Week 5 Reading Guide: Multiple Regression

|  |
| --- |
| Answers in different color text |
| It is **really** helpful for me if you make your answers in a different color, so I can easily pick them out! |

|  |
| --- |
| Tip |
| It is important you understand both types of multiple linear regression models, but I would highly recommend spending extra time understanding the particular type of model that your midterm project data will be analyzed with. For example, if you have two numerical explanatory variables, I would dedicate extra time to understanding the ins and outs of Section 6.2. |

# 6.1. – One numerical and one categorical explanatory variable

## 6.1.1 – Exploratory data analysis

**How was gender included in the scatterplot? Specifically, what was added to the R code to include this variable?**

**How were the standard error lines (grey bars) surrounding the regression lines removed? Specifically, what was added to the R code to accomplish this task?**

**Based on the colored scatterplot, how would you describe the relationship between teaching evaluation scores, age, and sex?**

## 6.1.2 – Interaction model

**What type of regression model is fit when you use a \* between your explanatory variables? (i.e., lm(response ~ var1\*var2))**

**In Table 6.3 what does the “Intercept” represent? Specifically, what does the estimate of 4.883 refer to?**

**In Table 6.3 what does the age row represent? Specifically, what does the estimate of -0.018 refer to?**

**In Table 6.3 what does the gender:male row represent? Specifically, what does the estimate of -0.446 refer to?**

**In Table 6.3 what does the age:gendermale represent? Specifically, what does the estimate of 0.014 refer to?**

**In Chapter 5, we saw indicator functions to adjust the mean life expectancy for each continent. How are indicator functions used in an interaction model? Specifically, where do they appear in the regression equation?**

## 6.1.3 – Parallel slopes model

**How is a “parallel slope” model different from an interaction model?**

**What R function do you use to plot lines with parallel slopes?**

**Compare the parallel slopes scatterplot (Figure 6.2) to the different slopes scatterplot (Figure 6.1). How different are they? Purely based on your eyes, which plot do you think better represents the relationships in the data?**

**How do you fit a parallel slopes model in R. Specifically, what symbol separates your two explanatory variables?**

**In a parallel slopes model, what line is missing from the coefficient table (compared to th interaction model)?**

# 6.2 – Two numerical explanatory variables

## 6.2.1 – Exploratory data analysis

**What does the cor() function give you?**

**What is “multicollinearity”?**

**The book makes two scatterplots, one for each variable. How could you have made a scatterplot with *both* variables?**

**Which variable has a stronger relationship with credit card debit?**

**When we interpret the slope for a model with more than one numerical variable, what additional statement must we make? In other words, can we interpret the impact of each variable separate from the others?**

# 6.3 – Related topics

## 6.3.1 – Model selection using visualization

**What is “Occam’s Razor”? How is it related to deciding what statistical model should be used?**

## 6.3.2 – Skip this section, we will discuss these ideas in Week 6

## 6.3.3 – Correlation coefficient

**If we were to change th units of a variable (e.g., grams to kilograms), would that change the correlation?**

## 6.3.4 – Simpson’s Paradox

**What is “Simpson’s Paradox”?**

**How does the visualization in Figure 6.11 exhibit Simpson’s Paradox?**