# Problem Set 06

WRITE YOUR NAME HERE

WRITE DATE HERE

## Contents

## Background

We will again use the hate crimes data we used in Problem Set 05. The FiveThirtyEight article article about those data are in the Jan. 23, 2017 article "Higher Rates Of Hate Crimes Are Tied To Income Inequality". This week, we will use these data to run regression models with a single **categorical** predictor (explanatory) variable **and** a single **numeric** predictor (explanatory) variable.

Remember you can knit this file to see the instructions. You can type your answers again right into this file, knit the final draft with all your answers. Please submit the `pdf` file on gradescope.

## Setup

First load the necessary packages

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

Copy paste and run the following in a code chunk to read in the data:

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

Next let's explore the `hate_crimes` data set using the `glimpse()` function from the `dplyr` package:

```r
glimpse(hate_crimes)
```

You should also take a look at the data in the **data viewer**.

Each case/row in these data is a state in the US. This week we will consider the response variable `income`, which is the numeric variable of median income of households in each state.

We will use

- A categorical explanatory variable `urbanization`: level of urbanization in a region
- A numerical explanatory variable `hs`: the percentage of adults 25 and older with a high school degree

# Income, education and urbanization

We will start by modeling the relationship between:

- $y$: Median household income in 2016
- $x_1$: numerical variable percent of adults 25 and older with a high-school degree, contained in the `hs` variable
- $x_2$: categorical variable level of urbanization in a state: `low`, or `high`, as contained in the variable `urbanization`

# Question 1

Create a data visualization comparing median household `income` at "low" and "high" levels of urbanization (you do not need to include the `hs` variable in this plot). Please include axis labels and title.

**Answer:**

# Question 2.A

Next, let's add the high-school degree variable into the mix by creating a scatterplot showing:

- Median household income on the $y$ axis
- Percent of adults 25 or older with a high school degree on the $x$ axis
- The points colored by the variable urbanization
- A line of best fit (regression line) for each level of the variable urbanization (one for "low", one for "high")

For this question, add the regression lines to the plot using the geom_parallel_slopes function from the moderndive package. This function will draw the regression lines based on fitting a regression model with parallel slopes (i.e., with no interaction between hs and urbanization).

## Question 2.B

Do you think the relationship between hs and income is strong or weak? linear or non-linear?

**Answer:**

## Question 2.C

Which regression line (high urbanization or low urbanization) appears to have the larger intercept?

**Answer:**

## Question 3.A

Now let's create a second scatterplot using the same variables, but this time draw the regression lines using geom_smooth, which will allow for separate, non-parallel slopes for each urbanization group.

## Question 3.B

How do the slopes show in the plot above compare to the slopes in Question 2? Are the two slopes fairly similar (parallel) for the two levels of urbanization, or do they differ now?

**Answer:**

## Question 3.C

Based on visually comparing the two models shown in Question 2 and Question 3, do you think it would be best to run a "parallel slopes" model (i.e. a model that estimates one shared slope for the two levels of urbanization), or a more complex "interaction model" (i.e. a model that estimates a separate slope for the two levels of urbanization)?

**Answer:**

## Question 4.A

Fit the following two regression models that examine the relationship between household income (as response variable), and high-school education (hs) and urbanization as explanatory variables:

1. A parallel slopes model (i.e., no interaction between hs and urbanization)
2. A non-parallel slopes model (i.e., allow hs and urbanization to interact in your model)

Be sure to save the output from the lm function for each model.

## Question 4.B

Use the `get_regression_summaries` function to find the unadjusted proportion of variance in `income` accounted for by each model, and report the value for each model

**Answer:**

## Question 4.C

Compare the **adjusted** proportion of variance account for each model. Based on this comparison, which model do you prefer? Does your preference here agree or disagree with your earlier preference based on visualizing the predictions of each model?

**Answer:**

**For Questions 5 though 10, base your answers on the model you've selected using visual and quantitative comparisons in Question 3 and 4.**

## Question 5.A

Generate the regression table for your preferred model using the `get_regression_table()` function from the `moderndive` package. Is the intercept the same for the states with a "low" and "high" level of urbanization? Is the slope the same?

**Answer:**

## Question 5.B

What is the slope for the regression line of the states with a "high" level of urbanization? What is the intercept?

**Answer:**

## Question 5.C

What is the slope for the regression line of the states with a "low" level of urbanization? What is the intercept?

**Answer:**

## Question 5.D

Based on your regression table output (and the data visualizations), is median household income greater in states that have lower or higher levels of urbanization? By how much?

**Answer:**

## Question 5.E

For every 1 percentage point increase of high-school educated adults in a state, what is the associated increase in the median household income?

**Answer:**

## Question 6

What would you predict as the median household income for a state with a **high** level of urbanization, in which 85% of adults have a high school degree? Careful with rounding!

**Answer:**

## Question 7

What would you predict as the median household income for a state with a **low** level of urbanization, in which 85% of the adults have a high school degree?

**Answer:**

## Question 8

What would you predict as the median household income for a state with a **low** level of urbanization in which 30% of adults have a high school degree?

**Answer:**

## Question 9

What was the observed `income` value for Maine (row 2)? What was the prediction for Maine according to your model? What is the residual? Did our model over or underestimate the median income for this state?

**Answer:**

## Independent analysis

You will now use the tools you have learned, and a new data set to solve a conservation problem.

Wildlife biologists are interested in managing/protecting habitats for a declining species of vole, but are not sure about what habitats it prefers. Two things that biologists can easily control with management is percent cover of vegetation, and where habitat improvements occur (i.e. is it important to create/protect habitat in moist or dry sites, etc). To help inform habitat management of this vole species, the researchers in this study counted the number of `voles` at 56 random study sites. At each site, they measured percent cover of `veg`etation, and recorded whether a site had moist or dry `soil`.

The data can be read in like so:

```
vole_trapping <- read_csv("http://bit.ly/2IgDF0E")
```

The data contains the variables:

- `site` for the id of each random study site (each case or row is a survey/trapping site)
- `voles` for the vole count at each site
- `veg` for the percent cover of vegetation at each site
- `soil` identifying a site as “moist” or “dry”

Generate an appropriate regression model with `voles` as the response variable `y`, and `veg` and `soil` as explanatory variables. Use the results of the model to answer the following questions **based on the available data**. A data visualization will probably also help you.

## Question 10

Would protecting a site with high vegetation cover be a more effective way to preserve the vole population than a site with low vegetation cover? Why?

**Answer:**

## Question 11

Dry sites typically cost a lot less to purchase and maintain for conservation organizations. Thus, if a conservation organization decides to purchase a few dry sites, roughly what percent cover of vegetation do they need to maintain on these sites (at a minimum) to support a population of about 30 voles at the site?

**Answer:**

## Question 12

The Nature Conservancy is looking at purchasing a site for this species (in the same study area) that has moist soil and 40% vegetation cover. **Using the regression equation** what would you predict as the possible vole population the site might be able to support?

**Answer:**