

# Lecture 14: Dummy variables

Economics 326 — Introduction to Econometrics II

Vadim Marmer, UBC

April 5, 2026

## Interval and ordinal variables

- An **interval** variable is one where the difference between two values is meaningful. Example: “Education” when measured in years. The difference between 12 and 10 years of education is meaningful.
- In some data sets, education is reported as an **ordinal** variable: only the order of its values matters, but the difference between values has no meaning. The following two variables are equivalent:

$$\text{Education}_i = \begin{cases} 1 & \text{if high-school graduate,} \\ 2 & \text{if college graduate,} \\ 3 & \text{if advanced degree.} \end{cases}$$

$$\text{Education}_i = \begin{cases} 1 & \text{if high-school graduate,} \\ 10 & \text{if college graduate,} \\ 234 & \text{if advanced degree.} \end{cases}$$

## Categorical variables

- A **categorical** variable has one or more categories, but there is no natural ordering to the categories. Examples: gender, race, marital status, geographic location.
- The following two variables are equivalent:

$$\text{Gender}_i = \begin{cases} 1 & \text{if observation } i \text{ corresponds to a woman,} \\ 2 & \text{if observation } i \text{ corresponds to a man.} \end{cases}$$

$$\text{Gender}_i = \begin{cases} 1 & \text{if observation } i \text{ corresponds to a man,} \\ 2 & \text{if observation } i \text{ corresponds to a woman.} \end{cases}$$

- Categorical and ordinal variables are also called **qualitative**.
- Qualitative variables cannot simply be included in a regression because the regression technique assumes that all variables are interval.

## Dummy variables

- A **dummy** variable is a binary zero-one variable that takes on the value one if some condition is satisfied and zero if that condition fails:

$$\text{Married}_i = \begin{cases} 1 & \text{if individual } i \text{ is married,} \\ 0 & \text{if individual } i \text{ is not married.} \end{cases}$$

- $\text{Unmarried}_i = \begin{cases} 1 & \text{if individual } i \text{ is not married,} \\ 0 & \text{if individual } i \text{ is married.} \end{cases}$
- Note that  $\text{Married}_i + \text{Unmarried}_i = 1$  for all observations  $i$ .

## Example

- Preview of the `wage1` data from the `wooldridge` package:

```
library(wooldridge)
data(wage1)
head(wage1[, c("wage", "female", "educ", "exper", "tenure")], n = 10)
```

	wage	female	educ	exper	tenure
1	3.10	1	11	2	0
2	3.24	1	12	22	2
3	3.00	0	11	2	0
4	6.00	0	8	44	28
5	5.30	0	12	7	2
6	8.75	0	16	9	8
7	11.25	0	18	15	7
8	5.00	1	12	5	3
9	3.60	1	12	26	4
10	18.18	0	17	22	21

- In this dataset:
  - `wage` (hourly wage) — **interval** variable.
  - `educ` (years of education), `exper` (years of experience), `tenure` (years at current firm) — **interval** variables.
  - `female` (1 if woman, 0 if man) — **dummy** (categorical) variable.

## Single dummy independent variable

- Consider the following regression:

$$\text{Wage}_i = \beta_0 + \delta_0 \text{Female}_i + \beta_1 \text{Educ}_i + \beta_3 \text{Exper}_i + \beta_4 \text{Tenure}_i + U_i,$$

and assume that  $E[U_i | \text{Female}_i, \text{Educ}_i, \text{Exper}_i, \text{Tenure}_i] = 0$ .

- Here, `tenure` refers to the number of years the worker has been employed at their current firm.
- If observation  $i$  corresponds to a woman,  $\text{Female}_i = 1$ , and

$$\begin{aligned} E[\text{Wage}_i | \text{Female}_i = 1, \text{Educ}_i, \text{Exper}_i, \text{Tenure}_i] \\ = \beta_0 + \delta_0 + \beta_1 \text{Educ}_i + \beta_3 \text{Exper}_i + \beta_4 \text{Tenure}_i. \end{aligned}$$

- If observation  $i$  corresponds to a man,  $\text{Female}_i = 0$ , and

$$\begin{aligned} E[\text{Wage}_i | \text{Female}_i = 0, \text{Educ}_i, \text{Exper}_i, \text{Tenure}_i] \\ = \beta_0 + \beta_1 \text{Educ}_i + \beta_3 \text{Exper}_i + \beta_4 \text{Tenure}_i. \end{aligned}$$

- Thus,

$$\begin{aligned} \delta_0 = E[\text{Wage}_i | \text{Female}_i = 1, \text{Educ}_i, \text{Exper}_i, \text{Tenure}_i] \\ - E[\text{Wage}_i | \text{Female}_i = 0, \text{Educ}_i, \text{Exper}_i, \text{Tenure}_i]. \end{aligned}$$

## Intercept shift

- The model:

$$\text{Wage}_i = \beta_0 + \delta_0 \text{Female}_i + \beta_1 \text{Educ}_i + \beta_3 \text{Exper}_i + \beta_4 \text{Tenure}_i + U_i.$$

- For men ( $\text{Female}_i = 0$ ):

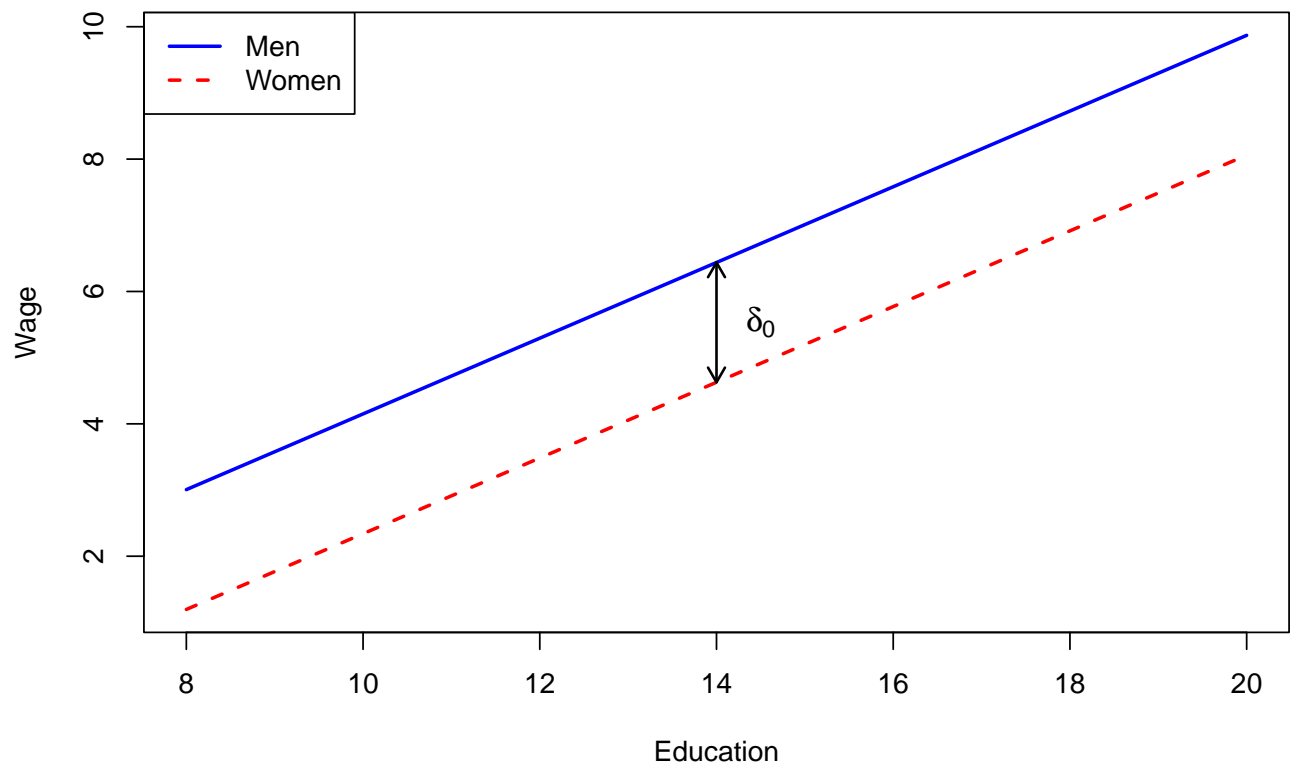
$$\text{Wage}_i^M = \beta_0 + \beta_1 \text{Educ}_i + \beta_3 \text{Exper}_i + \beta_4 \text{Tenure}_i + U_i.$$

- For women ( $\text{Female}_i = 1$ ):

$$\text{Wage}_i^F = (\beta_0 + \delta_0) + \beta_1 \text{Educ}_i + \beta_3 \text{Exper}_i + \beta_4 \text{Tenure}_i + U_i.$$

- In this case, men play the role of the **base** group.
- $\delta_0$  measures the wage difference relative to the base group.

## Intercept shift



## Example

- Estimated equation:

$$\widehat{\text{Wage}}_i = \underset{(0.72)}{-1.57} \underset{(0.26)}{-1.81} \text{Female}_i + \underset{(0.049)}{0.572} \text{Educ}_i \\ + \underset{(0.012)}{0.025} \text{Exper}_i + \underset{(0.021)}{0.141} \text{Tenure}_i.$$

- The dependent variable is the wage per hour.
- $\hat{\delta}_0 = -1.81$  implies that a woman earns \$1.81 less per hour than a man with the same level of education, experience, and tenure. (These are 1976 wages.)
- The difference is also statistically significant.

## Log dependent variable

- The model:

$$\ln(\text{Wage}_i) = \beta_0 + \delta_0 \text{Female}_i + \beta_1 \text{Educ}_i + \beta_3 \text{Exper}_i + \beta_4 \text{Tenure}_i + U_i.$$

- In this case,

$$\begin{aligned} \delta_0 &= \ln(\text{Wage}^F) - \ln(\text{Wage}^M) \\ &= \ln\left(\frac{\text{Wage}^F}{\text{Wage}^M}\right) \\ &= \ln\left(\frac{\text{Wage}^M + (\text{Wage}^F - \text{Wage}^M)}{\text{Wage}^M}\right) \\ &= \ln\left(1 + \frac{\text{Wage}^F - \text{Wage}^M}{\text{Wage}^M}\right) \\ &\approx \frac{\text{Wage}^F - \text{Wage}^M}{\text{Wage}^M}. \end{aligned}$$

- When the dependent variable is in the log form,  $\delta_0$  has a **percentage** interpretation.

## Example

- Estimated equation:

$$\begin{aligned} \ln(\widehat{\text{Wage}}_i) &= 0.417 \underset{(0.099)}{-0.297} \text{Female}_i + 0.080 \underset{(0.007)}{\text{Educ}_i} \\ &\quad + 0.029 \underset{(0.005)}{\text{Exper}_i} - 0.00058 \underset{(0.00010)}{\text{Exper}_i^2} \\ &\quad + 0.032 \underset{(0.007)}{\text{Tenure}_i} - 0.00059 \underset{(0.00023)}{\text{Tenure}_i^2}. \end{aligned}$$

- $\hat{\delta}_0 = -0.297$  implies that a woman earns 29.7% less than a man with the same level of education, experience, and tenure.

## Changing the base group

- Instead of

$$\begin{aligned} \ln(\text{Wage}_i) &= \beta_0 + \delta_0 \text{Female}_i + \beta_1 \text{Educ}_i \\ &\quad + \beta_3 \text{Exper}_i + \beta_4 \text{Tenure}_i + U_i, \end{aligned}$$

consider:

$$\ln(\text{Wage}_i) = \theta_0 + \gamma_0 \text{Male}_i + \theta_1 \text{Educ}_i + \theta_3 \text{Exper}_i + \theta_4 \text{Tenure}_i + U_i.$$

- Since  $\text{Male}_i = 1 - \text{Female}_i$ ,

$$\begin{aligned} \ln(\text{Wage}_i) &= \theta_0 + \gamma_0 \text{Male}_i + \theta_1 \text{Educ}_i + \theta_3 \text{Exper}_i + \theta_4 \text{Tenure}_i + U_i \\ &= \theta_0 + \gamma_0 (1 - \text{Female}_i) + \theta_1 \text{Educ}_i + \theta_3 \text{Exper}_i + \theta_4 \text{Tenure}_i + U_i \\ &= (\theta_0 + \gamma_0) - \gamma_0 \text{Female}_i + \theta_1 \text{Educ}_i + \theta_3 \text{Exper}_i + \theta_4 \text{Tenure}_i + U_i. \end{aligned}$$

- We conclude that  $\delta_0 = -\gamma_0$ ,  $\beta_0 = \theta_0 + \gamma_0$ ,  $\beta_1 = \theta_1$ , etc.:

$$\ln(\text{Wage}_i) = (\beta_0 + \delta_0) - \delta_0 \text{Female}_i + \beta_1 \text{Educ}_i + \beta_3 \text{Exper}_i + \beta_4 \text{Tenure}_i + U_i.$$

- Thus, changing the base group has no effect on the conclusions.
- In this dataset, gender is recorded as a binary variable (female/male). The dummy variable approach shown here applies to any binary grouping.

## Dummy variable trap

- Consider the equation:

$$\ln(\text{Wage}_i) = \beta_0 + \delta_0 \text{Female}_i + \gamma_0 \text{Male}_i + \beta_1 \text{Educ}_i + \beta_3 \text{Exper}_i + \beta_4 \text{Tenure}_i + U_i.$$

- Recall that the intercept is a regressor that takes the value one for all observations.
- In this dataset,  $\text{Female}_i + \text{Male}_i = 1$  for all observations  $i$ , so we have **perfect multicollinearity**. Such an equation cannot be estimated.
- **One cannot include an intercept and dummies for all the groups!**

## Dummy variable trap

- One of the dummies has to be omitted and the corresponding group becomes the **base** group:
  - Men are the base group:  $\ln(\text{Wage}_i) = \beta_0 + \delta_0 \text{Female}_i + \beta_1 \text{Educ}_i + \beta_3 \text{Exper}_i + \beta_4 \text{Tenure}_i + U_i$ .
  - Women are the base group:  $\ln(\text{Wage}_i) = \theta_0 + \gamma_0 \text{Male}_i + \beta_1 \text{Educ}_i + \beta_3 \text{Exper}_i + \beta_4 \text{Tenure}_i + U_i$ .
- Alternatively, one can include both dummies **without** the intercept:

$$\ln(\text{Wage}_i) = \pi_0 \text{Female}_i + \pi_1 \text{Male}_i + \beta_1 \text{Educ}_i + \beta_3 \text{Exper}_i + \beta_4 \text{Tenure}_i + U_i.$$

- In R, a regression without an intercept can be estimated by adding `+ 0` or `- 1` to the formula:

```
lm(Y ~ X + 0)
```

or equivalently:

`lm(Y ~ X - 1)`

- The coefficients on the dummy variables lose the difference interpretation.

## Slope changes and interactions

- We can also allow the returns to education to be different for men and women:

$$\ln(\text{Wage}_i) = \beta_0 + \delta_0 \text{Female}_i + \beta_1 \text{Educ}_i + \delta_1 (\text{Female}_i \cdot \text{Educ}_i) + \beta_3 \text{Exper}_i + \beta_4 \text{Tenure}_i + U_i.$$

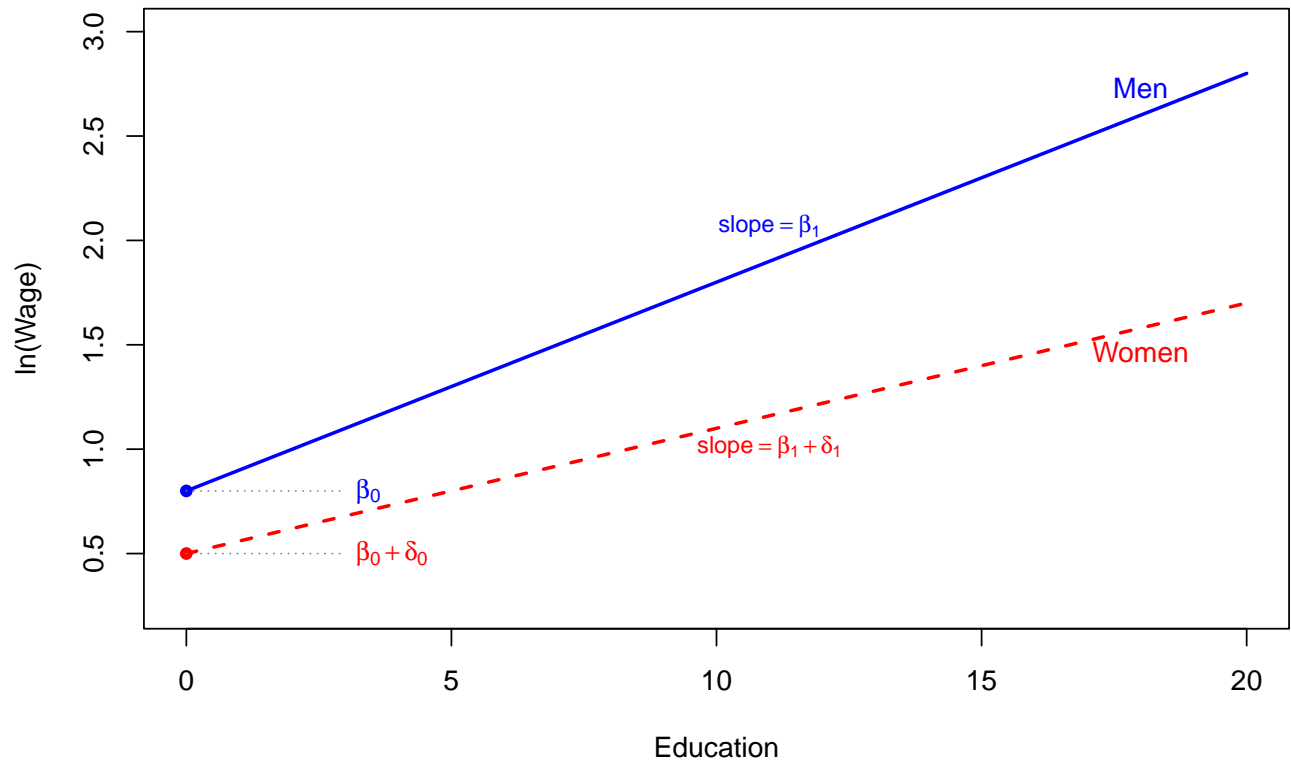
- The variable  $(\text{Female}_i \cdot \text{Educ}_i)$  is called an **interaction**.
- The equation for men ( $\text{Female}_i = 0$ ):

$$\ln(\text{Wage}_i^M) = \beta_0 + \beta_1 \text{Educ}_i + \beta_3 \text{Exper}_i + \beta_4 \text{Tenure}_i + U_i.$$

- The equation for women ( $\text{Female}_i = 1$ ):

$$\ln(\text{Wage}_i^F) = (\beta_0 + \delta_0) + (\beta_1 + \delta_1) \text{Educ}_i + \beta_3 \text{Exper}_i + \beta_4 \text{Tenure}_i + U_i.$$

- $\delta_1$  can be interpreted as the difference in the return to education between women and men (the base group) after controlling for experience and tenure.



## Example

- Estimated equation:

$$\begin{aligned}\ln(\widehat{\text{Wage}}_i) &= 0.389 - 0.227 \text{Female}_i \\ &\quad \begin{matrix} (0.119) & (0.168) \end{matrix} \\ &+ 0.082 \text{Educ}_i - 0.0056 \text{Female}_i \cdot \text{Educ}_i \\ &\quad \begin{matrix} (0.008) & (0.0131) \end{matrix} \\ &+ 0.029 \text{Exper}_i - 0.00058 \text{Exper}_i^2 \\ &\quad \begin{matrix} (0.005) & (0.00011) \end{matrix} \\ &+ 0.032 \text{Tenure}_i - 0.00059 \text{Tenure}_i^2 \\ &\quad \begin{matrix} (0.007) & (0.00024) \end{matrix} \end{aligned}$$

- $\hat{\delta}_1 = -0.0056$ , suggesting that the return to education for women is 0.56 percentage points less than for men; however, this difference is not statistically significant. We cannot reject the hypothesis that the return to education is the same for men and women.

## Multiple categories

- In the previous examples, Educ was a quantitative variable: years of education.
- Suppose now that instead the education variable is **ordinal**:

$$\text{Education}_i = \begin{cases} 1 & \text{if high-school dropout,} \\ 2 & \text{if high-school graduate,} \\ 3 & \text{if some college,} \\ 4 & \text{if college graduate,} \\ 5 & \text{if advanced degree.} \end{cases}$$

- Only the order is important, and there is no meaning to the **distance** between the values.
- Adding such a variable to the regression will give a meaningless result.

## Multiple categories

- Recall the ordinal education variable:

$$\text{Education}_i = \begin{cases} 1 & \text{if high-school dropout,} \\ 2 & \text{if high-school graduate,} \\ 3 & \text{if some college,} \\ 4 & \text{if college graduate,} \\ 5 & \text{if advanced degree.} \end{cases}$$

- Define 5 new dummy variables:

$$E_{1,i} = \begin{cases} 1 & \text{if high-school dropout,} \\ 0 & \text{otherwise.} \end{cases}$$

$$E_{2,i} = \begin{cases} 1 & \text{if high-school graduate,} \\ 0 & \text{otherwise.} \end{cases}$$

$$E_{3,i} = \begin{cases} 1 & \text{if some college,} \\ 0 & \text{otherwise.} \end{cases}$$

$$E_{4,i} = \begin{cases} 1 & \text{if college graduate,} \\ 0 & \text{otherwise.} \end{cases}$$

$$E_{5,i} = \begin{cases} 1 & \text{if advanced degree,} \\ 0 & \text{otherwise.} \end{cases}$$

- To avoid the dummy variable trap, one of the dummies has to be omitted:

$$\begin{aligned} \text{Wage}_i = & \beta_0 + \delta_0 \text{Female}_i + \delta_2 E_{2,i} + \delta_3 E_{3,i} \\ & + \delta_4 E_{4,i} + \delta_5 E_{5,i} + \text{Other Factors} \end{aligned}$$

- Group 1 (high-school dropout) becomes the base group.
- $\delta_2$  measures the wage difference between high-school graduates and high-school dropouts.
- $\delta_3$  measures the wage difference between individuals with some college education and high-school dropouts.

### Comparing consecutive groups

- The previous definitions compare each group to the **base** group (high-school dropouts). Alternatively, we can define dummies that compare each group to the **previous** one:

$$D_{2,i} = \begin{cases} 1 & \text{if high-school graduate or higher,} \\ 0 & \text{otherwise.} \end{cases}$$

$$D_{3,i} = \begin{cases} 1 & \text{if some college or higher,} \\ 0 & \text{otherwise.} \end{cases}$$

$$D_{4,i} = \begin{cases} 1 & \text{if college graduate or higher,} \\ 0 & \text{otherwise.} \end{cases}$$

$$D_{5,i} = \begin{cases} 1 & \text{if advanced degree,} \\ 0 & \text{otherwise.} \end{cases}$$

- The model:

$$\begin{aligned} \text{Wage}_i = & \beta_0 + \delta_0 \text{Female}_i + \gamma_2 D_{2,i} + \gamma_3 D_{3,i} \\ & + \gamma_4 D_{4,i} + \gamma_5 D_{5,i} + \text{Other Factors} \end{aligned}$$

- $\gamma_2$  measures the wage difference between high-school graduates and high-school dropouts.
- $\gamma_3$  measures the wage difference between individuals with some college and high-school graduates.
- $\gamma_4$  measures the wage difference between college graduates and individuals with some college.
- $\gamma_5$  measures the wage difference between individuals with advanced degrees and college graduates.