taking log does that are thinking that the fn is non-linear

4 Testing Hypotheses about a Single Linear Combination of the Parameter

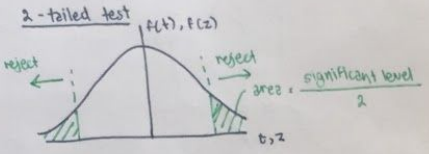
Consider

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

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 Junior college is the same as that of the university

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

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



$$t = \frac{(\hat{\beta}_1 - \hat{\beta}_2) - 0}{s.e\,(\hat{\beta}_1 - \hat{\beta}_2)}$$ ← we compute this t-statistic and compare with the critical value

where $s.e.(\hat{\beta}_1 - \hat{\beta}_2) = \sqrt{var(\hat{\beta}_1 - \hat{\beta}_2)}$

"not very straight forward to calculate"

$$= \sqrt{var(\hat{\beta}_1) + var(\hat{\beta}_2) - 2\,cov(\hat{\beta}_1, \beta_2)}$$

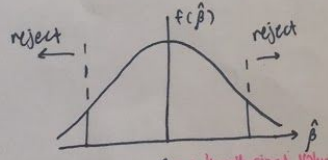
↓ we use a variable transformation trick → see notes

March 12 **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)

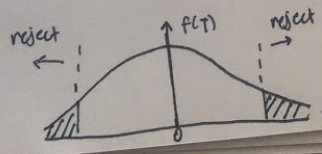
BUT. we don't know what the true β are. So we use $\hat{\beta}$ (estimator) and $se(\hat{\beta})$ to test the hypothesis



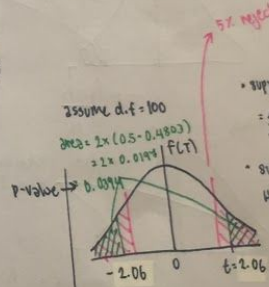
β = a hypothesized value
ex. $\beta = 0$ or $\beta = 1$ etc.

① test if β = same 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

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



significant level = total area in the rejection region



- suppose we calculate a t-statistic $= \frac{\hat{\beta}_j - \beta_j}{se\,\hat{\beta}_j} = 2.06$
- suppose, we are testing $H_0: \beta_j = 0$ $H_a: \beta_j \neq 0$
  ↳ 2-tailed test
- p-value $= \frac{(\hat{\beta}_j - \beta_j)}{se(\hat{\beta}_j)}$ • p-value = total shaded area

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

If the p-value < significance level → reject H_0

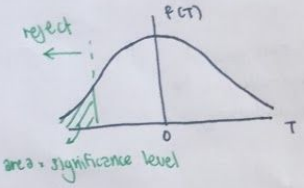
rule of thumb

another possible hypothesis test (one-tailed alternative)

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

$H_2: \beta_1 < \beta_2 \rightarrow H_2: \beta_1 - \beta_2 < 0$

· H is assumed that β_1 would not be more than β_2 (returns to a 2-year college would never be more than returns to university education)



$$t = \frac{(\hat{\beta}_1 - \beta_2) - 0}{s.e(\hat{\beta}_1 - \hat{\beta}_2)}$$

" then go to extra note"

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

1st) write the t-statistic for testing H_0

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

2nd) Define $\theta_1 = \hat{\beta}_1 - 3\hat{\beta}_2 \rightarrow H_0: \theta_1 = 1, H_0: \theta_1 \neq 1$

$t = \frac{\hat{\theta}_1 - 1}{se(\theta_1)}$ → we need our regression to have θ_1 in it. So, STATA or OLS estimation will automatically give $\hat{\theta}_1$ & 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

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

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

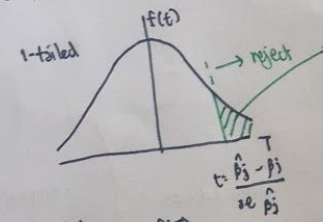
$= \beta_0 + \theta_1 x_1 + \beta_2(x_2 - 3x_1) + \beta_3 x_3 + u$

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

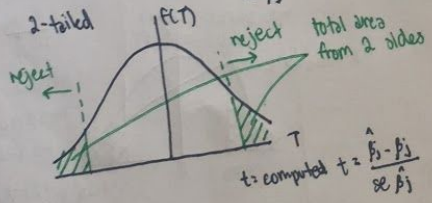
· we can calculate $t = \frac{\hat{\theta}_1 - 1}{se\,\theta}$

5 Computing p-Values for t-Tests

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



this shaded area in the rejection region is the p-value



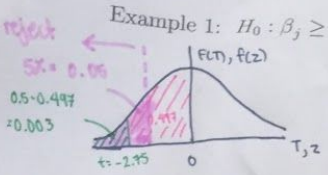
- p-value : $P(|T| > |t|)$

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 (in the || term) than our t in the H_0 test

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



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

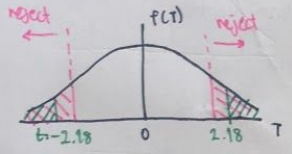
suppose the calculated $t_{\hat{\beta}_j} = -2.75$

$t_{\hat{\beta}_j} = \frac{(\hat{\beta}_j - \beta_j)}{se(\hat{\beta}_j)}$

- 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%? Yes

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

Example 2: $H_0 : \beta_j = a_j, \quad H_a : \beta_j \neq a_j, \quad$ d.f.$= 18$. ← use 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%?

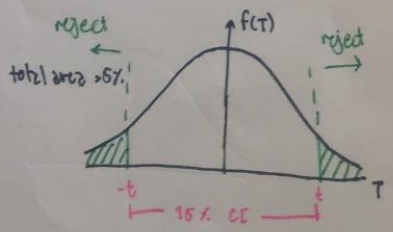
Yes, reject H_0 because the area is less than 0.05 or p-value < 0.05

6 Confidence Intervals (CI)

- Confidence Intervales for the POPULATION PARAMETER (β_j)

→ the range of values that would capture the true β_j at a 15% chance

- A 95% CI of β_j is given by

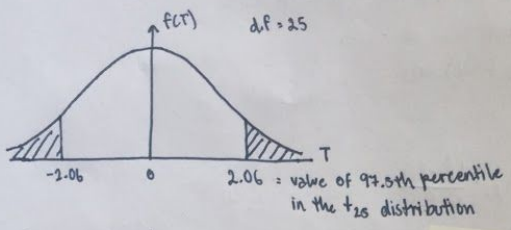


CI = $\hat{\beta}_j \pm c \times se(\hat{\beta}_j)$

↑ → 97.5 (in this case)

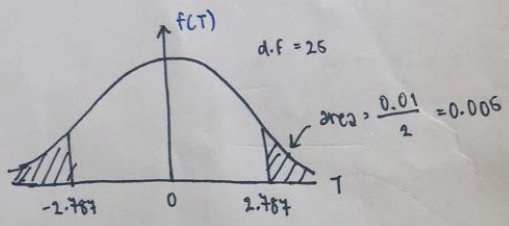
c is the □ percentile in the t-distribution with n-k-1 d.f.

Example 1: **95% CI**



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**



The value of 99.5 percentile or area = 1 - 0.005 = 0.995

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

7 Testing Multiple Linear Restrictions: The F-test

Suppose the model is specified by

$$\begin{aligned} y &= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + u \\ H_0 &: \beta_1 = 0 \text{ and } \beta_2 = 0 \quad \rightarrow \text{want to test if } x_1 \text{ and } x_2 \text{ BOTH have no impact on } y \\ H_2, H_1 &: H_0 \text{ is not true} \end{aligned}$$

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

1. Our full model is called the "unrestricted" (big model) 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 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 → don't reject H_0

- suppose there are "q" number of β that we would like to perform a joint -test of =0
- eg. in this model $q = 2$

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

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

(the last q $\beta_s = 0$)

H_2: H_0 is not true

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 \dots + \beta_{k-q} x_{k-q} + \beta_{k-q+1} x_{k-q+1} + \beta_{k-q+2} x_{k-q+2} + \beta_k x_k + u$$

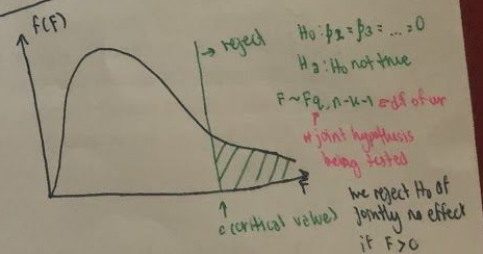
r

ur

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

this is always (+) because $SSR_{ur} < SSR_r$ everytime you add 1 more x_i the model will be better explained

↑ d.f in ur model



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

H_2: H_0 not true

$F \sim F_{q, n-k-1}$ ← df of ur

↑ joint hypothesis being tested

we reject H_0 of jointly no effect if $F > c$

- 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 everytime we add 1 more x, the $var(\hat{\beta}_s)$ will increase making the prediction of β_s less precise. So, we only keep the additional x_s if it/they can improve the model enough → can↓SSR (↑R^2) enough or can significantly↓SSR and ↑R^2

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:

→ From $R^2 = 1 - \frac{SSR}{SST}$

we have $F = \frac{(R^2_{ur} - R^2_r)/q}{(1-R^2_{ur})/n-k-1}$ ← q: # of β that are set to "0"; k+1: intercept + # of slope β; n: no. obs

→ 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

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

y – *salary* = season salary

r – *years* = years in major leagues

r – *gamesyr* = games per year in the league

ur – *bavg* = career batting average

ur – *hrunsyr* = homeruns per year

ur – *rbisyr* = runs batted in per year

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

$H_0: \beta_{bavg} = \beta_{hrunsyr} = \beta_{rbisyr} = 0$

H_a: otherwise is true

- the unrestricted model (ur) is defined by

ur model

```
. 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

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

k = 5

when considering each of the performance variable, none of them has a significant impact at 5%

- the restricted model (r) is defined by

```
. 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

SSE, SSR, SST

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

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

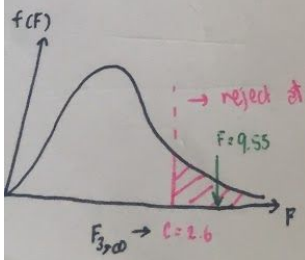
$$\equiv \frac{(198.311 - 183.186)/3}{(183.186)/(353-5-1)} \approx 9.55$$

HW

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

= ??

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



use 5% level of sig.

since F = 9.55 > 2.6, we reject H_0 at 5% level and conclude that performances have joint effects on salary

8 How the Hypothesis Testing is done in Practice

1. Check the values of $t-statistic$ reported by the statistical software (i.e. STATA, SPSS, SAS)

   $\Rightarrow$ These $t-statistics$ are to test $H_0 : \beta_i = 0$

   $\Rightarrow$ If the d.f. > 30, then when $t > 1.96$, we can reject H_0 with 5% sig. level (z-table)

   $\Rightarrow$ **When $t > 1.96$**, we can say that β_i is **statistically significant** at 5% level. (value of $\beta_i \neq 0$)

   $\Rightarrow$ **When $t < 1.96$** we can say that β_i is **not statistically significant** at 5% level.

   $\Rightarrow$ If $t < 1.96$ we can drop x_i from the model

   $\Rightarrow$ After we drop x_i, we estimate the new regression function and obtain a new set of $\hat{\beta}$.

2. We can also perform other hypothesis testings of interest.

   e.g. $H_0 : \beta_i = \beta_j$

   or $H_0 : \beta_i = 5$ etc.

   or perform an F-test for testing multiple linear restrictions

3. Usually, in economics, the estimation results are reported using this form

Dependent Variable: log(*salary*)			
Independent Variables	**(1)**	**(2)**	**(3)**
log(*sales*) (sales →)	.224 (.027)	.158 (.040)	.188 (.040)
log(*mktval*) (other company performance)	—	.112 (.050)	.100 (.049)
profmarg (other company performance)	—	−.0023 (.0022)	−.0022 (.0021)
ceoten (CEO characteristics)	—	—	.0171 (.0055)
comten (CEO characteristics)	—	—	−.0092 (.0033)
intercept	4.94 (0.20)	4.62 0.25)	4.57 (0.25)
Observations *R*-squared	177 .281	177 .304	177 .353

↑ like a simple regression with 1 x