Problem Set 03: Simple Linear Regression Solutions

Professor Bradley Warner

June, 2022

Documentation:

We used all the resources available to instructors from the authors of Modern Dive.

Introduction

In this problem set we will work with a data set concerning hate crimes occurring in the US. For more context about these data, you can read a FiveThirtyEight article about these data that appeared in January of 2017: "Higher Rates Of Hate Crimes Are Tied To Income Inequality"

Here are the first three paragraphs of the article for context:

In the 10 days after the 2016 election, nearly 900 hate incidents were reported to the Southern Poverty Law Center, averaging out to 90 per day. By comparison, about 36,000 hate crimes were reported to the FBI from 2010 through 2015 — an average of 16 per day.

The numbers we have are tricky; the data is limited by how it's collected and can't definitively tell us whether there were more hate incidents in the days after the election than is typical. What we can do, however, is look for trends within the numbers, such as how hate crimes vary by state, as well as what factors within those states might be tied to hate crime rates.

An analysis of FBI and Southern Poverty Law Center data revealed one factor that stood out as a predictor of hate crimes and hate incidents in a given state: income inequality. States with more inequality were more likely to have higher rates of hate incidents per capita. This was true both before and after the election, and the connection held even after we controlled for other relevant variables.

In this problem set, we will develop a model these data using simple linear regression models with a single explanatory variable. These models are basic, but will allow us to practice using the tools and developing the skills we are learning.

Markdown File

Just like the first problem set, you will create an rmarkdown file; you can refer to the first problem set for instruction. Name your file with the following format **PS03_lastname_firstname_section**

Fill in your answers to the exercises below with a corresponding exercise header in your rmarkdown file. Be sure to include text *and* code where necessary.

Deliverable

- You will turn in your completed problem set as a **knitted pdf file** to Gradescope.

- Include a documentation statement.
- Each exercise is worth 3 points.
- Be sure to **make a header to label each Exercise**, this means using # followed by the word Exercise and the exercise number.
- Please type your code to answer the questions in a code chunk (gray part), under the exercise headers and type (**BRIEF**) answers to any interpretation questions in the white part under the headers.

Setup

First load the necessary packages

```
library(ggplot2)
library(dplyr)
library(moderndive)
library(readr)
```

Next, load the data set from where it is stored on the web:

```
hate_crimes <- read_csv("http://bit.ly/2ItxYg3")
```

You can take a glimpse at the data like so:

```
glimpse(hate_crimes)
```

```
## Rows: 51
## Columns: 9
## $ state            <chr> "New Mexico", "Maine", "New York", "Illinois", "Delaw~
## $ median_house_inc <chr> "low", "low", "low", "low", "high", "high", "high", "~
## $ share_pop_metro  <dbl> 0.69, 0.54, 0.94, 0.90, 0.90, 1.00, 0.87, 0.86, 0.97,~
## $ hs               <dbl> 83, 90, 85, 86, 87, 85, 89, 90, 81, 91, 89, 89, 87, 8~
## $ hate_crimes      <dbl> 0.295, 0.616, 0.351, 0.195, 0.323, 0.095, 0.833, 0.67~
## $ trump_support    <chr> "low", "low", "low", "low", "low", "low", "low", "low~
## $ unemployment     <chr> "high", "low", "low", "high", "low", "high", "high", ~
## $ urbanization     <chr> "low", "low", "high", "high", "high", "high", "high",~
## $ income           <dbl> 46686, 51710, 54310, 54916, 57522, 58633, 58875, 5906~
```

Be sure to also examine the data in the RStudio data viewer.

```
summary(hate_crimes)
```

```
##     state           median_house_inc   share_pop_metro        hs
##  Length:51          Length:51          Min.   :0.3100   Min.   :80.00
##  Class :character   Class :character   1st Qu.:0.6300   1st Qu.:84.00
##  Mode  :character   Mode  :character   Median :0.7900   Median :87.00
##                                        Mean   :0.7502   Mean   :86.79
##                                        3rd Qu.:0.8950   3rd Qu.:90.00
##                                        Max.   :1.0000   Max.   :92.00
##                                                         NA’s   :3
##   hate_crimes     trump_support      unemployment       urbanization
##  Min.   :0.0670   Length:51          Length:51          Length:51
##  1st Qu.:0.1430   Class :character   Class :character   Class :character
##  Median :0.2260   Mode  :character   Mode  :character   Mode  :character
##  Mean   :0.3041
##  3rd Qu.:0.3570
##  Max.   :1.5220
##  NA's   :4
##      income
##  Min.   :35521
##  1st Qu.:48657
##  Median :54916
##  Mean   :55224
##  3rd Qu.:60719
##  Max.   :76165
##
```

Notice that 4 states do not have data for the response variable. As a simple fix, let's remove these values. In practice, we would want to investigate the nature of why these states are missing values.

```
hate_crimes_ps <- hate_crimes %>%
  select(state, hate_crimes, share_pop_metro, urbanization) %>%
  na.omit()
```

About the data set

Each case/row in these data represents a state in the US. The response variable we will consider is `hate_crimes`, which is the number of hate crimes per 100k individuals in the 10 days after the 2016 US election as measured by the Southern Poverty Law Center (SPLC).

For this project week we will examine the explanatory strength of variables that represent the urbanization of the state. The first is numeric, `share_pop_metro` and the second is a categorical variable generated from the first, `urbanization`:

- `share_pop_metro`: a numeric variable that is the proportion of the state population that lives in a urban metropolitan setting.
- `urbanization`: a categorical variable that is classifies the level of urbanization in a state (low or high; split below or above mean)

Hate Crimes and Proportion of Population Living in Metropolitan Area

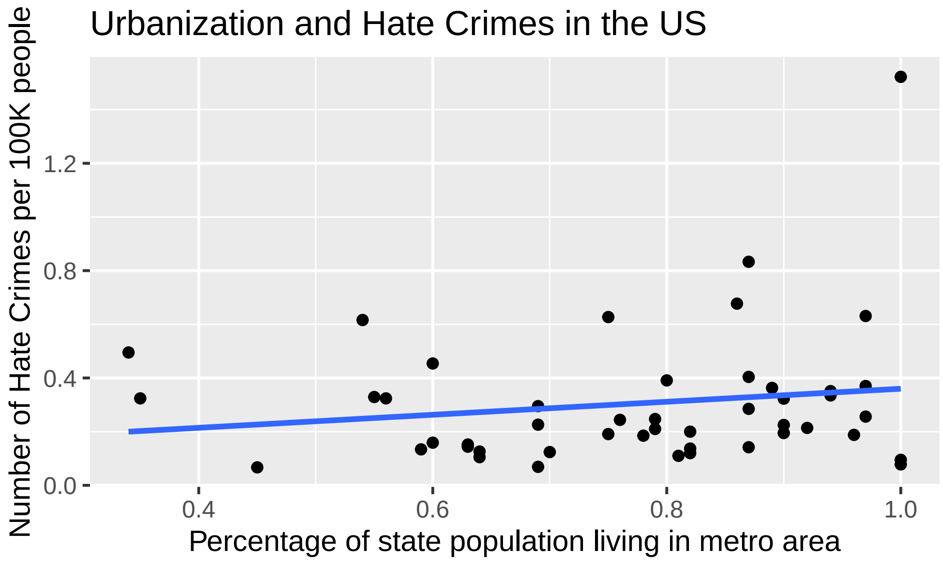
Let's start by modeling the relationship between:

- y: `hate_crimes` per 100K individuals
- x: Proportion of state's population living in a metropolitan area `share_pop_metro`

Exercise 1

Create a visualization that will allow you to conduct an "eyeball test" of the relationship between hate crimes per 100K and proportion of the population that lives in a metropolitan area. Include appropriate axes labels and a title.

```
ggplot(data = hate_crimes_ps, aes(y = hate_crimes, x = share_pop_metro)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE) +
  labs(x = "Percentage of state population living in metro area",
       y = "Number of Hate Crimes per 100K people",
       title = "Urbanization and Hate Crimes in the US")
```



Exercise 2

Now run a model that examines the relationship between hate crime rates and the proportion of the population living in a metro area. Generate a regression table.

```
urban_mod <- lm(hate_crimes ~ share_pop_metro, data = hate_crimes_ps)
```

```
get_regression_table(urban_mod)
```

```
## # 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          0.117     0.173     0.679   0.501   -0.231    0.466
## 2 share_pop_metro    0.242     0.22      1.10    0.276   -0.2      0.685
```

Exercise 3

The regression equation for this model is the following:

$$\widehat{y} = 0.117 + 0.242 \times \text{prop in metro area}$$

Interpret the slope coefficient in this regression table.

Answer:

- For every increase in proportion of population in metro area of 1 percentage point, there is an **associated increase** in hate crimes on average of .00242 per 100K people.

Exercise 4

What value does the model estimate for the number of hate crimes per 100,000 people in state with a proportion of 0.82 of the population living in a metropolitan area?

Answer:

```
y_hat = 0.117 + 0.242 * .82
y_hat
```

```
## [1] 0.31544
```

On average for a state with 82 percent of the population living in a metropolitan area, there are 0.31544 hate crimes per 100k people.

Exercise 5

Do you think it is a good idea to predict hate crimes in a state that has 25% of the population living in a metropolitan area, based on this regression equation? Explain your reasoning.

Answer:

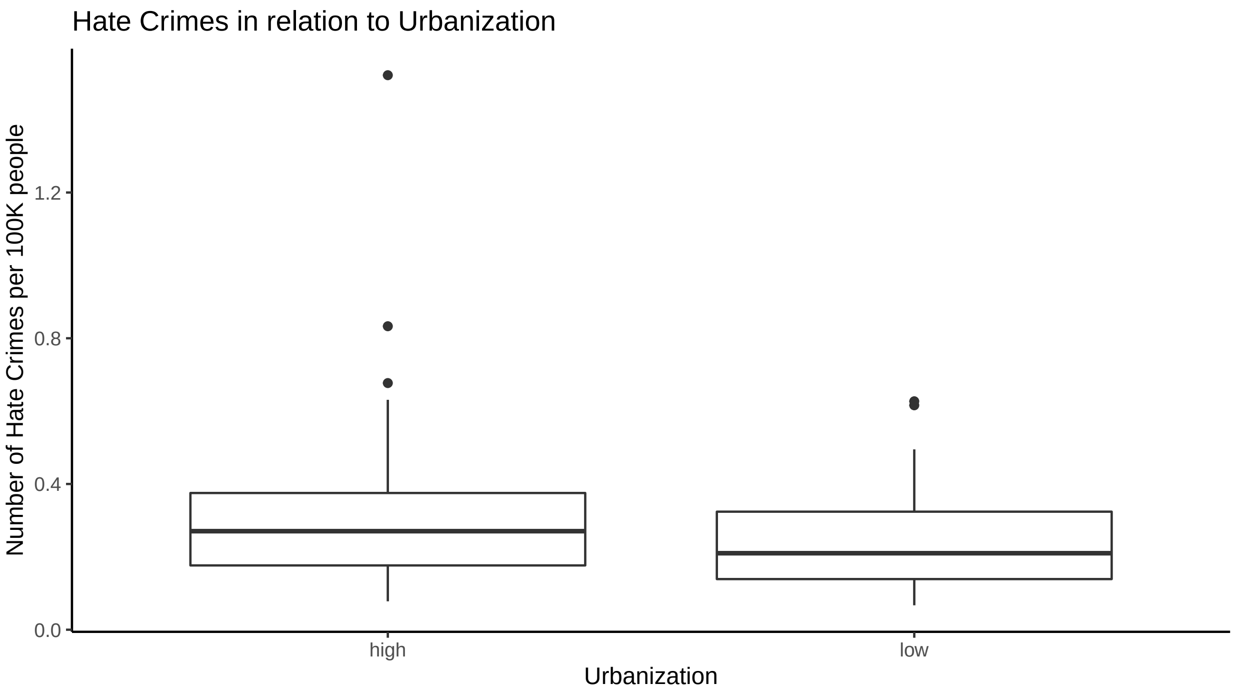
No, that would be extrapolation. Our data does not include values below 0.34 for the percent of the population living in a metropolitan area.

Hate Crimes and Urbanization

Exercise 6

Create a visualization, side-by-side boxplots, that will allow you to conduct an "eyeball test" of the relationship between hate crimes per 100K and urbanization. Include appropriate axes labels and a title.

```
ggplot(data = hate_crimes_ps, aes(x = urbanization, y = hate_crimes)) +
  geom_boxplot() +
  labs(x = "Urbanization",
       y = "Number of Hate Crimes per 100K people",
       title = "Hate Crimes in relation to Urbanization") +
  theme_classic()
```



Exercise 7

Now run a model that examines the relationship between hate crime rates and urbanization. Generate a regression table.

```
hate_mod <- lm(hate_crimes ~ urbanization, data = hate_crimes_ps)
```

```
get_regression_table(hate_mod)
```

```
## # 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            0.352     0.051      6.88   0        0.249    0.455
## 2 urbanization: low   -0.098     0.073     -1.34   0.188   -0.245    0.05
```

Exercise 8

What does the intercept mean in this regression table?

Answer:

The intercept is the average estimated hate crime rate for states with **high** urbanization and has the value 0.352.

Exercise 9

The regression equation for this model is the following:

$$\widehat{y} = 0.352 - 0.098 \times 1_{\text{low urbanization}}(x)$$

Find the average value of hate crimes for a state with low urbanization.

Answer:

```
y_hat = 0.352 - 0.098 * 1
y_hat
```

```
## [1] 0.254
```

Exercise 10

Based on our model, can we conclude that high level of urbanization cause higher levels of hate crime in a state?

Answer:

No there is an association but we don't know if there is a causal relationship. Correlation does not necessarily imply causation.