4 Testing Hypotheses about a Single Linear Combination of the Parameter

Consider

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

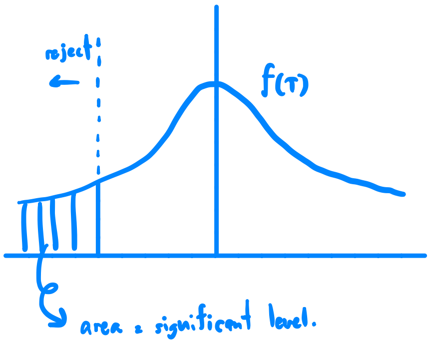
where $jc =$ number of years attending a two-year college
$univ =$ number of years at a four-year college
$\exp er =$ months in the workforce.
We want to test whether $\beta_1 = \beta_2$.

another possible hypothesis test (one-tailed alternative)

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

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

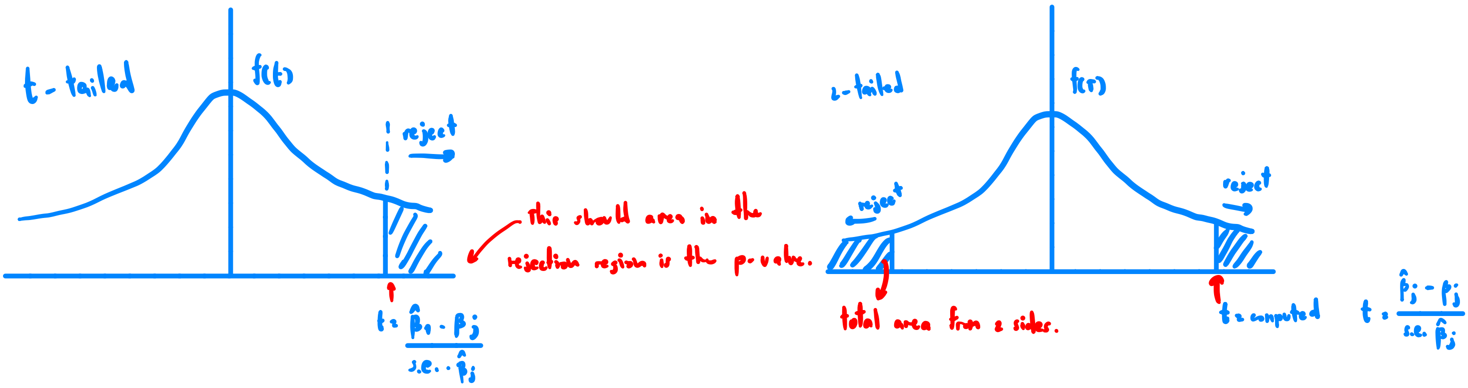
- It is assumed that β_1 would not be more than β_2 (returns to a 2-year college would never be more than return to university education)



$$t = \frac{(\hat\beta_1 - \hat\beta_2) - 0}{s.e.(\hat\beta_1 - \hat\beta_2)}$$

5 Computing p-Values for t-Tests

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



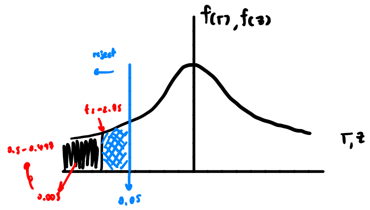
- p-value : $P(|T| > |t|)$

T = t-distributed random variable with d.f. = 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: $H_0 : \beta_j \geq 0, \quad H_a : \beta_j < 0, \quad$ d.f.$= 140$. → n-k-1 > 30 so we use z-table.



↝ P-value : what should be the significant level, given the critical value of -2.75?!

→ find the shaded area !

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

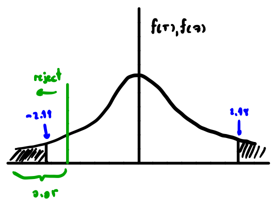
↪ $t_{\hat{\beta}_j} = \hat{\beta}_j - \beta_j / se(\hat{\beta}_j)$

- From the z-table, the value -2.75 corresponds to area = _ _ _ 0.003 _ _ _ _ _ _ _ _ _.
- Thus, p-value = _ _ _ 0.003 _ _ _ _ _ _.
- Would we reject H_0 if we use the significance level = 5%?

→ Rule! we reject H_0 if P-value < significant

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 $t_{\hat{\beta}_j} = -2.18$

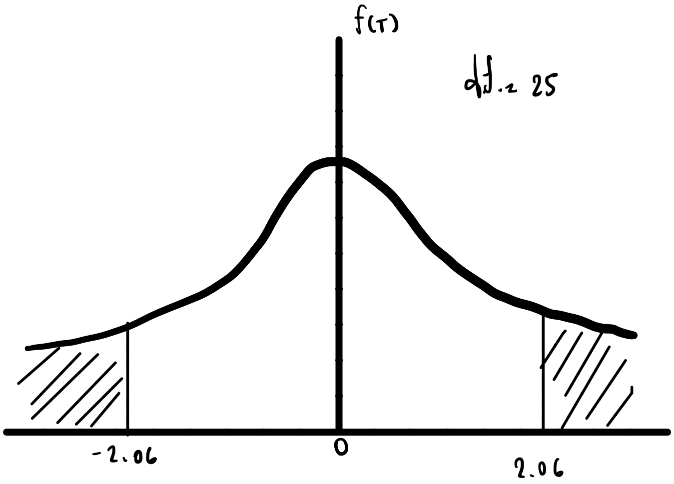
- From the t-table, the value -2.18 corresponds to area = _ _ _ (0.02, 0.05) _ _ _ _ _ _ _.
  (find from t-table)
- Thus, p-value = _ _ between 0.02 and 0.05 _ _ _.
- Would we reject H_0 if we use the significance level = 5%?

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

6 Confidence Intervals (CI)

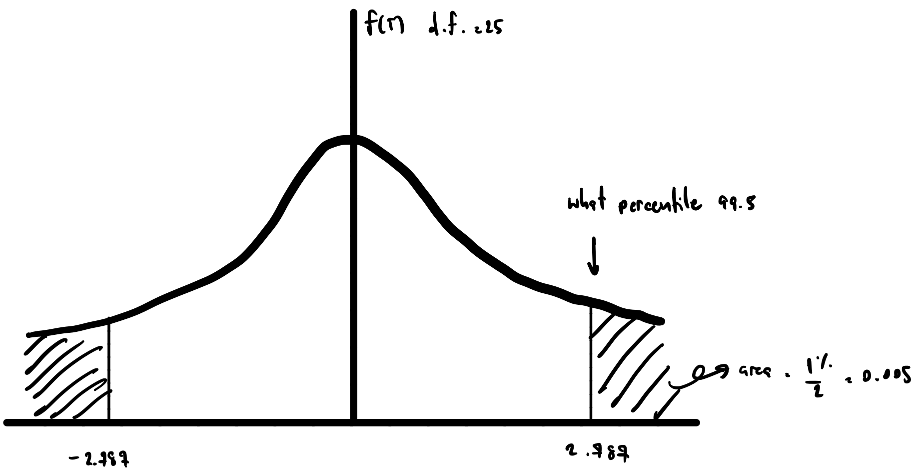
- Confidence Intervales for the POPULATION PARAMETER (β_j)
- A 95% CI of β_j is given by

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

7 Testing Multiple Linear Restrictions: The F-test

Suppose the model is specified by

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

→ want to test if x_2, x_3 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" (Big Model) model (ur). Suppose it can be expressed as:

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

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

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

   ↳ small Model

   $y = \beta_0 + \beta_1 x_1 + u$ is true → do not reject H_0.

- Suppose there are "q" number of β that we would like to perform a joint test of $= 0$

  e.g. in this model $q = 2$

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

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

  (the last q $\beta_q = 0$)

  H_a: H_0 is not true.

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

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

→ this term always positive and $SSR_{ur} < SSR_r$

Every time you add 1 more x, the model will be better explain.

$(n-k-1)$ df 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) → $SSR_{ur} > SSR_r$

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

doesn't improve the model enough.

4. Other ways to calculate the F-statistics:

→ From $R^2 = 1 - \frac{SSR \nearrow RSS}{SST \to TSS}$

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

# of β that are set to "0"

# of obs

slopes of β

intercept

⇒ 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

$F = \dfrac{R^2(\text{in the model})/k}{(1-R^2)/(n-k-1)}$

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.

	salary	= season salary
r	*years*	= years in major leagues
	gamesyr	= games per year in the league
ur	*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

More explanatory variables
↳ more R^2, adjusted R^2

ur model

y x

```
. regress log_salary years gamesyr bavg hrunsyr rbisyr

      Source |       SS       df       MS              Number of obs =     353
-------------+------------------------------           F(  5,   347) =  117.06
SSE    Model |  308.989208     5  61.7978416           Prob > F      =  0.0000
SSR Residual |  183.186327   347  .527914487           R-squared     =  0.6278
-------------+------------------------------           Adj R-squared =  0.6224 ← we look at
SST    Total |  492.175535   352  1.39822595           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
q=3     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
------------------------------------------------------------------------------
```

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

- the restricted model (r) is defined by

y x

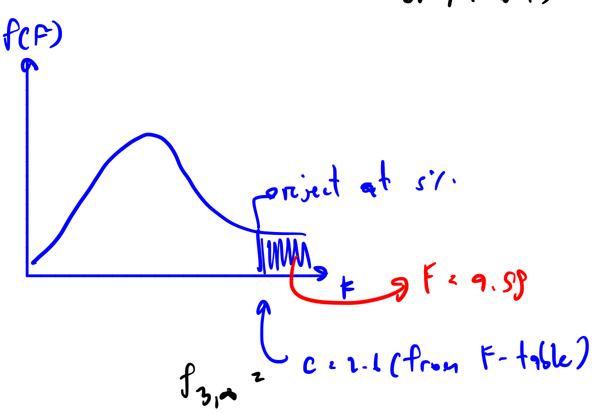
```
. regress log_salary years gamesyr

      Source |       SS       df       MS              Number of obs =     353
-------------+------------------------------           F(  2,   350) =  259.32
SSE    Model |  293.864058     2  146.932029           Prob > F      =  0.0000
SSR Residual |  198.311477   350  .566604221           R-squared     =  0.5971
-------------+------------------------------           Adj R-squared =  0.5948
SST    Total |  492.175535   352  1.39822595           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, performances has joint impact.

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

Now, our H_0 and H_a becomes

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



let's use 5% level of sig.

since F = 9.55 > 2.6, we reject H_0 at 5% level and 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 $H_0 : \beta_i = 0$ z-table

   $\Rightarrow$ If the d.f. > 30, then when $t \;\; > 1.96$, we can reject H_0 with 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 $\quad H_0 : \beta_i = 5 \quad$ etc.

   or perform an F-test for testing multiple linear restrictions

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

sales

other company performance

CEO characteristics

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

like a simple regression with 1x

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

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

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 bweight in kilograms ?

1kg = 1,000 g

$$\widehat{bweight}_{kg} = \frac{\widehat{bweight}_g}{1000} = \frac{\hat\beta_0}{1000} + \frac{\hat\beta_1}{1000} cigs + \frac{\hat\beta_2}{1000} faminc$$

$$= \hat\alpha_0 + \hat\alpha_1 cigs + \hat\alpha_2 faminc$$

$$\rightarrow \hat\alpha_0 = \frac{\hat\beta_0}{1000}, \quad \hat\alpha_1 = \hat\beta_1/1000, \quad \hat\alpha_2 = \frac{\hat\beta_2}{1000}$$

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

$$bweight_g = \hat\beta_0 + \hat\beta_1 cigs + \hat\beta_2 faminc_{USD}$$

the value of the variable is going to be 1000 times larger than faminc.

$$\rightarrow \hat\theta_2 = \hat\beta_2/1000$$

in other words θ_2 = impact of 1 USD ↑ income

$$\hat\beta_2 = " \longrightarrow 1000 \text{ USD} \uparrow \text{income}$$

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

$$bweight_{kg} = \frac{\hat\beta_0}{1000} + \frac{\hat\beta_1}{1000} cigs + \hat\beta_2 \div \frac{\hat\beta_2}{1000} faminc_{THB}$$

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

2 More on functional forms

- Logarithmic Functional Form

usually means natural log.

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

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

With the log y of log x format, the coefficient is going to be the elasticity!

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

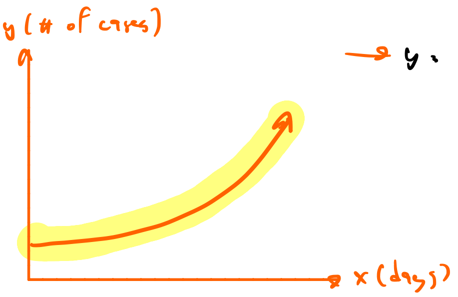
→ if we want the upper term to be % change, then $100\beta_2 = \%\Delta$ in y

$100\beta_2 = \dfrac{100\frac{1}{y}\Delta y}{\Delta x_2} = \dfrac{\%\Delta y}{\Delta x_2}$ given that x_2 increase by 1 unit.

- Models with Quadratics

→ Capture increasing / decreasing marginal effects (slope of the relationship btw x & y is not constant)

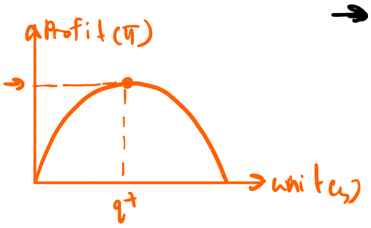
Covid-19 example



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

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

Decreasing Return



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

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

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

$\pi = (100-q-10)q$ Demand : $P = 100 - q$

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

Example : Effects of Pullution on Housing Prices

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

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 school in the area without having to take any test so the lower stratio, the better the school.

The estimation result is given by

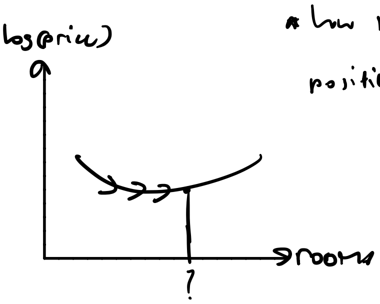
```
regress  lprice lnox dist rooms rooms_sq stratio

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

log(price) → lprice; log (nox) → lnox; β_1, β_2, β_3, β_4, β_5

all < 0.05 reject H_0

Consider the effect of "room" → all variables are significant

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



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

$0 < -0.553 + 2(0.062)\, rooms$

rooms < 4.4

Answer → at 4.4 rooms or more

at ~ 5 rooms or more.

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

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

total % change in price when # rooms ↑ from 5 to 7 is 6.7 + 19.1 % = 25.8%

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

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

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

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

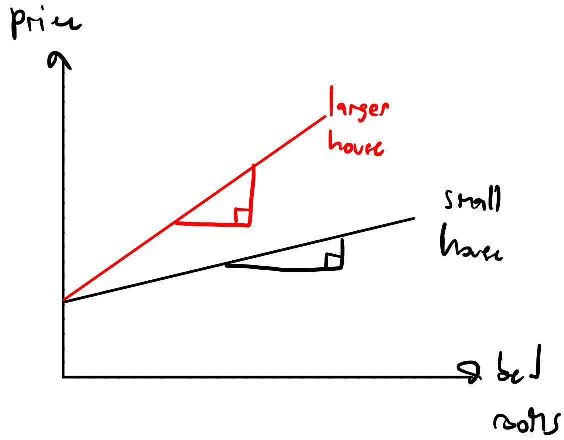
3 Models with Interaction Terms

→ used when the impact of one variable depend on the value (level) of another variable,

Consider

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

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



- $\dfrac{d\,price}{d\,bdrms} = \beta_2 + \beta_3\, sqrft$

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

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. $= n-k-1$ ↓

$SSR/(n-k-1)$ ↑, adj-R^2 ↓

$$\widehat{salary} = \begin{array}{lll} 830.63 & +0.0163 sales & +19.63 roe \\ (223.90) & (0.0089) & (11.08) \end{array}$$

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

Consider Model 2

$$\log(\widehat{salary}) = \begin{array}{lll} 4.36 & +0.2751 \log(sales) & +0.0179 roe \\ (0.29) & (0.033) & (0.004) \end{array}$$

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

29.5 % 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

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

$= \beta_0 + \delta_0 female + \beta_1 educ + E(u \mid female, educ)$ $\hookrightarrow = 0$

(ass MLR 1-4 holds)

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

② Thus ♀ : $E(wage \mid female = 1, educ) = \beta_0 + \delta_0(1) + \beta_1 educ = \beta_0 + \delta_0 + \beta_1 educ$

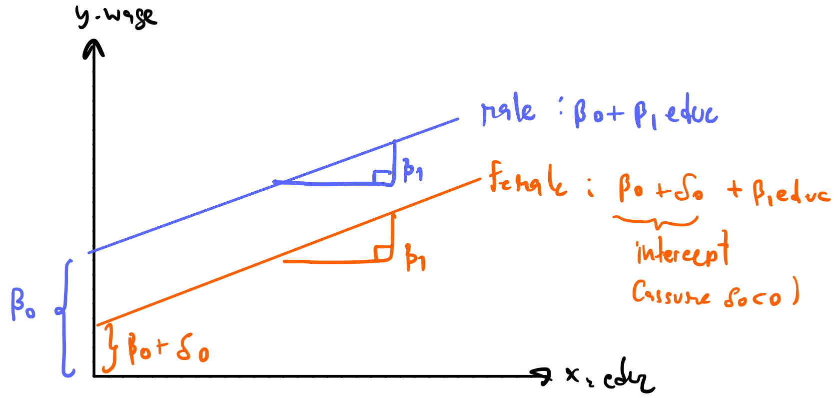
♂ : $E(wage \mid female = 0, educ)\ \beta_0 + \delta_0(0) + \beta_1 educ = \beta_0 + \beta_1 educ$

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

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

* given the same value of educ (same education level)

δ_0 is the difference in the expected wage of females and males



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

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

intercept (x_1) (x_2) (x_3)

$x_0 = x_1 + x_3$

$$1 = female + male$$
$$female = male + 1$$

id	female	male
1	1	0
2	1	0
3	0	1
4	0	1
⋮	0	1
⋮	1	1
⋮	1	0
95		

or

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

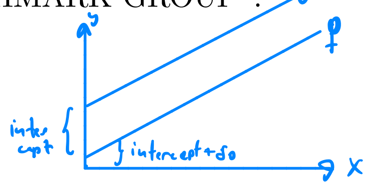
$$1 = winter + spring + summer + fall$$
$$winter = 1 - spring - summer - fall$$

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

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

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

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



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

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

      Source |       SS           df       MS      Number of obs   =       526
-------------+----------------------------------   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 worker 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*.

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

$$\begin{aligned} \log(wage) &= \beta_0 + \delta_0 female + \delta_1 married + \beta_1 educ + \beta_2 \exp er + \beta_3 \exp er^2 \\ &\quad + \beta_4 tenure + \beta_5 tenure^2 + u. \end{aligned}$$

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

      Source |       SS       df       MS              Number of obs =     526
-------------+------------------------------           F(  7,   518) =   58.76
       Model |  65.6482326     7  9.37831895           Prob > F      =  0.0000
    Residual |  82.6815188   518  .159616832           R-squared     =  0.4426
-------------+------------------------------           Adj R-squared =  0.4351
       Total |  148.329751   525  .28253286            Root MSE      =  .39952

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

2.) δ_1 measures the impact of be married. (marriage premium)

Comments: but since $|t| < 1.96$ or $p > 0.05$, we don't reject H_0 of no impact.

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

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

females workers are expected to earn less than males workers by 29.02%, holding other factors the same.

$$100 \cdot \frac{\frac{1}{wage}\, dwage}{d\, female} = 100(-0.29)$$

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

	♀	♂
msm	marfem	marrmale
sing	singfem	singmale

Consider a model which includes dummy variables for each gender/marital status combination– *marrmale*, *marrfem* and *singfem*. (on single male ← 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
Model	68.3617623	8	8.54522029
Residual	79.9679891	517	.154676961
Total	148.329751	525	.28253286

Number of obs =	526
F(8, 517) =	55.25
Prob > F =	0.0000
R-squared =	0.4609
Adj R-squared =	0.4525
Root MSE =	.39329

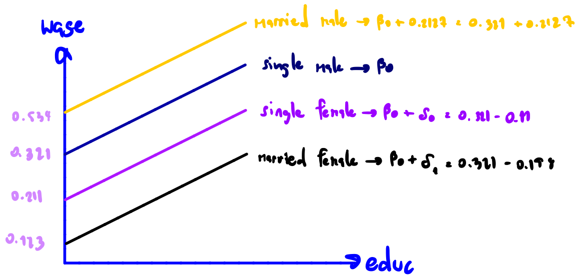
	lwage	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
δ_0	marrmale	.2126757	.0553572	3.84	0.000	.103923	.3214284
δ_1	marrfem	-.1982676	.0578355	-3.43	0.001	-.311889	-.0846462
δ_2	singfem	-.1103502	.0557421	-1.98	0.048	-.219859	-.0008414
$\hat{\beta}$	educ	.0789103	.0066945	11.79	0.000	.0657585	.092062
	exper	.0268006	.0052428	5.11	0.000	.0165007	.0371005
	expersq	-.0005352	.0001104	-4.85	0.000	-.0007522	-.0003183
	tenure	.0290875	.006762	4.30	0.000	.0158031	.0423719
	tenursq	-.0005331	.0002312	-2.31	0.022	-.0009874	-.0000789
	_cons	.3213781	.100009	3.21	0.001	.1249041	.5178521

Comments: This regression is not the same as the previous one.

It uses "single male" as the base group.

(The previous 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...
- $\delta_2 \rightsquigarrow$ same rational



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

e.g. age may explain the model better if split into generations young 0-15 gen Z 16-29 etc.

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

      Source |       SS       df       MS              Number of obs =     136
-------------+------------------------------           F(  8,   127) =  120.15
       Model |  9.16538532     8  1.14567316           Prob > F      =  0.0000
    Residual |   1.2109665   127  .009535169           R-squared     =  0.8833
-------------+------------------------------           Adj R-squared =  0.8759
       Total |  10.3763518   135  .076861865           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 base line is ranking 61th and wage.

Comments:

1.) δ_0 measures the difference 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 rank 61th and worse.

2.) $\delta_1 \rightarrow$ use the same rational.

rank	top10	r11-25	r26-40
1	1	0	0
2	1	0	0
3	1	0	0
.	1	0	0
.	1	0	0
.	1	0	0
.	0	1	0
10	0	1	0
11	0	1	0
12		1	0
|		0	0
|		0	1
25		0	1
26		0	1
.			
40			