Midterm Exam Prep

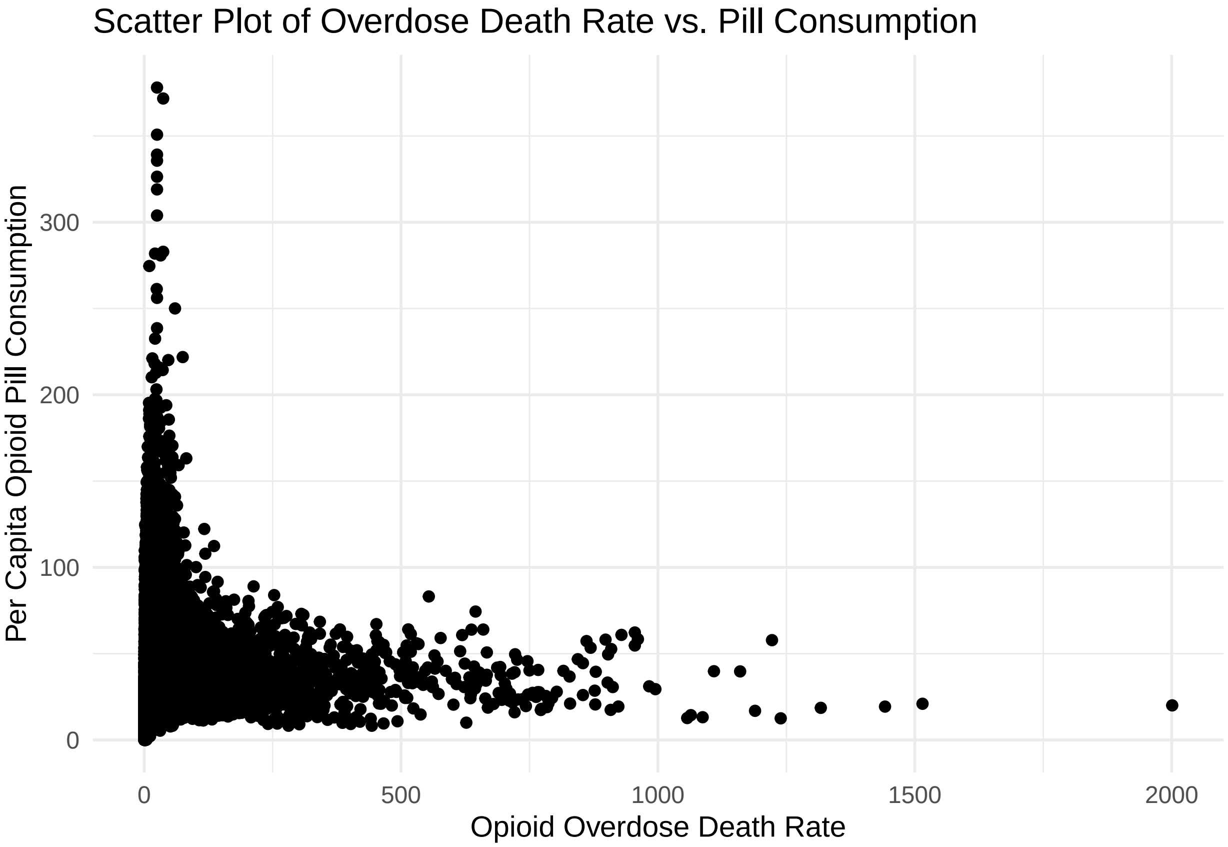
Dr. Soliman

1. What is the unit of observation for the countyopioids dataset you just read in?

The unit of observation in the dataset is a county. More specifically, each row represents a specific county in a given year.

2. Generate a scatter plot with the opioid overdose death rate (overdosedeaths) on the x-axis and per capita opioid pill consumption (percapitapills) on the y-axis.

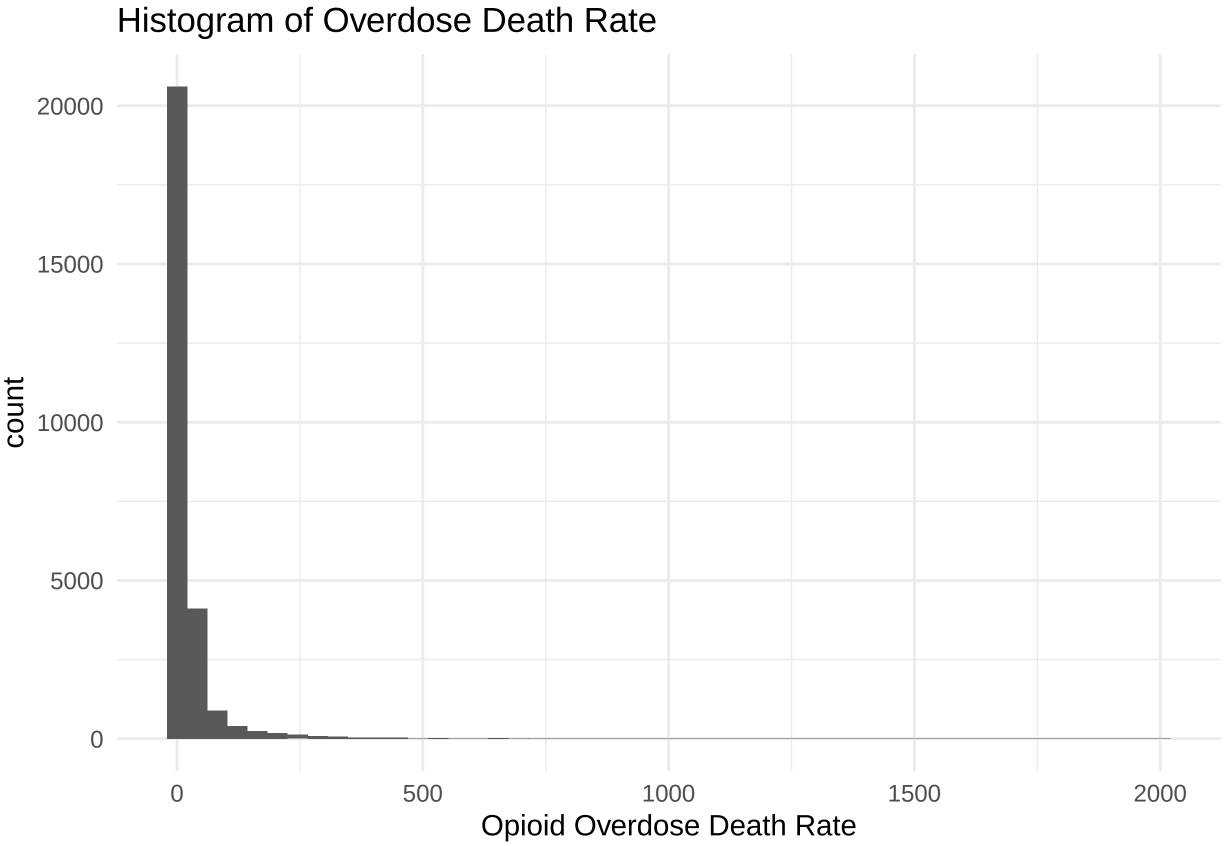
```
ggplot(countyopioids, aes(x = overdosedeaths, y = percapitapills)) +
  geom_point() +
  labs(x = "Opioid Overdose Death Rate", y = "Per Capita Opioid Pill Consumption",
       title = "Scatter Plot of Overdose Death Rate vs. Pill Consumption") +
  theme_minimal()
```



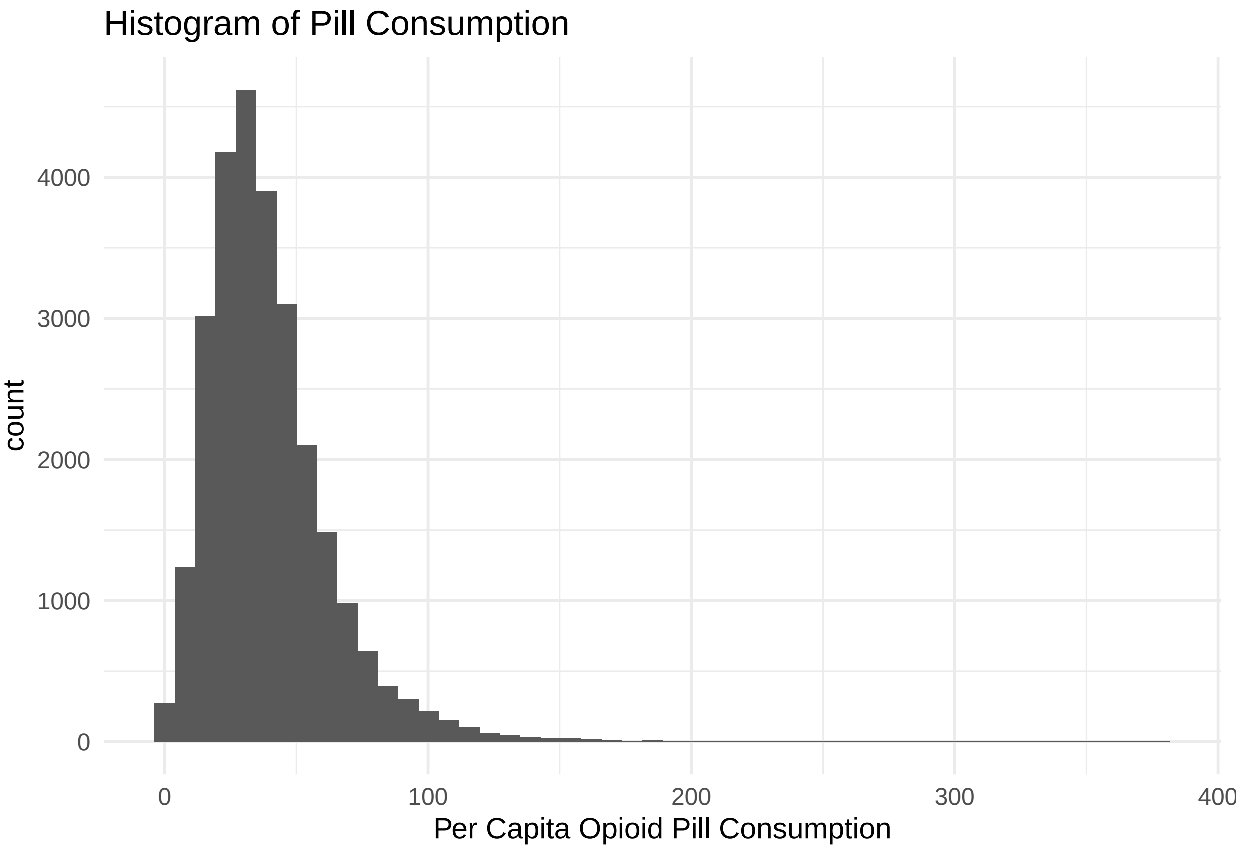
3. Generate histograms separately for the opioid overdose death rate (overdosedeaths) and per capita opioid pills (percapitapills). If you had to take the logarithm of one of these variables based on the distributions you just generated, which one would you transform and why?

If I had to take the logarithm of one of these variables based on the distributions/ histograms, I would transform the opioid overdose death rate because the distribution is more right skewed. Taking the log will render it more *normally distributed.*

```
# Opioid overdose death rate histogram
ggplot(countyopioids, aes(x = overdosedeaths)) +
  geom_histogram(bins = 50) +
  labs(x = "Opioid Overdose Death Rate", title = "Histogram of Overdose Death Rate") +
  theme_minimal()
```



```
# Per capita opioid pills histogram
ggplot(countyopioids, aes(x = percapitapills)) +
  geom_histogram(bins = 50) +
  labs(x = "Per Capita Opioid Pill Consumption", title = "Histogram of Pill Consumption") +
  theme_minimal()
```



4. Create two new variables by taking the logarithm of overdosedeaths and percapitapills (for the name, add log_ to the beginning), remove Alaska and Hawaii from the original dataset (i.e., filter them out), and assign/create this to a new object called contigiousopioiddata. You will use this for the remainder of the assignment.

```
contigiousopioiddata <- countyopioids %>%
  mutate(log_overdosedeaths = log(overdosedeaths),
         log_percapitapills = log(percapitapills)) %>%
  filter(state != "AK" & state != "HI")
```

5. Regress the opioid overdose death rate (level) on the per capita pill consumption (level). Interpret each coefficient. Remember that it is regressing the dependent variable on the independent variable (i.e., y on x). Does that seem reasonable?

On average, a county has a yearly overdose death rate of 24.50 when there are no opioid pills. The coefficient on percapitapills suggests that, on average, an increase of 1 opioid pill per capita is associated with a 0.05 increase in the opioid overdose death rate. It seems reasonable that the coefficient is positive, as we would expect more pills to be associated with more overdose deaths.

```
# one option
lm(overdosedeaths ~ percapitapills, data = contigiousopioiddata)
```

```
##
## Call:
## lm(formula = overdosedeaths ~ percapitapills, data = contigiousopioiddata)
##
## Coefficients:
##    (Intercept)  percapitapills
##       24.50483         0.04852
```

```r
# alternative option
model_level <- lm(overdosedeaths ~ percapitapills, data = contigiousopioiddata)
summary(model_level)
```

```
##
## Call:
## lm(formula = overdosedeaths ~ percapitapills, data = contigiousopioiddata)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
##  -28.77  -22.21  -17.60   -7.62 1975.52
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)    24.50483    0.83976  29.181  < 2e-16 ***
## percapitapills  0.04852    0.01808   2.684  0.00728 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 71.78 on 26781 degrees of freedom
## Multiple R-squared:  0.0002689,  Adjusted R-squared:  0.0002316
## F-statistic: 7.204 on 1 and 26781 DF,  p-value: 0.00728
```

6. Regress the opioid overdose death rate on the per capita pill consumption, but this time, do it three separate ways: log-level, level-log, and log-log. Interpret each coefficient below the code chunk for each way.

a. log-level

A one unit increase in per capita pill consumption is associated, on average, with a 1.39 *percent change in opioid overdose death rate. Because the coefficient is small, we can approximate it directly from the coefficient in the regression, i.e., a 1.4% increase in the overdose death rate.*

```r
lm(log(overdosedeaths) ~ percapitapills, data = contigiousopioiddata)
```

```
##
## Call:
## lm(formula = log(overdosedeaths) ~ percapitapills, data = contigiousopioiddata)
##
## Coefficients:
##    (Intercept)  percapitapills
##        1.68438         0.01383
```

b. level-log

A one percent increase in per capita pill consumption is associated, on average, with a 0.05 *unit change in opioid overdose death rate. As before, we can approximate this directly from the coefficient, i.e., 4.852/100 = 0.04852 increase in the overdose death rate.*

```r
lm(overdosedeaths ~ log(percapitapills), data = contigiousopioiddata)
```

```
##
## Call:
## lm(formula = overdosedeaths ~ log(percapitapills), data = contigiousopioiddata)
##
## Coefficients:
##         (Intercept)  log(percapitapills)
##               9.512                4.852
```

c. log-log

A one percent increase in per capita pill consumption is associated, on average, with a 0.59 *percent change in opioid overdose death rate.*

```
lm(log(overdosedeaths) ~ log(percapitapills), data = contigiousopioiddata)
```

```
##
## Call:
## lm(formula = log(overdosedeaths) ~ log(percapitapills), data = contigiousopioiddata)
##
## Coefficients:
##         (Intercept)  log(percapitapills)
##              0.1876               0.5866
```

7. Regress the opioid overdose death rate (log) on the per capita pill consumption (log), and add the variable rural. Interpret each coefficient. Use the summary function to obtain the standard errors, test statistics, and p-values. What can you infer about the statistical significance of each coefficient?

The intercept represents the expected log of opioid overdose deaths when all predictors (i.e., log(percapitapills) and rural) are zero, i.e., in urban areas with no pills. This is not super meaningful here, as similar to a class with no students, there are likely no counties with zero pills.

The coefficient for log(percapitapills) tells us a one percent increase in per capita pill consumption is associated, on average, with a 0.43 *percent change in opioid overdose death rate, holding rural constant.*

The coefficient for rural tells us compared with an urban area, the opioid overdose death rate in rural area is on average 72.98 *percent lower, holding pill consumption constant.*

All coefficients are statistically significant at traditional levels of significance (i.e., $\alpha = 0.05$ or 0.1*).*

```
model_log_log_rural <- lm(log(overdosedeaths) ~ log(percapitapills) + rural, data = contigiousopioiddata
summary(model_log_log_rural)
```

```
##
## Call:
## lm(formula = log(overdosedeaths) ~ log(percapitapills) + rural,
##     data = contigiousopioiddata)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.2331 -0.7575 -0.0964  0.5603  5.3427
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)          0.973601   0.035201   27.66   <2e-16 ***
## log(percapitapills)  0.428713   0.009639   44.48   <2e-16 ***
## rural               -1.308674   0.018446  -70.94   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.129 on 26780 degrees of freedom
## Multiple R-squared:  0.2504, Adjusted R-squared:  0.2503
## F-statistic:  4473 on 2 and 26780 DF,  p-value: < 2.2e-16
```

8. Generate a 95% confidence interval for the log(percapitapills) coefficient using information from the previous regression output in question 7. What does this confidence interval mean?

If we repeated our sampling procedure a large number of times, we expect about 95% of the resulting confidence interval [0.4054404,0.4418982] to capture the true value of the population parameter. We got this by using the formula [$\hat{\beta}$ - 1.96Xse($\hat{\beta}$), $\hat{\beta}$ + 1.96Xse($\hat{\beta}$)] = [0.42367 - 1.96X0.00930, 0.42367 + 1.96X0.00930].

```
# Alternative option in R is using the following command
confint(model_log_log_rural, level = 0.95)
```

```
##                          2.5 %     97.5 %
## (Intercept)          0.9046051  1.0425962
## log(percapitapills)  0.4098197  0.4476065
## rural               -1.3448298 -1.2725182
```

9. Regress the opioid overdose death rate (log) on the per capita pill consumption (log), but now interact the log of pill consumption with rural (use *) instead of adding it as a separate regressor. Interpret each coefficient. BE CAREFUL!

The intercept represents the expected log of opioid overdose deaths when all predictors are zero, so again, urban areas with no pills.

In urban areas, a one percent increase in per capita pill consumption is associated, on average, with a 0.49 *percent increase in the opioid overdose death rate.*

In rural areas, a one percent increase in per capita pill consumption is associated, on average, with a 0.34 *percent increase in the opioid overdose death rate. Remember that we can get this directly from the regression output adding 0.496 + -0.159…*

Finally, the negative coefficient on rural indicates that rural areas have an opioid overdose death rate that is about 54.32 *lower than that of urban areas, on average.*

```
model_interaction <- lm(log(overdosedeaths) ~ log(percapitapills) * rural, data = contigiousopioiddata)
summary(model_interaction)
```

```
## 
## Call:
## lm(formula = log(overdosedeaths) ~ log(percapitapills) * rural, 
##     data = contigiousopioiddata)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -5.5372 -0.7562 -0.1000  0.5585  5.3810 
## 
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)    
## (Intercept)                 0.73345    0.04589  15.983  < 2e-16 ***
## log(percapitapills)         0.49607    0.01269  39.082  < 2e-16 ***
## rural                      -0.78353    0.06707 -11.682  < 2e-16 ***
## log(percapitapills):rural -0.15860    0.01948  -8.143 4.03e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 1.127 on 26779 degrees of freedom
## Multiple R-squared:  0.2522, Adjusted R-squared:  0.2522 
## F-statistic:  3011 on 3 and 26779 DF,  p-value: < 2.2e-16
```