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 bweight in kilograms
1 kg = 1,000 g

bwght_kg^ = bwght_g^/1000 = (β^0/1000) + (β^1/1000) cigs + (β^2/1000) faminc
= α0^ + α1^ cigs + α2^ faminc
=> α0^ = (β^0/1000) , α1^ = (β^1/1000) , α2^ = (β^2/1000)

What if we use faminc in USD (instead of 1000 USD)
bwght_g = β^0 + β^1 cigs + β^2/1000 faminc_usd
= β^0 + β^1 cigs + θ^2 faminc_usd

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

=> θ^2 = β^2/1000
in other words θ^2 = impact of 1 USD increase in income
β^2 = impact of 1000 USD increase in income

What if we use bwght in kg & income in THB
bwght_kg = β^0/1000 + β^1/1000 cigs + β^2/1000 faminc_thb

2 More on functional forms

- Logarithmic Functional Form

△y = Y1 - Y2
△x1 = X11 - X12

usually means natural log

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

β1 = d log(y) / d log(x)
= (1/y)dy / (1/x1)dx1
= (1/y)△y / (1/x1)△x1
= [100(1/y)]△y / [100(1/x1)]△x1
= %y/%x

with the log y & log x format, the coefficient is going to be the elasticity.
(x1 elasticity of y)
price demand

β2 = d log(y) / dx2
= (1/y)dy / dx2
= (1/y)△y / △x2

-> If we want the upper term to be % change then

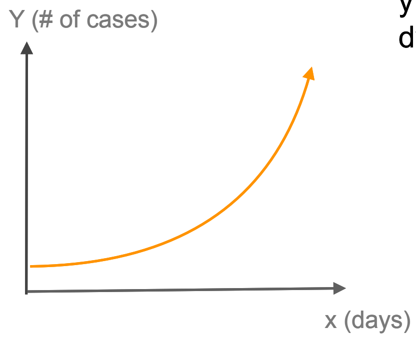
100 β2 = [100(1/y)]△y / △x2
100 β2 = % △ y / △ x2

100 β2 = %△ in y given that x2 increases by 1 unit

- Models with Quadratics

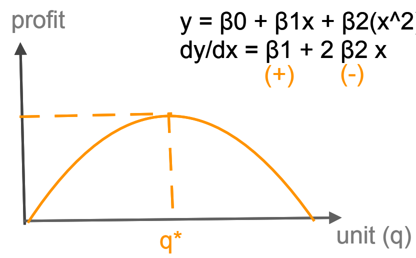
-> capture increasing / decreasing marginal effects (slope of the relationship between x & y is not constant)

COVID-19 example



y = β0 + β1x + β2(x^2) + u
dy/dx = β1 + 2 β2 x
(+) days

Decreasing returns



y = β0 + β1x + β2(x^2) + u
dy/dx = β1 + 2 β2 x
(+) (-)

π = (p-mc) q
π = (100-q(hat) - 10) q
F.O.C dπ/dq = 90 - 2Q

Assume mc=10
Demand: P=100-q

Example : Effects of Pullution on Housing Prices

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

page 76

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

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

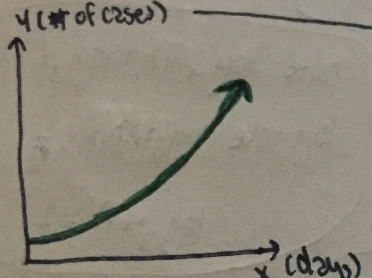
- $\beta_2 = \frac{d\log(y)}{dx_2} = \frac{\frac{1}{y}dy}{dx_2} = \frac{\frac{1}{y}\Delta y}{\Delta x_2}$

→ if we want the upper term to be % change,
then $100\beta_2 = \frac{100\frac{1}{y}\Delta y}{\Delta x_2}$

$100\beta_2 = \frac{\%\Delta y}{\Delta x_2}$ | $100\beta_2 = \%\Delta$ in y given that x_2 increases by 1 unit

→ capture increasing / decreasing marginal effects (slope of the relationship between x & y is not constant

COVID-19 example

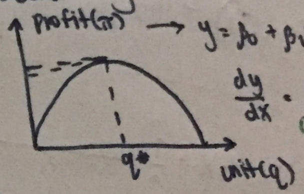


$y = \beta_0 + \beta_1 x + \beta_2 x^2 + u$

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

(+) (+) days

Decreasing returns



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

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

(+) (−)

$\pi = (p - mc)q$

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

$\boxed{\text{Assume } MC = 10 \quad \text{demand } p = 100 - q}$

FOC $\frac{d\pi}{dq} = 90 - 2Q$ → β_2 is (−)

β_1 is (+)

when adding quadratics to the same x this will make the relationship btw x & y be non linear and can capture the increasing and decreasing marginal effects 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 schools 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
Model	51.4933152	5	10.298663
Residual	33.0889098	500	.06617782
Total	84.582225	505	.167489554

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

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

|t| . 1.96
-> all variables are significant

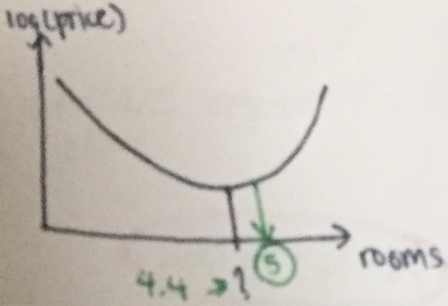
all < 0,05

When t statistic reject null P would also reject

Consider the effect of "room"

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

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



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

$0 = -0.553 + 2(0.062)\text{ rooms}$

$\text{rooms} = 4.4$

Answer → at 4.4 rooms or more

≈ 5 rooms or more

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

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

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

What about the % change in price when rooms increases from 5 to 7?

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

$= 19.1\%$

when # rooms ↑ from 5 → 7

total % Δ in price is 6.7 + 19.1%

= 25.8%

3 Models with Interaction Terms

=> used when the impact of one variable depends on the value (level) of another variable

Consider

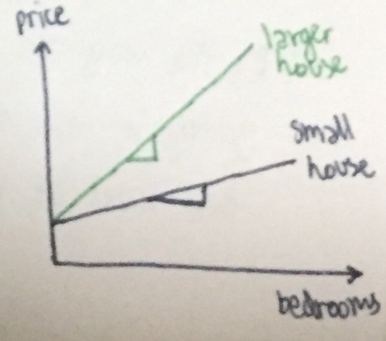
$$price = \beta_0 + \beta_1 \underbrace{sqrft}_{X1} + \beta_2 \underbrace{bdrms}_{X2} + \beta_3 \overbrace{\underbrace{sqrft \times bdrms}_{X1 * X2}}^{X3} + \beta_4 \underbrace{bthrms}_{X2} + u$$

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 increase

But if 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 R^-2 when we want to punish adding too many regressors
R^2 = 1 - (SSR/SST) = 1 - (SSR/n / SST/n)

Adj. R^2 = 1 - [SSR/(n-k-1) / SST/ (n-1)]
If we have more k, d.f. = n-k-1 decrease
SSR/ (n-k-1) increases, adj-R^2 decrease

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

$= \beta_0 + \delta_0 female + \beta_1 educ + E(u|female, educ)$

$= \beta_0 + \delta_0 female + \beta_0 educ$ $\quad \downarrow = 0$

(assumption MLR1-4) holds

② Thus,

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

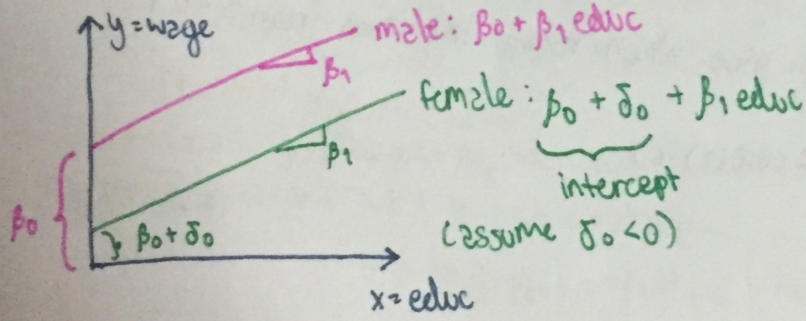
women ♂ : $E(wage|female = 0, educ) = \beta_0 + \delta_0(0) + \beta_1 educ = \beta_0 + \beta_1 educ$

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

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

* 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 educ

4 It is not possible to include all of the dummy alternatives in the same model (as long as there is an intercept in the model)

- If we include all alternatives of a dummy variable in the same model, we will face the "perfect collinearity" problem.

wage = β0X0+ δ0 female +β1 educ +δ1 male +u

(x1) (x2) (x3)

For example:

intercept = 1

$$
\begin{aligned}
x0 &= x1 + x3 \\
1 &= female + male \\
female &= male + 1
\end{aligned}
$$

or

If there are “n” categories, we omit “1” category to avoid multi collinearity

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

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

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

etc.

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

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

5 Using dummy variables for multiple categories

Case 1 We can use many dummy variables in the same model

Consider a model which includes 2 dummy variables– *female* and *married*.

1 if female
0 if otherwise

1 if married
0 if otherwise

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

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

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

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

$\hat{\beta}$

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

Comments:

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

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

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

$$\frac{100 \cdot \frac{1}{wage} dwage}{\partial female} = 100 \cdot -0.29$$

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

2) δ_0 measures the impact of be married (marriage premium)

But since $|t| < 1.96$ or $p > 0.05$, we do not reject H_0 of no impact

	male	female
marr	marfem	marmale
sing	singfem	singmale

basecase

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

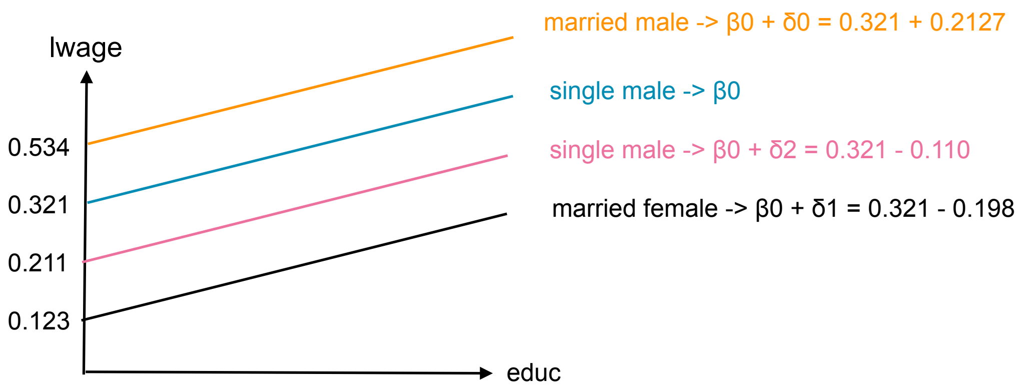
Comments:

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)

- $\delta 0$ measures 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 other factors constant.

- $\delta 2$ -> same rationale



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 cases the "range of value" serve as a better explanatory variable than the "value" itself. eg age may be explain the model better if split into generations young (0-15) genz (16 - 29) etc.

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

      Source |       SS       df       MS              Number of obs =     136
-------------+------------------------------           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 baseline is ranking 61st and worse

Comments:

1) δ0 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 61st and worse

2) δ1 —> use the same rationale