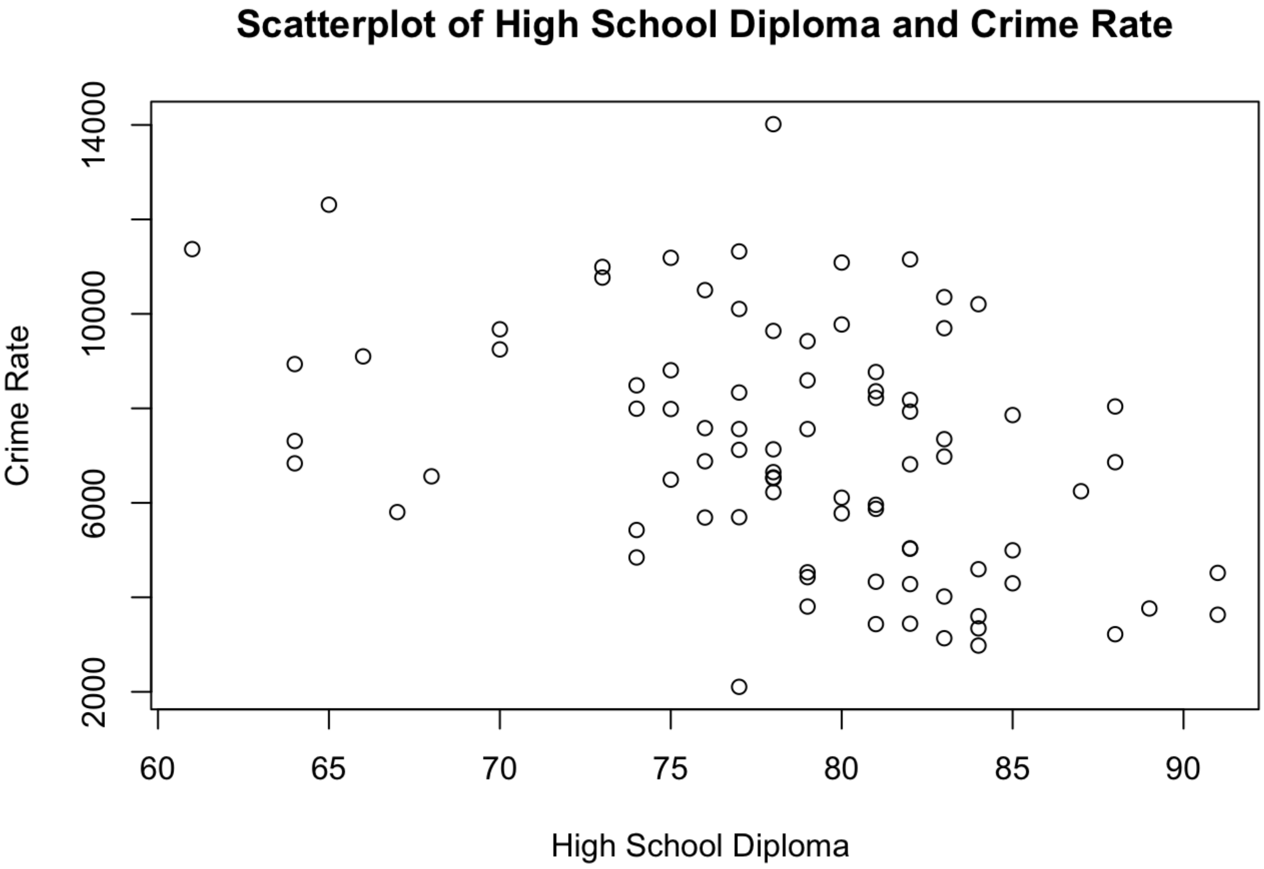
Crime Data Analysis

Oliver Hannaoui

6/13/2019

Crime.csv file is used for the dataset. It has two columns, one of which is the percentage of individuals in the county with at least a high-school diploma (columndip), and the other is the crime rate per 100,000 residents for thecounties (columnrate). ConsiderYto be crime rate, andXto be percentage with high school diploma.

I. a) Below is a scatterplot of the percent of people in each county with a high school diploma on the x-axis and the crime rate per 100,000 residents for the counties on the y-axis.

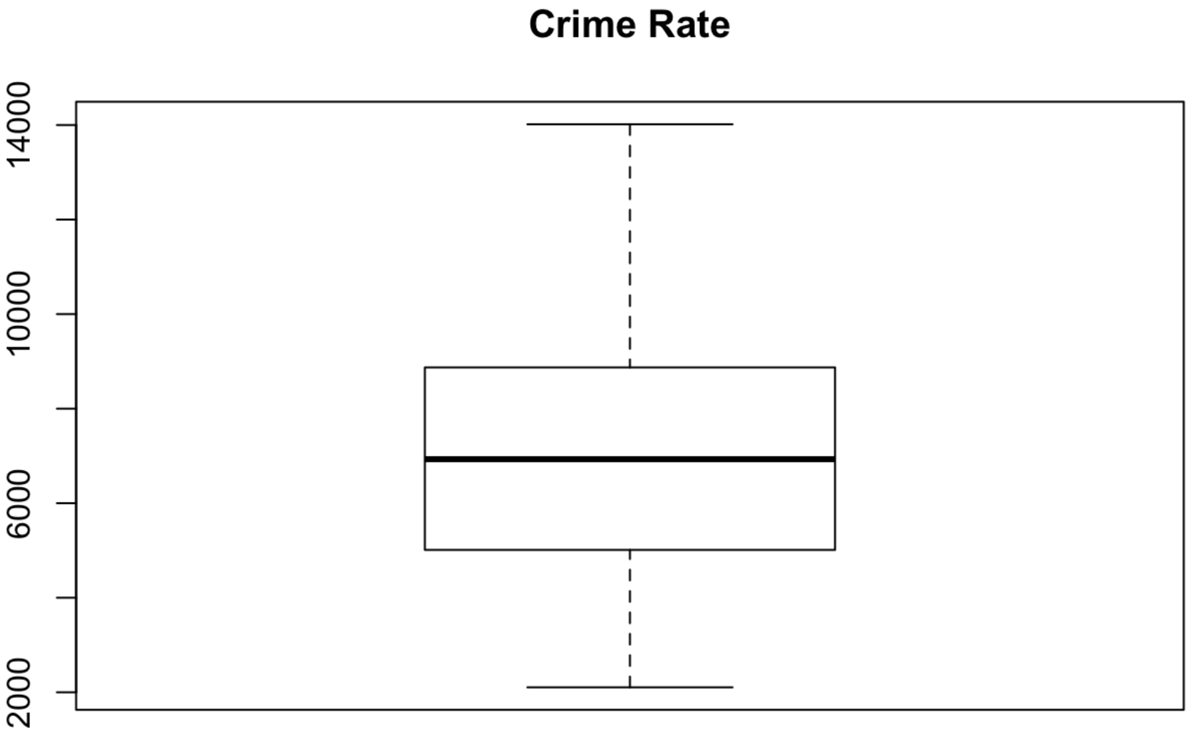


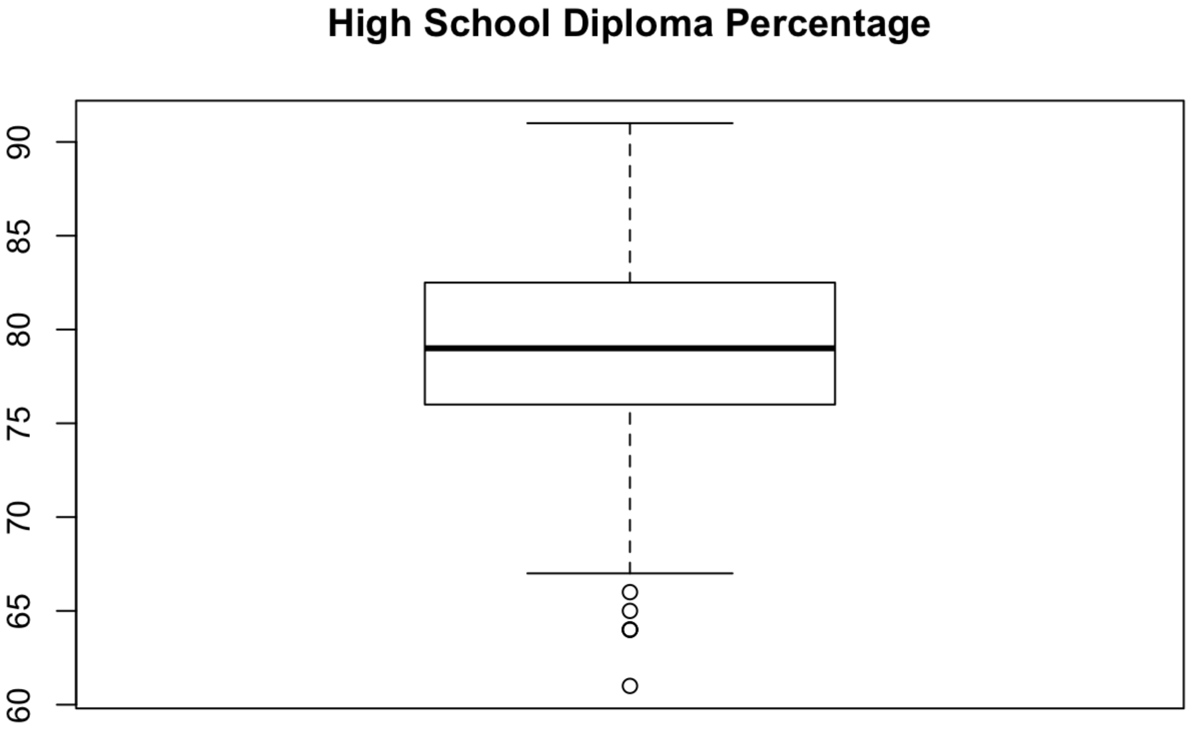
b. Estimating the regression model we find that the intercept is 20517.5999 and the slope is -170.5752.

```
## (Intercept)         dip
##  20517.5999   -170.5752
```

c. The slope depicts the estimated average decrease in crime rate is 170.5752 per 100,000 residents with a one percent increase in high school diploma attainment. The y intercept says that when the high school diploma attainment percentage is 0, the crime rate per 100,000 residents is 20,5017.99.

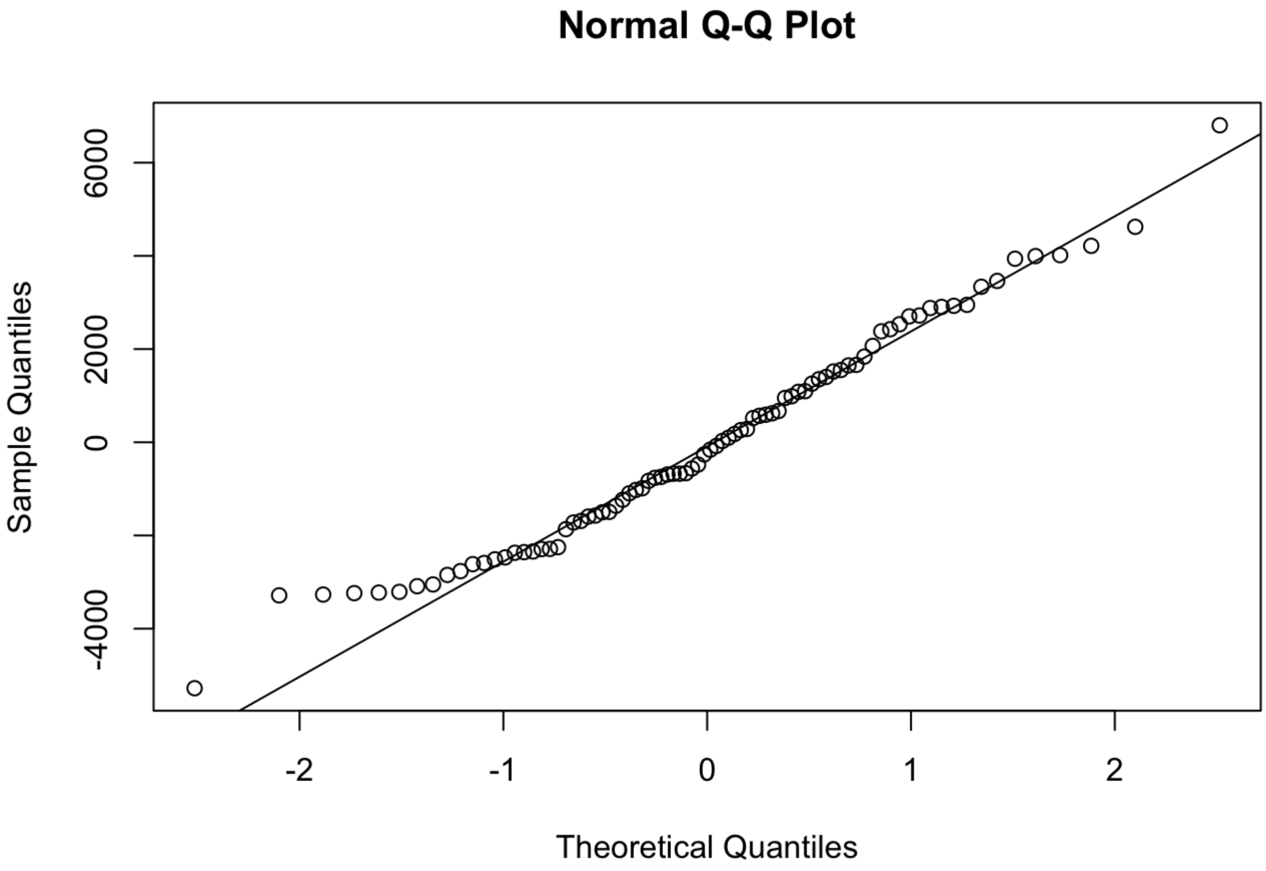
d. Based off the boxplots for crime rate and diploma percentage we can visually see that there are outliers in the boxplot for diploma percentage. Subsetting the data to identify the outliers we find that there are 6 outliers which are all below the value of 67.


Crime Rate

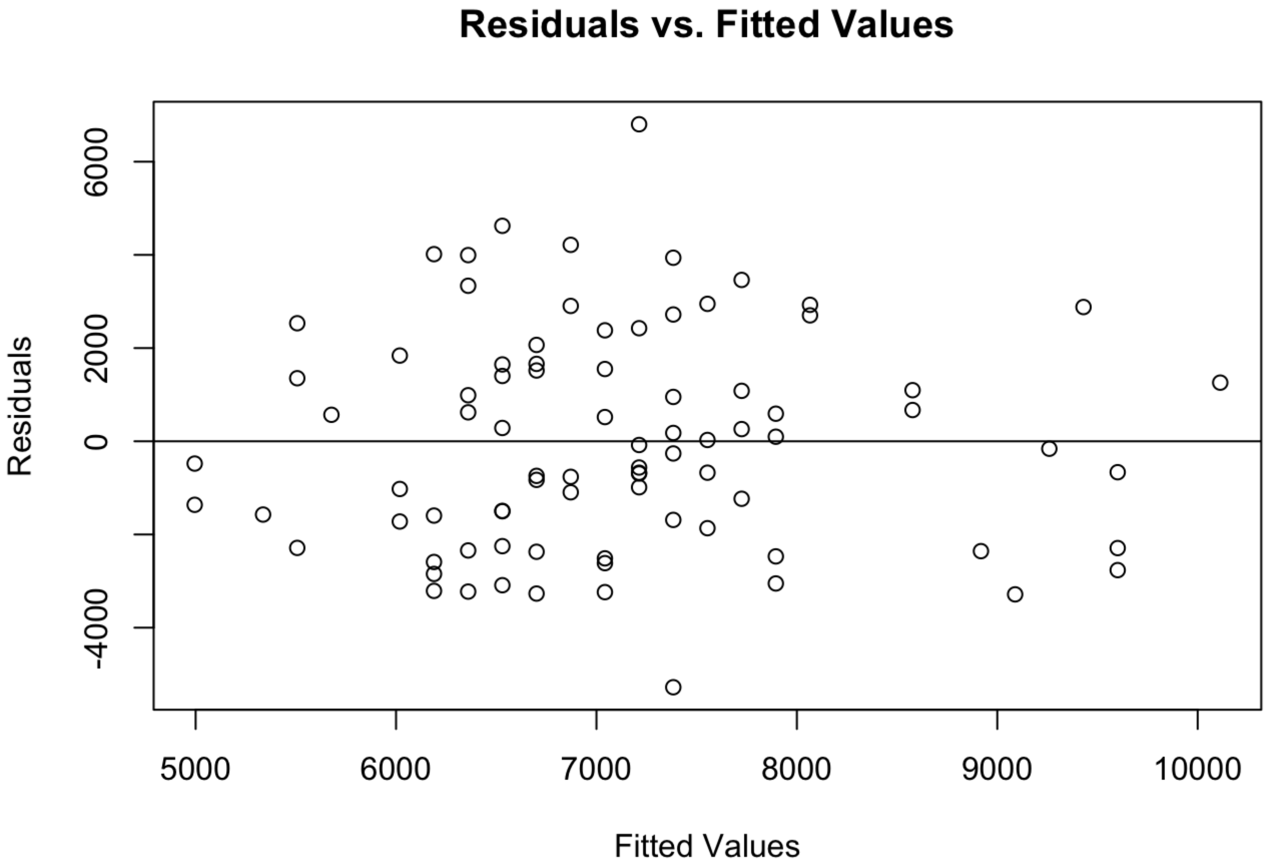

High School Diploma Percentage

```
## [1] 66 65 61 64 64 64
```

e. The Q-Q plot of the residuals appears to be normally distributed. We can observe the Normal Q-Q plot below and see that the relationship between the sample quantiles and the theoretical quantiles are approximately linear. Therefore we can say the normal probability plot is indeed normally distributed.



f. Plotted below are the errors vs fitted values. It does not appear that the variance of the errors is constant, this is because there is no clear trend in the values below therefore the variance is not explained and the variance of the errors is not constant.



g. Below is the 95% confidence interval for the slope. We are 95% confident that the average decrease in crime rarte per 100,000 people for every one percent increase in high school diploma attainment is between 87.87061 and 253.27977.

```
##      2.5 %     97.5 %
## -253.27977  -87.87061
```

II.

a. The null hypothesis is that the true average percent of people with a high school diploma is 80. The alternate hypothesis is that the true average percent of people with a high school diploma is not 80.

b. The test statistic is -2.096.

```
## 
##  One Sample t-test
## 
## data:  crime$dip
## t = -2.0696, df = 83, p-value = 0.0416
## alternative hypothesis: true mean is not equal to 80
## 90 percent confidence interval:
##  77.46615 79.72432
## sample estimates:
## mean of x 
##  78.59524
```

c. Based off the one sample t-test we calculated in part b) we found that the p-value is 0.0416. Given the null hypothesis is true that the average percent of people with a high school diploma is 80, the probability that we observe the sample data is 0.0416.

d. Since the p-value is less than the level of significance of 0.1, we chose to reject the alternate hypothesis that the true average percent of people with a high school diploma is 80.

e. We reject the null hypothesis that the true average is 80 and accept the alternate hypothesis that the true average percent of people with a high school diploma is not 80.

f. Based of the one sample t-test calculated in part b) we found a 90% confidence interval with a lower bound of 77.46615 and an upper bound of 79.72432. This confidence interval supports our decision that the true average percent of people with a high school diploma is not 80 since 80 does not lie within our lower or upper bounds of the confidence interval.

g. We could have made a Type I error. A Type I error occurs when we reject the null hypothesis when in reality the null hypothesis was correct. In the context of this hypothesis test a Type I error would occur if the true average percent of people with a high school diploma is 80 since we rejected this hypothesis based off the one sample t-test.

Appendix:

```
# Loading the data
crime <- read.csv("~/Desktop/crime.csv")
#1a
plot(crime$dip,crime$rate,main="Scatterplot of High School Diploma and Crime Rate",xlab
 = "High School Diploma ",ylab="Crime Rate")
#1b
crime.model = lm(rate~dip, data = crime)
crime.model$coefficients
#1d
boxplot(crime$rate, main='Crime Rate')

boxplot(crime$dip, main="High School Diploma Percentage")

outliers = boxplot(crime$dip, plot=FALSE)$out
# Our outliers from the High School Diploma Percentage boxplot
outliers
#1e
qqnorm(crime.model$residuals)
qqline(crime.model$residuals)
#1f
plot(crime.model$fitted.values,crime.model$residuals, main="Residuals vs. Fitted Values"
, xlab="Fitted Values", ylab='Residuals')
abline(h=0)
#1g
confint(crime.model)[2,]
#2b
t.test(crime$dip, mu = 80, alternative = "two.sided", conf.level = 0.9)
```