# Problem Set 04: Multiple Linear Regression Solutions

Professor Bradley Warner

June, 2022

### Documentation:

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

## Introduction

We will again use the hate crimes data we used in Problem Set 03. The FiveThirtyEight article article about those data are in the Jan. 23, 2017 article "Higher Rates Of Hate Crimes Are Tied To Income Inequality". For this project, we will use these data to run regression models with a single **categorical** predictor (explanatory) variable **and** a single **numeric** predictor (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)
```

```
## 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.

A summary of the data also helps us to understand our data.

```
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_ps4 <- hate_crimes %>%
  select(state, hs, income, urbanization) %>%
  na.omit()
```

**Use `hate_crimes_ps4` for all your work in this problem set.**

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

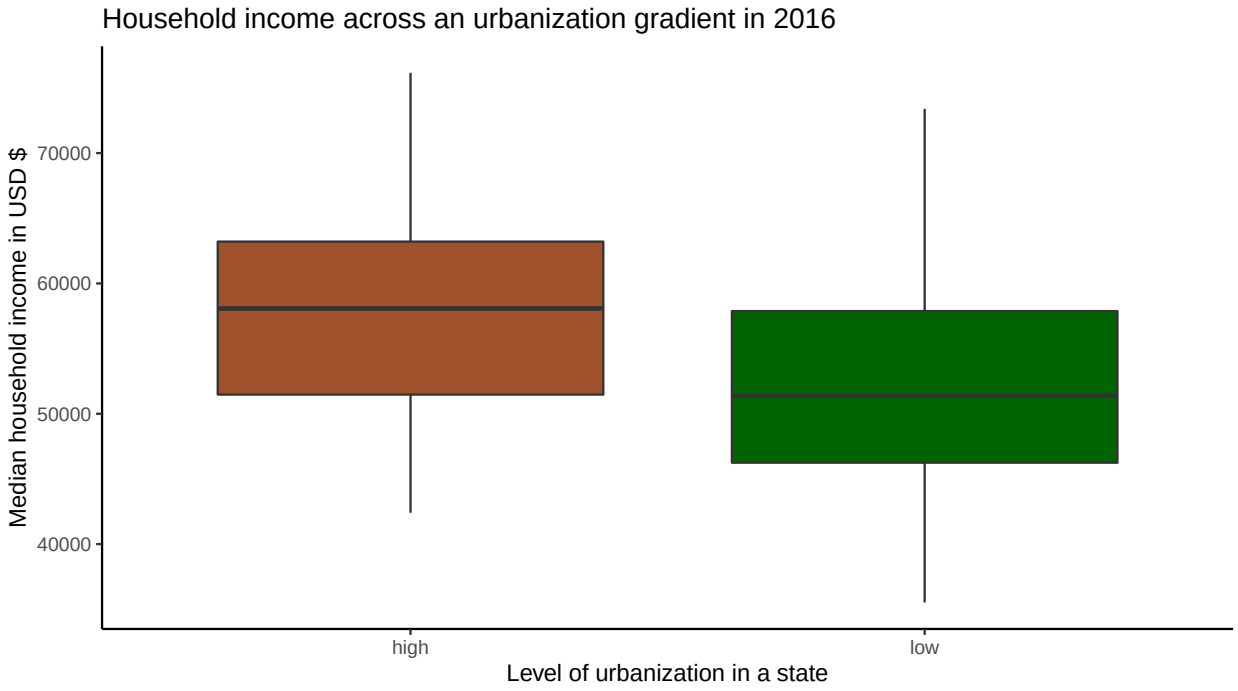
- $x_2$: categorical variable level of urbanization in a state: `low`, or `high`, as contained in the variable `urbanization`

## Exercise 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:**

```
ggplot(data = hate_crimes_ps4, aes(x = urbanization, y = income)) +
  geom_boxplot(fill = c("sienna", "darkgreen")) +
  labs(x = "Level of urbanization in a state", y = "Median household income in USD $",
       title = "Household income across an urbanization gradient in 2016") +
  theme_classic()
```




**States with a "high" level of urbanization have a higher median household income.**

## Exercise 2

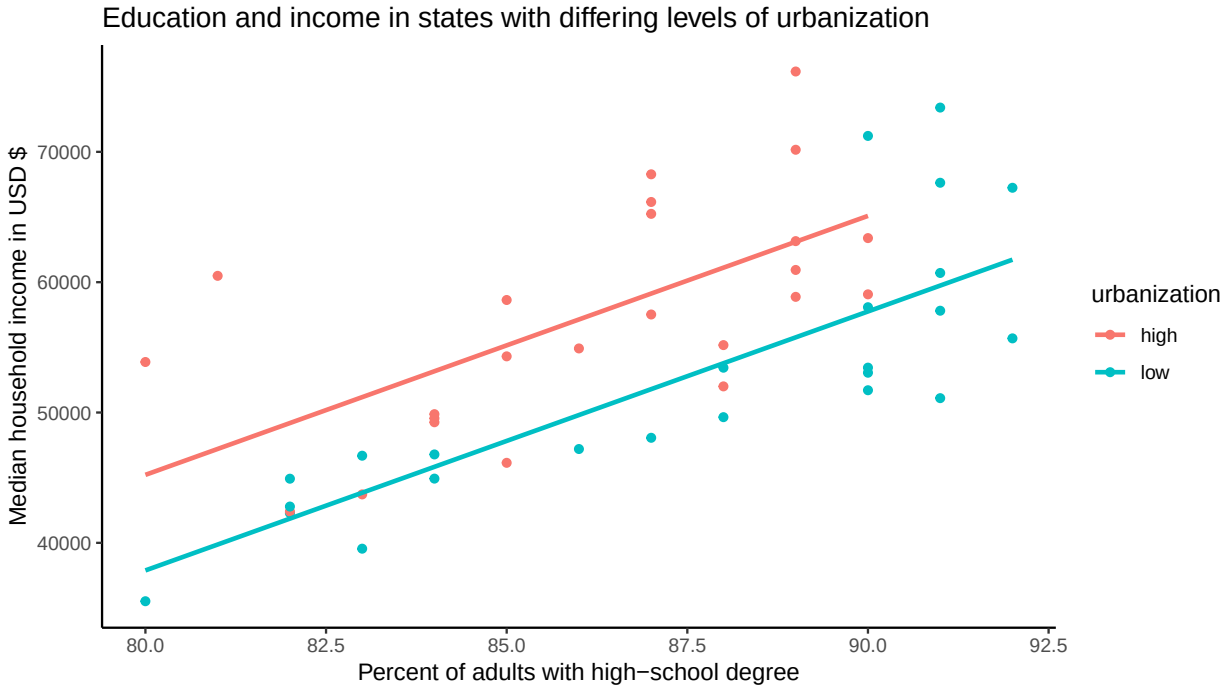
Next, let's add the high-school degree variable into the mix by creating a scatterplot showing:

- 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`).

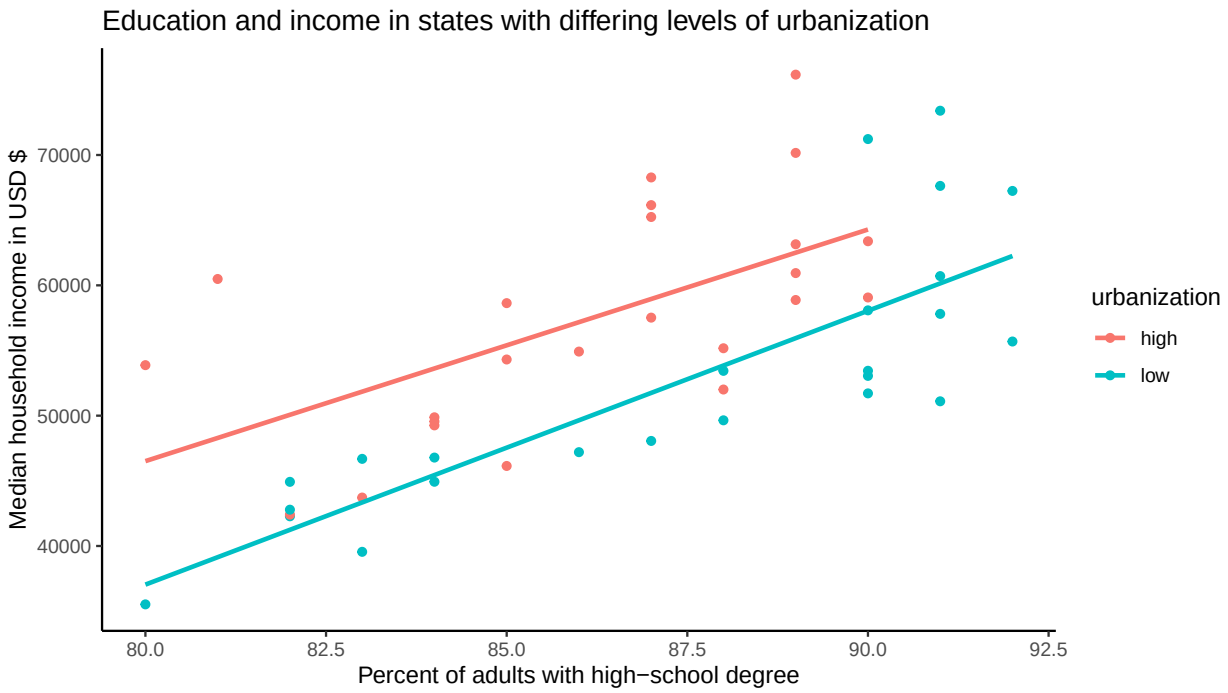
```
ggplot(data = hate_crimes_ps4, aes(y = income, x = hs, color = urbanization)) +
  geom_point()+
  geom_parallel_slopes(se = FALSE) +
  labs(x = "Percent of adults with high-school degree",
       y = "Median household income in USD $",
       title = "Education and income in states with differing levels of urbanization"
       ) +
  theme_classic()
```




## Exercise 3

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.

```
ggplot(data = hate_crimes_ps4, aes(y = income, x = hs, color = urbanization)) +
  geom_point()+
  geom_smooth(method = "lm", se = FALSE) +
  labs(
    x = "Percent of adults with high-school degree",
    y = "Median household income in USD $",
    title = "Education and income in states with differing levels of urbanization") +
  theme_classic()
```




## Exercise 4

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:**

The slopes do not appear to be much different, so we do not think an interaction model is warranted.

## Exercise 5

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.

```
income_parallel_model <- lm(income ~ hs + urbanization, data = hate_crimes_ps4)
```

```
income_interaction_model <- lm(income ~ hs * urbanization, data = hate_crimes_ps4)
```

## Exercise 6

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

```
get_regression_summaries(income_parallel_model)
```

```
## # A tibble: 1 x 9
##   r_squared adj_r_squared       mse  rmse sigma statistic p_value    df  nobs
##       <dbl>         <dbl>     <dbl> <dbl> <dbl>     <dbl>   <dbl> <dbl> <dbl>
## 1     0.572         0.553 37513189. 6125. 6326.      30.0       0     2    48
```

```
get_regression_summaries(income_interaction_model)
```

```
## # A tibble: 1 x 9
##   r_squared adj_r_squared       mse  rmse sigma statistic p_value    df  nobs
##       <dbl>         <dbl>     <dbl> <dbl> <dbl>     <dbl>   <dbl> <dbl> <dbl>
## 1     0.575         0.546 37245066. 6103. 6374.      19.8       0     3    48
```

**Answer:** For the parallel slopes model, the proportion of variance accounted for ($R^2$) is .572. For the interaction model, it is .575.

## Exercise 7

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 the parallel slopes model, the adjusted $R^2$ is .553. For the interaction model, it is .546. Since adjusted $R^2$ actually decreases for the more complex interaction model, we should prefer the simpler parallel slopes model.

**For Exercise 8 though 10, base your answers on the model you've selected in Exercise 7.**

### Note to Instructors

**Students *should* prefer the parallel slopes model, and should have said so in earlier questions. But if they don't, do not penalize them further on questions 8 through 10. For completeness, solutions for both the parallel slopes and interaction models are included for questions 8 through 10.**

## Exercise 8

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

**Answer:**

```
get_regression_table(income_parallel_model)
```

```
## # A tibble: 3 x 7
##   term              estimate std_error statistic p_value lower_ci upper_ci
##   <chr>                <dbl>     <dbl>     <dbl>   <dbl>    <dbl>    <dbl>
## 1 intercept         -113725.    23553.     -4.83       0 -161163.  -66287.
## 2 hs                   1987.      273.      7.28       0    1437.    2537.
## 3 urbanization: low   -7333.     1858.     -3.95       0  -11075.   -3592.
```

- **Parallel slopes: Slope is 1986.794, Intercept is -113725.193 (rounded is OK)**

```
get_regression_table(income_interaction_model)
```

```
## # A tibble: 4 x 7
##   term               estimate std_error statistic p_value lower_ci upper_ci
##   <chr>                 <dbl>     <dbl>     <dbl>   <dbl>    <dbl>    <dbl>
## 1 intercept           -95647.    39938.    -2.40    0.021 -176137.  -15158.
## 2 hs                    1777.      463.     3.84    0         843.    2711.
## 3 urbanization: low   -35394.    49894.    -0.709   0.482 -135948.   65160.
## 4 hs:urbanizationlow     324.      576.     0.563   0.576    -836.    1484.
```

- **Interaction Model: Slope is 1777 (rounded is OK), Intercept is -95647.4**

## Exercise 9

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:**

- **Parallel slopes: The associated increase is the slope: $1986**
- **Interaction Model: The associated increase for "low" states is the slope on `hs`: $1777. For "high" states, it is $2101**

## Exercise 10

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:**

- **Parallel slopes: -113725.193 + (-7333.326 * 0) + (1986.794 * 85) = 55152.28 (rounding OK)**
- **Interaction Model: -95647.4 + (-35394.0*0) + (1777.0 + (324 * 0)) * 85 = 55,397.6 (rounding OK) Note that answers may vary substantially based on rounding**

## Documenting software

- File creation date: 2022-06-18
- R version 4.1.3 (2022-03-10)
- `tidyverse` package version: 1.3.1
- `moderndive` package version: 0.5.4