

# Econometric Modeling

AECN 396/896-002

# Before we start

## Learning objectives

1. Enhance the understanding of the interpretation of various models
2. Post-estimation simulation

## Table of contents

1. Expanding on Simple Models
2. Interaction terms
3. Categorical variable
4. R coding tips: categorical variables and interaction terms
5. Other miscellaneous topics

## More on functional forms

---

# Various econometric models

## log-linear

$$\log(y_i) = \beta_0 + \beta_1 x_i + u_i$$

## linear-log

$$y_i = \beta_0 + \beta_1 \log(x_i) + u_i$$

## log-log

$$\log(y_i) = \beta_0 + \beta_1 \log(x_i) + u_i$$

## quadratic

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + u_i$$

# Quadratic

## Model

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + u_i$$

## Calculus

Differentiating the both sides wrt  $x_i$ ,

$$\frac{\partial y_i}{\partial x_i} = \beta_1 + 2 * \beta_2 x_i \Rightarrow \Delta y_i = (\beta_1 + 2 * \beta_2 x_i) \Delta x_i$$

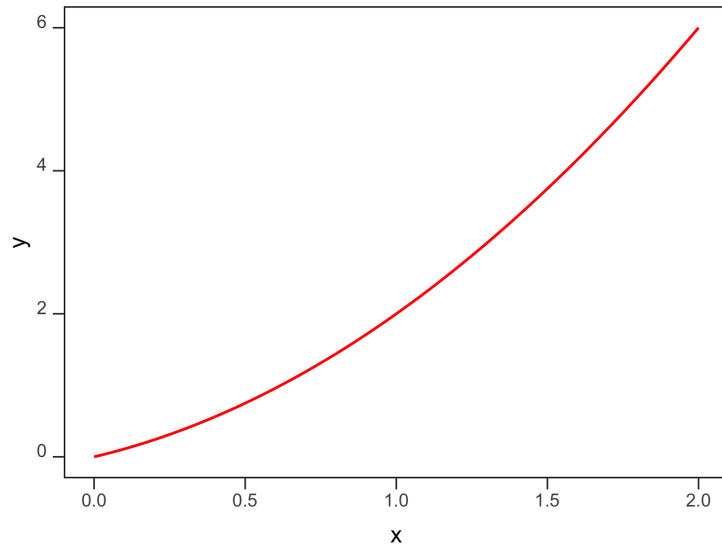
## Interpretation

When  $x$  increases by 1 unit ( $\Delta x_i = 1$ ),  $y$  increases by  $\beta_1 + 2 * \beta_2 x_i$

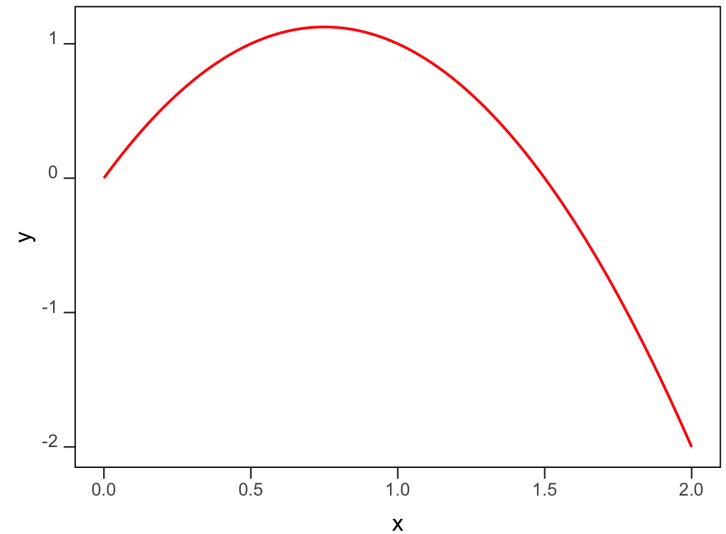
# Visualization

Quadratic functional form is quite flexible.

$$y = x + x^2 \quad (\beta_1 = 1, \beta_2 = 1)$$



$$y = 3x - 2x^2 \quad (\beta_1 = 3, \beta_2 = -2)$$



# Example

## Education impacts of income

The marginal impact of education (the impact of a small change in education on income) may differ what level of education you have had:

- How much does it help to have two more years of education when you have had education until elementary school?
- How much does it help to have two more years of education when you have graduated a college?
- How much does it help to spend two more years as a Ph.D student if you have already spent six years in a Ph.D program

## Observation

The marginal impact of education does not seem to be linear.

## Implementation in R

When you include a variable that is a transformation of an existing variable, use `I()` function in which you write the mathematical expression of the desired transformation.

```
### prepare a dataset ###
wage <- readRDS("wage1.rds")

### run a regression ###
quad_reg <- feols(wage ~ female + educ + I(educ^2), data = wage)

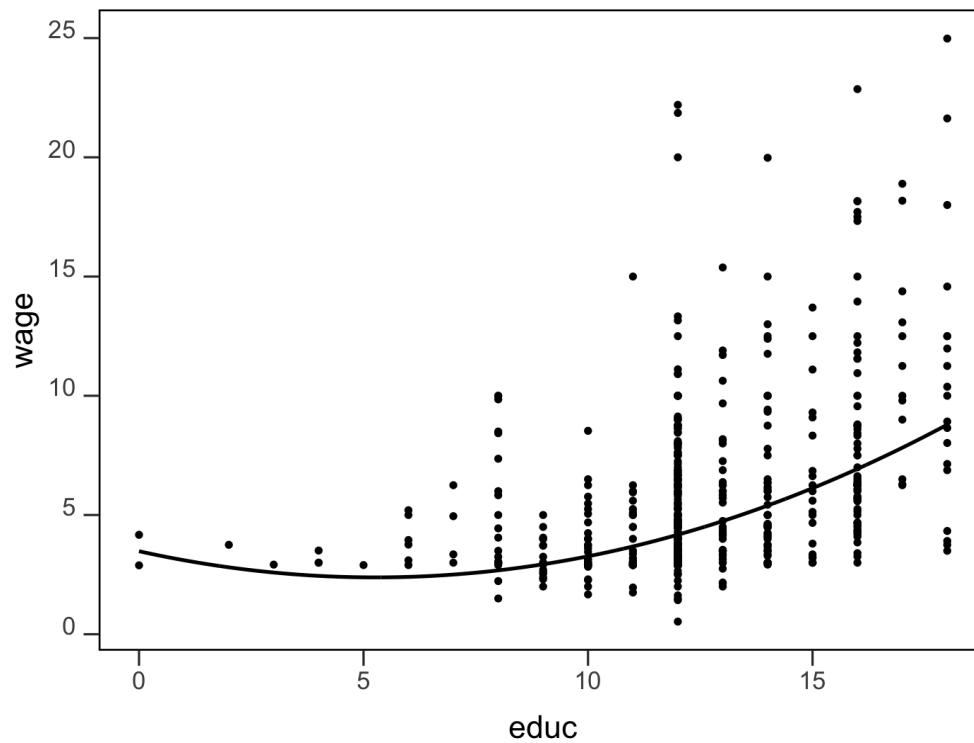
### look at the results ###
tidy(quad_reg)
```

```
## # A tibble: 4 × 5
##   term      estimate std.error statistic  p.value
##   <chr>      <dbl>    <dbl>    <dbl>    <dbl>
## 1 (Intercept)  5.61      1.38      4.05 5.91e- 5
## 2 female     -2.13     0.277    -7.67 8.50e-14
## 3 educ       -0.416    0.231    -1.81 7.14e- 2
## 4 I(educ^2)    0.0395   0.00964    4.10 4.80e- 5
```



### Estimated Model

$$wage = 5.60 - 2.12 \times female - 0.416 \times educ + 0.039 \times educ^2$$



### Estimated Model

$$wage = 5.60 - 2.12 \times female - 0.416 \times educ + 0.039 \times educ^2$$

### Problem

What is the marginal impact of *educ*?

$$\frac{\partial wage}{\partial educ} = ?$$

### Answer

$$\frac{\partial wage}{\partial educ} = -0.416 + 0.039 \times 2 \times educ$$

- When *educ* = 4, additional year of education is going to increase hourly wage by -0.104 on average
- When *educ* = 10, additional year of education is going to increase hourly wage by 0.364 on average

## Statistical significance of the marginal impact

The marginal impact of *educ* is:

$$\frac{\partial wage}{\partial educ} = -0.416 + 0.039 \times 2 \times educ$$

- *educ*:  $-0.416$  ( $t$ -stat =  $-1.80$ )
- *educ*<sup>2</sup>:  $0.039$  ( $t$ -stat =  $4.10$ )

### Question

So, is the marginal impact of *educ* statistically significantly different from 0?

## In the linear case

```
linear_reg <- feols(wage ~ female + educ, data = wage)
tidy(linear_reg)
```

```
## # A tibble: 3 × 5
##   term      estimate std.error statistic  p.value
##   <chr>      <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept)  0.623      0.673      0.926 3.55e- 1
## 2 female     -2.27      0.279     -8.15 2.76e-15
## 3 educ        0.506     0.0504     10.1 7.56e-22
```

### Estimated model

$$wage = 0.62 + 0.51 \times educ$$

### Estimated model

$$wage = 0.62 + 0.51 \times educ$$

### Question

- What is the marginal impact of *educ*?

0.51

- Does the marginal impact of education vary depending on the level of education?

No, the model we estimated assumed that the marginal impact of education is constant.

### Testing

You can just test if  $\hat{\beta}_{educ}$  (the marginal impact of education) is statistically significantly different from 0, which is just a t-test.

## Going back to the quadratic case

With the quadratic specification

- The marginal impact of education varies depending on your education level
- There is no single test that tells you whether the marginal impact of education is statistically significant universally
- Indeed, you need different tests for different values education levels

# Example 1

## Marginal impact of education

$$\hat{\beta}_{educ} + \hat{\beta}_{educ^2} \times 2 \times educ$$

## Hypothesis testing

Does additional year of education has a statistically significant impact (positive or negative) if your current education level is 4?

- $H_0: \hat{\beta}_{educ} + \hat{\beta}_{educ^2} \times 2 \times 4 = 0$
- $H_1: \hat{\beta}_{educ} + \hat{\beta}_{educ^2} \times 2 \times 4 \neq 0$

## Question

Is this

- test of a single coefficient? (t-test)
- test of a single equation with multiple coefficients? (t-test)
- test of multiples equations with multiple coefficients? (F-test)

### t-statistic

$$t = \frac{\hat{\beta}_{educ} + \hat{\beta}_{educ^2} \times 2 \times 4}{se(\hat{\beta}_{educ} + \hat{\beta}_{educ^2} \times 2 \times 4)} = \frac{\hat{\beta}_{educ} + \hat{\beta}_{educ^2} \times 8}{se(\hat{\beta}_{educ} + \hat{\beta}_{educ^2} \times 8)}$$

### R implementation

Remember, a trick to do this test using R is take advantage of the fact that  $F_{1,n-k-1} \sim t_{n-k-1}^2$ .

```
linearHypothesis(quad_reg, "educ + 8*I(educ^2)=0")
```

```
## Linear hypothesis test
##
## Hypothesis:
## educ + 8 I(educ^2) = 0
##
## Model 1: restricted model
## Model 2: wage ~ female + educ + I(educ^2)
##
##   Df  Chisq Pr(>Chisq)
## 1
## 2   1 0.4126    0.5207
```

Since the p-value is 0.529, we do not reject the null.



## Example 2

### Marginal impact of education

$$\hat{\beta}_{educ} + \hat{\beta}_{educ^2} \times 2 \times educ$$

### Hypothesis testing

Does additional year of education has a statistically significant impact (positive or negative) if your current education level is 4?

- $H_0: \hat{\beta}_{educ} + \hat{\beta}_{educ^2} \times 2 \times 10 = 0$
- $H_1: \hat{\beta}_{educ} + \hat{\beta}_{educ^2} \times 2 \times 10 \neq 0$

### Question

Is this

- test of a single coefficient? (t-test)
- test of a single equation with multiple coefficients? (t-test)
- test of multiples equations with multiple coefficients? (F-test)

### t-statistic

$$t = \frac{\hat{\beta}_{educ} + \hat{\beta}_{educ^2} \times 2 \times 10}{se(\hat{\beta}_{educ} + \hat{\beta}_{educ^2} \times 2 \times 10)} = \frac{\hat{\beta}_{educ} + \hat{\beta}_{educ^2} \times 20}{se(\hat{\beta}_{educ} + \hat{\beta}_{educ^2} \times 20)}$$

### R implementation

```
linearHypothesis(quad_reg, "educ + 20*I(educ^2)=0")
```

```
## Linear hypothesis test
##
## Hypothesis:
## educ + 20 I(educ^2) = 0
##
## Model 1: restricted model
## Model 2: wage ~ female + educ + I(educ^2)
##
## Df Chisq Pr(>Chisq)
## 1
## 2 1 39.831 2.769e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Since the much lower than is 0.01, we can reject the null at the 1% level.

## Interaction terms

---

### An interaction term

A variable that is a multiplication of two variables

### Example

$educ \times exper$

### A model with an interaction term

$$wage = \beta_0 + \beta_1 exper + \beta_2 educ \times exper + u$$

### Marginal impact of experience

$$\frac{\partial wage}{\partial exper} = \beta_1 + \beta_2 \times educ$$

### Implications

The marginal impact of experience depends on education

- $\beta_1$ : the marginal impact of experience when  $educ = ?$
- if  $\beta_2 > 0$ : additional year of experience is worth more when you have more years of education

## Regression with interaction terms

Just like the quadratic case with  $educ^2$ , you can use `I()`.

```
reg_int <- feols(wage ~ female + exper + I(exper * educ), data = wage)
```

	Model 1
(Intercept)	6.121***
	(0.267)
exper	-0.188***
	(0.024)
female	-2.418***
	(0.277)
I(exper * educ)	0.020***
	(0.002)
Std. errors	IID
* p < 0.1, ** p < 0.05, *** p < 0.01	

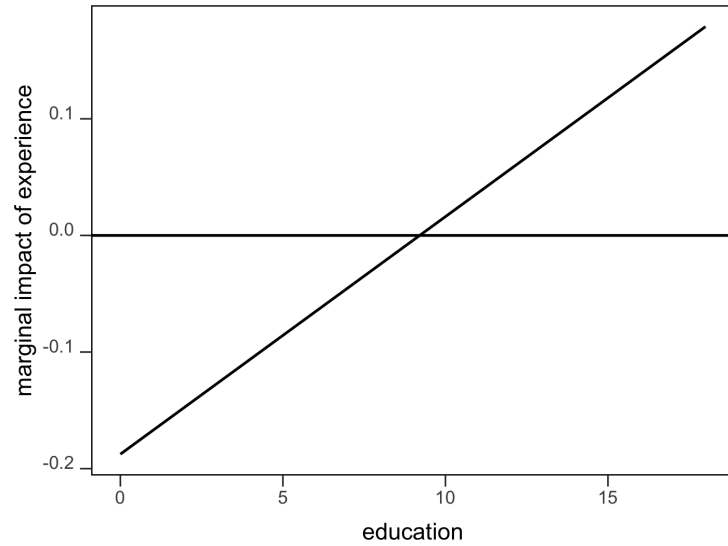
### Estimated Model

$$wage = 6.121 - 2.418 \times female - 0.188 \times exper + 0.020 \times educ \times exper$$

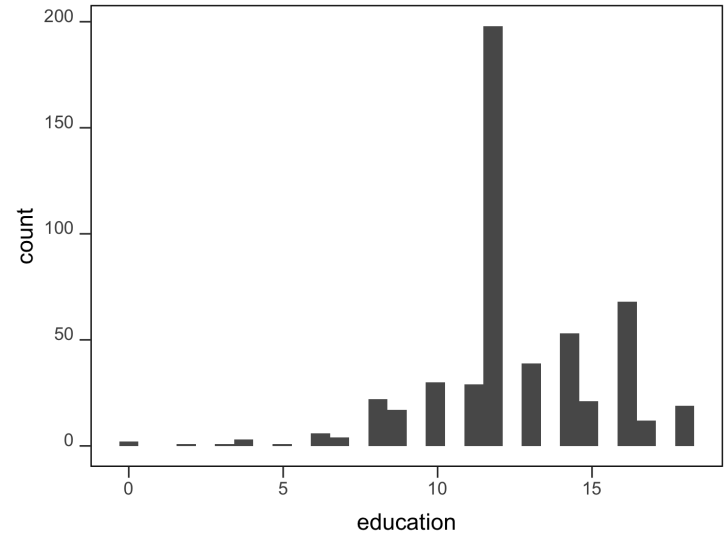
### Marginal impact of experience

$$\frac{\partial wage}{\partial exper} = -0.188 + 0.020 \times educ$$

Marginal impact of *exper*:



Histogram of education:





### Testing of the marginal impact

- Just like the case of the quadratic specification of education, marginal impact of experience is not constant
- We can test if the marginal impact of experience is statistically significant for a given level of education
  - When  $educ = 10$ ,  $\frac{\partial wage}{\partial exper} = -0.188 + 0.020 \times 10 = 0.012$
  - When  $educ = 15$ ,  $\frac{\partial wage}{\partial exper} = -0.188 + 0.020 \times 15 = 0.112$

### Question

Does additional year of experience has a statistically significant impact (positive or negative) if your current education level is 10

### Hypothesis

- $H_0: \hat{\beta}_{exper} + \hat{\beta}_{exper\_educ} \times 10 = 0$
- $H_1: \hat{\beta}_{exper} + \hat{\beta}_{exper\_educ} \times 10 \neq 0$

## R implementation

```
linearHypothesis(reg_int, "exper+10*I(exper * educ)=0")
```

```
## Linear hypothesis test
##
## Hypothesis:
## exper + 10 I(exper * educ) = 0
##
## Model 1: restricted model
## Model 2: wage ~ female + exper + I(exper * educ)
##
##      Df  Chisq Pr(>Chisq)
## 1
## 2  1 2.4627    0.1166
```

## Including qualitative information

---

# Qualitative information

## Issue

How do we include qualitative information as an independent variable?

## Examples

- male or female (binary)
- married or single (binary)
- high-school, college, masters, or Ph.D (more than two states)

# Binary variables

## Dummy variable

- Relevant information in binary variables can be captured by a **zero-one** variable that takes the value of 1 for one state and 0 for the other state
- We use "dummy variable" to refer to a binary (zero-one) variable

## Example

```
wage <- readRDS("wage1.rds")  
dplyr::select(wage, wage, educ, exper, female, married) %>%  
  head()
```

```
##   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  
## 6 8.75   16     9       0        1
```

### Model with dummy a variable

$$wage = \beta_0 + \sigma_f female + \beta_2 educ + u$$

### Interpretation

- **female**:  $E[wage|female = 1, educ] = \beta_0 + \sigma_f + \beta_2 educ$
- **male**:  $E[wage|female = 0, educ] = \beta_0 + \beta_2 educ$

This means that

$$\sigma_f = E[wage|female = 1, educ] - E[wage|female = 0, educ]$$

$$\sigma_f = E[wage|female = 1, educ] - E[wage|female = 0, educ]$$

Verbally,

- $\sigma_f$  is the difference in the expected wage conditional on education between female and male
- $\sigma_f$  measures how much more (less) female workers make compared to male workers (baseline) if they were to have the same education level



## Regression with a dummy variable

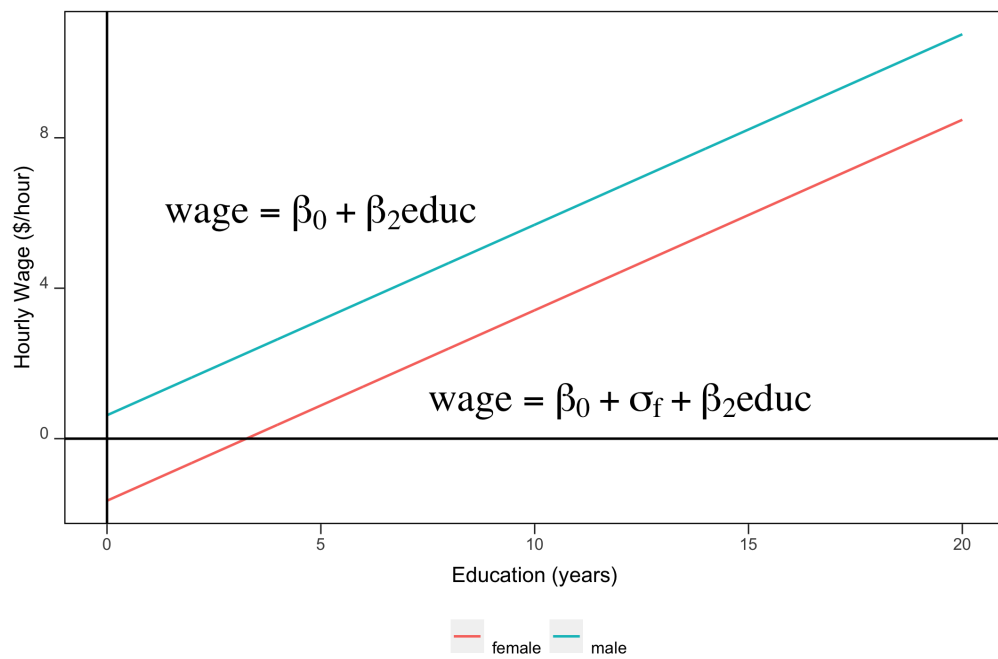
```
reg_df <- feols(wage ~ female + educ, data = wage)
reg_df
```

```
## OLS estimation, Dep. Var.: wage
## Observations: 526
## Standard-errors: IID
##           Estimate Std. Error   t value   Pr(>|t|)
## (Intercept)  0.622817   0.672533   0.926076 3.5483e-01
## female      -2.273362   0.279044  -8.146954 2.7642e-15 ***
## educ         0.506452   0.050391 10.050520 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## RMSE: 3.17642   Adj. R2: 0.255985
```

## Interpretation

Female workers make -2.2733619 (\$/hour) less than male workers on average even though they have the same education level.

## Visualization of the estimated model



### Model with dummy a variable

$$wage = \beta_0 + \sigma_m male + \beta_2 educ + u$$

### Interpretation

- **male**:  $E[wage|male = 1, educ] = \beta_0 + \sigma_m + \beta_2 educ$
- **female**:  $E[wage|male = 0, educ] = \beta_0 + \beta_2 educ$

This means that

$$\sigma_m = E[wage|male = 1, educ] - E[wage|male = 0, educ]$$

$$\sigma_m = E[wage|male = 1, educ] - E[wage|male = 0, educ]$$

Verbally,

- $\sigma_m$  is the difference in the expected wage conditional on education between female and male
- $\sigma_m$  measures how much more (less) male workers make compared to female workers ([baseline](#)) if they were to have the same education level

**Important:** whichever status that is given the value of 0 becomes the baseline

## Regression with a dummy variable

```
wage <- mutate(wage, male = 1 - female)
reg_df <- feols(wage ~ male + educ, data = wage)
reg_df
```

```
## OLS estimation, Dep. Var.: wage
## Observations: 526
## Standard-errors: IID
##           Estimate Std. Error  t value  Pr(>|t|)
## (Intercept) -1.650545    0.652317  -2.53028 1.1689e-02 *
## male         2.273362    0.279044   8.14695 2.7642e-15 ***
## educ         0.506452    0.050391  10.05052 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## RMSE: 3.17642   Adj. R2: 0.255985
```

## Interpretation

Female workers make 2.2733619 (\$/hour) more than male workers on average even though they have the same education level.

### Question

What do you think will happen if we include both male and female dummy variables?

### Answer

- They contain redundant information
- Indeed, including both of them along with the intercept would cause perfect collinearity problem
- So, you need to drop either one of them

### Perfect Collinearity

intercept = male + female

Here is what happens if you include both:

```
reg_dmf <- feols(wage ~ male + female + educ, data = wage)
reg_dmf
```

```
## OLS estimation, Dep. Var.: wage
## Observations: 526
## Standard-errors: IID
##      Estimate Std. Error  t value  Pr(>|t|)
## (Intercept) -1.650545    0.652317  -2.53028 1.1689e-02 *
## male         2.273362    0.279044   8.14695 2.7642e-15 ***
## educ         0.506452    0.050391  10.05052 < 2.2e-16 ***
## ... 1 variable was removed because of collinearity (female)
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## RMSE: 3.17642   Adj. R2: 0.255985
```

## Interactions with a dummy variable

### Issue

- In the previous example, the impact of education on wage was modeled to be exactly the same
- Can we build a more flexible model that allows us to estimate the differential impacts of education on wage between male and female?



### A more flexible model

$$wage = \beta_0 + \sigma_f female + \beta_2 educ + \gamma female \times educ + u$$

- [female]:  $E[wage|female = 1, educ] = \beta_0 + \sigma_f + (\beta_2 + \gamma)educ$
- [male]:  $E[wage|female = 0, educ] = \beta_0 + \beta_2 educ$

### Interpretation

For female, education is more effective by  $\gamma$  than it is for male.

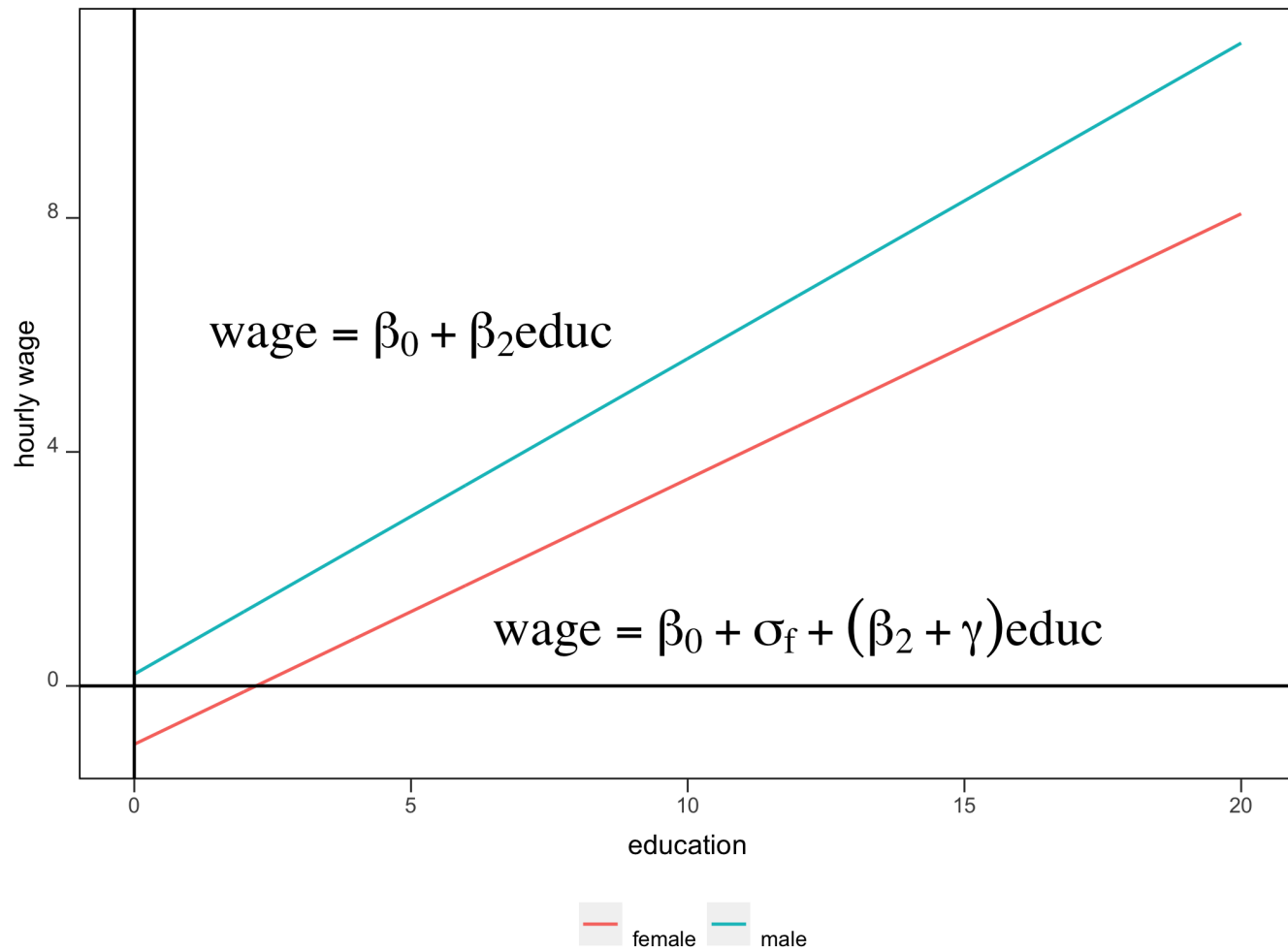
### Example using R

```
reg_di <- lm(wage ~ female + educ + I(female * educ), data = wage)
reg_di
```

```
##
## Call:
## lm(formula = wage ~ female + educ + I(female * educ), data = wage)
##
## Coefficients:
##      (Intercept)          female          educ  I(female * educ)
##           0.2005          -1.1985           0.5395          -0.0860
```

### Interpretation

The marginal benefit of education is 0.086 (\$/hour) less for females workers than for male workers on average.



## Categorical variable: more than two states

### Issue

- Consider a variable called *degree* which has three status values: college, master, and doctor.
- Unlike a binary variable, there are three status values.
- How do we include a categorical variable like this in a model?

### What do we do about this?

You can create three dummy variables likes below:

- `college`: 1 if the highest degree is college, 0 otherwise
- `master`: 1 if the highest degree is Master's, 0 otherwise
- `doctor`: 1 if the highest degree is Ph.D., 0 otherwise

You then include two (the number of status values - 1) of the three dummy variables:

### Model

$$wage = \beta_0 + \sigma_m master + \sigma_d doctor + \beta_1 educ + u$$

- [college]:  $E[wage|master = 0, doctor = 0, educ] = \beta_0 + \beta_1 educ$
- [master]:  $E[wage|master = 1, doctor = 0, educ] = \beta_0 + \sigma_m + \beta_1 educ$
- [doctor]:  $E[wage|master = 0, doctor = 1, educ] = \beta_0 + \sigma_d + \beta_1 educ$

### Interpretation

$\sigma_m$ : the impact of having a MS degree **relative to** having a **college degree**

$\sigma_d$ : the impact of having a Ph.D. degree **relative to** having a **college degree**

### Important

The omitted category (here, **college**) becomes the baseline.

# Structural differences across groups

## Definition

Structural difference refers to the fundamental differences in the model of a phenomenon in the population:

### Example

Male:  $cumgpa = \alpha_0 + \alpha_1 sat + \alpha_2 hsperc + \alpha_3 tothrs + u$

Female:  $cumgpa = \beta_0 + \beta_1 sat + \beta_2 hsperc + \beta_3 tothrs + u$

- *cumgpa*: college grade points averages for male and female college athletes
- *sat*: SAT score
- *hsperc*: high school rank percentile
- *tothrs*: total hours of college courses

### In this example,

*cumgpa* are determined in a fundamentally different manner between female and male students.

You do not want to run a single regression that fits a single model for both female and male students.



### What to do?

If you suspect that the underlying process of how the dependent variable is determined vary across groups, then you should test that hypothesis!

### To do so,

You estimate the model that allows to estimate separate models across groups within a single regression analysis.

$$\begin{aligned} cumgpa = & \beta_0 + \sigma_0 female + \beta_1 sat + \sigma_1 (sat \times female) \\ & + \beta_2 hsperc + \sigma_2 (hsperc \times female) \\ & + \beta_3 tothrs + \sigma_3 (tothrs \times female) + u \end{aligned}$$

### The flexible model

$$\begin{aligned} cumgpa = & \beta_0 + \sigma_0 female + \beta_1 sat + \sigma_1 (sat \times female) \\ & + \beta_2 hsperc + \sigma_2 (hsperc \times female) \\ & + \beta_3 tothrs + \sigma_3 (tothrs \times female) + u \end{aligned}$$

### Male

$$E[cumgpa] = \beta_0 + \beta_1 sat + \beta_2 hsperc + \beta_3 tothrs$$

### Female

$$E[cumgpa] = (\beta_0 + \sigma_0) + (\beta_1 + \sigma_1) sat + (\beta_2 + \sigma_2) hsperc + (\beta_3 + \sigma_3) tothrs$$

### Interpretation

- $\beta$ s are commonly shared by female and male students
- $\sigma$ s capture the differences between female and male students

### Null Hypothesis (verbal)

The model of GPA for male and female students are not structurally different.

### Null Hypothesis

$$H_0 : \sigma_0 = 0, \sigma_1 = 0, \sigma_2 = 0, \text{ and } \sigma_3 = 0$$

### Question

What test do we do? t-test or F-test?

## R code

Run the unrestricted model with all the interaction terms:

```
gpa <-  
  read.dta13("GPA3.dta") %>%  
  filter(!is.na(ctothrs)) %>%  
  #--- create interaction terms ---#  
  mutate(  
    female_sat := female * sat,  
    female_hspc := female * hspc,  
    female_tothrs := female * tothrs  
  )  
  
#--- regression with female dummy ---#  
reg_full <-  
  feols(  
    cumgpa ~  
    female + sat + female_sat + hspc + female_hspc +  
    tothrs + female_tothrs,  
    data = gpa  
  )
```

### What do you see?

- None of the variables that involve *female* are statistically significant at the 5% level individually.
- Does this mean that *male* and *female* students have the same regression function?
- No, we are testing the joint significance of the coefficients. We need to do an  $F$ -test!

	Model 1
(Intercept)	1.481***
	(0.207)
female	-0.353
	(0.411)
female_hspc	-0.001
	(0.003)
female_sat	0.001*
	(0.000)
female_tothrs	-0.000
	(0.002)
hspc	-0.008***
	(0.001)
sat	0.001***
	(0.000)
tothrs	0.002***
	(0.001)
* p < 0.1, ** p < 0.05, *** p < 0.01	

```
linearHypothesis(
  reg_full,
  c(
    "female = 0",
    "female_hspc = 0",
    "female_sat = 0",
    "female_tothrs = 0"
  )
)
```

```
## Linear hypothesis test
##
## Hypothesis:
## female = 0
## female_hspc = 0
## female_sat = 0
## female_tothrs = 0
##
## Model 1: restricted model
## Model 2: cumgpa ~ female + sat + female_sat + hspc + female_hspc +
##         tothrs + female_tothrs
##
##      Df  Chisq Pr(>Chisq)
## 1
## 2   4 32.716  1.365e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

## R coding tips: categorical variables and interaction terms

---

## R coding tips: categorical variables and interaction terms

```
## load the package to access the data we want
library(wooldridge)

## get big9salary
data("big9salary")

## creat a variable that indicates university
## this is how the data would like most of the time (instead of having bunch of dummy variables)
big9salary_c <-
  as_tibble(big9salary) %>%
  mutate(
    university =
      case_when(
        osu == 1 ~ "Ohio State U",
        iowa == 1 ~ "U of Iowa",
        indiana == 1 ~ "Indiana U",
        purdue == 1 ~ "Purdue U",
        msu == 1 ~ "Michigan State U",
        mich == 1 ~ "Michigan U",
        wisc == 1 ~ "U of Wisconsin",
        illinois == 1 ~ "U of Illinois"
      )
  ) %>%
  relocate(id, year, salary, pubindx, university)
```



Take a look at the data,

```
head(big9salary_c)
```

```
## # A tibble: 6 × 31
##       id year salary pubindx university totpge assist assoc prof chair top20phd yearphd female osu iowa
##   <int> <int> <int>   <dbl> <chr>      <dbl>   <int> <int> <int> <int>   <int>   <int> <int> <int> <int>
## 1   101   92    NA    30.5 Indiana U    92.7     0     0     1     0       0     73     0     0     0
## 2   101   95    NA    31.0 Indiana U   107.     0     0     1     0       0     73     0     0     0
## 3   101   99 107100   40.5 Indiana U   186.     0     0     1     0       0     73     0     0     0
## 4   102   92  79420   33.5 Indiana U   128.     0     0     1     0       0     76     0     0     0
## 5   102   95  88239   33.9 Indiana U   133.     0     0     1     0       0     76     0     0     0
## 6   102   99 100450   36.2 Indiana U   192.     0     0     1     0       0     76     0     0     0
```

```
tail(big9salary_c)
```

```
## # A tibble: 6 × 31
##       id year salary pubindx university totpge assist assoc prof chair top20phd yearphd female osu iowa
##   <int> <int> <int>   <dbl> <chr>      <dbl>   <int> <int> <int> <int>   <int>   <int> <int> <int> <int>
## 1   932   92  90856   72.7 U of Wisconsin 269.     0     0     1     0       1     73     1     0     0
## 2   932   95 110090   73.5 U of Wisconsin 294.     0     0     1     0       1     73     1     0     0
## 3   932   99 122397   75.2 U of Wisconsin 315.     0     0     1     0       1     73     1     0     0
## 4   933   92  45755    2.19 U of Wisconsin   9.5     1     0     0     0       1     91     0     0     0
## 5   933   95  51846    8.11 U of Wisconsin   88.     1     0     0     0       1     92     0     0     0
## 6   933   99  69630   59.5 U of Wisconsin 208.     0     1     0     0       1     93     0     0     0
```

You can use the `i()` function inside `feols()` like below:

```
feols(salary ~ pubindx + female + i(university, ref = "Indiana U"), data = big9salary_c) %>%  
  tidy()
```

```
## # A tibble: 10 × 5  
##   term                estimate std.error statistic  p.value  
##   <chr>              <dbl>    <dbl>    <dbl>    <dbl>  
## 1 (Intercept)        74544.    3001.     24.8  9.34e-94  
## 2 pubindx             346.      26.6     13.0  3.18e-34  
## 3 female             -5877.    3067.     -1.92  5.59e- 2  
## 4 university::Michigan State U -9188.    3631.     -2.53  1.17e- 2  
## 5 university::Michigan U    -11561.   3833.     -3.02  2.67e- 3  
## 6 university::Ohio State U  -4707.    3790.     -1.24  2.15e- 1  
## 7 university::Purdue U     -10517.   4310.     -2.44  1.50e- 2  
## 8 university::U of Illinois -1809.    3686.     -0.491 6.24e- 1  
## 9 university::U of Iowa    -519.     3951.     -0.131 8.95e- 1  
## 10 university::U of Wisconsin -6840.    4186.     -1.63  1.03e- 1
```

`ref = "Indiana U"` sets the base category to `"Indiana U"`.

So, for example, the highlighted line means that faculty members at Michigan State U make 9,118 USD less annually than those at Indiana U.

### Key

You do not have to make bunch of dummy variables like the original dataset. Just use `i(category_variable)`.

# Interactions terms

You can use `i()` for creating interactions of a categorical variable and a continuous variable.

Suppose you are interested in understanding the impact of `pubindx` (continuous) by `university` (categorical), then

```
feols(salary ~ female + pubindx + i(university, ref = "Indiana U") + i(university, totpge, ref = "Indiana U"))
tidy()
```

```
## # A tibble: 17 × 5
##   term                                estimate std.error statistic  p.value
##   <chr>                                <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept)                        79593.    4267.     18.7    3.02e-61
## 2 female                             -3782.    3113.     -1.21   2.25e- 1
## 3 pubindx                             42.5      172.      0.247  8.05e- 1
## 4 university::Michigan State U       -17995.    5190.     -3.47   5.65e- 4
## 5 university::Michigan U             -13162.    5577.     -2.36   1.86e- 2
## 6 university::Ohio State U           -10073.    5633.     -1.79   7.42e- 2
## 7 university::Purdue U               -19022.    6291.     -3.02   2.61e- 3
## 8 university::U of Illinois          -12818.    5568.     -2.30   2.17e- 2
## 9 university::U of Iowa              -11785.    5510.     -2.14   3.29e- 2
## 10 university::U of Wisconsin        -8197.    6132.     -1.34   1.82e- 1
## 11 university::Michigan State U:pubindx  436.      191.      2.29   2.25e- 2
## 12 university::Michigan U:pubindx      253.      177.      1.43   1.54e- 1
## 13 university::Ohio State U:pubindx     305.      185.      1.65   9.96e- 2
## 14 university::Purdue U:pubindx         422.      212.      2.00   4.65e- 2
## 15 university::U of Illinois:pubindx     594.      225.      2.64   8.44e- 3
## 16 university::U of Iowa:pubindx        588.      206.      2.85   4.50e- 3
## 17 university::U of Wisconsin:pubindx   247.      180.      1.37   1.70e- 1
```

So, the marginal impact of `pubindex` is 436 greater for those at Michigan State U than those at Indiana U.

Other miscellaneous topic

---

## Goodness of fit: $R^2$

### Important

Small value of  $R^2$  does not mean the end of the world (In fact, we could not care less about it.)

### Example

$$ecolabs = \beta_0 + \beta_1 regprc + \beta_2 ecoprc$$

- *ecolabs*: the (hypothetical) pounds of ecologically friendly (ecolabled) apples a family would demand
- *regprc*: prices of regular apples
- *ecoprc*: prices of the hypothetical ecolabled apples

### Key

- The data was obtained via survey and *ecoprc* was set randomly (So, we know  $E[u|x] = 0$ ) by the researcher.
- The (only) objective of the study is to understand the impact of the price of ecolabled apple on the demand for ecolabled apples.

<i>Dependent variable:</i>	
	ecolbs
regprc	3.029*** (0.711)
ecoprc	-2.926*** (0.588)
Constant	1.965*** (0.380)
Observations	660
R <sup>2</sup>	0.036

Suppose you are challenged by somebody who claim that your regression is not good because the  $R^2$  is tiny. How would your respond to his/her attack?



# Scaling

---

### Questions

What happens if you scale up/down variables used in regression?

- coefficients
- standard errors
- t-statistics
- $R^2$

```
#--- regression with female dummy ---#  
reg_no_scale <- lm(wage ~ female + educ, data = wage)  
reg_scale <- lm(wage ~ female + I(educ * 12), data = wage)
```

```
tidy(reg_no_scale)
```

```
## # A tibble: 3 × 5
##   term      estimate std.error statistic  p.value
##   <chr>      <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept)  0.623     0.673      0.926 3.55e- 1
## 2 female     -2.27     0.279     -8.15 2.76e-15
## 3 educ        0.506     0.0504     10.1  7.56e-22
```

```
tidy(reg_scale)
```

```
## # A tibble: 3 × 5
##   term      estimate std.error statistic  p.value
##   <chr>      <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept)  0.623     0.673      0.926 3.55e- 1
## 2 female     -2.27     0.279     -8.15 2.76e-15
## 3 I(educ * 12)  0.0422    0.00420     10.1  7.56e-22
```

So,

- coefficient: 1/12
- standard error: 1/12
- t-stat: the same

### Interpretation

- Regression **without** scaling

hourly wage increases by 0.506 if education increases by a **year**

- Regression **with** scaling (e.g., 48 means 4 years)

hourly wage increases by 0.0422 if education increases by a **month**

### Note

According to the scaled model, hourly wage increases by  $0.0422 * 12$  if education increases by a year (12 months).

That is, the estimated marginal impact of education on wage from the scaled model is the same as that from the non-scaled model.

## Summary

When an independent variable is scaled,

- its coefficient estimate and standard error are going to be scaled up/back to the exact degree the variable is scaled up/back
- t-statistics stays the same (as it should be)
- $R^2$  stays the same (the model does not improve by simply scaling independent variables)