

4 Testing Hypotheses about a Single Linear Combination of the Parameter

Consider

take log to make it $\left\{ \begin{array}{l} \text{linear} \\ \text{better } R^2 \end{array} \right.$

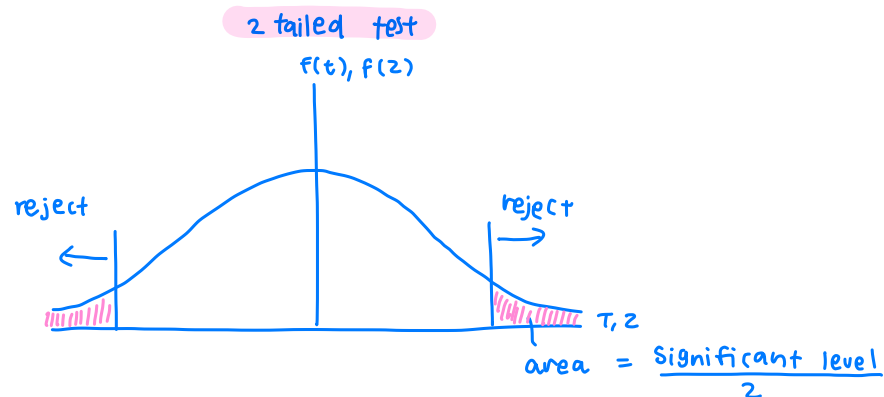
$$\log(\text{wage}) = \beta_0 + \beta_1 jc + \beta_2 \text{univ} + \beta_3 \text{exper} + u$$

where jc = number of years attending a two-year college $univ$ = number of years at a four-year collegeexperience $\rightarrow$ exper = months in the workforce.We want to test whether $\beta_1 = \beta_2$.we want to see whether the reward from JC ^{wage} equal to uni

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

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

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



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

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

?

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

Not very straight forward
to calculate $\rightarrow$ we use
a variable transformation trick
see note!

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

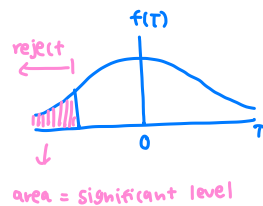
another possible hypothesis test (one-tailed alternative)

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

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

↑
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)

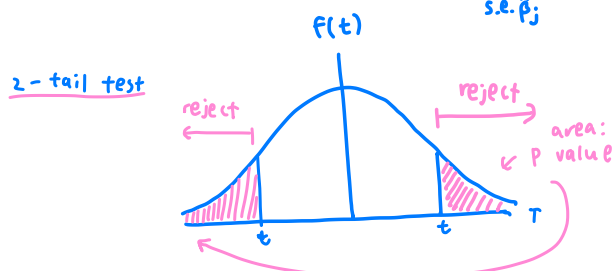
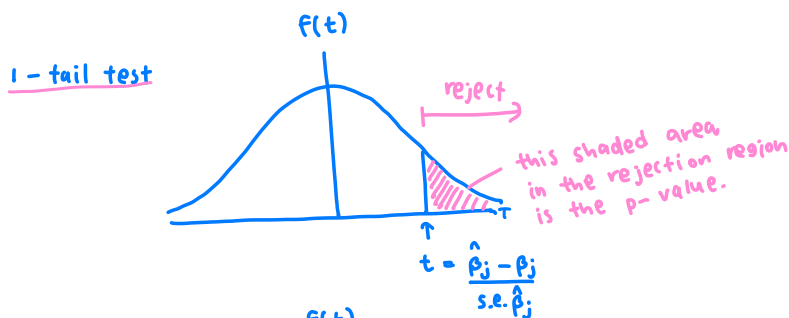


$$t = \frac{(\hat{\beta}_1 - \hat{\beta}_2) - 0}{\text{s.e.}(\hat{\beta}_1 - \hat{\beta}_2)}$$

* 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?



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

In class exercise

In 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) write the t statistic for testing H_0

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

* 2) Define $\theta_1 = \beta_1 - 3\beta_2 \Rightarrow H_0 = \theta_1 = 1$

$H_a = \theta_1 \neq 1$

$t = \frac{\hat{\theta}_1 - 1}{\text{s.e.}(\hat{\theta}_1)}$ — 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

$$y = \beta_0 + (\theta_1 + 3\beta_2)X_1 + \beta_2 X_2 + \beta_3 X_3 + u$$

$$= \beta_0 + \theta_1 X_1 + 3\beta_2 X_1 + \beta_2 X_2 + \beta_3 X_3 + u$$

$$Y = \beta_0 + \theta_1 X_1 + \beta_2 (X_2 + 3X_1) + \beta_3 X_3 + u$$

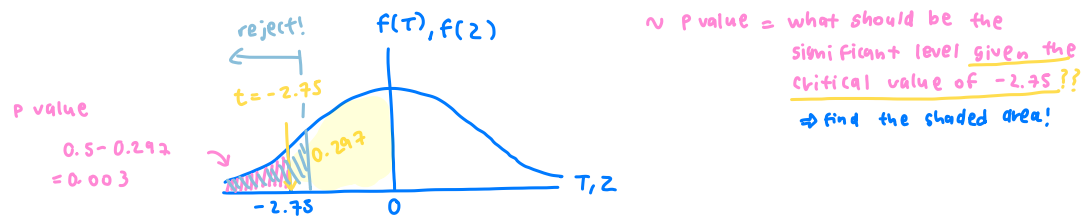
Now, the explanatory variables are going to be

X_1 , $X_2 + 3X_1$ and X_3

→ we can calculate

$$t = \frac{\hat{\theta}_1 - 1}{\text{s.e.}(\hat{\theta}_1)}$$

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



$\rightarrow$ 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

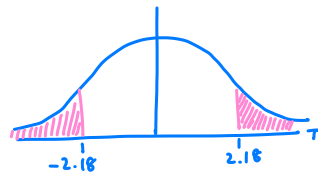
• Thus, p-value = 0.003

• Would we reject H_0 if we use the significance level = 5%? **Yes!**

**** rule!** we reject H_0 if p value < sig. level

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

2 tail test τ_{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 H_0 because the area is less than 0.05 or $p < 0.05$

6 Confidence Intervals (CI)

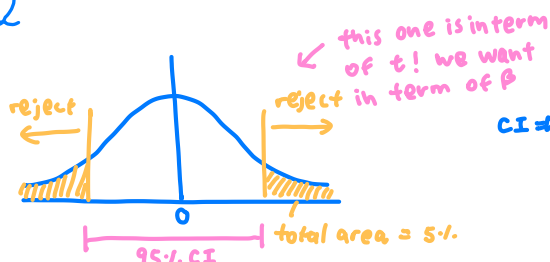
• Confidence Intervals for the POPULATION PARAMETER (β_j)

the range of values that would capture the true β_j 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."



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

this one is in term of t! we want reject in term of β
c is the **99.5** percentile in the t-distribution with $n-k-1$ d.f.
Significant level in 1 side! bc. we have to look at the table!
100 - 2.5

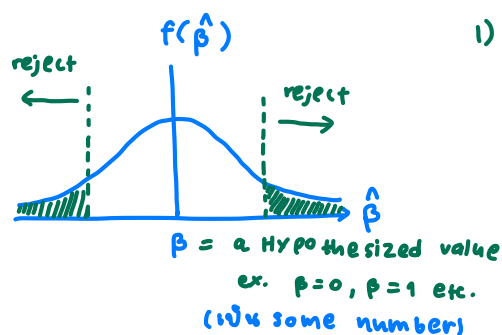
Inference

Hypothesis testing about " β " the true parameter.

$$\text{wage} = \beta_0 + \beta_1 \text{educ} + \beta_2 \text{experience} + \dots + u$$

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

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



1) test if $\beta =$ some number

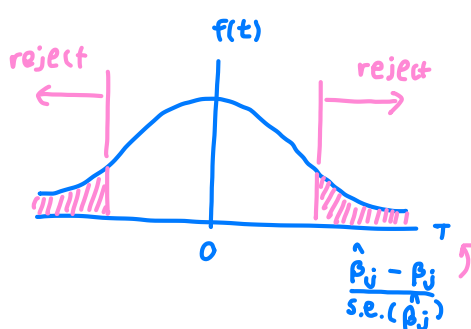
e.g. $\beta_j = 0 \rightarrow x_j$ has no impact on y

$\beta_j = 1 \rightarrow 1$ unit $\uparrow$ in x_j correspond to 1 unit $\uparrow$ in y

$\Rightarrow$ t-test!

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

$\downarrow$ change this to t standardize



5% significant level

= total area in rejection region.

ass. df: 100

$$\text{area} = 2 \times (0.5 - 0.4894)$$

$$= 0.0394$$

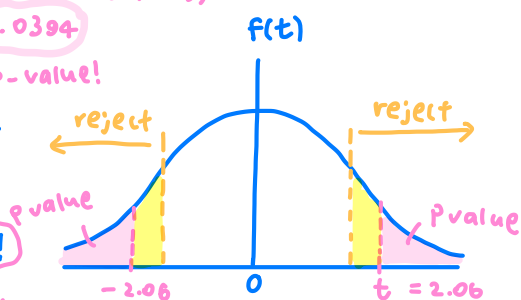
p-value!

in 5% reject region.

we reject H_0 !

$$0.5 > 0.0394$$

we reject H_0 !



• suppose, we calculate a

$$t\text{-statistic} = \frac{\hat{\beta}_j - \beta_j}{\text{s.e.}(\hat{\beta}_j)} = 2.06$$

• suppose, we are testing

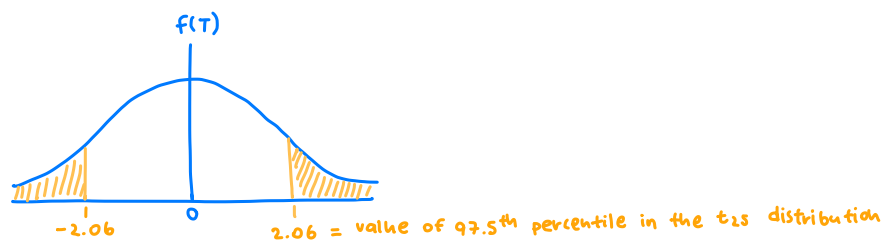
$$H_0: \beta_j = 0$$

$$H_a: \beta_j \neq 0$$

2 tail test = total shaded area.

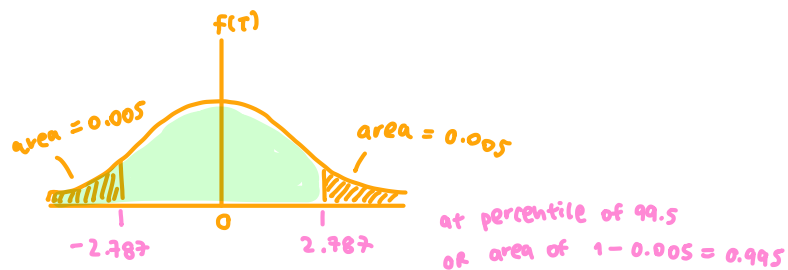
P-value = significant level which we will reject the H_0 or prob that we will reject H_0

if p-value < significant level $\Rightarrow$ reject H_0 (regardless the # tail :))

Example 1: **95% CI**

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

↑
it's a range!
|-----|

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

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

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

- F - test motivation -

$$\tilde{Q} \beta_1, \beta_0 = 0$$

are we still add them in our UR?

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

$$\text{Grade}_{325} = \beta_0 + \beta_1 \# \text{ times - front} + \beta_2 \# \text{ times - back} + \beta_3 \text{ hr_Study} + \beta_4 \text{ past_GPA} + \beta_5 \text{ gender} + u$$

↓

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

$$\left. \begin{array}{l} \text{OR } \beta_1 \neq 0 \quad \text{and} \quad \beta_2 \neq 0 \\ \text{OR } \beta_1 \neq 0 \quad \text{and} \quad \beta_2 = 0 \\ \text{OR } \beta_1 = 0 \quad \text{and} \quad \beta_2 \neq 0 \end{array} \right\} \text{at least 1 of the } \beta_1, \beta_2 \neq 0.$$

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

$$F \equiv \frac{(\text{SSR}_r - \text{SSR}_{ur})}{q} \cdot \frac{n-k-1}{\text{SSR}_{ur}}$$

always has lower u bc. it has more x ! → better explained
So $\text{SSR}_{ur} < \text{SSR}_r$ it always ⊕

d.f. of the "ur" model

ur - more x

→ better explained

→ $R^2 \downarrow$

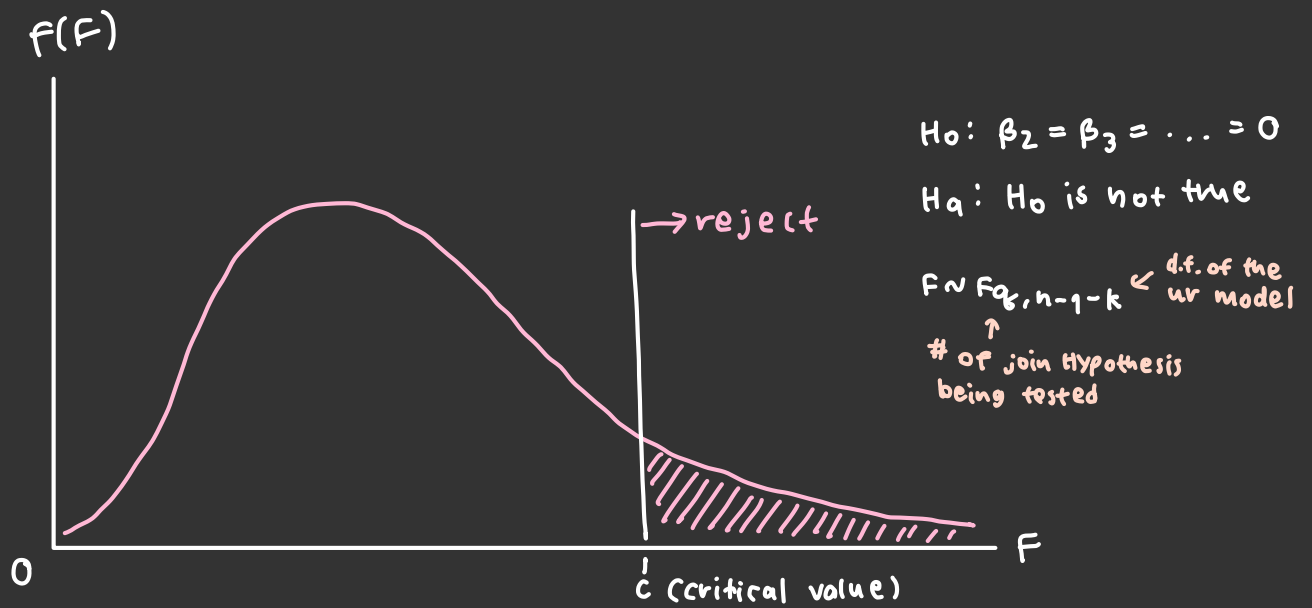
→ $\text{SSR} \downarrow$

→ But variance ↑

hard to predict precisely

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

can $\downarrow R^2$, $\downarrow \text{SSR}$ enough



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

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

7 Testing Multiple Linear Restrictions: The F-test

Suppose the model is specified by

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

$$H_0 : \beta_2 = 0 \text{ and } \beta_3 = 0 \rightarrow \text{want to test if } x_1 \text{ and } x_2 \text{ BOTH have no impact on } y!$$

$$H_a, H_1 : H_0 \text{ is not true}$$

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

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

- 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

$$y = \beta_0 + \beta_1 x_1 + u \text{ — is true } \Rightarrow \text{do not reject } H_0$$

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 model our $q = 2$

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

↑ $\beta_{k-q+1}, \dots, \beta_k$

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

(the last q $\beta = 0$)

$H_a : H_0$ is not true.

So, we test whether we want to include $x_{k-q+1}, x_{k-q+2}, \dots, x_k$ variable in our model

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

restricted model.

unrestricted model.

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

3. Some useful facts

① $R_{ur}^2 > R_r^2$ 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 does not improve the model enough

4. Other ways to calculate the F-statistics:

→ From $R^2 = 1 - \frac{SSR}{SST}$ (RSS / TSS)

we have $F \equiv \frac{(R_{ur}^2 - R_r^2)}{\frac{q_f \in \# \text{ of } \beta \text{ that set to 0}}{(1 - R_{ur}^2)}}$
n-k-1 int. # Slope of β # of observation

if we want to test the overall significant of the model

F always given by $H_0 : \beta_1 = \beta_2 = \beta_3 = \dots = \beta_k = 0$, H_a : other wise

$F \equiv \frac{R^2/k}{1 - R^2/(n-k-1)}$ (R^2 of the model $\approx UR$ 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)

indicator of performance	{	<i>salary</i>	= season salary
	{	<i>years</i>	= years in major leagues
		<i>gamesyr</i>	= games per year in the league
		<i>bavg</i>	= career batting average
		<i>hrunsyr</i>	= homeruns per year
		<i>rbisyr</i>	= 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

K=5

this is lower

not excess 1.96

Each of them don't excess 1.96
 none of them has a significant impact at 5%, we cannot reject H_0 (that say $H_0: \beta = 0$)

So, each of them cannot explain salary base on our methode (cannot reject H_0) but we want to test whether they can jointly explain $Y \rightarrow F$ test. $\rightarrow$ if jointly we include in the model. (UR??)

R model

↓

- 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

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

rule: we reject H_0 if p value < sig. level

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

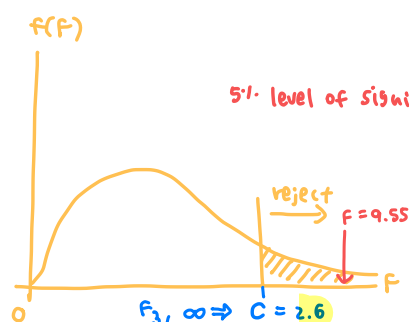
$$F \equiv \frac{(SSR_r - SSR_{UR}) / q}{SSR_{UR} / (n - k - 1)}$$

$$F \equiv \frac{(198.311477 - 183.186327) / 3}{183.186327 / (353 - 5 - 1)} \approx 9.55$$

 R^2 from UR

HW:

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

 $\equiv ??$ 

5% level of significant

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% 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 don't put it on model
 drop bc. 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

very UR of $F \rightarrow$ 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)	value of β changes
log(mktval)	—	.112 (.050)	.100 (.049)	
profmarg	—	-.0023 (.0022)	-.0022 (.0021)	
ceoten	—	—	.0171 (.0055)	impact on sale (depend on direction of bias)
comten (# year work with comp.)	—	—	-.0092 (.0033)	
intercept	4.94 (0.20)	4.62 (0.25)	4.57 (0.25)	
Observations	177	177	177	
R-squared	.281	.304	.353	

most restricted unrestricted

like a simple regression (only 1 x)

other com. performance {

CEO characteristics {

Sales $\rightarrow$

Multiple Regression Analysis : Further Issues

1 Data scaling on OLS statistics

When we change the **unit of measurement** of a variable, the value of estimators would change accordingly. For example

$$\widehat{bwght}_g = \hat{\beta}_0 + \hat{\beta}_1 \text{cigs} + \hat{\beta}_2 \text{faminc},$$

$\nwarrow$ weight
 $\rightarrow$ ex. currency

where

$bwght$ = child birth weight, in grams.

$cigs$ = number of cigarettes smoked by the mother while pregnant, per day.

$faminc$ = annual family income, in thousands of dollars.

- what if we use $bwght$ in kilograms?

$$\begin{aligned}
 1 \text{ kg.} &= 1,000 \text{ g} \\
 \widehat{bwght}_{kg} &= \frac{\widehat{bwght}_g}{1,000} = \frac{\hat{\beta}_0}{1,000} + \frac{\hat{\beta}_1 \text{cigs}}{1,000} + \frac{\hat{\beta}_2 \text{faminc}}{1,000} \\
 &= \hat{\alpha}_0 + \hat{\alpha}_1 \text{cigs} + \hat{\alpha}_2 \text{faminc} \\
 \hat{\alpha}_0 &= \frac{\hat{\beta}_0}{1,000}, \quad \hat{\alpha}_1 = \frac{\hat{\beta}_1}{1,000}, \quad \hat{\alpha}_2 = \frac{\hat{\beta}_2}{1,000}
 \end{aligned}$$

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)

$$\begin{aligned}
 \widehat{bwght}_g &= \hat{\beta}_0 + \hat{\beta}_1 \text{cigs} + \frac{\hat{\beta}_2 \text{faminc}}{1,000} \text{ USD} \\
 \Rightarrow \hat{\theta}_2 &= \frac{\hat{\beta}_2}{1,000}
 \end{aligned}$$

$\uparrow$
 the value of this variable is going to be 1000 times larger
 the effect will be lesser bc. we use 1 dollar

in other words

$\hat{\theta}_2$ = impact of 1USD $\uparrow$ income
 $\hat{\beta}_2$ = " " " 1,000 USD $\uparrow$ income

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

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

$\downarrow$
 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 } will not change.
significant

2 More on functional forms

- Logarithmic Functional Form

↳ usually means natural log

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

$$\begin{aligned}\beta_1 &= \frac{d \log(y)}{d \log(x_1)} \\ &= \frac{\frac{1}{y} dy}{\frac{1}{x_1} dx_1} \\ &= \frac{\frac{1}{y} \Delta y}{\frac{1}{x_1} \Delta x_1} \\ &= \frac{\frac{1}{y} (y_1 - y_2)}{\frac{1}{x_1} (x_1 - x_2)} \\ &= \frac{100 \times \frac{1}{y} \Delta y}{100 \times \frac{1}{x} \Delta x} \\ &= \frac{\% \Delta y}{\% \Delta x} \leftarrow \text{elasticity}\end{aligned}$$

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

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

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

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

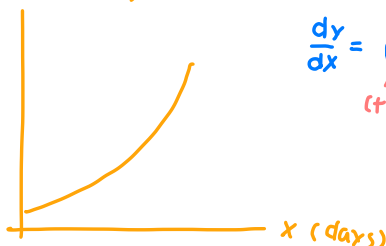
$100 \beta_2 = \% \Delta \text{ in } y$ given that $x_2 \uparrow$ by 1 unit.

with the log y & log x format, the coefficient is elasticity
(x_1 elasticity of y)
P
Q demand

- 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)



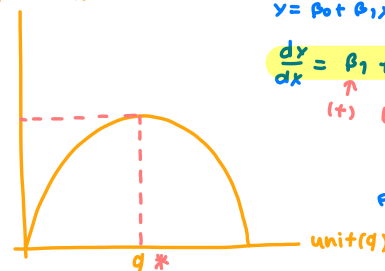
$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + u$$

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

(+) (+) days

Decreasing in marginal effect.

profit (π)



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

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

(+) (-)

$$\pi = (p - mc) q_f ; mc = 10$$

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

$$f.o.c \quad \frac{d\pi}{dq_f} = 0 = \beta_1 (+) + \beta_2 (-)$$

$$90 - 2q_f$$

same x !

Example : Effects of Pollution on Housing Prices

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

where

$price$ = housing price

nox = level of pollution

$dist$ = distance from downtown

$rooms$ = number of rooms

$stratio$ = average student per teacher ratio (low $\rightarrow$ good school $\rightarrow$ ↑ P. of Houses)

The estimation result is given by

regress lprice lnox dist rooms rooms_sq stratio

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

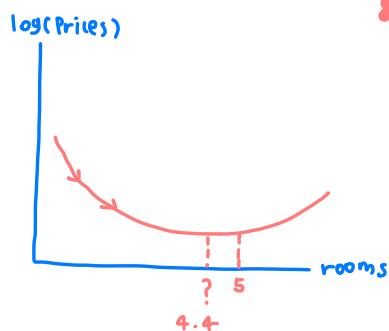
log(price) lprice	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
log(nox) - lnox	-.9767545	.0995938	-9.81	0.000	-1.172429	-.7810806
dist	-.0321972	.0094013	-3.42	0.001	-.050668	-.0137264
rooms	-.5528032	.1612965	-3.43	0.001	-.8697056	-.2359007
rooms_sq	.0624697	.0124867	5.00	0.000	.0379368	.0870025
stratio	-.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 significance.

Consider the effect of "room"

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



At How many rooms does an additional room have a positive impact on log price?

$$0 = -0.553 + 2(0.062) \text{rooms}$$

$$\text{rooms} = 4.4$$

At 4.4 rooms or more #
rooms or more

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

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

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

$$= 100 \times 0.067 = 6.7\% \text{ increase}$$

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

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

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

3 Models with Interaction Terms → used when the impact of one variable depends on the value (value) of another variable.

Consider

$$price = \beta_0 + \beta_1 \underset{x_1}{sqrft} + \beta_2 \underset{x_2}{bdrms} + \beta_3 \overset{x_3}{\underbrace{sqrft \times bdrms}_{x_1 \times x_2}} + \beta_4 \underset{x_4}{bthrms} + u$$

where

$price$ = housing price

$sqrft$ = house size (square feet)

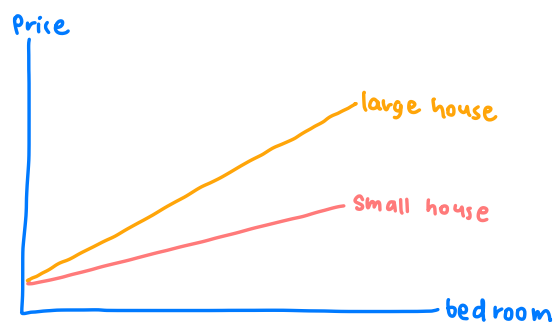
$bdrms$ = number of bedrooms

$bthrms$ = number of bathrooms

$$\frac{dprice}{dbdrms} = \beta_2 + \beta_3 \text{ sqrft} \leftarrow \begin{array}{l} \text{+ ; add bedroom} \rightarrow \uparrow \text{ price} \\ \text{also the size of the House} \end{array}$$

it's ⊕ so if ⊕ bedroom and 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 $\rightarrow R^2$ always increases

Trade off ; we lose "the degree of freedom"

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

$\rightarrow$ 1 data point is sacrificed everytime we estimate the parameter.

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

- we use adjusted R^2 or 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. = $n-k-1 \downarrow$
 $SSR/(n-k-1) \uparrow$, $\text{adj. } R^2 \uparrow$

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

Consider Model 1

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

Consider Model 2

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

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

$\uparrow$
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 arbitrarily assign a dummy variable to describe them.

$$\begin{aligned} \text{dummy variable } female &= \begin{cases} 1 & \text{if female} \\ 0 & \text{otherwise (or if male)} \end{cases} \\ married &= \begin{cases} 1 & \text{if married} \\ 0 & \text{otherwise (of if single)} \end{cases} \end{aligned}$$

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

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

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.

$$\begin{aligned} \textcircled{1} \quad E(wage | female, educ) &= E(\beta_0 + \delta_0 female + \beta_1 educ + u | female, educ) \\ &= \beta_0 + \delta_0 female + \beta_1 educ + E(u | female, educ) \\ &= \beta_0 + \delta_0 female + \beta_1 educ \end{aligned}$$

all MLR 1-4 satisfy $\downarrow = 0$

② Thus

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

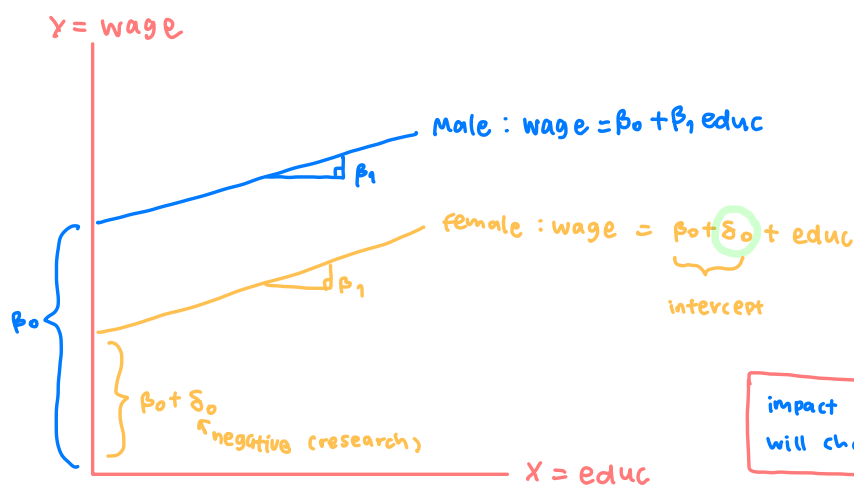
$$m: E(wage | male = 0, educ) = \beta_0 + \delta_0(0) + \beta_1 educ = \beta_0 + \beta_1 educ.$$

= 0

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

$$\text{OR } \delta_0 = E(wage | female, educ) - E(wage | male, educ)$$

* given the same value of educ (same educ level),

 δ_0 is the difference in the expected wage of females and males.

other will be the same, slope not change

Female variable give a constant impact on wage!

- 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.

L P in \$ and \$ ~ you have to choose 1.

For example:

$$\text{wage} = \beta_0 x_0 + \beta_1 \text{female} + \beta_2 \text{educ} + \beta_3 \text{male} + u$$

$\downarrow$ (x₁) (x₂) (x₃)
 intercept x₁

pick 1 →

$$x_0 = x_1 + x_3$$

$$1 = \text{female} + \text{male}$$

$$\text{female} = \text{male} + 1$$

or multiple categories.

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

$$x_0 = \text{winter} + \text{spring} + \text{summer} + \text{fall}$$

$$\text{winter} = 1 - \text{spring} - \text{summer} - \text{fall}$$

$$\text{winter} = \begin{cases} 1 & \text{if winter} \\ 0 & \text{otherwise} \end{cases}$$

$$\text{spring} = \begin{cases} 1 & \text{if spring} \\ 0 & \text{otherwise} \end{cases}$$

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	0	1	0	1
...	0	1	0	0	1
...	0	1	0	0	1

omit fall to don't have perfect collinearity.
(it might add up to 0 if we in fall season.)

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 "BASELINE" or "BENCHMARK GROUP".

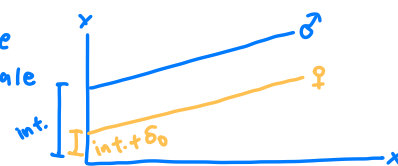
. regress lwage female male married educ exper
note: male omitted because of collinearity

Source	SS	df	MS
Model	54.3265253	4	13.5816313
Residual	94.0032262	521	.180428457
Total	148.329751	525	.28253286

Number of obs = 526
F(4, 521) = 75.27
Prob > F = 0.0000
R-squared = 0.3663
Adj R-squared = 0.3614
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*.

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

$\begin{cases} 1 & \text{if female} \\ 0 & \text{otherwise} \end{cases}$
 $\begin{cases} 1 & \text{if married} \\ 0 & \text{otherwise} \end{cases}$

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

Source	SS	df	MS	Number of obs = 526		
Model	65.6482326	7	9.37831895	F(7, 518) = 58.76		
Residual	82.6815188	518	.159616832	Prob > F = 0.0000		
Total	148.329751	525	.28253286	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

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 their same marital status and other factors.

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

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

∴ change

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

$$= \frac{1 \cdot \Delta \text{wage}}{d \text{female}} = 29.02\%$$

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

	♀	♂
marr	marrfem	marrmale
Sing	sigfem	Sig male

base line

8. Multiple Regression Analysis with Qualitative Information: 85

Consider a model which includes dummy variables for each gender/marital status combination— *marrmale*, *marrfem* and *sigfem*.

$$\log(\text{wage}) = \beta_0 + \delta_0 \text{marrmale} + \delta_1 \text{marrfem} + \delta_2 \text{sigfem} + \beta_1 \text{educ} + \beta_2 \text{exper} + \beta_3 \text{exper}^2 + \beta_4 \text{tenure} + \beta_5 \text{tenure}^2 + u. \quad (8.1)$$

(or sigmale ← used as the base case)

regress lwage marrmale marrfem singfem educ exper expersq tenure tenursq

Source	SS	df	MS	Number of obs =	526
Model	68.3617623	8	8.54522029	F(8, 517) =	55.25
Residual	79.9679891	517	.154676961	Prob > F =	0.0000
Total	148.329751	525	.28253286	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
sigfem	-.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 →

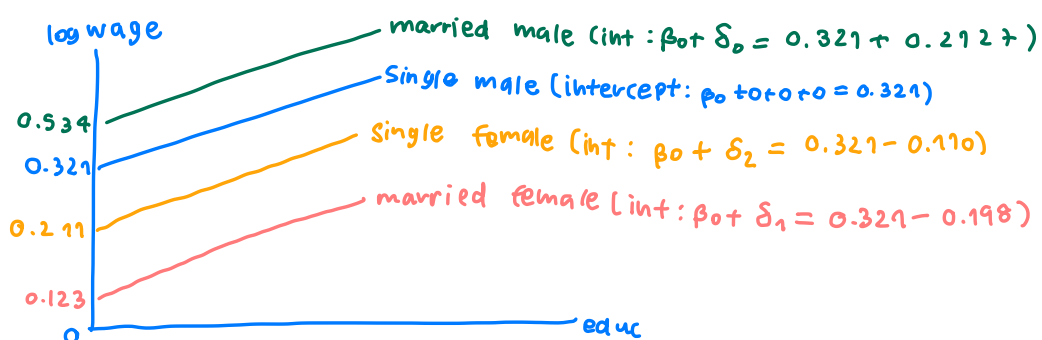
Same as the previous case,

just intercept that change (qualitative breakdown)

Comments:

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

- 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.

$$\log(\text{salary}) = \beta_0 + \delta_0 \text{top10} + \delta_1 r11_25 + \delta_2 r26_40 + \delta_3 r41_60 + \beta_1 \text{LSAT} + \beta_2 \text{GPA} + \beta_3 \log(\text{libvol}) + \beta_4 \log(\text{cost}) + u.$$

* In many cases the "range of value" serve as a better explanatory variable than "value" itself 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.

eg. age may explain the model better if spite in to generation like young (0-15) gen 2 (16-29) etc.

. regress lsalary top10 r11_25 r26_40 r41_60 LSAT GPA llibvol lcost

Source	SS	df	MS	Number of obs =	136
Model	9.16538532	8	1.14567316	F(8, 127) =	120.15
Residual	1.2109665	127	.009535169	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

Comments:

baseline is ranking 61th and worse

Rank	top 10	top 11-25	top 26-40	etc.
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 $\log(\text{salary})$ of law-school grad from top 10 u compare to expected $\log(\text{salary})$ of law-school grad from 61th → worse
- 2) $\delta_1 \sim$ use the same rational! :)