# Practice Final Exam

### What You have Learned about Exploring Variation

Let’s start with a data set called **bad\_drivers** from [FiveThirtyEight](https://fivethirtyeight.com/features/higher-rates-of-hate-crimes-are-tied-to-income-inequality/) (you’ll need to add the code require(fivethirtyeight) to DataCamp sandbox). If you want to know more information about these variables at any time, google: **bad\_drivers r documentation**. Go ahead and use R to look at the first six lines of this data frame. Write code here:

What are the cases in this data frame?

Take a look at the variation in **insurance\_premiums** (the average car insurance premiums in dollars). Write R code here:

Which state had the highest **num\_drivers**?

Does that mean that state had the most fatal collisions? (Hint: Take a look at the documentation for this data.)

What does the **num\_drivers** number tell you?

As statisticians (you are a statistician now!), we want to do is **explain** the variation we see here. Let’s take the question of insurance premiums -- why is it more expensive to have car insurance in some states and cheaper in others? Perhaps **num\_drivers** explains some of the variation we see in insurance premiums. Write this idea as a word equation.

In your word equation (above), label which is the outcome and which is the explanatory variable.

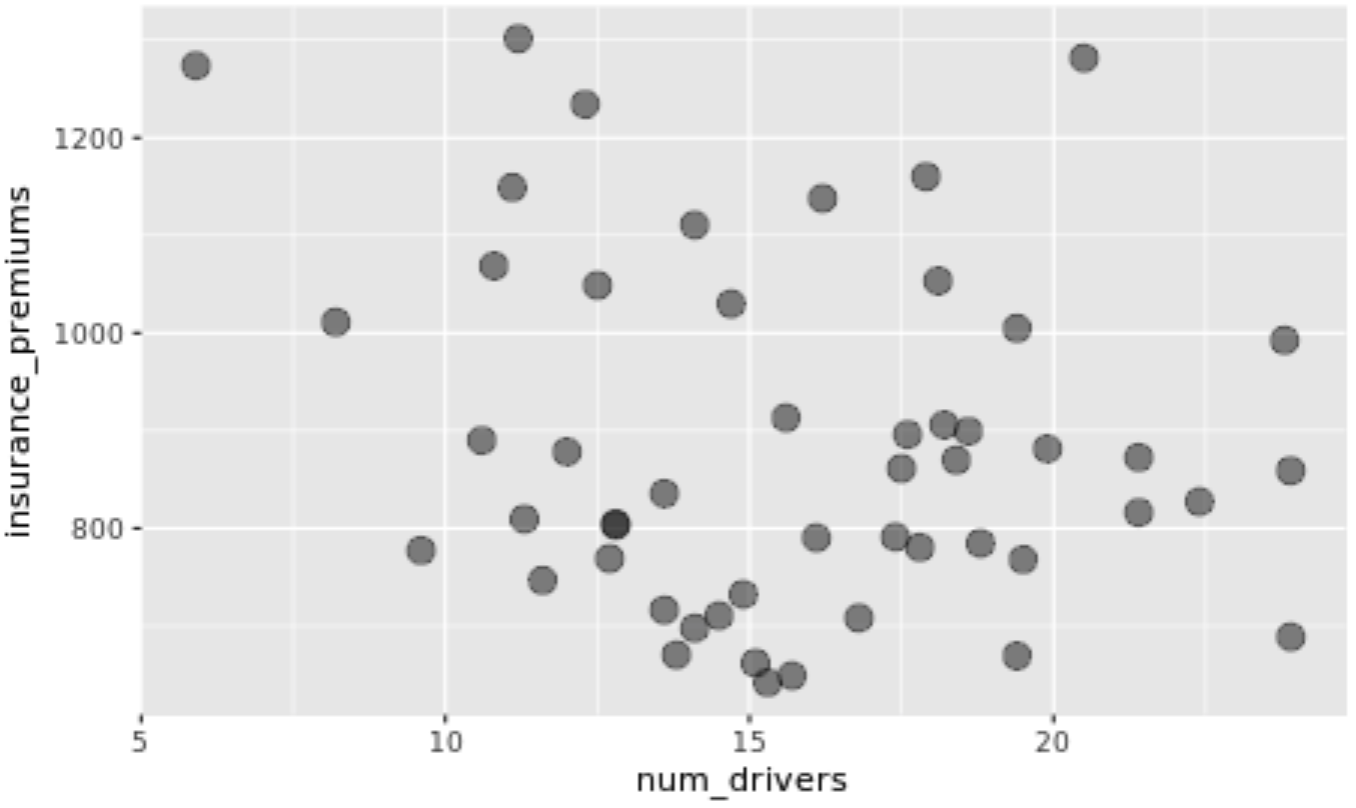
Let’s explore the relationships between the outcome and explanatory variable in the data. Make a visualization that would help you explore this relationship and write the R code here:

In this situation, what would it mean to “explain variation”?

Just from eyeballing your visualization, do you get the sense that your explanatory variable “explains” variation in the outcome variable well? Why or why not?

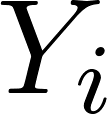
### What You have Learned about Modeling Variation

Here is a scatterplot that depicts **insurance\_premiums** by **num\_drivers**. (1) Draw in the empty model in blue. (2) Draw in your complex model in another color. (3) Draw a negative and positive residual from the empty model in blue. (4) Draw a negative and positive residual from the complex model in another color.

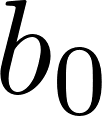


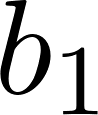
Write the R code below to find the best fitting model.

Write the best fitting model in General Linear Model (GLM) format.

What does your [](about:blank) mean?

What does your [](about:blank) mean?

What does your [](about:blank) mean?

What does your [](about:blank) mean?

What percentage of the leftover variation from the empty model is explained by the model that includes **num\_drivers**?

Dr. Ji surmises that the percentage of fluke accidents might explain a lot of the variation in **insurance\_premiums**. So she creates a model using **perc\_no\_previous** (percentage of fatal collisions involving drivers that had never been involved in accidents before)and reports that the PRE for that model is .0057. She argues that is smaller than .05 so the empty model is unlikely. You suspect that she is wrong. What is odd about her thinking?

|  |
| --- |
| BONUS SECTION! To check out Dr. Ji’s story, you decide to explore the data to see if her story might be worth digging into. What pattern do you see in the data?  Should she create a regression model to explain this pattern? Why or why not?  What should she do instead? |

### What You have Learned about Evaluating Models

The **num\_drivers** model tells us about the relationship between **insurance\_premiums** and **num\_drivers** *in the sample data.* But we want to know if this relationship exists in the Data Generating Process.

The sample F for this model is 2.03. Is it possible that we could have gotten this F value if there was no relationship between insurance premiums and number of drivers involved in fatal collisions per billion miles in the DGP?

Let’s use randomization (**shuffle**) to create a data generating process where there is no relationship between **insurance\_premiums** and **num\_drivers**. To start, let’s create a sampling distribution with 10,000 randomized Fs. Write that code below. Also save the sample F as **sampleF**.

Draw a picture of the distribution triad here. Include a drawing of the resulting sampling distribution of Fs and color in the p-value in a different color.

What does your p-value mean?

Write up the results of your analysis here in APA style.

Which of the following statements is supported by the data analysis you conducted (check all that apply):

[ ] If the p-value was less than .05 we could have concluded that more drivers involved in fatal collisions causes insurance premiums to go down.

[ ] More drivers involved in fatal collisions causes insurance premiums to go down.

[ ] Bad drivers tend to have lower insurance premiums.

[ ] States with more drivers involved in fatal collisions tend to have lower insurance premiums.

[ ] States with more drivers involved in fatal collisions per billion miles tend to have lower insurance premiums.

[ ] We do not have a good reason to think that there is a relationship between a state’s number of drivers involved in fatal collisions per billion miles and insurance premiums.

[ ] The complex model shown in the sampling distribution of the data generating process of the sample data of the drivers involved in fatal collisions in this analysis with costs going down in the graph.

[ ] Fatal collisions are randomly caused in each state.

[ ] Knowing a state’s number of fatal collisions per billion miles does not help us make a better prediction about their insurance premiums. Predicting the mean insurance premium would be a perfectly fine guess.