Class Project Part 2

Sean Leggett - BDA201 Winter 2020
March 29, 2020

Questions

1. Investigate whether there is correlation between X3 and X2, What is Pearson correlation coefficient, How do you interpret it.
2. Does a linear relationship exist between X3 and X2 ? What about X3 and X1 ? How can you determine that definitively? Hint: residuals
3. Make a linear model with X1 as the target variable. How good is the model? What measure of goodness did you use and why? Plot the regression line on the data as a dashed line.
4. Extend you model to be multiple linear regression. Which variables/features should be included in the model and why? Are there any features that should be ignored?
5. Find the mean violent crime rate per 100,000 residents (X2), of our sample of 50 cities, and use it to estimate the mean crime rate of all cities in US, with a confidence interval of 98%. Interpret your estimation.
6. Mean overall reported crime rate per 1 million residents (X1) in cities across the US, is claimed to be 520 by the federal government. Based on the sample of 50 cities that you have in the dataset, can you say with 5% significance level that the federal government claim is incorrect. Answer this question by conducting a hypothesis test.

Answers

First, read in data. We saved xls file as csv to maintain standards of the course although R does have packages enabling read direct from Excel files.

```
## load libraries
library(car)
```

```
## Loading required package: carData
```

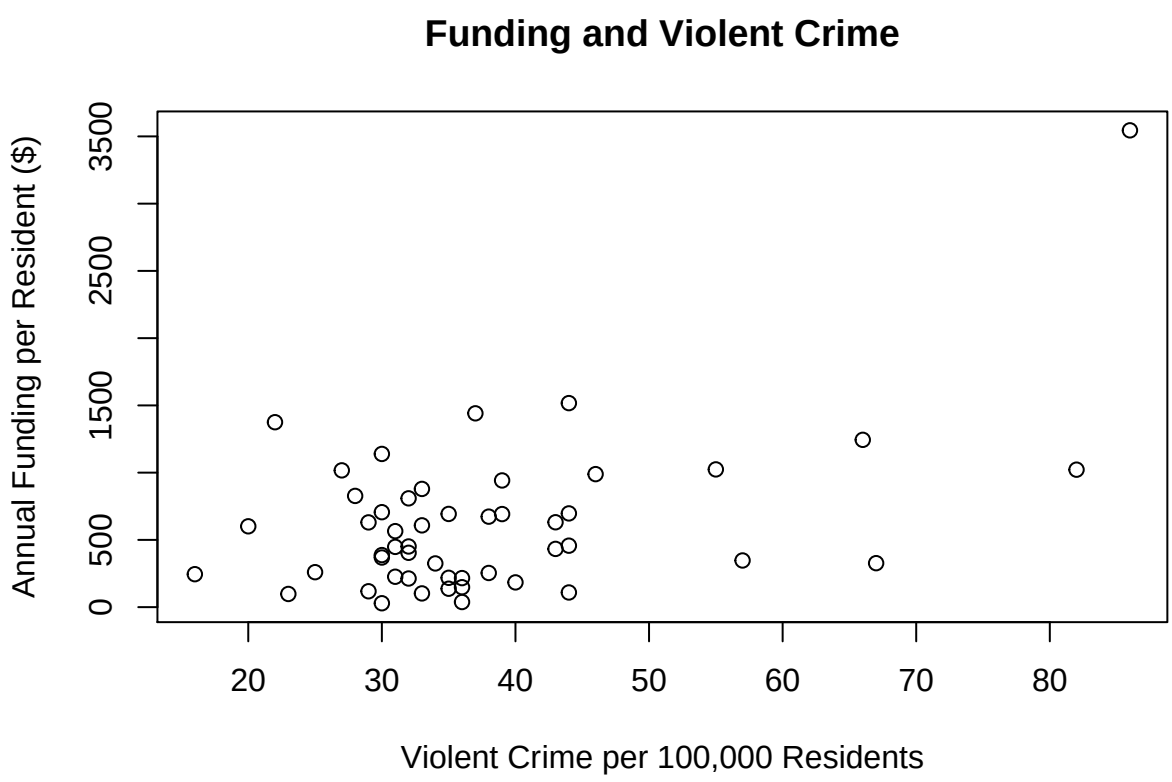
```
##import data to data frame
crimedata <- read.csv("crime_data_standard.csv")
crimedata
```

```
##      X1   X2 X3 X4 X5 X6 X7
## 1   478  184 40 74 11 31 20
## 2   494  213 32 72 11 43 18
## 3   643  347 57 70 18 16 16
## 4   341  565 31 71 11 25 19
## 5   773  327 67 72  9 29 24
## 6   603  260 25 68  8 32 15
## 7   484  325 34 68 12 24 14
## 8   546  102 33 62 13 28 11
```

```
## 9   424   38 36 69  7 25 12
## 10  548  226 31 66  9 58 15
## 11  506  137 35 60 13 21  9
## 12  819  369 30 81  4 77 36
## 13  541  109 44 66  9 37 12
## 14  491  809 32 67 11 37 16
## 15  514   29 30 65 12 35 11
## 16  371  245 16 64 10 42 14
## 17  457  118 29 64 12 21 10
## 18  437  148 36 62  7 81 27
## 19  570  387 30 59 15 31 16
## 20  432   98 23 56 15 50 15
## 21  619  608 33 46 22 24  8
## 22  357  218 35 54 14 27 13
## 23  623  254 38 54 20 22 11
## 24  547  697 44 45 26 18  8
## 25  792  827 28 57 12 23 11
## 26  799  693 35 57  9 60 18
## 27  439  448 31 61 19 14 12
## 28  867  942 39 52 17 31 10
## 29  912 1017 27 44 21 24  9
## 30  462  216 36 43 18 23  8
## 31  859  673 38 48 19 22 10
## 32  805  989 46 57 14 25 12
## 33  652  630 29 47 19 25  9
## 34  776  404 32 50 19 21  9
## 35  919  692 39 48 16 32 11
## 36  732 1517 44 49 13 31 14
## 37  657  879 33 72 13 13 22
## 38 1419  631 43 59 14 21 13
## 39  989 1375 22 49  9 46 13
## 40  821 1139 30 54 13 27 12
## 41 1740 3545 86 62 22 18 15
## 42  815  706 30 47 17 39 11
## 43  760  451 32 45 34 15 10
## 44  936  433 43 48 26 23 12
## 45  863  601 20 69 23  7 12
## 46  783 1024 55 42 23 23 11
## 47  715  457 44 49 18 30 12
## 48 1504 1441 37 57 15 35 13
## 49 1324 1022 82 72 22 15 16
## 50  940 1244 66 67 26 18 16
```

Question 1)
First we will plot the two variables to check visibly for correlation pattern. We will be looking at annual police funding data compared to violent crime rate.

```
q1plot <- plot(crimedata$X3, crimedata$X2,
               main = "Funding and Violent Crime",
               xlab = "Violent Crime per 100,000 Residents",
               ylab = "Annual Funding per Resident ($)")
```




```
q1plot
```

```
## NULL
```

Intuitively, we would expect a correlation between spending and crime rate. However, no obvious visual pattern emerges, however, we can test correlation coefficients.

```
## cor.test default method appears to be pearson but we specify for clarity
cor.test(crimedata$X2, crimedata$X3, method = "pearson")
```

```
##
##  Pearson's product-moment correlation
##
## data:  crimedata$X2 and crimedata$X3
## t = 4.1006, df = 48, p-value = 0.0001583
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.2691501 0.6898804
## sample estimates:
##       cor
## 0.5093395
```

We can infer a strong correlation here by evaluating p-value against α. A p-value lower than α implies a strong correlation and in this case p-value = 0.0001583 which is less than $\alpha = 0.05$. Having observed this,

we do not have a particularly strong coefficient. Coefficients closer to either 1 or -1 in absolute terms imply a very strong correlation. An absolute value approaching zero implies no correlation. The results here imply only a mild correlation between annual funding per resident and violent crime per 100,000 residents.

Question 2)
First, let's have a look at the X3 and X2 model..

```
model1 <- lm(crimedata$X3 ~ crimedata$X2)
summary(model1)
```

```
##
## Call:
## lm(formula = crimedata$X3 ~ crimedata$X2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -25.070  -7.521  -1.393   4.965  39.261
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  30.200033   2.507383  12.044 4.08e-16 ***
## crimedata$X2  0.012269   0.002992   4.101 0.000158 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12.02 on 48 degrees of freedom
## Multiple R-squared:  0.2594, Adjusted R-squared:  0.244
## F-statistic: 16.81 on 1 and 48 DF,  p-value: 0.0001583
```

In this case, r-squared tells us that only ~ 26% of variation is explained by the linear relationship. Not strong.

For the case of X3 vs X1, let's examine:

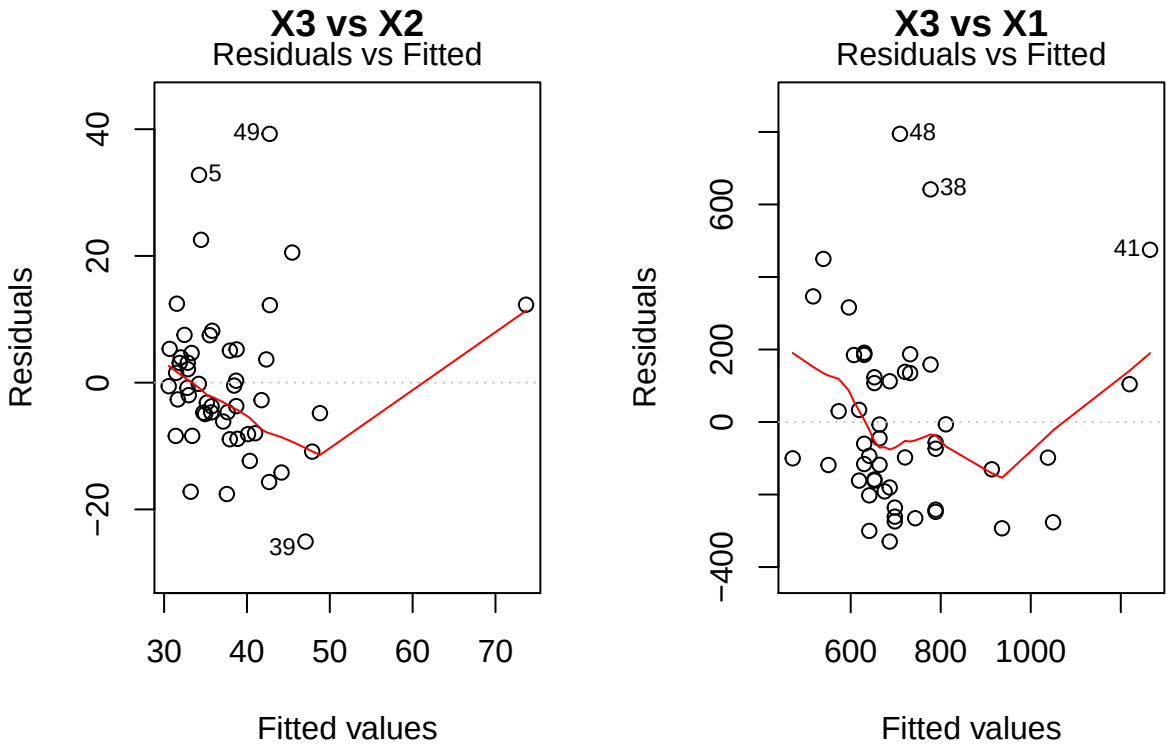
```
model2 <- lm(crimedata$X1 ~ crimedata$X3)
summary(model2)
```

```
##
## Call:
## lm(formula = crimedata$X1 ~ crimedata$X3)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -329.66 -175.91  -66.84  137.48  794.66
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   289.749    104.304   2.778  0.00778 **
## crimedata$X3   11.340      2.597   4.367  6.7e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 251.2 on 48 degrees of freedom
## Multiple R-squared:  0.2843, Adjusted R-squared:  0.2694
## F-statistic: 19.07 on 1 and 48 DF,  p-value: 6.699e-05
```

The residuals in both cases demonstrate the amount of data not explained by the linear model. Both sets of residuals are substantial but X1 vs X3 is slightly more linear in relationship when compared to X2 vs X3.

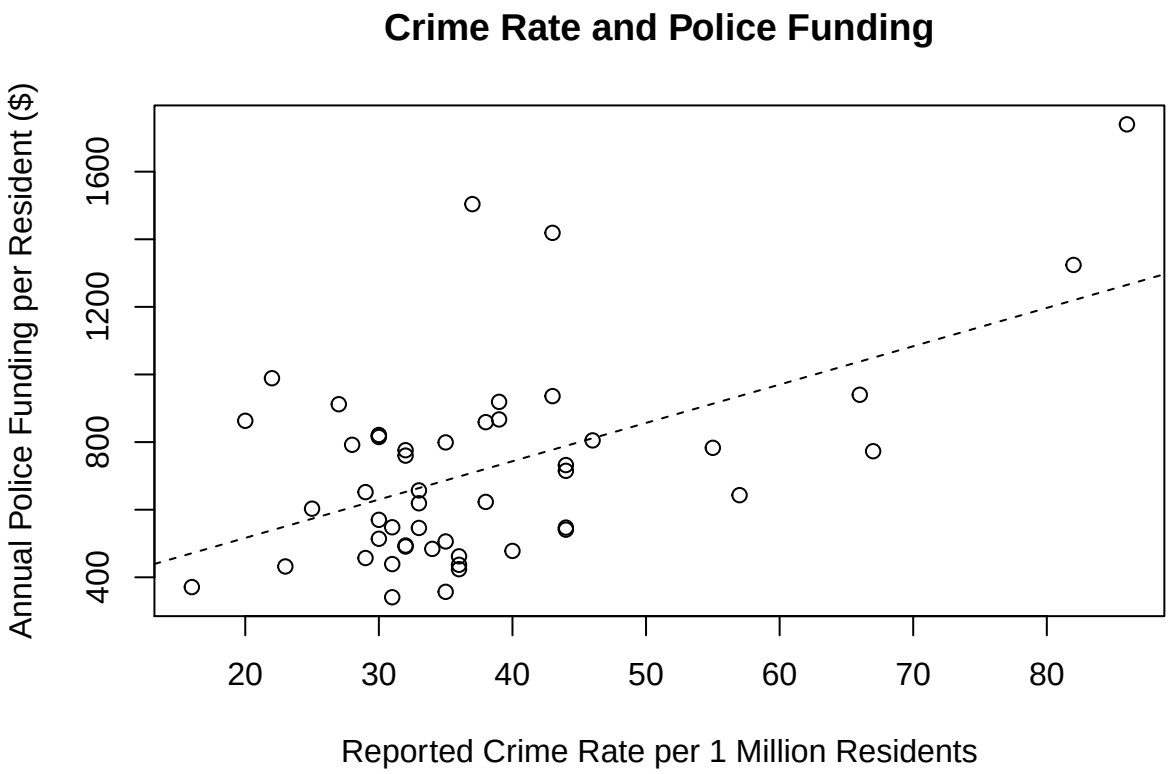
We can plot these for illustration purposes. These plots illustrate residuals vs line of fit.

```
par(mfrow = c(1, 2))
plot1 <- plot(model1, which = 1)
title("X3 vs X2", line = 1.2)
plot2 <- plot(model2, which = 1)
title("X3 vs X1", line = 1.2)
```



Question 3)
Model with X1 vs X3.

```
model3 <- lm(X1 ~ X3, data = crimedata)
## plot using model as line of fit
plot3 <- plot(crimedata$X3, crimedata$X1,
              main = "Crime Rate and Police Funding",
              xlab = "Reported Crime Rate per 1 Million Residents",
              ylab = "Annual Police Funding per Resident ($)")
abline(model3, lty =2)
```



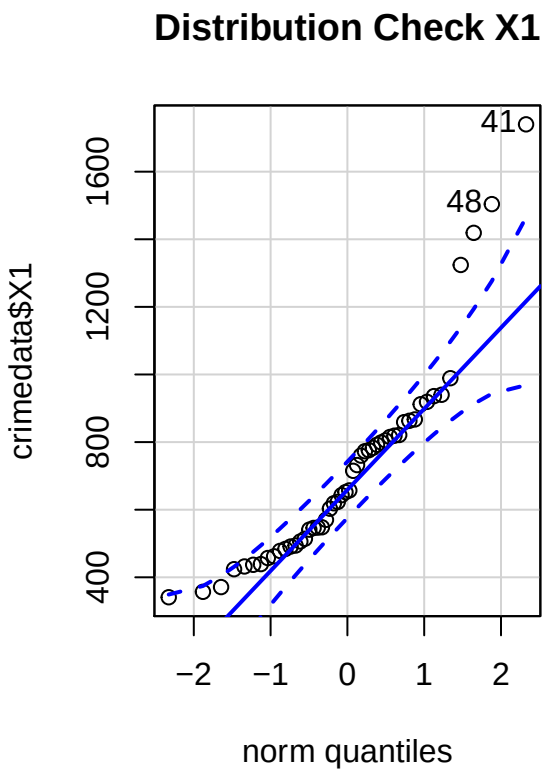
In Question 2, we actually built this model and reviewed the plot of residuals vs fit. We will not reproduce here for space but will repeat our observations. This particular linear model does not show great fit. We assessed R-Squared for goodness of fit and find it underwhelming at 0.2843. This means that only ~28% of variation is described by the linear relationship.
We took R-Squared as a measure of goodness of fit since we are dealing with a single linear model.

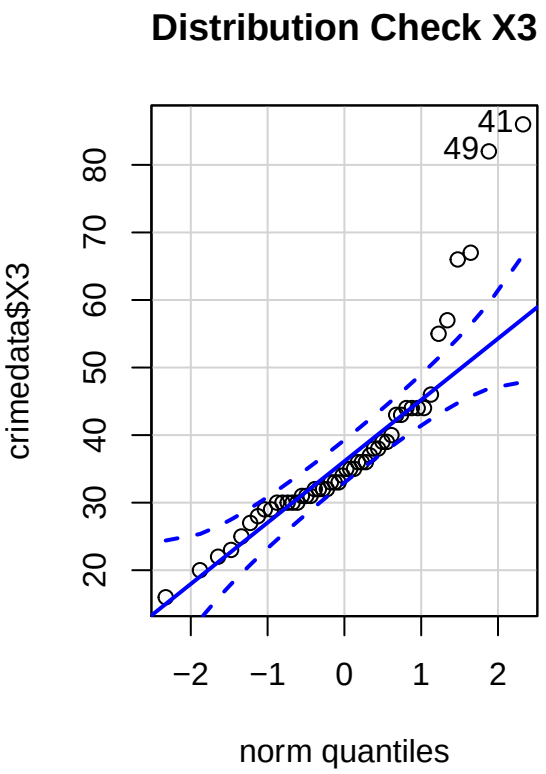
We will take a moment here to evaluate to the distribution of both variables (X1 and X3)

```
par(mfrow = c(1,2))
qqPlot(crimedata$X1,
       main = "Distribution Check X1")
```

```
## [1] 41 48
```

```
qqPlot(crimedata$X3,
       main = "Distribution Check X3")
```






```
## [1] 41 49
```

We can see that neither variable is particularly normal in its distribution which may cause problems in future calculations.

Question 4)
We are going to extend our model to include one variable, X5. We may need to add more later but intuitively, we believe that X5 may show the greatest impact. X5 shows the percent of 16 to 19 year old residents who are neither in highschool nor graduates of highschool. All other available variables basically extend education criteria beyond this. We expect X5 to show an impact on the model even if additional variables can refine it further. Hence...

```
model4 <- lm(X1 ~ X3 + X5, data = crimedata)
summary(model4)
```

```
##
## Call:
## lm(formula = X1 ~ X3 + X5, data = crimedata)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -324.2 -166.1  -72.5  119.2  797.2
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept)  202.915    121.521   1.670 0.101609
## X3            10.193      2.709   3.762 0.000467 ***
## X5             8.453      6.216   1.360 0.180350
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 249.1 on 47 degrees of freedom
## Multiple R-squared:  0.3114, Adjusted R-squared:  0.2821
## F-statistic: 10.63 on 2 and 47 DF,  p-value: 0.0001557
```

Since this model is now multi-faceted, we evaluated Adjusted R-Squared. We see little improvement by adding X5 variable, which surprises us in terms of what we expected intuitively.

By evaluating the F-statistic and its p-value, we can see that there is an inclination that adding X5 has improved the fit of the model.

Let's try expanding this. Rather than looking at young people not in highschool or graduated from high-school, we will look at educational attainment as opposed to lack thereof. Perhaps X4 will help improve the value of this model. Adults 25 years old and above who have completed four years of highschool which we interpret as the basis for all other variables related to education attainment.

```
model5 <- lm(X1 ~ X3 + X4, data = crimedata)
summary(model5)
```

```
##
## Call:
## lm(formula = X1 ~ X3 + X4, data = crimedata)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -356.90 -162.92  -60.86  100.69  784.30
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  621.426    222.685   2.791  0.00758 **
## X3            11.858      2.568   4.618 3.02e-05 ***
## X4            -5.973      3.561  -1.677  0.10013
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 246.6 on 47 degrees of freedom
## Multiple R-squared:  0.3247, Adjusted R-squared:  0.296
## F-statistic:  11.3 on 2 and 47 DF,  p-value: 9.838e-05
```

A marked improvement but not a particularly strong relationship. However, the addition of no other variable produces a better R-Squared score except one, X2. X2 is the reported violent crime rate, which we completely expect to be related to overall crime rates. We view X2 as a function of X1 rather than a predictor. The signifcance of the relationship is obvious intuitively but produced below for illustration. We see that 56% of variation is explained by this relationship.

The improvement on the p-value of the F-statistic is also substantial (0.00009838 vs previous attempt at 0.0001557)

```r
model6 <- lm(X1 ~ X2, data = crimedata)
summary(model6)
```

```
##
## Call:
## lm(formula = X1 ~ X2, data = crimedata)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -357.12  -98.58  -22.17   56.87  695.30
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 479.14487   40.52693  11.823 7.99e-16 ***
## X2            0.38757    0.04836   8.014 2.10e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 194.2 on 48 degrees of freedom
## Multiple R-squared:  0.5723, Adjusted R-squared:  0.5634
## F-statistic: 64.23 on 1 and 48 DF,  p-value: 2.096e-10
```

Therefore, we choose X1~X3 + X4 as the best possible fit for predicting future values of X1.

Question 5)
Use mean crime rate per 100,000 residents to forecast mean violent crime rate for all cities with 98% confidence interval.

```r
crimemean <- mean(crimedata$X2)
## we will also require standard deviation of our sample
crimesd <- sd(crimedata$X2)

xbar = crimemean
z_alpha <- qnorm(0.98)
n <- 50
s <- crimesd

lowerbound <- crimemean - z_alpha * crimesd/sqrt(n)
upperbound <- crimemean + z_alpha * crimesd/sqrt(n)
meanrange <- (upperbound - lowerbound) /2
meanrange
```

```
## [1] 166.6391
```

```r
crimemean
```

```
## [1] 616.18
```

Which tells us the mean estimate for all cities is 616.18 +/- 166.64 violent crimes per 100,000 residents, 98 percent of the time.

Question 6)
Prove that government claim of mean crime rate per 1 million residents of 520 is incorrect.

H_0: $\mu = 520$ H_a μ != 520

```r
totalcrime <- mean(crimedata$X1)
sdtotalcrime <- sd(crimedata$X1)
x = 520
m = totalcrime
n = 50
s = sdtotalcrime

z_stat <- (x - m) / (s/sqrt(n))
z_stat
```

```
## [1] -4.762178
```

```r
alpha03 <- 0.05 ## distribution of 95% confidence interval
qnorm(alpha03)
```

```
## [1] -1.644854
```

```r
qnorm(1-alpha03)
```

```
## [1] 1.644854
```

```r
pnorm(z_stat)
```

```
## [1] 9.575755e-07
```

The calculated z-score test statistic and the comparison of p-value vs α of 0.05 tells us that we can reject the null hypothesis $\mu = 520$ within a required 95% confidence interval. That is to say, there is not sufficient statistical evidence to confirm the government's claim. We can accept the alternate hypothesis μ != 520