Homework 4

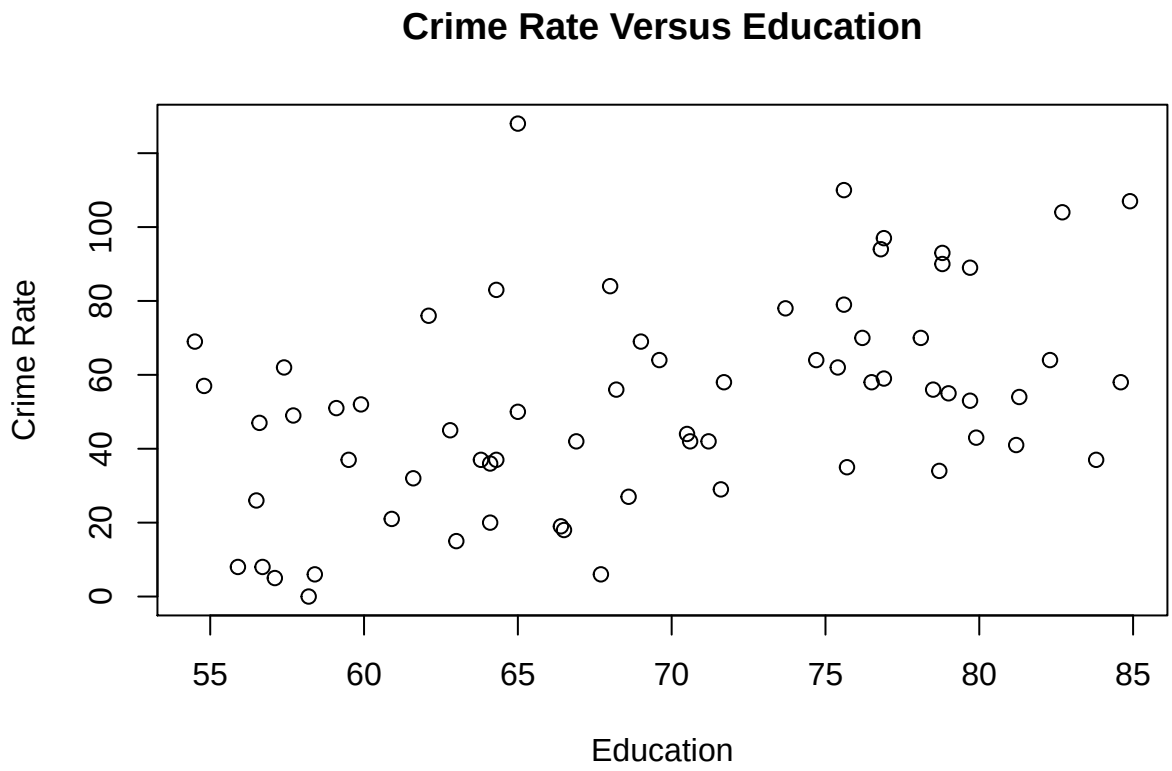
Miles Tweed

Problem 1

1.1

The plot of crime rate versus education contain quite a bit of spread, but overall there seems to be a linear relationship. Even though at any one level of education there is a range in associated crime rate, the crime rate seems to increase with education. It could be possible to describe this relationship with a straight line.

```
crime <- read.csv("http://sites.williams.edu/bklingen/files/2015/06/fl_crime.csv")
attach(crime)
names(crime) <- c("county", "crime.rate", "education", "urbanization", "median.income")
#str(crime)
plot(crime.rate ~ education, data = crime,
     xlab = "Education", ylab = "Crime Rate",
     main = "Crime Rate Versus Education")
```



1.2

The correlation between crime rate and education is:

```
cor(crime$crime.rate, crime$education)
```

```
## [1] 0.4669119
```

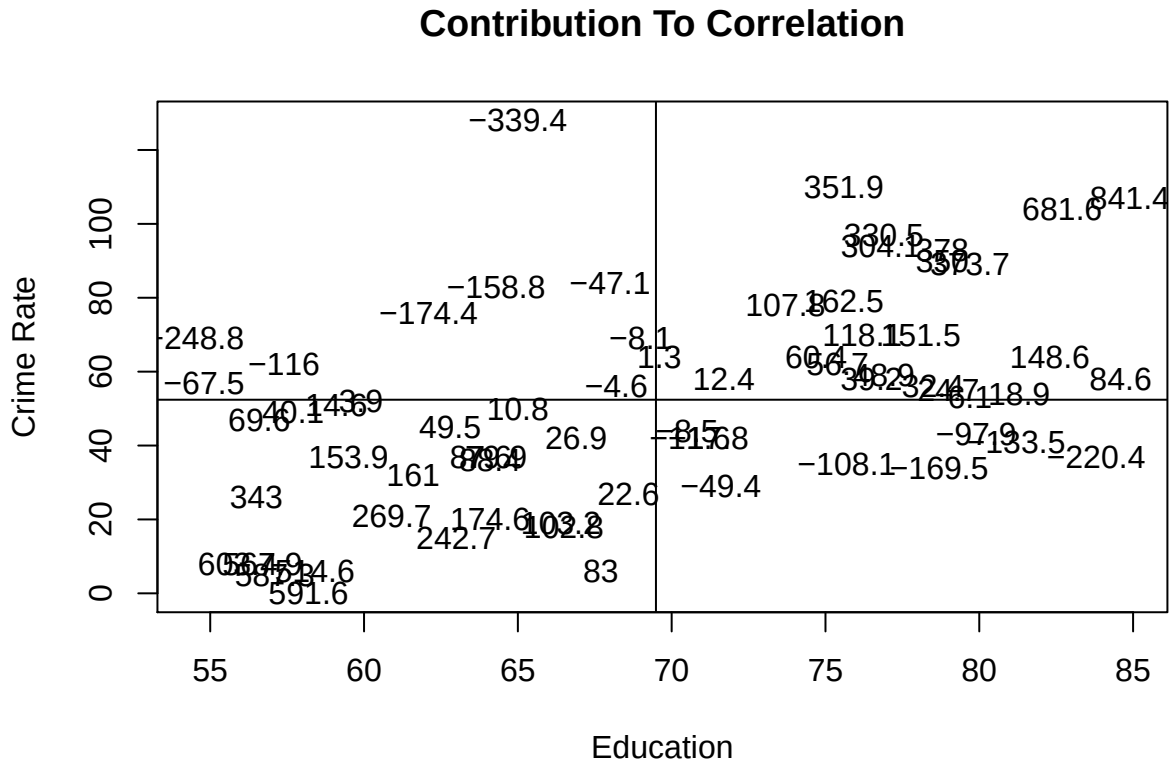
Since this number is between one and zero, it indicates that there is some positive correlation between crime rate and education. The value is closer to zero than it is to one which indicates that the correlation is not

very strong. Viewing a graph in which the means are plotted can help to elucidate which values contribute positively to this correlation and which contribute negatively.

```
x <- crime$education
y <- crime$crime.rate

x.score <- x-mean(x)
y.score <- y-mean(y)

plot(y~x, type="n", ylab = "Crime Rate", xlab = "Education", main = "Contribution To Correlation")
text(y~x, labels = round(x.score*y.score,1))
abline(h=mean(y), v=mean(x))
```



1.3

Fitting a linear regression model to the data yields:

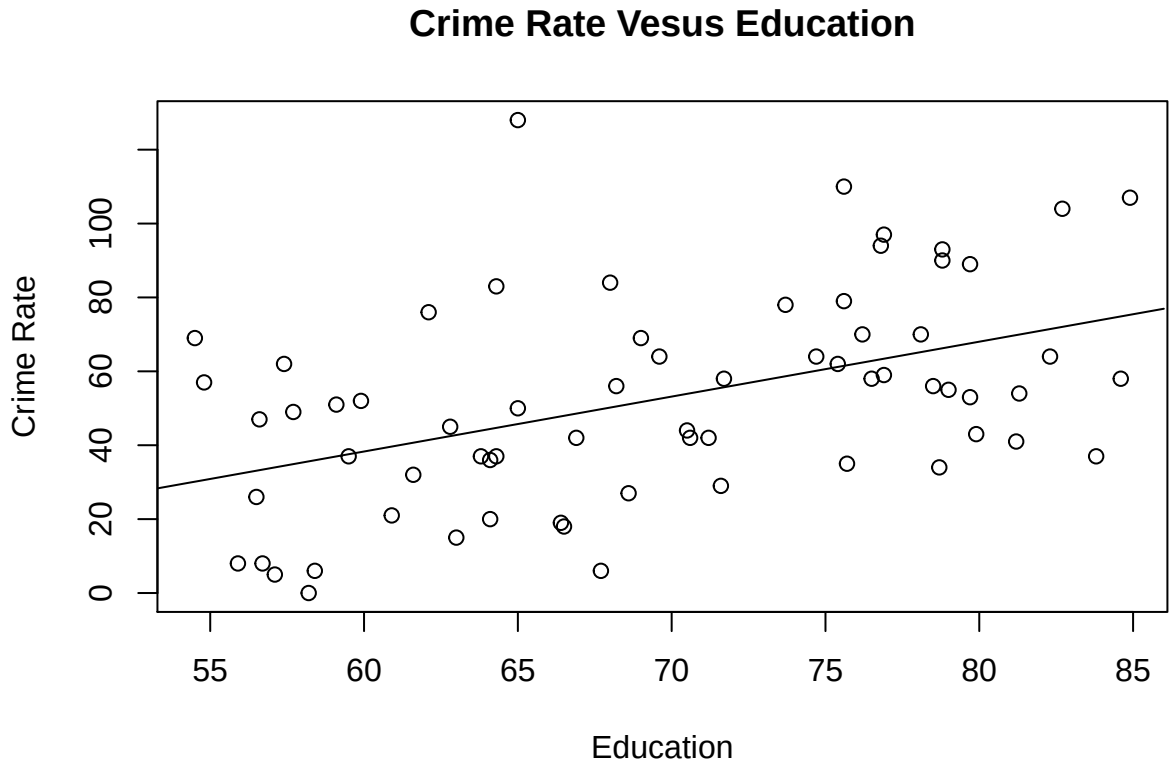
```
lm.obj <- lm(crime.rate ~ education, data = crime)
lm.obj
```

```
##
## Call:
## lm(formula = crime.rate ~ education, data = crime)
##
## Coefficients:
## (Intercept)    education
##     -50.857        1.486
```

a)

According to the linear fit the equation for the fitted linear regression of crime rate versus education is $\hat{y} = 1.486x - 50.857$. Looking at the overlapping graphs, This equation does seem to fit the data.

```
plot(crime.rate~education, data = crime,
     ylab = "Crime Rate", xlab = "Education",
     main = "Crime Rate Vesus Education")
curve(1.486*x - 50.857, from  = 53, to = 86, add = T)
```



b)

The intercept is -50.857 which would suggest a negative crime rate in a population with no education. A negative crime rate would be nonsensical because the lowest level of crime that could exist would be none.

c)

The slope indicates that a unit increase of education corresponds to a 1.486 unit increase in crime rate.

d)

```
summary(lm.obj)
```

```
##
## Call:
## lm(formula = crime.rate ~ education, data = crime)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -43.74 -21.36  -4.82  17.42  82.27
##
```

```
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -50.8569    24.4507  -2.080   0.0415 *
## education     1.4860     0.3491   4.257 6.81e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 25.12 on 65 degrees of freedom
## Multiple R-squared:  0.218,  Adjusted R-squared:  0.206
## F-statistic: 18.12 on 1 and 65 DF,  p-value: 6.806e-05
```

The R-squared value for crime rate versus education is approximately 0.22 which indicates that 22% of the variation in crime rate can be explained by education.

1.4

```
y <- crime$crime.rate
y.hat <- predict(lm.obj)
y.bar <- mean(y)

r.sq <- (sum((y-y.bar)^2)-sum((y-y.hat)^2))/sum((y-y.bar)^2)
cat(paste("The hard coded calculation for r-squared is",
           round(r.sq, digits = 2),".\nThis is identical for the value in part 3."))
```

```
## The hard coded calculation for r-squared is 0.22 .
## This is identical for the value in part 3.
```

1.5

```
y.hat.70 <- 1.486*70 - 50.857
y.hat.35 <- 1.486*35 - 50.857

cat(paste("The pridicted crime rate at 70 is", y.hat.70,
          "\nand the predicted crime rate at 35 is", y.hat.35))
```

```
## The pridicted crime rate at 70 is 53.163
## and the predicted crime rate at 35 is 1.153
```

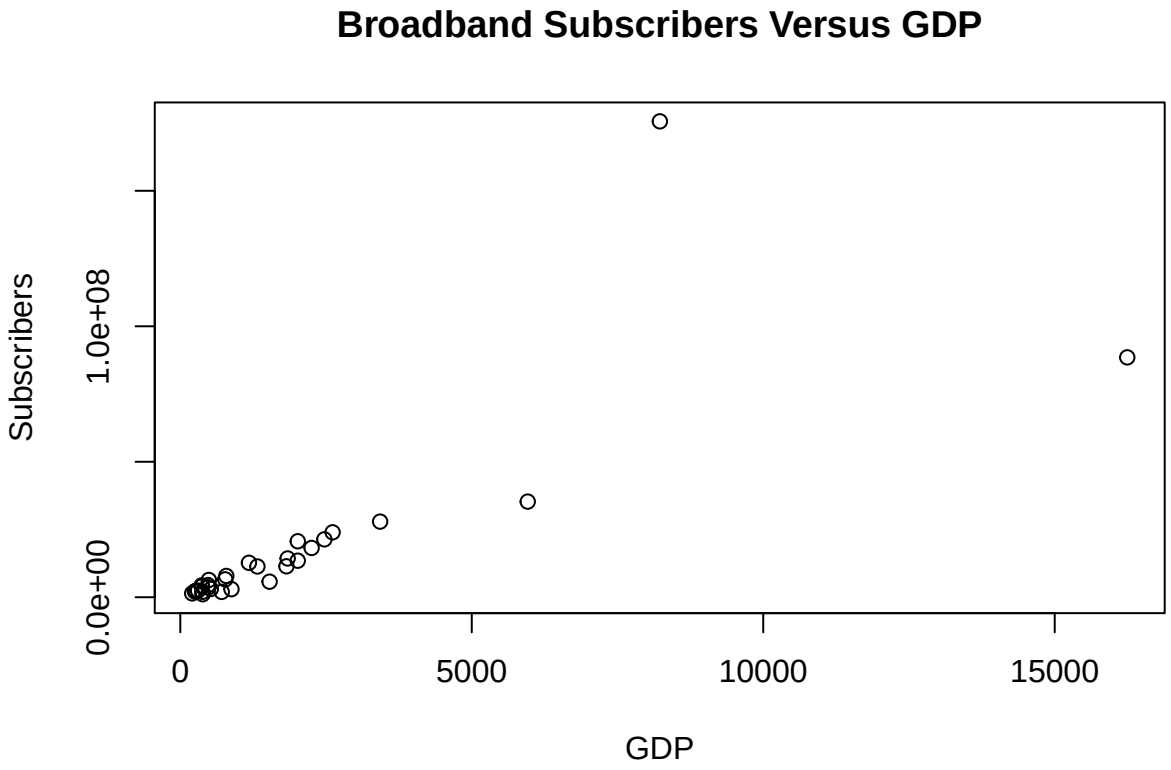
In general, these would likely be poor predictions of the actual crime rate. The prediction at 70 would be the better of the two prediction since it is within the range of the data on hand, however only 22% of the variability in crime rate is explained by education so any prediction that only used education will not be very accurate. The prediction at 35, however, would be extrapolation since it is below the range of the given data and it is dangerous to extrapolate because the predictions are likely to be incorrect.

Problem 2

2.1

Plot of broadband subscribers versus GDP:

```
broadband <- read.csv(file.path(project.dir,dataset.dir,"Broadband_data.csv"))
#str(broadband)
plot(Broadband ~ GDP, data = broadband,
     ylab = "Subscribers", xlab = "GDP",
     main = "Broadband Subscribers Versus GDP")
```



The correlation between crime rate and education is:

```
cor(broadband$Broadband, broadband$GDP)
```

```
## [1] 0.7706488
```

This correlation is positive and close to one, therefore, the GDP of a country and the number of broadband subscribers is highly correlated. In other words it is highly likely that countries with higher GDP will have a greater number of broadband subscribers. This makes sense because increased wealth would allow for increased access to luxuries like broadband internet.

2.2

```
bb.lm.obj <- lm(Broadband ~ GDP, data = broadband)
bb.lm.obj
```

```
##
## Call:
## lm(formula = Broadband ~ GDP, data = broadband)
##
## Coefficients:
## (Intercept)          GDP
##     1292551         8189
```

a)

The equation of the regression line is $\hat{y} = 1292551 + 8189x$.

b)

The intercept suggests that a country with a GDP of 0 would have 1,292,551 broadband internet subscribers which would not make sense. This intercept is not useful information and should not be trusted because the

smallest value for GDP in the sample is 204 billion which would indicate that a prediction at 0 GDP would be a drastic extrapolation.

c)

The slope suggests that an increase of 1 billion dollars in GDP corresponds to 8189 more broadband subscribers.

d)

```
summary(bb.lm.obj)
```

```
##
## Call:
## lm(formula = Broadband ~ GDP, data = broadband)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -45798862  -2728052  -1333465   -301567 106963369
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1292551    4487513   0.288    0.775
## GDP             8189       1236   6.624 2.47e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 21590000 on 30 degrees of freedom
## Multiple R-squared:  0.5939, Adjusted R-squared:  0.5804
## F-statistic: 43.87 on 1 and 30 DF,  p-value: 2.475e-07
```

The r-squared values for broadband subscribers versus GDP is 0.59 which indicates that 59% of the variation in broadband subscribers is explained by the countries GDP.

3.26

a)

```
a <- 9.2 + 77*2000
b <- 9.2 + 77*3000

cat(paste("The predicted value of a 2000 sq. ft. home is", a,
            "\nThe predicted value of a 3000 sq. ft. home is", b))
```

```
## The predicted value of a 2000 sq. ft. home is 154009.2
## The predicted value of a 3000 sq. ft. home is 231009.2
```

b)

The slope suggests that a one square foot increase in house size corresponds to a $77 increase in price.

c)

The correlation between these variables is positive since an increase in square footage correlates to an increase in price.

d)

```
c <- 300000 - 231009.2
print(paste("The residual is", c))
```

```
## [1] "The residual is 68990.8"
```

Since the residual is positive, the prediction was an under estimate of the actual value. The regression predicts that the values of the home would be almost $70,000 less that it sold for.

3.70

a)

(i) b
(ii) a
(iii) c

b)

Based on the plots I would conclude that there seems to be a relationship between the amount of dust and rainfall in Africa where more rain corresponds to less dust traveling over the Atlantic Ocean.

3.83

a)

```
d <- -20000/1.25

print(paste("The intercept for the equation in euros is", d))
```

```
## [1] "The intercept for the equation in euros is -16000"
```

b)

```
e <- 4000/1.25

print(paste("The slope for the equation in euros is", e))
```

```
## [1] "The slope for the equation in euros is 3200"
```

c)

The correlation when the regression is measured in euros is still 0.50 because correlation is not dependent on the units used to measure either of the quantities.