# 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 ?

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

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

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

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

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

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

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

(The value of this variable is going to be 1000 times larger than faminc.

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

$\hat{\beta}_2 =$ " ——— " 1,000 USD ↑ income,

- What if we use bweght in kg & income in THB.

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

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

## 2 More on functional forms

- Logarithmic Functional Form

$\Delta y = y_1 - y_2$

$\Delta x = x_1 - x_2$

usually means natural log

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

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

with the log y & log x format, the coefficient is going to be the elasticity! ($x_1$ elasticity of y)

(price) (demand)

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

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

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

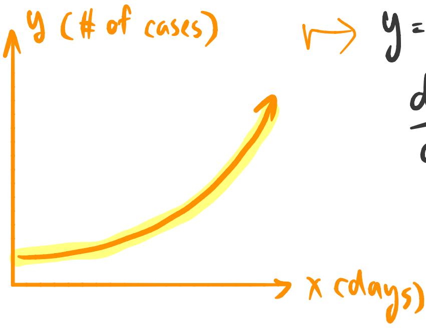
$100\beta_2 = \%\Delta$ in y given that $x_2$ increase by 1 unit

- Models with Quadratics

(squares)

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

Covid-19 example

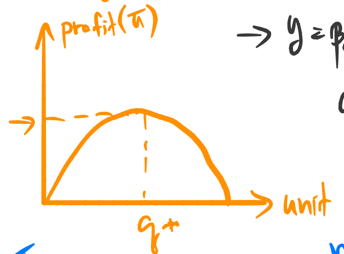



$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + u$$

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

(+) (+) days.

Decreasing returns.




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

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

(+) (−)

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

Demand : $P = 100 - q$

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

F.O.C $\frac{d\pi}{dq} = 0 = 90 - 2Q$

$\beta_1$ is positive

**Example :** Effects of Pullution on Housing Prices

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

$\beta_2$ is (−)

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

$price$ = 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 other 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)

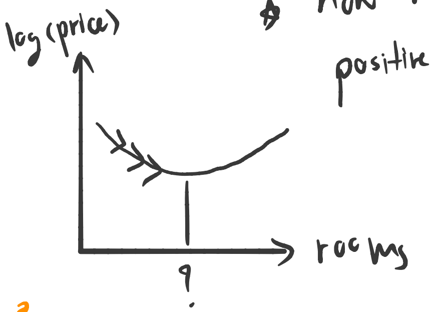
log(nox

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

$|t| > 1.96$  all $< 0.05$ reject $H_0$

Consider the effect of "room"  ⇝ all variables are significant.

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




* how many rooms dose 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

* ? ล่า

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

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

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

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

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

when about % in price when # rooms increases 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 depends on the value levels of another variable.

Consider

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

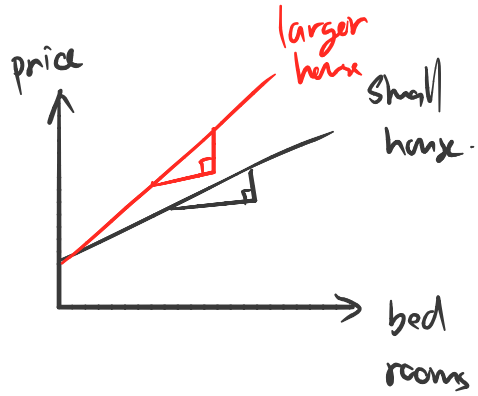
where

$price$ = housing price

$sqrft$ = house size (square feet)

$bdrms$ = number of bedrooms

$bthrms$ = number of bathrooms




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

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

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

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

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

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

27.5% of Varaition 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 $\delta_0$ 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 $E(wage \mid female, edu) = E(\beta_0 + \delta_0 female + \beta_1 educ + u \mid female, educ)$

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

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

② Thus

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

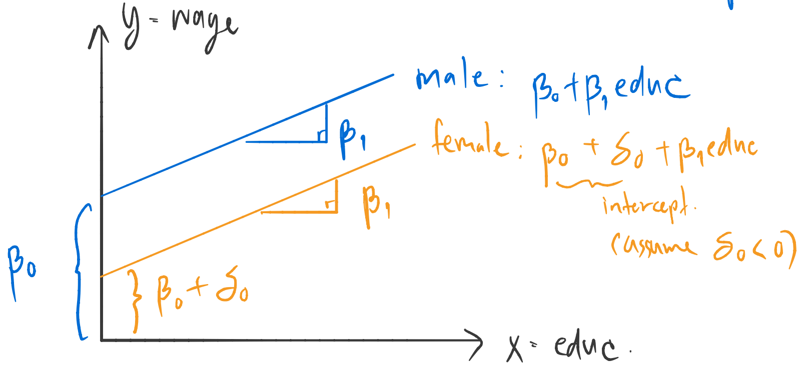
♂: $E(wage \mid female = 0, educ)$ $\beta_0 + \delta_0(0) + \beta_1 educ = \beta_0 + \beta_1 educ$

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

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

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

$\delta_0$ 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.

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

intercept $x_0$, $(x_1)$, $(x_2)$, $(x_3)$

$x_0 = x_1 + x_3$

| id | female | male |
|---|---|---|
| 1 | 1 | 0 |
| 2 | 1 | 0 |
| 3 | 0 | 1 |
| 4 | 0 | 1 |
| . | 0 | 1 |
| . | 0 | 1 |
| . | 1 | 0 |
| | 1 | 0 |
| 99 | | |

For example:

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

or

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

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

winter = { 1 if winter; 0 if otherwise

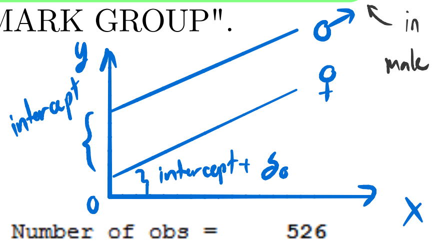
spring = { 1 if spring; 0 otherwise

etc.

| id | winter | spring | summer | fall |
|---|---|---|---|---|
| | | 0 | 0 | 0 |
| 1 | 1 | 0 | 1 | 0 |
| 2 | 0 | 0 | 0 | 0 |
| 3 | 1 | 0 | 0 | 0 |
| : | 0 | 1 | 0 | 0 |
| | 0 | : | 1 | 0 |
| | 1 | : | 0 | 0 |

- 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.

## 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*.

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

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

```
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 |

$\hat{\beta}$

| 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.) $\delta_1$ measures the impact of be married. (marriage premium)
but since $|t|<1.96$ or $p>0.05$, we don't reject $H_0$ of no impact.

Comments:

1.) $\delta_0$ measures the expected difference between female & male workers given the same marital status and other factors.

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

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

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

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

| | ♀ | ♂ |
|---|---|---|
| marr | marrfem | marrmale |
| sing | singfem | singmale |



Consider a model which includes dummy variables for each gender/marital status combination– $marrmale$, $marrfem$ and $singfem$. (or singmale ← used as the basecase)

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

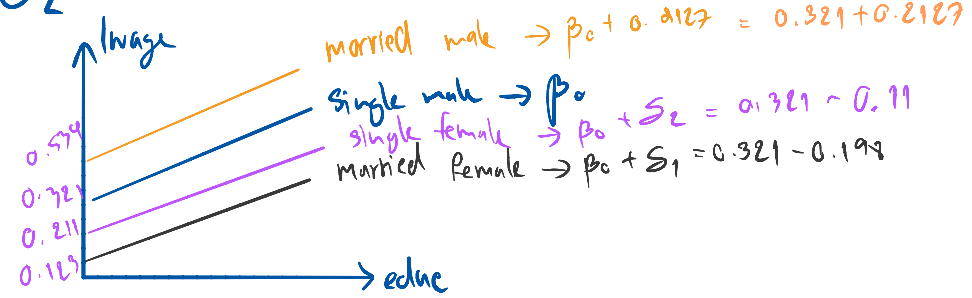
```
regress lwage marrmale marrfem singfem educ exper expersq tenure tenursq

      Source |       SS       df       MS              Number of obs =     526
-------------+------------------------------           F(  8,   517) =   55.25
       Model |  68.3617623     8  8.54522029           Prob > F      =  0.0000
    Residual |  79.9679891   517  .154676961           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
------------------------------------------------------------------------------
```

$\delta_0$, $\delta_1$, $\delta_2$ — $\hat{\beta}$

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.)

- $\delta_0$ measure the expected diff. in wage of married male as compared with single males, holding other factors constant.
- $\delta_1$ measures the expected diff. in wage of married female as compared with single males, holding ...
- $\delta_2$ ~> same rational

intercept.



focus on dummy variables.



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

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

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.

the baseline is ranking 61th and worse.

```
. 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
------------------------------------------------------------------------------
```

Comments:

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

1.) $\delta_0$ 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 ranked 61th and worse.

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