STA Project 1

Ishita Dutta, Devika Sunil Kumar, Fernanda Serna Godoy

4/28/2021

Intro

This project focuses on taking in the county demographic information, also known as CDI, for 440 of the most populous counties in the United States between the years 1990 and 1992. There are a total of 17 variables provided from this data from which our main focus for this project is centered around total population, number of hospital beds, per capita income, total personal income, number of active physicians, percent bachelor's degree, and geographic region.

We will analyze the data in two major ways. The first way we will analyze the data is through regressing the number of active physicians against the total population, number of beds, and personal income indicators to see which one best fits the data overall. The second way is to split our data into the four specific regions each county has been classified into, then taking the regression of the percent bachelor's degrees in that county against the per capita income of the physicians to see which region has the best fit for this indicator.

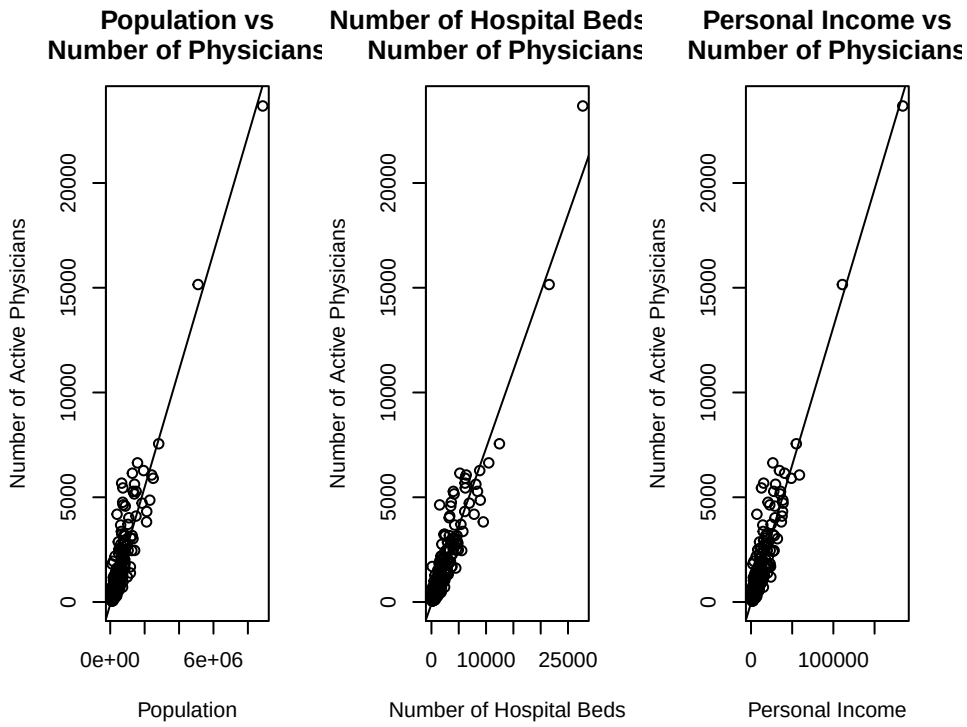
For both indicators, we will further analyze the data by taking measures such as the mean square estimates, residuals, confidence intervals, etc. From this, we will conclude which predictor(s) is best suited for our data based on the three indicators in our first way of analysis, as well as which region(s) is really the best fit when we analyze our data through the second method.

Problem 1 - Fitting Regression Models

1.43 a)

```
## Population(In Hundred Thousands):

## Y =  -110.6348  +  0.002795425 x

## Number of Hospital Beds(In Thousands):

## Y =  -95.93218  +  0.7431164 x

## Personal Income(In Ten Thousands):

## Y =  -48.39485  +  0.1317012 x
```

1.43 b)



A linear regression relation appears to provide a good fit for each of the predictor variables (Total population, Personal Income, and Number of Hospital beds).

1.43 c)

calculate MSE

```
## MSE for Population Measured in Hundred Thousands:  372203.5

## MSE for Number of Hospital Beds Measured in Thousands:  310191.9

## MSE for Personal Income Measured in Ten Thousands:  324539.4
```

By comparing the three MSE obtained above, it is observed that the smallest value corresponds to the number of hospital beds and thus, this predictor variable leads to the smallest variability around the fitted regression line.

1.44 a)

```
## Y = 9223.82 + 522.16x

## Y = 13581.41 + 238.67x

## Y = 10529.79 + 330.61x

## Y = 8615.05 + 440.32x
```

1.44 b)

As it is observed in part a), the regression functions for all of the regions (1-4) have both a positive slope and intercept but they are very different in terms of the values for each one. The variable with the greatest magnitude of slope is region 1 and the one with the smallest magnitude corresponds to region 2.

1.44 c)

```
## MSE for Region 1:  7335008

## MSE for Region 2:  4411341

## MSE for Region 3:  7474349

## MSE for Region 4:  8214318
```

The variability around the fitted regression line, estimated by the MSE for each region, is approximately the same for regions 1 and 3. There's a considerable difference between region 2's variability (which is significantly smaller than the other regions) and region 4 (which is the largest among all regions).

2.62

```
## R ^ 2 =  0.8840674

## R ^ 2 =  0.9033826

## R ^ 2 =  0.8989137
```

The predictor variable that accounts for the largest reduction in variability in the number of active physicians corresponds to the number of hospital beds. Approximately 90% of the variation in Number of physicians (Y) is explained/reduced by the use of "Number of Hospital beds" as predictor variable.

2.63

```
##          5 %  95 %
## RX1 460.5177 583.8

##          5 %    95 %
## RX2 193.4858 283.853

##          5 %     95 %
## RX3 285.7076 375.5158
```

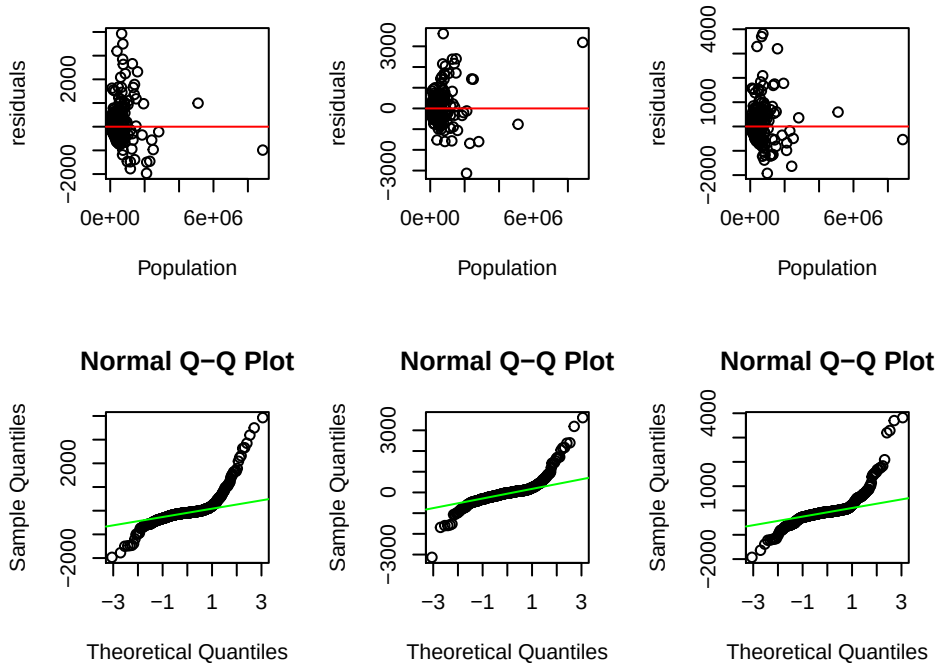
	Df	Sum Sq	Mean Sq	F value	Pr(>F)
RX2	1	338907694	338907694	76.82646	0
Residuals	106	467602149	4411341		

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
RX3	1	1109873245	1109873245	148.491	0
Residuals	150	1121152411	7474349		

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
RX4	1	773745787	773745787	94.19477	0
Residuals	75	616073841	8214318		

F-value for region 1: 197.7527 F-value for region 2: 76.82646 F-value for region 3: 148.491 F-value for region 4: 94.19477 We are 90% confident that B1 in region 1 is between 460.5177 and 583.8. We are 90% confident that the B1 in region 2 is between 193.4858 and 283.853. We are 90% confident that the B1 in region 3 is between 285.7076 and 375.5158. We are 90% confident that the B1 in region 4 is between 364.7585 and 515.8729. We can conclude that region 2 is the best fitted regression model as it carries out the smallest F-value. The regression lines for the different regions do not appear to have similar slopes but there is a significant overlap in the interval estimates for the slope of regions 1 and 4 which would require further analysis.

3.25



Conclusions: The comparison between the MSE for each of the predictor variables (Total population, Number of Hospital beds, and personal income), as well as the coefficients of correlation obtained in the previous sections demonstrate that the linear regression model would be more appropriate for the predictor variable "Number of Hospital Beds" than for the other two variables.

Part V - Discussion

As stated in the introduction, we analyzed the CDI information in two ways, first by regressing the number of active physicians(AP) against three indicators (population, number of beds, and personal income), and second by splitting the data into the four regions each county is categorized in and comparing the percent of the population with bachelor's degrees against the AP.

In the first method, we first obtained the functions for the relationship (regression line) of each indicator compared with the AP and noted that there is an increase with the number of active physicians each time there is an increase in any of the three indicators for question 1.43 in Part I. Here, we also noted that comparing against the number of hospital beds against the AP gave a smaller variation than the other indicators, suggesting that it might be the best indicator. We were further able to suggest this when we found the correlation coefficient for each of the 3 indicators when the coefficient was the largest for the comparison between hospital beds and AP during question 2.62 in Part II. This is because the closer this coefficient is to 1, the closer the points are to the regression line. We officially concluded this in question 3.25 from Part IV, since the residual plots and qqplots showed the least variety in the comparison between the number of beds against the AP.

In the second method, we first sorted the data based on which region of four total regions each county is in. We then found the regression lines comparing the percentage of people with a bachelor's degree in that county with the AP for that county. Here we noted that the highest amount of change in AP from a change in the percent of bachelor's degrees was in region 1, while the smallest amount of change for the same was in region 2. This brought us to think that region 2 might be the region that best fits this indicator relationship at the end of question 1.44 in Part I. We concluded the same in question 2.63 from Part III, when we found the confidence intervals and f-values for each of the four regions. The confidence interval gives a range of values for the true change per percent increase in bachelor's degrees, while the f-value gives us a measure of error. Region 2 has the smallest range for the interval as well as the smallest f-value, meaning that it is the best fit for the indicator between the percent of bachelor's degrees compared with the AP.

To additionally improve our analysis, we can consider looking at how combinations of variables affect the AP, since it is rarely ever the case in real life that only one variable can best indicate a certain measure as opposed to a combination of variables. We should also consider conducting other tests, such as looking at a relationship that is not linear for each comparison.

Appendix

```
library(knitr)
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")
#"ID", "County", "State", "Area", "Population", "18_to_34", "65+", "Active_Physicians", "Beds", "Seriou
X1 = CDI[,5] #Population in Hundred Thousands
X1 = X1
X2 = CDI[,9] #Number of Hospital Beds in Thousands
```

```
X2 = X2
X3 = CDI[,16] #Personal Income in Ten Thousands
X3 = X3
Y = CDI[,8]
n = length(X1)
cat("Population(In Hundred Thousands): \n")
b1hat1 = t(X1-mean(X1))%*%(Y-mean(Y))/sum((X1-mean(X1))^2)
b0hat1 = mean(Y) - b1hat1*mean(X1)
cat("Y = ", b0hat1, " + ", b1hat1, "x\n")

cat("Number of Hospital Beds(In Thousands): \n")
b1hat2 = t(X2-mean(X2))%*%(Y-mean(Y))/sum((X2-mean(X2))^2)
b0hat2 = mean(Y) - b1hat2*mean(X2)
cat("Y = ", b0hat2, " + ", b1hat2, "x\n")

cat("Personal Income(In Ten Thousands): \n")
b1hat3 = t(X3-mean(X3))%*%(Y-mean(Y))/sum((X3-mean(X3))^2)
b0hat3 = mean(Y) - b1hat3*mean(X3)
cat("Y = ", b0hat3, " + ", b1hat3, "x\n")
par(mfrow=c(1,3))
plot(X1, Y, xlab="Population", ylab="Number of Active Physicians" , main = "Population vs\nNumber of Phy
abline(lm(Y ~ X1))

plot(X2, Y, xlab="Number of Hospital Beds", ylab="Number of Active Physicians" , main = "Number of Hospi
abline(lm(Y ~ X2))

plot(X3, Y, xlab="Personal Income", ylab="Number of Active Physicians" , main = "Personal Income vs\nNum
abline(lm(Y ~ X3))
fit.y1 = b0hat1[1] + b1hat1[1]*X1
mse1 = 1/(n-2)*sum((Y - fit.y1)^2)
cat("MSE for Population Measured in Hundred Thousands: ", mse1, "\n")

fit.y2 = b0hat2[1] + b1hat2[1]*X2
mse2 = 1/(n-2)*sum((Y - fit.y2)^2)
cat("MSE for Number of Hospital Beds Measured in Thousands: ", mse2, "\n")

fit.y3 = b0hat3[1] + b1hat3[1]*X3
mse3 = 1/(n-2)*sum((Y - fit.y3)^2)
cat("MSE for Personal Income Measured in Ten Thousands: ", mse3, "\n")

Reg_1 <- subset(CDI, CDI[,17] == 1)
RX1 = Reg_1[,12]
RY1 = Reg_1[,15]
R1n = length(RX1)
fitr1 = lm(RY1 ~ RX1)
cat("Y = 9223.82 + 522.16x\n")

Reg_2 <- subset(CDI, CDI[,17] == 2)
RX2 = Reg_2[,12]
RY2 = Reg_2[,15]
R2n = length(RX2)
fitr2 = lm(RY2 ~ RX2)
cat("Y = 13581.41 + 238.67x\n")
```

```r
Reg_3 <- subset(CDI, CDI[,17] == 3)
RX3 = Reg_3[,12]
RY3 = Reg_3[,15]
R3n = length(RX3)
fitr3 = lm(RY3 ~ RX3)
cat("Y = 10529.79 + 330.61x\n")

Reg_4 <- subset(CDI, CDI[,17] == 4)
RX4 = Reg_4[,12]
RY4 = Reg_4[,15]
R4n = length(RX4)
fitr4 = lm(RY4 ~ RX4)
cat("Y = 8615.05 + 440.32x\n")

R1b1hat1 = t(RX1-mean(RX1))%*%(RY1-mean(RY1))/sum((RX1-mean(RX1))^2)
R1b0hat1 = mean(RY1) - R1b1hat1*mean(RX1)
R1fit = R1b0hat1[1] + R1b1hat1[1]*RX1
R1mse = 1/(R1n-2)*sum((RY1 - R1fit)^2)
cat("MSE for Region 1: ", R1mse, "\n")

R2b1hat1 = t(RX2-mean(RX2))%*%(RY2-mean(RY2))/sum((RX2-mean(RX2))^2)
R2b0hat1 = mean(RY2) - R2b1hat1*mean(RX2)
R2fit = R2b0hat1[1] + R2b1hat1[1]*RX2
R2mse = 1/(R2n-2)*sum((RY2 - R2fit)^2)
cat("MSE for Region 2: ", R2mse, "\n")

R3b1hat1 = t(RX3-mean(RX3))%*%(RY3-mean(RY3))/sum((RX3-mean(RX3))^2)
R3b0hat1 = mean(RY3) - R3b1hat1*mean(RX3)
R3fit = R3b0hat1[1] + R3b1hat1[1]*RX3
R3mse = 1/(R3n-2)*sum((RY3 - R3fit)^2)
cat("MSE for Region 3: ", R3mse, "\n")

R4b1hat1 = t(RX4-mean(RX4))%*%(RY4-mean(RY4))/sum((RX4-mean(RX4))^2)
R4b0hat1 = mean(RY4) - R4b1hat1*mean(RX4)
R4fit = R4b0hat1[1] + R4b1hat1[1]*RX4
R4mse = 1/(R4n-2)*sum((RY4 - R4fit)^2)
cat("MSE for Region 4: ", R4mse, "\n")
fit1 = lm(Y~X1)
y_hat1 = fit1$fitted.values
SSR = sum((y_hat1 - mean(Y))^2)
SSTO = sum (((Y - mean(Y)))^2)
R2a = SSR / SSTO
cat("R ^ 2 = ", R2a, "\n")

fit2 = lm(Y~X2)
y_hat2 = fit2$fitted.values
SSR = sum((y_hat2 - mean(Y))^2)
SSTO = sum (((Y - mean(Y)))^2)
R2b = SSR / SSTO
cat("R ^ 2 = ", R2b, "\n")

fit3 = lm(Y~X3)
y_hat3 = fit3$fitted.values
```

```r
SSR = sum((y_hat3 - mean(Y))^2)
SSTO = sum (((Y - mean(Y)))^2)
R2c = SSR / SSTO
cat("R ^ 2 = ", R2c, "\n")
library (knitr)
confint(fitr1, parm = "RX1",  level = 0.90)
confint(fitr2, parm = "RX2",  level = 0.90)
confint(fitr3, parm = "RX3",  level = 0.90)

kable(anova(fitr2))
kable(anova(fitr3))
kable(anova(fitr4))
par(mfcol = c(2, 3), no.readonly = TRUE)
R2a = fit1$residuals
plot(X1, y = R2a, xlab='Population', ylab='residuals')
abline(h=0, col='red')
qqnorm(R2a)
qqline(R2a, col='green')

R2b = fit2$residuals
plot(X1, y = R2b, xlab='Population', ylab='residuals')
abline(h=0, col='red')
qqnorm(R2b)
qqline(R2b, col='green')

R2c = fit3$residuals
plot(X1, y = R2c, xlab='Population', ylab='residuals')
abline(h=0, col='red')
qqnorm(R2c)
qqline(R2c, col='green')
```