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

```
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
------------------------------------------------------------------------------
```

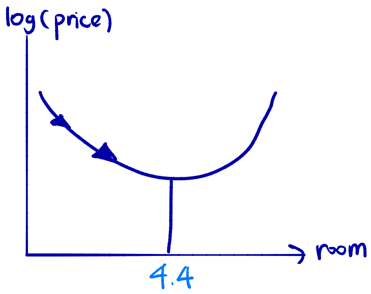
log(price) ← lprice; log(nox) ← lnox; β_1, β_2, β_3, β_4, β_5 (coefficients of lnox, dist, rooms, rooms_sq, stratio)

$|t| > 1.96$ < 0.05

→ all variables all significant

Consider the effect of "room"

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



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

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

$rooms = 4.4$

At 4.4 rooms or more

At $\overline{5}$ rooms or more

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

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

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

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

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

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

$\%\Delta price = 100(-0.0553 + 2(0.062)\cdot 6) = 19.1\%$

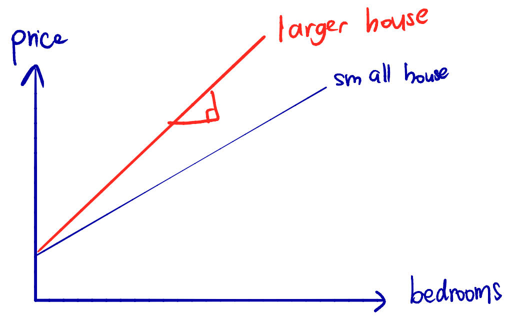
3 Models with Interaction Terms

→ used when the impact of one variable depends on the value of another variable

Consider

$$price = \beta_0 + \beta_1 \underbrace{sqrft}_{X_1} + \beta_2 \underbrace{bdrms}_{X_2} + \beta_3 \overbrace{\underbrace{sqrft}_{X_1} \times \underbrace{bdrms}_{X_2}}^{X_3} + \beta_4 bthrms + u$$

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



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

→ 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 $R^2 \uparrow$

- 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 to many regressors

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

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

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

d.f. $= n-k-1 \downarrow$

Consider Model 1

If we have more k, d.f. $\downarrow$, $\frac{SSR}{n-k-1} \uparrow$, adj $-R^2 \downarrow$

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

Consider Model 2

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

27.5% of variation in Y is explained. 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) $E(w \mid female, educ) = E(\beta_0 + \delta_0 female + \beta_1 educ + u \mid female, educ)$

$= \beta_0 + \delta_0 female + \beta_1 educ + \underbrace{E(u \mid female, educ)}_{=0}$

assume MLR 1-4 hold

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

2) Thus

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

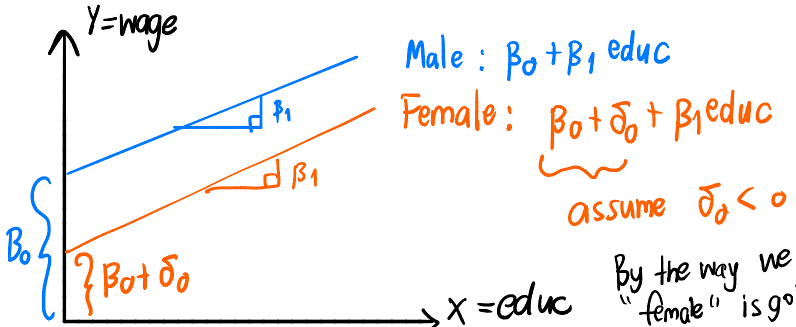
♂ : $E(wage \mid female = 0, educ) = \beta_0 + \delta_0(0) + \beta_1 educ = \beta_0 + \beta_1 educ$

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

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

* given the same value of educ (given 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 educ.

4 It is not possible to include all of the dummy alternatives in the same model

- If we include all alternatives of a dummy variable in the same model, we will face the "perfect collinearity" problem. (as long as there is an intercept in the model)

For example: 2 variables can be expressed as a linear relationship → values of 2 variables exactly the same
e.g. use price in THB and usp in the same equation

$$1 = female + male \quad X_0 = X_1 + X_3$$

$$female = male + 1$$

or

$$wage = \beta_0 X_0 + \delta_0 \underset{(X_1)}{female} + \beta_1 \underset{(X_2)}{educ} + \delta_1 \underset{(X_3)}{male} + u$$

intercept = 1

id	F	M	X_0
1	1	0	1
2	1	0	1
3	0	1	1
4	0	1	1
⋮			
99			

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

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

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

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

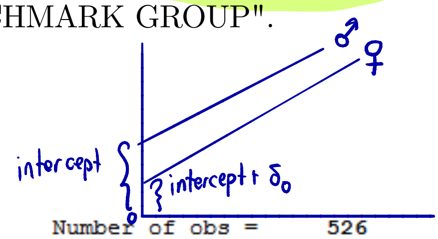
id	winter	spring	summer	fall
1	1	0	0	0
2	1	0	0	0
3	0	0	1	0
4	0	1	0	0
5	0	1	0	0

Add to 1

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

- At least one alternative has to be dropped. We treat the dropped alternative (in this case, male) as the "BASE GROUP" or "BASELINE" or "BENCHMARK GROUP".

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




```
. 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 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*.

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

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

```
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

2) δ_1 measure the impact of being married (marriage premium)
but since $|t| < 1.96$ or $p > 0.05$, we do not reject H_0 of no impact.

Comments:

1) δ_0 measures the expected difference between males and female given the same marital status and other factors

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

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

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

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

♀	♂
marr fem	marr male
singfem	singmale

base case

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

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

```
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
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 as it uses "single male" as base group (The previous one uses male & single as 2 base groups)

- δ_0 measures 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 male