p - value$

$p - value$

- 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)

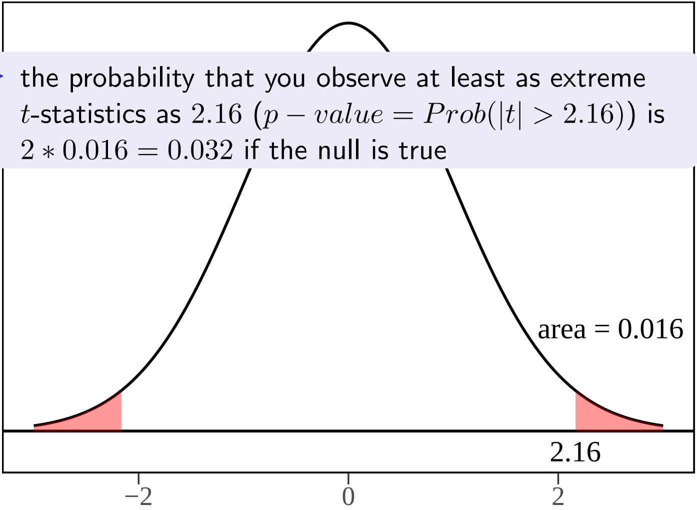
p-value: two-sided alternative

Figure: $t = 2.16$, $df = 522$



p-value: two-sided alternative

Figure: $t = 2.16$, $df = 522$



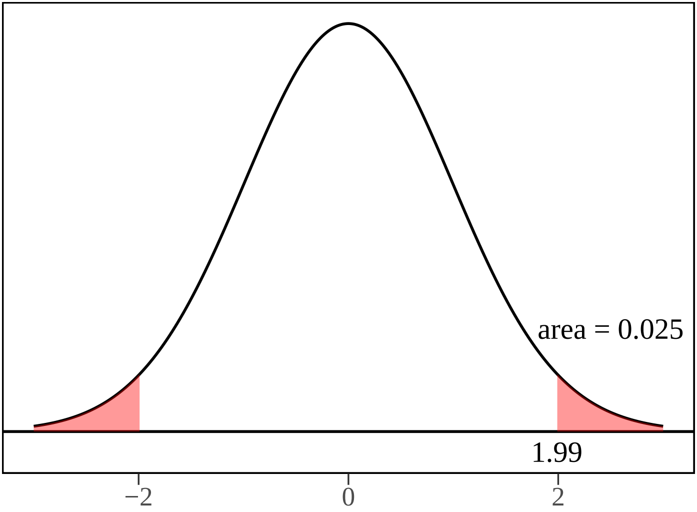
- the probability that you observe at least as extreme t-statistics as 2.16 ($p - value = Prob(|t| > 2.16)$) is $2 * 0.016 = 0.032$ if the null is true

Decision rule using p-value

- if the $p-value$ is smaller (greater) than the significance level, you reject (not reject) the null hypothesis
- whether you use $t-statistics$ or $p-value$, you are going to reach the same conclusion

Decision rule using p-value

Figure: $t = 2.16$, $df = 522$



Statistical vs. Economic Significance

Important

Statistical significance is NOT economic significance

Statistical vs. Economic Significance

R code: statistical and economic significance

```
set.seed(23478)
N <- 300000
glasses <- runif(N)*40 # years wearing glasses
u <- 0.1*rnorm(N) # error
income <- 0.001*glasses+u # annual income
data <- data.table(x=glasses,y=income)
reg <- summary(lm(y~x,data=data))
reg$coef
```

```
                Estimate   Std. Error   t value   Pr(>|t|)
(Intercept) 0.0006568871 3.649371e-04  1.800001 0.07186155
x           0.0009793425 1.580148e-05 61.977892 0.00000000
```

Statistical vs. Economic Significance

So,

- If you have been wearing glasses for 10 years, your annual income is higher by $\$0.009793$, which is statistically significant at the 0.00000001% level!!
- Would you care?

Statistical vs. Economic Significance

So,

- If you have been wearing glasses for 10 years, your annual income is higher by $\$0.009793$, which is statistically significant at the 0.00000001% level!!
- Would you care?

Important

Do not confuse statistical significance with economic significance!

Statistical vs. Economic Significance

R code: statistical and economic significance

```
set.seed(23478)
N <- 300000
glasses <- runif(N)*40 # years wearing glasses
u <- 0.1*rnorm(N) # error
income <- 0.001*glasses+u # annual income
data <- data.table(x=glasses,y=income)
reg <- summary(lm(y~x,data=data))
```

Question

What do you notice about the data generating process?

Confidence Intervals (CI)

Confidence Interval (CI)

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

Confidence Intervals (CI)

Confidence Interval (CI)

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

What CI is NOT

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

Confidence Interval

Under $MLR.1$ through $MLR.6$

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

Steps to calculate confidence interval

1. set the confidence level, say $\alpha\%$ (95%)
2. find the $1 - (1 - \alpha)/2$ (97.5%) quantile of t_{n-k-1}, call it c
3. set the upper and lower bounds as $\hat{\beta}_j + c \times se(\hat{\beta}_j)$ and $\hat{\beta}_j - c \times se(\hat{\beta}_j)$

Confidence Interval

R code: Confidence Interval

```
#--- 1. set the CI level ---#
alpha <- 0.95

#--- 2. find the appropriate percentile ---#
df <- reg_wage$df[2] # n-k-1
cons <- qt(1-(1-alpha)/2,df=df)

#--- 3. calculate the upper and lower bounds of the CI ---#
beta_hat <- reg_wage$coef[2,1] # coef on educ
se_beta <- reg_wage$coef[2,2] # se of the coef on educ
upper <- beta_hat+cons*se_beta
lower <- beta_hat-cons*se_beta
c(lower,upper)
```

```
[1] 0.4982176 0.6997126
```

Testing Hypotheses about a Single Linear Combination of the Parameters

Consider the following model

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

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

Testing Hypotheses about a Single Linear Combination of the Parameters

Consider the following model

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

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

Question

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

Testing Hypotheses about a Single Linear Combination of the Parameters

Hypothesis

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

Testing Hypotheses about a Single Linear Combination of the Parameters

Hypothesis

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

Rewriting the hypotheses

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

Testing Hypotheses about a Single Linear Combination of the Parameters

Hypothesis

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

Rewriting the hypotheses

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

Let α denote $\beta_1 - \beta_2$ (looks a lot more familiar?)

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

Testing Hypotheses: Linear Combination of the Parameters

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

Testing Hypotheses: Linear Combination of the Parameters

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

numerator

It is easy to calculate it. Just plug in the coefficient estimates

Testing Hypotheses: Linear Combination of the Parameters

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

numerator

It is easy to calculate it. Just plug in the coefficient estimates

denominator

math aside

$$\begin{aligned} se(\hat{\beta}_1 - \hat{\beta}_2) =& \sqrt{Var(\hat{\beta}_1 - \hat{\beta}_2)} \left(\neq \sqrt{Var(\hat{\beta}_1) + Var(\hat{\beta}_2)} \right) \\ =& \sqrt{Var(\hat{\beta}_1) - 2Cov(\hat{\beta}_1, \hat{\beta}_2) + Var(\hat{\beta}_2)} \end{aligned}$$

Testing Hypotheses: Linear Combination of the Parameters

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

numerator

It is easy to calculate it. Just plug in the coefficient estimates

denominator

math aside

$$\begin{aligned} se(\hat{\beta}_1 - \hat{\beta}_2) =& \sqrt{Var(\hat{\beta}_1 - \hat{\beta}_2)}\Big(\neq \sqrt{Var(\hat{\beta}_1) + Var(\hat{\beta}_2)}\Big) \\ =& \sqrt{Var(\hat{\beta}_1) - 2Cov(\hat{\beta}_1, \hat{\beta}_2) + Var(\hat{\beta}_2)} \end{aligned}$$

test

if the calculated $t\text{-}statistics$ is smaller (greater) than the critical value for your choice of significance level, you reject (do not reject) the null.

Testing Hypotheses: Linear Combination of the Parameters

R code: Testing

```
twoyear <- readRDS('twoyear.rds') # import data
reg_sc <- lm(lwage~jc+univ+exper,data=twoyear) # OLS

#--- get the variance covariance matrix of coefficient estimators ---#
vcov_sc <- vcov(reg_sc) # variance covariance matrix
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
```

Testing Hypotheses: Linear Combination of the Parameters

R code: Testing

```
twoyear <- readRDS('twoyear.rds') # import data
reg_sc <- lm(lwage~jc+univ+exper,data=twoyear) # OLS

#--- get the variance covariance matrix of coefficient estimators ---#
vcov_sc <- vcov(reg_sc) # variance covariance matrix
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
```

Variance Covariance Matrix

- $VCOV_{i,i}$: the variance of ith variable
- $VCOV_{i,j}$: the covariance between ith and $j$the variables

Testing Hypotheses: Linear Combination of the Parameters

denominator

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

R code: Testing

```
numerator <- reg_sc$coef['jc']-reg_sc$coef['univ']
denomenator <- sqrt(
  vcov_sc['jc','jc']-2*vcov_sc['jc','univ']+vcov_sc['univ','univ']
  )
t_stat <- numerator/denomenator
t_stat
```

```
       jc
-1.467657
```

Testing Multiple Linear Restrictions: The F-test

When we want to test multiple hypotheses at the same time, we use F-test.

An Example

Salary of major league baseball players

$$
\begin{aligned}
log(salary) = &\beta_0 + \beta_1 years + \beta_2 gamesyr + \beta_3 bavg \\
&+ \beta_4 hrunsyr + \beta_5 rbisyr + u
\end{aligned}
$$

- $salary$: 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

An Example

Hypotheses

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.

Mathematically

H_0: $\beta_3 = 0$, $\beta_4 = 0$, and $\beta_5 = 0$ H_1: H_0 is not true

An Example

Mathematically

H_0: $\beta_3 = 0$, $\beta_4 = 0$, and $\beta_5 = 0$ H_1: H_0 is not true

How do we test this?

- The alternative H_1 holds if at least one of β_3, β_4, or β_5 is different from zero.
- Conduct t-test for each coefficient individually?

Regression

R code: MLB salary

```r
library(readstata13)
temp_data <- read.dta13('MLB1.dta')
mlb_data <- data.table(temp_data)
reg_1 <- summary(lm(log(salary)~years+gamesyr+bavg
  +hrunsyr+rbisyr,data=mlb_data))
reg_1$coef
```

```
              Estimate  Std. Error    t value      Pr(>|t|)
(Intercept) 1.119242e+01 0.288822851 38.7518448 4.187260e-128
years       6.886264e-02 0.012114543  5.6842952  2.787579e-08
gamesyr     1.255212e-02 0.002646763  4.7424408  3.088620e-06
bavg        9.786036e-04 0.001103509  0.8868108  3.757950e-01
hrunsyr     1.442947e-02 0.016056977  0.8986417  3.694667e-01
rbisyr      1.076573e-02 0.007174960  1.5004590  1.344049e-01
```

Individually,

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

F-test

Individually,

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

F-test

Individually,

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

But,

- 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

We compare sum of squared residuals (SSR) of two models:

Unrestricted Model

$$\begin{aligned} log(salary) =& \beta_0 + \beta_1 years + \beta_2 gamesyr + \beta_3 bavg \\ &+ \beta_4 hrunsyr + \beta_5 rbisyr + u \end{aligned}$$

Restricted

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

The coefficients on $bavg$, $hrunsyr$, and $rbisyr$ are restricted to be 0.

F-test

- Let SSR_u and SSR_r denote the SSR for the unrestricted and restricted models, respectively

R code: SSR comparison

```
res_u <- lm(log(salary)~years+gamesyr+bavg
  +hrunsyr+rbisyr,data=mlb_data)$residuals
sum(res_u^2)
```

```
[1] 183.1863
```

```
res_r <- lm(log(salary)~years+gamesyr,data=mlb_data)$residuals
sum(res_r^2)
```

```
[1] 198.3115
```

F-test

- Let SSR_u and SSR_r denote the SSR for the unrestricted and restricted models, respectively
- Which SSR_u or SSR_r is larger?

R code: SSR comparison

```
res_u <- lm(log(salary)~years+gamesyr+bavg
  +hrunsyr+rbisyr,data=mlb_data)$residuals
sum(res_u^2)
```

```
[1] 183.1863
```

```
res_r <- lm(log(salary)~years+gamesyr,data=mlb_data)$residuals
sum(res_r^2)
```

```
[1] 198.3115
```

F-test

- Let SSR_u and SSR_r denote the SSR for the unrestricted and restricted models, respectively
- Which SSR_u or SSR_r is larger?

R code: SSR comparison

```
res_u <- lm(log(salary)~years+gamesyr+bavg
  +hrunsyr+rbisyr,data=mlb_data)$residuals
sum(res_u^2)
```

```
[1] 183.1863
```

```
res_r <- lm(log(salary)~years+gamesyr,data=mlb_data)$residuals
sum(res_r^2)
```

```
[1] 198.3115
```

- What does SSR_r - SSR_u measure?

F-test

- Let SSR_u and SSR_r denote the SSR for the unrestricted and restricted models, respectively
- Which SSR_u or SSR_r is larger?

R code: SSR comparison

```
res_u <- lm(log(salary)~years+gamesyr+bavg
  +hrunsyr+rbisyr,data=mlb_data)$residuals
sum(res_u^2)
```

```
[1] 183.1863
```

```
res_r <- lm(log(salary)~years+gamesyr,data=mlb_data)$residuals
sum(res_r^2)
```

```
[1] 198.3115
```

- What does SSR_r - SSR_u measure?
- Is the contribution from the restricted variables big enough?

F-test: generally

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-test: generally

F-statistic

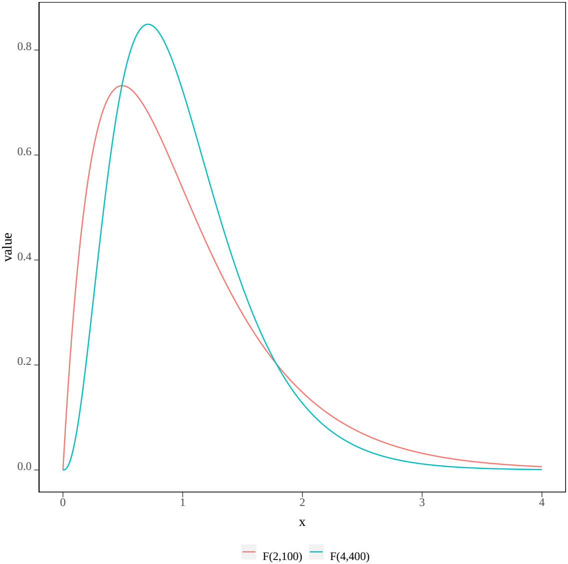
$$F = \frac{(SSR_r - SSR_u)/q}{SSR_u/(n-k-1)} \sim F_{q,n-k-1}$$

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

Questions

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

F-distribution



F-test: Steps

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

Going back to the example,

R code: Steps 1-3

```
#--- Step 2: unrestricted model estimation ---#
reg_u <- lm(log(salary)~years+gamesyr+
  bavg+hrunsyr+rbisyr,data=mlb_data)
SSR_u <- sum(reg_u$residuals^2)

#--- Step 2: restricted model estimation ---#
reg_r <- lm(log(salary)~years+gamesyr,data=mlb_data)
SSR_r <- sum(reg_r$residuals^2)

#--- Step 3: calculate F-stat ---#
df_q <- 3 # the number of restrictions
df_ur <- reg_u$df.residual # degrees of freedom for the unrestricted model
F_stat_num <- (SSR_r-SSR_u)/df_q
F_stat_denom <- SSR_u/df_ur
F_sta <- F_stat_num/F_stat_denom
F_sta
```

```
[1] 9.550254
```

Going back to the example,

R code: Steps 1-3

```
#--- Step 4: 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 5: F> critical value? ---#
F_sta > c_value
```

```
[1] TRUE
```

So,

The performance variables have statistically significant impacts on salary jointly

R code: F-test

```r
set.seed(48937) # set seed
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
cor(x1,x2) # correlation between x1 and x2
```

```
[1] 0.9977728
```

```r
u <- rnorm(N) # error
y <- 1 + x1 + x2 + x3 + u # generate y
data <- data.table(y=y,x1=x1,x2=x2) # combine into a data.table
reg_u <- lm(y~x1+x2+x3,data=data) # OLS
summary(reg_u)$coef # results
```

```
             Estimate Std. Error   t value     Pr(>|t|)
(Intercept) 0.9302513  0.1530575 6.0777898 3.745233e-09
x1          1.2921380  1.5295875 0.8447624 3.989258e-01
x2          0.8783883  1.5135746 0.5803403 5.621267e-01
x3          1.0827919  0.2105412 5.1428970 4.929468e-07
```

R code: F-test

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

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

#--- F ---#
F_stat <- ((SSR_r-SSR_u)/2)/(SSR_u/reg_u$df.residual)
F_stat
```

```
[1] 227.7407
```

```r
#--- critical value ---#
alpha <- 0.05
c_value <- qf(1-alpha,df1=2,df2=reg_u$df.residual)
c_value
```

```
[1] 3.026257
```

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

```
[1] 227.7407
```

```r
F_stat > c_value
```

```
[1] TRUE
```

What happened?

- Due to multicollinearity between x_1 and x_2, it is hard to distinguish their impacts individually
- But, collectively, they have large impacts. F-test was able to detect the statistical significance of their impacts collectively

MLB example

R code: Correlations

```
dplyr::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
```

R-squared form of F-statistic

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

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

Remember $R^2 = 1 - SSR/SST \Rightarrow SSR = SST(1-R^2)$. So,

$$\begin{aligned}
F =& \frac{(SSR_r - SSR_u)/q}{SSR_u/(n-k-1)} \\
=& \frac{(SST(1-R_r^2) - SST(1-R_u^2))/q}{SST(1-R_u^2)/(n-k-1)} \\
=& \frac{((1-R_r^2) - (1-R_u^2))/q}{(1-R_u^2)/(n-k-1)} \\
=& \frac{(R_u^2 - R_r^2)/q}{(1-R_u^2)/(n-k-1)}
\end{aligned}$$

F-test using R^2

R code: F-test

```
#--- unrestricted ---#
reg_u_sum <- summary(reg_u)
R2_u <- reg_u_sum$r.squared

#--- restricted ---#
reg_r_sum <- summary(reg_r)
R2_r <- reg_r_sum$r.squared

#--- F ---#
F_stat <- ((R2_u-R2_r)/2)/((1-R2_u)/reg_u$df.residual)
F_stat
```

```
[1] 227.7407
```

Simpler implementation of F-test in R

You can use the **linearHypothesis()** function from the **car** package

$$linearHypothesis(regression, hypothesis)$$

- $regression$: the name of regression results (unrestricted model)
- $hypothesis$: a text of null hypothesis:
  - Ex. c('x1=0','x2=1') means the coefficients on $x1$ and $x2$ are 0 and 1, respectively

R code: F-test using the *car* package

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

#--- unrestricted regression ---#
reg_u <- lm(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

  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1    298 773.37
2    296 304.62  2    468.75 227.74 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Linear combination of parameters, 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
- We can still do F-test for this type of hypothesis testing
- Indeed, $F_{1,t-n-k} \sim t^2_{t-n-k}$.

R code: multiple coefficients (1 restriction)

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

#--- F-test ---#
F_res <- linearHypothesis(reg_sc,c('jc-univ=0'))
F_res
```

```
Linear hypothesis test

Hypothesis:
jc - univ = 0

Model 1: restricted model
Model 2: lwage ~ jc + univ + exper

  Res.Df    RSS Df Sum of Sq     F Pr(>F)
1   6760 1250.9
2   6759 1250.5  1   0.39853 2.154 0.1422
```

```
#--- F-stat ---#
sqrt(F_res$F)
```

```
[1]       NA 1.467657
```

Math Aside 1

go back

Variance

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

Example

$a = 2$ and $b = -1$,

$$Var(x - y) = 4Var(x) - 4Cov(x, y) + Var(y)$$