Math 300 NTI Lesson 14

Multiple Linear Regression - Numerical & Discrete

Professor Bradley Warner

June, 2022

Contents

Objectives

1. Generate, plot, and explain the interaction model for two explanatory variables (one numerical and one categorical).
2. Generate, plot, and explain the parallel slopes model for two explanatory variables (one numerical and one categorical).
3. Generate a table of observations, fitted values, and residuals from a linear regression object.

Reading

Chapter 6 - 6.1

Lesson

Remember that you will be running this more like a lab than a lecture. You want them using `R` and answering questions. Have them open the notes rmd and work through it together.

Work through the learning check LC6.1.

- LC6.1 is not enough. We need to work through the example of the book and recreate the work.

- Explaining the terms in the interaction model is difficult. Spend time on this.
- Notice that the line for females stops at the extremes of the observed data in `ggplot2()`. You have to be careful about extrapolating. The assumption of linearity outside of the observed data is risky at best.
- The use of `+` and `*` in the regression formulas confuses students. These are not arithmetical operations but formula operations. If we want to arithmetically add the predictors we would have to use the identity function `I()`.

Setup

```
library(tidyverse)
library(moderndive)
library(skimr)
library(ISLR)
```

Recreate the analysis done in the book.

```
evals_ch6 <- evals %>%
  select(ID, score, age, gender)
```

Let's look at 5 random rows of data.

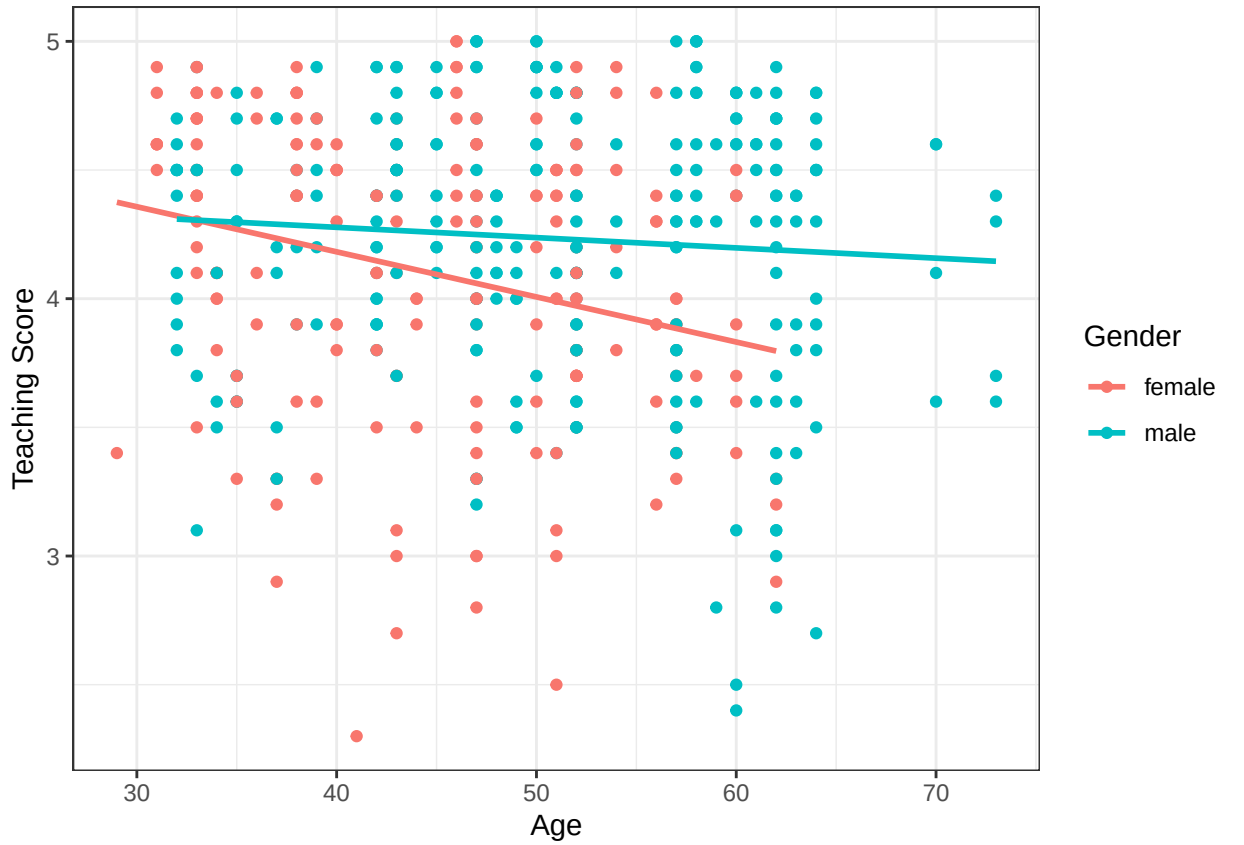
```
set.seed(941)
evals_ch6 %>%
  sample_n(size = 5)
```

```
## # A tibble: 5 x 4
##      ID score   age gender
##   <int> <dbl> <int> <fct>
## 1    61   3.7    35 male
## 2    15   3.9    40 female
## 3   309   3.6    35 male
## 4   274   4.2    57 male
## 5   256   4.1    52 male
```

- Interaction Model (Objective 1)

In this model we allow a different slope and intercept for each gender.

```
ggplot(evals_ch6, aes(x = age, y = score, color = gender)) +
  geom_point() +
  labs(x = "Age", y = "Teaching Score", color = "Gender") +
  geom_smooth(method = "lm", se = FALSE) +
  theme_bw()
```



To get the model in R, we use the * which is not multiplication but an interaction term in the model formula.

```
# Fit regression model:
score_model_interaction <- lm(score ~ age * gender, data = evals_ch6)
```

```
# Get regression table:
get_regression_table(score_model_interaction)
```

```
## # A tibble: 4 x 7
##   term           estimate std_error statistic p_value lower_ci upper_ci
##   <chr>             <dbl>     <dbl>     <dbl>   <dbl>    <dbl>    <dbl>
## 1 intercept          4.88      0.205     23.8    0         4.48     5.29
## 2 age               -0.018     0.004     -3.92   0        -0.026   -0.009
## 3 gender: male      -0.446     0.265     -1.68   0.094    -0.968    0.076
## 4 age:gendermale     0.014     0.006      2.45   0.015     0.003    0.024
```

Go through and explain this table. From the reading: *We say there is an interaction effect if the associated effect of one variable depends on the value of another variable.*

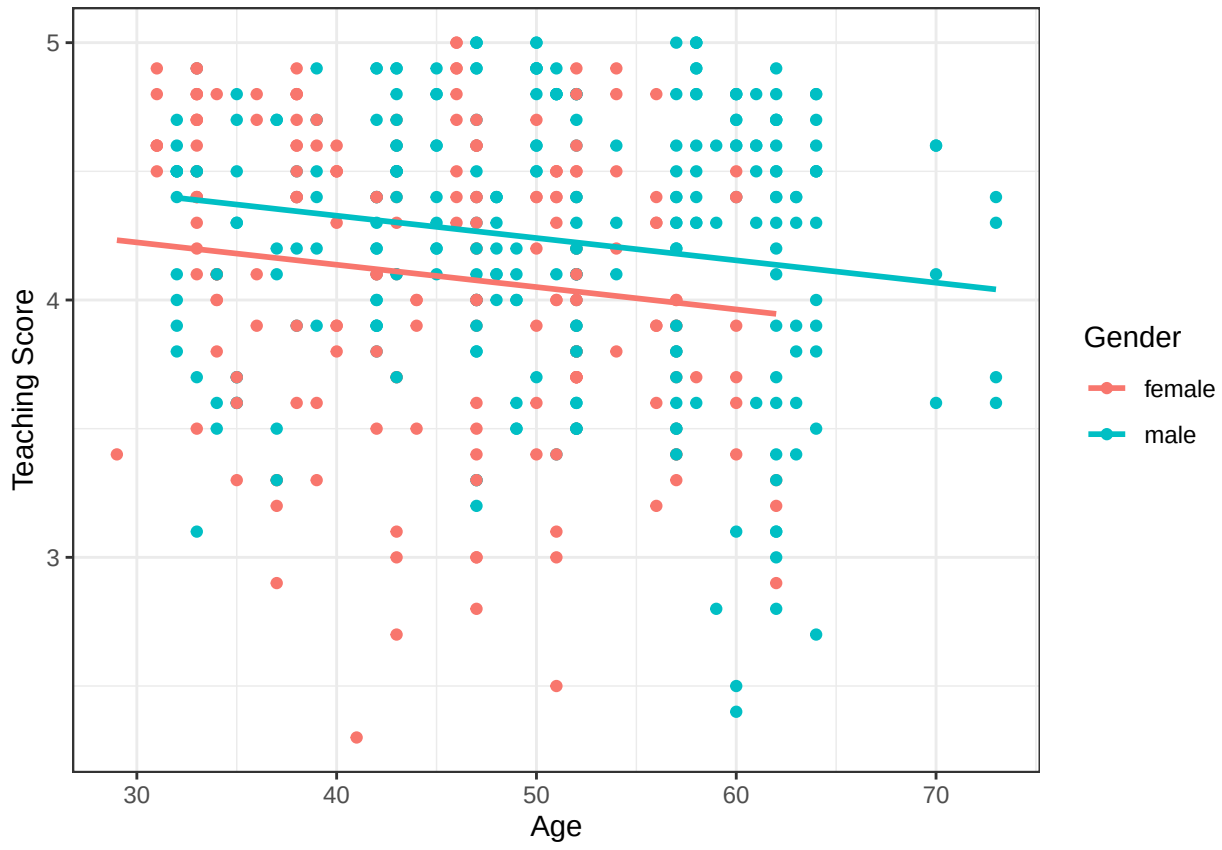
- What is the intercept for the males?
- What is the slope for the males?
- Interpret the slope for the female instructors?

- Parallel Slopes Model (Objective 2)

The parallel slopes model assumes that there is no interaction between the two explanatory variables. Their impact on the response is not related to the values of the other variable.

We will use the same data, but just build a different model.

```
ggplot(evals_ch6, aes(x = age, y = score, color = gender)) +
  geom_point() +
  labs(x = "Age", y = "Teaching Score", color = "Gender") +
  geom_parallel_slopes(se = FALSE) +
  theme_bw()
```



- Notice that the line for females stops at the extremes of the observed data.

```
# Fit regression model:
score_model_parallel_slopes <- lm(score ~ age + gender, data = evals_ch6)
```

```
# Get regression table:
get_regression_table(score_model_parallel_slopes)
```

```
## # A tibble: 3 x 7
```

```
##   term          estimate std_error statistic p_value lower_ci upper_ci
##   <chr>            <dbl>     <dbl>     <dbl>   <dbl>    <dbl>    <dbl>
## 1 intercept       4.48      0.125     35.8    0        4.24     4.73
## 2 age            -0.009     0.003     -3.28   0.001   -0.014   -0.003
## 3 gender: male    0.191     0.052      3.63   0        0.087    0.294
```

- What is the intercept for the males?
- What is the slope for the males?
- Interpret the slope for the female instructors?

Why would we do this? It seems like we lose information if we force the two slopes to be the same. This is true, and for these data, we would not want to make this assumption. In a later section, we will discuss a case where we would want to force the slopes to be the same.

LC 6.1 (Objective 3)

(LC6.1) Compute the observed values, fitted values, and residuals not for the interaction model as we just did, but rather for the parallel slopes model we saved in `score_model_parallel_slopes`.

Solution:

```
regression_points_parallel <- get_regression_points(score_model_parallel_slopes)
```

```
head(regression_points_parallel)
```

```
## # A tibble: 6 x 6
##      ID score   age gender score_hat residual
##   <int> <dbl> <int> <fct>      <dbl>    <dbl>
## 1     1   4.7    36 female      4.17    0.528
## 2     2   4.1    36 female      4.17   -0.072
## 3     3   3.9    36 female      4.17   -0.272
## 4     4   4.8    36 female      4.17    0.628
## 5     5   4.6    59 male        4.16    0.437
## 6     6   4.3    59 male        4.16    0.137
```

Documenting software

- File creation date: 2022-06-24
- R version 4.1.3 (2022-03-10)
- `tidyverse` package version: 1.3.1
- `skimr` package version: 2.1.4
- `ISLR` package version: 1.4
- `moderndive` package version: 0.5.4