# Homework 5

Miles Tweed

October 12, 2020

## Problem 1

```
crime <- read.csv(file.path(project.dir,database.dir,"fl_crime.csv"))
str(crime)
```

```
## 'data.frame':    67 obs. of  5 variables:
##  $ county                   : Factor w/ 67 levels "Alachua","Baker",..: 1 2 3 4 5 6 7 8 9 10 ...
##  $ crime.rate..per.1000.    : int  104 20 64 50 64 94 8 35 27 41 ...
##  $ education....            : num  82.7 64.1 74.7 65 82.3 76.8 55.9 75.7 68.6 81.2 ...
##  $ urbanization....         : num  73.2 21.5 85 23.2 91.9 98.9 0 80.2 31 65.8 ...
##  $ income..median..in.1000.: num  22.1 25.8 24.7 24.6 30.5 30.6 18.6 25.7 21.3 34.9 ...
```

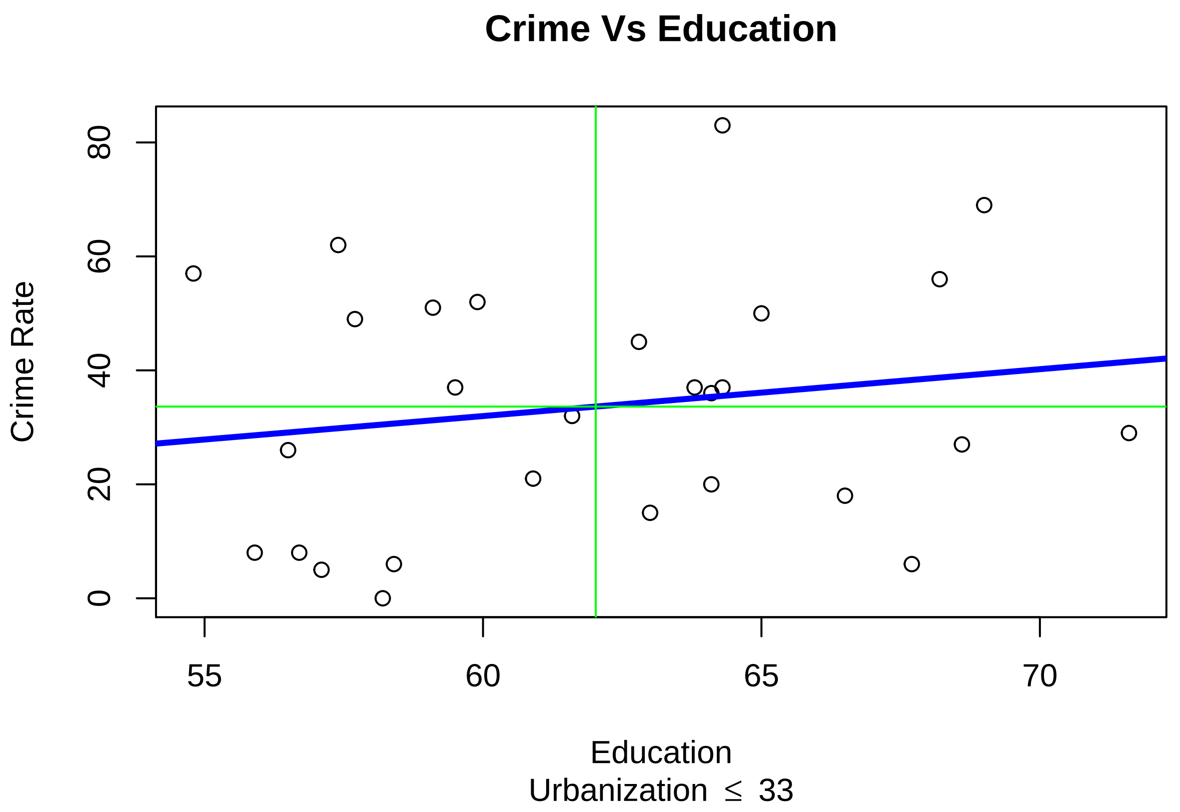
### Part 1

```
urb.le.33 <- crime[crime$urbanization.... <= 33,]
urb.btw.33.66 <- crime[crime$urbanization.... > 33 & crime$urbanization.... <=66,]
urb.btw.66.100 <- crime[crime$urbanization.... > 66 & crime$urbanization.... <=100,]
```

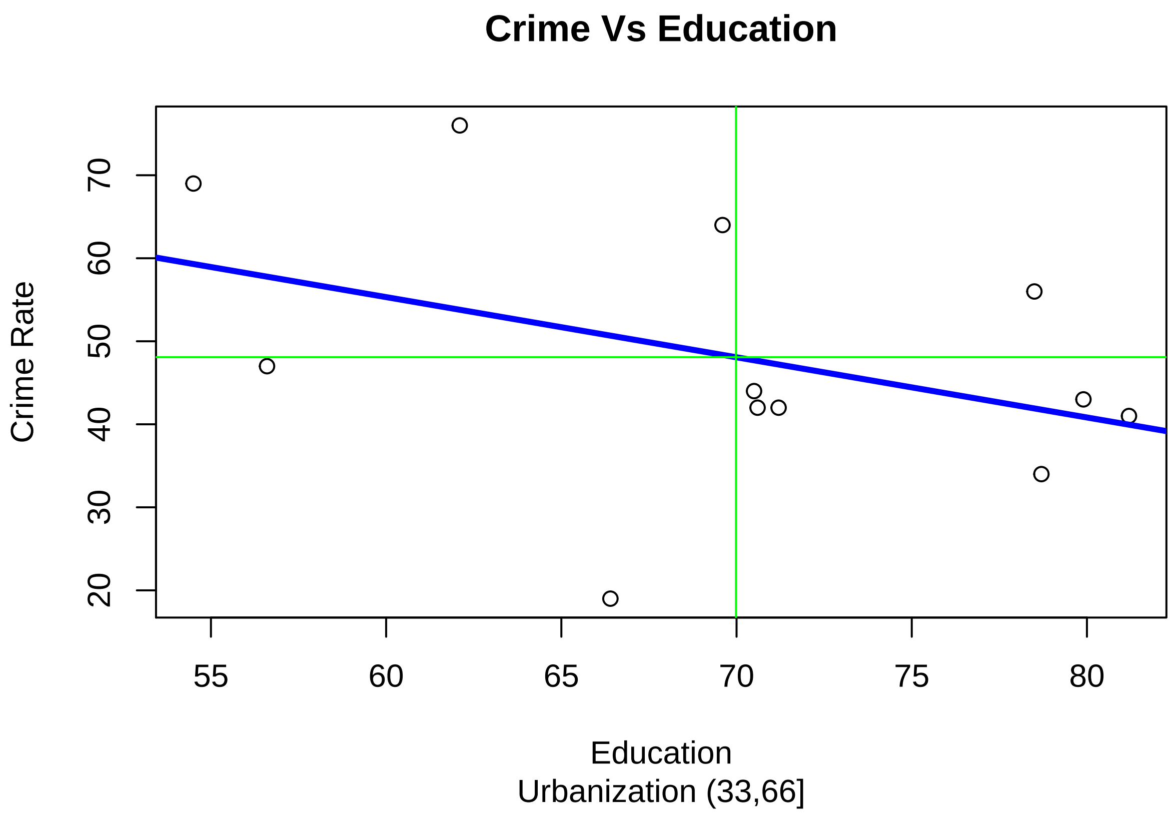
### Part 2

#### a

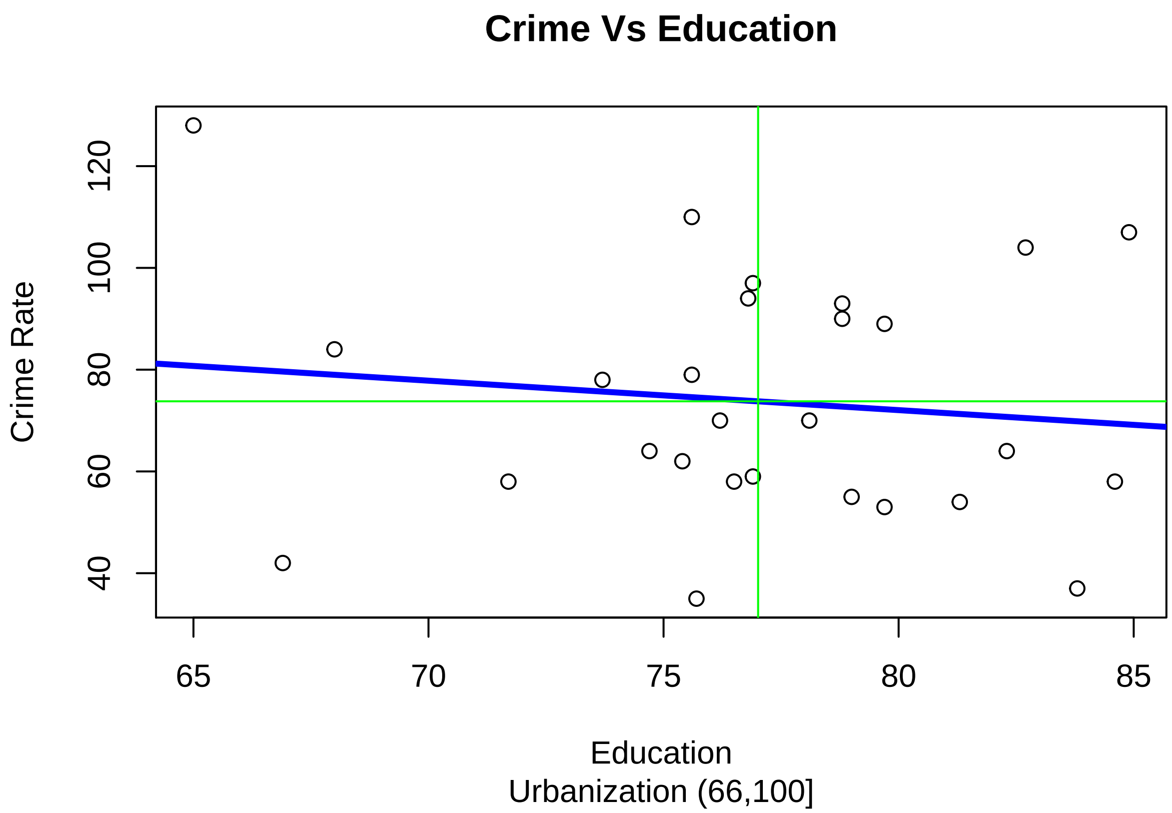
```
plot.new()
plot(crime.rate..per.1000. ~ education...., data = urb.le.33,
     main = "Crime Vs Education", sub = TeX("Urbanization $\\leq$ 33"),
     ylab = "Crime Rate", xlab = "Education")
le33.regrline <- lm(crime.rate..per.1000. ~ education...., data = urb.le.33)
abline(le33.regrline, lwd = 3, col = "blue")
abline(v = mean(urb.le.33$education....), lwd = 1, col = "green")
abline(h = mean(urb.le.33$crime.rate..per.1000.), lwd = 1, col = "green")
```



```r
plot.new()
plot(crime.rate..per.1000. ~ education...., data = urb.btw.33.66,
     main = "Crime Vs Education", sub = "Urbanization (33,66]",
     ylab = "Crime Rate", xlab = "Education")
btw33.66.regrline <- lm(crime.rate..per.1000. ~ education...., data = urb.btw.33.66)
abline(btw33.66.regrline, lwd = 3, col = "blue")
abline(v = mean(urb.btw.33.66$education....), lwd = 1, col = "green")
abline(h = mean(urb.btw.33.66$crime.rate..per.1000.), lwd = 1, col = "green")
```



```r
plot.new()
plot(crime.rate..per.1000. ~ education...., data = urb.btw.66.100,
     main = "Crime Vs Education", sub = "Urbanization (66,100]",
     ylab = "Crime Rate", xlab = "Education")
btw66.100.regrline <- lm(crime.rate..per.1000. ~ education...., data = urb.btw.66.100)
abline(btw66.100.regrline, lwd = 3, col = "blue")
abline(v = mean(urb.btw.66.100$education....), lwd = 1, col = "green")
abline(h = mean(urb.btw.66.100$crime.rate..per.1000.), lwd = 1, col = "green")
```



###

b

```
leq33.corr <- cor(urb.le.33$education...., urb.le.33$crime.rate..per.1000.)
btw33.66.corr <- cor(urb.btw.33.66$education...., urb.btw.33.66$crime.rate..per.1000.)
btw66.100.cor <- cor(urb.btw.66.100$education...., urb.btw.66.100$crime.rate..per.1000.)
print(paste("Correlation for urbanization less than or equal to 33:", leq33.corr))
```

```
## [1] "Correlation for urbanization less than or equal to 33: 0.174029179998998"
```

```
print(paste("Correlation for urbanization (33,66]:", btw33.66.corr))
```

```
## [1] "Correlation for urbanization (33,66]: -0.407823743695125"
```

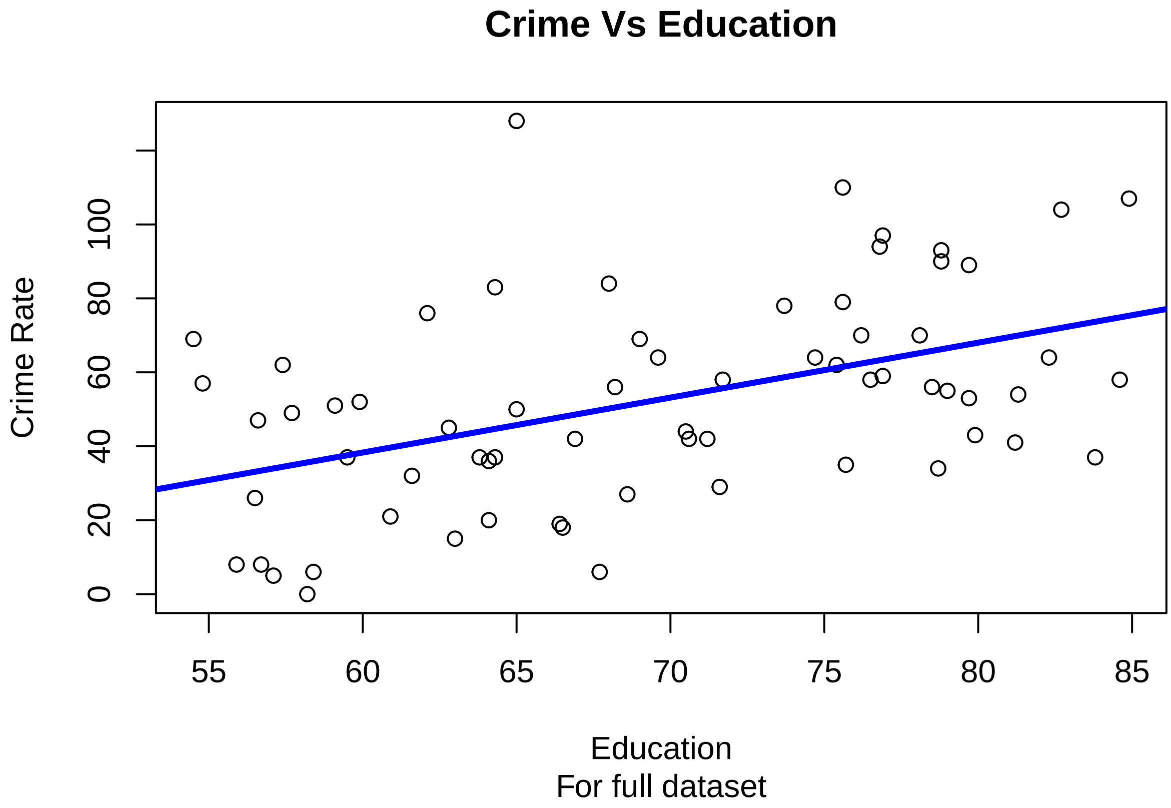
```
print(paste("Correlation for urbanization (66,100]:", btw66.100.cor))
```

```
## [1] "Correlation for urbanization (66,100]: -0.12229978329291"
```

## Part 3

Looking at the plot of the regression line for the whole data set, the direction of the linear relationship did change for the urbanization groups (33,66] and (66,100].

```
plot.new()
plot(crime.rate..per.1000. ~ education...., data = crime,
     main = "Crime Vs Education", sub = TeX("For full dataset"),
     ylab = "Crime Rate", xlab = "Education")
le33.regrline <- lm(crime.rate..per.1000. ~ education...., data = crime)
abline(le33.regrline, lwd = 3, col = "blue")
```



The correlation for every urbanization group is less than the overall correlation. The urbanization group that has a correlation closer to the overall in magnitude is (33,66] (-0.408), however the correlation for this group is negative (it has a different direction). The correlation for urbanization less than or equal to 33 had the same direction (positive) as the overall correlation while the correlation for the (66,100] group had a different direction (negative).

```
overall.corr <- cor(crime$education...., crime$crime.rate..per.1000.)
print(paste("Correlation overall:", overall.corr))
```

```
## [1] "Correlation overall: 0.466911880227914"
```

## Part 4

This phenomenon is called Simpson's Paradox and the urbanization variable is a confounding variable with respect to the crime-education relationship.

## Part 5

```
summary(btw66.100.regrline)
```

```
##
## Call:
## lm(formula = crime.rate..per.1000. ~ education...., data = urb.btw.66.100)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -39.536 -16.685  -4.247  18.679  47.273
##
## Coefficients:
```

```
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)     118.3366    72.4682   1.633    0.115
## education....    -0.5786     0.9391  -0.616    0.543
##
## Residual standard error: 23.98 on 25 degrees of freedom
## Multiple R-squared:  0.01496,    Adjusted R-squared:  -0.02444
## F-statistic: 0.3796 on 1 and 25 DF,  p-value: 0.5434
```

The formula for this trend line would be y = 118.3366 - 0.5786*x and the intercept would suggest that the number of crimes per 1000 residents was 118 in counties in which at least 66% of residents lived in urban areas and where no one had a high school diploma or more education. This intercept could make sense if there were such a county. The slope indicated that as the percentage of people with at least a high school diploma goes up by two the number of crimes per 1000 residents goes down by about one on average (worded with the consideration that 0.57 of a crime is nonsensical). The r-squared values suggests that only about 1.5% of the variation in crime rate could be explained by changes in education.

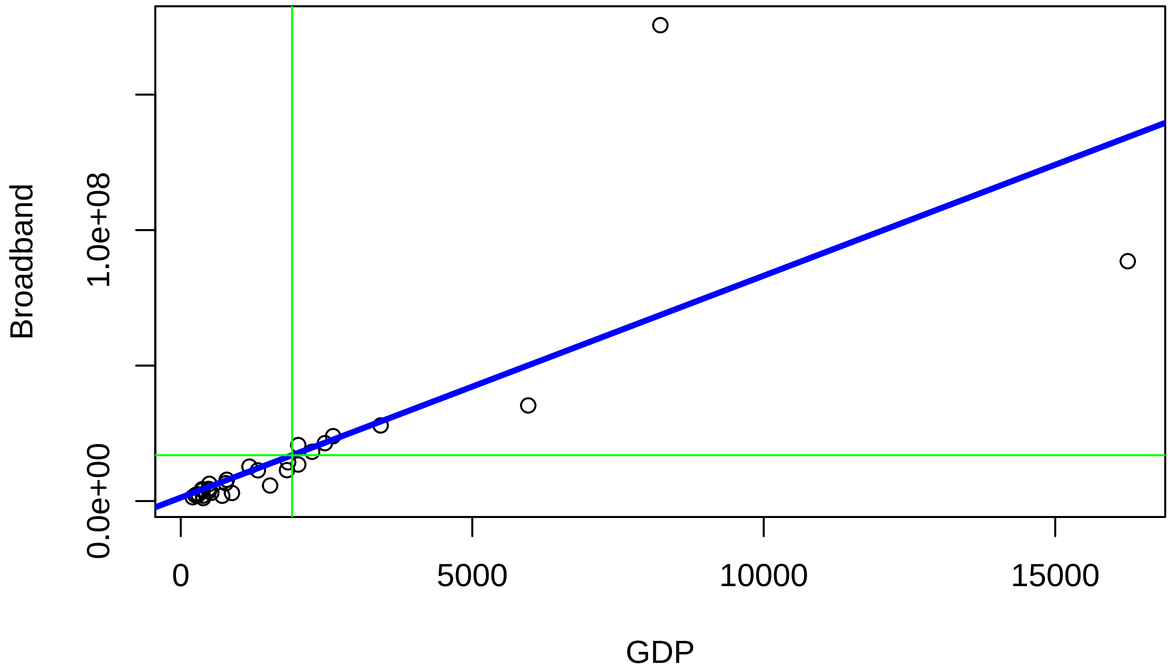
### Part 6

Since the data is quite scattered and does not closely follow the trend line, I would say that there are not any influential observation. In other words, none of the observations that are relatively low or high in the education range are much further away from the trend line as compared to the other data points. However, if one was identified it could be removed from the data set or the regression could be recalculated using the least absolute deviations method which is more resistant to outliers than the least squares method.

## Problem 2

### Part 1

```
broadband = read.csv(file.path(project.dir,database.dir,"Broadband_data.csv"))
#str(broadband)
plot.new()
plot(Broadband ~ GDP, broadband)
broad.regrline <- lm(Broadband ~ GDP, broadband)
abline(broad.regrline, lwd = 3, col = "blue")
abline(v=mean(broadband$GDP),col = "green")
abline(h=mean(broadband$Broadband),col = "green")
```



```
broadband.corr <- cor(broadband$GDP, broadband$Broadband)
print(paste("Correlation:", broadband.corr))
```

```
## [1] "Correlation: 0.770648779292529"
```

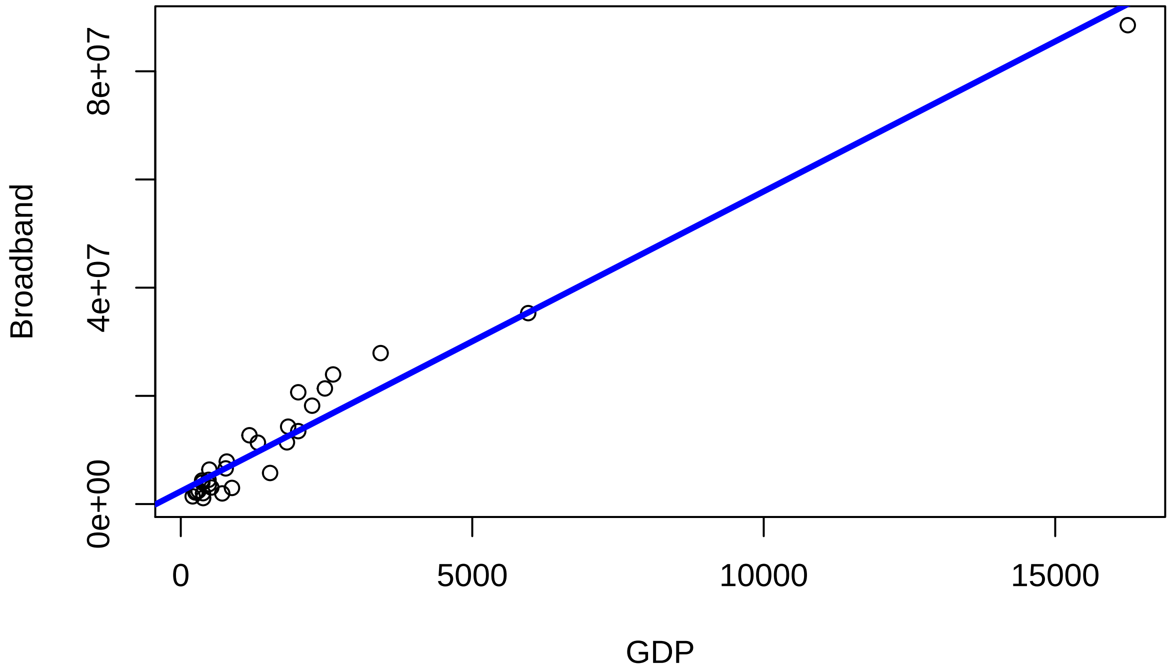
China seemed to be the regression outlier with a GDP of 8227 (billion) and 175624800 broadband subscribers. This is because the x value is well beyond the average value for GDP and the y value is well above the trend in the data for broadband subscribers.

```
broadband[broadband$Broadband == max(broadband$Broadband),]
```

```
##   Country  GDP Broadband
## 7   China 8227 175624800
```

The linear regression fit the data much more closely after removing the outlier and the correlation moved much closer to one. Because of these facts I would trust the regression line more for data within this range after removing the outlier than I would before.

```
broadband2 <- broadband[!broadband$Country=="China",]
plot.new()
plot(Broadband ~ GDP, broadband2)
broad2.regrline <- lm(Broadband ~ GDP, broadband2)
abline(broad2.regrline, lwd = 3, col = "blue")
```



```
broadband2.corr <- cor(broadband2$GDP, broadband2$Broadband)
print(paste("Correlation:", broadband2.corr))
```

```
## [1] "Correlation: 0.980381655409952"
```

## Part 2

```
library(L1pack)
lad.regrline <- lad(Broadband ~ GDP, broadband)
print(paste("The slope using LAD regression",lad.regrline$coefficients[2]))
```

```
## [1] "The slope using LAD regression 6390.26904296875"
```

```
print(paste("The slope using least squares regression before removing outlier",broad.regrline$coefficie
```

```
## [1] "The slope using least squares regression before removing outlier 8188.75414344233"
```

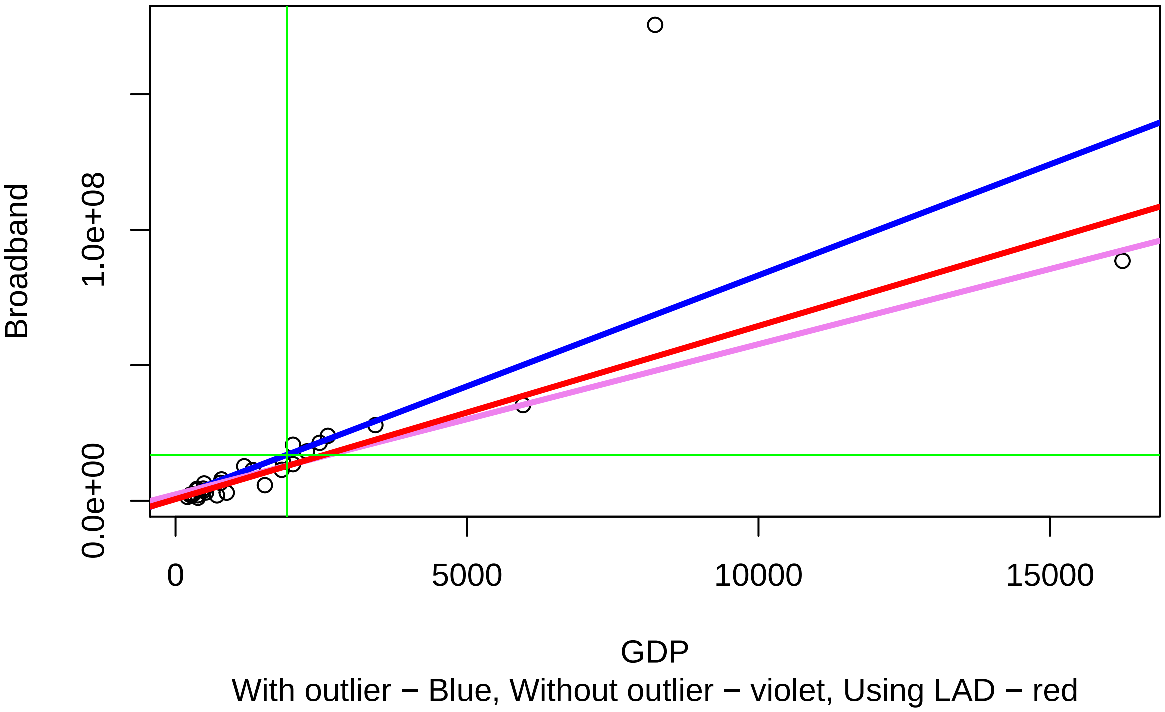
```
print(paste("The slope using least squares regression after removing outlier",broad2.regrline$coefficie
```

```
## [1] "The slope using least squares regression after removing outlier 5544.08616246035"
```

The slope of the regression line using the least absolute deviations method while including the outlier is much closer to the slope of the regression line after removing the outlier than it is to the regression line before removing the outlier. This demonstrates that this method is more more resistant to outliers that the least squares method. For the LAD method, the absolute value of the difference between the actual and predicted result is optimized while the least squares method optimizes the squared difference. As a result, outliers have a much greater impact on the squared difference than the absolute difference causing a much bigger change in the slope.

```
plot.new()
plot(Broadband ~ GDP, broadband, sub = "With outlier - Blue, Without outlier - violet, Using LAD - red")
broad.regrline <- lm(Broadband ~ GDP, broadband)
abline(broad.regrline, lwd = 3, col = "blue")
abline(broad2.regrline, lwd = 3, col = "violet")
abline(lad.regrline, lwd = 3, col = "red")
```

```r
abline(v=mean(broadband$GDP),col = "green")
abline(h=mean(broadband$Broadband),col = "green")
```



## Problem 3

### Part 1

The goal would be to show that the presence of COVID did not have an impact on the death rate of people with certain underlying conditions. We would need to collect information about the death rates of people with similar conditions in time periods similar and prior to the time of the COVID outbreak. If possible, we could collect the death rate of people with those conditions that also had other flu like diseases simultaneously.

### Part 2

In this scenario, which is concerned with states in which weed had been legalized, the response variable seems to be the number of fatal traffic accidents and the explanatory variable would be the presence of marijuana in the drivers system. I would imagine the primary lurking variable is the increase in tourism to the state for recreational marijuana. The increase in weed tourism would have a causal relationship to both an increase in traffic, as well as traffic accidents, and an increase in people with weed in their systems. Other, lurking variables could include the the weather, traffic conditions, or even the maintenance of the vehicles involved.

## Problem 4

### Part 1

She should assign students to the groups randomly. By using random selection she can ensure that the groups are balanced between gender, GPA, age, and other variable. This way the effects of lurking variables that could interfere with inferring causality are accounted for.

```r
student <- read.csv(file.path(project.dir,database.dir,"fl_student_survey.csv"))
group1 <- student[student$subject %in% sample(student$subject, 30),]
group2 <- student[!(student$subject %in% group1$subject),]
summary(group1[,c("gender", "age", "college_GPA")])
```

```
##  gender      age         college_GPA   
##  f:16   Min.   :22.00   Min.   :2.600  
##  m:14   1st Qu.:23.00   1st Qu.:3.125  
##         Median :26.00   Median :3.500  
##         Mean   :27.57   Mean   :3.433  
##         3rd Qu.:30.25   3rd Qu.:3.700  
##         Max.   :50.00   Max.   :4.000
```

```r
summary(group2[,c("gender", "age", "college_GPA")])
```

```
##  gender      age         college_GPA   
##  f:15   Min.   :22.00   Min.   :2.800  
##  m:15   1st Qu.:25.25   1st Qu.:3.200  
##         Median :27.50   Median :3.500  
##         Mean   :30.77   Mean   :3.473  
##         3rd Qu.:31.75   3rd Qu.:3.800  
##         Max.   :71.00   Max.   :4.000
```

As shown in the summary, by taking a random sample of the students, the two groups are roughly balanced between age, gender, and college GPA.

### Part 2

I would assign one group per teaching method throughout one semester and compare the group's grades. Perhaps the teaching methods could be switch the following semester to ensure that any noticeable differences were not due to the makeup of the group, however, this should have been accounted for in the sampling method. Additionally, I believe it would be important that the two groups are in courses of the same subject matter so that the perceived differences were not confounded by that fact.

### Part 3

I believe she could generalize the measured effects onto the wider student population because of the care taken to minimize bias by selecting the groups using a randomized approach. This is because the groups are relatively balanced between age, sex, and college GPA. It is important to note that the median GPA for both groups is around 3.5,so generalization would be okay as long as this was also the case for the general student body. The primary concern I would have is that the students were a volunteer sample which may introduce bias based on the students that are likely to volunteer for such a study. These students may be ones that are more motivated to try new approaches to learning or they may be students of the professor who are looking to gain favor with her by volunteering.

## Problem 5

3.91

a) Gender could be a potential lurking variable because, on average, males are taller than females and also tend to make more money (on average for no good reason).

b) If gender was included in the study it would be a confounding variable. This is because it has association with both income and height which would make it difficult to determine what the cause of increased income really is.

3.92

a) No, correlation does not imply causation, and there may be lurking variables.

b) One possible lurking variable could be development status. More developed nations, which may have more per capita television ownership, could have lower birth rates due to less reliance on large families for labor as well as better birth control methods.

4.2

a) This is an observational study.

b) The explanatory variable is the combination of high blood pressure and binge drinking. The response variable is dying from a stroke or heart attack.

c) The study does not prove that a combination of high blood pressure and binge drinking causes increased death from heart attack or stroke because there could be unaccounted for lurking variables such as genetics or other dietary intakes.

4.3

a) The response variable is the amount of weight lost and the explanatory variable is the type of diet followed.

b) This is an experimental study because the participants were assigned to certain diets which means that the researchers had control over the types of diet and groups associated with each.

c) It would not be appropriate to recommend a low-carb diet to everyone who wished to lose weight for several reason. First, the participants in this study were specifically selected to be without heart disease or diabetes and it would be ill advised to recommend a low-carb diet over a low-fat to someone who did have heart disease. Additionally, there may have been lurking variables that contributed to the differences in weight loss over one year, for example, perhaps the low-carb diet was easier to maintain while those in the low-fat group tended to stray from the diet.

4.9

a) An observational study would be better for this since smoking is hazardous and it would be amoral to force participants to smoke.

b) An observational study would be best because it would impossible to control the participants' performance on the SAT. A related study that could be experimental would be to track the performance of a group that took part in a specific SAT preparation course and those that did not.

c) This would be an experimental study because it would be feasible and moral to randomly send catalogs with and without the coupon and track those that used it to order products.

4.22

a) The population of interest in this study is all employers of those with a college degree.

b) In order to calculate a non-response rate we would need to know how many employers were given the survey in total.

c) This could have sampling bias due to it being a nonrandom sample since it was a volunteer sample that only included employers that interacted with at least one of 300 college career service centers. There could also be non-response bias if the number of employers offered the survey was much higher that the numbers that responded. For example, perhaps only those hiring people with a masters degree or a PhD were likely to respond. Depending on the working of the survey, there could also be response bias. The questions could have been worded for responses about recent hires as opposed to most common hires.

4.26

a) Sampling bias would be introduced because it is an online survey. This means that only teenagers likely to see and respond to a survey online would respond thus the survey is not sampling a sufficiently random sample frame.

b) Non-response bias would occur because the teenagers would be unlikely to admit to doing something that would be a crime and would likely not respond to the survey

c) Since the action in question would be a crime, those who did respond may not respond truthfully in order to avoid implication.