4 Testing Hypotheses about a Single Linear Combination of the Parameter

take log to make it ┌ linear
└ better R^2

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
experience → $\exp er$ = months in the workforce.
We want to test whether $\beta_1 = \beta_2$.

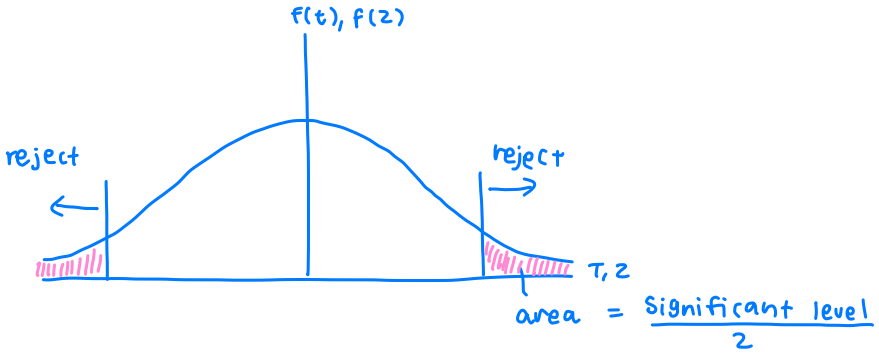
we want to see whether the reward (wage) from JC equal to uni

→ if the returns from 1 more year of edu at JC is the same as 1 more year of the uni.

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

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

2 tailed test



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

→ we compute this t statistic and compare with the critical value!

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

when $\hat{\beta}_1$ and $\hat{\beta}_2$ are correlated → we have covariance

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

Not very straight forward to calculate → we use a variable transformation trick see note!

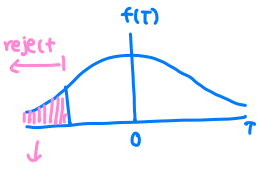
another possible hypothesis test (one-tailed alternative)

$H_0 : \beta_1 = \beta_2 \Rightarrow H_0 : \beta_1 - \beta_0 = 0$

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

1 more y. at JC gives lesser w

- it is assumed that β_1 would not be more than β_2 (return to JC would less than returns to uni)



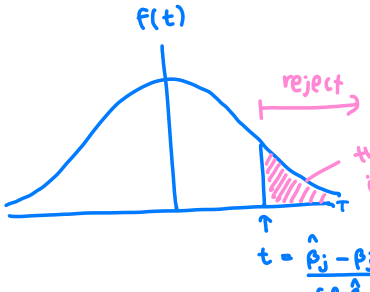
$$t = \frac{(\hat{\beta}_1 - \hat{\beta}_2) - 0}{s.e.(\hat{\beta}_1 - \hat{\beta}_2)}$$

* Then, go to the extra note to find $s.e.(\hat{\beta}_1 - \hat{\beta}_2)$

5 Computing p-Values for t-Tests

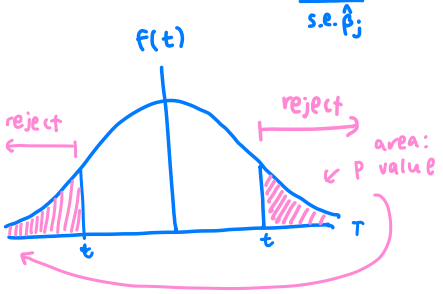
- What is the significance level given the computed t-statistics?

1-tail test



$$t = \frac{\hat{\beta}_j - \beta_j}{s.e.\hat{\beta}_j}$$

2-tail test



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

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

t = computed t-statistic

→ p-value = prob that a random T value will be greater (in 11 term) than our t in the Hypothesis test (H_0)

In class exercise

In multiple regression model, assume MLR 1-6 are statisfied.

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + u$$

You would like to test the $H_0: \beta_1 - 3\beta_2 = 1$
H_a : otherwise is true

1) write the t statistic for testing H_0

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

* 2) Define $\theta_1 = \hat{\beta}_1 - 3\hat{\beta}_2 \Rightarrow$ $H_0 = \theta_1 = 1$

$H_a = \theta_1 \neq 1$

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

now, $\hat{\beta}_1 = \hat{\theta}_1 + 3\hat{\beta}_2$

OR $\beta_1 = \theta_1 + 3\beta_2$

sub in main regression and get

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

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

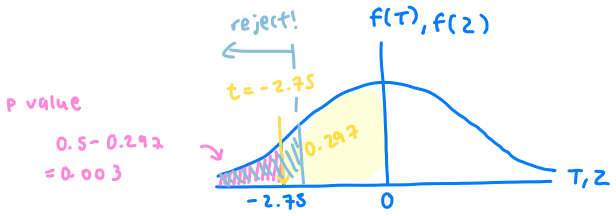
$$Y = \beta_0 + \theta_1 x_1 + \beta_2 (x_2 + 3x_1) + \beta_3 x_3 + u$$

Now, the explanatory variables are going to be

x_1, $x_2 + 3x_1$ and x_3

→ we can calculate $t = \frac{\hat{\theta}_1 - 1}{s.e.\hat{\theta}_1}$

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 $t_{\hat{\beta}_j} = -2.75$

↘ meaning that we did $t_{\hat{\beta}_j} = \frac{(\hat{\beta}_j - \beta_j)}{s.e.(\hat{\beta}_j)}$

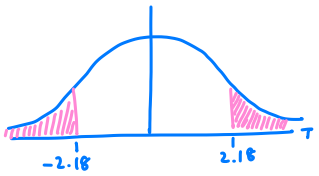
- From the z-table, the value -2.75 corresponds to area = _ _ _ _ _ _ _ _ _ _ _ _ 0.003 _ _ _ _ _ _ _ _ _ _ _ _ _.
- Thus, p-value = _ _ _ _ _ _ _ _ _ _ 0.003 _ _ _ _ _ _ _ _ _ _.
- Would we reject H_0 if we use the significance level = 5%? Yes!

** Rule! we reject Ho if p value < sig. level

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

2 tail test

T constant



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 = _ _ _ _ _ _ _ btw 0.02 - 0.05 _ _ _ _ _ _.
- Would we reject H_0 if we use the significance level = 5%?

Yes! reject Ho because the area is less than 0.05 or p < 0.05

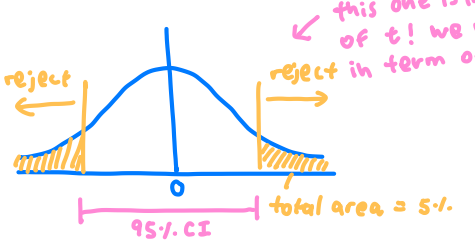
6 Confidence Intervals (CI)

- Confidence Intervals for the POPULATION PARAMETER (β_j)

the range of values that would capture the true βj at a 95% chance.

- A 95% CI of β_j is given by

"we capture the true value that we want 95% of the time."



this one is in term of t! we want reject in term of β

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

c is the 97.5 percentile in the t-distribution with n-k-1 d.f.

100 - 2.5 Significant level in 1 side!

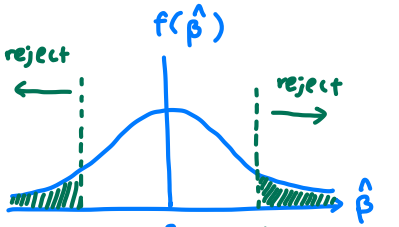
bc. we have to look at the table!

Inference Hypothesis testing about "β" the true parameter.

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

we want to test the true impact (β) of each x variables (educ, exper) on the dependent variable (y)

BUT we dont know what the true β are. so, we use $\hat{\beta}$ (estimator) and s.e.($\hat{\beta}$) to test the hypothesis.



β = a Hypothesized value
ex. β=0, β=1 etc.
(ไม่ใช่ some number)

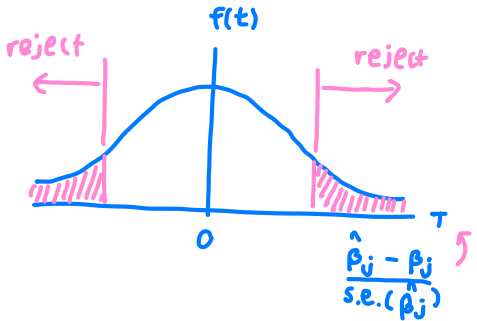
1) test if β = some number
e.g. $\beta_j = 0 \rightarrow x_j$ has no impact on y

$\beta_j = 1 \rightarrow$ 1 unit ↑ in x_j correstpond to 1 unit ↑ in y

⇒ t-test!

$$\frac{\hat{\beta}_j - \beta_j}{s.e.(\hat{\beta}_j)} \sim t_{d.f.}$$

Change this to t standarize



5% significant level

= total area in rejection region.

---

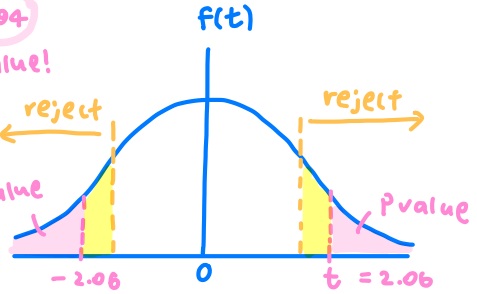
ass. d.f. 100

area = 2 × (0.5 − 0.4803)
= 0.0394
p-value!

in 5% reject region.

we reject Ho!

0.5 > 0.0394



- suppose, we calculate a t-statistic = $\frac{\hat{\beta}_j - \beta_j}{s.e.(\hat{\beta}_j)} = 2.06$
- suppose, we are Testing

$H_0: \beta_j = 0$

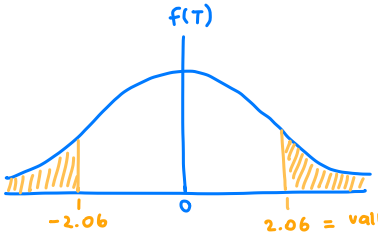
$H_a: \beta_j \neq 0$

2 tail test = total shaded area.

P-value = significant level which we will reject the Ho OR prob that we will reject Ho

if p-value < Significant level ⇒ reject Ho (regradless the # tail :))

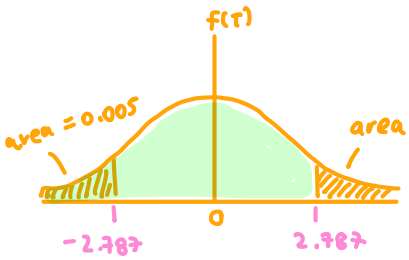
Example 1: **95% CI**



the 95% CI for $\hat{\beta}_j = [\hat{\beta}_j - 2.06 \cdot s.e.(\hat{\beta}_j), \hat{\beta}_j + 2.06 \cdot s.e.(\hat{\beta}_j)]$

↑ it's a range!

Example 2: **99% CI** d.f. = 25



at percentile of 99.5
or area of $1 - 0.005 = 0.995$

* t table don't care about the pentile but z does!

the 99% CI for $\hat{\beta}_j = [\hat{\beta}_j - \underset{2.787}{\square} \cdot s.e.(\hat{\beta}_j), \hat{\beta}_j + \underset{2.787}{\square} \cdot s.e.(\hat{\beta}_j)]$

F-test motivation

* * * * *

ถ้า $\beta_1, \beta_0 = 0$ are we still add them in our UR?

→ we want to test the significance of a group of Hypothesis (multiple Hypotheses)

$$Grade_{325} = \beta_0 + \beta_1 \#times\_front + \beta_2 \#times\_back + \beta_3 hr\_study + \beta_4 past\_GPA + \beta_5 gender + u$$

entire equation = unrestricted one w/o seating = restricted.

H_0: seat position doesnot have impact on GPA

$$\beta_1 = 0 \quad \text{and} \quad \beta_2 = 0 \Rightarrow \beta_1 = \beta_2 = 0$$

H_a: seat position does matters

$$\begin{cases} \beta_1 \neq 0 \text{ and } \beta_2 \neq 0 \\ \text{OR } \beta_1 \neq 0 \text{ and } \beta_2 = 0 \\ \text{OR } \beta_1 = 0 \text{ and } \beta_2 \neq 0 \end{cases} \} \text{ at least 1 of the } \beta_1, \beta_2 \neq 0.$$

we are testing whether we want to include β_1 and β_2 in our model

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

always has lower u bc. it has more x! → better explained

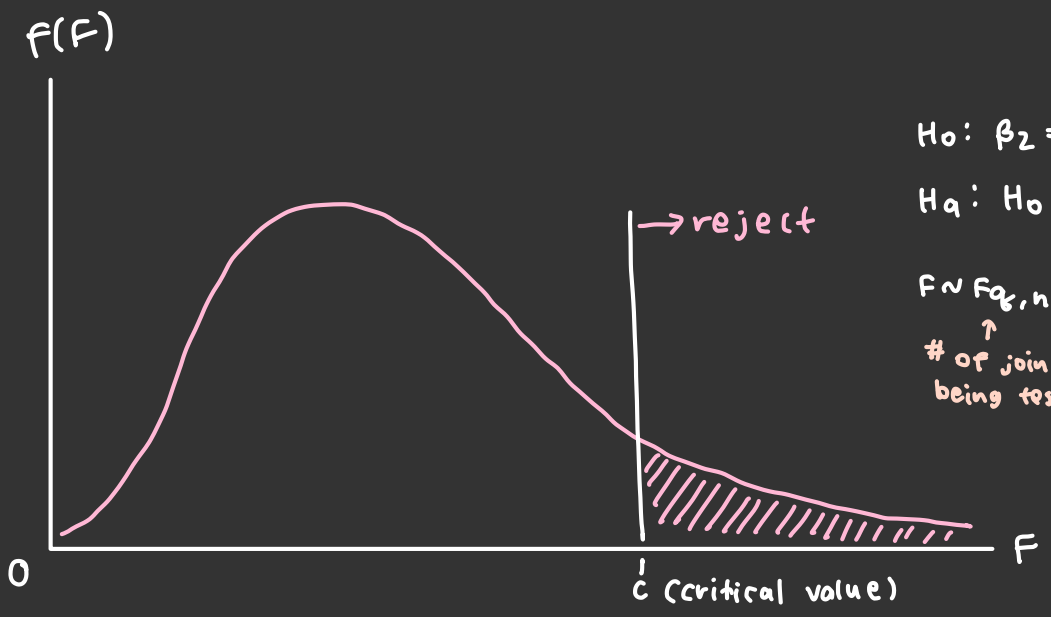
so $SSR_{ur} < SSR_r$ it always ⊕

$(n-k-1)$ — d.f. of the "ur" model

ur - more x
→ better explained
→ R^2 ↓
→ SSR ↓
→ But variance ↑
hard to predict percisly

every time we add 1 more x, $var(\hat{\beta}_s)$ (all of $\hat{\beta}$) will increase, making the prediction of β less precise, so we only keep the additional xs if it/ they can improve the model enough

can ↓ R^2, ↓ SSR enough



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

H_a: H_0 is not true

$F \sim F_{q, n-1-k}$ ← d.f. of the ur model

↑ # of join Hypothesis being tested

- always 1 tailed test
- rejection region always away from 0

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

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".

2 models we can use!

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 H_0

to k number

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

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

less than K

$y = \beta_0 + \beta_1 x_1 + u$ – is true ⇒ do not reject H_0

which one to use? :)

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

$y = \beta_0 + \beta_1 x_1 + u$

e.g. in this mode our $q = 2$

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

ตัวที่เหลือ

$H_0 : \beta_{K-q+1} = \beta_{K-q+2} = \ldots = \beta_K = 0$

(the last q $\beta_0 = 0$)

H_a : H_0 is not true.

So, we test whether we want to include $x_{K-q+1}, x_{K-q+2}, \ldots x_K$ variable in our model

$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots \beta_{K-q} x_{K-q} + \beta_{K-q+1} x_{K-q+1} + \beta_{K-q+2} x_{K-q+2} + \ldots + \beta_K x_K + u$

restricted model.

unrestricted model.

F test is testing do we need to include all variables or not!

3. Some useful facts

① $R^2_{ur} > R^2_r$ because additional x will increase R^2 (improve fit)
→ $SSR_{ur} < SSR_r$

② By including more x, the model is certainly explained. However, we wanna reject H_0 if the inclusion of extra variables doesnot improve the model enough

4. Other ways to calculate the F-statistics:

→ from $R^2 = 1 - \frac{SSR}{SST}$ (SSR: RSS, SST: TSS)

we have $F = \dfrac{(R^2_{ur} - R^2_r)/q}{(1 - R^2_{ur})/(n-k-1)}$ (q = # of β that set to 0; k = # slope of β int.; n = # of observation)

if we want to test the overall significant of the model

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

F always given

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

we want to test that x can explain the model!

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 (indicator of performance)	*bavg*	= career batting average
UR (indicator of performance)	*hrunsyr*	= homeruns per year
UR (indicator of performance)	*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

UR has higher R^2

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

this case lower

not excess 1.96 and > 0.05

So, each of them cannot explain salary base on our methode (cannot reject Ho) but we want to test wether they can jointly explain y → f test. → if joinly ✓ we include in the model. (UR??)

each of them dont excess 1.96 none of them has a significant impact at 5%, we cannot reject Ho (that say Ho: $\beta = 0$)

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

rule! : we reject Ho if p value < sig. level

R model

- the restricted model (r) is defined by

```
. regress log_salary years gamesyr
```

Source	SS	df	MS
SSE Model	293.864058	2	146.932029
SSR Residual	198.311477	350	.566604221
SST 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

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

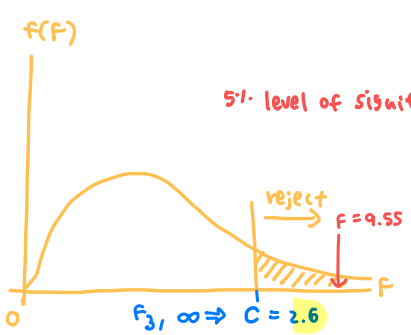
R^2 from UR

HW:

$$F \equiv \frac{R^2/q}{1-R^2/n-k-1}$$

$\equiv$??

$$F \equiv \frac{(193.311477 - 183.186327)/3}{183.186327/353-5-1} \approx 9.55$$



Since $F = 9.55 > 2.6$, we reject Ho at 5% level and conclude that proformances 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% sign. level

   z-table

   5% significant level

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

   $\Rightarrow$ **When $t < 1.96$** we can say that β_i is **not statistically significant** at 5% level. s.t. people dont put it on model

   drop dc. it is not show significant impact on explaining y

   $\Rightarrow$ 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

   get UR or R → compare SSR or R^2

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

we have 3 models, but which one to use?

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

sales → value of β changes

other com. performance

CEO characteristics

(# year work with compa.)

impact on sale (depend on direction of bias)

most restricted

like a simple regression (only 1 x)

unrestricted

Multiple Regression Analysis : Further Issues

1 Data scaling on OLS statistics

When we change the unit of measurment (weight → ex. currency) of a variable, the value of estimators would chance accordingly. For example

$$\widehat{bweght}_g = \widehat{\beta}_0 + \widehat{\beta}_1 cigs + \widehat{\beta}_2 fa\min c,$$

where
$bwght$ = child birth weight, in grams.
$cigs$ = number of cigarettes smoked by the mother while pregnant, per day.
$fa\min c$ = annual family income, in thousands of dollars.

- what if we use bweght in kilograms?

1 kg. = 1,000 g

$$\widehat{bweght}_{kg} = \frac{\widehat{bweght}_g}{1{,}000} = \frac{\hat{\beta}_0}{1{,}000} + \frac{\hat{\beta}_1 cigs}{1{,}000} + \frac{\hat{\beta}_2 faminc}{1{,}000}$$

$$= \hat{\alpha}_0 + \hat{\alpha}_1 cigs + \hat{\alpha}_2 faminc$$

$$\hat{\alpha}_0 = \frac{\hat{\beta}_0}{1000}, \quad \hat{\alpha}_1 = \frac{\hat{\beta}_1}{1000}, \quad \hat{\alpha}_2 = \frac{\hat{\beta}_2}{1000}$$

all intercept and slope will be less than the old one by a thousand time.

- what if we use faminc in USD (instead of 1,000 USD)

$$\widehat{bweght}_g = \hat{\beta}_0 + \hat{\beta}_1 cigs + \frac{\hat{\beta}_2}{1000} faminc_{USD}$$

the value of this variable is going to be 1000 times larger

the effect will be lesser bc. we use 1 dollar

$$\Rightarrow \hat{\theta}_2 = \frac{\hat{\beta}_2}{1{,}000}$$

in other words $\hat{\theta}_2$ = impact of 1 USD ↑ income
$\hat{\beta}_2$ = " —— " 1,000 USD ↑ income

- what if we use bweght in kg & income in THB

$$bweght_{kg} = \frac{\hat{\beta}_0}{1000} + \frac{\hat{\beta}_1}{1000} cigs + \frac{\hat{\beta}_2}{1000} faminc_{THB}$$

30,000 → effect get smaller :)

this value is going to be 30,000 times more than famine (1K USD)

change in unit of measurement will not change t-Stat
P-value

implication
significant
} will not change.

2 More on functional forms

- Logarithmic Functional Form

usually means natural log

$$\log(y) = \beta_0 + \beta_1 \log(x_1) + \beta_2 x_2 + u$$

$\frac{d}{dx} \ln x = \frac{1}{x}$

$d \ln x = \frac{1}{x} dx$

$$\beta_1 = \frac{d \log(y)}{d \log(x_1)} = \frac{\frac{1}{y} dy}{\frac{1}{x_1} dx_1} = \frac{\frac{1}{y} \Delta y}{\frac{1}{x_1} \Delta x_1} = \frac{\frac{1}{y}(y_1 - y_2)}{\frac{1}{x_1}(x_1 - x_2)} = \frac{100 \times \frac{1}{y} \Delta y}{100 \times \frac{1}{x} \Delta x} = \frac{\%\Delta y}{\%\Delta x}$$ ← elasticity

with the log y & log x format, the coffient is elasticity (x_1 elasticity of y)

p / Q demand

$$\beta_2 = \frac{d \log(y)}{d x_2} = \frac{\frac{1}{y} dy}{d x_2} = \frac{\frac{1}{y} \Delta y}{\Delta x_2}$$

if we want the upper term to be 100% change then

$$100 \beta_2 = \frac{100 \frac{1}{y} \Delta y}{\Delta x_2}$$

$$100 \beta_2 = \frac{\%\Delta y}{\Delta x_2}$$

$100 \beta_2 = \%\Delta$ in y given that $x_2 \uparrow$ by 1 unit.

- Models with Quadratics (squares)

→ capture increasing / decreasing marginal effects (slope of the relationship btw x and y is not constant)

COVID-19 example

Y (# cases)

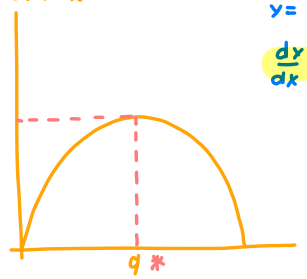
$y = \beta_0 + \beta_1 x + \beta_2 x^2 + u$

$\frac{dy}{dx} = \beta_1 + 2\beta_2 x$

(+) (+) days

Decreasing in marginal effect.

profit (π)



same x!

$y = \beta_0 + \beta_1 x + \beta_2 x^2 + u$

$\frac{dy}{dx} = \beta_1 + 2\beta_2 x$

(+) (−)

$\pi = (P - MC) q$; $MC = 10$

$\pi = (100 - q - 10) q$

f.o.c $\frac{d\pi}{dq} = 0 = 90 - 2q$ (β_1 (+), β_2 (−))

Example : Effects of Pullution on Housing Prices

$$\log(price) = \beta_0 + \beta_1 \log(nox) + \beta_2 \log(dist) + \beta_3 rooms + \beta_4 room^2 + \beta_5 stratio + u$$

where $\log(price) = \beta_0 + \beta_1 \log(nox) + \beta_2 \log(dist) + \beta_3 rooms + \beta_4 room^2 + \beta_5 stratio + u$

$price$ = housing price
nox = level of pollution
$dist$ = distance from downtown
$rooms$ = number of rooms
$stratio$ = average student per teacher ratio (low → good school → P. of Houses)
The estimation result is given by

```
. regress lprice lnox dist rooms rooms_sq stratio
```

Source	SS	df	MS		
Model	51.4933152	5	10.298663	Number of obs	= 506
Residual	33.0889098	500	.06617782	F(5, 500)	= 155.62
				Prob > F	= 0.0000
Total	84.582225	505	.167489554	R-squared	= 0.6088
				Adj R-squared	= 0.6049
				Root MSE	= .25725

log(price)

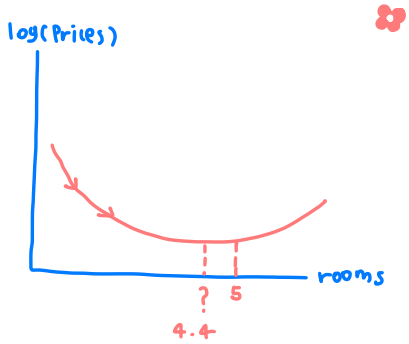
lprice	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
log(nox) — lnox	-.9767545	.0995938	-9.81	0.000	-1.172429	-.7810806
dist	-.0321972	.0094013	-3.42	0.001	-.050668	-.0137264
rooms	-.5528032	.1612965	-3.43	0.001	-.8697056	-.2359007
rooms_sq	.0624697	.0124867	5.00	0.000	.0379368	.0870025
stratio	-.0486667	.0058131	-8.37	0.000	-.0600879	-.0372455
_cons	13.59154	.5650901	24.05	0.000	12.4813	14.70178

|t| > 1.96
all variables are significant.

p < 0.05
we can reject at 5% level of significants.

Consider the effect of "room"

$$\frac{d\log(price)}{d\,rooms} = \beta_3 + 2\beta_4 rooms = -0.553 + 2(0.062)\,rooms$$



At how many rooms does 1 additional room has a possitive impact on log price?

$0 = -0.553 + 2(0.062)\,rooms$

$rooms = 4.4$

at 4.4 rooms or more # ≈ 5 rooms or more

* What wold be the % change in price when the number of room increases from 5 to 6?

$$\frac{d\log(Price)}{d\,rooms} = -0.553 + 2(0.062)\cdot rooms$$

$$\frac{100 \cdot \frac{1}{price}\, price}{d\,rooms} = 100(-0.553) + 100(2(0.062))\cdot(5)$$

$= 100 \times 0.067 = 6.7\%$ increase

what about % in price when # rooms increases 5 to 7 ??

$\%\,\Delta price = 100(-0.553 + 2(0.062)\cdot 6) = 19.1\%$

total impact : total % Δ in price when # ↑ from 5 to 7 is
6.7 + 19.1 = 25.8 %

3 Models with Interaction Terms

→ used when the impact of one variable depends one the value (value) of another variable.

Consider

$$price = \beta_0 + \beta_1 sqrft + \beta_2 bdrms + \beta_3 sqrft \times bdrms + \beta_4 bthrms + u$$

(x_1 = $sqrft$, x_2 = $bdrms$, x_3 = $sqrft \times bdrms$, x_4 = $bthrms$)

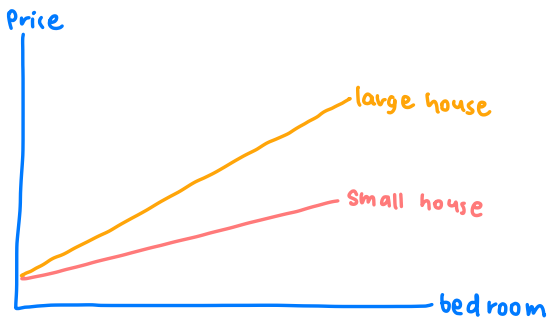
where

$price$ = housing price
$sqrft$ = house size (square feet)
$bdrms$ = number of bedrooms
$bthrms$ = number of bathrooms

$\frac{\Delta price}{\Delta bdrms} = \beta_2 + \beta_3 sqrft$; + add bedroom → ↑ price

← also the size of the House เกี่ยวด้วย it's ⊕ so if ⊕ bedroom มาก ๆ ⇒ Big House → Higher Price.

→ if $\beta_2 > 0$ then, an additional bedroom would ↑ price more for the larger house.



4 More on the Goodness-of-Fit and Selection of Regressors

** • Adding more regressors ALWAYS improve fit → R^2 always increases

Trade off ; we lose "the degree of freedom"

(d.f. = free data point used to est. the parameter)

↘ 1 data point is sacrified everytime we estimate the parameter.

• using R^2 would not punish "Having too many regressors"

• we use adjusted R^2 or $\bar{R}^2$ where we want to punish "Having too many regressors"

$$R^2 = 1 - \frac{SSR}{SST} = 1 - \frac{SSR/n}{SST/n}$$

$$\text{adj. } R^2 = 1 - \frac{SSR/(n-k-1)}{SST/(n-1)}$$

↑ if we have more k, d.f. $= n-k-1$ ↓, $SSR/(n-k-1)$ ↑, adj R^2 ↑

Using adjusted R-squared to choose between non-nested models (one model is not a subset of another).

Consider Model 1

$$\begin{aligned} \widehat{salary} &= \underset{(223.90)}{830.63} + \underset{(0.0089)}{0.0163}sales + \underset{(11.08)}{19.63}roe \\ n &= 209, \quad R^2 = 0.029, \quad \bar{R}^2 = 0.020 \end{aligned}$$

Consider Model 2

$$\begin{aligned} \widehat{\log(salary)} &= \underset{(0.29)}{4.36} + \underset{(0.033)}{0.2751}\log(sales) + \underset{(0.004)}{0.0179}roe \\ n &= 209, \quad R^2 = 0.282, \quad \bar{R}^2 = 0.275 \end{aligned}$$

27.5% of variatio in y is explained, this is better model!

↑ the second is better

Multiple Regression Analysis with Qualitative Information:

1 Outline

- Describing qualitative information (male/Female, seasons)
- Using a single dummy independent variable
- Using dummy variables for multiple categories
- Interactions involving dummy variables
- A binary dependent variable (Y variable): The linear probability model (what happen if y is still qualitative)

2 Describing Qualitative Information

- "Female" and "Married" are qualitative variable.
- We arbitarily assign a dummy variable to decribe them.

(dummy variable)

$$famale = \begin{cases} 1 & \text{if female} \\ 0 & \text{otherwise (or if male)} \end{cases}$$

$$married = \begin{cases} 1 & \text{if married} \\ 0 & \text{otherwise (of if single)} \end{cases}$$

TABLE 7.1

A Partial Listing of the Data in WAGE1.RAW

person	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
⋮	⋮	⋮	⋮	⋮	⋮
525	11.56	16	5	0	1
526	3.50	14	5	1	0

3 Models with a single dummy independent variable

Consider

$$wage = \beta_0 + \delta_0 female + \beta_1 educ + u. \quad (1)$$

where

$$famale = \begin{cases} 1 & \text{if female} \\ 0 & \text{otherwise (or if male)} \end{cases}$$

use with dummy v.

In this case, the δ_0 notation is used to highlight the interpretation of the parameters multiplying dummy variables. In other cases, we can use any notation that is the most convenient.

① $E(wage \mid female, educ) = E(\beta_0 + \delta_0 female + \beta_1 educ + u \mid female, educ)$

$= \beta_0 + \delta_0 female + \beta_1 educ + E(u \mid female, educ)$

all MLR 1-4 statisfy $\rightarrow = 0$

$= \beta_0 + \delta_0 female + \beta_0 educ$

② Thus

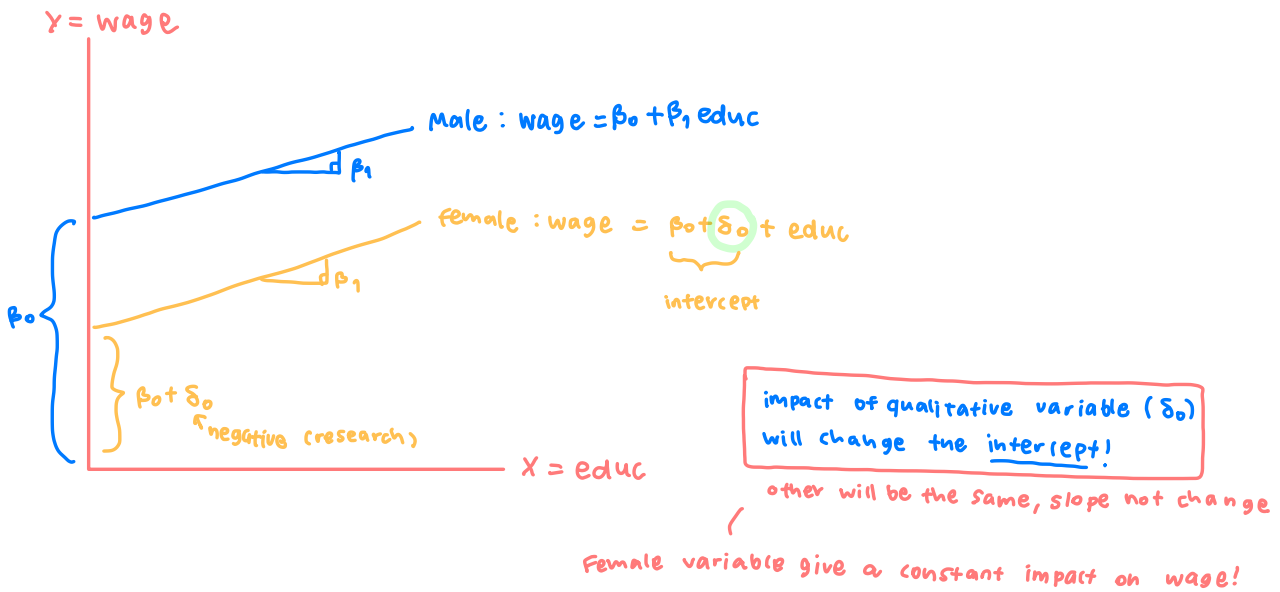
♀ : $E(wage \mid female = 1, educ) = \beta_0 + \delta_0(1) + \beta_1 educ = \beta_0 + \delta_0 + \beta_1 educ.$

♂ : $E(wage \mid male = 0, educ) = \beta_0 + \delta_0(0) + \beta_1 educ = \beta_0 + \beta_1 educ.$ ($= 0$)

$\delta_0 = E(wage \mid female = 1, educ) - E(wage \mid male = 0, educ)$

OR $\delta_0 = E(wage \mid female, educ) - E(wage \mid male, educ)$

* given the same value of educ (same educ level), δ_0 is the diffence in the expected wage of females and males.



4 It is not possible to include all of the dummy alternatives in the same model (as long as there is an intercept in the model)

- If we include all alternatives of a dummy variable in the same model, we will face the "perfect collinearity" problem.

When value of 2 variable are exactly the same.
↳ P in $ and ₿ ~ you have to choose 1.

For example:

$$wage = \beta_0 x_0 + \delta_0 female + \beta_1 educ + \delta_1 male + u$$
$$(x_1) \qquad (x_2) \qquad (x_3)$$

pick 1 → intercept ×1

$$x_0 = x_1 + x_3$$
$$1 = female + male$$
$$female = male + 1$$

or multiple categories. → if there are n categories, we omit 1 category to avoid multicollinearity.

Id	female	Male	
1	1	0	1
2	1	0	1
3	0	1	1
4	0	1	1
⋮	0	1	1
	1	0	1
⋮			
99			

it always add up to 1 !

x_0

$$1 = winter + spring + summer + fall$$
$$winter = 1 - spring - summer - fall$$

omit fall to dont have perfect collinearity.
(it might add up to 0 if we in fall season.)

winter = { 1 if winter, 0 otherwise

spring = { 1 if spring, 0 otherwise

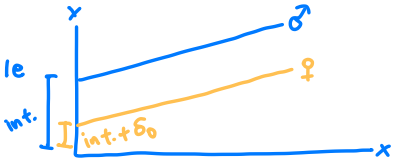
etc.

id	winter	spring	summer	fall	x_0
1	1	0	0	0	1
2	1	0	0	0	1
3	0	0	1	0	1
4	0	0	1	0	1
⋮	0	1	0	0	1
⋮	0	1	0	0	1
	0	1	0	0	1

in this case, male (female is dummy)

- At least one alternative has to be dropped. We treat the dropped alternative as the "BASE GROUP" or "BENCHMARK GROUP".

{ 1 if female, 0 if male } → { 1 if male, 0 if female }




```
. regress lwage female male married educ exper
note: male omitted because of collinearity

      Source |       SS       df       MS              Number of obs =     526
-------------+------------------------------           F(  4,   521) =   75.27
       Model |  54.3265253     4  13.5816313            Prob > F      =  0.0000
    Residual |  94.0032262   521  .180428457           R-squared     =  0.3663
-------------+------------------------------           Adj R-squared =  0.3614
       Total |  148.329751   525  .28253286            Root MSE      =  .42477

------------------------------------------------------------------------------
       lwage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      female |  -.3251146   .0377061    -8.62   0.000    -.3991892    -.25104
        male |           0  (omitted)
     married |   .1380145   .0411197     3.36   0.001     .0572338    .2187953
        educ |   .0872644   .0071554    12.20   0.000     .0732075    .1013213
       exper |   .0076213   .0015314     4.98   0.000     .0046129    .0106297
       _cons |   .4690918   .1040575     4.51   0.000      .264668    .6735156
------------------------------------------------------------------------------
```

being a female workers are expected to have less wage compare to male workers.

5 Using dummy variables for multiple categories

Case 1 We can use many dummy variables in the same model

Consider a model which includes 2 dummy variables– $female$ and $married$.

$$\log(wage) = \beta_0 + \delta_0 female + \delta_1 married + \beta_1 educ + \beta_2 \exp er + \beta_3 \exp er^2 + \beta_4 tenure + \beta_5 tenure^2 + u.$$

1 if female, 0 otherwise; 1 if married, 0 otherwise

```
regress lwage female married educ exper expersq tenure tenursq
```

Source	SS	df	MS
Model	65.6482326	7	9.37831895
Residual	82.6815188	518	.159616832
Total	148.329751	525	.28253286

Number of obs	= 526
F(7, 518)	= 58.76
Prob > F	= 0.0000
R-squared	= 0.4426
Adj R-squared	= 0.4351
Root MSE	= .39952

lwage	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
female	-.2901838	.0361121	-8.04	0.000	-.3611279	-.2192396
married	.0529219	.0407561	1.30	0.195	-.0271456	.1329894
educ	.0791547	.0068003	11.64	0.000	.0657952	.0925143
exper	.0269535	.0053258	5.06	0.000	.0164907	.0374163
expersq	-.0005399	.0001122	-4.81	0.000	-.0007603	-.0003196
tenure	.0312962	.0068482	4.57	0.000	.0178426	.0447499
tenursq	-.0005744	.0002347	-2.45	0.015	-.0010355	-.0001134
_cons	.4177837	.0988662	4.23	0.000	.2235557	.6120116

δ_0, δ_1; not reject H_0

2) δ_1 measures the impact of being married

Comments: (marriage premium) But since $|t| < 1.96$ or $p > 0.05$, we do not reject H_0 of marriage has no impact.

1) δ_0 measures the expected difference btw female & male workers given them same marital status and other factors.

$$\frac{d \log(wage)}{d\, female} = \frac{\frac{1}{wage} d\, wage}{d\, female} = -0.29$$

$\frac{d}{dx} \ln x = \frac{1}{x}$

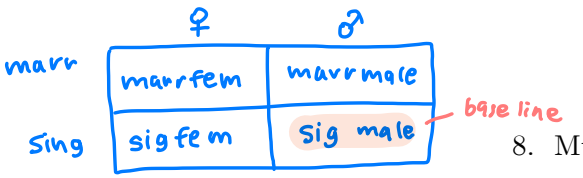
$d \ln x = \frac{1}{x} dx$

% change

$$= \frac{100 \cdot \frac{1}{wage} d\, wage}{d\, female} = \frac{100 \cdot -0.29}{d\, female}$$

$$= \frac{\%\ \Delta\ wage}{d\, female} = 29.02\%$$

female workers are expected to earn less than male workers by 29.02% holding other factors the same.



Consider a model which includes dummy variables for each gender/marital status combination– $marrmale$, $marrfem$ and $singfem$.

$$\log(wage) = \beta_0 + \delta_0 marrmale + \delta_1 marrfem + \delta_3 singfem + \beta_1 educ + \beta_2 exper + \beta_3 exper^2 + \beta_4 tenure + \beta_5 tenure^2 + u. \quad (8.1)$$

(or sigmale is used as the base case)

```
regress lwage marrmale marrfem singfem educ exper expersq tenure tenursq
```

Source	SS	df	MS
Model	68.3617623	8	8.54522029
Residual	79.9679891	517	.154676961
Total	148.329751	525	.28253286

Number of obs	= 526
F(8, 517)	= 55.25
Prob > F	= 0.0000
R-squared	= 0.4609
Adj R-squared	= 0.4525
Root MSE	= .39329

lwage	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
marrmale	.2126757	.0553572	3.84	0.000	.103923	.3214284
marrfem	-.1982676	.0578355	-3.43	0.001	-.311889	-.0846462
singfem	-.1103502	.0557421	-1.98	0.048	-.219859	-.0008414
educ	.0789103	.0066945	11.79	0.000	.0657585	.092062
exper	.0268006	.0052428	5.11	0.000	.0165007	.0371005
expersq	-.0005352	.0001104	-4.85	0.000	-.0007522	-.0003183
tenure	.0290875	.006762	4.30	0.000	.0158031	.0423719
tenursq	-.0005331	.0002312	-2.31	0.022	-.0009874	-.0000789
_cons	.3213781	.100009	3.21	0.001	.1249041	.5178521

no sigmale →

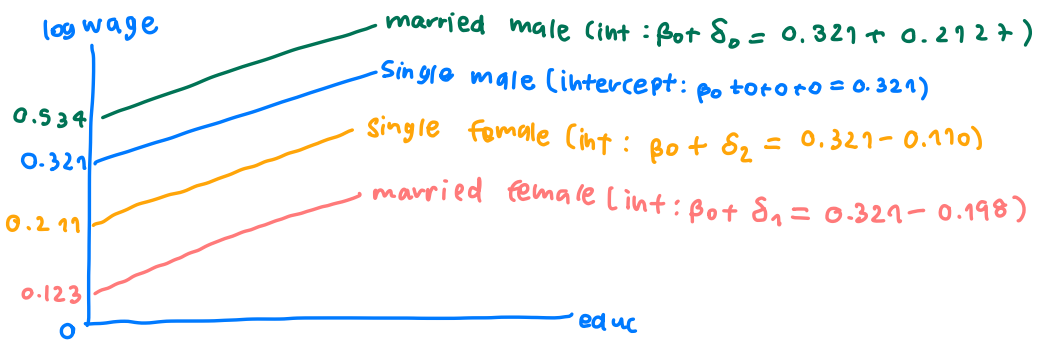
$\hat{\beta}$

Same as the previous case, just intecept that change (qualitative break down)

This regression are not the same as the previous one. It uses "single male" as the base group. (the prevous one use male & single as 2 base group!)

Comments:

- δ_0 measures the expected diff. in wage of married male as compared with single males, holding other factors constants.
- δ_1 measures the expected diff. in wage of married female as compared with single males, holding other factors constants.
- δ_2 measures the expected diff. in wage of single female as compared with single males, holding other factors constants.



Case 2 We can use dummy variables to represent multiple categories of a variable
Consider the relationship between law school rankings and starting salaries

$$\log(salary) = \beta_0 + \delta_0 top10 + \delta_1 r11\_25 + \delta_3 r26\_40 + \delta_4 r41\_60 + \beta_1 LSAT + \beta_2 GPA + \beta_3 \log(libvol) + \beta_4 \log(cost) + u.$$

ranking

* In many cases the "rage of value" serve as a better explanatory variable than "value" itself eg. age may explain the model better if spite in to generation like young (0-15) genz (16-29) etc.

where $top10$, $r11\_25$, $r26\_40$, $r41\_60$ would be equal to 1 when the variable $rank$ falls into the appropriate range.

** Rank below 60 would be the base case.

```
. regress lsalary top10 r11_25 r26_40 r41_60 LSAT GPA llibvol lcost
```

Source	SS	df	MS		
Model	9.16538532	8	1.14567316	Number of obs	= 136
Residual	1.2109665	127	.009535169	F(8, 127)	= 120.15
				Prob > F	= 0.0000
Total	10.3763518	135	.076861865	R-squared	= 0.8833
				Adj R-squared	= 0.8759
				Root MSE	= .09765

lsalary	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
top10	.5393428 δ_0	.053542	10.07	0.000	.4333927	.6452928
r11_25	.4716199 δ_1	.0390921	12.06	0.000	.3942637	.548976
r26_40	.2790977 δ_2	.0346972	8.04	0.000	.2104383	.3477571
r41_60	.182382 δ_3	.0283098	6.44	0.000	.126362	.238402
LSAT	.0060482	.0034919	1.73	0.086	-.0008616	.012958
GPA	.1305893	.0818678	1.60	0.113	-.0314122	.2925908
llibvol	.0725522	.0289213	2.51	0.013	.0153221	.1297824
lcost	.0249169	.0283224	0.88	0.381	-.031128	.0809619
_cons	8.363103	.4457314	18.76	0.000	7.481081	9.245125

baseline is ranking 61th and worse

Comments:

baseline ↓ etc.

Rank	top 10	top 11-25	top 26-40
1	1	0	0
2	1	0	0
3	1	0	0
⋮	⋮	⋮	⋮
10	1	0	0
11	0	1	0
12	0	1	0
⋮	⋮	⋮	⋮
25	0	1	0
26	0	0	1
⋮	⋮	⋮	⋮
40	0	0	1

1) δ_0 measure the difference in expected log(salary) of law-school grad from top 10 u compare to expected log(salary) of law-school grad from 61th → worse

2) δ_1 ~ use the same rational ! ☺