

### 4 Testing Hypotheses about a Single Linear Combination of the Parameter

Consider

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

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

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

$exper$  = months in the workforce.

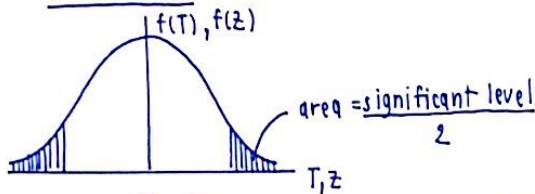
We want to test whether  $\beta_1 = \beta_2$ .  $\rightarrow$  Test if the returns from 1 more year of education at a Junior college is the same as that of the university

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

against

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

2-tailed test

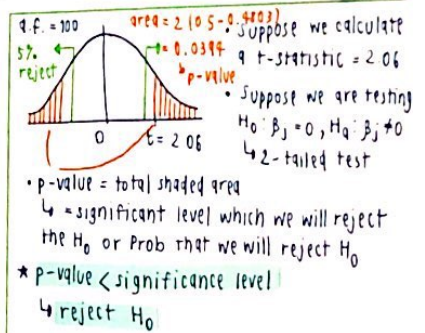


$$t = \frac{(\hat{\beta}_1 - \hat{\beta}_2) - 0}{\text{s.e.}(\hat{\beta}_1 - \hat{\beta}_2)}$$

$\rightarrow$  we compute this t-statistic and compare with the critical value

$$\begin{aligned} \text{where s.e.}(\hat{\beta}_1 - \hat{\beta}_2) &= \sqrt{\text{var}(\hat{\beta}_1 - \hat{\beta}_2)} \\ &= \sqrt{\text{var}(\hat{\beta}_1) + \text{var}(\hat{\beta}_2) - 2 \text{cov}(\hat{\beta}_1, \hat{\beta}_2)} \end{aligned}$$

not very straight forward to calculate  
 $\rightarrow$  we use a variable transformation trick

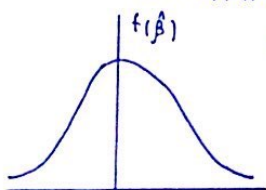


**Inference**  $\rightarrow$  Hypothesis testing about  $\beta$  (true parameter)

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

We want to test hypothesis about the true impact ( $\beta$ ) of each X variables (educ, experience) on the dependent variable (Y)

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

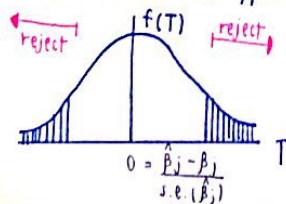


$\beta = a$  hypothesized value  
 ex.  $\beta = 0$  or  $\beta = 1$

- Test if  $\beta = \text{some number}$   
 e.g.  $\beta_j = 0 \rightarrow X_j$  has no impact on Y  
 $\beta_j = 1 \rightarrow 1$  unit in  $X_j$  correspond to 1 unit in Y

$\rightarrow$  t-test  

$$\frac{\hat{\beta}_j - \beta_j}{\text{s.e.}(\hat{\beta}_j)} \sim t_{d.f.}$$
 significant level = total area in the rejection region

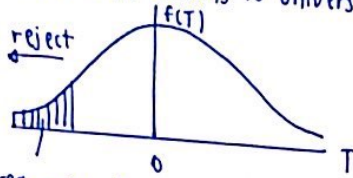


another possible hypothesis test (one-tailed alternative)

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

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

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



area = sig. level

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

Consider the multiple regression model, assume MLR 1-6 are satisfied.

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + u$$

$\rightarrow H_a$ : otherwise is true

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

① Write the t-statistic for testing  $H_0$

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

② Define  $\theta_1 = \hat{\beta}_1 - 3\hat{\beta}_2$

$$H_0: \theta_1 = 1 \quad t = \frac{\hat{\theta}_1 - 1}{s.e.(\hat{\theta}_1)} \rightarrow \text{we need our regression to have } \theta_1 \text{ in it.}$$

$H_a: \theta_1 \neq 1$  So, STATA or OLS estimation will automatically give  $\hat{\theta}_1$  and  $s.e. \hat{\theta}_1$

Now,  $\hat{\beta}_1 = \theta_1 + 3\hat{\beta}_2$  or  $\beta_1 = \theta_1 + 3\beta_2$  sub in the main regression and get

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

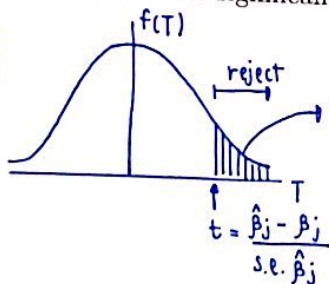
Now, the explanatory variables are going to be  $X_1, X_2 + 3X_1$  and  $X_3$

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

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

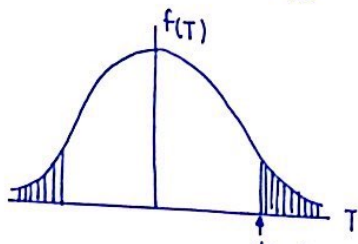
- What is the significance level given the computed t-statistics?

1-tailed



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

2-tailed



• p-value:  $P(|T| > |t|)$

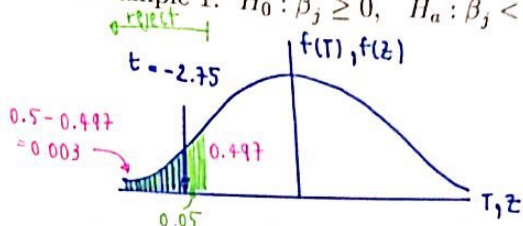
$$t = \text{computed } t = \frac{\hat{\beta}_j - \beta_j}{s.e. \hat{\beta}_j}$$

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

$t =$  computed t-statistic

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

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



$\rightarrow$  p-value = what should be the significant level, given the critical value of -2.75  
 $\hookrightarrow$  find the shaded area

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

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

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

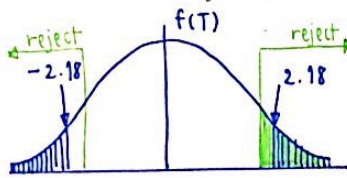
• Thus, p-value = 0.003

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

$\hookrightarrow$  use t-table



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

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

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

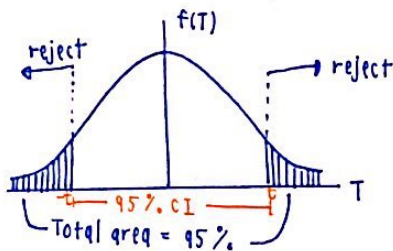
• Would we reject  $H_0$  if we use the significance level = 5%? **Yes, reject  $H_0$  because the area is less than 0.05 or p-value < 0.05**

### 6 Confidence Intervals (CI)

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

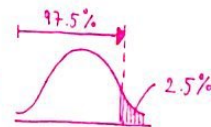
$\rightarrow$  The range of values that would capture the true  $\beta_j$  at a 95% chance

• A 95% CI of  $\beta_j$  is given by



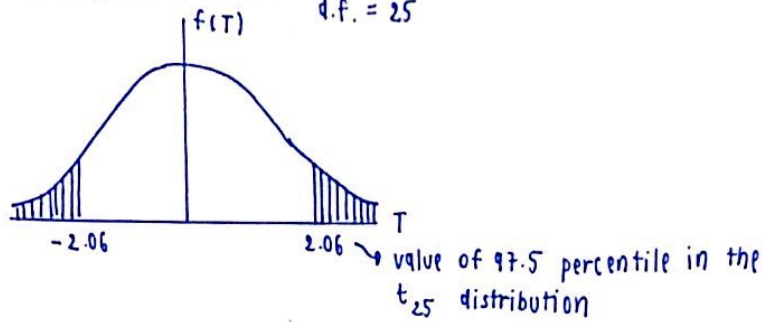
$$CI \Rightarrow \hat{\beta}_j \pm C \times s.e.(\hat{\beta}_j)$$

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



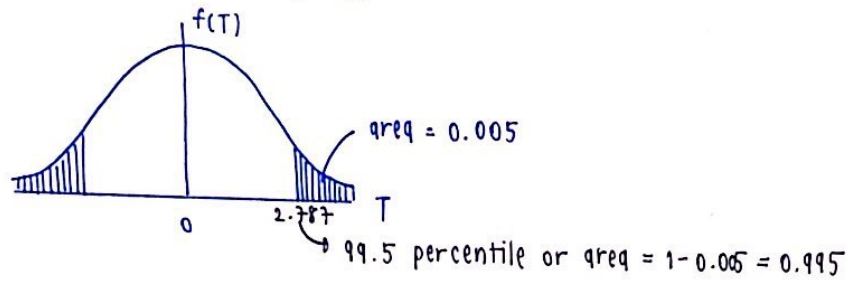
Example 1: 95% CI

d.f. = 25



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

**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 →  $\beta_1 = 0$  and  $\beta_2 = 0$  →  $\beta_1 = \beta_2 = 0$

$H_a$ : seat position matters →  $\beta_1 \neq 0$  and  $\beta_2 \neq 0$  +  $\beta_1 \neq 0$  or  $\beta_2 \neq 0$  → at least one of the  $\beta_1, \beta_2$  not equal to 0

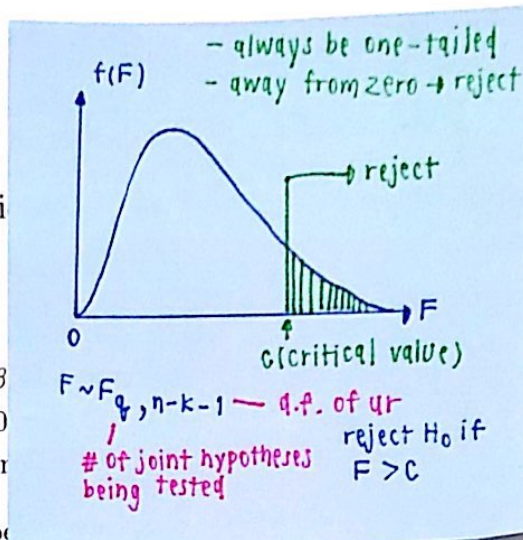
## 7 Testing Multiple Linear Restriction

Suppose the model is specified by

$$y = \beta_0 + \beta_1 x_1 + \dots + \beta_k x_k + u$$

$$H_0 : \beta_2 = 0$$

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



We can use the F-test to test this type of model

1. Our full model is called the "unrestricted" model (ur). Suppose it can be expressed as:
 
$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + u \rightarrow \text{is true} \rightarrow \text{Reject } H_0$$

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

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

$$Y = \beta_0 + \beta_1 x_1 + u \text{ is true} \rightarrow \text{donot reject } H_0$$

Suppose there are  $q$  number of  $\beta$  that we should 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)

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

$$Y = \underbrace{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_{k-q} x_{k-q}}_{(r)} + \underbrace{\beta_{k-q+1} x_{k-q+1} + \beta_{k-q+2} x_{k-q+2} + \dots + \beta_k x_k + u}_{ur}$$

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

This is always  $\oplus$   
 b/c  $SSR_{ur} < SSR_r$   
 (Every time you add one more  $X$ , the model will be better explained)  
 d.f. of the ur model

• So, if every time you add 1 more  $X$  variable, the  $SSR \downarrow$  and  $R^2 \uparrow$ , why don't we just keep the additional  $X$  in the model?

→ Because every time we add 1 more  $X$ ,  $Var(\hat{\beta}_s)$  will increase, making the prediction of  $\beta$  less precise. So, we only keep the additional  $X$ , if it/they can improve the model enough.

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

3. Some useful facts

①  $R^2_{ur} > R^2_r$  because any additional X would increase  $R^2$  (improve fit)  
 $SSR_{ur} < SSR_r$

② By including more X, the model is certainly better explained. However, we would like to reject  $H_0$  if the inclusion of extra variables does not improve the model enough

4. Other ways to calculate the F-statistics:

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

We have  $F = \frac{R^2_{ur} - R^2_r}{\frac{1 - R^2_{ur}}{n - k - 1}}$

$\underbrace{\hspace{10em}}_{\text{# of } \beta \text{ that are set to 0}}$ 
 $\underbrace{\hspace{10em}}_{\text{intercept}}$   
 $\underbrace{\hspace{10em}}_{\text{# of obs}}$ 
 $\underbrace{\hspace{10em}}_{\text{# of slope } \beta}$

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

$F = \frac{R^2_{ur}}{(1 - R^2_{ur}) / (n - k - 1)}$  →  $R^2$  of ur → r model has no X at all

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

- |               |                           |                                |
|---------------|---------------------------|--------------------------------|
| $r$ {<br>ur { | $Y$ salary                | = season salary                |
|               | years                     | = years in major leagues       |
|               | gamesyr                   | = games per year in the league |
|               | bavg                      | = career batting average       |
|               | hrunsyr                   | = homeruns per year            |
| rbisyr        | = runs batted in per year |                                |

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

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

$H_a$ : otherwise is true

- the unrestricted model (ur) is defined by

"ur model"

. regress log\_salary years gamesyr bavg hrunsyr rbisyr

Source	SS	df	MS	
Model	308.989208	5	61.7978416	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

have  
may not be individually  
impact but must be  
three factors combined  
to create impact on  
salary

"Use F-test"

to find whether  
three factors have  
jointly impact  
or not

• 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

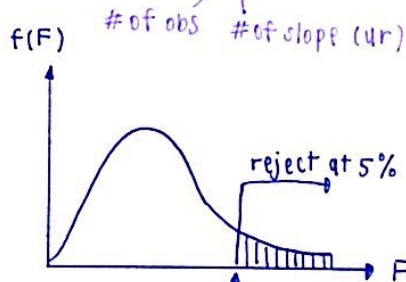
log_salary	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
years	.071318	.012505	5.70	0.000	.0467236	.0959124
gamesyr	.0201745	.0013429	15.02	0.000	.0175334	.0228156
_cons	11.2238	.108312	103.62	0.000	11.01078	11.43683

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

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

Now, our  $H_0$  and  $H_a$  becomes

$$F = \frac{\left( \frac{SSR_r - SSR_{ur}}{q} \right)}{\left( \frac{SSR_{ur}}{n-k-1} \right)} = \frac{\frac{198.311 - 183.186}{3}}{\frac{183.186}{353-5-1}} \approx 9.55$$



$C = 2.6$   
F table page 966  
 $F_{3, \infty}$   
Pr = 0.05  
df  $N_1 = 3$   
df  $N_2 = \infty$

allow 5% mistake  
Let's use 5% level of significance  
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% significant 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

sales ←

other company performance {

CEO characteristics {

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

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

↳ in gram

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 kg = 1000 g

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

$$\widehat{bwght}_{kg} = \hat{\alpha}_0 + \hat{\alpha}_1 cigs + \hat{\alpha}_2 faminc$$

$gen\ bwght\_g = bwght \times 28.35$   
 $gen\ bwght\_kg = bwght\_g / 1000$  } generate new variable

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

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

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

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

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

in other words  $\hat{\theta}_2$  = impact of 1 USD ↑ in income       $gen\ faminc\_usd = faminc \times 1000$   
 $\hat{\beta}_2$  = impact of 1000 USD ↑ in income

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

$$bwght_{kg} = \frac{\hat{\beta}_0}{1000} + \frac{\hat{\beta}_1}{1000} cigs + \frac{\hat{\beta}_2}{30,000} faminc_{THB}$$

← This value is going to be 30,000 times larger than  $faminc$  (assume 30 THB = 1 USD)

2 More on functional forms

- Logarithmic Functional Form

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

usually means natural log in econometrics

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

$$\Delta y = y_1 - y_2, \Delta x_1 = x_{11} - x_{12}$$

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

with the log y and log x format, the coefficient is going to be the elasticity ( $x_1$  elasticity of y)  
 (price) (demand)

$$\beta_2 = \frac{d \log(y)}{d x_2} = \frac{\frac{1}{y} dy}{dx_2} = \frac{\frac{1}{y} \Delta y}{\Delta x_2}$$

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

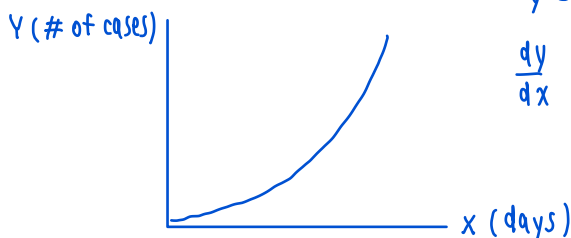
$$100 \beta_2 = \frac{100 \frac{1}{y} \Delta y}{\Delta x_2}$$

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

- Models with Quadratics → squares

capture increasing/decreasing marginal effects (slope of the relationship between X & Y is not constant)

COVID-19 example

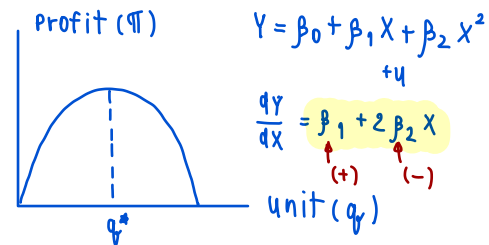


$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + u$$

$$\frac{dY}{dX} = \beta_1 + 2\beta_2 X$$

↑ ↑ ↑  
(+) (+) days

Decreasing returns



$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + u$$

$$\frac{dY}{dX} = \beta_1 + 2\beta_2 X$$

↑ ↑  
(+) (-)

$$\pi = (P - MC)q \leftarrow \text{Assume } mc = 10 \text{ Demand: } P = 100 - q$$

$$\pi = (100 - q - 10)q \leftarrow \beta_1 \text{ is positive?}$$

$$FOC \quad \frac{\partial \pi}{\partial q} = 0 = 90 - 2q \leftarrow \beta_2 \text{ is negative}$$

● same result

Example : Effects of Pullution 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 US or many countries, students can apply to schools in the area without having to take any test. So, the lower stratio, the better the school.

regress lprice lnox dist rooms rooms\_sq stratio

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

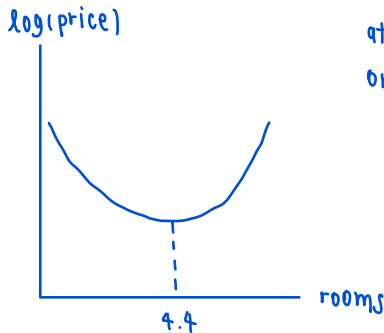
	lprice	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
lnox	$\beta_1$	-.9767545	.0995938	-9.81	0.000	[-1.172429, -.7810806]
dist	$\beta_2$	-.0321972	.0094013	-3.42	0.001	[-.050668, -.0137264]
rooms	$\beta_3$	-.5528032	.1612965	-3.43	0.001	[-.8697056, -.2359007]
rooms_sq	$\beta_4$	.0624697	.0124867	5.00	0.000	[.0379368, .0870025]
stratio	$\beta_5$	-.0486667	.0058131	-8.37	0.000	[-.0600879, -.0372455]
_cons		13.59154	.5650901	24.05	0.000	[12.4813, 14.70178]

log(price)  
log(nox)

$|t| > 1.96$  for all variables  $\rightarrow$  all  $< 0.05$   
 $\hookrightarrow$  All variables are significant

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 1 additional room has a positive impact on log(price)??

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

$$\text{rooms} = 4.4$$

At 4.4 rooms or more

$\hookrightarrow$  round up to 5 rooms or more

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

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

$$\frac{100 \frac{1}{\text{price}} d \text{price}}{d \text{rooms}} = 100(-0.553 + 2(0.062)(5)) = 100(0.067) = 6.7\% \text{ increase}$$

What about %  $\Delta$  in price when # rooms increases from 5 to 7?

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

$$\text{total } \% \Delta \text{ in price when \# rooms } \uparrow \text{ from 5 to 7 is } 6.7 + 19.1 = 25.8\%$$

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

Consider

$$price = \beta_0 + \beta_1 \underset{X_1}{sqr\ ft} + \beta_2 \underset{X_2}{bdrms} + \beta_3 \overbrace{sqr\ ft \times bdrms}^{X_3} + \beta_4 \underset{X_4}{bthrms} + u$$

where

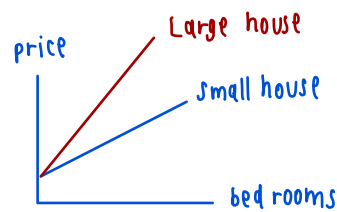
$price$  = housing price

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

$bdrms$  = number of bedrooms

$bthrms$  = number of bathrooms

$$\frac{dprice}{dbdrms} = \beta_2 + \beta_3 \text{ sqr ft}$$



→ if  $\beta_3 > 0$  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 increase
- $\rightarrow$  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.
- $\rightarrow$  Using  $R^2$  would not punish "having too many regressors"
- $\rightarrow$  We use adjusted- $R^2$  or  $\bar{R}^2$  when we want to punish adding too many regressors

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

$$\text{adj. } R^2 = 1 - \frac{SSR/(n-k-1)}{SST/(n-1)} \rightarrow \text{If we have more } k, \text{ d.f.} = n-k-1 \downarrow, \text{ } SSR/(n-k-1) \uparrow, \text{ adj. } R^2 \downarrow$$

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

Consider Model 1

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

Consider Model 2

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



# 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 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