4 Testing Hypotheses about a Single Linear Combination of the Parameter

Consider

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

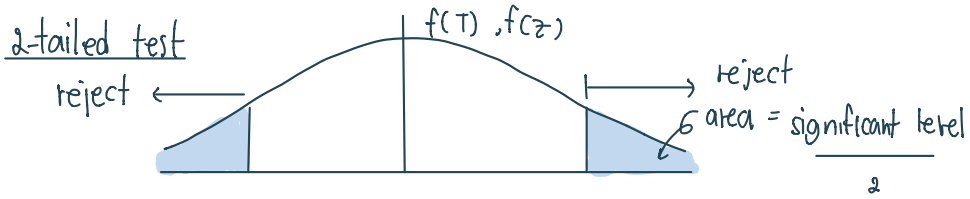
where jc = number of years attending a two-year college
$univ$ = number of years at a four-year college
$\exp er$ = months in the workforce.
We want to test whether $\beta_1 = \beta_2$.

If the returns from 1 more years of education at a junior college is the same as that of the university

$H_0 : \beta_1 = \beta_2 \Rightarrow H_0 : \beta_1 - \beta_2 = 0$

against

$H_a : \beta_1 \neq \beta_2 \Rightarrow H_a : \beta_1 - \beta_2 \neq 0$



$t = \dfrac{(\hat{\beta}_1 - \hat{\beta}_2) - 0}{se(\hat{\beta}_1 - \hat{\beta}_2)}$ } we compute this t statistic and compare with the critical value.

where $se(\hat{\beta}_1 - \hat{\beta}_2) = \sqrt{var(\hat{\beta}_1 - \hat{\beta}_2)}$

$= \sqrt{var(\hat{\beta}_1) + var(\hat{\beta}_2) - 2\,cov(\hat{\beta}_1, \hat{\beta}_2)}$

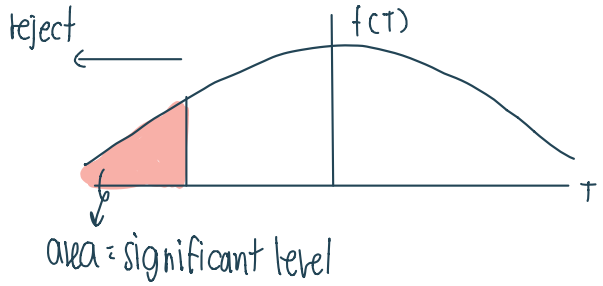
Not very straightforward to calculate
→ We use a variable transformation trick

another possible hypothesis test (one-tailed alternative)

$H_0 : \beta_1 - \beta_2 \Rightarrow H_0 : \beta_1 - \beta_2 = 0$

$H_a : \beta_1 < \beta_2 \Rightarrow H_a : \beta_1 - \beta_2 < 0$

• it is assume that β_1 would not be more than β_2 (return to 2-years college would never be more than return to university education)

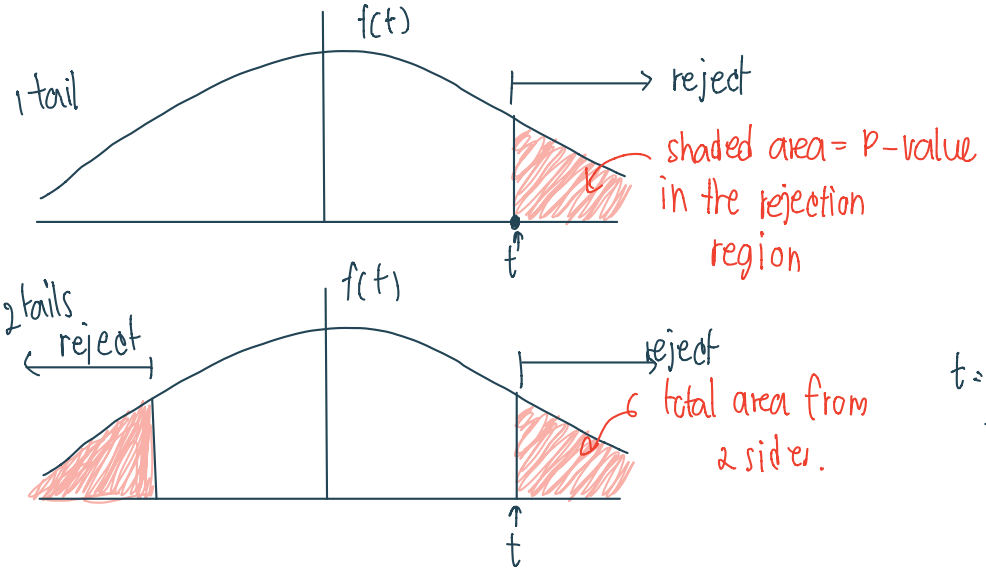


$$t = \frac{(\hat{\beta}_1 - \hat{\beta}_2) - 0}{se(\hat{\beta}_1 - \hat{\beta}_2)}$$

* go to extra note 11

5 Computing p-Values for t-Tests

- What is the significance level given the computed t-statistics?



$$t = \frac{\hat{\beta}_j - \beta_j}{se\,\hat{\beta}_j}$$

- p-value : $P(|T| > |t|)$

T : t-distributed random variable with d.f. = n-k-1

t : computed t-statistic

⇝ p-value = probability that a random T value will greater (in the || term) than our t in H_0 testing.

In class exercise

Consider the multiple regression model, assume MLR 1-6 are satisfied

$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + u$: you would like to test $H_0 : \beta_1 - 3\beta_2 = 1$

you would like to test the null hypothesis (H_0) H_a : otherwise is true.

1. Write the T statistic for testing the null hypothesis

$$t = \frac{(\hat{\beta}_1 - 3\hat{\beta}_2) - 1}{se(\hat{\beta}_1 - 3\hat{\beta}_2)}$$

2. Define $\theta_1 = \hat{\beta}_1 - 3\hat{\beta}_2 \Rightarrow H_0 : \theta_1 - 1 , H_a : \theta \neq 1$

$t = \frac{\hat{\theta}_1 - 1}{se(\hat{\theta}_1)}$ ⇒ we need our regression to have θ_1 in it. So, STATA or OLS estimation will automatically give $\hat{\theta}_1$ & $se\hat{\theta}_1$

Now, $\hat{\beta}_1 = \hat{\theta}_1 + 3\hat{\beta}_2$ or $\hat{\beta}_1 = \theta_1 + 3\beta_2$ | sub in the main regression and get

$$y = \beta_0 + (\theta_1 + 3\beta_2)x_1 + \beta_2 x_2 + \beta_3 x_3 + u$$
$$= \beta_0 + \theta_1 x_1 + 3\beta_2 x_2 + \beta_2 x_2 + \beta_3 x_3 + u$$
$$= \beta_0 + \theta_1 x_1 + \beta_2 (x_2 + 3x_1) + \beta_3 x_3 + u$$

* Now, the explanatory variables are going to be x_1, $x_2 + 3x_1$, and x_3

- we can calculate the T statistic equal to the estimate of β_1

$$t = \frac{\hat{\theta}_1 - 1}{se(\hat{\theta}_1)}$$ ↓ ⇒ min / max.

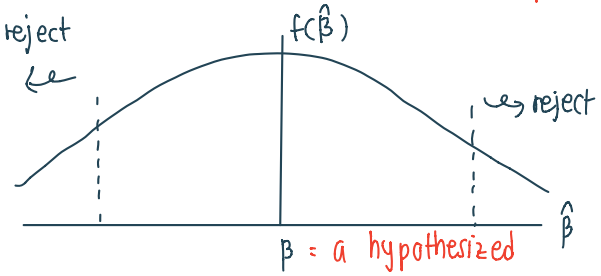
Inference → Hypothesis testing about "β" the true parameter.

$wage = \beta_0 + \beta_1 educ + \beta_2 experience + \ldots + u$

↓

we want to test hypothesis about the true impact (β) of each x variables (educ, experience) on the dependent variable (Y)

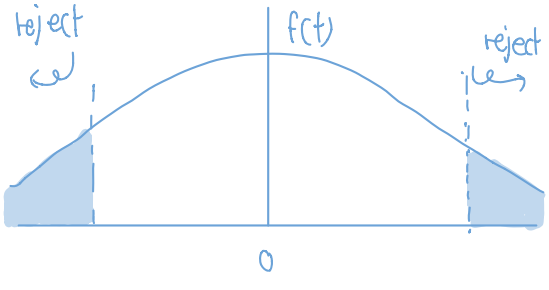
But we don't know what the true β are. So, we use $\hat{\beta}$ (estimator) and $se(\hat{\beta})$ to test the hypothesis.



1.) Test if β = some number

eg. $\beta_j = 0 \Rightarrow X_j$ has no impact on Y.

$\beta_j = 1 \Rightarrow$ 1 unit ↑ in X_j correspond to 1 unit ↑ in Y.



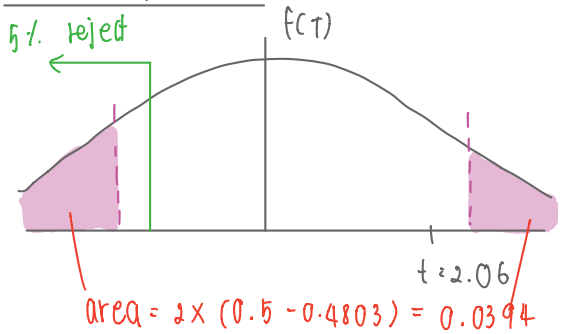
2.) ⇒ t-test

$t = \frac{\hat{\beta}_j - \beta_j}{se(\hat{\beta}_j)} \sim t\ d.f.$

the rejection region area = significant level area.

Significant level = total area in the rejection region

What is p-value?



- suppose, we calculate a t-statistic = $\frac{\hat{\beta}_j - \beta_j}{se(\hat{\beta}_j)} = 2.06$
- suppose, we are testing

$H_0: \beta_j = 0$ } 2-tailed test

$H_a: \beta_j \neq 0$

- p-value = total shaded area
  ↳ significant level which we will reject the H_0 or probability that we reject H_0

∴ If p-value < significance level ⇒ reject H_0.

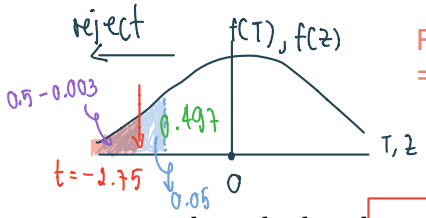
if we use 5% significant level

0.0394 < 0.05 or

0.0197 < 0.025

1 tail test.

Example 1: $H_0 : \beta_j \geq 0, \quad H_a : \beta_j < 0,$ d.f.= 140 → z-table



P-value = what should be the significant level, given that critical value of -2.75 => find shaded area

suppose the calculated $t_{\hat{\beta}_j} = -2.75$ → $t_{\hat{\beta}_j} = \frac{(\hat{\beta}_j - \beta_j)}{se\hat{\beta}_j}$

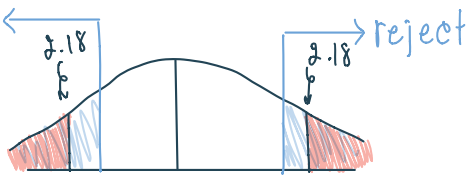
- From the z-table, the value -2.75 corresponds to area = 0.4970.
- Thus, p-value = 0.003.
- Would we reject H_0 if we use the significance level = 5%? 0.05 yes

Rule! we reject H_0 if p-value < sig. level.

2 tails test

Example 2: $H_0 : \beta_j = a_j, \quad H_a : \beta_j \neq a_j,$ d.f.= 18 → t-table

reject



There

suppose the calculated $t_{\hat{\beta}_j} = -2.18$

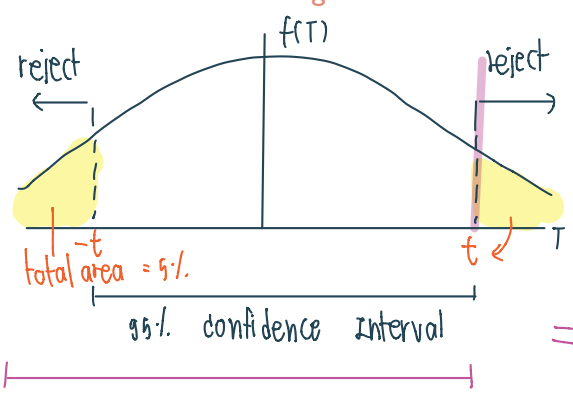
- From the t-table, the value -2.18 corresponds to area = 0.02 to 0.05.
- Thus, p-value = is between 0.02-0.05
- Would we reject H_0 if we use the significance level = 5%?

yes, reject H_0 because the area is less than 0.05. or p-value is less than 0.05

6 Confidence Intervals (CI)

- Confidence Intervales for the POPULATION PARAMETER (β_j)
- A 95% CI of β_j is given by

The range of values that would capture the true Bj at 95% chance

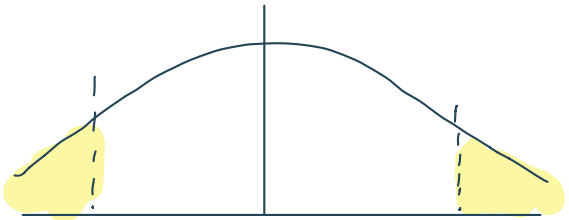


⇒ CI = $\hat{\beta}_j \pm c \times se(\hat{\beta}_j)$, c is the 97.5 percentile in the T-distribution with n-k-1 df.

ที่ระดับ 2.5% or 0.025
ก็ 97.5

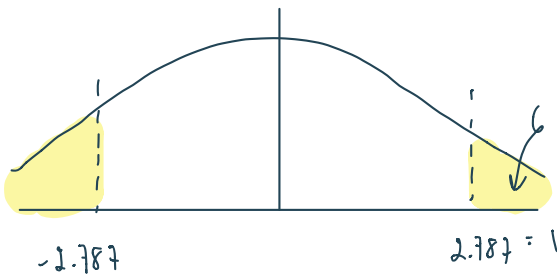
When we read the table

Example 1: **95% CI** df : 25



The 95% CI for $\hat{\beta}_j = (\hat{\beta}_j - 2.06 \cdot se(\hat{\beta}_j),\ \hat{\beta}_j + 2.06 \cdot se(\hat{\beta}_j))$

Example 2: **99% CI** df. 25



The 99% CI for $\hat{\beta}_j = (\hat{\beta}_j - 2.787 \cdot se(\hat{\beta}_j),\ \hat{\beta}_j + 2.787 \cdot se(\hat{\beta}_j))$

F test motivation

We want to test the significance of a group of hypothesis (multiple hypothesis)

$Grade_{325} = \beta_0 + \beta_1$ # times_front $+ \beta_2$ #times_back $+ \beta_3$ hr_study $+ \beta_4$ past_GPA $+ \beta_5$ gender $+ u$

H_0 : seat position doesn't have impact on GPA ; $\beta_1 = 0$ and $\beta_2 = 0 \Rightarrow \beta_1 = \beta_2 = 0$

H_a : seat position matters ; $\beta_1 \neq 0$ and $\beta_2 \neq 0$ or $\beta_1 \neq 0$ and $\beta_2 = 0$ or $\beta_1 = 0$ and $\beta_2 \neq 0$ } at least one of the $\beta_1, \beta_2 \neq 0$

7 Testing Multiple Linear Restrictions: The F-test

Suppose the model is specified by

$$\begin{aligned} y &= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + u \\ H_0 &: \beta_1 = 0 \text{ and } \beta_2 = 0 \\ H_a &: H_0 \text{ is not true} \end{aligned}$$

→ want to test if x_1 and x_2 both have no impact on y.

We can use the F-test to test this type of "multiple hypotheses".

big model

1. Our full model is called the "unrestricted" model (ur). Suppose it can be expressed as:

   $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + u$ is true ⇒ reject null

   $$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \ldots + \beta_k x_k + u$$

2. The model which takes out x (which we think its associated $\beta = 0$) is called the restricted model (r).

   small model $y = \beta_0 + \beta_1 x_1 + u$ is true ⇒ do not reject null.

Suppose there are "q" number of β that we would like to perform a joint-test of $=0$

eg. in this model $q = 2$

$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_{k-q} + u$

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

H_a : H_0 is not true.

$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 \ldots + \beta_{k-q} X_{k-q} + \beta_{k-q+1} + \beta_{k-q+2} X_{k-q+2} \ldots + \beta_k X_k + u$

(r)

(ur)

$$F = \frac{(SSR_r - SSR_{ur})/q}{SSR_{ur}/(n-k-1)}$$

this is always positive (+) because $SSR_{ur} < SSR_r$.

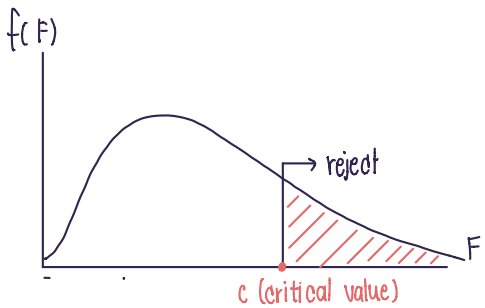
every time you add more X, the model will be better explained.

d.f. of the ur model

- so, if everytime you add 1 more X variable, the SSR ↓ and R^2 ↑, why don't we just keep the additional X in the model ⇒ because every time we add 1 more X, $var(\beta_3)$ will increase, making the prediction of β less precise. So, we only keep the addition X_3 if it / they can improve the model enough.

  can ↓ SSR (↑ R^2) enough.

  can significantly ↓ SSR and ↑ R^2



if t statistic is close to zero = fail to reject H_0

away to zero = all the β are likely to be jointly non zero

$H_0 : \beta_2 = \beta_3 = \ldots = 0$

H_a : H_0 not true

joint Hypothesis being tested

- $F \sim F_{q, n-k-1}$ (d.f. of the ur model)
- we reject H_0 of jointly no effect if $F > c$

3. Some useful facts

① $R^2_{ur} > R^2_r$ because any additional x would increase R^2 (improve fit). $\Rightarrow SSR_{ur} < SSR_r$

② By including more x, the model is certainly better explained.
However, we would like to reject H_0 if the inclusion of extra variables does not improve the model enough.

4. Other ways to calculate the F-statistics:

$\Rightarrow$ From $R^2 = 1 - \frac{SSR}{SST}$ (RSS/TSS); we have $F = \dfrac{\frac{(R^2_{ur} - R^2_r)}{q}}{\frac{(1 - R^2_{ur})}{n-k-1}}$

q ← # of β that are set to "0"

$n-k-1$ ← intercept

n = # of obs. k = # of slope β

if we want to test the overall significance of the model
$H_0: \beta_1 = \beta_2 = \beta_3 = \dots = \beta_k = 0$, H_a = otherwise

$F = \dfrac{R^2/k}{(1-R^2)/(n-k-1)}$ R^2 of the model = UR, the r model has no x at all

Example: Suppose we are interested in understanding the determinant of a baseball player's salary.

salary	= season salary
years	= years in major leagues
gamesyr	= games per year in the league
bavg	= career batting average
hrunsyr	= homeruns per year
rbisyr	= runs batted in per year

If we want to test whether "performance" has any impact on salary

- the unrestricted model (ur) is defined by

$H_0: \beta_{bavg} = \beta_{hrunsyr} = \beta_{rbisyr} = 0$
H_a: otherwise is true

• F > C , reject H_0 • ถ้าใช้ SSR ในการหา F ต้อง ur
• ถ้าใช้ R^2 ใช้ R^2 ของ ur (normal r-squared)

y X In MRL, we focus more on adjusted R^2

(ur)

```
. regress log_salary years gamesyr bavg hrunsyr rbisyr
```

Source	SS	df	MS
Model	308.989208	5	61.7978416
Residual	183.186327	347	.527914487
Total	492.175535	352	1.39822595

SSE SSR SST

Number of obs	= 353
F(5, 347)	= 117.06
Prob > F	= 0.0000
R-squared	= 0.6278
Adj R-squared	= 0.6224
Root MSE	= .72658

log_salary	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
years 1	.0688626	.0121145	5.68	0.000	.0450355	.0926898
gamesyr 2	.0125521	.0026468	4.74	0.000	.0073464	.0177578
bavg 3	.0009786	.0011035	0.89	0.376	-.0011918	.003149
hrunsyr 4	.0144295	.016057	0.90	0.369	-.0171518	.0460107
rbisyr 5	.0107657	.007175	1.50	0.134	-.0033462	.0248776
_cons	11.19242	.2888229	38.75	0.000	10.62435	11.76048

↑ intercept

the more variable we put in the higher R^2 we get

- the restricted model (r) is defined by

y X

(r)

```
. regress log_salary years gamesyr
```

Source	SS	df	MS
Model	293.864058	2	146.932029
Residual	198.311477	350	.566604221
Total	492.175535	352	1.39822595

Number of obs	= 353
F(2, 350)	= 259.32
Prob > F	= 0.0000
R-squared	= 0.5971
Adj R-squared	= 0.5948
Root MSE	= .75273

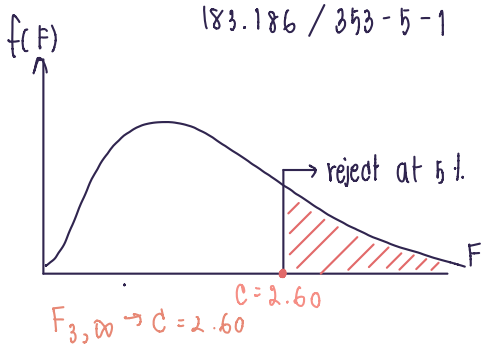
log_salary	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
years	.071318	.012505	5.70	0.000	.0467236	.0959124
gamesyr	.0201745	.0013429	15.02	0.000	.0175334	.0228156
_cons	11.2238	.108312	103.62	0.000	11.01078	11.43683

Now, our H_0 and H_a becomes

$$F = \frac{(SSR_r - SSR_{ur})/q}{SSR_{ur}/(n-k-1)}$$

q = # joint hypothesis we are testing (in this case 3 → (gamesyr , bavg, hrunsyr))

$$\frac{198.311 - 183.186 / 3}{183.186 / 353 - 5 - 1} \simeq 9.55$$



$N_1 = q$ / the numerator / joint hypothesis tested.
$N_2 = N-k-1$
in this case
$q = 3$
$N_2 = 353 - 5 - 1 = 347 > 200 = \infty$
at $N_2 = \infty$, $N_1 = 3$, 5% significant, the critical value is 2.60

So, if calculated F statistic is more than 2.60 then we reject null. In this case F statistic is 9.55 which is greater than 2.60 (critical value). So we reject H_0.

• ans Aj. version ⇒ Since $F = 9.55 > 2.6$, we reject H_0 at 5% level and conclude that performances have joint effect on salary.

HW. : use R^2 from ur

$$F = \frac{(R^2/q)}{(1-R^2)/n-k-1}$$

$$\frac{0.6278/3}{(1-0.6278)/353-5-1} = 195.098$$

$\downarrow$

347

- when considering each of the performance x one-by-one, none of them has a significant impact at 5%.
- but when performing F test, performances have joint impact