Math 300 NTI Lesson 11

Simple Linear Regression - Continuous Predictor

Professor Bradley Warner

June, 2022

Contents

Objectives

1. Use the `skimr` package to summarize multiple numerical variables in a data frame.
2. Build a scatterplot to describe the relationship between two continuous, numerical variables; use `geom_smooth()` to visualize the best fit line.
3. Fit a linear regression model between two variables using the `lm()` function and interpret the output. This includes the interpretation of slope and the use of *association* and not *causation*.
4. Generate a table of observations, fitted values, and residuals from a linear regression object.

Reading

Chapter 5 - 5.1

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 checks LC5.1 - LC5.3.

- Regression can be used for explanatory and predictive purposes. It falls on that line between traditional statistics/econometrics and machine learning. In this course we focus on its more traditional use to interpret the relationship between predictors and a response. Math 378 is our machine learning course and expands on linear regression in this framework.

- Note the many different terms for `x` and `y` in regression. These names come from different fields. For example, `y` is called the response, dependent variable, outcome, and output. Meanwhile, `x` is called input, predictor, independent variable, and explanatory variable. Also point out that in linear regression, `y` is numerical while `x` can be numerical or categorical.
- We are using new packages. The `tidyverse` package is a wrapper and actually loads `readr`, `dplyr`, `ggplot2`, and `tidyr`.
- In the reading, the authors setup the problem with instructor teaching score as the response and beauty score as the explanatory variable. What is the research question?
- The reading introduces tilde `~` as a formula. You might want to talk about this as we use it in LC 5.1.
- The interpretation of the slope has the key phrase **average**. For a one unit change in `x`, the average value of `y` changes by the value of the slope.

Setup

```
library(tidyverse)
library(moderndive)
library(skimr)
library(gapminder)
```

Create the data needed for the exercises.

```
evals_ch5 <- evals %>%
  select(ID, score, bty_avg, age)
```

Let's look at 5 random rows of data.

```
set.seed(1234)
evals_ch5 %>%
  sample_n(size = 5)
```

```
## # A tibble: 5 x 4
##      ID score bty_avg   age
##   <int> <dbl>   <dbl> <int>
## 1   284   4      1.67    34
## 2   336   3.1    1.67    60
## 3   406   5      2.83    57
## 4   101   4.4    4.33    48
## 5   111   3.5    4.33    57
```

LC 5.1 (Objective 1)

(LC5.1) Refer to the example in section 5.1.1. Conduct a new exploratory data analysis with the same outcome variable y being `score` but with `age` as the new explanatory variable x. Remember, this involves three things:

- Looking at the raw data values.
- Computing summary statistics.

- Creating data visualizations.

What can you say about the relationship between age and teaching scores based on this exploration?

Solution:

- Looking at the raw data values:

```
glimpse(evals_ch5)
```

```
## Rows: 463
## Columns: 4
## $ ID      <int> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18,~
## $ score   <dbl> 4.7, 4.1, 3.9, 4.8, 4.6, 4.3, 2.8, 4.1, 3.4, 4.5, 3.8, 4.5, 4.~
## $ bty_avg <dbl> 5.000, 5.000, 5.000, 5.000, 3.000, 3.000, 3.000, 3.333, 3.333,~
## $ age     <int> 36, 36, 36, 36, 59, 59, 59, 51, 51, 40, 40, 40, 40, 40, 40, 40~
```

- Computing summary statistics:

```
my_skim<-skim_with(numeric = sfl(hist = NULL))
```

```
evals_ch5 %>%
  select(score, age) %>%
  my_skim() %>%
  print()
```

```
## -- Data Summary ------------------------
##                            Values
## Name                       Piped data
## Number of rows             463
## Number of columns          2
## _______________________
## Column type frequency:
##   numeric                  2
## ________________________
## Group variables            None
##
## -- Variable type: numeric ------------------------------------------------------
##   skim_variable n_missing complete_rate  mean    sd   p0  p25  p50  p75 p100
## 1 score                 0             1  4.17 0.544  2.3  3.8  4.3  4.6    5
## 2 age                   0             1 48.4  9.80  29   42   48   57     73
```

```
## $numeric
##
## -- Variable type: numeric ------------------------------------------------------
##   skim_variable n_missing complete_rate  mean    sd   p0  p25  p50  p75 p100
## 1 score                 0             1  4.17 0.544  2.3  3.8  4.3  4.6    5
## 2 age                   0             1 48.4  9.80  29   42   48   57     73
```

(Note that for formatting purposes, the inline histogram that is usually printed with skim() has been removed.)

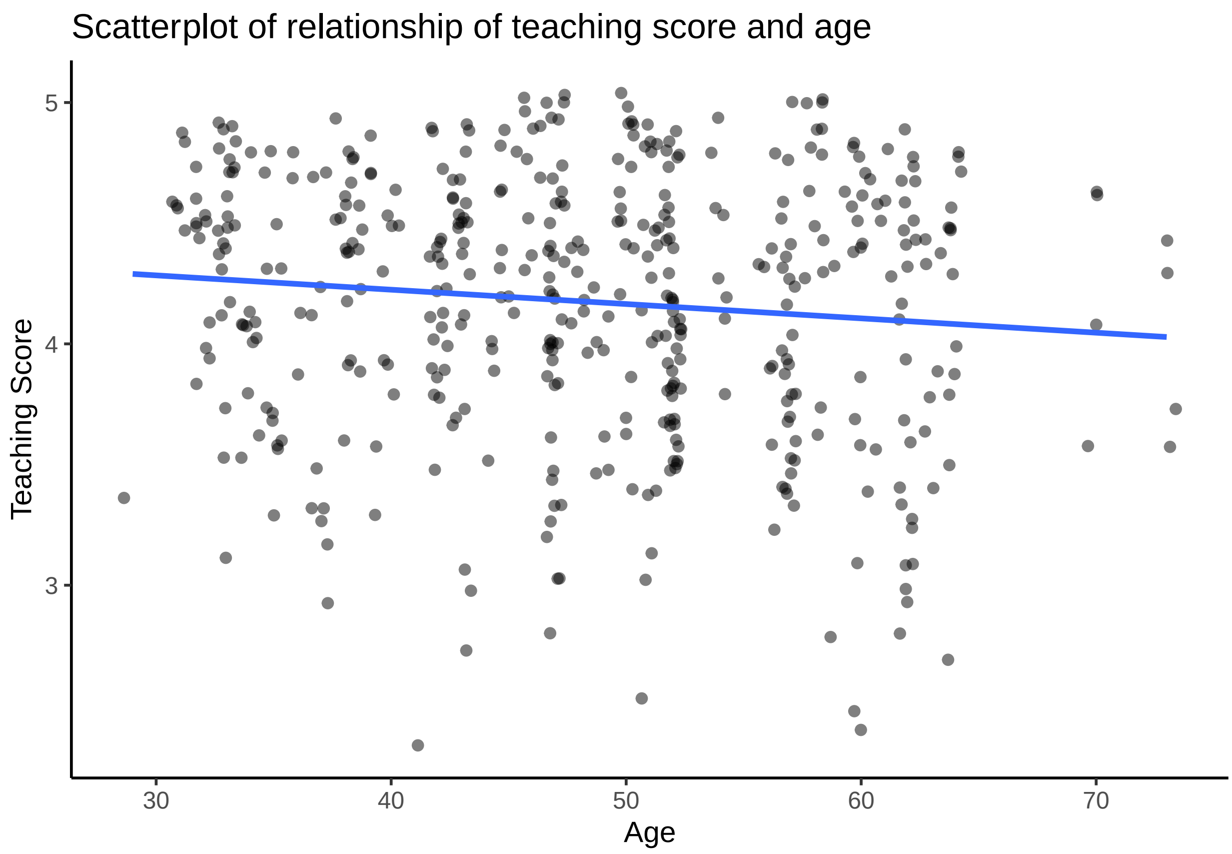
- Bivariate summary:

```
evals_ch5 %>%
  get_correlation(formula = score ~ age)
```

```
## # A tibble: 1 x 1
##      cor
##    <dbl>
## 1 -0.107
```

- Creating data visualizations:

```
ggplot(evals_ch5, aes(x = age, y = score)) +
  geom_jitter(alpha=0.5) +
  labs(
    x = "Age", y = "Teaching Score",
    title = "Scatterplot of relationship of teaching score and age") +
  geom_smooth(method = "lm", se = FALSE) +
  theme_classic()
```



Based on the scatterplot, there does not appear to be a relationship between age and teaching score. If anything, there might be a slight negative linear trend. That is, as age increases, the **average** teaching score decreases slightly. Even thought the correlation coefficient is negative, it is small in absolute value and thus there may be no relationship between the variables.

LC 5.2 (Objective 2)

(LC5.2) Fit a new simple linear regression using `lm(score ~ age, data = evals_ch5)` where `age` is the new explanatory variable `x`. Get information about the "best-fitting" line from the regression table by applying the `get_regression_table()` function. How do the regression results match up with the results from your earlier exploratory data analysis?

Solution:

```
# Fit regression model:
score_age_model <- lm(score ~ age, data = evals_ch5)
```

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

```
## # A tibble: 2 x 7
##   term      estimate std_error statistic p_value lower_ci upper_ci
##   <chr>        <dbl>     <dbl>     <dbl>   <dbl>    <dbl>    <dbl>
## 1 intercept    4.46      0.127     35.2    0         4.21     4.71
## 2 age         -0.006     0.003     -2.31   0.021    -0.011   -0.001
```

$$
\begin{aligned}
\widehat{y} &= b_0 + b_1 \cdot x \\
\widehat{\text{score}} &= b_0 + b_{\text{age}} \cdot \text{age} \\
&= 4.462 - 0.006 \cdot \text{age}
\end{aligned}
$$

For every increase of 1 year in `age`, there is an *associated* decrease of 0.006 units of the **average** teaching `score`. It matches with the results from our earlier exploratory data analysis.

LC 5.3 (Objective 3)

(LC5.3) Generate a data frame of the residuals of the model where you used `age` as the explanatory x variable.

Solution:

```
score_age_regression_points <- get_regression_points(score_age_model)
```

```
head(score_age_regression_points)
```

```
## # A tibble: 6 x 5
##      ID score   age score_hat residual
##   <int> <dbl> <int>     <dbl>    <dbl>
## 1     1   4.7    36      4.25    0.452
## 2     2   4.1    36      4.25   -0.148
## 3     3   3.9    36      4.25   -0.348
## 4     4   4.8    36      4.25    0.552
## 5     5   4.6    59      4.11    0.488
## 6     6   4.3    59      4.11    0.188
```

Documenting software

- File creation date: 2022-06-21
- R version 4.1.3 (2022-03-10)
- `tidyverse` package version: 1.3.1
- `skimr` package version: 2.1.4
- `gapminder` package version: 0.3.0

- `moderndive` package version: 0.5.4