

4 Testing Hypotheses about a Single Linear Combination of the Parameter

Consider

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

where jc = number of years attending a two-year college

$univ$ = number of years at a four-year college

exper = months in the workforce.

We want to test whether $\beta_1 = \beta_2$.

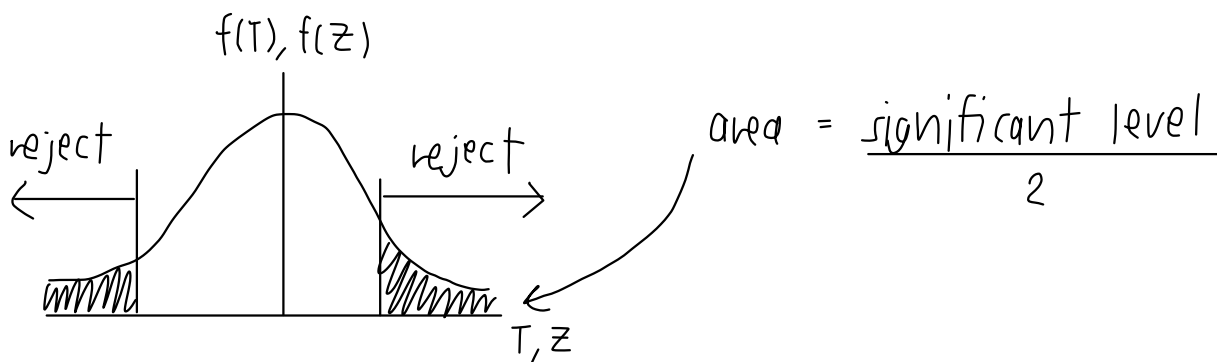
—————→ if the return from 1 more year of education at junior college is the same as that of university

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

against

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

2-tailed test



$$t = \frac{(\hat{\beta}_2 - \hat{\beta}_1) - 0}{2}$$

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

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

Not very straight forward to calculate

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

we use variable transformation tricks

↓
see notes

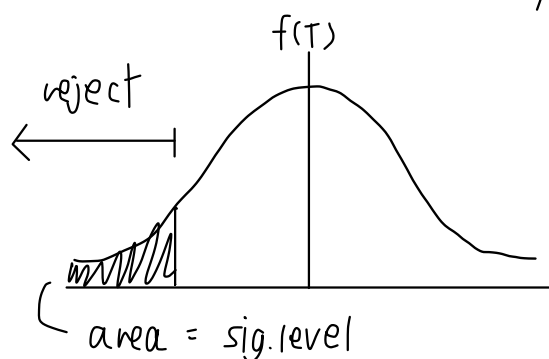
another possible hypothesis test (one-tailed alternative)

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

$$H_a: \beta_1 = \beta_2 \rightarrow H_a: \beta_1 - \beta_2 < 0$$

It would assume that β_1 would not be more than β_2

(Returns to a 2-year college would never be more than return to university education)

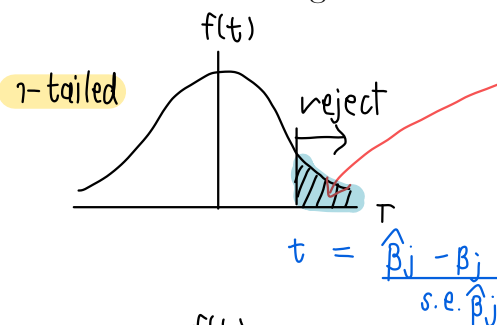


$$t = \frac{(\hat{\beta}_1 - \hat{\beta}_2) - 0}{\text{s.e.}(\hat{\beta}_1 - \hat{\beta}_2)}$$

★ Then, go to the extra note

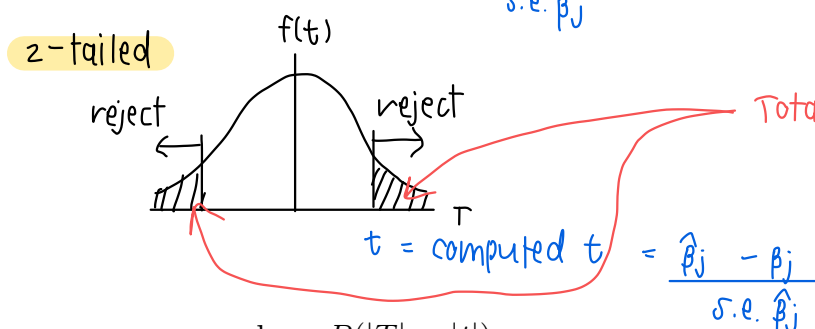
5 Computing p-Values for t-Tests

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



This shaded area in the rejection region is the

p-value



Total area from 2 sides

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

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

t = computed t-statistics

→ p-value = probability that a random T value will be greater (in the || term) than our t 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

1st write t-statistics for testing H_0

$$t = \frac{(\hat{\beta}_1 - 3\hat{\beta}_2) - 0}{\text{s.e.}(\hat{\beta}_1 - 3\hat{\beta}_2)}$$

★ ★ ★

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

$$t = \frac{\hat{\theta}_1 - 1}{\text{s.e.}(\hat{\theta}_1)} \rightarrow \text{we need our regression to have } \theta_1 \text{ in it.}$$

so, STATA or OLS estimation will automatically give $\hat{\theta}_1$ & $\text{s.e.}(\hat{\theta}_1)$

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

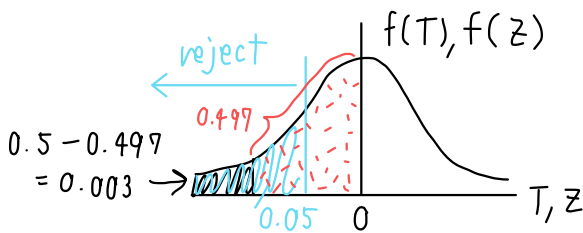
substitute in the main regression and get

$$\begin{aligned} Y &= \beta_0 + (\theta_1 + 3\beta_2)X_1 + \beta_2 X_2 + \beta_3 X_3 + u \\ &= \beta_0 + \theta_1 X_1 + 3\beta_2 X_1 + \beta_2 X_2 + \beta_3 X_3 + u \\ &= \beta_0 + \theta_1 X_1 + \beta_2 (X_2 + 3X_1) + \beta_3 X_3 + u \end{aligned}$$

★ 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



$\rightarrow$ p-value = what should be the significant level, given the critical value of -2.75 ??

$\rightarrow$ Find the shaded area

suppose the calculated $t_{\hat{\beta}_j} = -2.75$
 $= (\hat{\beta}_j - \beta_j) / \text{s.e.}(\hat{\beta}_j)$

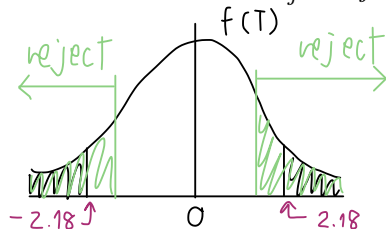
• From the z-table, the value -2.75 corresponds to area = 0.003

• Thus, p-value = 0.003

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

★ 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.



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 + 0.05

• Thus, p-value = is between 0.02 - 0.05

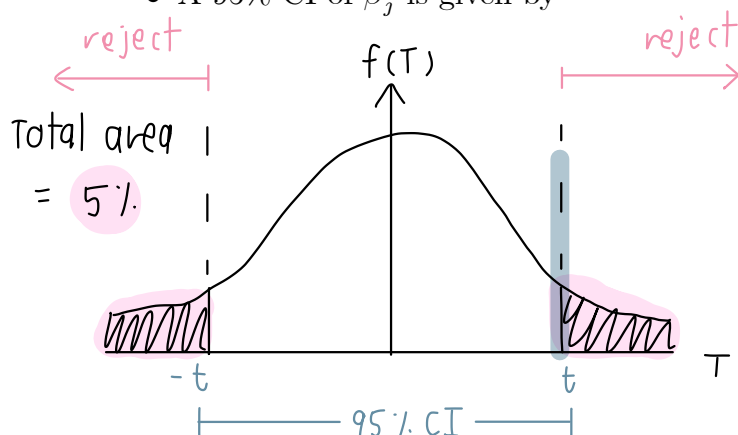
• Would we reject H_0 if we use the significance level = 5%?

6 Confidence Intervals (CI)

• Confidence Intervals for the POPULATION PARAMETER (β_j)

$\hookrightarrow$ the range of value that would capture the true β_j at a 95% chance

• A 95% CI of β_j is given by



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

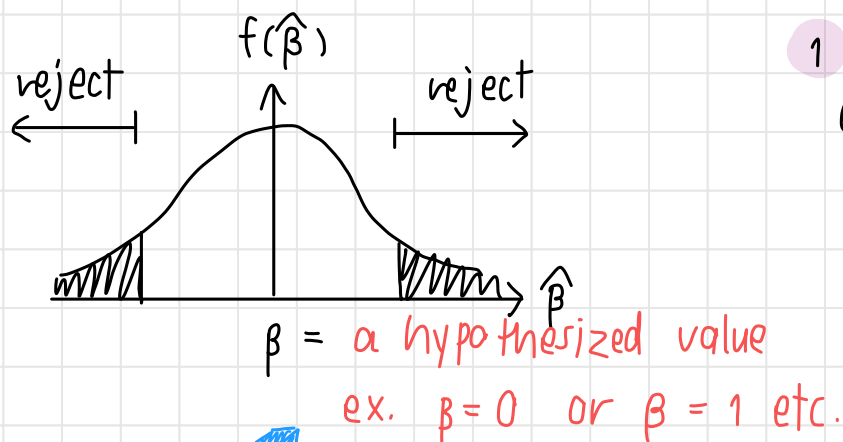
c is the 97.5 percentile in the t-distribution with $n-k-1$ d.f.

Inference → Hypothesis testing about " β " the true parameter

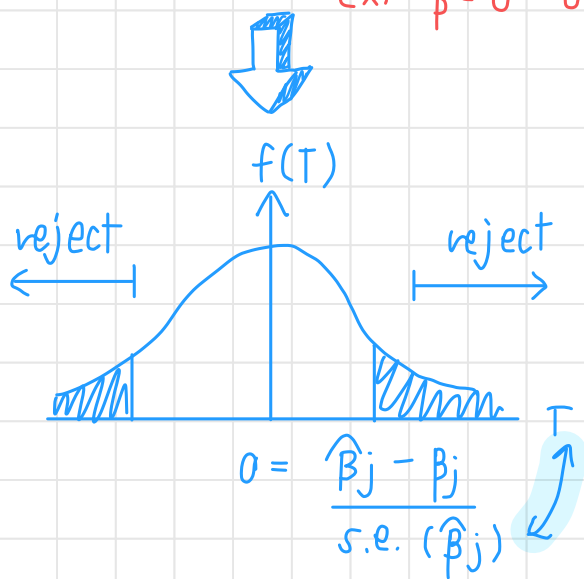
$$\text{wage} = \beta_0 + \beta_1 \text{educ} + \beta_2 \text{experience} + \dots + w$$

We want to test hypotheses about the true impact (β) of each x variables (educ, experience) on the dependent variables (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 hypotheses



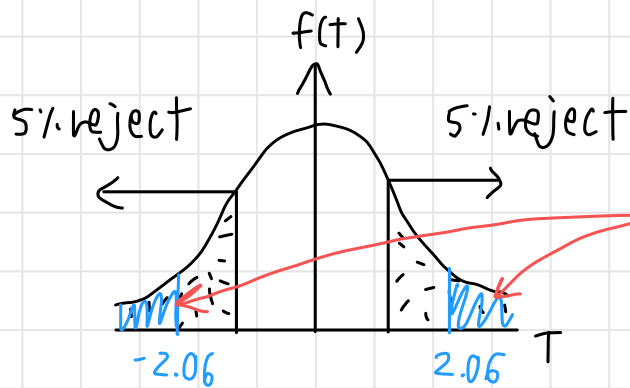
- 1 Test if $\beta =$ same number
eg. $\beta_j = 0 \rightarrow X_j$ has no impact on y
 $\beta_j = 1 \rightarrow 1 \text{ unit } \uparrow \text{ in } X_j \text{ correspond to } 1 \text{ unit } \uparrow \text{ in } y$



⇒ t-test ★ → How ??

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

Significant Level = total area in rejection region



ass. df. = 100

$$\begin{aligned}
 \text{area} &= 2(0.5 - 0.4803) \\
 &= 2 \times 0.0197 \\
 &= 0.0394 = \text{p-value!}
 \end{aligned}$$

◦ Suppose, we calculate a t-statistic = $\frac{\hat{\beta}_j - \beta_j}{\text{s.e.}(\hat{\beta}_j)} = 2.06$

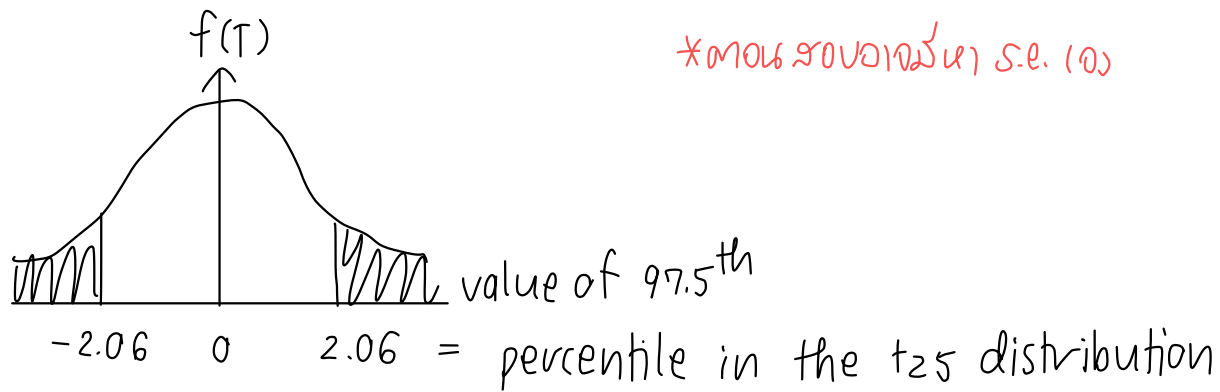
◦ Suppose, we are testing $H_0 : \beta_j = 0, H_a : \beta_j \neq 0$
2-tailed test

◦ p-value = total shaded area

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

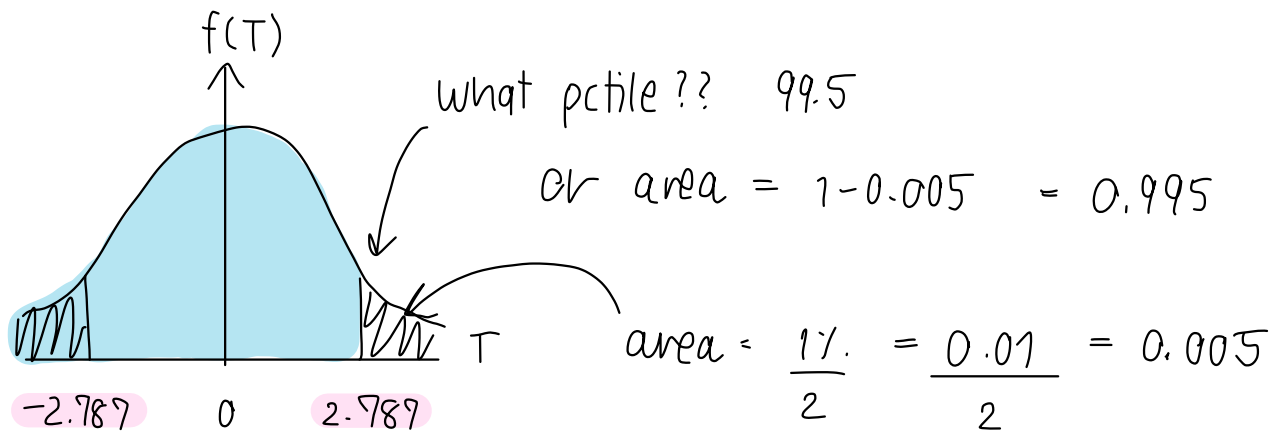
* If p-value < significance level $\rightarrow$ reject H_0

Example 1: **95% CI** d.f. = 25



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

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



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

12 Mar 2020

F-test motivation

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

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

H_0 : seat position doesn't have impact on GPA

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

H_a : seat position matters
 $\beta_1 \neq 0$ and $\beta_2 \neq 0$

or $\beta_1 \neq 0$ and $\beta_2 = 0$

or $\beta_1 = 0$ and $\beta_2 \neq 0$

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

↑ testing 2 β s at the same time → F test

7 Testing Multiple Linear Restrictions: The F-test

Suppose the model is specified by

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

$$H_0 : \beta_1 = 0 \text{ and } \beta_3 = 0$$

$$H_a, H_1 : H_0 \text{ is not true} \rightarrow \text{want to test if } x_1 \text{ and } x_2 \text{ BOTH have no impact on } y$$

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:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + u \text{ is true} \Rightarrow \text{Reject } H_0$$

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

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

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u \text{ is true} \Rightarrow \text{Reject } H_0$$

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

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

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

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

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

ur

$$F \equiv \frac{(SSR_r - SSR_{ur})}{q}$$

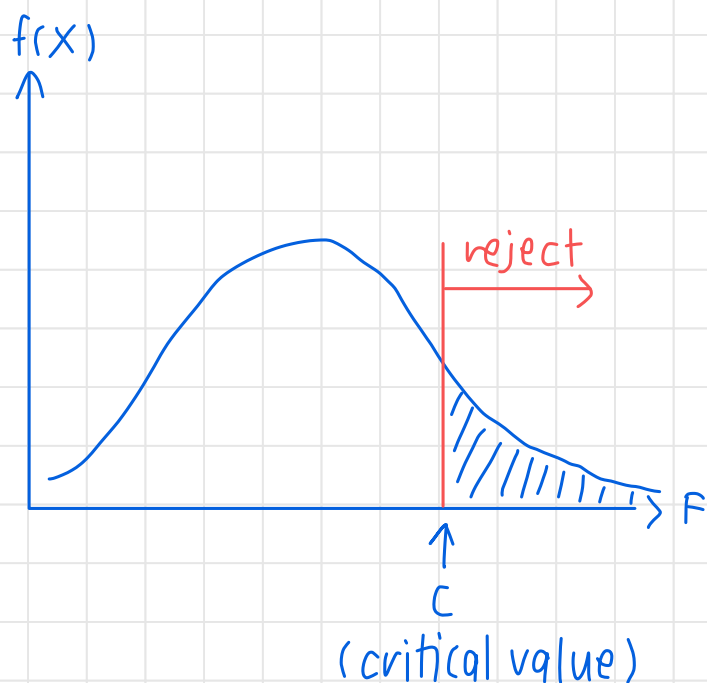
$$\frac{SSR_{ur}}{(n-k-1)}$$

d.f. of the ur. model

This whole term is always $\oplus$ because $SSR_{ur} < SSR_r$. Every time you add 1 more x , the model will be better explained

so, if everytime you add 1 more X variable, the $SSR \downarrow$ and $R^2 \uparrow$,
 why don't we just keep an additional X in the model
 → because everytime we add 1 more X , $\text{Var}(\hat{\beta}_3)$ will increase,
 making the prediction of β less precise. so, we only keep
 the additional X , if it/they can improve the model enough

Can reduce SSR ($\uparrow R^2$) enough
 can significantly reduce SSR and $\uparrow R^2$



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

H_a : H_0 not true

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

of joint hypotheses being tested $\uparrow$ d.f. of the ur model



we reject H_0 of

jointly no effect if $F > c$

3. Some useful facts

- 1 $R^2_{ur} > R^2_r$ because any additional X would increase R^2 (improve fit) $\rightarrow SSR_{ur} < SSR_r$
- 2 By including more X , the model is certainly better explained. However, we would like to reject H_0 if the inclusion of extra variables does not improve the model enough

4. Other ways to calculate the F-statistics:

$$\Rightarrow \text{From } R^2 = 1 - \frac{SSR}{SST} \rightarrow \frac{RSS}{TSS}$$

$$\text{we have } F \equiv \frac{(R^2_{ur} - R^2_r) / q}{(1 - R^2_{ur}) / (n - k - 1)}$$

$\swarrow$ $\nearrow$
 # of β that are set to "0" # of obs. # of slope β intercept

$\rightarrow$ 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 $\approx$ ur
 the " r " model has no X at all

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

$ur \left\{ \begin{array}{l} \\ \\ \\ \\ \end{array} \right. u$	{	<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

The more explanatory variables = The more R^2

6. Multiple Regression Analysis (Inference) 73

ur model

↪ . regress $\log_salary$ years gamesyr bavg hrunsyr rbisyr

	Source	SS	df	MS		
SSE	Model	308.989208	5	61.7978416	Number of obs =	353
SSR	Residual	183.186327	347	.527914487	F(5, 347) =	117.06
					Prob > F	= 0.0000
					R-squared	= 0.6278
SST	Total	492.175535	352	1.39822595	Adj R-squared =	0.6224
					Root MSE	= .72658

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

- the restricted model (r) is defined by

. regress $\log_salary$ years gamesyr

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

	Source	SS	df	MS		
	Model	293.864058	2	146.932029	Number of obs =	353
	Residual	198.311477	350	.566604221	F(2, 350) =	259.32
					Prob > F	= 0.0000
					R-squared	= 0.5971
	Total	492.175535	352	1.39822595	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

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

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

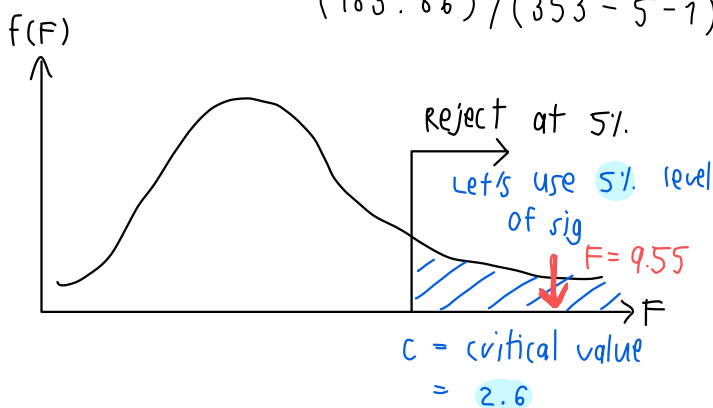
$$= \frac{(198.311 - 183.186) / 3}{(183.186) / (353 - 5 - 1)}$$

≈ 9.55

HW:

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

$$= ???$$



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)

⇒ These t - *statistics* are to test $H_0 : \beta_i = 0$ ↖ z table

⇒ If the d.f. > 30 , then when $t > 1.96$, we can reject H_0 with 5% sig. level

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

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

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

⇒ 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

sales →

company performance {

CEO characteristics {

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	177	177	177
R-squared	.281	.304	.353

↑
like a simple regression with 1 x



Multiple Regression Analysis : Further Issues

1 Data scaling on OLS statistics Different unit of measurement

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

$$\widehat{bweight} = \hat{\beta}_0 + \hat{\beta}_1 cigs + \hat{\beta}_2 faminc,$$

where

$bweight$ = 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 $bweight$ in kilogram?

$$1 \text{ kg.} = 1,000 \text{ g}$$

$$\widehat{bweight}_{\text{kg}} = \frac{\widehat{bweight}_{\text{g}}}{1000} = \frac{\hat{\beta}_0}{1000} + \frac{\hat{\beta}_1}{1000} cigs + \frac{\hat{\beta}_2}{1000} faminc$$

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

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

• what if we use $faminc$ in USD (instead of 1000 USD)

$$bweight_g = \hat{\beta}_0 + \hat{\beta}_1 cigs + \frac{\hat{\beta}_2}{1000} faminc_{\text{USD}}$$

The value of this variable is going to be 1000 times larger than $faminc$

$$= \hat{\beta}_0 + \hat{\beta}_1 cigs + \hat{\theta}_2 faminc_{\text{USD}}$$

$$\Rightarrow \hat{\theta}_2 = \frac{\hat{\beta}_2}{1000}$$

in other words $\hat{\theta}_2$ = impact of 1 USD ↑

$\hat{\beta}_2$ = " ——— " 1,000 USD ↑ in income

◦ what if we use bwght in kg and income in THB

$$\text{bwght}_{\text{kg}} = \frac{\hat{\beta}_0}{1000} + \frac{\hat{\beta}_1}{1000} \text{cigs} + \left(\frac{\hat{\beta}_2}{1000} \right) \text{faminc}_{\text{THB}}$$

30,000 ↑
This value is going to be 30,000 times more than faminc

2 More on functional forms

• Logarithmic Functional Form

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

↓

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

usually means natural log

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

$$\Delta Y = Y_1 - Y_2$$

$$\Delta X = X_{n1} - X_{n2}$$

$$\beta_1 = \frac{d \log(Y)}{d \log(X_1)} = \frac{\frac{1}{Y} dY}{\frac{1}{X_1} dX_1} = \frac{\frac{1}{Y} \Delta Y}{\frac{1}{X_1} \Delta X_1} = \frac{100 \times \frac{1}{Y} \Delta Y}{100 \times \frac{1}{X_1} \Delta X_1} = \frac{\% \Delta Y}{\% \Delta X_1}$$

with the log y and log x format, the coefficient is going to be the elasticity % (x elasticity of y)

$$\beta_2 = \frac{d \log(Y)}{d X_2} = \frac{\frac{1}{Y} dY}{dX_2} = \frac{\frac{1}{Y} \Delta Y}{\Delta X_2}$$

→ if we want the upper term to be % change,

$$\text{then } 100 \beta_2 = \frac{100 \frac{1}{Y} \Delta Y}{\Delta X_2}$$

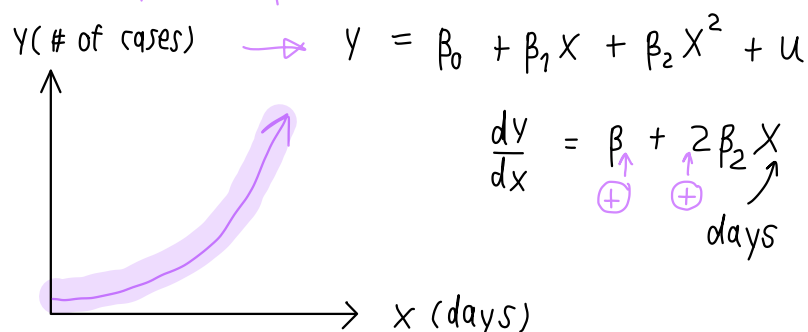
$$100 \beta_2 = \frac{\% \Delta Y}{\Delta X_2}$$

$100 \beta_2 = \% \Delta$ in y given that X_2 increases by 1 unit

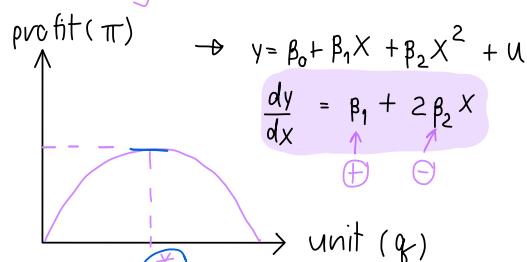
• Models with Quadratics (squares)

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

COVID-19 example



Decreasing returns



Assume $\pi = (P - MC)q$; $MC = 10$
 Demand: $P = 100 - q$
 F.O.C. $= \frac{\partial \pi}{\partial q} = 0 = 90 - 2q$
 β_1 is $\oplus$ β_2 is $\ominus$

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

The estimation result is given by

→ In the U.S. or many other countries, student can apply to schools in the area without having to test any test.
So, the lower # student ratio, the better the school

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
				R-squared =	0.6088
				Adj R-squared =	0.6049
Total	84.582225	505	.167489554	Root MSE =	.25725

log (price)

log (nox)

	lprice	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
lnox	β_1	-.9767545	.0995938	-9.81	0.000	-1.172429 - .7810806
dist	β_2	-.0321972	.0094013	-3.42	0.001	-.050668 - .0137264
rooms	β_3	-.5528032	.1612965	-3.43	0.001	-.8697056 - .2359007
rooms_sq	β_4	.0624697	.0124867	5.00	0.000	.0379368 .0870025
stratio	β_5	-.0486667	.0058131	-8.37	0.000	-.0600879 - .0372455
_cons		13.59154	.5650901	24.05	0.000	12.4813 14.70178

None of them have t less than 1.96, so each of

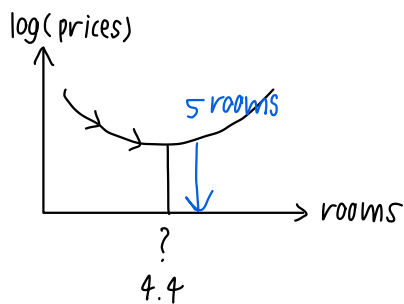
them is statistically significant
 $|t| > 1.96$

All of them are less than 0.05, so we can reject the null of the

coeff = 0 with 5% confident level
 $\alpha < 0.05$

Consider the effect of "room"

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



★ at how many rooms does 1 additional room has a positive impact on log (price)??

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

$$\text{rooms} = 4.4$$

Answer → at 4.4 rooms or more
round up → at 5 rooms or more

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

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

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

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

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

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

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

3 Models with Interaction Terms $\Rightarrow$ use when the impact of one variable depends on the value (level) of another variables

Consider

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

where

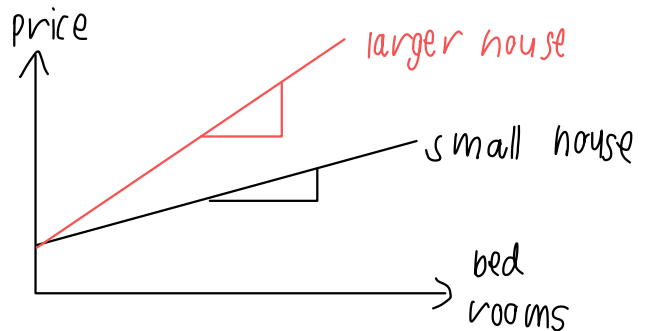
$price$ = housing price

$sqrft$ = house size (square feet)

$bdrms$ = number of bedrooms

$bthrms$ = number of bathrooms

$$\frac{\partial price}{\partial bdrms} = \beta_2 + \beta_3 sqrft$$



$\Rightarrow$ if $\beta_2 > 0$ then, an additional bedroom would increase price more for a larger house!

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

- Adding more regressors ALWAYS improve fit $\rightarrow R^2$ always $\uparrow$
 - But we lose the "degree of freedom"
(d.f. = free data point used to estimate the parameter)
 $\rightarrow$ 1 data point is sacrificed every time we estimate a parameter
 - using R^2 would not punish "having too many regressors"
 - we use $\text{adjusted-}R^2$ or $\bar{R}^2$ when we want to punish adding too many regressors
- $$R^2 = 1 - \frac{SSR}{SST} = 1 - \frac{SSR/n}{SST/n}$$
- $$\text{adj } R^2 = 1 - \frac{SSR/(n-k-1)}{SST/(n-1)} \quad \left\{ \begin{array}{l} \text{If we have more } k, \text{ d.f.} = n-k-1 \downarrow, \\ SSR/(n-k-1) \uparrow, \text{ adj-}R^2 \downarrow \end{array} \right.$$

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

Consider Model 1 None of these model are subset of another

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

Consider Model 2 much superior performance

$$\begin{aligned} \log(\widehat{\text{salary}}) &= 4.36 + 0.2751 \log(\text{sales}) + 0.0179\text{roe} \\ &\quad (0.29) \quad (0.033) \quad (0.004) \\ n &= 209, \quad R^2 = 0.282, \quad \bar{R}^2 = 0.275 \rightarrow 27.5\% \text{ of variation in } y \\ &\quad \text{is explained.} \\ &\quad \text{So, this model is} \\ &\quad \text{better } \nabla \end{aligned}$$

Multiple Regression Analysis with Qualitative Information:

1 Outline

- Describing qualitative information
- 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

2 Describing Qualitative Information

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

$famale$

$=$

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

$married$

$=$

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

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

Assume that this
is a PRF
where

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

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

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.

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

↓ (MLR 1-4 are satisfied)
= 0

2. Thus

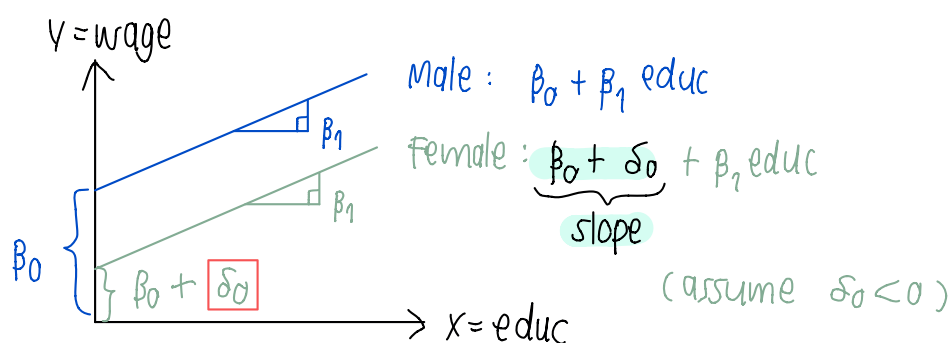
♀: $E(\text{wage} | \text{female} = 1, \text{educ}) = \beta_0 + \delta_0(1) + \beta_1 \text{educ} = \beta_0 + \delta_0 + \beta_1 \text{educ}$

♂: $E(\text{wage} | \text{female} = 0, \text{educ}) = \beta_0 + \delta_0(0) + \beta_1 \text{educ} = \beta_0 + \beta_1 \text{educ}$

$$\delta_0 = E(\text{wage} | \text{female} = 1, \text{educ}) - E(\text{wage} | \text{female} = 0, \text{educ})$$

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

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



→ By the way we model this regression function, "female" is going to give a constant impact on wage, regardless of the level of educ.

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.

$$\text{wage} = \beta_0 X_0 + \delta_0 \text{female} + \beta_1 \text{educ} + \delta_1 \text{male} + u$$

For example: $\text{intercept} \times 1$

(X_1) (X_2) (X_3)

ID	female	male	X_0
1	1	0	1
2	1	0	1
3	0	1	1
4	0	1	1
⋮			
99			

$$X_0 = X_1 + X_3$$

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

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

or If there are "n" categories, we omit "1" categories to avoid multicollinearity

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

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

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

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

etc.

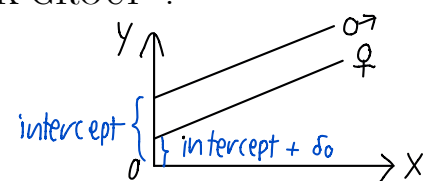
ID	winter	spring	summer	fall	X_0
1	1	0	0	0	1
2	0	1	0	0	1
3	0	0	1	0	1
4	0	0	0	1	1
⋮					

In this case, male

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

$$\begin{cases} 1 & \text{if female} \\ 0 & \text{if male} \end{cases} \quad \begin{cases} 1 & \text{if male} \\ 0 & \text{if female} \end{cases}$$

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



Source	SS	df	MS	Number of obs =	526
Model	54.3265253	4	13.5816313	F(4, 521) =	75.27
Residual	94.0032262	521	.180428457	Prob > F =	0.0000
				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

Female workers are expected to have less wage compared to male workers

5 Using dummy variables for multiple categories

Case 1 We can use many dummy variables in the same modelConsider a model which includes 2 dummy variables– *female* and *married*.

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

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

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

Comments:

- 1 δ_0 measures the expected difference between female and male workers given the same marital status and other factors

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

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

$$\frac{\% \Delta \text{wage}}{\partial \text{female}} = 29.02 \%$$

- Female workers are expected to earn less than male workers by 29.02%, holding other factors the same

	♀	♂
marr	marrfem	marrmale
sing	singfem	singmale

Base case

8. Multiple Regression Analysis with Qualitative Information: 85

Consider a model which includes dummy variables for each gender/marital status combination— *marrmale*, *marrfem* and *singfem*. (or *singmale* ← used as the base case)

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

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
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

Comments:

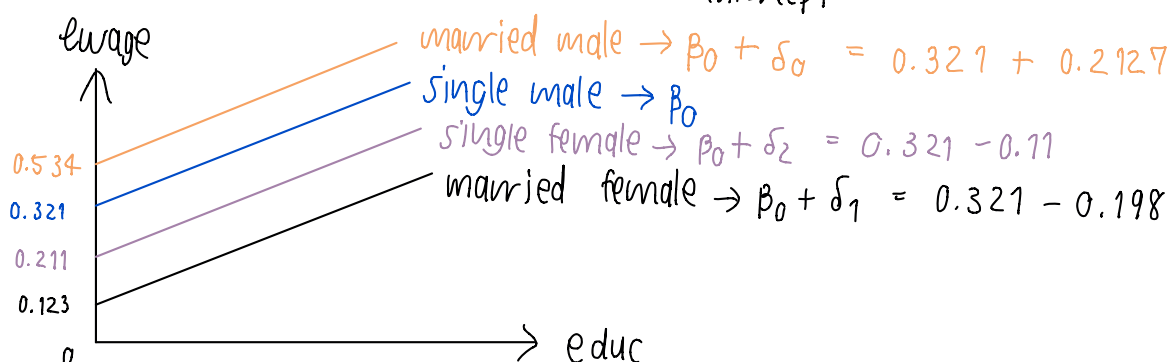
This regression is not the same as the previous one. It uses "singmale" as the base group. (The previous use male and single as 2 base group)

• δ_0 measures the expected different in wage of married male as compared with single males, holding other factors constant

• δ_1 measures the expected different in wage of married female as compared with single males, holding other factors constant

• δ_2 → same rational

intercept



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

$$\log(\text{salary}) = \beta_0 + \delta_0 \text{top10} + \delta_1 r11_25 + \delta_3 r26_40 + \delta_4 r41_60 + \beta_1 \text{LSAT} + \beta_2 \text{GPA} + \beta_3 \log(\text{libvol}) + \beta_4 \log(\text{cost}) + u.$$

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.

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

. regress lsalary top10 r11_25 r26_40 r41_60 LSAT GPA llibvol lcost

age may explain the model better if split into generations.

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

young 0-15
gen Z 16-29
etc.

the baseline is ranking 61th and worse

lsalary	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
top10	.5393428	.053542	10.07	0.000	.4333927 .6452928
r11_25	.4716199	.0390921	12.06	0.000	.3942637 .548976
r26_40	.2790977	.0346972	8.04	0.000	.2104383 .3477571
r41_60	.182382	.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:

rank	top10	r11_25	r26_40
1	1	0	0
2	1	0	0
3	1	0	0
⋮	⋮	⋮	⋮
10	1	⋮	⋮
11	⋮	1	⋮
12	⋮	1	⋮
⋮	⋮	⋮	⋮
25	⋮	1	⋮
26	⋮	⋮	1
⋮	⋮	⋮	⋮
40	⋮	⋮	1
⋮	⋮	⋮	⋮
⋮	⋮	⋮	⋮

1 δ_0 measures the difference in expected $\log(\text{salary})$ of a law-school graduate from a top 10 university compared to expected $\log(\text{salary})$ of those who graduated from the school ranked 61th and worse

2 $\delta_1 \rightarrow$ use the same rational (Top 11-25)