

#### 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

$\text{exper}$  = months in the workforce.

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

if the returns from 1 more years of education at a Junior College is the same as that of the university.

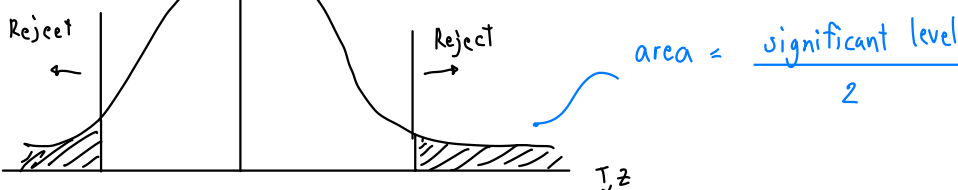
$$H_0: \beta_1 = \beta_2 \Rightarrow H_0: \beta_1 - \beta_2 = 0$$

against

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

2-Tailed Test

$f_{CT}$ ,  $f(z)$



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

We compute this t-statistic and compare with the critical value.

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

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

not very straight forward to calculate

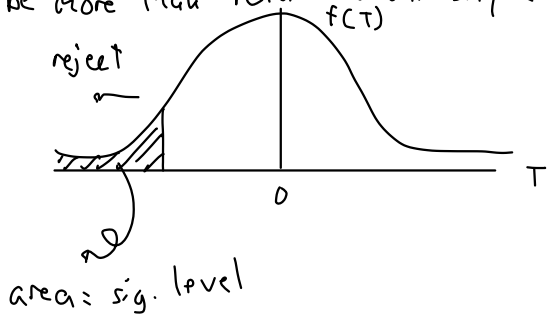
so we use a variable transformation trick → see notes

another possible hypothesis test (one-tailed alternative)

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

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

• It is assumed that  $\beta_1$  would not be more than  $\beta_2$  (returns to a 2-year college would never be more than returns to university education).

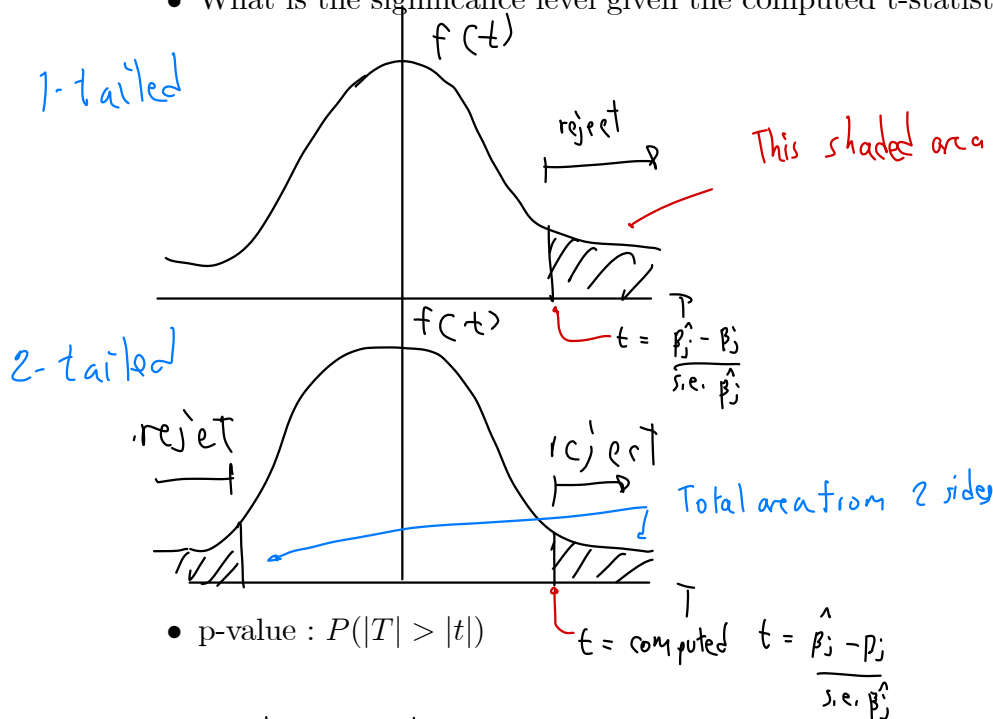


$$t = \frac{(\hat{\beta}_1 - \hat{\beta}_2) - 0}{\text{s.e.}(\hat{\beta}_1 - \hat{\beta}_2)}$$

\* They go to the extra note.

### 5 Computing p-Values for t-Tests

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



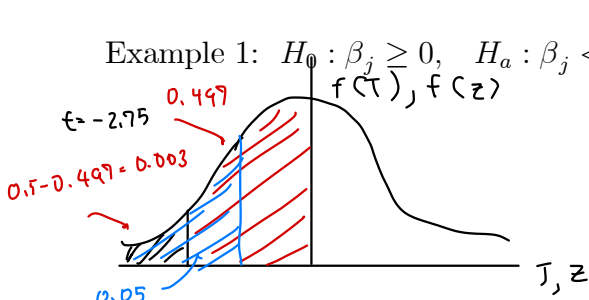
- p-value :  $P(|T| > |t|)$

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

$t =$  computed t-statistic

• p-value = probability that a random  $T$  value will be greater (in the  $| |$  term) than our  $t$  in the  $H_0$  test.

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



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

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

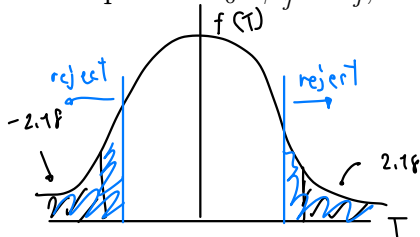
• From the z-table, the value -2.75 corresponds to area = 0.003

• 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:  $H_0 : \beta_j = a_j, H_a : \beta_j \neq a_j, d.f. = 18.$



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

• From the t-table, the value -2.18 corresponds to area = 0.02 to 0.05

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

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

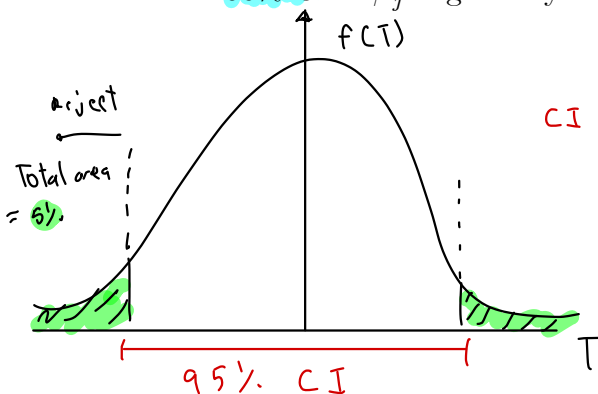
Yes, we reject  $H_0$  because the area is less than 0.05 or p-value < 0.05

## 6 Confidence Intervals (CI)

• Confidence Intervals for the POPULATION PARAMETER ( $\beta_j$ )

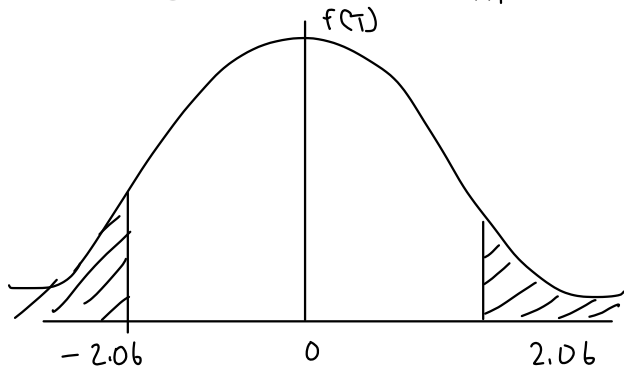
$\hookrightarrow$  The range of values that would capture the true  $\beta_j$  at a

• A 95% CI of  $\beta_j$  is given by 95% change.

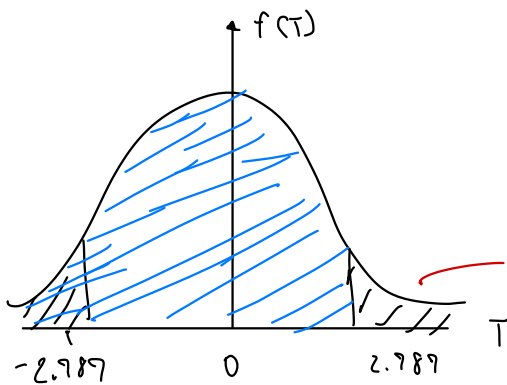


$$CI \Rightarrow \hat{\beta}_j \pm C \times s.e.(\hat{\beta}_j)$$

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

Example 1: 95% CI  $d.f. = 25$ value of 97.5<sup>th</sup> = percentile in the  $t_{25}$  distributionThe 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



What percentile? 99.5

or area =  $1 - 0.005 = 0.995$ 

$$\text{area} = \frac{1\%}{2} = \frac{0.01}{2} = 0.005$$

The 99% CI for  $\hat{\beta}_j = [\hat{\beta}_j - 2.987 \cdot \text{s.e.}(\hat{\beta}_j), \hat{\beta}_j + 2.987 \cdot \text{s.e.}(\hat{\beta}_j)]$

### 7 Testing Multiple Linear Restrictions: The F-test

Suppose the model is specified by

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

$H_0 : \beta_{i_2} = 0 \text{ and } \beta_{j_3} = 0 \rightarrow \text{Want to test if } x_1 \text{ and } x_2$   
 $H_a, H_1 : H_0 \text{ is not true} \quad \text{BOTH have no impact on } y.$

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:

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

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

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

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

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

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

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

(The last q  $\beta_s = 0$ )

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

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

(r) ur

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

This is always (+) because  $SSR_{ur} < SSR_r$ .  
 Everytime you add 1 more  $x$ , the model will be better explained.  
 dif. of the "ur" model.

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:

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

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

$q$  # of  $\beta$  that are set to "0".

$n - k - 1$  # of obs.     $k$  # of slope  $\beta$ .    Intercept

$\Rightarrow$  If we want to test the overall significance of the model

$H_0: \beta_1 = \beta_2 = \beta_3 = \dots = \beta_k = 0$  ,  $H_a = \text{otherwise}$

$F = \frac{R^2/k}{(1 - R^2)/(n - k - 1)}$   $R^2$  of the model =  $ur$   
the " $r$ " model has no  $x$  at all.

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

- $Y$  salary = season salary
- $x_1$  years = years in major leagues
- $x_2$  gamesyr = games per year in the league
- $x_3$  **bavg** = career batting average
- $x_4$  **hrunsyr** = homeruns per year
- $x_5$  **rbisyr** = runs batted in per year

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

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

$H_a: \text{otherwise}$

- the unrestricted model (ur) is defined by

our model

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

Source	SS	df	MS	
Model	308.989208	5	61.7978416	Number of obs = 353
Residual	183.186327	347	.527914487	F( 5, 347) = 117.06
Total	492.175535	352	1.39822595	Prob > F = 0.0000

R-squared = 0.6278  
Adj R-squared = 0.6224  
Root MSE = .72658

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

Intercept

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

- the restricted model (r) is defined by

```
. regress log_salary years gamesyr
```

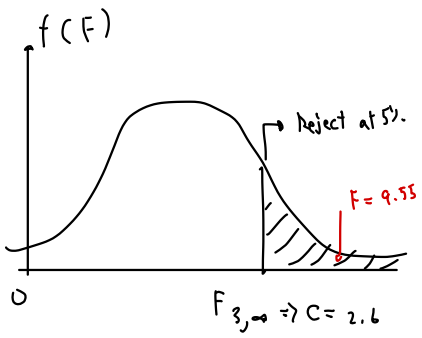
Source	SS	df	MS	
Model	293.864058	2	146.932029	Number of obs = 353
Residual	198.311477	350	.566604221	F( 2, 350) = 259.32
Total	492.175535	352	1.39822595	Prob > F = 0.0000

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

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



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

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  $H_0 : \beta_i = 0$

⇒ 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  $\beta_i$  is **statistically significant** at 5% level.  
(value of  $\beta_i \neq 0$ )

⇒ **When**  $t < 1.96$  we can say that  $\beta_i$  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

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 1X.

# 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

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

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 ?

$$1 \text{ kg} = 1000 \text{ g}$$

$$\widehat{bwght}_{kg} = \frac{\widehat{bwght}_g}{1000} = \frac{\hat{\beta}_0}{1000} + \frac{\hat{\beta}_1}{1000} cigs + \frac{\hat{\beta}_2}{1000} faminc$$

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

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

- What if we use  $faminc$  in USD (instead of 1000 USD)

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

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

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

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

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

$\hat{\beta}_2$  = " — " 1000 USD ↑ in income

- What if we use  $bwght$  in kg & income in THB

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

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

2 More on functional forms

$$\Delta y = y_1 - y_2$$

$$\Delta X = X_{11} - X_{12}$$

• Logarithmic Functional Form

usually means natural log

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

$$\bullet \beta_1 = \frac{d \log(y)}{d \log(x_1)} = \frac{\frac{1}{y} dy}{\frac{1}{x} dx_1} = \frac{\frac{1}{y} \Delta y}{\frac{1}{x} \Delta x_1} = \frac{100 \times \frac{1}{y} \Delta y}{100 \times \frac{1}{x_1} \Delta x_1} = \frac{\% \Delta y}{\% \Delta x}$$

with the log y & log x format, the coefficient is going to be the **elasticity!** ( $X_1$  elasticity of y)  
(price) (demand)

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

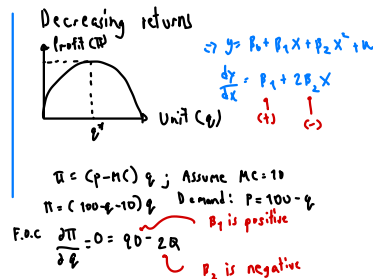
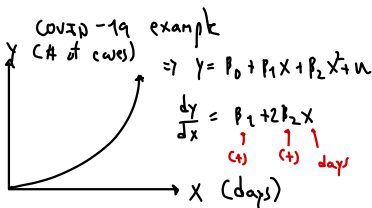
⇒ if we want the upper form to be % change,  $100 \beta_2 = \% \Delta y$  given that  $x_2$  increases by 1 unit.

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

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

• Models with Quadratics (Squares)

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



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

In the US or many other countries, students can apply to schools in the area without having to take any test. Therefore, the lower stratio, the better the school.

The estimation result is given by

regress lprice lnox dist rooms rooms\_sq stratio

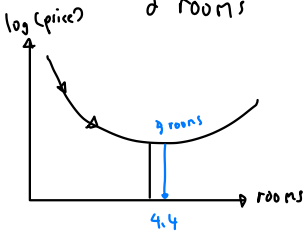
Source	SS	df	MS	Number of obs =	506
Model	51.4933152	5	10.298663	F( 5, 500) =	155.62
Residual	33.0889098	500	.06617782	Prob > F =	0.0000
				R-squared =	0.6088
				Adj R-squared =	0.6049
Total	84.582225	505	.167489554	Root MSE =	.25725

	lprice	lnox	dist	rooms	rooms_sq	stratio	_cons
log(price)	1price						
log(nox)		lnox					
		dist					
		rooms					
		rooms_sq					
		stratio					
		_cons					

|t| > 1.961 1 all < 0.05  
 ~ all variables are significant

Consider the effect of "room"

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



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

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

$$\text{rooms} = 4.4$$

Ans at 4.4 rooms or more  
 at ~ 5 rooms or more

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

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

total % Δ in price  
 when # rooms ↑ from 5 to 6 is 6.7 + 19.1% = 25.8%

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

$$= 100 \times 0.069 = 6.7\% \text{ increase}$$

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

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

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

Consider

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

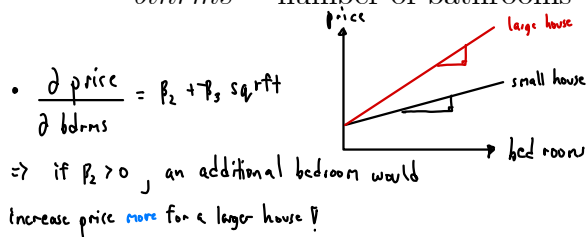
where

$price$  = housing price

$sqr\ ft$  = house size (square feet)

$bdrms$  = number of bedrooms

$bthrms$  = number of bathrooms



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

- Adding more regressors ALWAYS improve fit  $\leadsto R^2$  always  $\uparrow$
- But we lose the "degree of freedom"  
 (d.f. = free data point used to estimate the parameter)  
 $\hookrightarrow$  1 data point is sacrificed every time we estimate a parameter.
- Using  $R^2$  would not punish "having too many regressors"
- We use adjusted- $R^2$  or  $\bar{R}^2$  when we want to punish adding too many regressors.

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

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

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

Consider Model 1

If we have more  $k$ , d.f. =  $n-k-1 \downarrow$ ,

$SSR/(n-k-1) \uparrow$ ,  $\text{adj-}R^2 \downarrow$

$$\widehat{\text{salary}} = 830.63 + 0.0163\text{sales} + 19.63\text{roe}$$

$$= (223.90) \quad (0.0089) \quad (11.08)$$

$n = 209, R^2 = 0.029, \bar{R}^2 = 0.020$

Consider Model 2

$$\log(\widehat{\text{salary}}) = 4.36 + 0.2751 \log(\text{sales}) + 0.0179\text{roe}$$

$$= (0.29) \quad (0.033) \quad (0.004)$$

$n = 209, R^2 = 0.282, \bar{R}^2 = 0.275$

29.5% of variation in  $Y$  is explained. Therefore, this model is better!



# 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 describe them.

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

**TABLE 7.1**  
A Partial Listing of the Data in WAGE1.RAW

<i>person</i>	<i>wage</i>	<i>educ</i>	<i>exper</i>	<i>female</i>	<i>married</i>
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

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

where

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

In this case, the  $\delta_0$  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.

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

② Thus

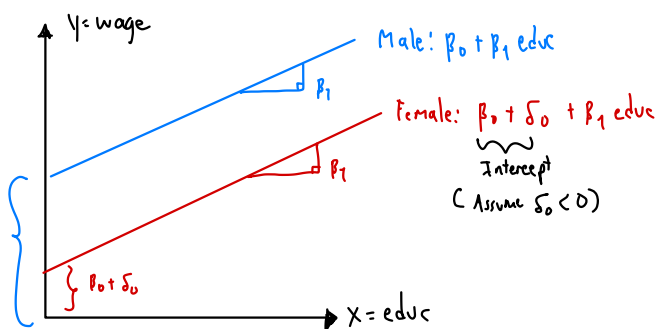
$$\textcircled{f} : E(\text{wage} \mid \text{female} = 1, \text{educ}) = \beta_0 + \delta_0(1) + \beta_1 \text{educ} = \beta_0 + \delta_0 + \beta_1 \text{educ}.$$

$$\textcircled{m} : E(\text{wage} \mid \text{female} = 0, \text{educ}) = \beta_0 + \delta_0(0) + \beta_1 \text{educ} = \beta_0 + \beta_1 \text{educ}$$

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

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

\* given the same value of educ (same education level),  $\delta_0$  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.

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

$\delta_0$  (x<sub>1</sub>)       $\beta_1$  (x<sub>2</sub>)       $\delta_1$  (x<sub>3</sub>)  
 } intercept \* 1

For example:

$$X_0 = X_1 + X_3$$

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

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

id.	female	male	X <sub>0</sub>
1	1	0	1
2	1	0	1
3	0	1	1
4	0	1	1
⋮	0	1	1
⋮	1	0	1
⋮	1	0	1
⋮	⋮	⋮	⋮
99	⋮	⋮	⋮

or If there are "n" categories, we omit "1" category to avoid multi collinearity.

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

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

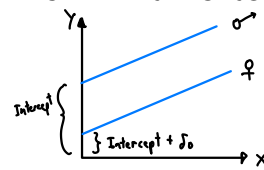
winter	spring	etc.
1 if winter	1 if spring	
0 otherwise	0 otherwise	

id	winter	spring	summer	fall	X <sub>0</sub>
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

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

$\left\{ \begin{array}{l} 1 \text{ if female} \\ 0 \text{ if male} \end{array} \right.$ 
  
 $\left\{ \begin{array}{l} 1 \text{ if male} \\ 0 \text{ if female} \end{array} \right.$



In this case, male

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

Source	SS	df	MS			
Model	54.3265253	4	13.5816313	Number of obs =	526	
Residual	94.0032262	521	.180428457	F( 4, 521) =	75.27	
Total	148.329751	525	.28253286	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

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 model

Consider a model which includes 2 dummy variables- *female* and *married*.

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

$\left\{ \begin{array}{l} 1 \text{ if female} \\ 0 \text{ if otherwise} \end{array} \right.$        $\left\{ \begin{array}{l} 1 \text{ if married} \\ 0 \text{ if otherwise} \end{array} \right.$

```
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

$\hat{\beta}$

2)  $\delta_1$  measures the impact of be married (marriage premium) But since  $|t| < 1.96$  or  $p > 0.05$ , we do not reject  $H_0$  of no impact.

Comments:

1)  $\delta_0$  measures the expected difference between female & male workers given the same marital status and other factors

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

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

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

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

	♀	♂
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)

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

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	
				R-squared =	0.4609	
				Adj R-squared =	0.4525	
Total	148.329751	525	.28253286	Root MSE =	.39329	

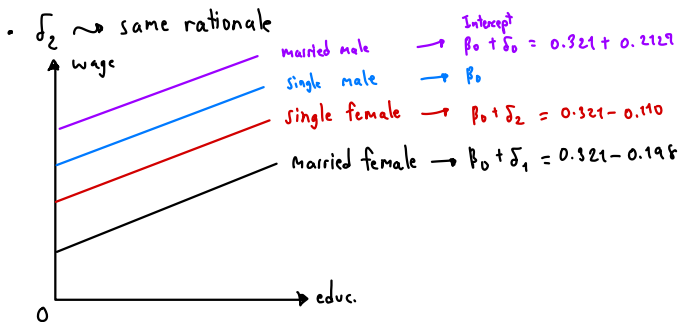
lwage	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
marrmale	.2126757	.0553572	3.84	0.000	.103923	.3214284
marrfem	-.1982676	.0578355	-3.43	0.001	-.311889	-.0846462
singfem	-.1103502	.0557421	-1.98	0.048	-.219859	-.0008414
educ	.0789103	.0066945	11.79	0.000	.0657585	.092062
exper	.0268006	.0052428	5.11	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

$\delta_0$   
 $\delta_1$   
 $\delta_2$

$\hat{\beta}$

Comments: This regression is not the same as the previous one. It uses "single male" as the base group. (The previous one we male & single as 2 base groups).

- $\delta_0$  measures the expected diff. in wage of married male as compared with single males, holding other factors constant.
- $\delta_1$  measures the expected diff. in "\_\_\_\_\_ " female "\_\_\_\_\_ ".



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

$$\log(\text{salary}) = \beta_0 + \delta_0 \text{top10} + \delta_1 r11\_25 + \delta_2 r26\_40 + \delta_3 r41\_60 + \beta_1 \text{LSAT} + \beta_2 \text{GPA} + \beta_3 \log(\text{libvol}) + \beta_4 \log(\text{cost}) + u.$$

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. age may explain. regress lsalary top10 r11\_25 r26\_40 r41\_60 LSAT GPA llibvol lcost  
 the model better if split into generations young (0-15), gen 2 (16-29) etc.

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

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

The baseline is ranking 61<sup>th</sup> and worse.

Comments: 1)  $\delta_0$  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 61<sup>th</sup> and worse.

2)  $\delta_1 \rightsquigarrow$  use the same rationale