Math 300 NTI Lesson 13

Simple Linear Regression - Related Topics

Professor Bradley Warner

June, 2022

Contents

Objectives 1
Reading 1
Lesson 1
Documenting software 3

Objectives

1. Explain and give an example of a confounding variable.
2. In linear regression, explain what best fit means and calculate the sum of squared errors.

Reading

Chapter 5.3 - 5.4

Lesson

This lesson gives you time to catch up on previous material.

Work through the learning check LC5.8.

- The correlation does not imply causation is an important idea, but decision makers and humans in general want to know causation. There are other courses that help, DOE and econometrics. We will not explore
- If there is time, play the correlation game.
- Answer LC questions and have them work on problem set.

Setup

```
library(tidyverse)
library(moderndive)
```

- Spurious correlations (Objective 1)

Spend some time talking about confounding variables. The spurious correlations website may give you some ideas.

LC 5.8 (Objective 2)

(LC5.8) Note in the following plot there are 3 points marked with dots along with:

- The "best" fitting solid regression line
- An arbitrarily chosen dotted line
- Another arbitrarily chosen dashed line

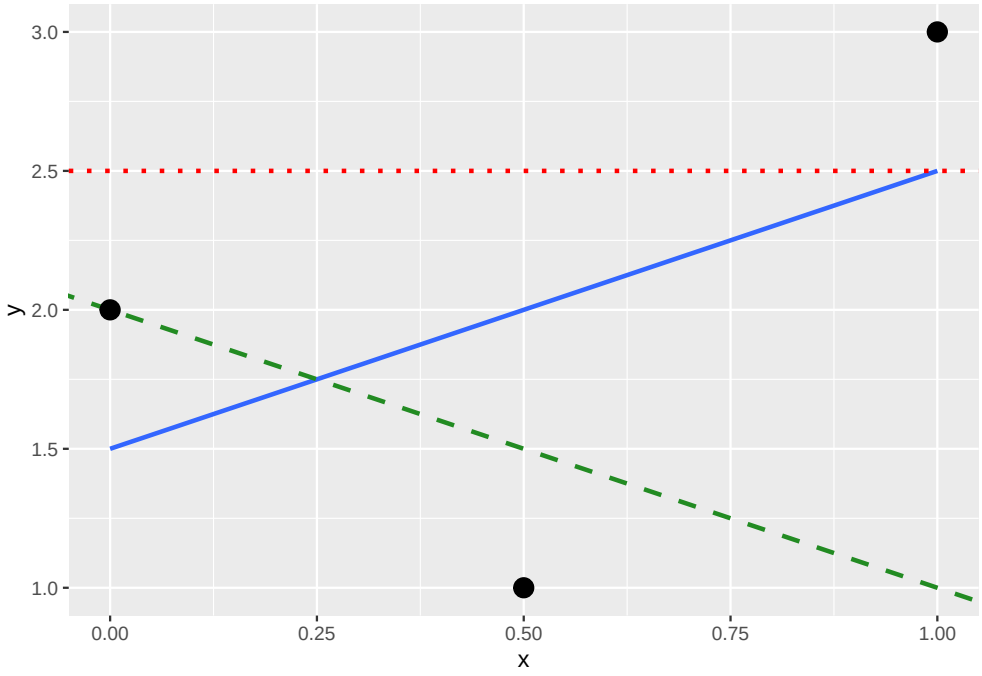

Figure 1: Regression line and two others.

Compute the sum of squared residuals for each line and show that of these three lines, the regression line has the smallest value.

Solution:

- The "best" fitting solid regression line :

$$\sum_{i=1}^{n}(y_i - \widehat{y}_i)^2 = (2.0 - 1.5)^2 + (1.0 - 2.0)^2 + (3.0 - 2.5)^2 = 1.5$$

```
sum((c(2, 1, 3)-c(1.5,2,2.5))^2)
```

```
## [1] 1.5
```

- An arbitrarily chosen dotted line:

$$\sum_{i=1}^{n}(y_i - \widehat{y}_i)^2 = (2.0 - 2.5)^2 + (1.00 - 2.5)^2 + (3.0 - 2.5)^2 = 2.75$$

```
sum((c(2, 1, 3)-c(2.5,2.5,2.5))^2)
```

```
## [1] 2.75
```

- Another arbitrarily chosen dashed line:

$$\sum_{i=1}^{n}(y_i - \widehat{y}_i)^2 = (2.0 - 2.0)^2 + (1.0 - 1.5)^2 + (3.0 - 1.0)^2 = 4.25$$

```
sum((c(2, 1, 3)-c(2,1.5,1))^2)
```

```
## [1] 4.25
```

As calculated, $1.5 < 2.75 < 4.25$. Therefore, we show that the regression line in blue has the smallest value of the residual sum of squares.

Documenting software

- File creation date: 2022-06-04
- R version 4.1.3 (2022-03-10)
- `tidyverse` package version: 1.3.1
- `moderndive` package version: 0.5.4