Supanuth Peganant
6104640161

4 Testing Hypotheses about a Single Linear Combination of the Parameter

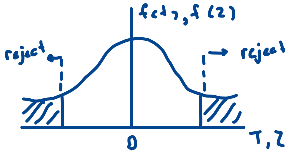
Consider

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

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 return from 1 more yr of educ at a junior colledge is the same as that of the uni.

$H_0 : \beta_1 = \beta_2$ >> $H_0 : \beta_1 - \beta_2 = 0$
against
$H_a : \beta_1 \neq \beta_2$ >> $H_0 : \beta_1 - \beta_2 \neq 0$

2 tailed test



$$\text{area} = \frac{\text{significant level}}{2}$$

$$\boxed{t = \frac{(\hat{\beta}_1 - \hat{\beta}_2) - 0}{se(\hat{\beta}_1 - \hat{\beta}_2)}}$$

} we compute this t-statistic and compare with the critical value

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

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

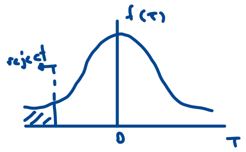
not very straight foward to calculate
>> we use a variable transformation trick → see notes

another possible hypothesis test (one-tailed alternative)

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

$H_a : \beta_1 < \beta_2 \gg H_0 \cdot \beta_1 - \beta_2 < 0$

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



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

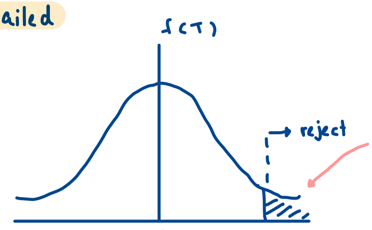
area = significant level

Then, go to the extra note.

5 Computing p-Values for t-Tests

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

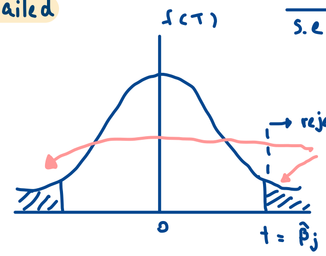
1 tailed



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 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 Ho test

EXTRA NOTES

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 $\theta_1 = \hat{\beta}_1 - 3\hat{\beta}_2$ » $H_0 : \theta_1 = 1$, $H_a : \theta_1 \neq 1$

$t = \frac{\hat{\theta}_1 - 1}{s.e(\hat{\theta}_1)}$ >> 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 $\beta_1 = \theta + 3\beta_2$

Sub in the main regression and get

$$Y = \beta_0 + (\theta_1 + 3\beta_2) X_1 + \beta_2 X_2 + \beta_3 X_3 + u$$

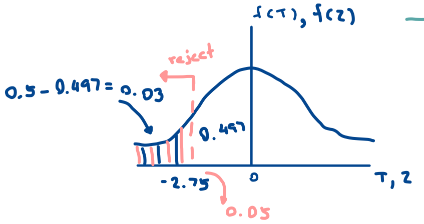
$$= \beta_0 + \theta_1 X_1 + 3\beta_2 X_1 + \beta_2 X_2 + \beta_3 X_3 + u$$

$$= \beta_0 + \theta_1 X_1 + \beta_2 (X_2 + 3X_1) + \beta_3 X_3 + u$$

*Now, the explanatory variables are going to be X_1, $X_2 + 3X_1$ and X_3

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

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



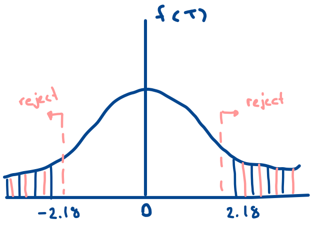
→ p - value = what should be the significant level, given the critical value of -2.75 ?
→ find the shaded area!

suppose the calculated $\boxed{t_{\hat{\beta}_j} = -2.75}$ >> $t_{\hat{\beta}_j} = \frac{(\hat{\beta}_j - \beta_j)}{s.e(\hat{\beta}_j)}$

- From the z-table, the value -2.75 corresponds to area = _ _ _0.03_ _ _ _ _ _ _ _ _ _.
- Thus, p-value = _ _ _0.03_ _ _ _ _ _ _.
- 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: $H_0 : \beta_j = a_j, \quad H_a : \beta_j \neq a_j, \quad$ d.f.= 18.



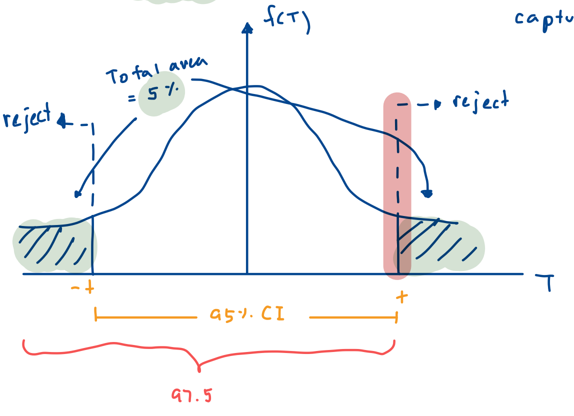
suppose the calculated $t_{\hat{\beta}_j} = -2.18$ >> use t-table

- From the t-table, the value -2.18 corresponds to area = _ _ _0.02_ _to_0.05_ _ _ _ _.
- Thus, p-value = _is_between_ 0.02 _to_0.05_.
- Would we reject H_0 if we use the significance level = 5%?

6 Confidence Intervals (CI)

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

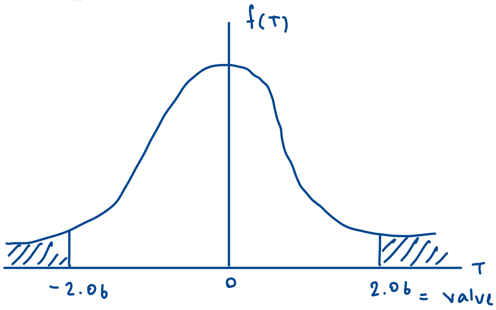
→ The range of values that would capture the true β_j at a 95% chance



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

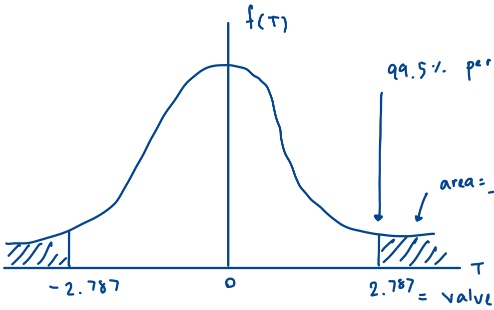
Example 1: **95% CI**



2.06 = value of 97.5th percentile in the t_{25} distribution

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**



99.5% percentile or area 1- 0.005 = 0.995

area = $\frac{1\%}{2} = \frac{0.01}{2} = 0.005$

2.787 = value of 99.5th percentile in the t_{25} distribution

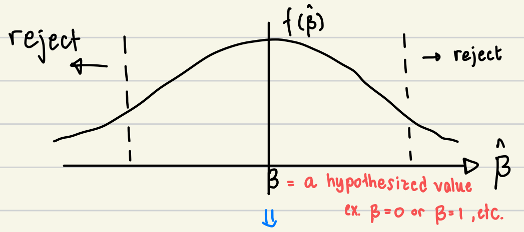
The 99% CI for $\hat{\beta}_j = [\hat{\beta}_j - 2.787 \cdot s.e(\hat{\beta}_j), \hat{\beta}_j + 2.787 \cdot s.e(\hat{\beta}_j)]$

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 hypotheses.

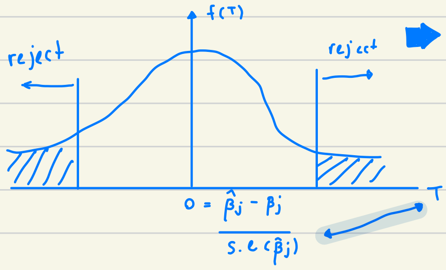


1) Test if β = some number

e.g. $\beta_j = 0 \rightarrow x_j$ has no impact on y.

$\beta_j = 1 \rightarrow$ 1 unit ↑ in X_j correspond to 1 unit ↑ in y.

⇓

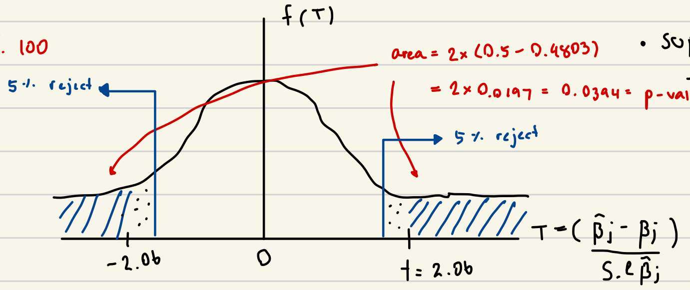


➡ t-test ⭑ => How??

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

$$\boxed{\text{Significant level = total area in the rejection region}}$$

ass. d.f. 100



area = $2 \times (0.5 - 0.4803)$

$= 2 \times 0.0197 = 0.0394 =$ p-value!

- Suppose, we calculate α t-statistic $= \frac{\hat{\beta}_j - \beta_j}{s.e.\,\hat{\beta}_j} = 2.06$
- Suppose we are testing $H_0: \beta_j = 0$, $H_a: \beta_j \neq 0$
  ↳ 2 tailed test
- p-value = total shaded area

$$\boxed{\begin{array}{l}\text{p-value = significant level which we will reject} \\ \text{the } H_0 \text{ or prob that we reject the } H_0\end{array}}$$

↳ If p-value < significant level => reject H_0

7 Testing Multiple Linear Restrictions: The F-test

Suppose the model is specified by

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

>> 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

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

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 the we would like to perform a joint-test of = 0

  e.g in this model $q = 2$

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

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

  (the last q βs = 0)

  Ha: H_0 is not true

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

$$F \equiv \frac{\dfrac{(\underbrace{SSR_r - SSR_{ur}})}{q}}{\dfrac{SSR_{ur}}{(n-k-1)}}$$

This is always (+) b.c. $SSR_{ur} < SSR_r$

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

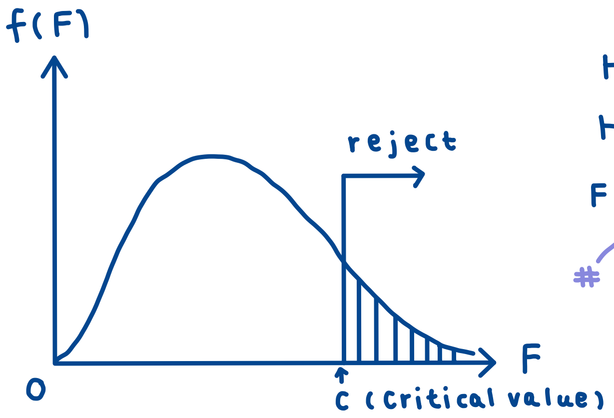
$(n-k-1)$ ↳ d.f. of the "ur" model

- 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, $var(\hat{\beta}_s)$ will increase, making the prediction of β less precise.
So, we only keep the addition x_s if it/they $\underbrace{\text{can improve the model enough}}$

can ↓ SSR (R^2↑) enough.

can significantly ↓ SSR and R^2↑



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

$H_a : H_0$ is not true.

$F \sim F_{q,\, n-k-1}$

d.f. of the ur model

# of joint Hypothesis being tested

3. Some useful facts

1) $R^2_{ur} > R^2_r$ because any additional x would increase R^2 (improve fit)

$\gg$ $SSR_{ur} < SSR_r$

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 does not improve the model enough.

4. Other ways to calculate the F-statistics:

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

$$\text{We have } F = \frac{\dfrac{(R^2_{ur} - R^2_r)}{\underbrace{q}_{\text{\# of } \beta \text{ that are set to "0"}}}}{\dfrac{(1 - R^2_{ur})}{\underbrace{n}_{\text{\# of obs}} - \underbrace{k}_{\text{\# of slope } \beta} - \overset{\text{intercept}}{1}}}$$

$\gg$ 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

$$F = \frac{\dfrac{R^2}{k}}{\dfrac{1-R^2}{n-k-1}}$$

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 { ur { *years* = years in major leagues
gamesyr = games per year in the league
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 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
       Model |  308.989208         5  61.7978416   Prob > F        =    0.0000
    Residual |  183.186327       347  .527914487   R-squared       =    0.6278
-------------+----------------------------------   Adj R-squared   =    0.6224
       Total |  492.175535       352  1.39822595   Root MSE        =    .72658

------------------------------------------------------------------------------
  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
q=3 { 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 intc.|   11.19242   .2888229    38.75   0.000     10.62435    11.76048
------------------------------------------------------------------------------
```

When considering each of the performance X 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 have joint impact.

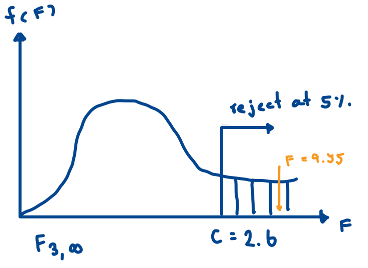
Now, our H_0 and H_a becomes

$$F \equiv \frac{\frac{SSR_r - SSR_{ur}}{q}}{\frac{SSR_{ur}}{n-k-1}}$$

$$\equiv \frac{\frac{198.311 - 183.86}{3}}{\frac{183.186}{353-5-1}} \approx 9.55$$

HW.

$$F \equiv \frac{\frac{R^2}{q}}{\frac{1-R^2}{n-k-1}} \equiv ?$$



since F = 9.55 > 2.6, we reject H_0 at 5% sig 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)

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

   z - table

   ⇒ If the d.f. > 30, then when $t \quad > 1.96$, we can reject H_0 with 5 % sig level

   ⇒ **When** $t \quad >$ **1.96**, we can say that β_i is **statistically significant** at 5% level. (value of $\beta_i \neq 0$)

   ⇒ **When** $t \quad <$ **1.96** we can say that β_i is **not statistically significant** at 5% level.

   ⇒ If $t \quad < 1.96$ we can drop x_i from the model

   ⇒ 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.

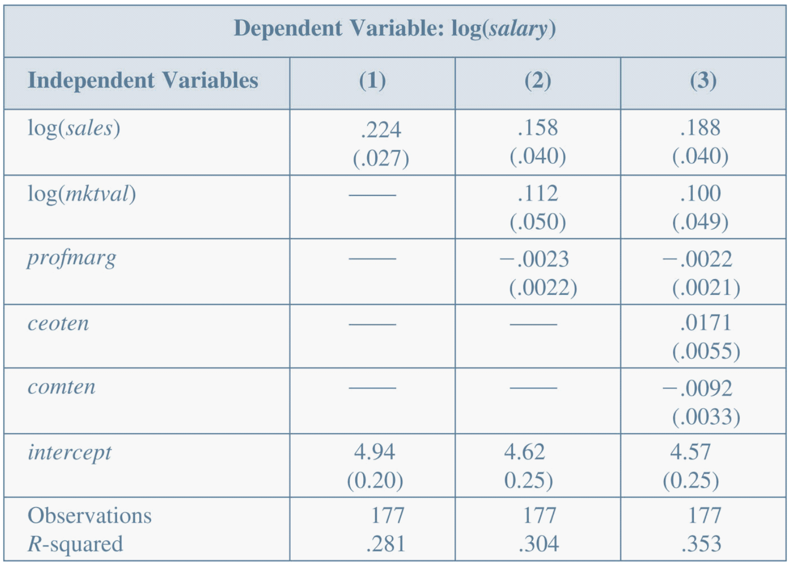
   or perform an F-test for testing multiple linear restrictions

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

sales

other company performance

CEO characteristic



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 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

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

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 b_{weight} in kilograms ?

$$1 \text{ kg} = 1000 \text{ g}$$

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

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

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

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

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

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

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

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 b_{weight} in kg & income in THB

$$b_{weight_{kg}} = \frac{\hat{\beta}_0}{1000} + \frac{\hat{\beta}_1}{1000} cigs + \frac{\left(\frac{\hat{\beta}_2}{1000}\right)}{1000} faminc_{THB}$$

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

2 More on functional forms

- Logarithmic Functional Form

$$\frac{d}{dx}c^x = c^x \ln c \quad (c>0)$$

$$\frac{d}{dx}e^x = e^x$$

$$\frac{d}{dx}\log_c x = \frac{1}{x \ln c} \quad (c>0,\ c\neq 1)$$

$$\frac{d}{dx}\ln x = \frac{1}{x} \quad (x>0) \longrightarrow \frac{d\ln X}{dX} = \frac{1}{X}$$

$$= d\ln(X) = \frac{1}{X}dX$$

$$\frac{d}{dx}\ln|x| = \frac{1}{x}$$

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

usually means natural log

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

$$\underbrace{\begin{matrix}\Delta Y = Y_1 - Y_2 \\ \Delta X_1 = X_{11} - X_{12}\end{matrix}}$$

$$\bullet\ \beta_1 = \frac{d\log(Y)}{d\log(X)} = \frac{\frac{1}{Y}dY}{\frac{1}{X_1}dX_1} = \frac{\frac{1}{Y}\Delta Y}{\frac{1}{X_1}\Delta X_1} = \frac{100\times\frac{1}{Y}\Delta Y}{100\times\frac{1}{X_1}\Delta X_1} = \frac{\%\,Y}{\%\,X}$$

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

(X, elasticity of Y)

(price) (demand)

$$\bullet\ \beta_2 = \frac{d\log(Y)}{dX_2} = \frac{\frac{1}{Y}dY}{dX_2} = \frac{\frac{1}{Y}\Delta Y}{\Delta X_2}$$

>> If we want the uppen term to be % change, then

$$100\beta_2 = \frac{100\frac{1}{Y}\Delta Y}{\Delta X_2}$$

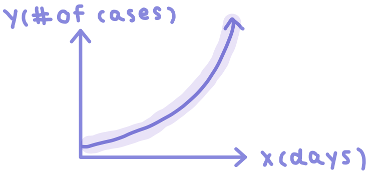
$$100\beta_2 = \frac{\%\,\Delta Y}{\Delta X_2}$$

$\therefore$ $100\beta_2$ = % in Y given that X_2 increases by 1 unit.

- Models with Quadratics (squares)

>> capture increasing / decreasing marginal effects (slope of the relationship between X & Y is not constant)

COVID-19 example

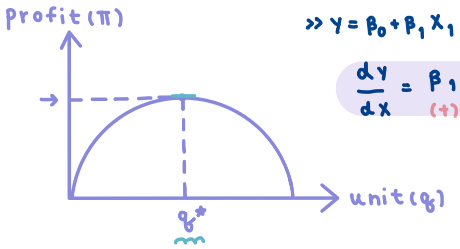


$$\gg Y = \beta_0 + \beta_1 X + \beta_2 X^2 + u$$

$$\frac{dY}{dX} = \underset{(+)}{\beta_1} + 2\underset{(+)}{\beta_2} X$$

↳ days

Decreasing returns.



$$\gg Y = \beta_0 + \beta_1 X_1 + \beta_2 X^2 + u$$

$$\frac{dY}{dX} = \underset{(+)}{\beta_1} + 2\underset{(-)}{\beta_2} X$$

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

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

$\text{F.o.C} = \frac{\partial\pi}{\partial q} = 0 = 90 - 2Q$

↗ β_1 is positive

β_2 is negative

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
The estimation result is given by

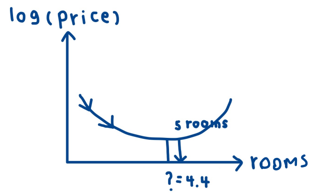
In the US or many other countries, students can apply to school in the area w/o having to take any test. So, the lower stratio, the better school

```
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

------------------------------------------------------------------------------
log price →  lprice |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
log (nox) →    lnox | β1 -.9767545   .0995938    -9.81   0.000    -1.172429   -.7810806
        dist | β2 -.0321972   .0094013    -3.42   0.001     -.050668   -.0137264
       rooms | β3 -.5528032   .1612965    -3.43   0.001    -.8697056   -.2359007
    rooms_sq | β4  .0624697   .0124867     5.00   0.000     .0379368    .0870025
     stratio | β5 -.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"

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



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

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

rooms = 4 4

ANS 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?

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

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

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

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

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

% Δ price = 100 (− 0.553 + 2(0.062) · 6) = 19.1 %

3 Models with Interaction Terms

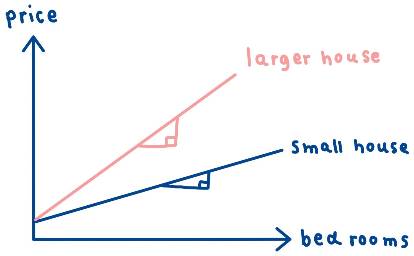
>> use when impact of one variable depends on the value (level) of another variable

Consider

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

where

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



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

>> 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 everytime we estimate a parameter

- using R^2 would not punish "having too many regressor"

- we use adjusted $-R^2$ or $\bar{R}^2$ when we want to punish adding too many regressor

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

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

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, $df = n-k-1 \downarrow$, $\frac{SSR}{n-k-1} \uparrow$, adj $R^2 \downarrow$

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

Consider Model 2

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

>> 27.5 % of variation in y is explined

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.

$$famale = \begin{cases} 1 & \text{if female} \\ 0 & \text{otherwise (or if male)} \end{cases}$$

$$married = \begin{cases} 1 & \text{if married} \\ 0 & \text{otherwise (of if single)} \end{cases}$$

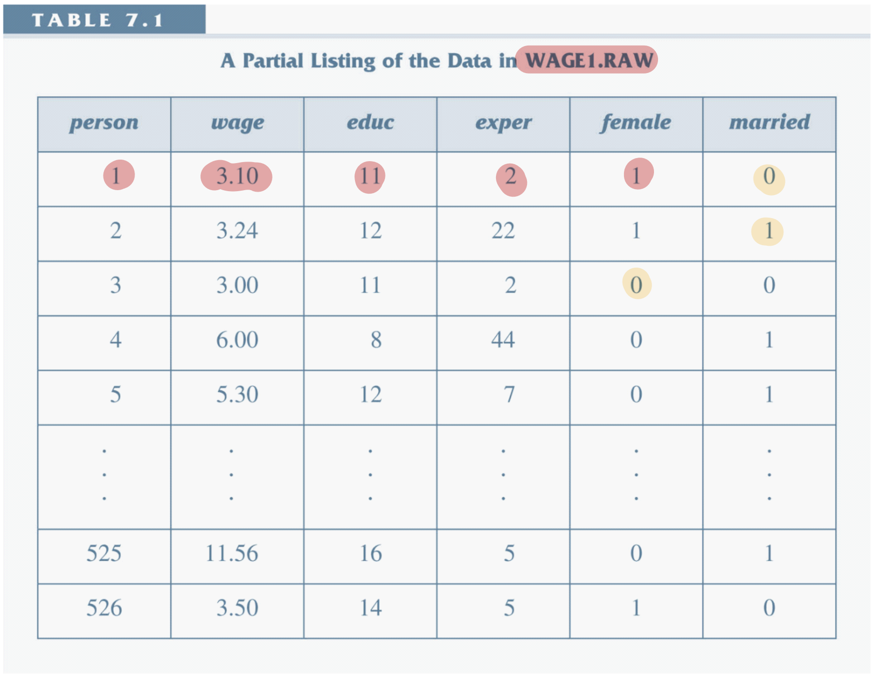


TABLE 7.1

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

$$wage = \beta_0 + \delta_0 \underline{female} + \beta_1 \underline{educ} + u.$$

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

$= \beta_0 + \delta female + \beta_1 edu + E(u \mid female, edu)$

$\downarrow = 0$ (ass MLR 1-4 holds)

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

2) Thus

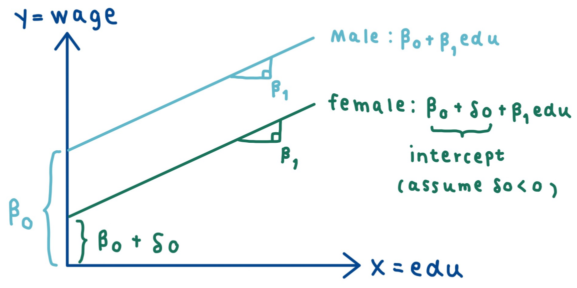
♀ : $E(wage \mid female = 1, edu) = \beta_0 + \delta_0(1) + \beta_1 edu = \beta_0 + \delta_0 + \beta_1 edu$

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

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

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

* given the same value of edu (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 edu

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:

$$wage = \beta_0 \underset{\substack{\uparrow \\ \text{Intercept} \times 1}}{x_0} + \delta_0 \underset{(x_1)}{female} + \beta_1 \underset{(x_2)}{edu} + \delta_1 \underset{(x_3)}{male} + u$$

id.	female	male	x_0
1	1	0	1
2	1	0	1
3	0	1	1
4	0	1	1
.	0	1	1
.	0	1	1
.	1	0	1
.	1	0	1
.			
.			
.			
.			
99			

$$X_0 = X_1 + X_3$$

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

or

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

$$\textcircled{1} = \overset{\checkmark}{winter} + \overset{\checkmark}{spring} + \overset{\checkmark}{summer} + \cancel{fall} \quad w$$
$$winter = 1 - spring - summer - fall$$

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

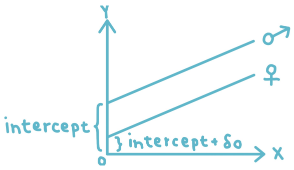
$$spring = \begin{cases} 1 & \text{if spring} \\ 0 & \text{otherwise} \end{cases}$$

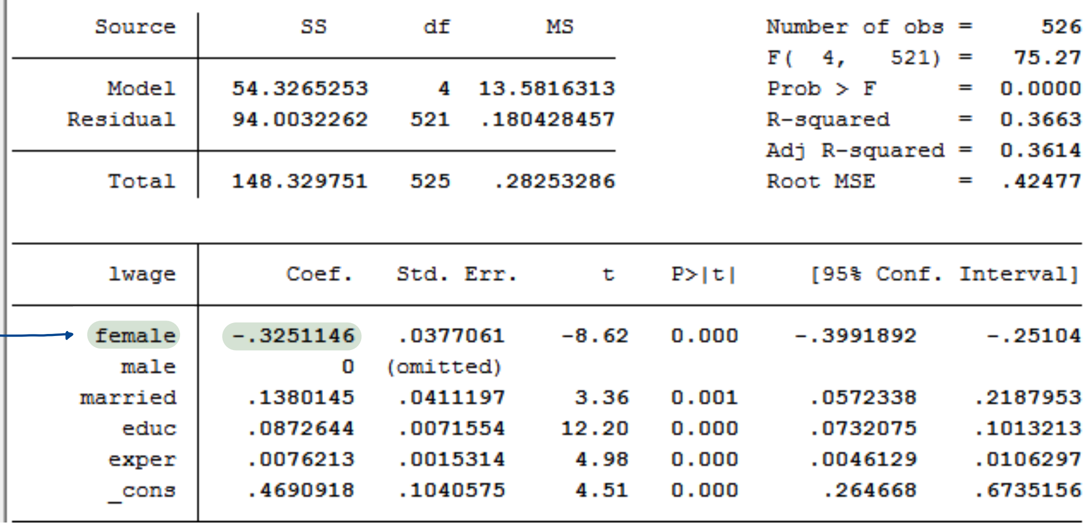
etc.

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
4	0	0	1	0	1
.	0	1	0	0	1
.	0	1	0	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

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





```
. 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 workers are expected to have less wage compared to male worker

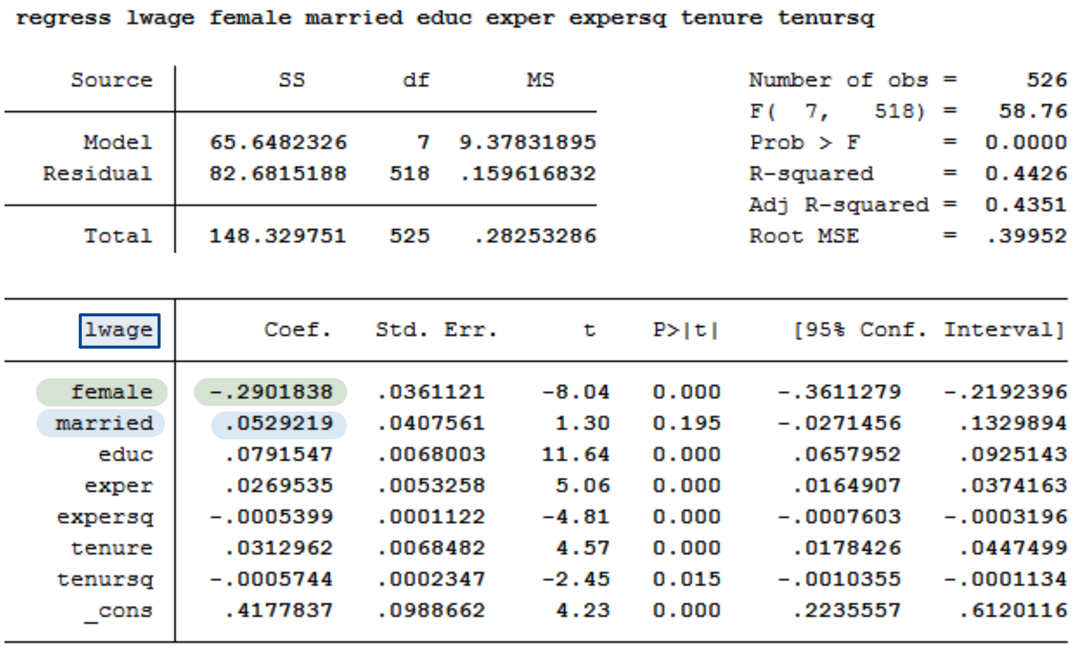
5 Using dummy variables for multiple categories

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

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

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

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



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

      Source |       SS       df       MS              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
------------------------------------------------------------------------------
```

Comments:

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

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

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

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

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

2) δ_1 measures the impact of be married

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

	♀	♂
marr	marrfem	marrmale
sing	singfem	singmale

because

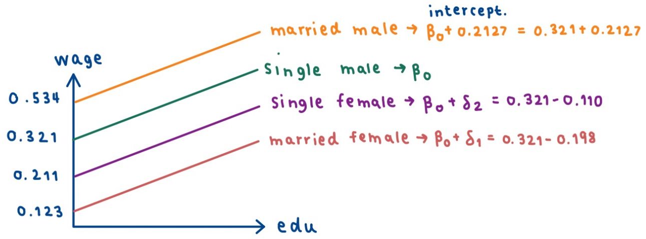
Consider a model which includes dummy variables for each gender/marital status combination– $marrmale$, $marrfem$ and $singfem$. (or singmale ⟵ used as the basecase)

$$\begin{aligned}\log(wage) \;=\;& \beta_0 + \delta_0 marrmale + \delta_1 marrfem + \delta_3 singfem + \beta_1 \underline{educ} + \beta_2 \underline{exper} \\ & + \beta_3 \underline{exper}^2 + \beta_4 \underline{tenure} + \beta_5 \underline{tenure}^2 + u. \qquad (8.1)\end{aligned}$$

```
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]
-------------+----------------------------------------------------------------
δ0  marrmale |   .2126757   .0553572     3.84   0.000      .103923    .3214284
δ1   marrfem |  -.1982676   .0578355    -3.43   0.001     -.311889   -.0846462
δ2   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
β̂  { 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 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 males, holding other factors constant.
- δ_2 → Some rationale



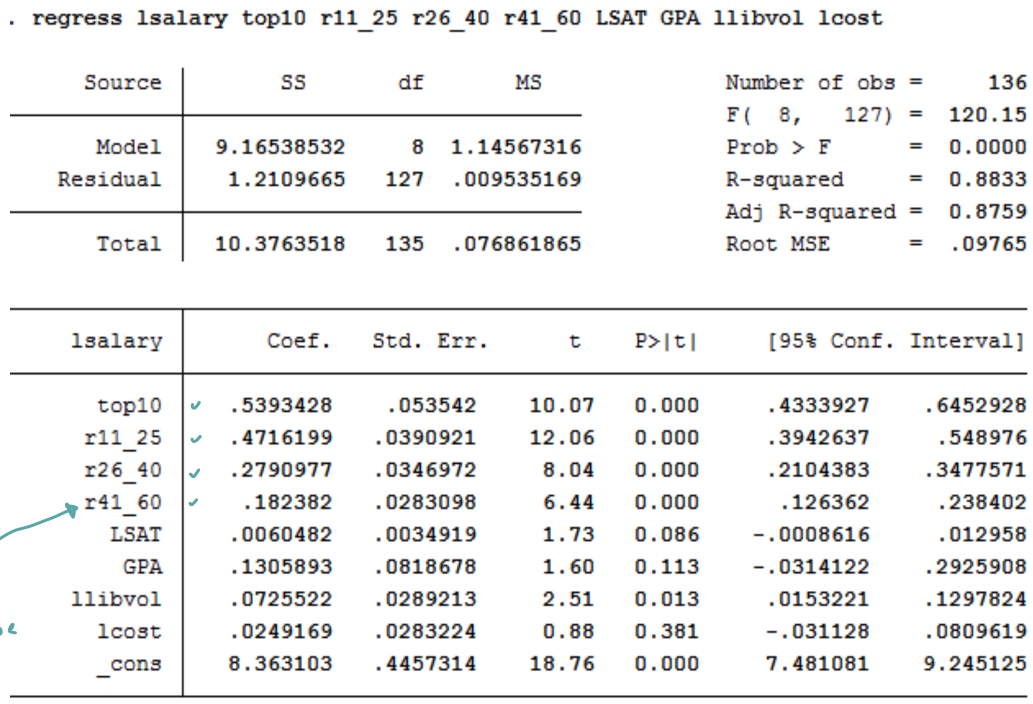
Case 2 We can use dummy variables to represent multiple categories of a variable Consider the relationship between law school rankings and starting salaries

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

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 case the "range of the value" Serve as a better explanatory variable than the "value itself" eg. age may explain the model better if split into generation young 0-15 gen 2 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 61th and worse

Comments:

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

etc.

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 ranked 61th and worse.

2) δ_1 → use the same rationale.