

3.3.1 Qualitative Predictors

Josie Clarke & Daniel Camacho

February 14th, 2020

Including Qualitative Variables into a Regression Model

Qualitative or factor variables as predictors:

- ▶ Gender
- ▶ Political affiliation
- ▶ Student status

They can be included in the model by creating an indicator or dummy variable.

Let's study the following scenarios:

- ▶ Predictors with only two levels
- ▶ Predictors with more than two levels

Predictors with Only Two Levels

Consider the *Credit* data set

- ▶ The aim here is to predict which customers will default on their credit card debt.

Based on the *gender* variable:

$$x_{i1} = \begin{cases} 1 & \text{if the } i\text{th person is female} \\ 0 & \text{if the } i\text{th person is male} \end{cases}$$

Regression equation:

$$y_i = \beta_0 + \beta_1 x_{i1} + \epsilon_i = \begin{cases} \beta_0 + \beta_1 + \epsilon_i & \text{if the } i\text{th person is female} \\ \beta_0 + \epsilon_i & \text{if the } i\text{th person is male} \end{cases}$$

Least squares coefficient estimates

Gender is encoded as a dummy variable in the model (Table 3.7)

```
## # A tibble: 2 x 5
```

##	term	estimate	std.error	statistic	p.value
##	<chr>	<dbl>	<dbl>	<dbl>	<dbl>
## 1	(Intercept)	510.	33.1	15.4	2.91e-42
## 2	GenderFemale	19.7	46.1	0.429	6.69e- 1

Average credit card debt:

- ▶ Males: \$509.80
- ▶ Females: $\$509.80 + \$19.73 = \$529.53$

p-value for the dummy variable is very high

Arbitrary Coding Scheme

The main difference is the way in which the coefficients are interpreted

$$x_{i1} = \begin{cases} 1 & \text{if the } i\text{th person is female} \\ 0 & \text{if the } i\text{th person is male} \end{cases}$$

$$x_{i1} = \begin{cases} 1 & \text{if the } i\text{th person is female} \\ -1 & \text{if the } i\text{th person is male} \end{cases}$$

Regression equation:

$$y_i = \beta_0 + \beta_1 x_{i1} + \epsilon_i = \begin{cases} \beta_0 + \beta_1 + \epsilon_i & \text{if the } i\text{th person is female} \\ \beta_0 - \beta_1 + \epsilon_i & \text{if the } i\text{th person is male} \end{cases}$$

Equations

$$x_{i1} = \begin{cases} 1 & \text{if the } i\text{th person is Asian} \\ 0 & \text{if the } i\text{th person is not Asian} \end{cases}$$

$$x_{i2} = \begin{cases} 1 & \text{if the } i\text{th person is Caucasian} \\ 0 & \text{if the } i\text{th person is not Caucasian} \end{cases}$$

$$y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \epsilon_i$$

$$= \begin{cases} \beta_0 + \beta_1 + \epsilon_i & \text{if the } i\text{th person is Asian} \\ \beta_0 + \beta_2 + \epsilon_i & \text{if the } i\text{th person is Caucasian} \\ \beta_0 + \epsilon_i & \text{if the } i\text{th person is African American} \end{cases}$$

Slide with Plot

