Ch5-3

Monday, February 10, 2025 11:27 AM

Ch5-3

# 5.3 – Dummy Variables

## Dummy variable

$-\pi$ 1,000,000

- Takes on one of two values (usually 0 or 1)
- Dichotomous variable, binary variable, categorical variable, factor

$$D_i = \begin{cases} 0, & \text{if individual } i \text{ belongs to group } A \\ 1, & \text{if individual } i \text{ belongs to group } B \end{cases}$$

male Female

1 if receives treatment
0 if control (placebo)

- Examples: gender, education, treatment, domestic, employed, insured, etc.

↳ 0 if no HS
1 if completed HS

↳ foreign/domestics

In this section, we consider that the "X" variable is a dummy.

1

## A population model with a dummy variable

0,1 true unobservable

$$Y_i = \beta_0 + \beta_1 D_i + \epsilon_i, \tag{5.13}$$

- What is the interpretation of $\beta_1$ here? slope? diff. in means
- Take a derivative? NOPE, D is not continuous
- What about $\beta_0$? still intercept
- Use *conditional expectations*

$$\mathrm{E}[Y_i|D_i = 1] - \mathrm{E}[Y_i|D_i = 0] = \beta_1 \tag{5.16}$$

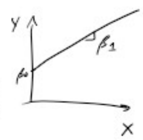
2

When X is continuous

$$y = \beta_0 + \beta_1 X + \epsilon$$

$$\frac{\partial y}{\partial x} = \beta_1 \rightarrow \text{"slope" "marginal effect"}$$

$$\frac{dy}{dx} = \lim_{\Delta x \to 0} \frac{f(x + \Delta x)}{\Delta x}$$




2 situations

when D=0 : $E[y] = \beta_0 + \beta_1(0) = \beta_0$ ← pop. mean for D=0 group

when D=1 : $E[y] = \beta_0 + \beta_1(1) = \beta_0 + \beta_1$ ← pop. mean for D=1 group

$(\beta_0 + \beta_1) - (\beta_0) = \beta_1$

D=1 D=0

↳ diff. in pop means btw. 2 groups

## An estimated model with a dummy variable

| TRUE | ESTIMATED |
| --- | --- |
| $\beta_0 \rightarrow \mu_{D=0}$ | $\bar{y}_{D=0}$ |
| $\beta_0 + \beta_1 \rightarrow \mu_{D=1}$ | $\bar{y}_{D=1}$ |

Use OLS as before. lm(y ~ D)

$$Y_i = b_0 + b_1 D_i + e_i, \quad \text{estimated} \tag{5.17}$$

- $b_0$ is the *sample* mean ($\bar{Y}$) for $D_i = 0$
- $b_0 + b_1$ is the *sample* mean for $D_i = 1$
- $b_1$ is the difference in sample means (be careful of the sign)

This means that, instead of using OLS, we could just divide the sample into two parts (using $D_i$), and calculate two sample averages! So why should we use OLS? At this stage, it looks like we are making things more complicated than they need to be. However, in the next chapter, we will add more $X$ variables, so that we will not be able to get the same results by dividing the sample into two.

3

## Example: Gender wage gap using CPS

The current population survey (CPS) is a monthly detailed survey conducted in the United States. It contains information on many labour market and demographic characteristics. In this section, we will use a subset of data from the 1985 CPS, to estimate the differences in wages between men and women.

You will see many variables in the dataset. For now, we look at only a few:

- wage - hourly wage
- education - number of years of education
- gender - dummy variable for gender



$\sqrt{\hat{var}(\bar{y}_m - \bar{y}_f)}$

Load the data:

```
cps <- read.csv("https://rtgodwin.com/data/cps1985.csv")
```

To run an OLS regression of *wage* on *gender*, use the following command:

gendermale = 0 if gender == "female"
= 1 "male"

```
summary(lm(wage ~ gender, data = cps))
```

men make $2.12 more on avg.

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   7.8789     0.3216   24.50  < 2e-16 ***
gendermale    2.1161     0.4372    4.84  1.7e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.034 on 532 degrees of freedom
Multiple R-squared:  0.04218,   Adjusted R-squared:  0.04038
F-statistic: 23.43 on 1 and 532 DF,  p-value: 1.703e-06
```

$\overline{wage}_{men} - \overline{wage}_{women}$

Test if there is a gender-wage gap.
Is there a diff. in wages of men and women?



If $D = 0 \rightarrow$ "male"
$D = 1 \rightarrow$ "female"

$b_1 = -2.12$

$\hookrightarrow \overline{wage}_{women} - \overline{wage}_{men}$

$wage = \beta_0 + \beta_1 gender + \epsilon$

$E(wage \mid gender=0) = \beta_0$
$E(wage \mid gender=1) = \beta_0 + \beta_1$ (true mean)

$\hat{wage} = b_0 + b_1 gender$

$\overline{wage}\mid_{gender=1}$

$\overline{wage}\mid_{gender=0}$

$H_0: \beta_1 = 0$ reject p-val < .001

$wage = \beta_0 + \beta_1 gender + \epsilon$

From this output, you should be able to answer the following questions:

- What is the sample mean wage for males and for females? $\rightarrow b_0 = \$7.88$ $\hookrightarrow b_0 + b_1 = \$10$
- What is the interpretation of $b_1$?

avg. diff. in wages

In class exercise: Test the hypothesis that there is no difference in the earnings of men and women.

$H_0: \beta_1 = 0 \Rightarrow t = \frac{b_1 - 0}{s.e.(b_1)}$



We stated earlier that the results we obtain from regressing on a dummy variable are equivalent to what we would obtain by dividing the sample into two parts (by gender). Let's verify this using the CPS data. In R, create subsets for men and women:

```
cps.m <- subset(cps, gender == 'male')
cps.f <- subset(cps, gender == 'female')
```

then take the difference in the sample mean wage between men and women:

```
mean(cps.m$wage) - mean(cps.f$wage)
```

```
[1] 2.116056
```

The difference is equal to $b_1$, which is 2.1161! Also, note that the sample mean wage for women is $b_0$:

Sample mean wage for women is $b_0$:

```
mean(cps.f$wage)
```

```
[1] 7.878857
```

and the sample mean wage for men is $b_0 - b_1$:

```
mean(cps.m$wage)
```

```
[1] 9.994913
```

Define: genderfemale = 0 if "male"
= 1 if "female"

Model: $wage = \beta_0 + \beta_1 genderfemale + \epsilon$

$b_0 = ? = \$10$

$b_1 = ? = -\$2.12$

Exercise: A researcher defines the dummy variable in the *opposite* way. What are the new values for $b_0$ and $b_1$?



## 5.4 Reporting regression results

$$\underset{(\,)}{\hat{wage}} = \underset{(0.32)}{\overset{b_0}{7.88}} + \underset{(0.44)}{\overset{b_1}{2.12}} \times gendermale, \quad \boxed{R^2 = 0.042}$$

s.e.$(b_1)$

$$t = \frac{2.12 - b_{1,0}}{0.44}$$

hypothesis

This equation contains:

- Estimated $\beta$s
- Estimated standard errors
- $R^2$
- Everything you need to do a hypothesis test
- Example: test the hypothesis that there is no wage-gender gap