Econ/Poli 5 Homework 5

Homework 5 is due Tuesday June 1st at 11:59 PM.

Setup

In this homework we will revisit an application we have already studied a few times in this class: What is the relationship between opioid usage and unemployment rates across U.S. counties.

The structure of the analysis in this homework will likely be helpful as you decide how to structure your final project. In particular we will first clearly explain the hypothesis, then provide some descriptive evidence, then estimate the relationship we are interested in and finally consider alternative hypotheses for whatever relationship we find.

As you will see the data for hw5 is named “hw5_data.dta”. This is a Stata dataset, but we will be using R for this dataset. The package “rio” allows you to import Stata datasets into R. You will need to (1) install rio, (2) load rio using the library() function and then (3) load in the data. The code to accomplish these steps has been added to the hw5_blank.R script.

Question 1

Clearly re-state the question we are studying. Explain why this question is important to explore and why we might expect a relationship between the two variables we are interested in. If you get stuck go back to the slides for stata3 and stata4 lectures to remind youreself about this application. **(4 points)**

Are poorer labor-market conditions related to higher opioid usage?

This is an important question to explore because of the dramatic increase in opioid usage in the United States starting in the 1990s.

We would expect a relationship because higher opioid usage may lead to people losing jobs, which leads to higher unemployment rates. Alternatively, these higher unemployment rates may in lead to increased substance abuse, which leads to higher opioid usage.

Question 2

Before moving onto analysis, it is important to understand the variables that we are analyzing.

(2a) Our first variable of interest in **urate**. This is the same variable that is described in the stata3 lecture. **Define this variable below** (in other words, write a description of what this variable represents): **(3 points)**

urate = Unemployment rate (percent of total labor force that is unemployed).

This variable is collected at the county level by the Bureau of Labor Statistics; this only accounts for the number of people actively searching for a job.

(2b) Our second variable of interest is **prescrip_rate**. This is the same variable that is described in the stata3 lecture. **Define this variable below** (in other words, write a description of what this variable represents) **(3 points)**

prescrip_rate = Prescription rate (the number of prescriptions per 100 people).

This variable is reported every year at the county level by the CDC.

Question 3

If you try to take the mean of the unemployment rate you will find that R reports NA. This is because a few of the values for the unemployment rate are missing. We are going to drop observations with missing values for this variable. The function (which is part of the dplyr package) is called "drop_na()". Search the help files and look at the examples to figure out how to drop observations that are missing**. Paste the code below** that drops counties with missing unemployment rates. **(5 points)**

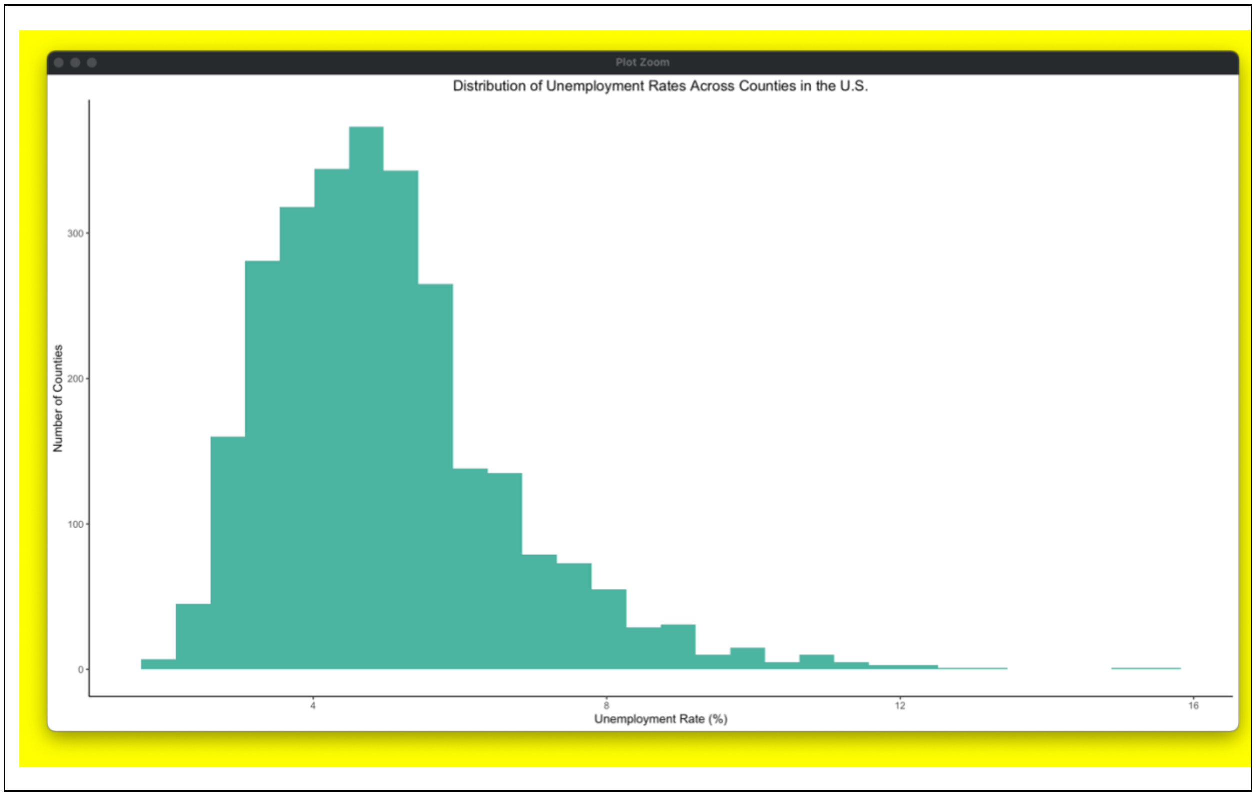
```
df <- df %>% drop_na(urate)
```

If you find a numeric variable for your final project is reporting that the mean is NA, it could be because some values are missing!

Question 4

Now that we understand our variables, it will be important to provide some descriptive visualization of our variables. To do this we will create histograms, which show the distribution of our variables across U.S. counties. Our variables are continuous, meaning they take on many values. If they were instead categorical, it might be appropriate to show a bar chart instead that shows counts within each category of the variable.

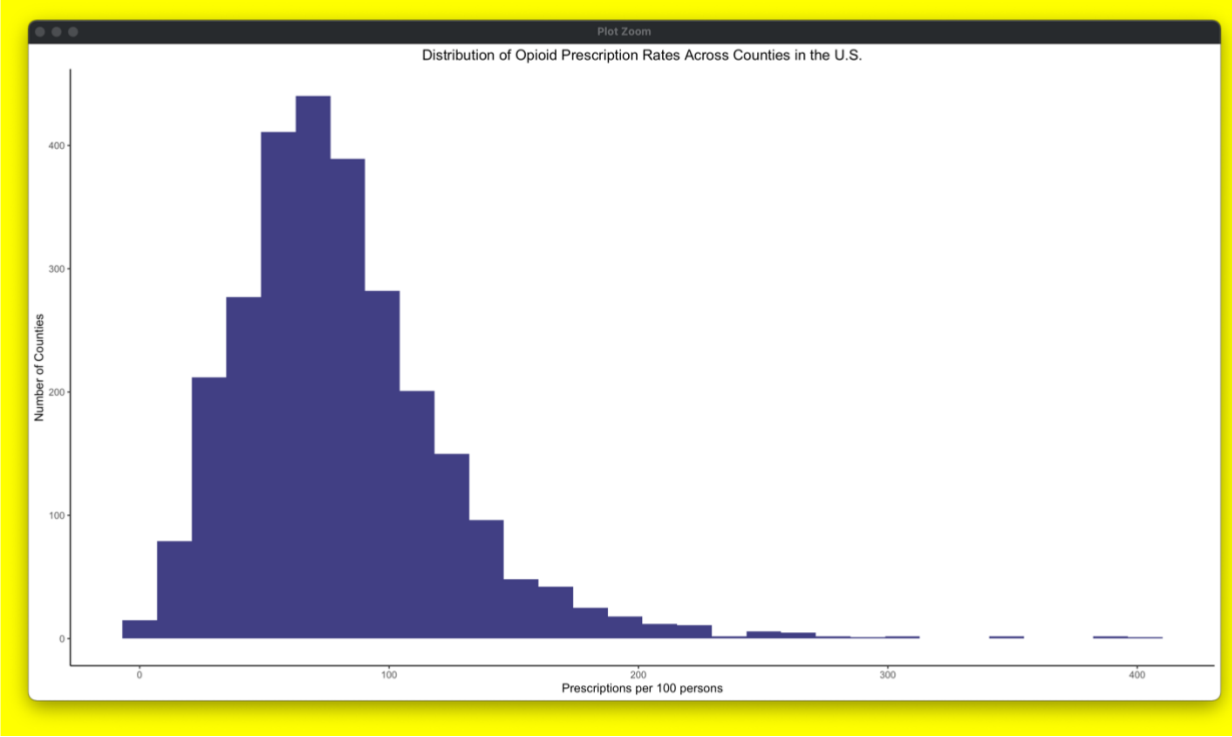
(4a) Create a properly labelled histogram of the **unemployment rate**. **Paste the histogram below. (5 points)**



(4b) Write a few sentences that describe the histogram. In other words, based on the histogram, what can you say about the distribution of unemployment rates across counties? Are there outliers? **(5 points)**

This histogram appears to be right skewed which suggests that a majority of the data is on the left side—the lower end. Based on the graph, we can see that the majority of counties had unemployment rates between 3% - 8%. There are some outliers in this data, in particular the counties whose unemployment rates are 12%+. The most common unemployment rate for counties appears to be around 5%.

(4c) Create a properly labelled histogram of the **prescription rate**. **Paste the histogram below. (5 points)**



(4d) Write a few sentences that describe the histogram. **(5 points)**

This histogram also appears to be skewed right, which suggests that a majority of the data falls on the left side—the lower end; this looks very similar to the unemployment rate histogram. As we can see from the histogram, a majority of the counties has prescription rates of 25-150 prescriptions per 100 persons. There are outliers, specifically the few counties whose prescription rates were 300+ per 100 persons. The most common prescription rate was approximately 80 prescriptions per 100 persons.

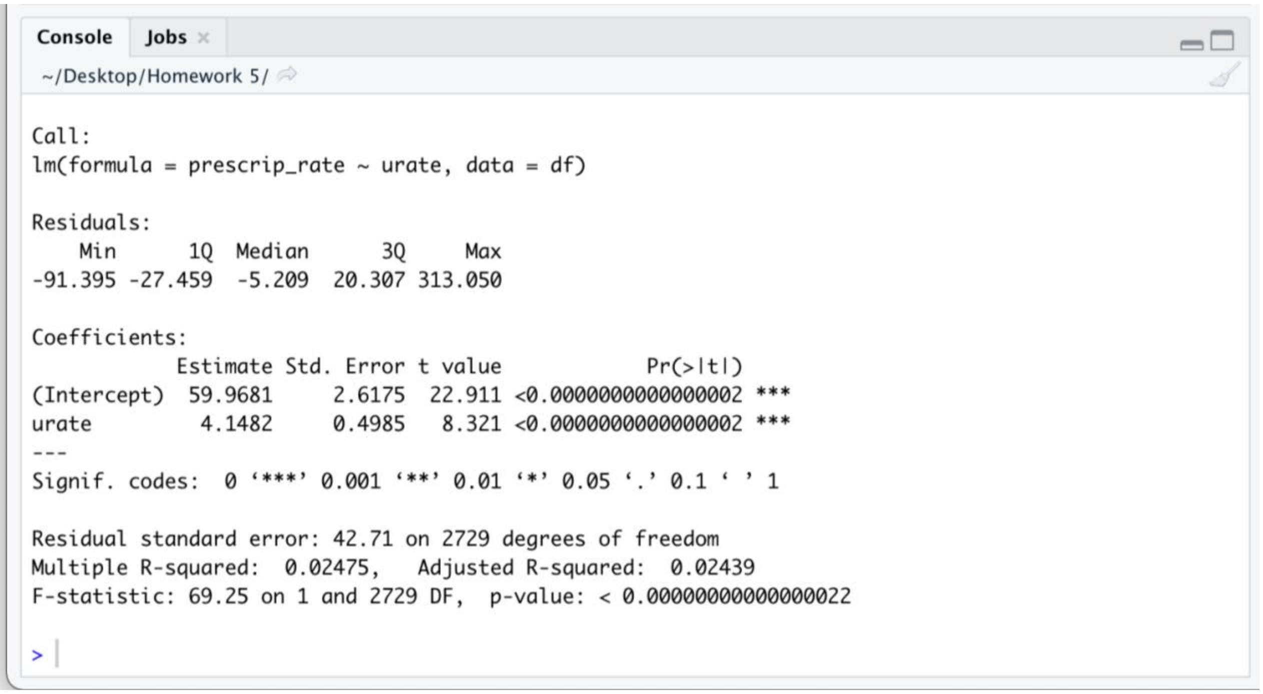
Question 5

Next, we will estimate the relationship between prescription rates and unemployment. To do so we will estimate a linear regression of the following form

$$Y_i = \beta_0 + \beta_1 X_i + \ \varepsilon_i$$

Where our **Y variable will be prescription rates** and **our X variable will be unemployment rates**.

(5a) Estimate this linear regression using the lm() function and save the results as an object called “lmfit”. Then type “summary(lmfit)” and **paste the resulting table** below. **(5 points)**



```
Call:
lm(formula = prescrip_rate ~ urate, data = df)

Residuals:
    Min      1Q  Median      3Q     Max 
-91.395 -27.459  -5.209  20.307 313.050 

Coefficients:
            Estimate Std. Error t value            Pr(>|t|)    
(Intercept)  59.9681     2.6175  22.911 <0.0000000000000002 ***
urate         4.1482     0.4985   8.321 <0.0000000000000002 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 42.71 on 2729 degrees of freedom
Multiple R-squared:  0.02475,	Adjusted R-squared:  0.02439 
F-statistic: 69.25 on 1 and 2729 DF,  p-value: < 0.00000000000000022

> |
```

(5b) Interpret the magnitude of the slope coefficient. In other words, a 1-unit increase in the unemployment rate is expected to increase (or decrease) prescriptions rates by what amount? **(5 points)**

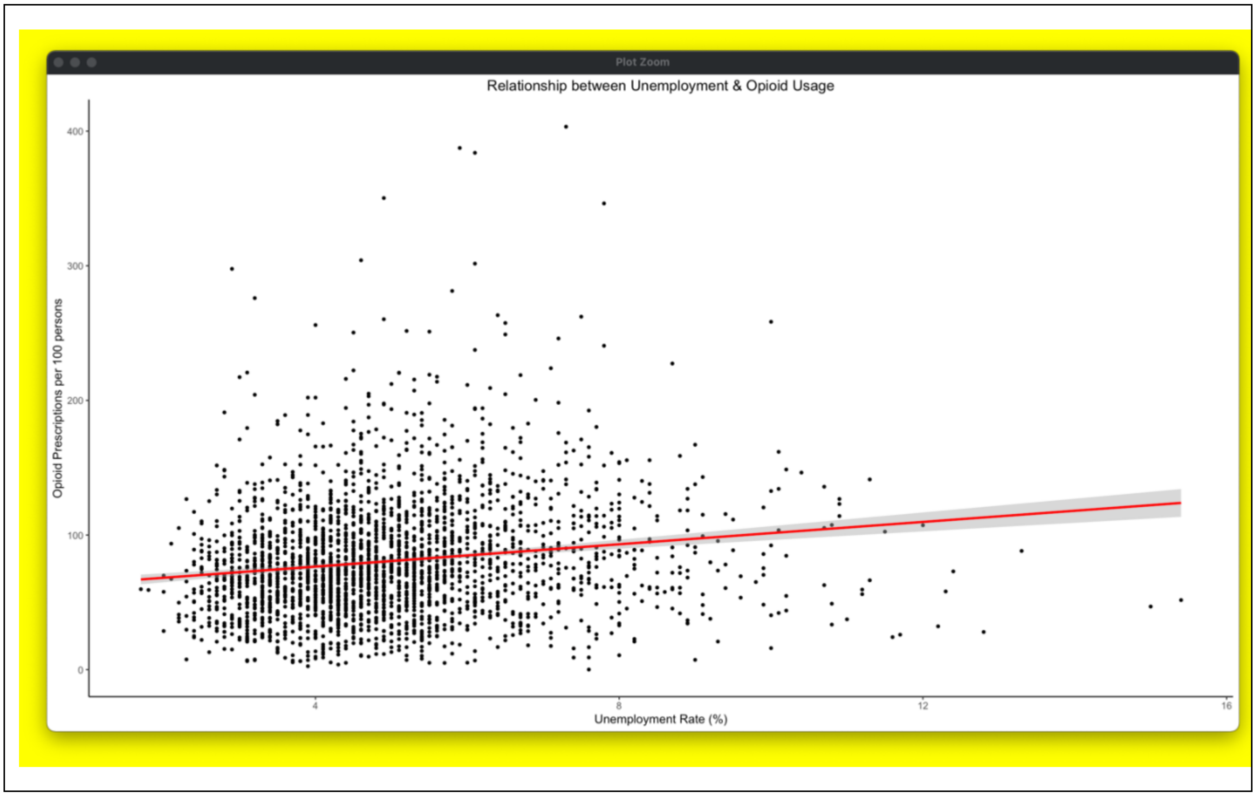
As the Unemployment Rate goes up by 1, we expect Opioid prescriptions per 100 persons to go up by 4.1482 prescriptions.

(5c) Discuss the statistical significance of the result. **(5 points)**

As we can see from our P-value of "<0.0000000000000002 ***", and the significance codes, this is statistically significant at the 0 % level.

Question 6

Next, add an appropriately labeled scatter plot with prescription rates on the Y-axis, unemployment rates on the X-axis, as well as a line of best-fit. **Paste the scatter plot below: (10 points)**



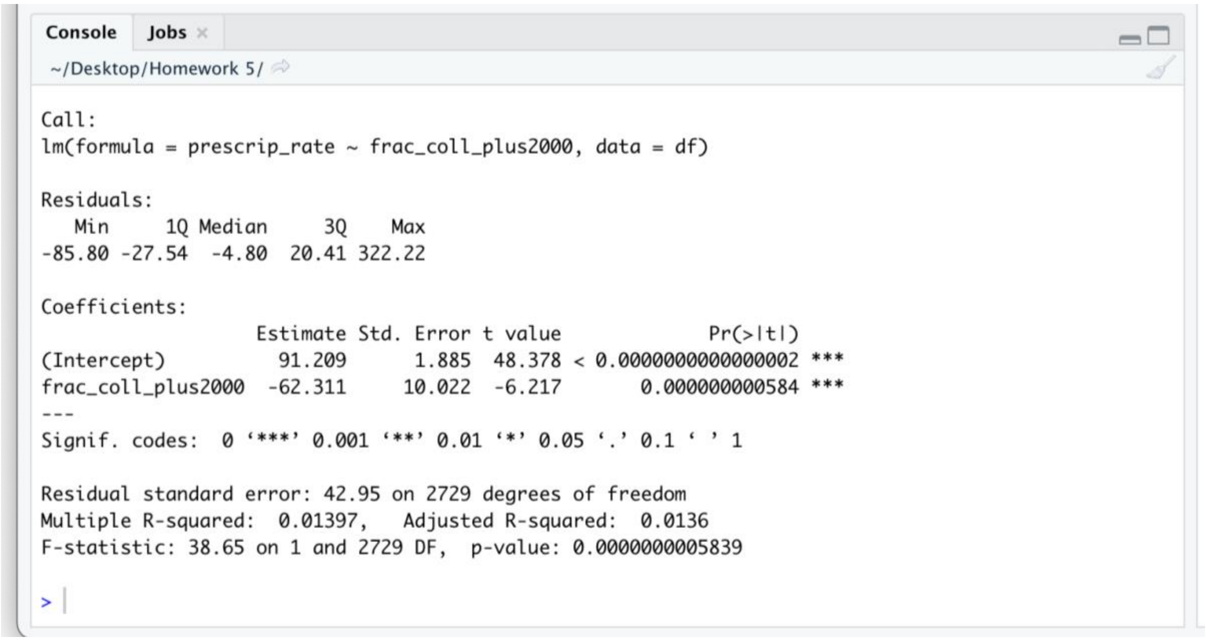
Question 7

Next we think “education” might be a confounding variable. In other words, we think it is possible that education both reduces the probability an individual is employed **and** decreases the probability an individual use opioids. Therefore, if we find a statistical relationship between opioids and unemployment, maybe it is being driven by educational differences across counties. In order to investigate this possibility, we are now going to estimate another regression model of the following form:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

Our Y-variable will remain prescrip_rate, but our new X variable will be frac_college_plus2000 which is equal to the fraction of residents in the county that have a college degree.

(7a) Estimate this linear regression using the lm() function and save the results as an object called "lmfit". Then type "summary(lmfit)" and **paste the resulting table** below.**(4 points)**



```
Call:
lm(formula = prescrip_rate ~ frac_coll_plus2000, data = df)

Residuals:
   Min     1Q Median     3Q    Max 
-85.80 -27.54  -4.80  20.41 322.22 

Coefficients:
                   Estimate Std. Error t value             Pr(>|t|)    
(Intercept)          91.209      1.885  48.378 < 0.0000000000000002 ***
frac_coll_plus2000  -62.311     10.022  -6.217       0.000000000584 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 42.95 on 2729 degrees of freedom
Multiple R-squared:  0.01397,	Adjusted R-squared:  0.0136 
F-statistic: 38.65 on 1 and 2729 DF,  p-value: 0.0000000005839

> 
```

(7b) Interpret the magnitude of the slope coefficient. **(4 points)**

As the Fraction of Residents w/ a College Degree or More goes up by 1, we expect the Opioid prescriptions per 100 persons to go down by 62.311 prescriptions.

(7c) Discuss the statistical significance of the slope coefficient. **(4 points)**

As we can see from our P-value of "0.000000000584 ***", and the significance codes, this is statistically significant at the 0 % level.

(7d) Given these results, do you think it is possible that education is a confounding variable? Why or why not? **(5 points)**

Yes, I believe that it is possible that education is a confounding variable. The fact that this is statistically significant at the 0% level, as well as the large slope coefficient, corroborates this claim. To verify this claim, we should check the effect of unemployment rate on opioid prescription rates, while controlling for education.

(7e) Try to think of another possible confounding variable. **You do not need to do anything with the data for this question**. Explain your reasoning behind choosing this particular variable. **This does not need to be a variable that is actually in the dataset**. **(5 points)**

This dataset has a variable called “**poor_share2000**” which measures the Share of the population Below Poverty Line. Given that numerous studies have shown a relationship between poverty and substance abuses, this might be a confounding variable in our analysis. The logic for this variable is similar to the one for unemployment. People living below the poverty line might face more economic insecurity & financial troubles, which could lead them to turn towards substance abuse. This in turn could lead to poorer job performance, which results in job loss, and eventually higher unemployment rates.

Question 8

If you are stuck with “what to do next” in your final project, one interesting thing to do is to explore heterogeneity. In other words, is the relationship between your variables of interest consistent or does it vary by some other characteristic in the data. For example, in stata6.pdf we explored whether the relationship between box office reviews and movie revenue is different for different movie genres.

One interesting thing to do in this dataset would be to explore different relationships by the size of the county (for example, is the relationship stronger or weaker in more rural areas?) To proxy for the size of the county we will use the variable “labor_force” which reports the number of individuals that are in the labor force in the county.

To do this, we will need to take a few steps.

First we will **coerce** the variable “labor_force” into a numeric variable. Because the code is somewhat complicated, I have provided it below and in the blank R script. Use this for a reference if you find yourself with a variable in the final project that you would like to turn numeric, but there are commas in the variable (in our case, labor force is stored as “1,000” for example instead of “1000” with the key difference being the comma). Copy and paste this code into R and run in order to convert the character variable “labor_force” into a numeric variable. You need to make sure that your dataframe is named “df” for this to run properly.

```
df <- df %>% mutate(labor_force =as.numeric(str_remove_all(labor_force,",")))
```

(8a) Use filter() to create a dataframe that restricts to counties in which the labor force is greater than 100,000 individuals. **Paste the code that creates this dataframe below: (4 points)**

```
df1 <- df %>% filter(labor_force > 100000)
```

(8b) Use filter() to create a dataframe that restricts to counties in which the labor force is less than or equal to 100,000 individuals. Make sure you give this dataframe a new name that is not the same as the dataframe you created in (6b). **Paste the code that creates this dataframe below: (4 points)**

```
df2 <- df %>% filter(labor_force <= 100000)
```

Question 9

Estimate a linear regression with prescription rate as the Y-variable and unemployment rate as the X-variable in each subset of the data created in (8b) and (8c).

(9a) Compare the results across the different subsets of the data. Make sure to comment on (1) the difference in the magnitude of the slope coefficient across subsets of the data and (2) the difference in statistical significance across subsets of the data. What have we learned by splitting by the size of the labor force? **(10 points)**

Labor Force is Greater than 100k individuals:
Slope Coefficient = 1.340; p-value = 0.273
As the Unemployment Rate goes up by 1, we expect Opioid prescriptions per 100 persons to go up by 1.340 prescriptions.
Given the p-value of 0.273, this is not statistically significant at any level.

Labor Force is Less than or Equal to 100k individuals:
Slope Coefficient = 4.1683; p-value = 1.03e-14 ***
As the Unemployment Rate goes up by 1, we expect Opioid prescriptions per 100 persons to go up by 4.1683 prescriptions.
Given the p-value of 1.03e-14 ***, and the significance code, this is statistically significant at the 0% level.

As we can see, not only does unemployment rate have a greater effect (3 more prescriptions) on counties with a labor force less than or equal to 100,000 individuals, but this effect is also statistically significant at all levels. Alternatively, in counties with a labor force greater than 100,000 individuals this effect is not statistically significant at any level. We can conclude that opioid usage is indeed related to unemployment in rural counties but not in urban counties.