5.3 – Dummy Variables

Dummy variable

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

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

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

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

A population model with a dummy variable

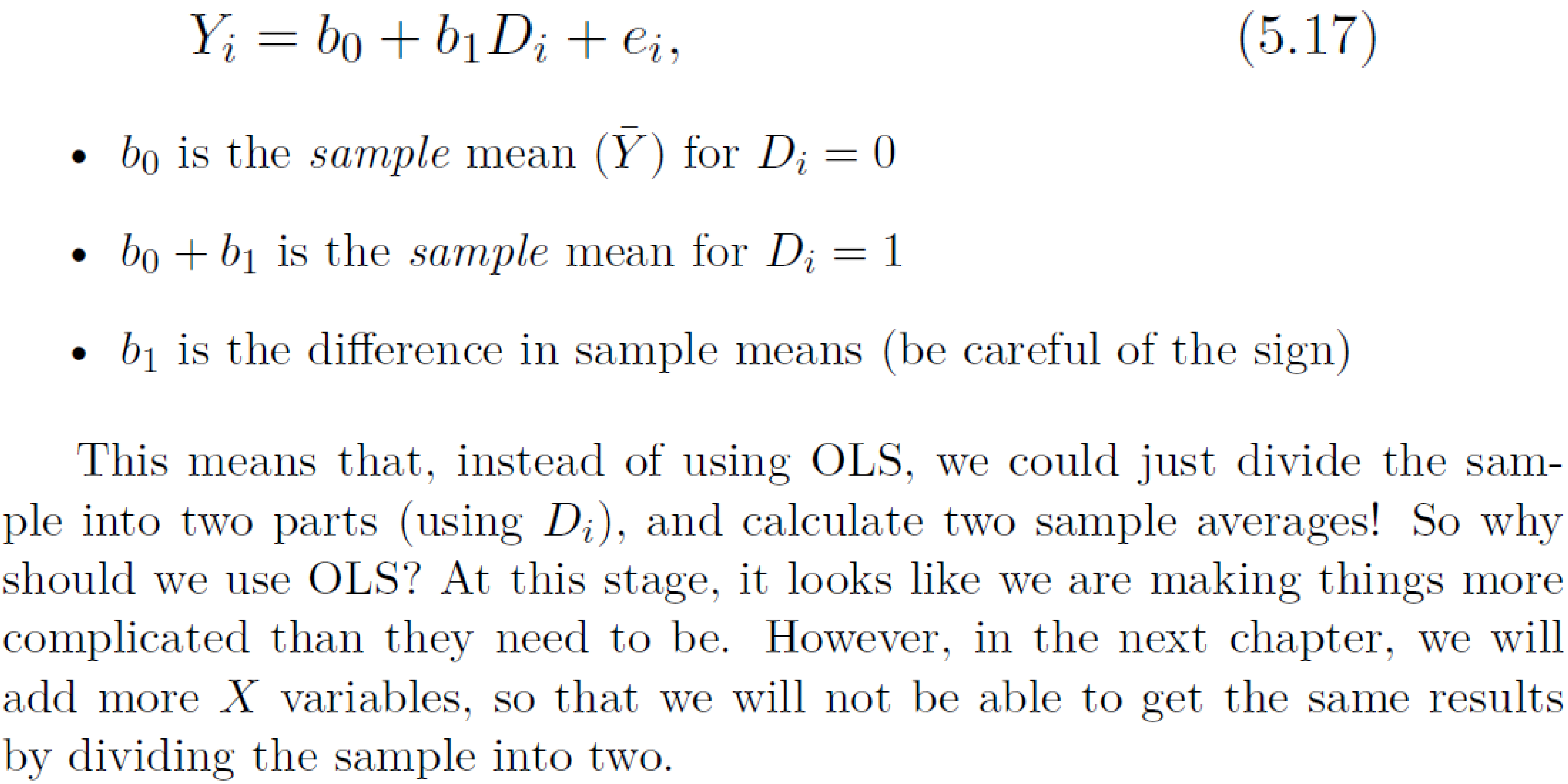
$$Y_i = \beta_0 + \beta_1 D_i + \epsilon_i, \tag{5.13}$$

- What is the interpretation of β_1 here?
- Take a derivative?
- What about β_0?
- Use *conditional expectations*

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

An estimated model with a dummy variable

Use OLS as before.

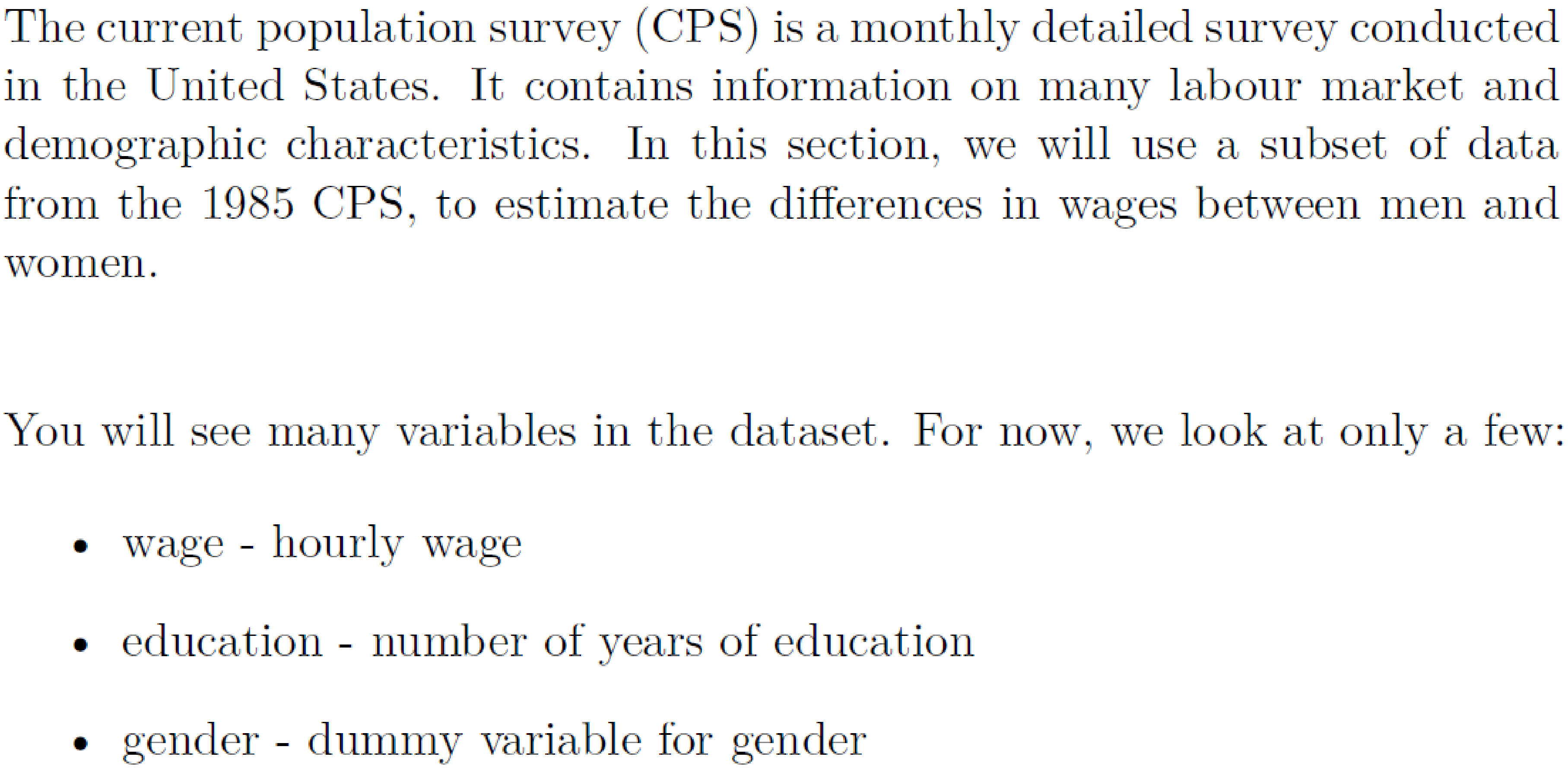


$$Y_i = b_0 + b_1 D_i + e_i, \tag{5.17}$$

- 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.

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

The data is available from the R package **AER** (Kleiber and Zeileis, 2008). To load this package, and the CPS data into R, use the following commands:

```
install.packages("AER")
library(AER)
data("CPS1985")
attach(CPS1985)
```

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

```
summary(lm(wage ~ gender))
```

You should see the following output:

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)    9.9949     0.2961   33.75  < 2e-16 ***
genderfemale  -2.1161     0.4372   -4.84  1.7e-06 ***
---
```

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

- What is the sample mean wage for males and for females?
- What is the interpretation of b_1?

In R, verify that OLS with a dummy variable is equivalent to taking the sample mean of the two groups.

```
mean(wage[gender=="male"])
```

```
mean(wage[gender=="female"])
```

5.4 Reporting regression results

$$\begin{aligned} \hat{wage} &= 10.00 - 2.12 \times gender, \;\; R^2 = 0.042 \\ &\;\;\;\,(0.30) \quad\;\; (0.44) \end{aligned} \tag{5.18}$$

This equation contains:

- Estimated *β*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