# 4 Testing Hypotheses about a Single Linear Combination of the Parameter

Consider

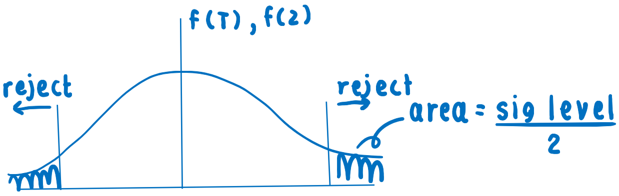
$$\log(wage) = \beta_0 + \beta_1 jc + \beta_2 univ + \beta_3 \exp er + u$$

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. if the return from 1 more years of at a junior college
We want to test whether $\beta_1 = \beta_2$. ⇝ is the same as that of the university

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

$H_0 : \beta_1 \neq \beta_2 \Rightarrow H_a : \beta_1 - \beta_2 \neq 0$

2-tailed test




$$t = \frac{(\hat{\beta}_1 - \hat{\beta}_2) - 0}{s.e.(\hat{\beta}_1 - \hat{\beta}_2)}$$ we compute this t-statistic and compare with the critical value

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

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

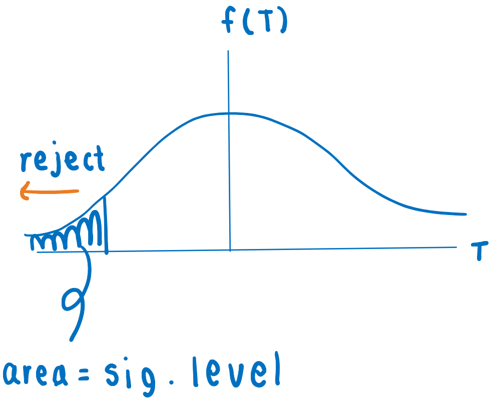
not very straight forward to calculate
normally we use a variable transformation tricks
└ see notes

another possible hypothesis test (one-tailed alternative)

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

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

- It is assume that $\beta_1$ will not be more than $\beta_2$ so returns to junior college will never be more than returns to university education



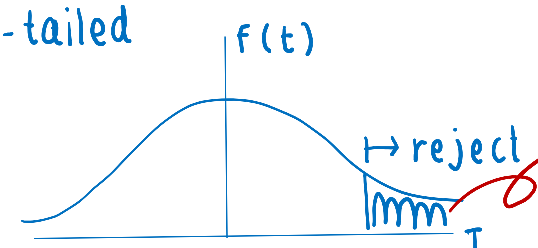

$$t = \frac{(\hat{\beta}_1 - \hat{\beta}_2) - 0}{s.e.(\hat{\beta}_1 - \hat{\beta}_2)}$$

* Then, go to the extra note.

## 5 Computing p-Values for t-Tests

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

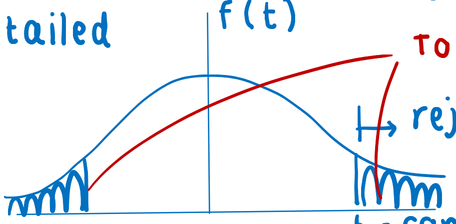
1-tailed




$t = \frac{\hat{\beta}_j - \beta_j}{s.e.\ \hat{\beta}_j}$

This shaded area in the rejection region is the p-value

2-tailed




Total area from 2 sides

t = computed $t = \frac{\hat{\beta}_j - \beta_j}{s.e.\ \hat{\beta}_j}$

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

T = t-distributed random variable with d.f. = n-k-1

t = computed t-stat

⇝ p-value = probability that a random T value will be greater in term of absolute than our t in $H_0$ test

## In-class exercise

consider the multiple reggression model, assume MLR 1-6 are satisfied

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

$H_a$ : otherwise is true

you would like to test the $H_0 : \beta_1 - 3\beta_2 = 1$

1) write the t-statistic for testing $H_0$

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

2) Define $\theta_1 = \hat\beta_1 - 3\hat\beta_2 \Rightarrow H_0 : \theta_1 = 1$, $H_a : \theta_1 \neq 1$

$$t = \frac{\hat\theta_1 - 1}{s.e.(\theta_1)}$$

$\Rightarrow$ we need our reggression to have $\theta_1$ in it. So, STATA or the OLS estimation will give $\hat\theta_1$ & se $\theta_1$

Now, $\hat\beta_1 = \hat\theta_1 + 3\hat\beta_2$

$\beta_1 = \theta_1 + 3\beta_2$

substitute in the main reggression and get

$$y = \beta_0 + (\theta_1 + 3\beta_2)x_1 + \beta_2 x_2 + \beta_3 x_3 + u$$

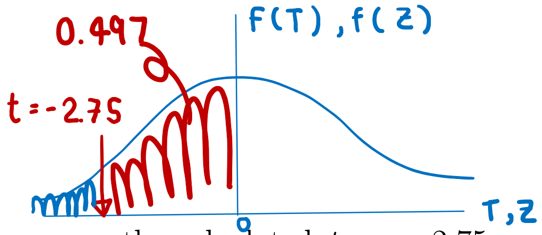
$$= \beta_0 + \theta_1 x_1 + 3\beta_2 x_2 + \beta_2 x_2 + \beta_3 x_3 + u$$

$$= \beta_0 + \theta_1 x_1 + \beta_2 (x_2 + 3x_1) + \beta_3 x_3 + u$$

☆ Now the explanatory variable 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: $H_0 : \beta_j \geq 0$, $H_a : \beta_j < 0$, d.f.= 140. ⤳ z-table




⤳ p-value = what should be the sig level given the critical value of -2.75 ⇒ find the shaded area!

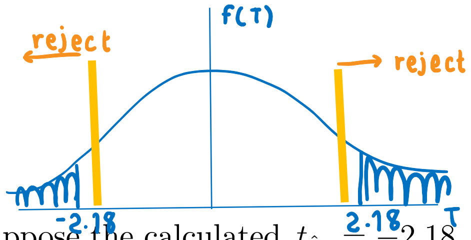
suppose the calculated $t_{\hat{\beta}_j} = -2.75$

$t_{\hat{\beta}_j} = (\hat{\beta}_j - \beta_j)/s.e.(\hat{\beta}_j)$

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

☆ RULE we reject $H_0$ if p-value < sig. level

Example 2: $H_0 : \beta_j = a_j$, $H_a : \beta_j \neq a_j$, d.f.= 18. ⤳ use t-table



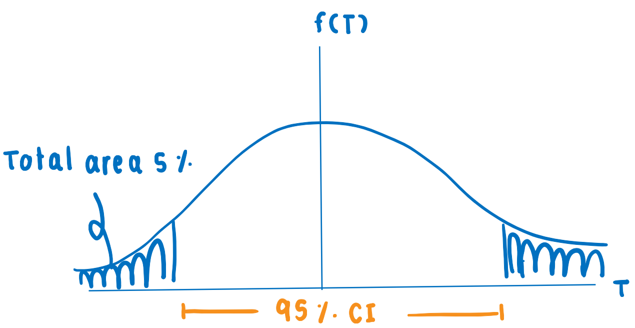

suppose the calculated $t_{\hat{\beta}_j} = -2.18$

- From the t-table, the value -2.18 corresponds to area = 0.02 to 0.05.
- Thus, p-value = is between 0.02-0.05.
- Would we reject $H_0$ if we use the significance level = 5%?

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

# 6 Confidence Intervals (CI)

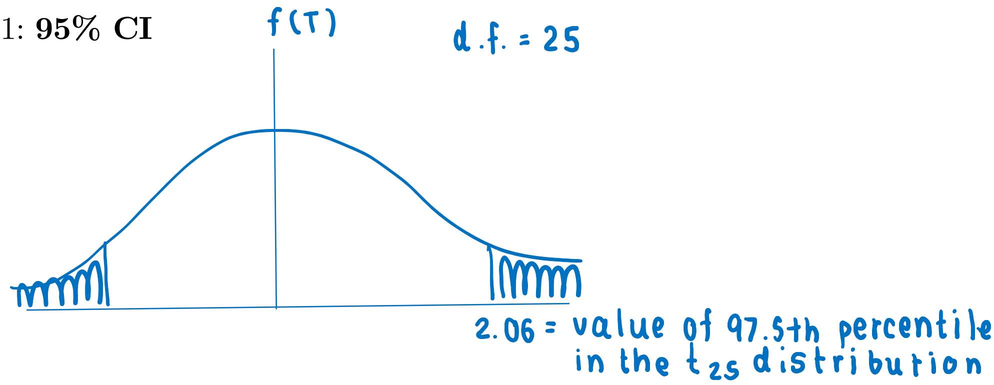
- Confidence Intervales for the POPULATION PARAMETER $(\beta_j)$
  └ The range of values that would capture the true $\beta_j$ at a 95% chance
- A 95% CI of $\beta_j$ is given by




CI ⇒ $\hat{\beta}_j \pm c \cdot s.e.(\hat{\beta}_j)$

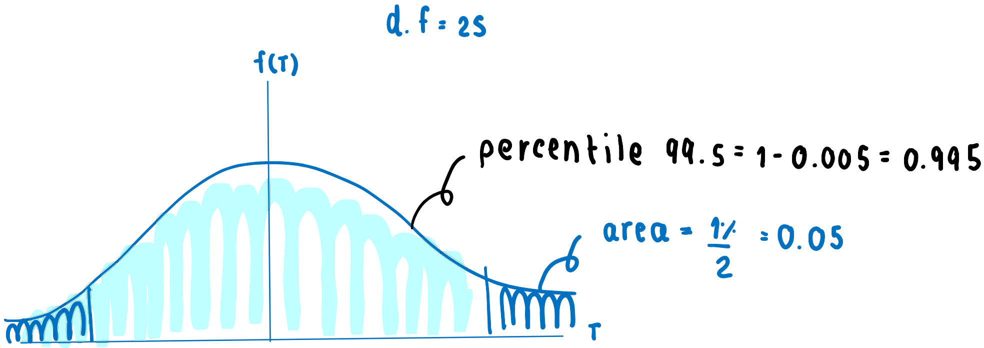
c is the 97.5 percentile in the t distribution with n-k-1 d.f

Example 1: **95% CI**




The 95% CI for $\hat{\beta}_j = [\hat{\beta}_j - 2.06 \cdot se(\hat{\beta}_j), \hat{\beta}_j + 2.06 \cdot se(\hat{\beta}_j)]$

Example 2: **99% CI**



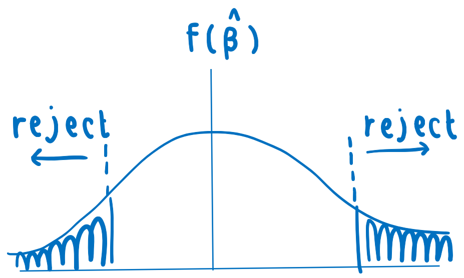

The 99% CI for $\hat{\beta}_j = [\hat{\beta}_j - 2.787 \cdot se(\hat{\beta}_j), \hat{\beta}_j + 2.787 \cdot se(\hat{\beta}_j)]$

# Inference >> Hypothesis testing abt "$\beta$" the true parameter

$$wage = \beta_0 + \beta_1 educ + \beta_2 experience + \dots + u$$

we want to test hypothesis about the true impact of ($\beta$) of each x variables (educ, experience) on the dependent variable (y)

but we don't know what true $\beta$ are so, we use $\hat{\beta}$ and s.e.($\hat{\beta}$) to test the hypothesis

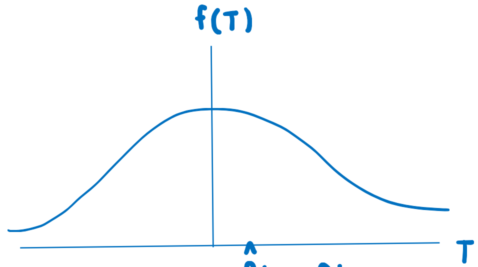



$\beta$ = a hypothesize value
ex) $\beta = 0$ or $\beta = 1$, etc.

1) Test if $\beta$ = same numbers

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



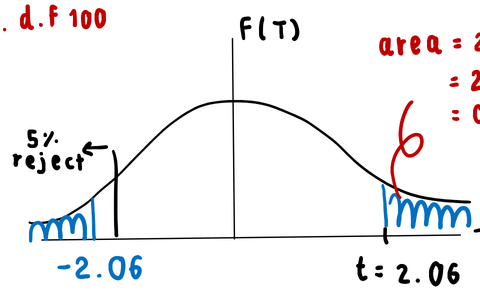

$$\frac{\hat{\beta}_j - \beta_j}{s.e.(\hat{\beta}_j)}$$

⇒ t-test ☆ ⇒ How??

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

Significant level = total area in the rejection region

ass. d.F 100




area = 2·(0.5 − 0.4803)
= 2 · 0.197
= 0.0394 → p-value

$$T = \frac{\hat{\beta}_j - \beta_j}{s.e.(\hat{\beta}_j)}$$

- Suppose we calculate a t-static $= \dfrac{\hat{\beta}_j - \beta_j}{s.e.(\hat{\beta}_j)} = 2.06$
- suppose, we are testing $H_0: \beta_j = 0$, $H_a: \beta_j \neq 0$
  └ 2-tailed
- p-value = total shaded area

p-value = significant level which we will reject the $H_0$ on the prob that we will reject $H_0$

* if p-value < significant level ⇒ reject $H_0$

## F-test motivation

(multiple hypothesis)

⇒ we want to test the significance of a group of hypothesis

$$Grade_{325} = \beta_0 + \beta_1 \#time\_front + \beta_2 \#time\_back + \beta_3 hr\_study + \beta_4 past\_GPA + \beta_5 gender + u$$

$H_0$ : seat position doesn't have impact on $GPA_{325}$

$\beta_1 = 0$ and $\beta_2 = 0$ ⇒ $\beta_1 = \beta_2 = 0$

$H_a$ : seat position matters

$$\left.\begin{array}{l} \beta_1 \neq 0 \text{ and } \beta_2 \neq 0 \text{ or} \\ \beta_1 \neq 0 \text{ and } \beta_2 = 0 \text{ or} \\ \beta_1 = 0 \text{ and } \beta_2 \neq 0 \end{array}\right\} \text{at least one of the } \beta_1, \beta_2 \neq 0$$

# 7 Testing Multiple Linear Restrictions: The F-test

Suppose the model is specified by

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

→ 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 <u>true</u>

   ⇒ Reject $H_0$

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

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 <u>true</u> ⇒ do not reject $H_0$

- suppose there are "q" number of $\beta$ that we want to perform a joint-test of = 0
  e.g. in this model $q = 2$

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

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

(the last $q$ $\beta$s $= 0$)

$H_a$ : $H_0$ is not true

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

$u(r)$

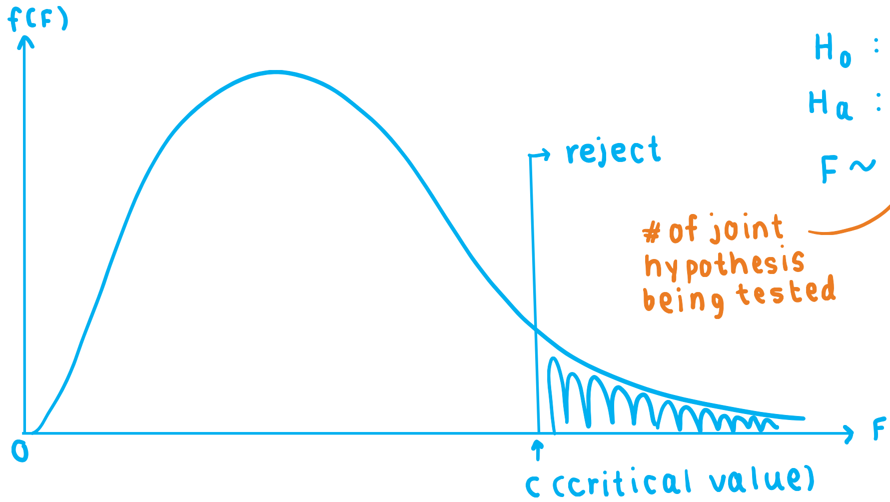
$$F = \frac{\dfrac{(SSR_r - SSR_{ur})}{q}}{\dfrac{SSR_{ur}}{(n-k-1)}}$$

This is always (+) b/c $SSR_{ur} < SSR_r$ everytime u add 1 more x the model will be better explained

d.f of the ur model

• So, if everytime you add one more x variable, the SSR ↓ and $R^2$ ↑, why don't we just keep additional x in the model ??

» Because every time we add 1 more x the var($\hat{\beta}_s$) will increase, making the prediction of β less precise. So, we only keep the additional xs if it can improve the model enough → can reduce SSR or ↑ $R^2$ enough




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

$H_a : H_0$ not true

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

# of joint hypothesis being tested

d.f. of the ur model

we reject $H_0$ of jointly no effect if $F > c$

3. Some useful facts

① $R^2_{ur} > R^2_r$ because additional x would increase $R^2$ (improve fit)
$\Rightarrow SSR_{ur} < SSR_r$

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:

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

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

q: # of β that are set to "0"
1: intercept
k: # of slope β
n: # of obs

$\Rightarrow$ if we want to test the overall significance of the model

$H_0: \beta_1 = \beta_2 = \beta_3 = \ldots = \beta_k = 0$, $H_a$ = Otherwise

$F \equiv \dfrac{R^2/k}{(1-R^2)/(n-k-1)}$

$R^2$ of the model ≈ UR
the "r" model has no x at all

**Example:** Suppose we are interested in understanding the determinant of a baseball player's salary.

| | | |
|---|---|---|
| y | *salary* | = season salary |
| r, ur | *years* | = years in major leagues |
| r, ur | *gamesyr* | = games per year in the league |
| ur | *bavg* | = career batting average |
| ur | *hrunsyr* | = homeruns per year |
| ur | *rbisyr* | = runs batted in per year |

performances

If we want to test whether performance has any impact on salary

$H_0: \beta_{bavg} = \beta_{hrunyr} = \beta_{rbisyr} = 0$

$H_a$: otherwise is true

- the unrestricted model (ur) is defined by ur model

. regress log_salary years gamesyr bavg hrunsyr rbisyr

·In multiple regress use adjusted r

| | Source | SS | df | MS |
|---|---|---|---|---|
| SSE | Model | 308.989208 | 5 | 61.7978416 |
| SSR | Residual | 183.186327 | 347 | .527914487 |
| SST | Total | 492.175535 | 352 | 1.39822595 |

| | | |
|---|---|---|
| Number of obs | = | 353 |
| F( 5, 347) | = | 117.06 |
| Prob > F | = | 0.0000 |
| R-squared | = | 0.6278 |
| Adj R-squared | = | 0.6224 |
| 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 |
| 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 | 11.19242 | .2888229 | 38.75 | 0.000 | 10.62435 | 11.76048 |

when considering each of the performance non of them has a significant impact at 5%

- the restricted model (r) is defined by

. regress log_salary years gamesyr

• But when performing an F-test performances have joint impact

| Source | SS | df | MS |
|---|---|---|---|
| Model | 293.864058 | 2 | 146.932029 |
| Residual | 198.311477 | 350 | .566604221 |
| Total | 492.175535 | 352 | 1.39822595 |

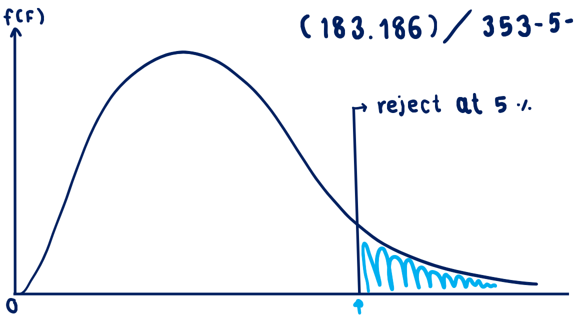
| | | |
|---|---|---|
| Number of obs | = | 353 |
| F( 2, 350) | = | 259.32 |
| Prob > F | = | 0.0000 |
| R-squared | = | 0.5971 |
| Adj R-squared | = | 0.5948 |
| Root MSE | = | .75273 |

| 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 |

Now, our $H_0$ and $H_a$ becomes

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

$$= \frac{(198.311 - 183.186)/3}{(183.186)/353-5-1} \approx 9.55$$




c = 2.6

$F_{3,\infty}$

ไปดู F table pg.966
แกน x เป็น q = 3
แกน y เป็น n-k-1 = ∞ greater than 200

since F = 9.55 > 2.6, we reject $H_0$ at 5% level and conclude that joint effects on salary

## 8 How the Hypothesis Testing is done in Practice

1. Check the values of $t - statistic$ reported by the statistical software (i.e. STATA, SPSS, SAS)

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

   z-table

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

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

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

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

   $\Rightarrow$ After we drop $x_i$, we estimate the new regression function and obtain a new set of $\hat{\beta}$.

2. We can also perform other hypothesis testings of interest.

   e.g. $H_0 : \beta_i = \beta_j$

   or $\quad H_0 : \beta_i = 5 \quad$ etc.

   or perform an F-test for testing multiple linear restrictions

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

sales

company performance

CEO characteris-

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

like simple reggression with 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

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

where
$bwght =$ child birth weight, in grams.
$cigs =$ number of cigarettes smoked by the mother while pregnant, per day.
$fa\min c =$ annual family income, in thousands of dollars.

- what if we use bweght in kilograms ??

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

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

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

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

- what if we used faminc in USD instead of 1000 USD

$$bweight_g = \hat{\beta}_0 + \hat{\beta}_1 cigs + \frac{\hat{\beta}_2}{1000} faminc_{USD}$$

The value of this variable is going to be 1,000 times larger than faminc

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

$$\Rightarrow \quad \hat{\theta}_2 = \frac{\hat{\beta}_2}{1000}$$

in other words $\hat{\theta}_2$ = impact of 1 USD ↑ in income
$\hat{\beta}_2$ = impact of 1,000 USD ↑ in income

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

$$bweight_{kg} = \frac{\hat{\beta}_0}{1000} + \frac{\hat{\beta}_1}{1000} cigs + \frac{\hat{\beta}_2}{\frac{1000}{30{,}000}} faminc_{THB}$$

this value going to be 30,000 times larger than faminc

## 2 More on functional forms

- Logarithmic Functional Form

note :
$\frac{d\ln x}{dx} = \frac{1}{x} \Rightarrow \boxed{d\ln(x) = \frac{1}{x}dx}$

usually means natural log

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

$\Delta y = y_1 - y_2$
$\Delta x = x_1 - x_2$

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

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

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

$\Rightarrow$ if we want the upper term to be % change, then

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

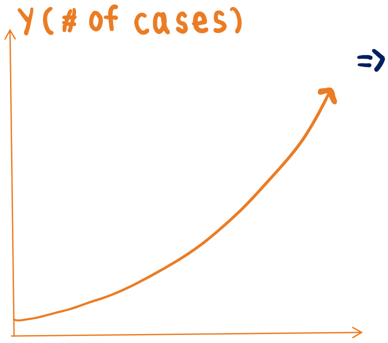
$100\beta_2 = \%\Delta$ in y given that $x_2$ increases by 1 unit

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

- Models with Quadratics (squares)

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

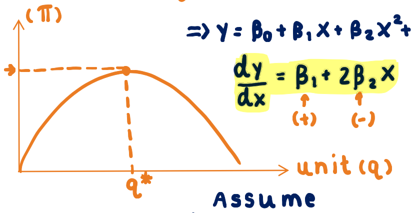
COVID-19 examples




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

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

Decreasing returns




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

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

Assume
$\pi = (P - MC)q$ ; $MC = 10$
Demand $P = 100 - q$

$= (100 - q - 10)q$

F.O.C

$\frac{\partial \pi}{\partial q} = 0 = 90 - 2Q$
(+) (-)

**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 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 |

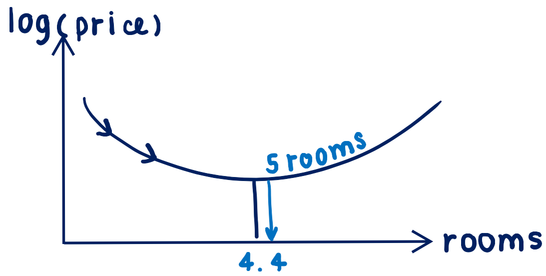
| | |
|---|---|
| Number of obs = | 506 |
| F( 5, 500) = | 155.62 |
| Prob > F = | 0.0000 |
| R-squared = | 0.6088 |
| Adj R-squared = | 0.6049 |
| Root MSE = | .25725 |

| 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 |

$|t| > 1.90$ ↑ ↑ all $< 0.05$
⇝ all variables are sig

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) ??

$0 = -0.553 + 2(.062)rooms)$

$rooms = 4.4$

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?

- $\frac{d\log(price)}{d\,rooms} = -0.533 + 2(0.062)\cdot rooms$

$$\frac{100 \cdot \frac{1}{price} \cdot d\,price}{d\,rooms} = 100 \cdot (-0.533 + 2(0.062)\cdot 5)$$

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

what about the change inprice if # rooms increase from 5 to 7 ??

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

$= 19.1\%$

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

## 3 Models with Interaction Terms

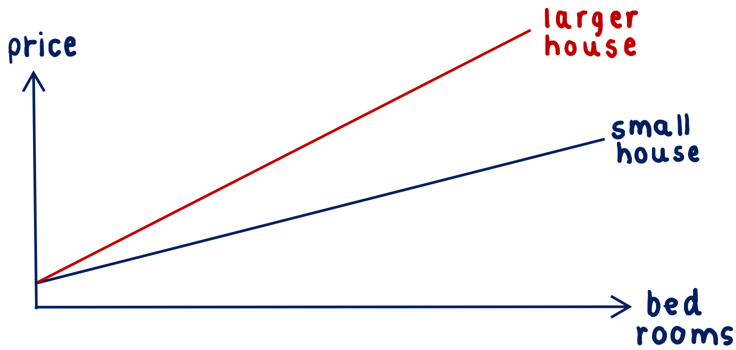
⇒ used when the impact of 1 variable depends on the value (level) of another variable

Consider

$$price = \beta_0 + \beta_1 \underbrace{sqrft}_{x_1} + \beta_2 \underbrace{bdrms}_{x_2} + \beta_3 \overbrace{\underbrace{sqrft}_{x_1} \times \underbrace{bdrms}_{x_2}}^{x_3} + \beta_4 \underbrace{bthrms}_{x_4} + u$$

where

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




$$\frac{dPrice}{dbdrms} = \beta_2 + \beta_3 sqrft$$

⇒ 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 increases

- But we lose the "degree of freedom"
  (d.f. = free data point used to estimate the parameter)
  ⇝ one data point is sacrificed everytime we estimate a parameter

- Using $R^2$ would not punish "having to many regressor"

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

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

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

→ If we have more $k$, $d.f = n-k-1$ ↓, $SSR/(n-k-1)$ ↑ and then adj $R^2$ ↓

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

Consider Model 1

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

Consider Model 2

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

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.

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

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

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

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

② Thus

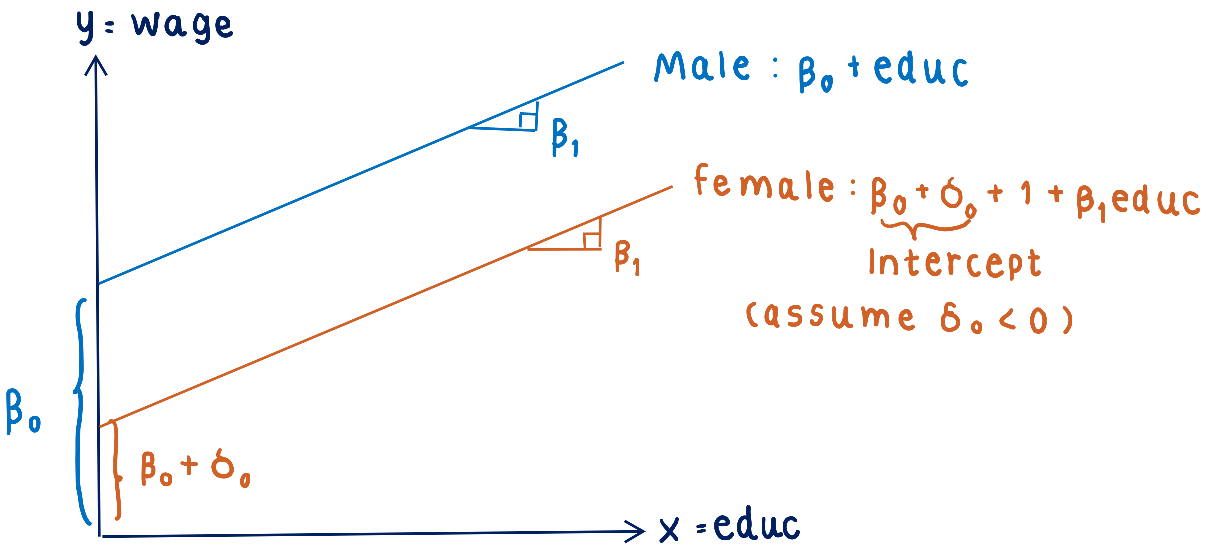
♀ : $E(wage|female = 1, educ) = \beta_0 + \delta_0(1) + \beta_1 educ = \beta_0 + \delta_0 + \beta_1 educ$

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

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

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

* given the same value of educ $\delta_0$ is the different 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. can only be 1 at a time

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

$x_0$: intercept × 1; female $(X_1)$; educ $(X_2)$; male $(X_3)$

For example:

$$X_0 = X_1 + X_3$$

$$1 = female + male$$

$$female = male + 1$$

| Id | female | male | $X_0$ |
|---|---|---|---|
| 1 | 1 | 0 | 1 |
| 2 | 1 | 0 | 1 |
| 3 | 0 | 1 | 1 |
| 4 | 0 | 1 | 1 |
| ⋮ | ⋮ | ⋮ | 1 |
| ⋮ | ⋮ | ⋮ | ⋮ |
| 99 | | | 1 |

or

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

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

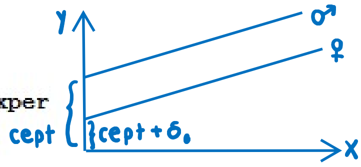
$$winter = 1 - spring - summer - fall$$

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

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

- 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

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




```
. 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 r expected to hv less wage compare to male

## 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*.

female: 1 female, 0 other

married: 1 married, 0 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 |

| 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 do not reject $H_0$ of no impact

Comments:

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

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

• female workers are expected to earn less than male workers by 29.02%, ceteris paribus

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

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

| | ♀ | ♂ |
|---|---|---|
| marr | marrfem | marrmal |
| sing | singfem | singmal |

Basecase



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
```

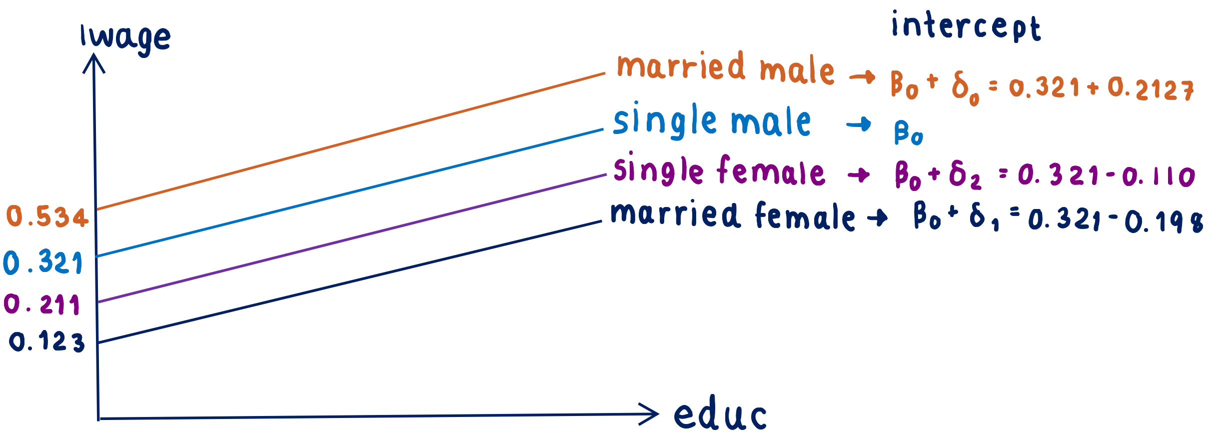
| Source | SS | df | MS |
|---|---|---|---|
| Model | 68.3617623 | 8 | 8.54522029 |
| Residual | 79.9679891 | 517 | .154676961 |
| Total | 148.329751 | 525 | .28253286 |

| | |
|---|---|
| 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] |
|---|---|---|---|---|---|---|---|
| $\delta_0$ | marrmale | .2126757 | .0553572 | 3.84 | 0.000 | .103923 | .3214284 |
| $\delta_1$ | marrfem | -.1982676 | .0578355 | -3.43 | 0.001 | -.311889 | -.0846462 |
| $\delta_2$ | singfem | -.1103502 | .0557421 | -1.98 | 0.048 | -.219859 | -.0008414 |
| $\hat{\beta}$ | educ | .0789103 | .0066945 | 11.79 | 0.000 | .0657585 | .092062 |
| | exper | .0268006 | .0052428 | 5.11 | 0.000 | .0165007 | .0371005 |
| | expersq | -.0005352 | .0001104 | -4.85 | 0.000 | -.0007522 | -.0003183 |
| | tenure | .0290875 | .006762 | 4.30 | 0.000 | .0158031 | .0423719 |
| | tenursq | -.0005331 | .0002312 | -2.31 | 0.022 | -.0009874 | -.0000789 |
| | _cons | .3213781 | .100009 | 3.21 | 0.001 | .1249041 | .5178521 |

Comments: This regression is not the same as the previous one. It uses "single male" as the base group (The previous one use male & single as 2 base group)

- $\delta_0$ measure the expected difference in wage of married male as compared with single males, holding other factors constant
- $\delta_1$ measures the expected difference in wage of married female as compared with single males, ceteris paribus
- $\delta_2$ ⇝ 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

1-10 : 1
other : 0

11-25 : 1
other : 0

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

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 e.g. age maybe explain the model better if split into generations.

young 0-15
gen z 15-29

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

| Source | SS | df | MS |
|---|---|---|---|
| Model | 9.16538532 | 8 | 1.14567316 |
| Residual | 1.2109665 | 127 | .009535169 |
| Total | 10.3763518 | 135 | .076861865 |

| | |
|---|---|
| Number of obs | = 136 |
| F( 8, 127) | = 120.15 |
| Prob > F | = 0.0000 |
| R-squared | = 0.8833 |
| Adj R-squared | = 0.8759 |
| 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 61 and worse

Comments:

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) those who graduate from the school ranked 61th and worse

2) $\delta_1$ ⤳ same rationale top 11-25