Project 1

Lillian Quynni ([redacted]), Marco Oviedo ([redacted]), Colin Lauffer ([redacted])

Introduction

.....

In this project, we were given a data set with 440 points and 17 different variables. All the data and graphs were calculated and made in R. For the first part of the project, we created three different first-order regression models and graphs. For all three regression models, the number of active physicians remained as Y. We then compared the number of active physicians to different X variables such as total population, number of hospital beds, and total personal income. For the second part we measured the linear association of the three previous graphs using the R^2 value. For the third part we created four regression models based off of the per capita income and the percent of the population that has a bachelor's degree. These models were separated based off of what region the data points were gathered in. We then calculated the confidence interval for 1and created ANOVA tables for each region. Lastly for the fourth part we created residual plots against x and normality plots from the three regression models that we created in first part of the project.

.....

Part I: Fitting regression models

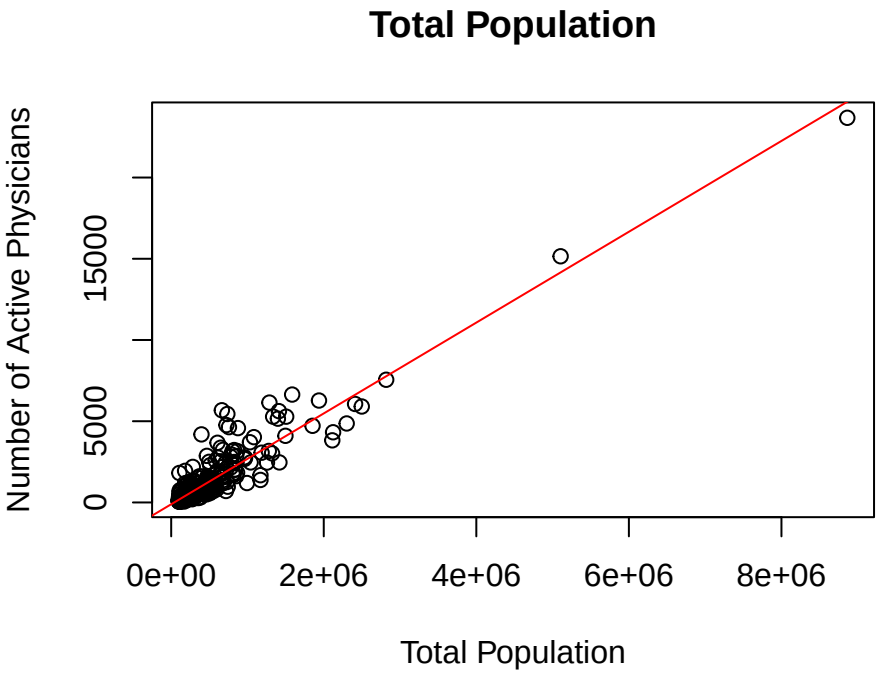
(a) Regress the number of active physicians in turn on each of the three predictor variables. State the estimated regression functions.

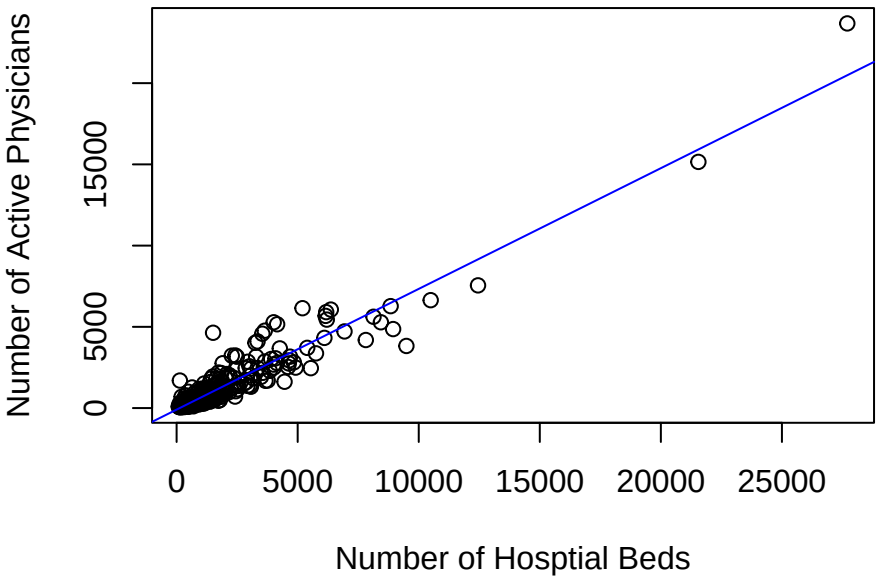
Let Y be breakage, X_1 be the total population The estimated regression function is $\hat{Y} = -110.6347772 + 0.0027954X_1$.

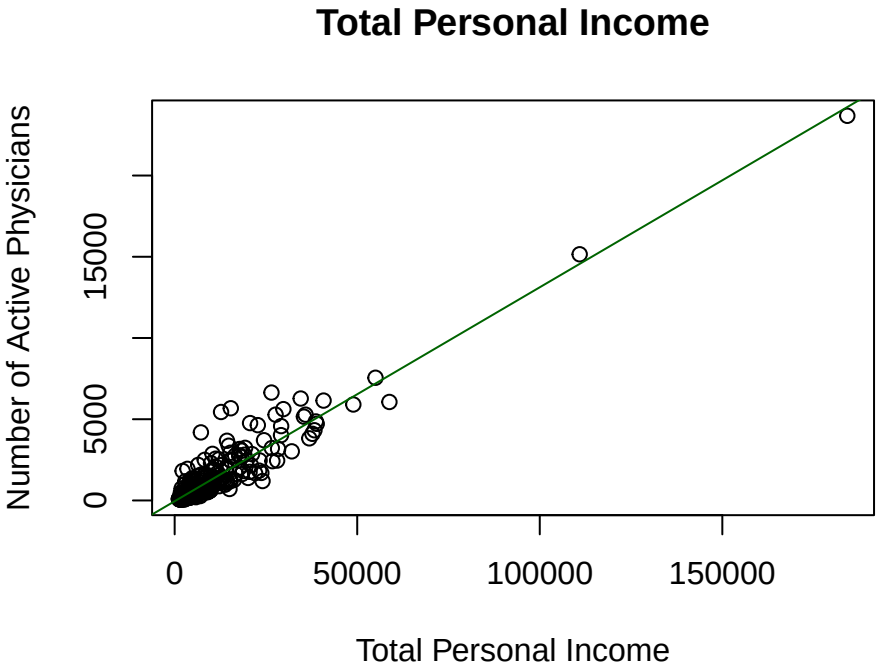
Let Y be breakage, X_1 be the number of hospital beds. The estimated regression function is $\hat{Y} = -95.9321847 + 0.7431164X_1$.

Let Y be breakage, X_1 be the total personal income of the population. The estimated regression function is $\hat{Y} = -48.3948489 + 0.1317012X_1$.

(b) Plot the three estimated regression functions and data on separate graphs. Does a linear regression relation appear to provide a good fit for each of the three predictor variables?


Total Population
Number of Active Physicians
0
5000
15000
0e+00
2e+06
4e+06
6e+06
8e+06
Total Population


Number of Hospital Beds
Number of Active Physicians
0
5000
15000
0
5000
10000
15000
20000
25000
Number of Hosptial Beds



Yes, a linear regression relation appears to be a good fit for each of the three predictor variables. This is because most of the residuals are close to the plotted line.

c)

The MSE for the total population is 3.722035×10^5.

The MSE for the number of hospital beds is 3.1019188×10^5.

The MSE for the total personal income is 3.2453939×10^5.

The total number of hospital beds results in the ssmallest variability around the fitted regression line since it has the smallest MSE.

Part II: Measuring linear associations

Y is the number of Active Physicians and X is the Total Population: $R^2 = 0.8840674$

Y is the number of Active Physicians, X is the number of Hospital Beds: $R^2 = 0.9033826$

Y is the number of Active Physicians, X is the Total Personal Income of the Population: $R^2 = 0.8989137$

The number of active physicians and the number of total hospital beds accounts for the largest reduction in variability since it has the closest R2 value to 1.

Part III: Inference about regression parameters

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
X	1	1450517671	1450517671	197.7527	0
Residuals	101	740835765	7335008		

Region 1: $\hat{B}_1 = 522.1588326$

$t_{103-2}(1 - \frac{.1}{2}) = 1.6600806$

s.e$(\hat{B_1}) = 37.1314065$

Confidence Interval for $B_1 = [460.52, 583.80]$

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
X	1	338907694	338907694	76.82646	0
Residuals	106	467602149	4411341		

Region 2: $\hat{B_1} = 238.6693985$

$t_{103-2}(1 - \frac{.1}{2}) = 1.659356$

s.e$(\hat{B_1}) = 27.229605$

Confidence Interval for $B_1 = [193.49, 283.85]$

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
X	1	1109873245	1109873245	148.491	0
Residuals	150	1121152411	7474349		

Region 3: $\hat{B_1} = 330.6117256$

$t_{103-2}(1 - \frac{.1}{2}) = 1.6550755$

s.e$(\hat{B_1}) = 27.1311535$

Confidence Interval for $B_1 = [285.71, 375.52]$

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
X	1	773745787	773745787	94.19477	0
Residuals	75	616073841	8214318		

Region 4: $\hat{B_1} = 440.3157072$

$t_{103-2}(1 - \frac{.1}{2}) = 1.6654254$

s.e$(\hat{B_1}) = 45.3681199$

Confidence Interval for $B_1 = [364.76, 515.87]$

We found that the regression lines appear to have different slopes. The confidence intervals vary between all 4 regions

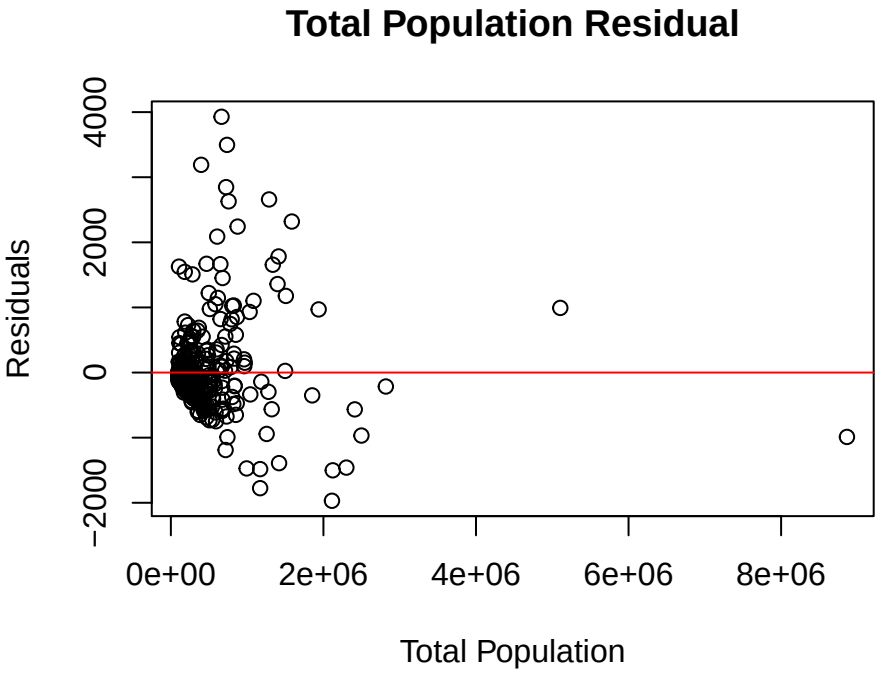
Region 1: our critical value is 2.76 and our f statistic is 197.75. Therefore, we reject our null hypothesis. There is a linear relationship between

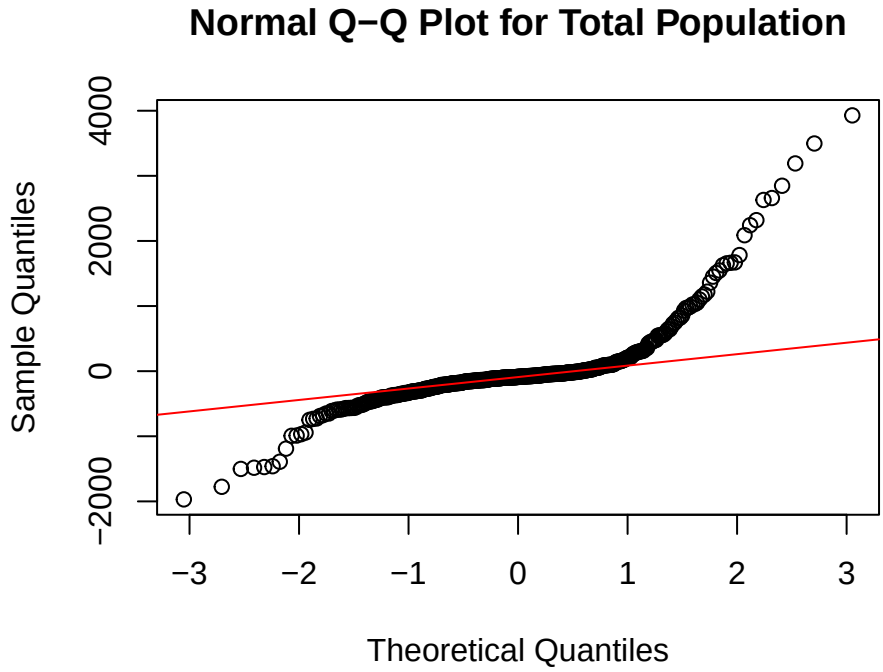
Region 2: our critical value is 2.75 and our f statistic is 76.83. Therefore, we reject our null hypothesis.

Region 3: our critical value is 2.74 and our f statistic is 148.49. Therefore, we reject our null hypothesis. There is a linear relationship between per capita income and individuals with a bachelor's degree.

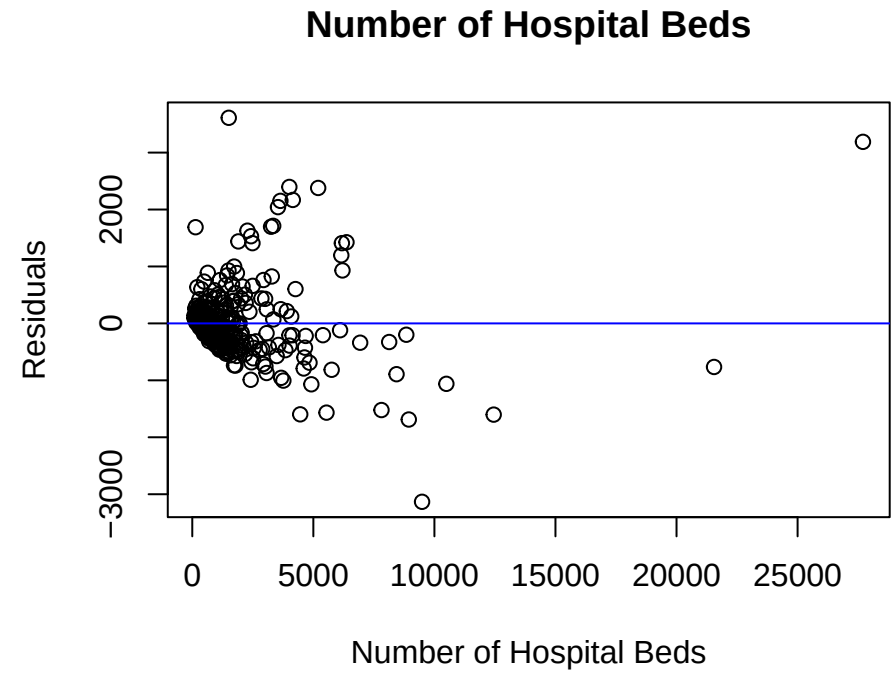
Region 4: Our critical value is 2.773642 and our F statistic is 94.19477. We therefore reject our null hypothesis.

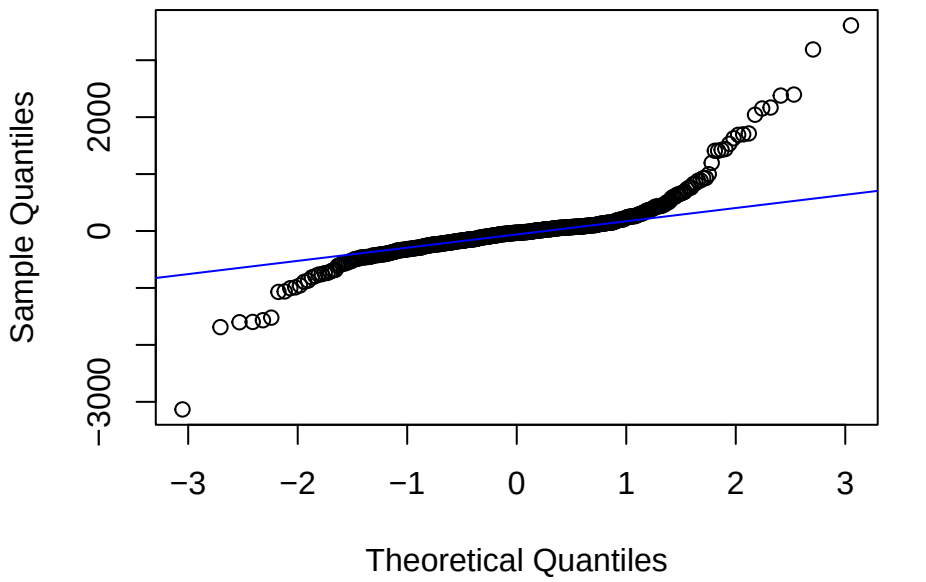
Part IV: Regression diagnostics



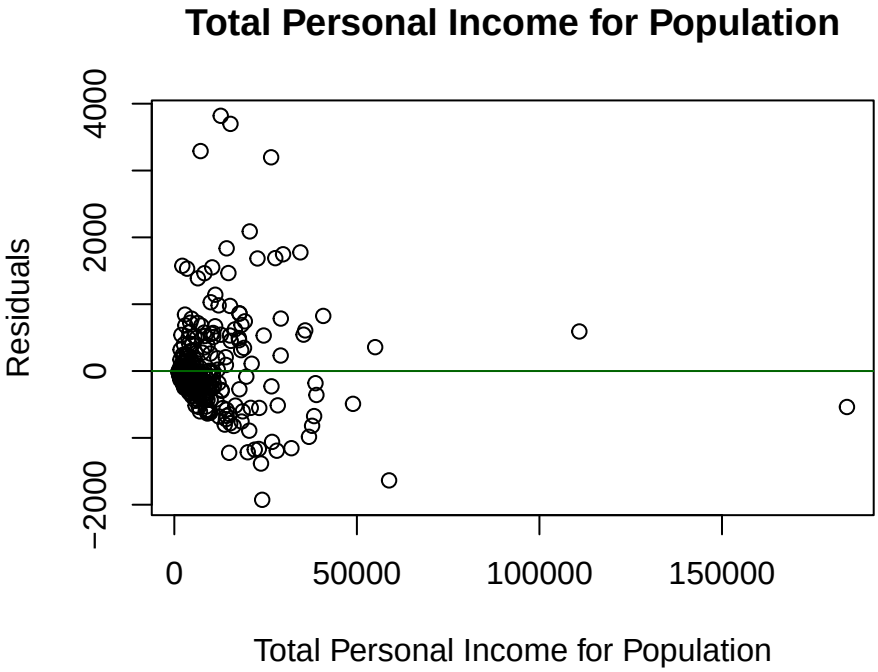


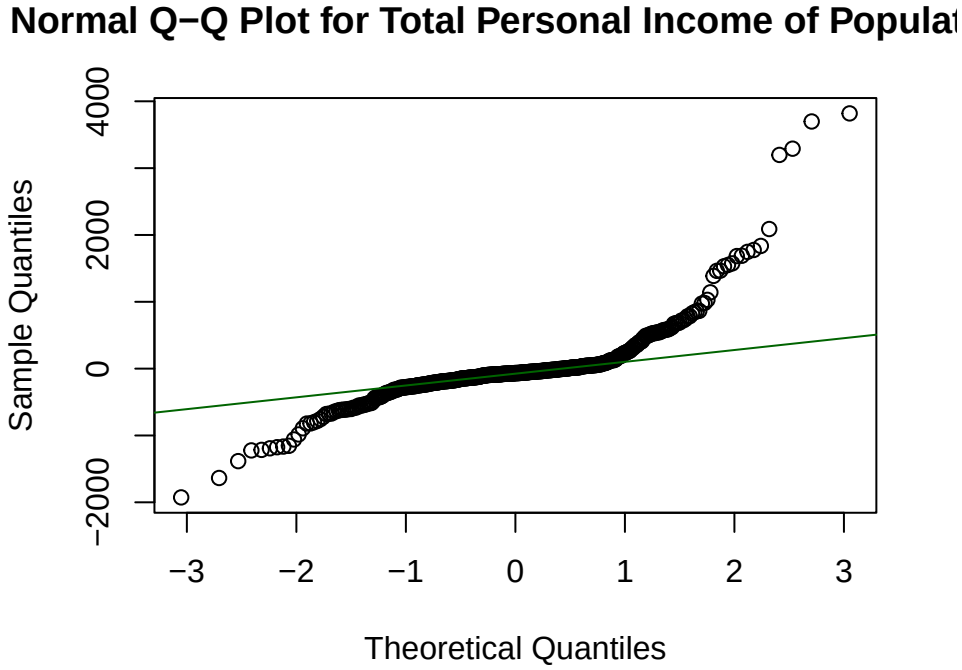
The plot seems to have no pattern, however there appears to be some outliers. Most of the data points appear to be near the line 0, and is split evenly above and below it, which is a good sign for a normal distribution. While the normal probability plot seems to be symmetrical with heavy tails.This means we can expect more outliers on both high and low ends of our data.





The plot seems to have no pattern, however there appears to be less outliers than the total population graph and our data has less range. Most of the data points appear to be near the line 0, and is split evenly above and below it, which is a good sign for a normal distribution While the normal probability plot seems to be symmetrical with heavy tails.This means we can expect more outliers on both high and low ends of our data.





The residual plot seems to have no pattern, however there appears to be less outliers but they appear to have a greater magnitude, otherwise the plots are bundled closer together. Most of the data points appear to be near the line 0, and is split evenly above and below, which is a good sign for a normal distribution. While the normal probability plot seems to be symmetrical with heavy tails. This means we can expect more outliers on both high and low ends of our data.

Each case should use linear model 2.1 as we are dealing with residuals and error we can't expect.

Part V: Discussion

We found that the wide range of the observational data made it difficult to get a clear look at the more clumped together data when viewing the data on a graph. If we had excluded the outliers we would have been able to see the more related data more closely.

As more future observable data is added, we could get a more clear look at the data. This could possibly allow us to bridge the large range difference in the data.

Appendix

```
knitr::opts_chunk$set(
    error = FALSE,
    message = FALSE,
    warning = FALSE,
    echo = FALSE, # hide all R codes!!
    fig.width=5, fig.height=4,#set figure size
    fig.align='center',#center plot
    options(knitr.kable.NA = ''), #do not print NA in knitr table
    tidy = FALSE #add line breaks in R codes
)
CDI <- read.table("CDI.txt")
library(knitr)
X1 <- CDI$V5
Y1 <- CDI$V8
fit1 <- lm(Y1 ~ X1)
coef = fit1$coefficients
b0hat = coef[1]
b1hat = coef[2]
X2 <- CDI$V9
Y2 <- CDI$V8
fit2 <- lm(Y2 ~ X2)
coef2 = fit2$coefficients
b0hat2 = coef2[1]
b1hat2 = coef2[2]
X3 <- CDI$V16
Y3 <- CDI$V8
fit3 <- lm(Y3 ~ X3)
coef3 = fit3$coefficients
b0hat3 = coef3[1]
b1hat3 = coef3[2]
plot(X1,Y1, ylab = "Number of Active Physicians", xlab = "Total Population", main = "Total Population")
abline(lm(Y1~X1), col = "red")
plot(X2,Y2, ylab = "Number of Active Physicians", xlab = "Number of Hosptial Beds", main = "Number of H
abline(lm(Y2~X2), col = "Blue")
plot(X3,Y3, ylab = "Number of Active Physicians", xlab = "Total Personal Income", main = "Total Personal
abline(lm(Y3~X3), col = "dark green")
MSE1 = summary(fit1)$sigma^2
MSE2 = summary(fit2)$sigma^2
MSE3 = summary(fit3)$sigma^2
r1 = summary(fit1)$r.squared

r2 = summary(fit2)$r.squared
```

```
r3 = summary(fit3)$r.squared
CDI<-CDI #region 1
W<-CDI$V17
df2<-subset(CDI, W == 1)
X<-df2$V12
Y<-df2$V15
fit.y=b0hat[1]+b1hat[1]*X
fit<-lm(Y~X)
coef=fit$coefficients
b0hat=coef[1]
b1hat=coef[2]
MSE = summary(fit)$sigma^2
se.b1hat = sqrt(MSE/sum((X-mean(X))^2))
ts = qt(.95,101, lower.tail=TRUE) #t critical value
kable(anova(fit))
CDI<-CDI #region 2
W<-CDI$V17
df2<-subset(CDI, W == 2)

X<-df2$V12
Y<-df2$V15
fit.y=b0hat[1]+b1hat[1]*X
fit<-lm(Y~X)
coef=fit$coefficients
b0hat=coef[1]
b1hat=coef[2]
MSE = summary(fit)$sigma^2

se.b1hat = sqrt(MSE/sum((X-mean(X))^2))
ts = qt(.95,106, lower.tail=TRUE) #t critical value
kable(anova(fit))
CDI<-CDI #region 3
W<-CDI$V17
df2<-subset(CDI, W == 3)
X<-df2$V12
Y<-df2$V15
fit.y=b0hat[1]+b1hat[1]*X
fit<-lm(Y~X)
coef=fit$coefficients
b0hat=coef[1]
b1hat=coef[2]
MSE = summary(fit)$sigma^2

se.b1hat = sqrt(MSE/sum((X-mean(X))^2))
ts = qt(.95,150, lower.tail=TRUE) #t critical value
kable(anova(fit))
CDI<-CDI #region 4
W<-CDI$V17
df2<-subset(CDI, W == 4)

X<-df2$V12
Y<-df2$V15
fit.y=b0hat[1]+b1hat[1]*X
```

```r
fit<-lm(Y~X)
coef=fit$coefficients
b0hat=coef[1]
b1hat=coef[2]

MSE = summary(fit)$sigma^2

se.b1hat = sqrt(MSE/sum((X-mean(X))^2))
ts = qt(.95,75, lower.tail=TRUE) #t critical value
kable(anova(fit))
residuals = fit1$residuals
plot(x = X1, y = residuals, ylab = "Residuals", xlab = "Total Population", main = "Total Population Resi
abline(h=0, col = "red")
qqnorm(residuals, main = "Normal Q-Q Plot for Total Population")
qqline(residuals, col="red")
residuals2 = fit2$residuals
plot(x = X2, y = residuals2, main = "Number of Hospital Beds", ylab = "Residuals", xlab = "Number of Hos
abline(h = 0, col = "blue")
qqnorm(residuals2, main = "Normal Q-Q Plot for Number of Hospital Beds")
qqline(residuals2, col="blue")
residuals3 = fit3$residuals
plot(x=X3, y = residuals3, ylab = "Residuals", xlab = "Total Personal Income for Population", main = "To
abline(h = 0, col = "dark green")
qqnorm(residuals3, main = "Normal Q-Q Plot for Total Personal Income of Population")
qqline(residuals3, col="dark green")
```