2.2 *Testing Against One-Sided Alternatives*

Suppose the rule for rejecting H_0 becomes

eg. years of edu has a positive impact on wage.

$$H_a : \beta_j > 0.$$

We could have the H_0 as

like in the hw 2 practice

$$H_0 : \beta_j \leq 0 \quad \text{or}$$
$$H_0 : \beta_j = 0$$

depending on the context. For $H_0 : \beta_j = 0$, it implies that we are ruling out population values of β_j less than zero. For $H_0 : \beta_j \leq 0$, it implies that we are not ruling out population values of β_j less than zero. To given an example of a test against the one-sided alternative hypothesis, consider the wage equation again :

want to test if $\beta_2 > 0$

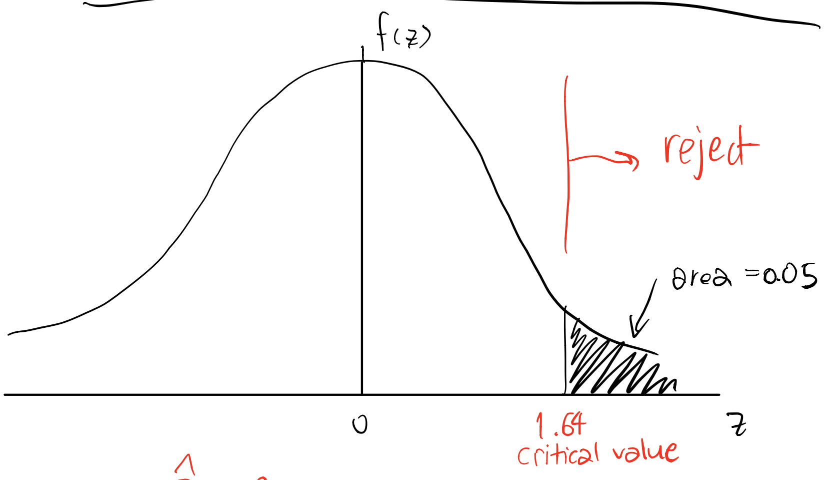
$$\log(wage) = \beta_0 + \beta_1 educ + \beta_2 \exp er + \beta_3 tenure + u$$

If we want to test whether experience has a positive partial effect on wage

suppose $n = 526$

$$H_0 : \beta_2 = 0$$
$$H_a : \beta_2 > 0.$$

We implicitly rule out the possibility that $\beta_2 < 0$ here.



either $H_0: \beta_2 = 0$

or $H_a: \beta_2 \leqslant 0$

will be fine

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

$= 526-3-1 = 522 \Rightarrow$ use z

→ let's test with 5% significant level.

$$z = \frac{\hat{\beta}_2 - \beta_2}{se.\ \hat{\beta}_2} = \frac{0.004}{0.0017} \approx 2.39 > 1.64$$

So, reject H_0 at 5% sig. level

→ exp. has a positive impact on wage!

$wage = \beta_0 + \beta_1 educ + \beta_2 exper + \ldots + u.$

3 Testing other hypotheses about β_j

- Most common H_0 is $H_0 : \beta_j = 0$. Test whether X_j has a statistically significant impact on y.

- However, we can test other types of hypotheses. For example, $H_0 : \beta_j = a_j$ where a_j is a constant number.

$$t_{d.f.} = \frac{\hat{\beta}_j - a_j}{s.e.\ \hat{\beta}_j} = \frac{\text{estimate} - \text{hypothesized value}}{s.e.}$$

$= n-k-1$

Example : husband income $= \hat{\beta}_0 + \hat{\beta}_1$ wife income + other factors.

- If we want to test whether 1 point increase in wife's income corresponds to 1 point increase in husband income, we test..

$H_0 : \beta_1 = 1$

$H_a : \beta_1 \neq 1$

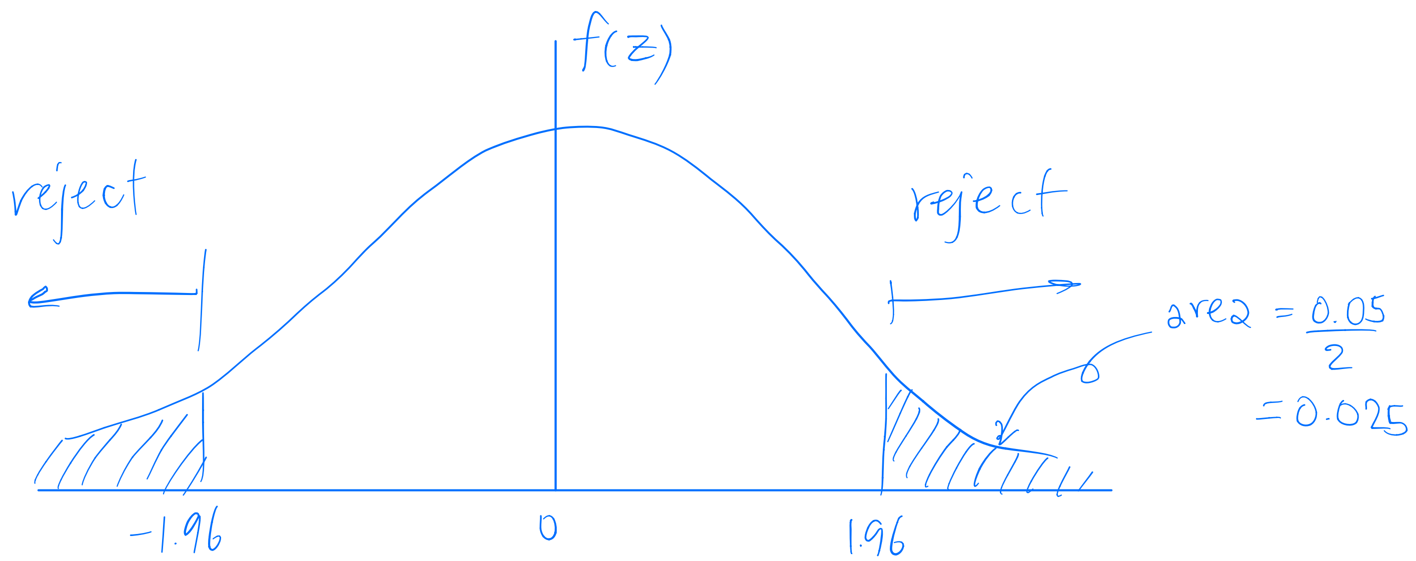
$$t_{wife\text{-}income} = \frac{\hat{\beta}_1 - 1}{s.e.\ \hat{\beta}_1}$$

- We test hypothesis about the population parameters ONLY. So, the β_j in the H_0, H_a cannot be $\hat{\beta}_j$
- This is a 2-Tailed test.
- The = sign is always with H_0!

Suppose d.f. = 250 ⇝ use z-table.

If we want to test the H_0 using a 5% significance level, then we reject H_0 if

$t_{wife\_income} >$ 1.96 or
$t_{wife\_income} <$ -1.96



4 Testing Hypotheses about a Single Linear Combination of the Parameter

Consider

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

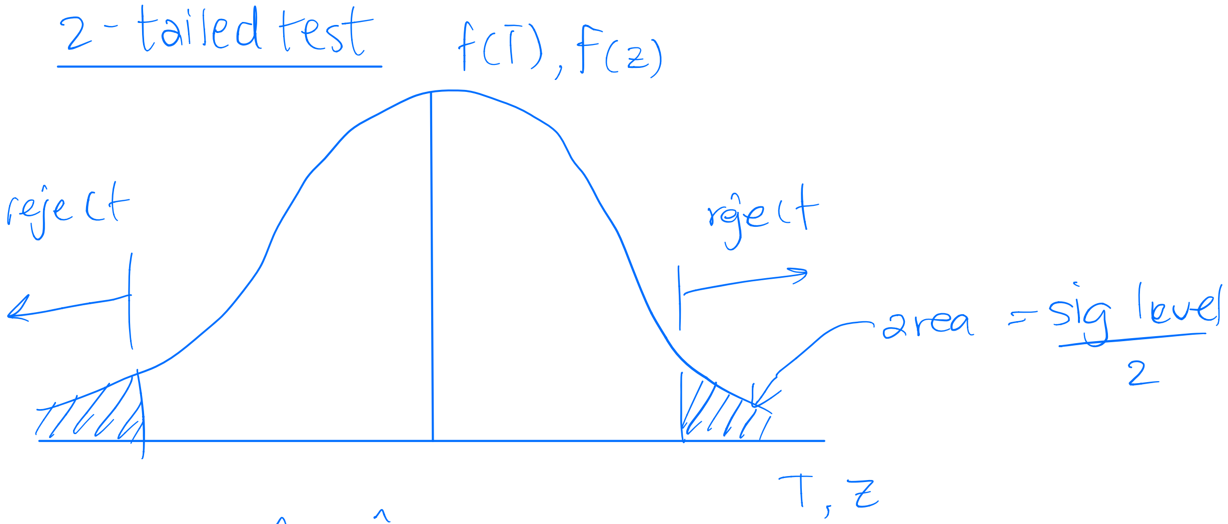
where jc = number of years attending a two-year college
$univ$ = number of years at a four-year college
$\exp er$ = months in the workforce.
We want to test whether $\beta_1 = \beta_2$. → if the returns from 1 more year of education at a junior college is the same as that of the university.

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

against

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

<u>2-tailed test</u>



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

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

⇝ we use a variable transformation trick

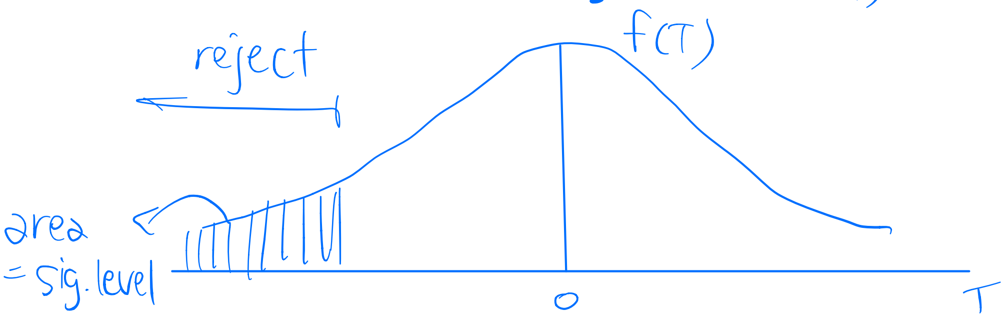
↳ see notes.

another possible hypothesis test (one-tailed alternative)

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

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

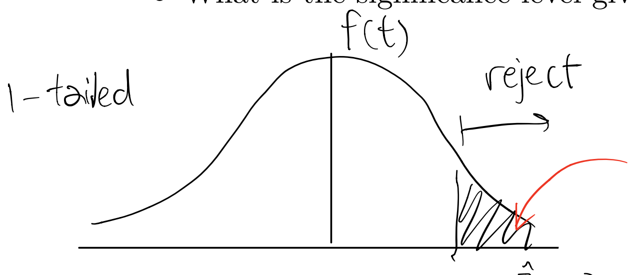
• It is assumed that β_1 would not be more than β_2 (returns to a 2-year college would never be more than returns to university education)



* Then, go to the extra note.

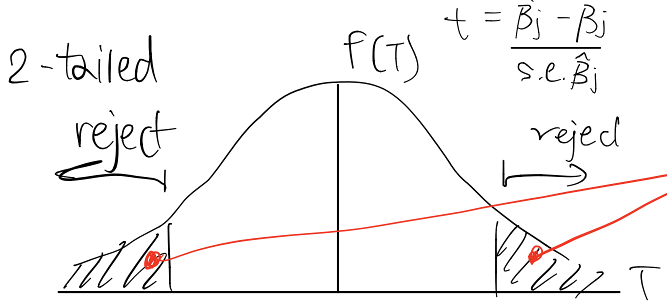
5 Computing p-Values for t-Tests

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



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

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



Total area from 2 sides

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

- 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 $|\ |_{abs}$ term) than our t in the H_0 test.

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 $H_0: \beta_1 - 3\beta_2 = 1$

H_a: Otherwise is true

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

$$t = \frac{\hat{\theta}_1 - 1}{s.e.(\hat{\theta}_1)}$$

$\Rightarrow$ 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 $\beta_1 = \theta_1 + 3\beta_2$

Sub in the main regression and get

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

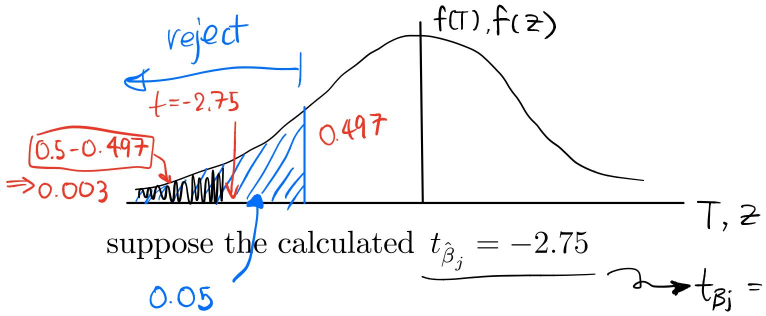
$$= \beta_0 + \theta_1 X_1 + 3\beta_2 X_1 + \beta_2 X_2 + \beta_3 X_3 + u$$

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

* 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: $H_0 : \beta_j \geq 0$, $H_a : \beta_j < 0$, d.f.= 140. → z-table.



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

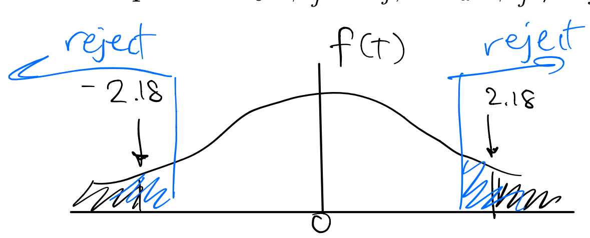
⇒ find the shaded area!

suppose the calculated $t_{\hat{\beta}_j} = -2.75$ → $t_{\beta_j} = \dfrac{(\hat{\beta}_j - \beta_j)}{se.(\hat{\beta}_j)}$

- 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.



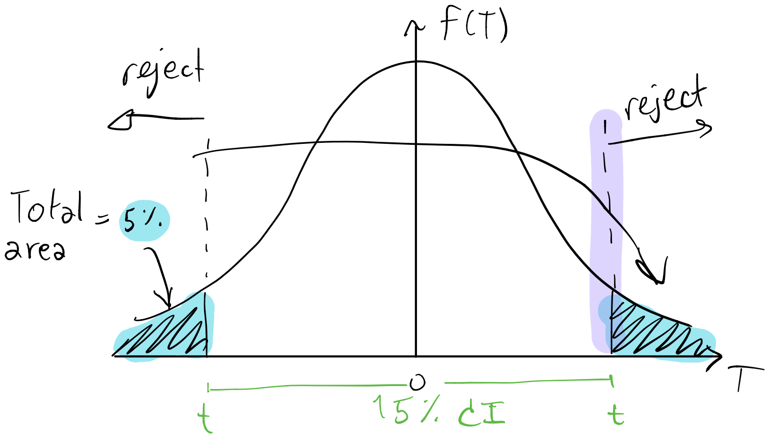
suppose the calculated $t_{\hat{\beta}_j} = -2.18$

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

Yes, reject H_0 b/c the area is less than 0.05 or p-value < 0.05

6 Confidence Intervals (CI)

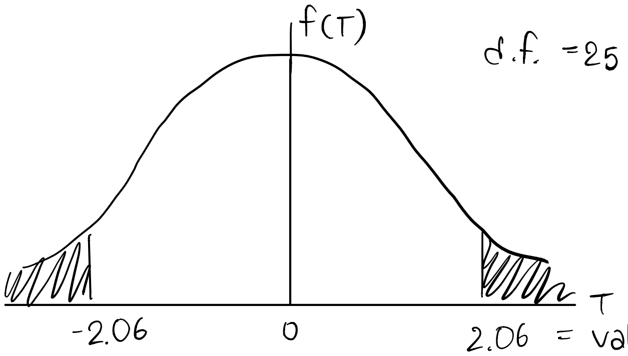
- Confidence Intervales for the POPULATION PARAMETER (β_j) → The range of values that would capture the true β_j at a 15% chance.
- A 95% CI of β_j is given by



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

c is the $\boxed{97.5}$ percentile in the t-distribution with n-k-1 d.f.

Example 1: **95% CI**



d.f. = 25

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

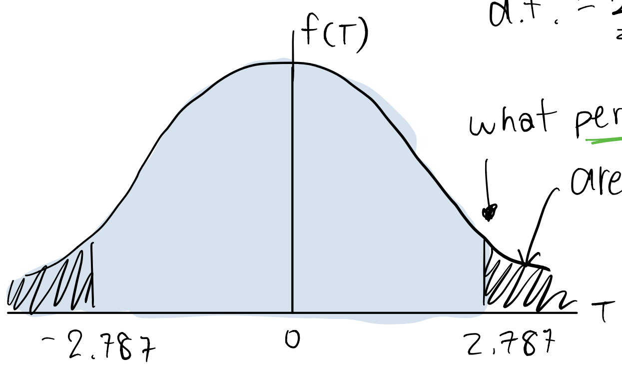
The 95% CI for $\hat{\beta}_j = [\hat{\beta}_j - 2.06 \cdot s.e.(\hat{\beta}_j), \hat{\beta}_j + 2.06 \cdot s.e.(\hat{\beta}_j)]$

Example 2: **99% CI**

d.f. = 25

what percentile? 99.5 or area = 1 − 0.005 = 0.995

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



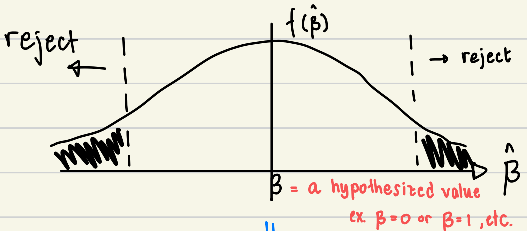
The 99% CI for $\hat{\beta}_j = [\hat{\beta}_j - 2.787 \cdot s.e.\, \hat{\beta}_j, \hat{\beta}_j + 2.787 \cdot s.e.\, \hat{\beta}_j]$

Inference → Hypothesis testing about "β" the true parameter

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

We want to test hypotheses about the true impact (β) of each x variables (educ, experience) on the dependent variable (y)

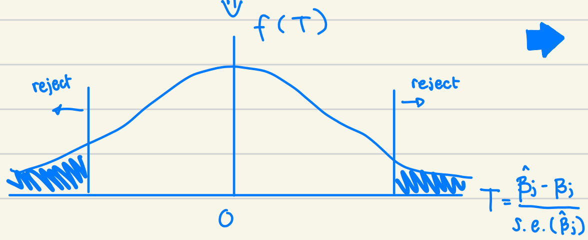
BUT. We don't know what the true β are. So, we use $\hat{\beta}$ (estimator) and s.e. ($\hat{\beta}$) to test hypotheses.



1) Test if β = some number

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

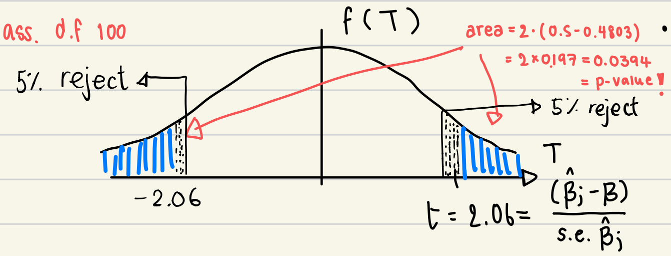
$\beta_j = 1 \rightarrow$ 1 unit ↑ in X_j correspond to 1 unit ↑ in y.



t-test ☆ ⇒ How??

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

Significance level = total area in the rejection region



- Suppose, we calculate a 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

P-value = significant level which we will reject the H_0 or prob that we will reject H_0.

✳ if p-value < significance level ⇒ reject H_0

F-test motivation

- we want to test the significance of a group of hypothesis or multiple hypothesis

$$\text{Grade}_{325} = \beta_0 + \beta_1 \#\text{times\_front} + \beta_2 \#\text{times\_back}$$
$$+ \beta_3 \text{hr\_study} + \beta_4 \text{Past\_GPA} + \beta_5 \text{gender} + u.$$

H_0: seat position doesn't have impact on GPA

$\beta_1 = 0$ and $\beta_2 = 0$ >> $\beta_1 = \beta_2 = 0$

H_a: seat position matters

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

7 Testing Multiple Linear Restrictions: The F-test

Suppose the model is specified by

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

>> 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 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 that 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 $= 0$)

$H_a : H_0$ isn't true.

$$\underbrace{\underbrace{Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_{k-q} X_{k-q}}_{(r)} + \beta_{k-q+1} X_{k-q+1} + \beta_{k-q+2} X_{k-q+2} + \dots + \beta_k X_k + u}_{ur}$$

This is always (+) b.c. $SSR_{ur} < SSR_r$
Everytime you add 1 more x,
the model will be better explained.

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

↳ d.f. of the "ur" model

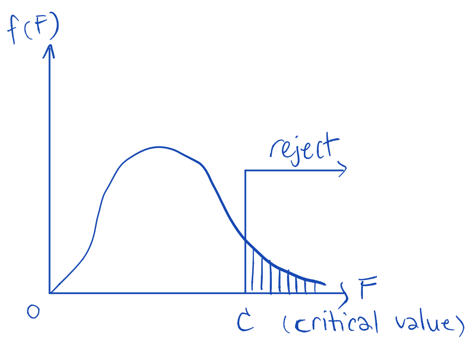
- So, if everytime you add 1 more x variable, the SSR ↓ and R^2 ↑, why don't we just keep the additional x in the model ??

>> because everytime 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 & R^2 ↑



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

H_a : H_0 isn't 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)
>> $SSR_{ur} < SSR_r$

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 doesn't improve the model enough.

4. Other ways to calculate the F-statistics:

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

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

# of β that are set to "0" → q

# of obs → n, # of slope β → k, 1 → intercept

>> 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

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

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

q = 3

1
2
3
4
5
intercept

- the restricted model (r) is defined by

when considering each of the performance x one by one, none of them has a significant impact at 5%.

y x

SSE
SSR
SST

Now, our H_0 and H_a becomes

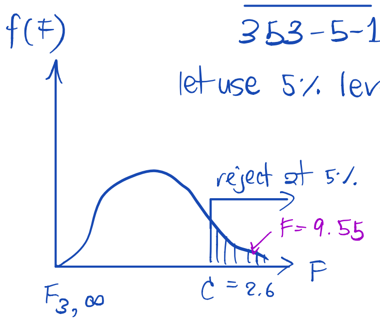
$$F \equiv \frac{\frac{SSR_r - SSR_{ur}}{q}}{\frac{SSR_{ur}}{n-k-1}} = \frac{\frac{198.311 - 183.186}{3}}{\frac{183.186}{353-5-1}} \approx 9.55$$

- But, when performing an F-test, performances have joint impact.

HW:

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

let use 5% level of sig.



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 $t \;\; > 1.96$, we can reject H_0 with 5% sig. level

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

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

   ⇒ If $t \;\; < 1.96$ 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 characteristics

like a simple regression w/1 x

Multiple Regression Analysis : Further Issues

1 Data scaling on OLS statistics

When we change the unit of measurement 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 b_{weght} in kilograms ??

$1\,kg = 1000\,g$

$$\widehat{bweght}_{kg} = \frac{\widehat{bweght}_g}{1000} = \frac{\hat{\beta}_0}{1{,}000} + \frac{\hat{\beta}_1}{1{,}000} cigs + \frac{\hat{\beta}_2}{1000} faminc$$

$$= \hat{\alpha}_0 + \hat{\alpha}_1 cigs + \hat{\alpha}_2 fa \cdot c$$

$$\gg \hat{\alpha}_0 = \frac{\hat{\beta}_0}{1{,}000}, \quad \hat{\alpha}_1 = \frac{\hat{\beta}_1}{1{,}000}, \quad \hat{\alpha}_2 = \frac{\hat{\beta}_2}{1{,}000}$$

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

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

$$\gg \hat{\theta}_2 = \frac{\hat{\beta}_2}{1{,}000}$$

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

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

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

$$b_{weght_{kg}} = \frac{\hat{\beta}_0}{1{,}000} + \frac{\hat{\beta}_1}{1{,}000} cigs + \frac{\left(\frac{\hat{\beta}_2}{1{,}000}\right)}{1{,}000} faminc_{THB}$$

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

2 More on functional forms

- Logarithmic Functional Form

usually means natural log

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

$\Delta y = y_1 - y_2$

$\Delta x_1 = x_{11} - x_{12}$

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

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

(price) (x, elasticity of y) (demand)

$\bullet\ \beta_2 = \dfrac{\partial \log(y)}{\partial x_2} = \dfrac{\frac{1}{y}\partial y}{\partial x_2} = \dfrac{\frac{1}{y}\Delta y}{\Delta x_2}$

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

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

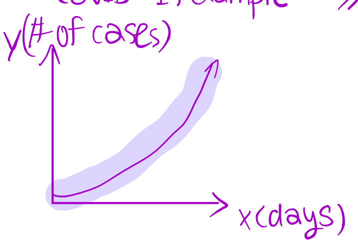
$100\beta_2 = \dfrac{\% \Delta y}{\Delta x_2}$

∴ $100\beta_2$ = % 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 isn't constant)

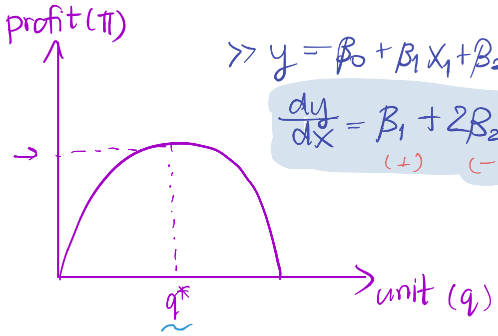
COVID-19 example



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

$\dfrac{dy}{dx} = \underset{(+)}{\beta_1} + \underset{(+)}{2\beta_2} x$ ↳ days

Decreasing returns.



>> $y = \beta_0 + \beta_1 x_1 + \beta_2 x^2 + u$

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

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

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

F.O.C. $= \dfrac{\partial \pi}{\partial q} = 0 = 90 - 2Q$

Example : Effects of Pullution on Housing Prices

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

β_1 is positive

β_2 is negative

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

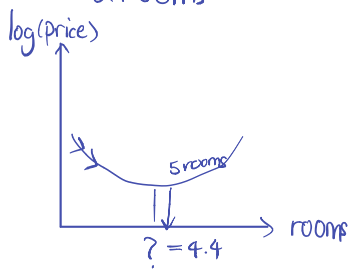
{ In the US or many other countries, students can apply to school in the area w/o having to take any test. So, the lower stratio, the better school.

log(price)
log(nox)
β_1
β_2
β_3
β_4
β_5
intc

$|t| > 1.96$ all < 0.05
>> all variables are significant

Consider the effect of "room"

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



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

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

$rooms = 4.4$

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?

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

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

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

• total % Δ in price when # rooms ↑ from 5 to 7 is

$6.7 + 19.1 = \boxed{25.8\%}$

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

$\% \Delta price = 100(-0.553 + 2(0.062) \cdot 6) = 19.1\%$

3 Models with Interaction Terms

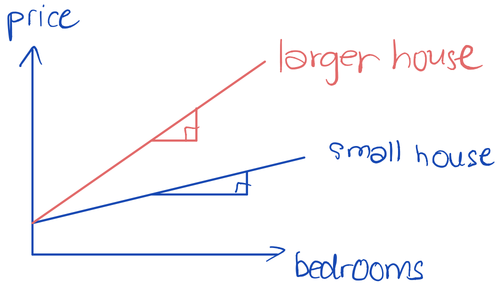
used when the impact of one variable depends on the value (level) of another variable.

Consider

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

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

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

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

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, d.f. $= n-k-1$ ↓, $\frac{SSR}{(n-k-1)}$ ↑, adj. R^2 ↓

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

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.

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

3 Models with a single dummy independent variable

Consider

$$wage = \beta_0 + \delta_0 female + \beta_1 educ + u. \quad (1)$$

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

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

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

2) Thus

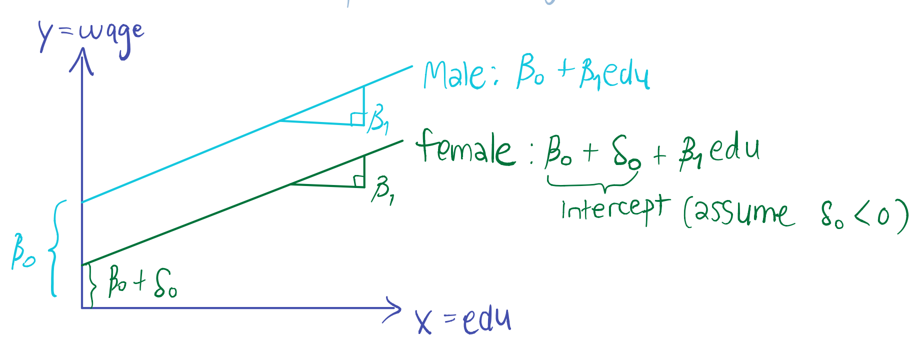
♀ : $E(wage \mid female = 1, edu) = \beta_0 + \delta_0(1) + \beta_1 edu = \beta_0 + \delta_0 + \beta_1 edu$

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

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

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.

$$wage = \beta_0 X_0 + \delta_0 female + \beta_1 edu + \delta_1 male + u$$

X_0: int = 1; female (x_1); edu (x_2); male (x_3)

For example:

$$X_0 = x_1 + x_3$$

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

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			

or

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

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

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

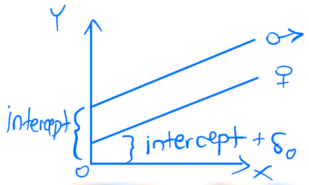
spring = $\begin{cases} 1 \text{ if spring} \\ 0 \text{ otherwise} \end{cases}$

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".

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



Female workers are expected to have less wage compared to male workers.

5 Using dummy variables for multiple categories

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

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

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

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

Comments:

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

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

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

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

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

2) δ_1 measures the impact of be married (marriage premium)

But since $|t| < 1.96$ or $p > 0.05$, we don't reject H_0 of no impact.

	♀	♂
marr	marrfem	marrmale
sing	singfem	singmale

→ basecase

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

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

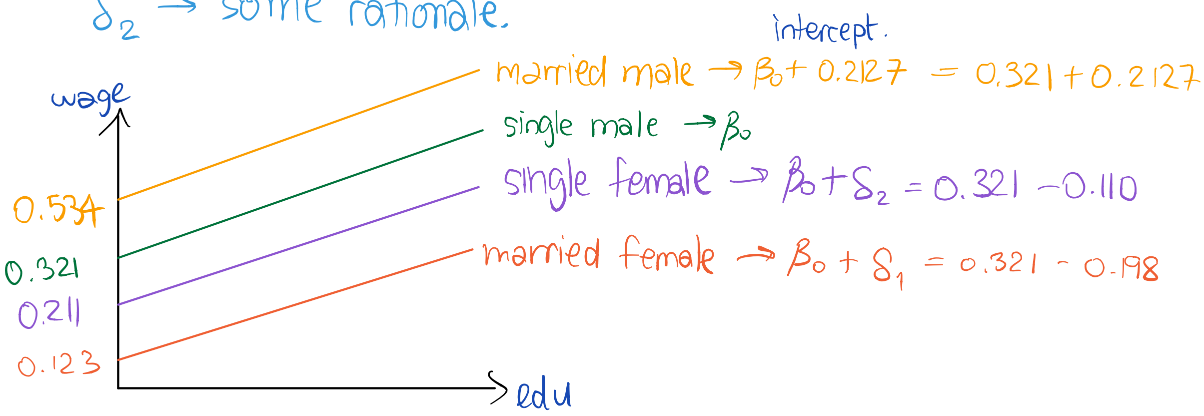
[illegible]

δ_0
δ_1
δ_2

$\hat{\beta}$

Comments: This regression isn't 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.



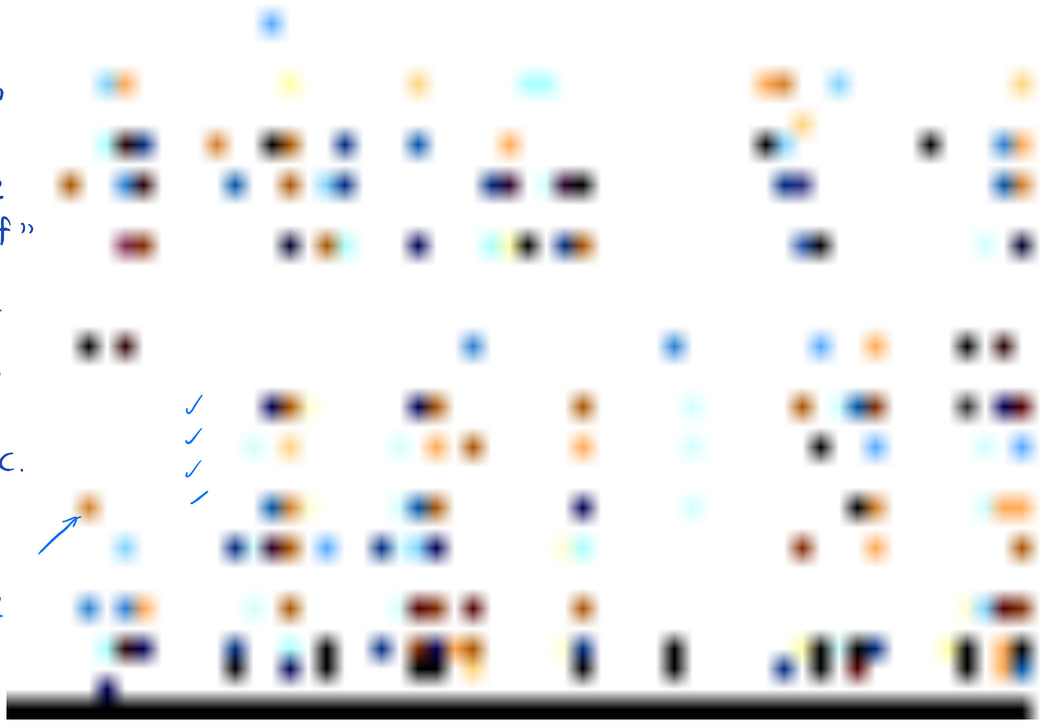
Case 2 We can use dummy variables to represent multiple categories of a variable
Consider the relationship between law school rankings and starting salaries

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

where $top10$, $r11\_25$, $r26\_40$, $r41\_60$ would be equal to 1 when the variable $rank$ falls into the appropriate range.

** Rank below 60 would be the base case.

In many case the "range of the value" serve as a better explanatory variable than the "value itself" eg. age may explain the model better if split into generations
young : 0-15
gen z: 16-29 etc.



the baseline is ranking 61th and worse

Comments:

rank	top 10	r 11 - 25	r 26 - 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
40		0	1

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.