4 Testing Hypotheses about a Single Linear Combination of the Parameter

Consider

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

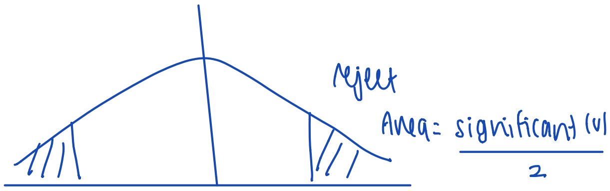
where $jc =$ number of years attending a two-year college
$univ =$ number of years at a four-year college
$\exp er =$ months in the workforce.
We want to test whether $\beta_1 = \beta_2$. → if the return from 1 more year of education at a jc is the same as that of the university

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

against

$H_a : \beta_1 \neq \beta_2 \Rightarrow H_0 : \beta_1 - \beta_2 \neq 0$

2-tailed test



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

we compute this t-statistics and compare it with the critical value.

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

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

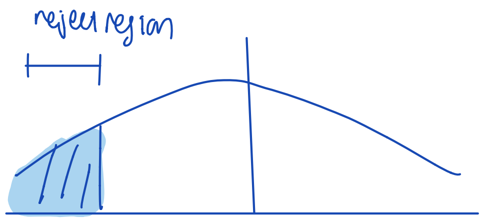
→ not very straight forward to calculate
→ we use a variable transformation trick → see notes

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 assumed that β_1 would not be more than β_2, so return to a JC would never be more than return to uni educ.

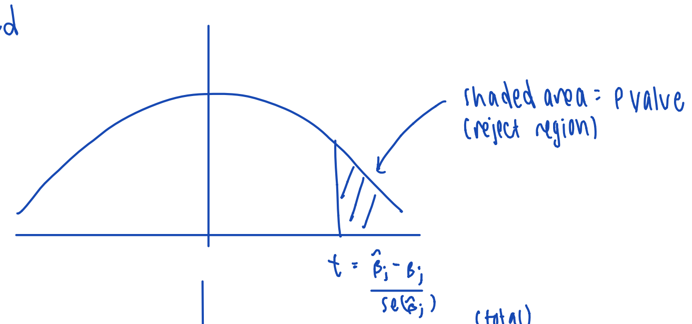


$$t = \frac{(\hat{\beta}_1 - \hat{\beta}_2) - 0}{se(\hat{\beta}_1 - \hat{\beta}_2)}$$

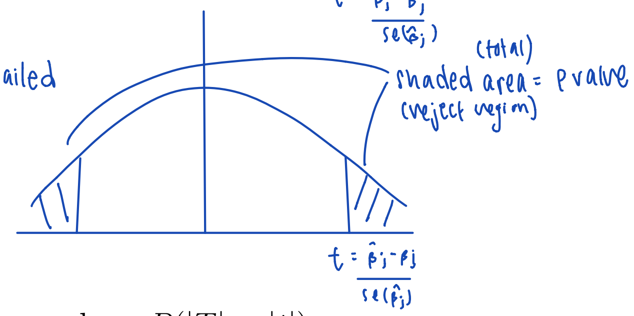
5 Computing p-Values for t-Tests

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

1-tailed



2-tailed



- 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 be greater than our t-statistic in the H_0 test

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 the $H_0: \beta_1 - 3\beta_2 = 1$

H_a: otherwise is true

1^{st}: write the t-satistic for testing H_0

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

2^{nd}: define $\hat{\theta}_1 = \hat{\beta}_1 - 3\hat{\beta}_2$ = $H_0: \theta_1 = 1$

$H_a: \theta_1 \neq 1$

$t = \frac{\hat{\theta}_1 - 1}{se(\hat{\theta}_1)}$ → we need our regression to have θ_1 in it so STATA or OLS estimation will automatically give $\hat{\theta}_1$ a se θ_1

Now, $\hat{\beta}_1 = \hat{\theta}_1 + 3\hat{\beta}_2$ or $\beta_1 = \theta_1 + 3\beta_2$

sub in the main regression and get

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

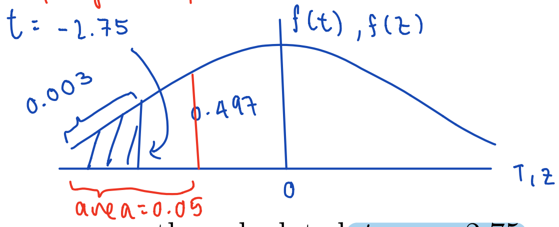
$= \beta_0 + \theta_1 x_1 + 3\beta_2 x_1 + \beta_2 x_2 + \beta_3 x_3 + u$

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

* Now, the explanatory variable are going to be x_1, $x_2 + 3x$ and x_3
- we can calculate $t = \frac{\hat{\theta}_1 - 1}{se(\hat{\theta}_1)}$

1-tailed test

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



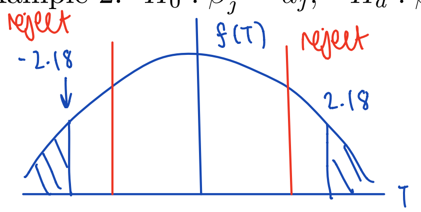
p-value = what should be the significant level given the critical value of -2.75
⇒ find the shaded area

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

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

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

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



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

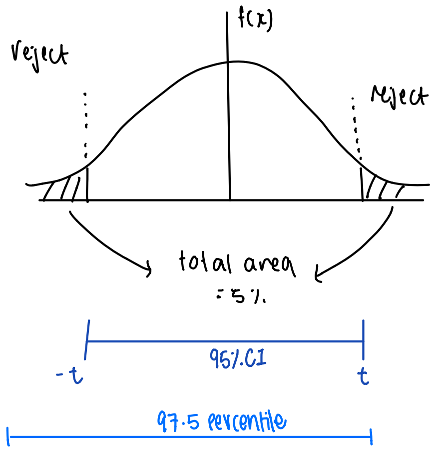
- From the t-table, the value -2.18 corresponds to area = 0.02 to 0.05.
- Thus, p-value = is between 0.02 – 0.05.
- Would we reject H_0 if we use the significance level = 5%?

  Yes, reject H_0 because the area is less than 0.05 or p-value < 0.05

6 Confidence Intervals (CI)

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

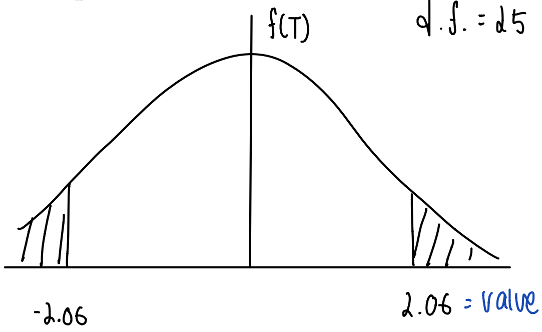
The range of values that would capture the true β_j at a 95% chance



CI ⇒ $\hat{\beta}_j \pm c \times s.e(\hat{\beta}_j)$

c is the $\boxed{97.5}$ percentile in the t-distribution with n-k-1 df

Example 1: **95% CI**

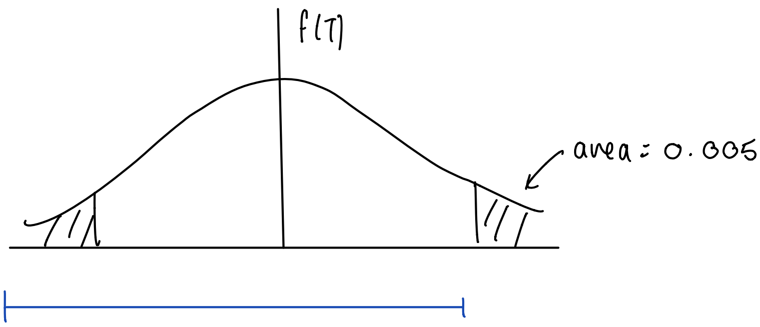


= value of 97.5^{th} percentile in the t_{25} distribution

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**



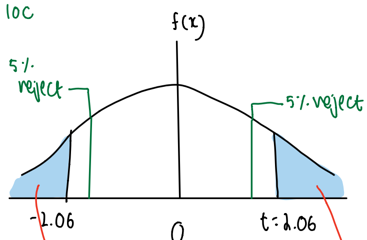
99.5^{th} percentile

or area $= 1 - 0.005 = 0.995$

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

significant level: total area in the rejection region

ass df loc



area = 2x(0.5-0.4803)

= 0.0394 = pvalue

- suppose we calculate a t-statistic $= \frac{\hat{\beta} - \beta}{se_{\hat{\beta}}} = 2.06$
- Suppose, we are testing $H_0: \beta_j = 0$; $H_a: \beta_j \neq 0$

two tail testing

pvalue = total shaded area

P-value = significant level which we will reject the H_0 or prob that we will reject H_0

* If pvalue < significant level ⇒ reject

confidence that we will capture

95% confident interval → true value of the parameter?

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 $\beta_2 = 0$ ⇒ $\beta_1 = \beta_2 = 0$

H_a: seat position matters

$\beta_1 \neq 0$ and $\beta_2 \neq 0$ or $\beta_1 = 0$ and $\beta_2 \neq 0$ or $\beta_1 \neq 0$ and $\beta_2 = 0$ } at least one of the $\beta_1, \beta_2 \neq 0$

cannot use the normal test

thus the F-test

7 Testing Multiple Linear Restrictions: The F-test

Suppose the model is specified by

$$
\begin{aligned}
y &= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + u \\
H_0 &: \beta_2 = 0 \text{ and } \beta_3 = 0 \\
H_a,\ H_1 &: H_0 \text{ is not true}
\end{aligned}
$$

→ 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".

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

   Big model

   $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + u$ is true or reject H_0

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

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 H_0

○ suppose there are "q" number of β that we would like to perform a joint test of "= 0".

e.g. in this model $q = 2$

$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_{k-q} x_{k-q} + u$

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

(the last q βs $= 0$)

H_a: H_0 is not true

$$y = \underbrace{\underbrace{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_{k-q} x_{k-q}}_{\text{restricted}} + \beta_{k-q+1} x_{k-q+1} + ... + \beta_k x_k + u}_{\text{unrestricted}}$$

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

— always positive; $SSR_{ur} < SSR_r$

everytime you add 1 more x, the model will be better explained | variances will also increase ↳ won't be predicted accurately

← d.f. of the "ur" model

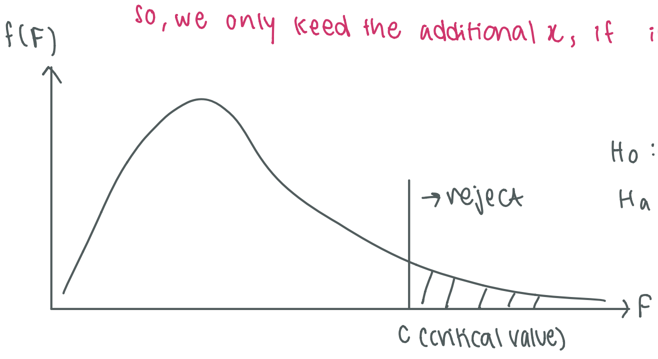
so, if every time you add 1 more x variable, the SSR↓ & R^2↑, why don't we just keep the additional x in the model?

⇒ Because every time we add 1 more x, $Var(\hat{\beta}_j)$ will increase, making the prediction of β less precise.

so, we only keed the additional x, if it/they can improve the model enough

can ↓SSR (↑R^2) enough

can significantly ↓SSR & R^2↑



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

H_a: H_0 is not true

$F \sim F_{q, n-k-1}$ (df)

#no of joint Hypothesis being tested

We reject H_0 of jointly no effect if $F > C$

3. Some useful facts

① $R^2_{ur} > R^2_r$ → any addition x would increase R^2 (improve the fit)
→ $SSR_{UR} < SSR_r$

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

4. Other ways to calculate the F-statistics:

⇒ From $R^2 = 1 - \frac{SSR}{SST}$

We have $F = \frac{(R^2_{ur} - R^2_r)}{q} \Big/ \frac{1 - R^2_{ur}}{n-k-1}$

If we want to test the overall significance of model

$H_0: \beta_1 = \beta_2 = \beta_3 = \ldots \beta_k = 0$, H_a: otherwise

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

$H_0: \beta_{bavg} = \beta_{hrunsyr} = \beta_{rbisyr}$

H_a: otherwise is true

- the unrestricted model (ur) is defined by ur model

y x

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

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

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	.0688626	.0121145	5.68	0.000	.0450355	.0926898
gamesyr	.0125521	.0026468	4.74	0.000	.0073464	.0177578
bavg	.0009786	.0011035	0.89	0.376	-.0011918	.003149
hrunsyr	.0144295	.016057	0.90	0.369	-.0171518	.0460107
rbisyr	.0107657	.007175	1.50	0.134	-.0033462	.0248776
_cons	11.19242	.2888229	38.75	0.000	10.62435	11.76048

when considering each of the performance x one-by-one, none of them has a significant impact at 5%.

- the restricted model (r) is defined by

y x

```
. 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

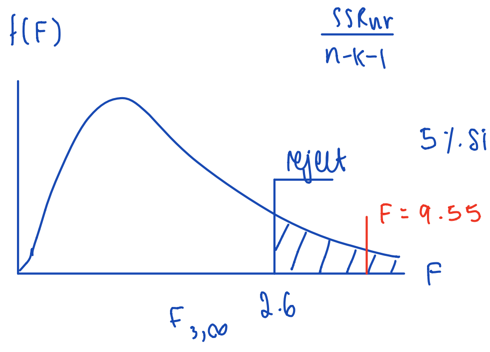
- But when performing an F test, performance has joint impact

Now, our H_0 and H_a becomes

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

$$F = \frac{\frac{SSR_r - SSR_{ur}}{q}}{\frac{SSR_{ur}}{n-k-1}} = \frac{(198.311 - 183.186)/3}{183.186/(353-5-1)} = 9.55017$$

$$= 9.550$$



5% significant level

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