Athicha Korkiertsatean
6104640062

4 Testing Hypotheses about a Single Linear Combination of the Parameter

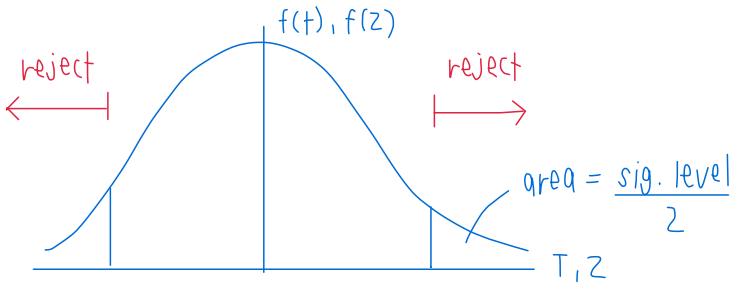
Consider

maybe non-linear
so, take log to be more linear

$$\log(wage) = \beta_0 + \beta_1 jc + \beta_2 univ + \beta_3 \exp er + u$$

where jc = number of years attending a two-year college (อาชีวะ)
$univ$ = number of years at a four-year college
$\exp er$ = months in the workforce.
We want to test whether $\beta_1 = \beta_2$. If the return from 1 more year of education at a JC is the same as that of the university.

$H_0 : \beta_1 = \beta_2 \rightarrow H_0 : \beta_1 - \beta_2 = 0$
$H_a : \beta_1 \neq \beta_2 \rightarrow H_a : \beta_1 - \beta_2 \neq 0$
} two-tailed test



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

Then compute t and compare w/ the critical value.

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

(use variable transformation)

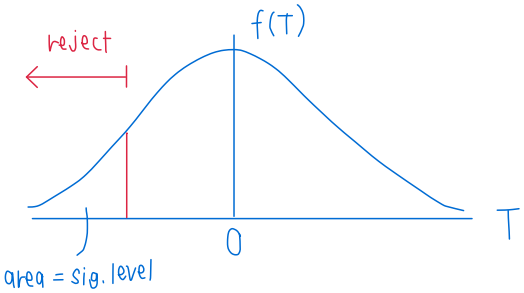
$$= \sqrt{var(\hat{\beta}_1) + var(\hat{\beta}_2) + 2cov(\hat{\beta}_1, \hat{\beta}_2)}$$

another possible hypothesis test (one-tailed alternative)

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

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

It's assume that β_1 would not be more than β_2 (return to a 2-yr will never be more than university)

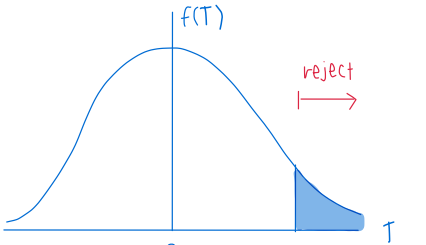


$$t = \frac{(\hat{\beta}_1 - \hat{\beta}_2) - 0}{s.e.(\hat{\beta}_1 - \hat{\beta}_2)}$$

go to extra note.

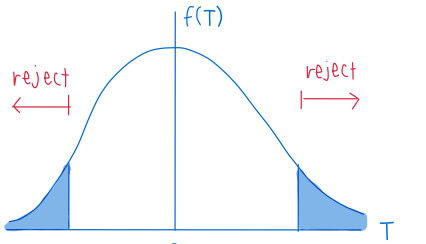
5 Computing p-Values for t-Tests

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



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

The shaded area in the rejection region is the p-value.



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

Total area from 2-sides is the p-value

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

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

t = computed t-statistic.

p-value = prob. 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, assume that MLR 1-6 are satisfied.

$$Y = \beta_0 + \beta_1 X1 + \beta_2 X2 + \beta_3 X3 + u$$

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

H_a: otherwise

1st write t-statistic for testing H_0 : $t = \dfrac{(\hat{\beta}_1 - 3\hat{\beta}_2) - 1}{s.e.(\hat{\beta}_1 - 3\hat{\beta}_2)}$

2nd Define $\hat{\theta}_1 = \hat{\beta}_1 - 3\hat{\beta}_2 \longrightarrow$ $H_0: \theta_1 = 1$, $H_a: \theta_1 \neq 1$ $\quad t = \dfrac{\hat{\theta}_1 - 1}{s.e.(\hat{\theta}_1)} \longrightarrow$ we need our regression to have θ_1 in it.

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

Now, if rearrange $\theta_1 = \beta_1 - 3\beta_2$, we have, $\hat{\beta}_1 = \hat{\theta}_1 + 3\hat{\beta}_2$

$\beta_1 = \theta_1 + 3\beta_2$

Then, sub. in the main regression and get

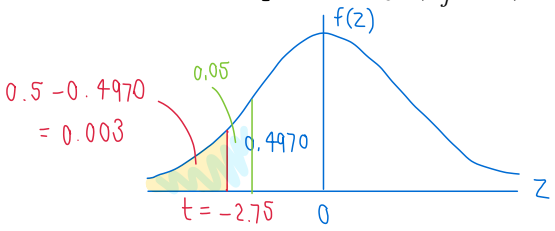
$$Y = \beta_0 + (\theta_1 + 3\beta_2)X1 + \beta_2 X2 + \beta_3 X3 + u$$
$$Y = \beta_0 + \theta_1 X1 + 3\beta_2 X1 + \beta_2 X2 + \beta_3 X3 + u$$
$$Y = \beta_0 + \theta_1 X1 + \beta_2 (X2 + 3X1) + \beta_3 X3 + u$$

Now the explanatory variables are X1, X2+3X1 and X3

so, we can calculate

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

Example 1: $H_0 : \beta_j \geq 0, \quad H_a : \beta_j < 0, \quad$ d.f.= 140. → z-table



p-value = what should be the sig. level given the critical value of -2.75

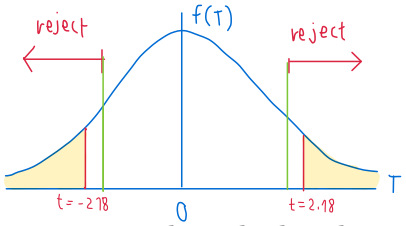
find the shaded area
= 0.003

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

- 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

reject Ho if p-value < sig. level

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



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

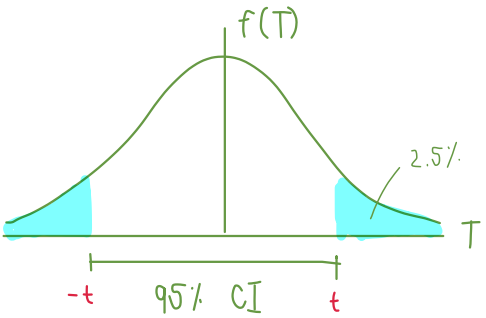
yes, reject Ho because the area is less than 0.05 or p-value < 0.05

6 Confidence Intervals (CI)

- Confidence Intervales for the POPULATION PARAMETER (β_j)

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

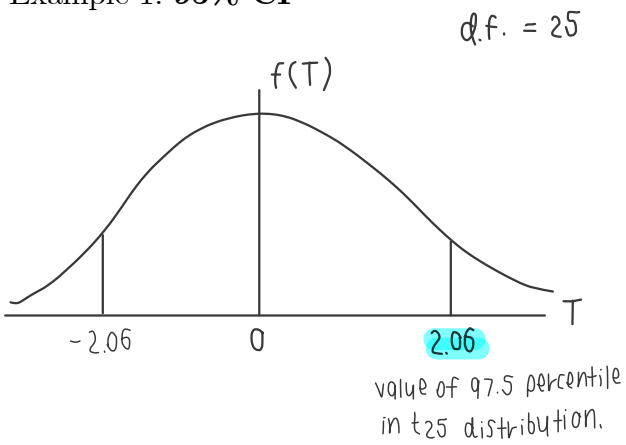
- A 95% CI of β_j is given by



CI → $\hat{\beta}j \pm c \cdot s.e.(\hat{\beta}j)$

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

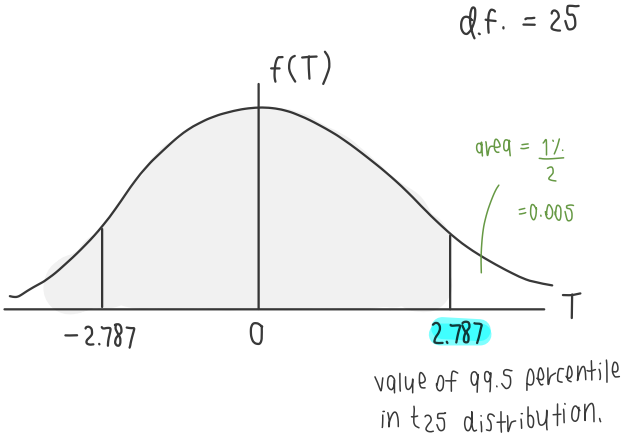
Example 1: **95% CI**



The 95% CI for $\hat{\beta}_j$

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

Example 2: **99% CI**



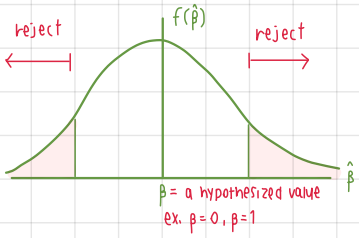
The 99% CI for $\hat{\beta}_j$

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

Inference → Hypothesis testing about "β", the true parameter

wage = β0 + β1edu + β2exp + ... + u

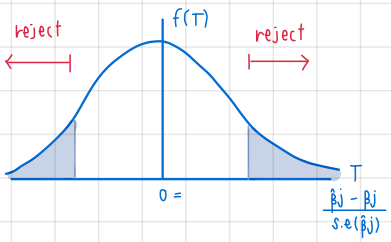
we want to test the hypothesis about true impact (β) of each X variables (edu, exp) on the dependent variable (Y)
BUT, we don't know what the true β are, so, we use β̂1 (estimator) and s.e.(β̂) to test the hypothesis.



1 Test if β = same number

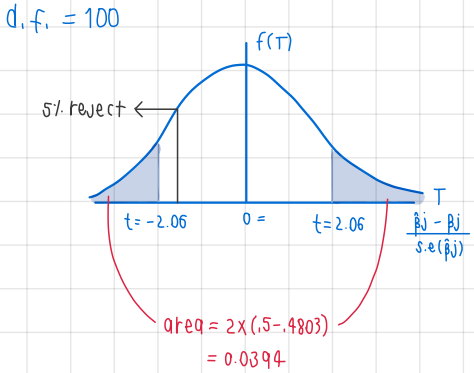
e.g. βj = 0, Xj has no impact on Y.

βj = 1, 1 unit increase in Xj corresponded to 1 unit increase in Y.



$$t\text{-test}: \frac{\hat{\beta}_j - \beta_j}{s.e.\hat{\beta}_j} \sim t_{d.f.}$$

Significant level = total area in the rejection region



- suppose, we calculate a t-statistic = $\frac{\hat{\beta}_j - \beta_j}{s.e.\hat{\beta}_j} = 5.78$
- suppose, we are testing Ho: $\hat{\beta}_j = 0$, Ha: $\hat{\beta}_j \neq 0$
- p-value = total shaded area
  = significant level or prob which we'll reject Ho.

If p-value < significant level ⟶ Reject Ho.

Ftest

test the sig. of a group of hypotheses (multiple hyp.)

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

Ho : seat position doesn't have impact on GPA

$\beta_1 = 0$, $\beta_2 = 0$; $\beta_1 = \beta_2 = 0$

Ha : seat position matters

$\beta_1 \neq 0$, $\beta_2 \neq 0$

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

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

} at least 1 of $\beta_1, \beta_2 \neq 0$

7 Testing Multiple Linear Restrictions: The F-test

Suppose the model is specified by

$$\begin{aligned} y &= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + u \\ H_0 &: \beta_2 = 0 \text{ and } \beta_3 = 0 \quad \text{if both X2 \& X3 don't have impact on Y} \\ H_1 &: H_0 \text{ is not true} \quad \text{\# otherwise} \end{aligned}$$

We can use the F-test to test this type of "multiple hypotheses".

1. Our full model is called the "unrestricted" model (ur). (big model : contain all variables) Suppose it can be expressed as:

   $y = \beta 0 + \beta 1 X1 + \beta 2 X2 + \beta 3 X3 + u$ is true (reject H0)

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

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

   small model

   $y = \beta 0 + \beta 1 X1 + u$ is true (not reject H0)

suppose there're q no. of β that we would like to perform a joint test of = 0

e.g. q = 2

$y = \beta 0 + \beta 1 X1 + \beta 2 X2 + \cdots + \beta_{K-q} X_{K-q}$

H0 : $\beta_{K-q+1} = \beta_{K-q+2} = \beta_K = 0$, last q βs = 0

Ha : H0 is not true.

$y = \beta 0 + \beta 1 X1 + \beta 2 X2 + \cdots + \beta_{K-q} X_{K-q} + \beta_{K-q+1} + \beta_{K-q+2} + \ldots + \beta_K X$

r

K + 1

ur

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

Always positive as $SSR_{ur} < SSR_r$ b/c more explanatory var., less residual. } SSR ↓, r^2 improve

so, everytime you add 1 more X, the model will be better explained.

So, why don't keep adding X variables AMAP

B/C, everytime we add 1 more X, var($\hat{\beta}$s) will increase making the prediction of β less precise. So, we only keep the addition of Xs if it/they can improve the model enough

can significantly ↓ SSR and ↑ r^2

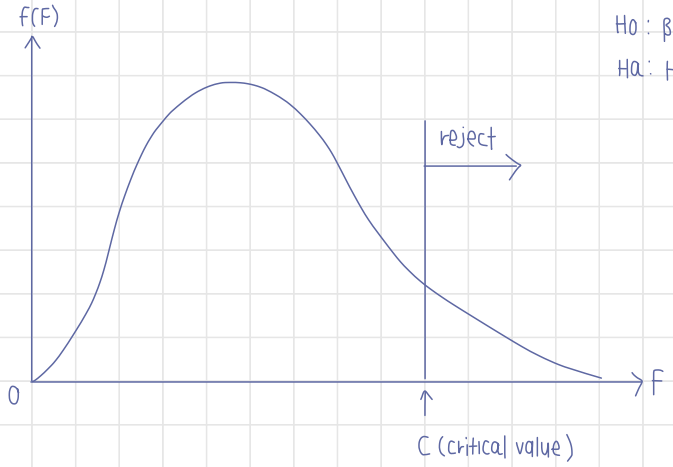
d.f. of ur model

SSR_r = some of sq. residual in restricted model

SSR_{ur} = some of sq. residual in unrestricted model

q = no. of explanatory variables we want to test = 0

Ftest always have 1 tail



if F statistic is close to zero then we fail to reject Ho

if F statistic is far from zero then we reject Ho

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

Ha : Ho not true

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

q: no. of joint hyp being tested

we reject Ho of jointly no effect if $F > C$

3. Some useful facts

1. $R^2_{ur} > R^2_r$ # better b/c additional X will increase R^2 (improve fit) $\longrightarrow$ $SSR_{ur} < SSR_r$

2. Add more X the model is certainly explained. However we would like to reject H0 if the inclusion of extra variables doesn't improve the model enough.

4. Other ways to calculate the F-statistics: in terms of R^2

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

we have $F \equiv \dfrac{\dfrac{(R^2_r - R^2_{ur})}{q \text{ — no. of } \beta \text{ that set to zero}}}{\dfrac{1-R^2_{ur}}{n-k-1}}$

obs. slope coefficient intercept

wn. diag test if $\beta = 0$
↳ don't know that x putting in the regression is making sense.
↳ test whether any β are non zero.

If we want to test the overall sig. of the model

H0 : $\beta_1 = \beta_2 = \beta_3 = \ldots = \beta_K = 0$

Ha : otherwise

$\longrightarrow$ $F \equiv \dfrac{\dfrac{R^2}{K}}{(1-R^2)/(n-K-1)}$

R of ur that 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

r { ur { $years$ = 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.

H0 : $\beta_{bavg} = \beta_{hrunsyr} = \beta_{rbisyr} = 0$

Ha : otherwise is true

- the unrestricted model (ur) is defined by

y

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

Source	SS	df	MS
Model	308.989208	5	61.7978416
Residual	183.186327	347	.527914487
Total	492.175535	352	1.39822595

SSE / SSR / SST

Number of obs	353
F(5, 347)	117.06
Prob > F	0.0000
R-squared	0.6278
Adj R-squared	0.6224
Root MSE	.72658

more explanatory variables = more r^2

use in multiple regression

lower S.e.

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

q

$t < 1.96$

when considering each of the performance X one-by-one. None of them has a significant impact at 5%. However, there is a small impact.

But when performing a F-test, performances have joint impact.

- the restricted model (r) is defined by

r model

```
. regress log_salary years gamesyr
```

Source	SS	df	MS
Model	293.864058	2	146.932029
Residual	198.311477	350	.566604221
Total	492.175535	352	1.39822595

SSE / SSR / SST

Number of obs	353
F(2, 350)	259.32
Prob > F	0.0000
R-squared	0.5971
Adj R-squared	0.5948
Root MSE	.75273

higher S.e.

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

Now, our H_0 and H_a becomes

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

no. of joint hyp. we are testing

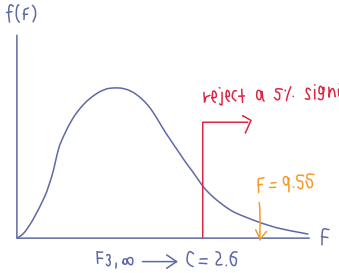
$$\equiv \frac{(198.311 - 183.186)/3}{183.186/(353-5-1)}$$

$$\equiv 9.55$$

HW.

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

$$\equiv ?$$



Since our F is 9.55 > 2.6, we reject Ho at 5% significant level.

And conclude that performances have joint effects on salary.

8 How the Hypothesis Testing is done in Practice

1. Check the values of $t-statistic$ reported by the statistical software (i.e. STATA, SPSS, SAS)

   $\Rightarrow$ These $t-statistics$ are to test $H_0 : \beta_i = 0$

   $\Rightarrow$ If the d.f. > 30, then when $t > 1.96$, we can reject H_0 (z-table; w/ 5% sig.)

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

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

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

   $\Rightarrow$ After we drop x_i, we estimate the new regression function and obtain a new set of $\hat{\beta}$.

2. We can also perform other hypothesis testings of interest.

   e.g. $H_0 : \beta_i = \beta_j$

   or $H_0 : \beta_i = 5$ etc.

   or perform an F-test for testing multiple linear restrictions

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

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

simple regression w/ 1X variable

Multiple Regression Analysis : Further Issues

1 Data scaling on OLS statistics

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

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

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

- What if we use bweght in kilograms ; 1 kg = 1000 g

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

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

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

$$bweght_g = \hat{\beta}_0 + \hat{\beta}_1 cigs + \frac{\hat{\beta}_2}{1000} faminc_{USD}$$ ⟵ will be 1000 times larger than actual faminc.

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

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

$\hat{\beta}_2$ = impact of 1000 USD ↑ in income.

- What if we use bweght in kg and income in Thb

$$bweght_{kg} = \hat{\beta}_0 + \frac{\hat{\beta}_1}{1000} cigs + \frac{\frac{\hat{\beta}_2}{1000}}{30,000} faminc_{THB}$$ ⟵ will be 30,000 times more than actual faminc.

2 More on functional forms

- Logarithmic Functional Form

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

(natural log)

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

($\Delta y = y_1 - y_2$; $\Delta x_1 = x_{11} - x_{12}$)

with the log(Y) and log(X1), the coefficient is going to be the elasticity. (X1 elasticity of Y) (price) (demand)

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

we want this to be %Δ, then :

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

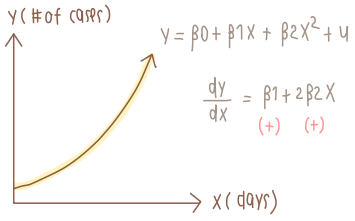
$$100\beta_2 = \frac{\%\Delta y}{\Delta x_2}$$

∴ so, $100\beta_2 = \%\Delta y$ given that x_2 increases by 1 unit

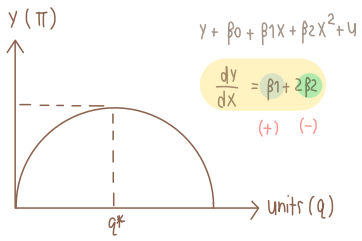
- Models with Quadratics (squares)

capture inc/dec marginal effects. (slope of the relationship btw X,Y is not constant.

COVID-19



Decreasing returns.



Assume that MC = 10, P = 100 − Q

$\pi = (P - MC)Q$

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

FOC : $\frac{d\pi}{dQ} = 0 = 90 - 2Q$

Example : Effects of Pullution on Housing Prices

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

where
price = housing price
nox = level of pollution
dist = distance from downtown
rooms = number of rooms
stratio = average student per teacher ratio (in US or many other country, student can apply to school in the area w/o any competition. 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
Model	51.4933152	5	10.298663
Residual	33.0889098	500	.06617782
Total	84.582225	505	.167489554

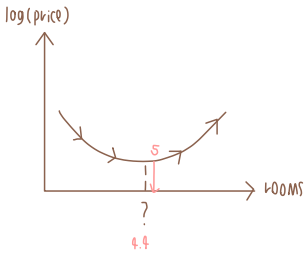
Number of obs = 506
F(5, 500) = 155.62
Prob > F = 0.0000
R-squared = 0.6088
Adj R-squared = 0.6049
Root MSE = .25725

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

$|t| > 1.96$ all are significant
$p < 0.05$

Consider the effect of "room"

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



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

$$0 = -0.553 + 2(0.062) \cdot rooms$$
$$rooms = 4.4 \leftarrow \text{round up}$$

∴ At 5 rooms or more

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

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

$$100 \cdot \frac{\frac{1}{price}\, dprice}{drooms} = 100\left[-0.553 + 2(0.062) \cdot 5\right]$$
$$= 100 \cdot 0.067$$
$$= 6.7\% \text{ increase}$$

What abt % change in price when #rooms increase from 5 to 7

$$\%\Delta price = 100\left[-0.553 + 2(0.062) \cdot 6\right]$$
$$= 100\,(0.191)$$
$$= 19.1\% \text{ increase}$$

∴ Total %Δprice when #rooms increase from 5 to 7 is $6.7 + 19.1 = 25.8\%$

3 Models with Interaction Terms

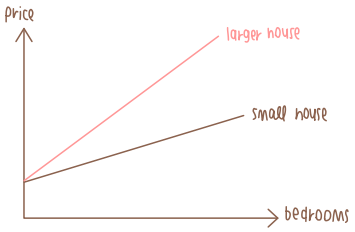
use when the impact of one variable depends on the value (level) of another variable.

Consider

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

X1 X2 X3 = X1X2 X4

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



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

if $\beta_2 > 0$ then, an additional bedroom would inc. price more for a larger house.

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

- Adding more regressors ALWAYS improve fit R^2 always ↑

  - But we lose the "degree of freedom"
    - free data point used to estimate the parameter
    - 1 data point is sacrified everytime we estimate a parameter.

  - using R^2 would not punish having to many regressors

  - use adjusted R^2 or $\bar{R}^2$ when we want to punish adding too many regressors.

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

Adjusted R^2 ⟶ $\bar{R}^2 = 1 - \frac{SSR/(n-k-1)}{SST/(n-1)}$

If we have 1 more k, the d.f. ↓.
making $SSR/(n-k-1)$ ↑
result in the dec in $\bar{R}^2$

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

Consider Model 1

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

Consider Model 2

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

non-restricted model.

27.5% of the variation in y is explained. So, this model is better.

Multiple Regression Analysis with Qualitative Information:

dummy variables

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 arbitarily 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

$$wage = \beta_0 + \delta_0 female + \beta_1 educ + u.$$

where

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

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.

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

$= \beta_0 + \delta_0 female + \beta_1 educ + \underbrace{E(u \mid female, educ)}_{0}$ Assume that MLR 1-4 hold

$= \beta_0 + \delta_0 female + \beta_1 educ$

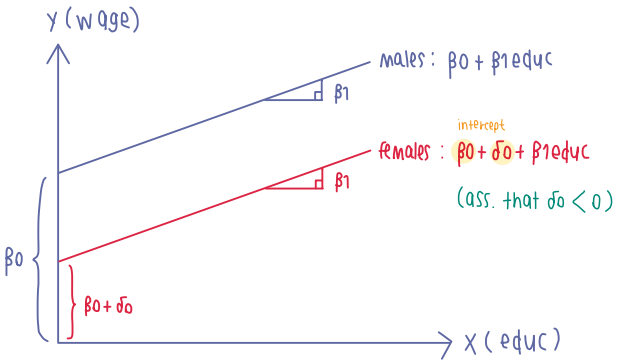
2. Thus

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

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

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

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



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

For example:

$$wage = \beta_0 x_0 + \delta_0 \underset{x_1}{female} + \beta_1 educ + \delta_1 \underset{x_3}{male} + u$$

x_0 = intercept · 1

$x_0 = x_1 + x_3$

$$\begin{aligned} 1 &= female + male \\ female &= male + 1 \end{aligned}$$

or

$$\begin{aligned} 1 &= winter + spring + summer + fall \\ winter &= 1 - spring - summer - fall \end{aligned}$$

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

If there are n categories we omit 1 categories to avoid multicollinearity

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

1 if female, 0 if male — 1 if male, 0 if female

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

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

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

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*.

1 if female, 0 if otherwise — 1 if married, 0 if otherwise

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

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

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

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

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

($\hat{\beta}$: educ, exper, expersq, tenure, tenursq)

|t| < 1.96

Comments:

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

$$\frac{d\log(wage)}{dfemale} = \frac{\frac{1}{wage}dwage}{dfemale} = -0.29$$

$$= 100\,\frac{\frac{1}{wage}dwage}{dfemale} = -0.29\,(100)$$

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

∴ female workers are expected to earn less than male by 29.02%, holding other factors to be constant.

2. δ_1 measure the impact of being married (marriage premium)

But since |t| < 1.96 or p > 0.05, we don't reject H_0 of no impact.

can't conclude that have impact

	♀	♂
marr	marrfem	marrmale
sing	singfem	singmale → basecase

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

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

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

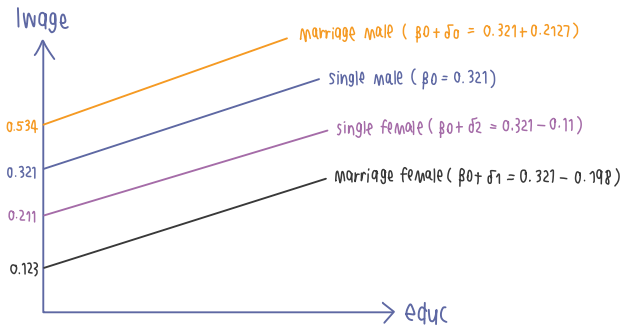
Source	SS	df	MS		
Model	68.3617623	8	8.54522029	Number of obs	= 526
Residual	79.9679891	517	.154676961	F(8, 517)	= 55.25
				Prob > F	= 0.0000
				R-squared	= 0.4609
				Adj R-squared	= 0.4525
Total	148.329751	525	.28253286	Root MSE	= .39329

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

($\hat{\beta}$: educ, exper, expersq, tenure, tenursq)

Comments: Not the same as the prev. one. It use "single male" as the base group
But the prev. one use male & single as 2 base groups.

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



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

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

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 case the range of value serve as a better explanatory variable than the value itself.

eg. age may explain the model better if split into generations.

young 0-15
genz 16-29
etc.

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

Source	SS	df	MS	Number of obs =	136
Model	9.16538532	8	1.14567316	F(8, 127) =	120.15
Residual	1.2109665	127	.009535169	Prob > F =	0.0000
				R-squared =	0.8833
Total	10.3763518	135	.076861865	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 61th and worse

Comments:

δ_0 measures the diff. in expected log(salary) of a law-school graduate from a top 10 university compared to expected log(salary) of those who graduated from the school ranked 61th and worse.

δ_1 → same rationale