Problem Set 05: Regression with One Categorical Variable

Your Name

Last modified on August 24, 2024 10:29:36 Eastern Daylight Time

Background

In this problem set, hate crimes data from the US will be used. The FiveThirtyEight article about the data appears in the Jan 23, 2017 "Higher Rates Of Hate Crimes Are Tied To Income Inequality."

The crimes data will be used to run regression models with a single categorical predictor (explanatory) variable.

Setup

First load the necessary packages:

R Code

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

Next, the data is read into the object `hate_crimes` from where it is stored on the web using the `read_csv()` function.

R Code

```r
url <- "http://bit.ly/2ItxYg3"
if(!dir.exists("./data/")){
  dir.create("./data/")
}
if(!file.exists("./data/hate_crimes.csv")){
    download.file(url, destfile = "./data/hate_crimes.csv")}
hate_crimes <- read_csv("./data/hate_crimes.csv")
```

```r
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 **ALSO** to examine the data in the viewer.

Each case/row in these data is 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 use three categorical explanatory variables in this data set:

- `trump_support`: level of Trump support in 2016 election (low, medium or high - split into roughly equal number of cases)
- `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 median)

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`

Problem 1

Create a visual model of these data (a graph) 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 ordered `high`, `low`, `medium`, and hence the baseline group is `high`. Also note that we could "reorder" the levels to `low`, `medium`, `high`....though we will leave the levels as is for this Problem Set.

Problem 1 Answers

```
# Type your code and comments inside the code chunk
```

Problem 2

2. Comment on the relationship between `hate_crimes`, and `trump_support`. Is this what you would've expected?

Problem 2 Answers

- Delete this and put your text answer here.

Problem 3

Create a model that examines the relationship between hate crime rates and the level of Trump support. Store the results of your model in an object named `hate_mod`. Generate a regression table using `hate_mod`.

Problem 3 Answers

```
# Type your code and comments inside the code chunk
```

Problem 4

What does the intercept mean in this regression table?

Problem 4 Answers

- Delete this and put your text answer here.

Problem 5

What does the model estimate as the number of hate crimes per 100,000 people in states with “low” Trump support?

Problem 5 Answers

- Delete this and put your text answer here.

```
# Type your code and comments inside the code chunk
```

Problem 6

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

Problem 6 Answers

```
# Type your code and comments inside the code chunk
```

- Delete this and put your text answer here.

Problem 7

How much greater were hate crimes in “medium” Trump-support states compared to “high” Trump-support states?

Problem 7 Answers

```
# Type your code and comments inside the code chunk
```

- Delete this and put your text answer here.

Problem 8

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!

Problem 8 Answers

```
# Type your code and comments inside the code chunk
```

- Delete this and put your text answer here.
- Delete this and put your text answer here.
- Delete this and put your text answer here.

Problem 9

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 = TRUE))
```

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

Problem 9 Answers

- Delete this and put your text answer here.

The Regression Equation

The regression equation for this model is the following (render the document and look at the output!)

$$\widehat{\text{hate\_crimes}} = 0.191 + 0.2691111 \times 1_{\text{low support}}(x) + 0.0314 \times 1_{\text{med support}}(x)$$

Another notation is to use a dummy variables (D_1, and D_2) and write the equation as

$$\hat{y} = 0.191 + 0.2691111 \times D_1 + 0.0314 \times D_2$$

where

$$D_1 = \begin{cases} 1 & \text{for low Trump support} \\ 0 & \text{for medium Trump support} \\ 0 & \text{for high Trump support} \end{cases}$$

$$D_2 = \begin{cases} 0 & \text{for low Trump support} \\ 1 & \text{for medium Trump support} \\ 0 & \text{for high Trump support} \end{cases}$$

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 as follows:

$$\begin{aligned} \hat{y} &= 0.191 + 0.2691111 \times 1 + 0.0314 \times 0 \\ \hat{y} &= 0.191 + 0.2691111 + 0 \\ \hat{y} &= 0.4601111 \end{aligned}$$

Problem 10

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

Problem 10 Answers

Your $LaTeX$ here

```
# Type your code and comments inside the code chunk
```

- Delete this and put your text answer here.

Problem 11

Which 5 states had the highest rate of hate crimes? What was the level of Trump support in these 5 states? Solve this question programmatically. Note: The District of Columbia

is not actually a state!
Do these results surprise you? (There is no right answer to this question.)

Problem 11 Answers

```
# Type your code and comments inside the code chunk
```

- Delete this and put your text answer here.

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`)

Problem 12

Create a visual model of this data (a graph) 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.

Problem 12 Answers

```
# Type your code and comments inside the code chunk
```

Problem 13

Create a model that examines the relationship between hate crime rates and the unemployment level. Name this model `job_mod`. Generate a regression table for `job_mod`.

Problem 13 Answers

```
# Type your code and comments inside the code chunk
```

Problem 14

Write out the regression equation for `job_mod`.

Problem 14 Answers

Write your equation with $LaTeX$ here.

Problem 15

Interpret the estimate of the intercept from the table below.

Problem 15 Answers

- Delete this and put your text answer here.

Problem 16

What does the model estimate the number of hate crimes per 100,000 people to be in a state with “low” unemployment?

Problem 16 Answers

```
# Type your code and comments inside the code chunk
```

- Delete this and put your text answer here.

Problem 17

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

Problem 17 Answers

```
# Type your code and comments inside the code chunk
```

- Delete this and put your text answer here.

Problem 18

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 (`job_mod`). Examine the first row: How are the residuals calculated here?

Problem 18 Answers

```
# Type your code and comments inside the code chunk
```

- Delete this and put your text answer here.

Hate Crimes and Household Income

Problem 19

Create a model that examines the relationship between `hate_crimes` and median household income in the state `median_house_inc`. Name the model `job_med_mod`. Run the function `summary()` on the model `job_med_mod`.

Problem 19 Answers

```
# Type your code and comments inside the code chunk
```

Problem 20

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?

Problem 20 Answers

- Delete this and put your text answer here.

Problem 21

Show the first six rows of the output created from running the `get_regression_points()` function on `job_med_mod`.

Problem 21 Answers

```
# Type your code and comments inside the code chunk
```

Problem 22

Take a look at data for Maine (row 2 of `hate_crimes`). Did the model (`job_med_mod`) **overpredict** or **underpredict** the `hate_crimes` level, compared to what was observed in the data?

Problem 22 Answers

```
# Type your code and comments inside the code chunk
```

- Delete this and put your text answer here.

EXTRA

Figure 1 reorders the levels of the variable `trump_support` using the `fct_relevel()` function from the `forcats` package then displays the data with side-by-side boxplots.

```
# Reordering trump_support
library(forcats)
hate_crimes <- hate_crimes |>
  mutate(trump_support = fct_relevel(trump_support,
            "low", "medium", "high"))
ggplot(data = hate_crimes, aes(x = trump_support, y = hate_crimes)) +
  geom_boxplot(fill = rainbow(3)) +
  labs(x = "Voter Support of Trump",
      y = "Number of Hate Crimes per 100K people",
      title = "Hate Crimes in relation to Trump Support") +
  theme_bw() +
  theme(plot.title = element_text(hjust = 0.5))
```

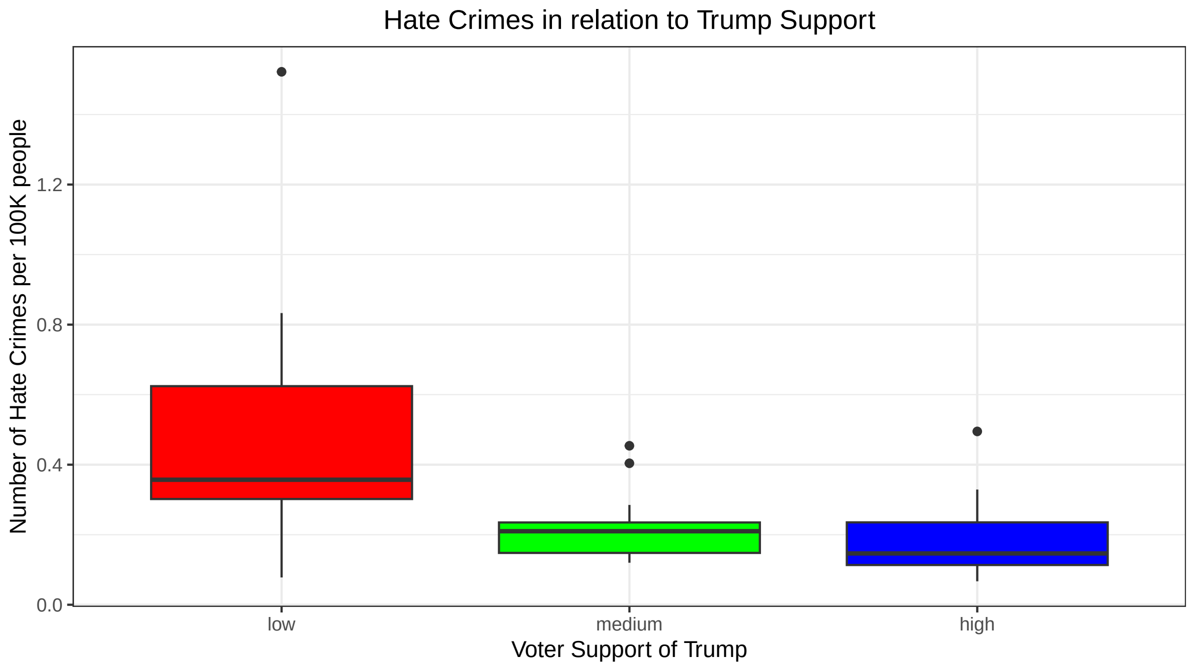

Figure 1: Hate crimes in relation to Trump support

Turning in Your Work

You will need to make sure you commit and push all of your changes to the github education repository where you obtained the lab.

Tip

- Make sure you **render a final copy with all your changes** and work.
- Look at your final html file to make sure it contains the work you expect and is formatted properly.

Logging out of the Server

There are many statistics classes and students using the Server. To keep the server running as fast as possible, it is best to sign out when you are done. To do so, follow all the same steps for closing Quarto document:

Tip

- Save all your work.
- Click on the orange button in the far right corner of the screen to quit `R`
- Choose **don't save** for the **Workspace image**
- When the browser refreshes, you can click on the sign out next to your name in the top right.
- You are signed out.

```r
sessionInfo()
```

```
R version 4.4.1 (2024-06-14)
Platform: x86_64-redhat-linux-gnu
Running under: Red Hat Enterprise Linux 9.4 (Plow)

Matrix products: default
BLAS/LAPACK: FlexiBLAS OPENBLAS-OPENMP;  LAPACK version 3.9.0

locale:
 [1] LC_CTYPE=en_US.UTF-8       LC_NUMERIC=C
 [3] LC_TIME=en_US.UTF-8        LC_COLLATE=en_US.UTF-8
 [5] LC_MONETARY=en_US.UTF-8    LC_MESSAGES=en_US.UTF-8
 [7] LC_PAPER=en_US.UTF-8       LC_NAME=C
 [9] LC_ADDRESS=C               LC_TELEPHONE=C
[11] LC_MEASUREMENT=en_US.UTF-8 LC_IDENTIFICATION=C

time zone: America/New_York
tzcode source: system (glibc)

attached base packages:
[1] stats     graphics  grDevices utils     datasets  methods   base

other attached packages:
[1] forcats_1.0.0    readr_2.1.5      moderndive_0.6.1 dplyr_1.1.4
[5] ggplot2_3.5.1    knitr_1.48

loaded via a namespace (and not attached):
 [1] utf8_1.2.4          generics_0.1.3      tidyr_1.3.1
 [4] stringi_1.8.4       hms_1.1.3           digest_0.6.36
 [7] magrittr_2.0.3      evaluate_0.24.0     grid_4.4.1
[10] timechange_0.3.0    fastmap_1.2.0       operator.tools_1.6.3
[13] jsonlite_1.8.8      backports_1.5.0     tinytex_0.52
```

```
[16] purrr_1.0.2          fansi_1.0.6          scales_1.3.0
[19] infer_1.0.7          cli_3.6.3            rlang_1.1.4
[22] crayon_1.5.3         bit64_4.0.5          munsell_0.5.1
[25] withr_3.0.1          yaml_2.3.10          tools_4.4.1
[28] parallel_4.4.1       tzdb_0.4.0           colorspace_2.1-1
[31] broom_1.0.6          vctrs_0.6.5          R6_2.5.1
[34] lifecycle_1.0.4      lubridate_1.9.3      snakecase_0.11.1
[37] stringr_1.5.1        bit_4.0.5            vroom_1.6.5
[40] janitor_2.2.0        pkgconfig_2.0.3      pillar_1.9.0
[43] gtable_0.3.5         glue_1.7.0           xfun_0.47
[46] tibble_3.2.1         tidyselect_1.2.1     rstudioapi_0.16.0
[49] farver_2.1.2         htmltools_0.5.8.1    labeling_0.4.3
[52] rmarkdown_2.28       formula.tools_1.7.1  compiler_4.4.1
```