Problem Set 05

WRITE YOUR NAME HERE

WRITE DATE HERE

Contents

Question 10 **5**

Collaboration

Please indicate who, if anyone, you collaborated with on this problem set:

Background

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"

In this problem set, we will model these data using regression models with a single categorical predictor variable (i.e., explanatory variable).

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)
```

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

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).

This week we will examine the explanatory strength of three categorical variables in the data set:

- `trump_support`: level of Trump support in 2016 election (low, medium or high; roughly equal number of observation at each category level)
- `unemployment`: level of unemployment in a state (low or high; split below or above mean)
- `median_house_inc`: median household income in the state (low or high; split below or above mean)

Hate Crimes and Trump Support

Let's start by modeling the relationship between:

- y: `hate_crimes` per 100K individuals
- x: Level of `trump_support` in the state: `low`, `medium`, or `high`

Question 1.A

Create a visualization that will allow you to conduct an "eyeball test" of the relationship between hate crimes per 100K and level of Trump support. Include appropriate axes labels and a title. Please note that because of alphanumeric ordering, the levels of

`trump_support` are alphabetically ordered `high`, `low`, `medium`, and hence the baseline group is `high`. Although this default ordering is counter-intuitive, we will use the ordering as-is for this problem set.

Question 1.B

Comment on the relationship between these two variables. Is this what you would've expected?

Answer:

Question 2

Now run a model that examines the relationship between hate crime rates and the level of Trump support. Generate a regression table.

Question 3.A

What does the intercept mean in this regression table?

Answer:

Question 3.B

What value does the model estimate for the number of hate crimes per 100,000 people in states with "low" Trump support?

Answer:

Question 3.C

Does the model estimate that hate crimes are more frequent in states that show "low" or "high" support for Trump?

Answer:

Question 3.D

How much greater were hate crimes in "medium" trump-support states compared to "high" trump-support states?

Answer:

Question 3.E

What are the three possible fitted values $\hat{y}$ for this model? Hint: use the `get_regression_points()` function to explore this if you are not sure!

Answer:

Question 3.F

Below we calculate the group means of hate crimes for the `high`, `medium` and `low` levels of Trump support. How do these numbers compare to the three possible fitted values $\hat{y}$ for this model?

```
hate_crimes %>%
group_by(trump_support) %>%
  summarize(mean_hate_crimes = mean(hate_crimes, na.rm = T))
```

```
## # A tibble: 3 x 2
##   trump_support mean_hate_crimes
##   <chr>                    <dbl>
## 1 high                     0.191
## 2 low                      0.460
## 3 medium                   0.222
```

Answer:

The regression equation

The regression equation for this model is the following:

$$\hat{y} = 0.191 + 0.269 \times 1_{\text{low support}}(x) + 0.031 \times 1_{\text{med support}}(x)$$

So for instance, in a state in which `trump_support` is "low" you would plug in 1 for $1_{\text{low support}}(x)$, and 0 in for $1_{\text{med support}}(x)$ and solve like so:

$$\begin{aligned}
\hat{y} &= 0.191 + 0.269 \times 1 + 0.031 \times 0 \\
\hat{y} &= 0.191 + 0.269 + 0 \\
\hat{y} &= 0.460
\end{aligned}$$

Question 4

Solve the regression equation for a state in which `trump_support` is "high"

Answer:

Question 5

Which 5 states had the highest rate of hate crimes? What was the level of Trump support in these 5 states? You can solve this any way you choose, using code or not…

Do these results surprise you? (There is no right answer to this question.)

Answers:

Hate crimes and unemployment

We will next model the relationship between:

- y: `hate_crimes` per 100K individuals after the 2016 US election
- x: Level of unemployment in the state (`low`, or `high`)

Question 6

Create a visualization that will allow you to conduct an "eyeball test" of the relationship between hate crimes per 100K and unemployment level. Include appropriate axes labels and a title.

Question 7.A

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

Question 7.B

Write out the regression equation for the above model. Try writing out the equation using the same LaTeX formatting used to write the equation for the first hate crime ~ Trump support model. You can copy the example LaTeX code that appears just after "The regression equation for this model is the following:" in this document, and edit it to describe the hate crime ~ unemployment model you just fit. You don't *have* to write the equation using LaTeX, but this is a great statistical communication skill to practice!

Answer:

Question 7.C

What does the intercept mean in this regression table?

Answer:

Question 7.D

What does the model estimate as the number of hate crimes per 100,000 people in states with "low" unemployment?

Answer:

Question 7.E

What are the two possible fitted values $\hat{y}$ for this model? Why are there only two this time, instead of the three like the previous model?

Answer:

Question 8

Use the `get_regression_points()` function to generate a table showing the fitted values and the residuals for the model relating `hate_crimes` to unemployment. Examine the first row: How are the residuals calculated here?

Answer:

Hate crimes and household income

Question 9A

Now run a model that examines the relationship between `hate_crimes` and median household income in the state `median_house_inc`.

Question 9B

Were there more hate crimes in areas with high or low median household incomes? How large was the difference between states with "low" and "high" levels of household income?

Answer:

Question 10

Run the `get_regression_points()` function for the `hate_crimes` and `median_house_inc` model, and examine the data for Maine (row 2). Did the model **overpredict** or **underpredict** the `hate_crimes` level, compared to what was observed in the data?

Answer: