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

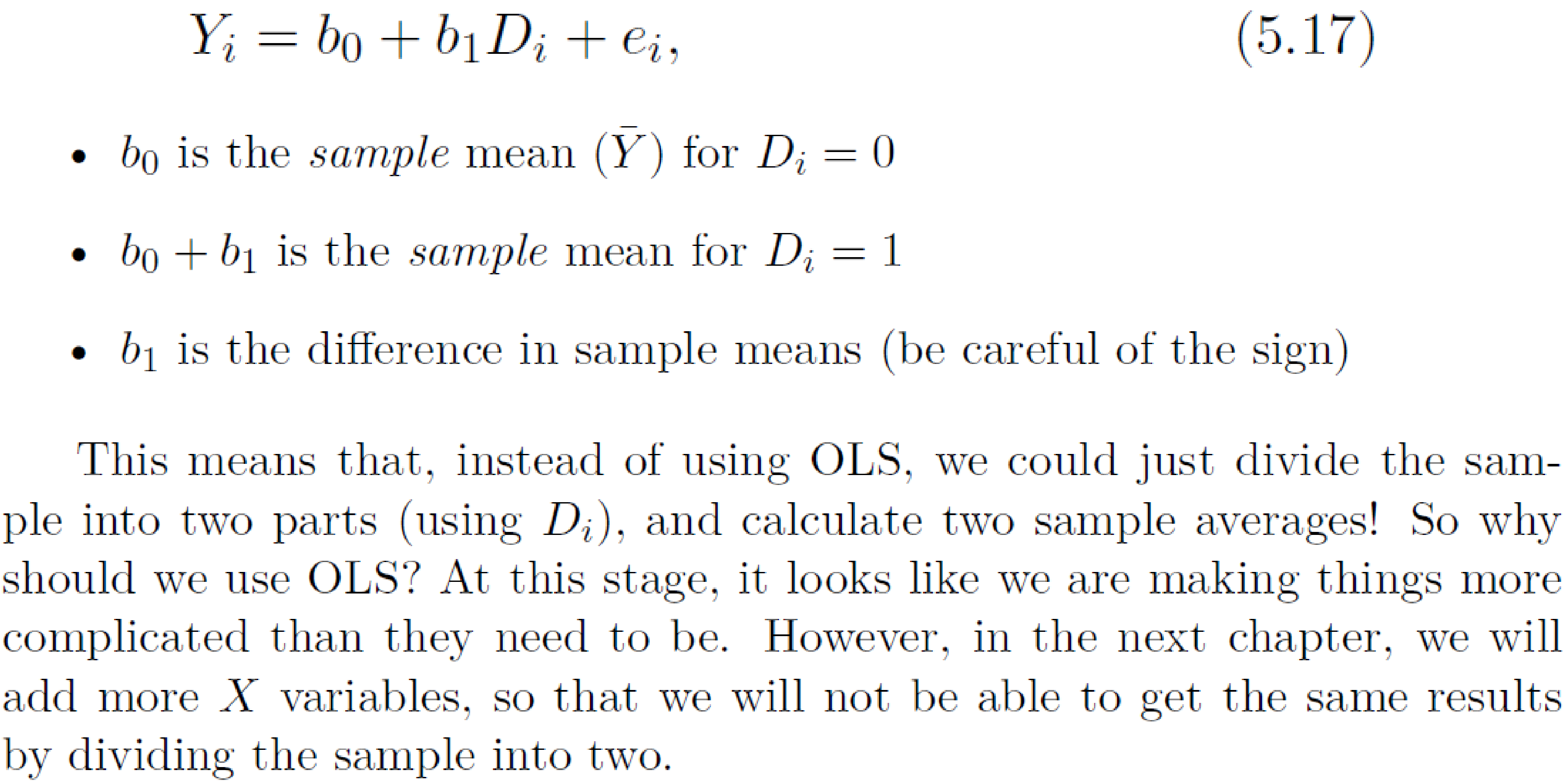
$$Y_i = \beta_0 + \beta_1 D_i + \epsilon_i, \tag{5.13}$$

- 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.

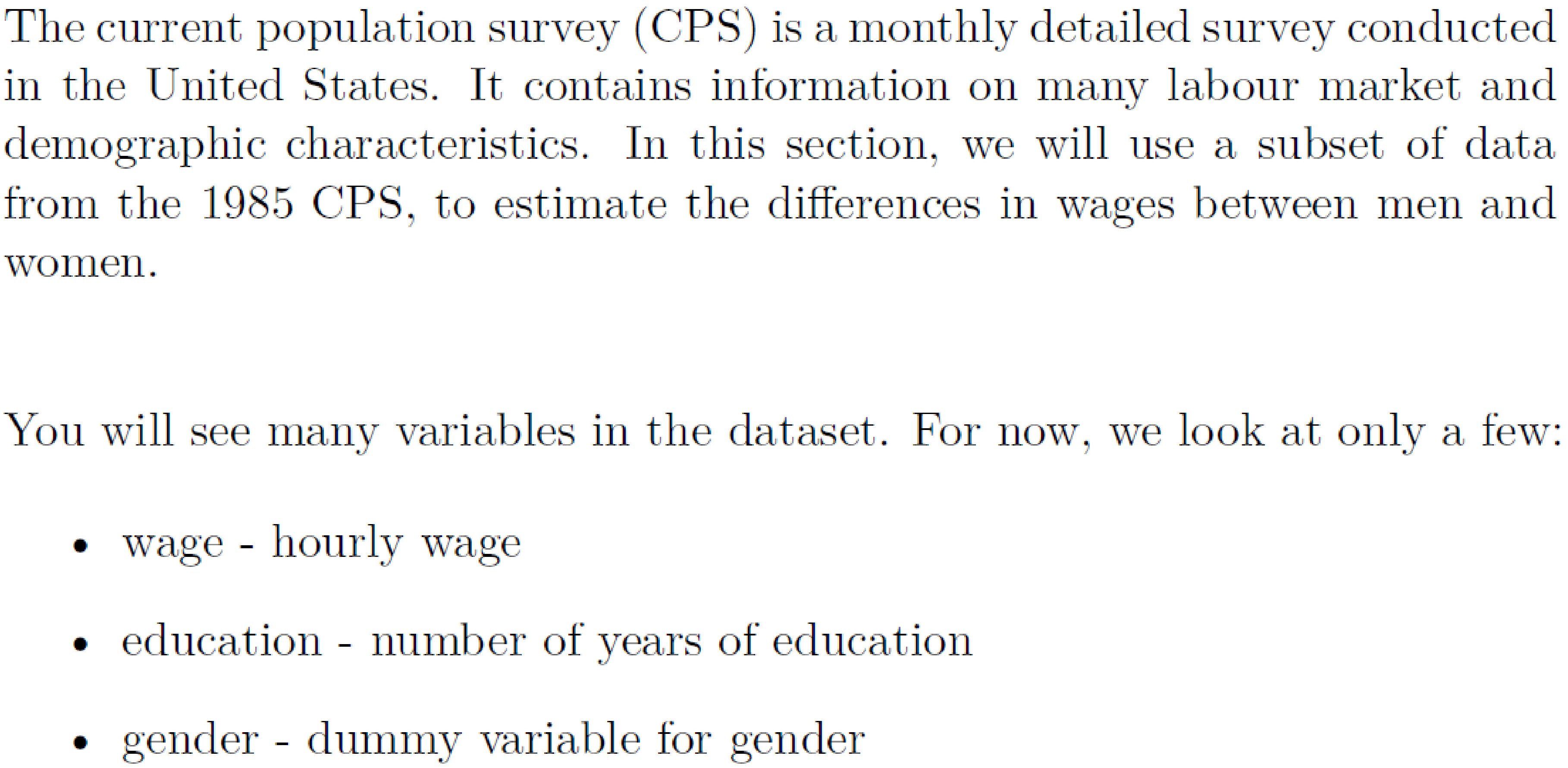


$$Y_i = b_0 + b_1 D_i + e_i, \tag{5.17}$$

- 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

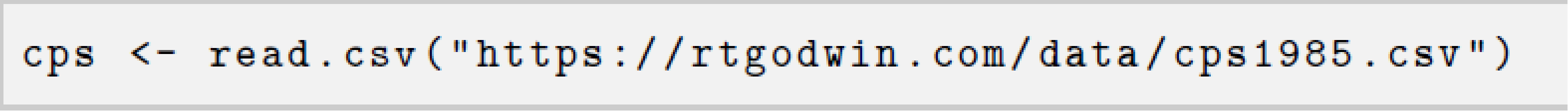


The current population survey (CPS) is a monthly detailed survey conducted in the United States. It contains information on many labour market and demographic characteristics. In this section, we will use a subset of data from the 1985 CPS, to estimate the differences in wages between men and women.

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

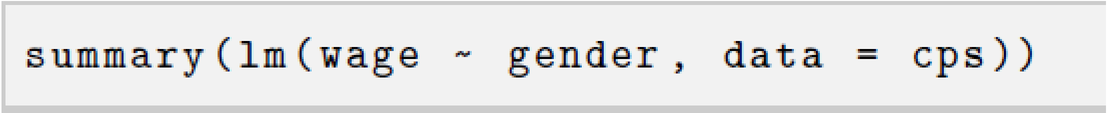
- wage - hourly wage
- education - number of years of education
- gender - dummy variable for gender

Load the data:

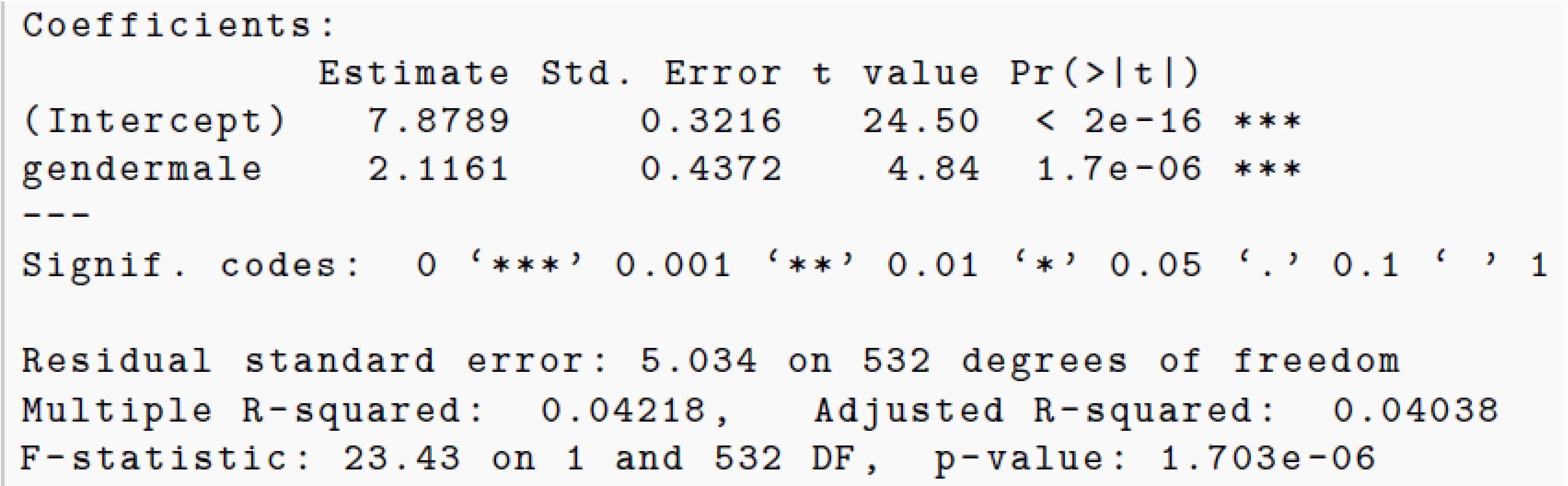


```
cps <- read.csv("https://rtgodwin.com/data/cps1985.csv")
```

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

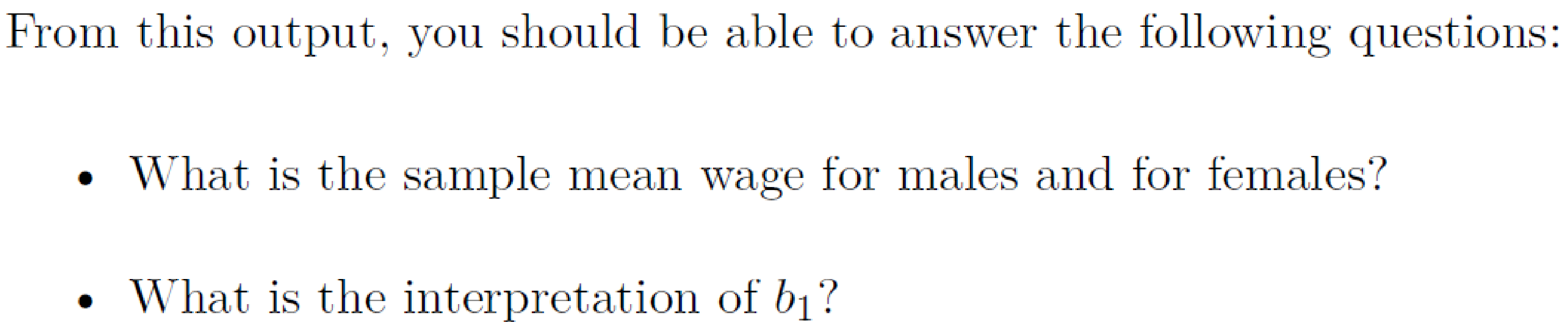


```
summary(lm(wage ~ gender, data = cps))
```



```
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
```



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

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:

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

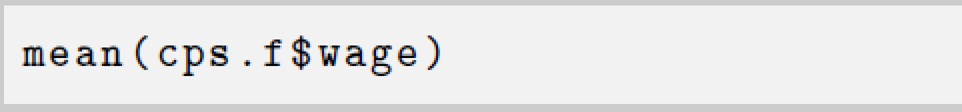
then take the difference in the sample mean wage between men and women:

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

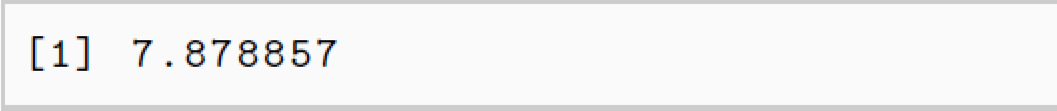
```
[1] 2.116056
```

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:

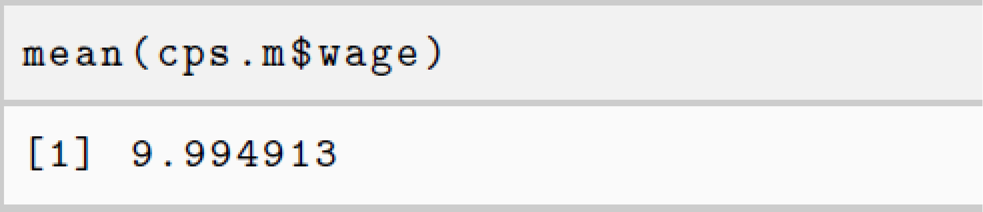


```
mean(cps.f$wage)
```



```
[1] 7.878857
```

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



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

```
[1] 9.994913
```

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

$$\begin{array}{l} \hat{wage} = \underset{(0.32)}{7.88} + \underset{(0.44)}{2.12} \times gendermale, \ R^2 = 0.042 \end{array}$$

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