

Math 300 Lesson 11 Notes

Simple Linear Regression - Continuous x

YOUR NAME HERE

June, 2022

Contents

Objectives	1
Reading	1
Lesson	1
Documenting software	3

Objectives

1. Explore the relationship between 2 variables using summary statistics and visualizations in R. This includes using the **skimr** package and fitting a best fit line in a scatterplot.
2. Fit a linear regression model to two variables using the **lm()** function and interpret the output. This includes the interpretation of slope and the use of *association* and not *cause*.
3. Generate a table of observations, fitted values, and residuals from a linear regression object.

Reading

Chapter 5 - 5.1

Lesson

Work through the learning checks LC5.1 - LC5.3. Complete the code as necessary.

- 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. While **x** is called input, predictor, independent variable, and explanatory variable. Also point out that **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**.

- 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(moderndiver)
library(skimr)
library(gapminder)
```

Create the data needed for the exercises.

```
# Select only the columns needed. Complete the code and remove comment labels
#evals_ch5 <- evals %>%
# select(ID, _____, bty_avg, _____)
```

Let's look at 5 random rows of data.

```
# Complete the code and remove comment labels
#set.seed(1234)
#evals_ch5 %>%
# sample_n(size = _____)
```

LC 5.1 (Objective 1)

(LC 5.1) 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:

```
# Complete the code and remove comment labels
#glimpse(_____)
```

- Computing summary statistics:

```
# Complete the code and remove comment labels
#evals_ch5 %>%
# select(_____, _____) %>%
# skim()
```

- Bivariate summary:

```
# Complete the code and remove comment labels
#evals_ch5 %>%
# get_correlation(formula = _____ ~ _____)
```

- Creating data visualizations:

```
# Create scatterplot. Complete the code and remove comment labels
#ggplot(_____, aes(x = _____, y = score)) +
# geom_jitter(alpha=_____) +
# labs(
#   x = "_____", y = "Teaching Score",
#   title = "_____") +
# geom_smooth(method = "lm", se = _____) +
# theme_classic()
```

LC 5.2 (Objective 2)

(LC 5.2) 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:

```
# Complete the code and remove comment labels
# Fit regression model:
#score_age_model <- lm(_____ ~ _____, data = _____)
# Get regression table:
#get_regression_table(score_age_model)
```

LC 5.3 (Objective 3)

(LC 5.3) Generate a data frame of the residuals of the model where you used `age` as the explanatory x variable.

Solution:

Documenting software

- File creation date: 2022-06-01
- 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