4 Testing Hypotheses about a Single Linear Combination of the Parameter

Consider

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

where jc = number of years attending a two-year college
$univ$ = 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 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
$H_a: \beta_1 \neq \beta_2 \rightarrow H_a: \beta_1 - \beta_2 \neq 0$

depending on degree of freedome,
area = significant level / 2

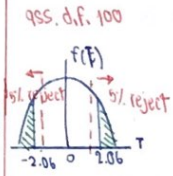
$$t = \frac{(\hat{\beta}_1 - \hat{\beta}_2) - 0}{s.e.(\hat{\beta}_1 - \hat{\beta}_2)}$$

we compute this t-statistic and compare w/ critical value

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

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

not very straight forward to calculate
↝ we use a variable transformation trick
↝ see notes



- suppose, we calculate a t-statistic $= \frac{\hat{\beta}_j - \beta_j}{s.e.\hat{\beta}_j} < -2.06$ (2 tailed test)
- suppose, we are testing $H_0: \beta_j = 0$ $H_a: \beta_j \neq 0$
- p-value = total shaded area
  p-value = significant level which we will reject the H_0
  or prob that we will reject H_0

area = 2 × (0.5 − 0.4803)
= 0.0394 < p-value> If p-value < significance level → reject H_0

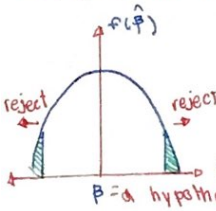
Note union Page 69

Inference ↝ Hypothesis testing about "β" the true parameter

$wage = \beta_0 + \beta_1 educ + \beta_2 experience + \ldots + u$

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

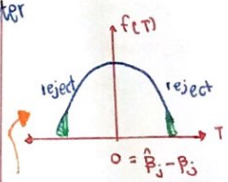
But we don't know what the true β are so, we use $\hat{\beta}$ (estimator) and $s.e.(\hat{\beta})$ to test the hypotheses



1) Test if β = some number
   e.g. $\beta_j = 0$ ↝ x_j has no impact on y
   $\beta_j = 1$ ↝ 1 unit↑ in x_j correspond to 1 unit↑ in y

continue from illustrate

↑ continuous



t-test ≈ how 99.
$\frac{\hat{\beta}_j - \beta_j}{s.e.(\hat{\beta}_j)} \sim t_{d.f.}$

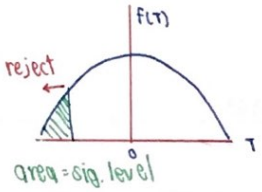
significant level = total area in the rejection region

another possible hypothesis test (one-tailed alternative)

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

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

- It is assumed that β_1 would not be more than β_2 (return to a 2-year college would never be more than returns to university education



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

* Then go to the extra note!

Sub in the main regression and get

$y = \beta_0 + (\theta_1 + 3\beta_2)x_1 + \beta_2 x_2 + \beta_3 x_3 + u$

$= \beta_0 + \theta_1 x_1 + 3\beta_2 x_1 + \beta_2 x_2 + \beta_3 x_3 + u$

$= \beta_0 + \theta_1 x_1 + \beta_2(3x_1 + x_2) + \beta_2 x_3 + u$

* Now, the explanatory variables are going to be x_1, $x_2 + 3x_1$ and x_3

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

In class exercise

Consider the multiple regression model

assume MLR 1-6 are satisfied

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

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

1st) write the t-statistic for testing H_0

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

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

$H_0: \theta_1 = 0$ $\quad H_a: \theta_1 \neq 1$

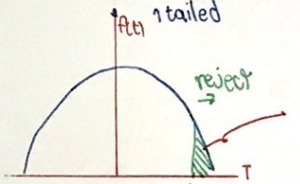
$t = \frac{\hat{\theta}_1 - 1}{se(\hat{\theta}_1)}$ → we need our regression to have θ_1 in it. So stata or OLS estimation will automatically give $\hat{\theta}_1$ & s.e. $\hat{\theta}_1$

Now, $\hat{\beta}_1 = \hat{\theta}_1 + 3\hat{\beta}_2$

or $\beta_1 = \theta_1 + 3\beta_2$

5 Computing p-Values for t-Tests

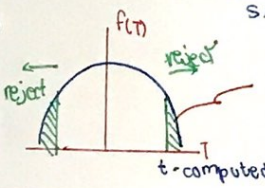
- What is the significance level given the computed t-statistics?



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

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

2 tailed



Total area from 2 sides

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

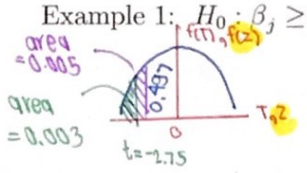
- p-value : $P(|T| > |t|)$

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

t = computed t-statistic

p-value = probability that a random T value will be greater (in the 1-1 term) than our t in H_0 test

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



P-value = what should be the significant level given the critical value of -2.75 ?? → find the shaded area!

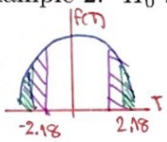
suppose the calculated $t_{\hat{\beta}_j} = -2.75$

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

- 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 < significant level

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



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

- From the t-table, the value -2.18 corresponds to area = 0.02 to 0.05.
- Thus, p-value = is between 0.02 - 0.05.
- Would we reject H_0 if we use the significance level = 5%?

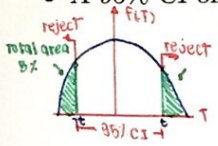
Yes, reject H_0 because the area is less than 0.05 or p-value

6 Confidence Intervals (CI)

- Confidence Intervales for the POPULATION PARAMETER (β_j)

The range of value that would capture the true β_j at a 95% chance

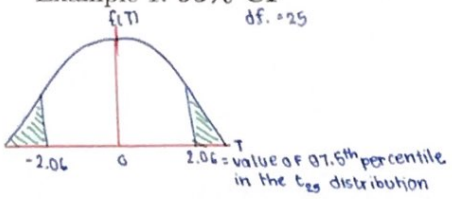
- A 95% CI of β_j is given by



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

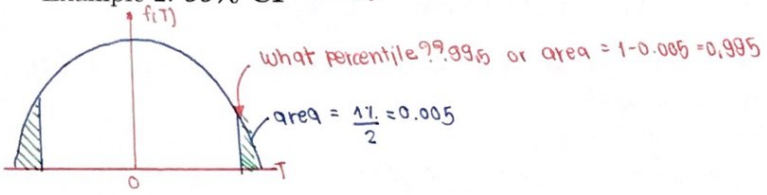
C is the 97.5 percentile in the t-distribution with n-k-1 d.f.

Example 1: **95% CI**



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** df = 25



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

Note F-test motivation

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

$Grade_{325} = \beta_0 + \beta_1$ # times_front + β_2 # times_back + β_3 hr. study + β_4 past! GPA + β_5 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 $\beta_2 \neq 0$

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

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

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

7 Testing Multiple Linear Restrictions: The F-test

Suppose the model is specified by

$$\begin{aligned} y &= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + u \\ H_0 &: \beta_2 = 0 \text{ and } \beta_3 = 0 \\ H_a, H_1 &: H_0 \text{ is not true} \end{aligned}$$

we want to test if x_1 and x_2 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: (Big model)

$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + u$ is true → reject H_0

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

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

e.g. in this model $q = 2$ → $y = \beta_0 + \beta_1 x_1 + u$ is true → don't reject H_0

- suppose these are "q" number of β that we would like to perform a joint-test of =0

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

$H_0: \beta_{k-q+1} = \beta_{k-q+2} = \ldots = \beta_k = 0$ (the last q $\beta_j = 0$)

H_a : H_0 is not true

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

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

This is always (+) b/c $SSR_{ur} < SSR_r$. Every time you add 1 more x_j, the model will be better explained.

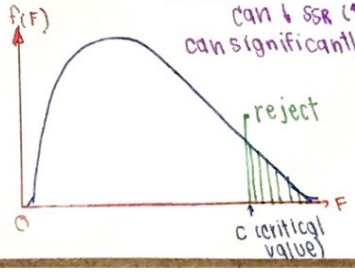
$n-k-1$: d.f. of the model "ur"

Note

So if every time you add 1 more x variable, the SSR↓ and R^2↑, why don't we just keep the additional x in the model ??

Because every time we add 1 more x_j, $var(\hat{\beta}_j)$ will increase, making the prediction of β less precise. So, we only keep the addition x_js if it/they can improve the model enough

can ↓ SSR (↑R^2) enough.
can significantly ↓ SSR and ↑ R^2



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

H_a : H_0 not true

$F \sim F_{q, n-k-1}$ (q: # of joint hypotheses being tested; $n-k-1$: d.f. of the ur. model)

we reject H_0 of jointly no effect if $F > \alpha$

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

we have $F = \frac{(R^2_{UR} - R^2_r)}{q} \div \frac{1 - R^2_{ur}}{n-k-1}$

q = # of β that are set to "0"

n = # of observations
k = # of slope β
1 = intercept

if we want to test overall significance of the model

$H_0 : \beta_1 = \beta_2 = \beta_3 = \ldots = \beta_k = 0$

H_a : otherwise

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

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
r	*years*	= years in major leagues
ur	*gamesyr*	= games per year in the league
ur	*bavg*	= career batting average
ur	*hrunsyr*	= homeruns per year
ur	*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
Residual	183.186327	347	.527914487
Total	492.175535	352	1.39822595

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

HW: $F \equiv \frac{R^2/q}{(1-R^2)/(n-k-1)}$ ≡ 99

log_salary	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
years 1	.0688626	.0121145	5.68	0.000	.0450355	.0926898
gamesyr 2	.0125521	.0026468	4.74	0.000	.0073464	.0177578
bavg 3	.0009786	.0011035	0.89	0.376	-.0011918	.003149
hrunsyr 4	.0144295	.016057	0.90	0.369	-.0171518	.0460107
rbisyr 5	.0107657	.007175	1.50	0.134	-.0033462	.0248776
_cons 0 intercept	11.19242	.2888229	38.75	0.000	10.62435	11.76048

- the restricted model (r) is defined by

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

. regress log_salary years gamesyr

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

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

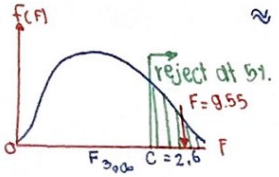
log_salary	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
years	.071318	.012505	5.70	0.000	.0467236	.0959124
gamesyr	.0201745	.0013429	15.02	0.000	.0175334	.0228156
_cons	11.2238	.108312	103.62	0.000	11.01078	11.43683

∘ But when performing an F-test, Performance have joint impact

Now, our H_0 and H_a becomes

$$F \equiv \frac{SSR_r - SSR_{ur}/q}{SSR_{ur}/(n-k-1)} = \frac{(198.311 - 183.186)/3}{183.86/353-5-1} \approx 9.55$$

let's use 5% level of sig.



since F = 9.55 > 2.6
we reject H_0 at 5% level
and conclude that performances have joint effects on salary.

8 How the Hypothesis Testing is done in Practice

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

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

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

   $\Rightarrow$ **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)**	**(2)**	**(3)**
log(*sales*) (sales)	.224 (.027)	.158 (.040)	.188 (.040)
log(*mktval*) (other company performance)	—	.112 (.050)	.100 (.049)
profmarg (other company performance)	—	−.0023 (.0022)	−.0022 (.0021)
ceoten (ceo characteristics)	—	—	.0171 (.0055)
comten (ceo characteristics)	—	—	−.0092 (.0033)
intercept	4.94 (0.20)	4.62 0.25)	4.57 (0.25)
Observations *R*-squared	177 .281	177 .304	177 .353

(like a simple regression w/ 1 x)

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

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

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 ??

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

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

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

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

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

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

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

The value of this variable is going to be 1000 times large than Faminc

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

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

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

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

$$bweght_{kg} = \frac{\hat{\beta}_0}{1000} + \frac{\hat{\beta}_1}{1000} cigs + \frac{\hat{\beta}_2}{\frac{1000}{30000}} faminc_{THB}$$

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

2 More on functional forms

- Logarithmic Functional Form

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

usually means natural log

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

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

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

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

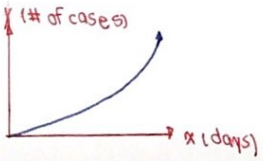
➡ if we want the upper term to be % change, then

$100\beta_2 = \frac{100\,\frac{1}{y}\Delta y}{\Delta x_2}$ $\quad 100\beta_2 = \%\,\Delta$ in y given that x_2 ↑ by 1 unit

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

- Models with Quadratics (squres)

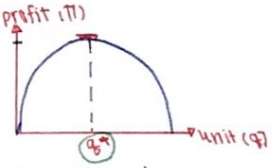
capture ↑ / ↓ marginal effects (slope of the relationship between x & y isn't constant.



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 return

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

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

Assume

$\pi = (P - MC)q$; MC=10 Demand ; P=100−q

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

F.O.C $\frac{d\pi}{dq} = 0 = 90 - 2q$ (β_1 is (+) β_2 is (−))

Example : Effects of Pullution on Housing Prices

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

0. providing note [illegible] log [illegible] wikipedia

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

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
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

log(price) → lprice; log(nox) → lnox

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

$|t| > 1.96$ ↑ ↑ all < 0.05
⤷ all variables are significant

Consider the effect of "room"

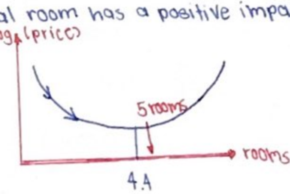
$\frac{d\log(price)}{d\,rooms} = \beta_3 + 2\beta_4\,rooms = -0.553 + 2(0.062)\cdot rooms$

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

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

$rooms = 4.4$

∴ at 4.4 room or more
at 5 room or more



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

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

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

$= 100 \times 0.067 = 6.7\%$ increase

total % Δ in price when # rooms ↑ from 5 to 7 is $(6.7 + 19.1) \times 100 = 25.8\%$.

○ what about % in price when # rooms ↑ from 6 to 7 ??

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

3 Models with Interaction Terms

→ used when the impact of one variable depends on the value (level)

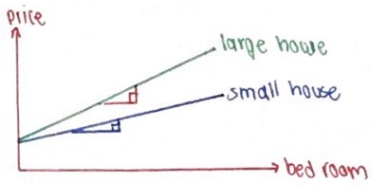
Consider

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

x_1 x_2 x_3 ($x_1 \cdot x_2$) x_2

where

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



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

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

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

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

- But we lose the "degree of freedom"
  (d.f. = free data point used to estimate the parameter)
  ⇝ 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.

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

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

if we have more k, d.f. $= n-k-1$ ↓, $SSR/(n-k-1)$ ↑, adj-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

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

27.9% of variation in Y is explained. So, 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 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

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.

$wage = \beta_0 x_0 + \delta_0 female + \beta_1 educ + \delta_1 male + u$

(x_0: intercept x_1; x_2; x_3)

For example:

id	female	male	x_0
1	1	0	1
	1	0	1
	0	1	1
	0	1	1
			1
99			1

$$x_0 = x_1 + x_2$$
$$1 = female + male$$
$$female = male + 1$$

or

if there are "n" categories, we omit "1" category to avoid multi collinearity

$$1 = winter + spring + summer + fall$$
$$winter = 1 - spring - summer - fall$$

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

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

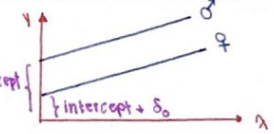
etc.

id	winter	spring	summer	fall	x_0
1	1	0	0	0	1
2	1	0	0	0	1
3	0	0	1	0	1
⋮	⋮	⋮	⋮	⋮	⋮

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

in this case, male

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