CDI ANALYSIS
PART 1

INTRO:

With rising awareness of socioeconomic inequalities, it is imperative to examine possible causes or symptoms of inequity. This project examines aspects county hospitals and how they correlate with aspects of the socioeconomic status of their counties. The data that is used in this project is the County Demographic Information (CDI) for the 440 most populous counties in the United States. Each line of data is identified by county and state, and has information about 14 variables. In this project, we focus on four of them: the number of active physicians, the number of hospital beds, total population, and the total personal income. The main objective of this project is to find how the number of active physicians in an area is influenced by the total population, the number of active hospital beds, and the total personal income.The main tool used in this project is R. R was used to calculate statistics and create the graphs that follow. Part 1 consists of regression plots for total population, hospital beds, and variable total personal income plotted against number of active physicians, as well as per capita income plotted against the percentage of individuals in the county with at least a bachelor's degree. In Part 2, we determine which predictor has the greatest effect on the number of active physicians. Part 3 is a confidence interval and analysis of variance for the results from the regression of per capita income plotted against percentage of individuals with at least bachelor's degrees. Part 4 is testing the appropriateness of the linear regression model applied in Part 1. Part 5 is a discussion of the findings in this project.

PART 1

The Number of active physicians is denoted by Y, the variable Total population is denoted by X1, the variable Number of hospital beds is denoted by X2, and the variable Total personal income is denoted by X3

1.43

Part a) - Regression Functions for Number of active physicians (Y):

The Estimated Regression function for Y and Total population (X1):

$$\hat{Y} = -110.6348 + 0.002795425X1$$

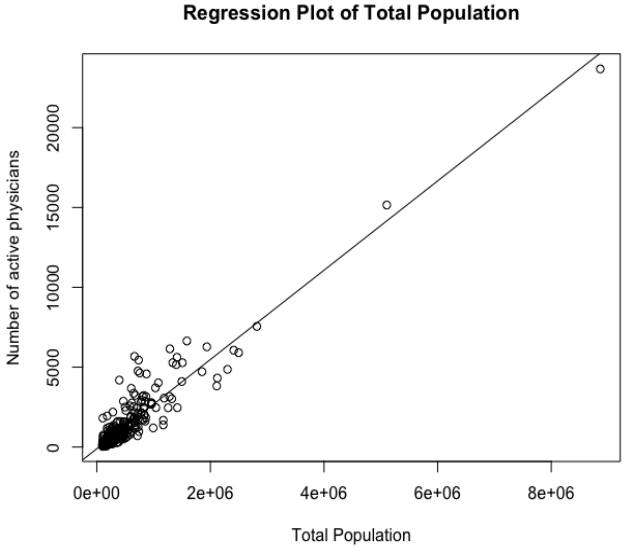
The Estimated Regression function for Y and Number of hospital beds (X2):

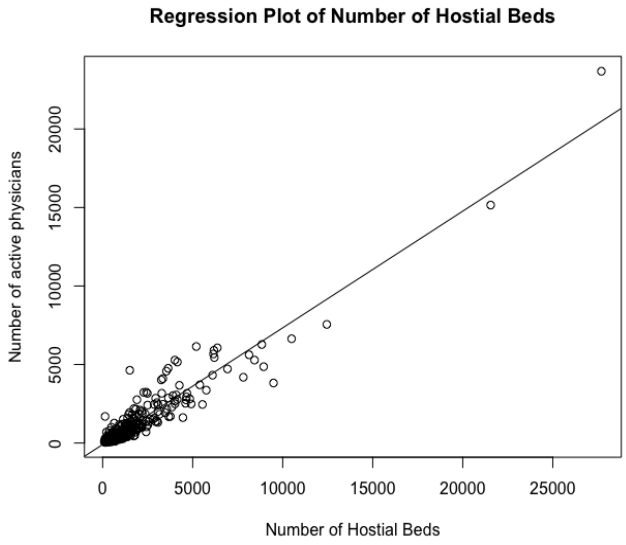
$$\hat{Y} = -95.93218 + 0.7431164X2$$

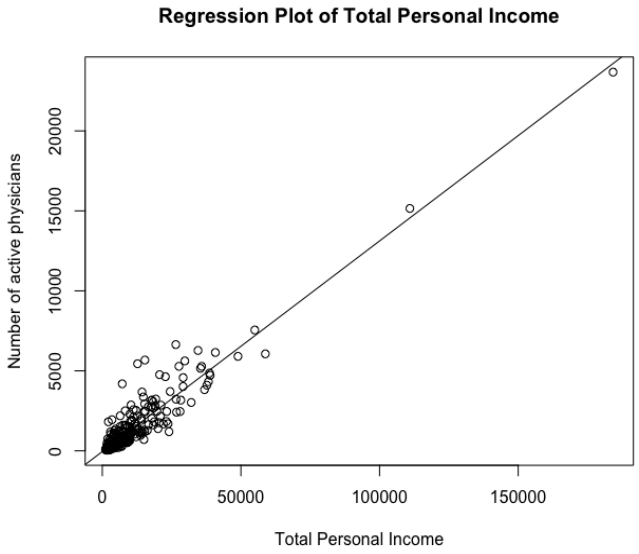
The Estimated Regression function for Y and Total personal income (X3):

$$\hat{Y} = -48.39485 + 0.1317012X3$$

Part b) - Regression Plots for each of the three variables







Based off the graphs above a linear regression model seems like a good fit. The data for each of the graphs hovers around the regression line

Part c)
For Y and X1 MSE for 372203.5
For Y and X2 MSE for 310191.9
For Y and X3 MSE for 324539.4
Smallest is X2, number of hospital beds.

1.44
Let Y = per capita income and X = Percentage of the population with a Bachelor's degree
There are four different geographical regions being overseved. Region 1 = NE Region 2 = NC Region 3 = S, and Region 4 = W
part a)
The Regression Models for X and Y for each of the 4 different Regions
The relationship between X and Y for Region 1: $\hat{Y} = 9223.816 + 522.1588X$
The relationship between X and Y for Region 2: $\hat{Y} = 13581.41 + 238.6694X$
The relationship between X and Y for Region 3: $\hat{Y} = 10529.79 + 330.6117X$
The relationship between X and Y for Region 4: $\hat{Y} = 8615.053 + 440.3157X$

Part b)

The estimator functions are not similar for all regions. There is a large difference between the highest and lowest values of both β_0 and β_1. The function with the highest value of β_0 does not have the highest value of β_1, so they are not proportionally similar either.

Part 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 regression line is very similar for regions 1 and 3. Region 4 is also similar to regions 1 and 3. Region 2 is very dissimilar from the other regions. The data indicates that the regression function fits best for region 2.

PART 2
Y = Number of active physicians , X1 = Total population, X2 = Number of hospital beds, and X3 = Total personal income

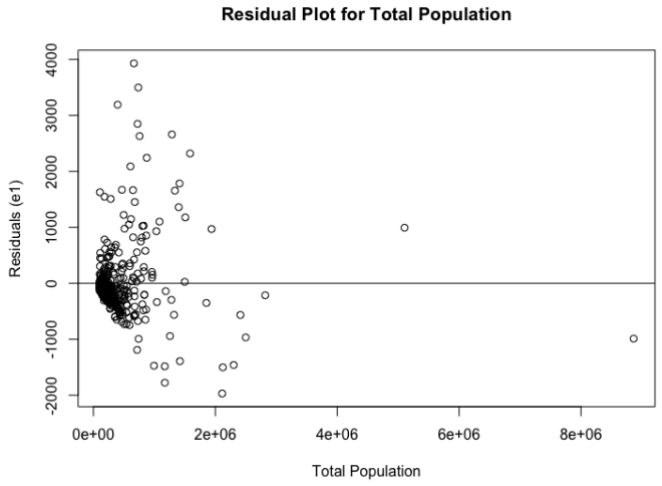
For Y SSTO = 1406206299

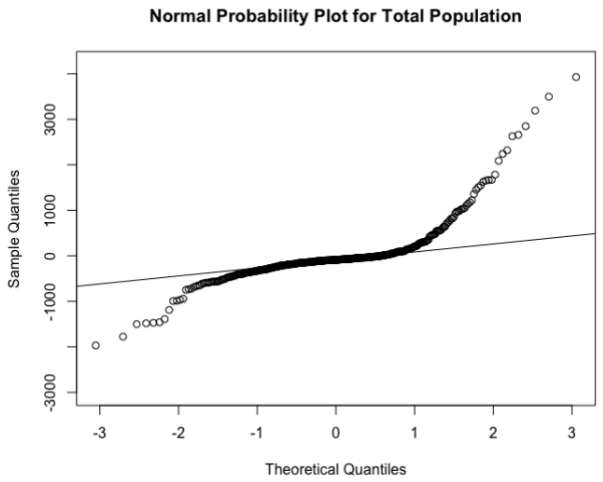
For Y and X1 SSR = 1243181164
For Y and X2 SSR = 1270342254
For Y and X3 SSR = 1264058045

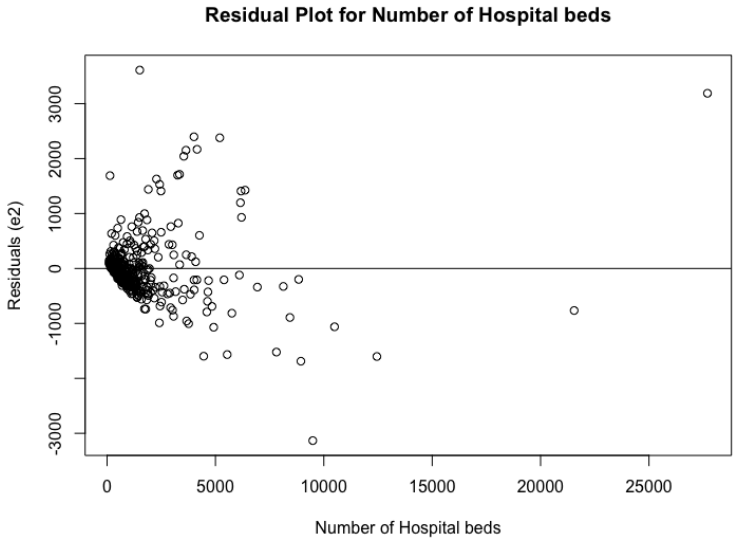
For Y and X1 R^2 = $\frac{1243181164}{1406206299}$= 0.8840674
For Y and X2 R^2 = $\frac{1270342254}{1406206299}$= 0.9033826
For Y and X3 R^2 = $\frac{1264058045}{1406206299}$= 0.8989137
The predictor variable that accounts for the largest reduction in the variability in active number of physicians is the number of active hospital beds (X2).

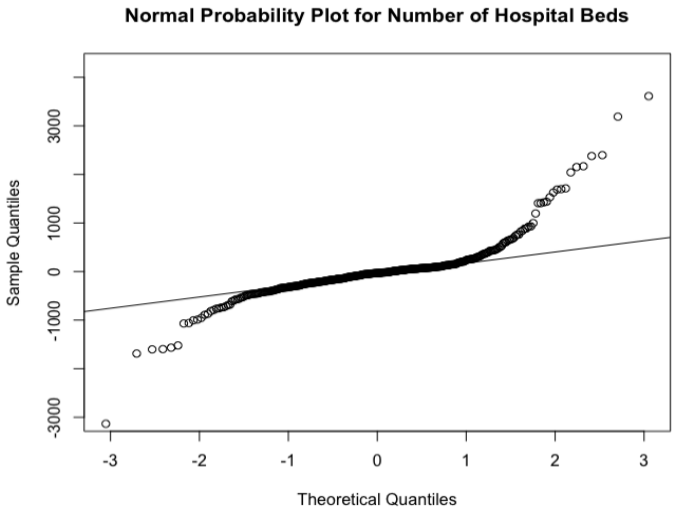
PART 3
Confidence intervals for $\hat{\beta}1$by region
CI for Region 1: [460.9537, 583.3640]
CI for Region 2: [193.7858, 283.5530]
CI for Region 3: [285.8904, 375.3331]
CI for Region 4: [365.5336, 515.0978]
The regression lines for each region do not appear to have similar slopes.

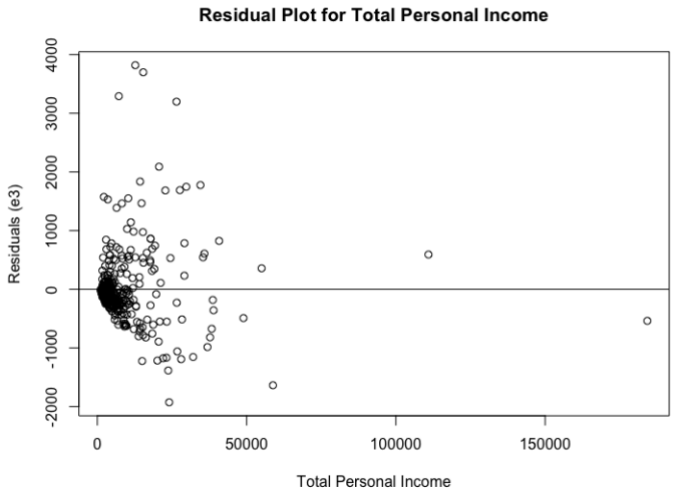
PART 4

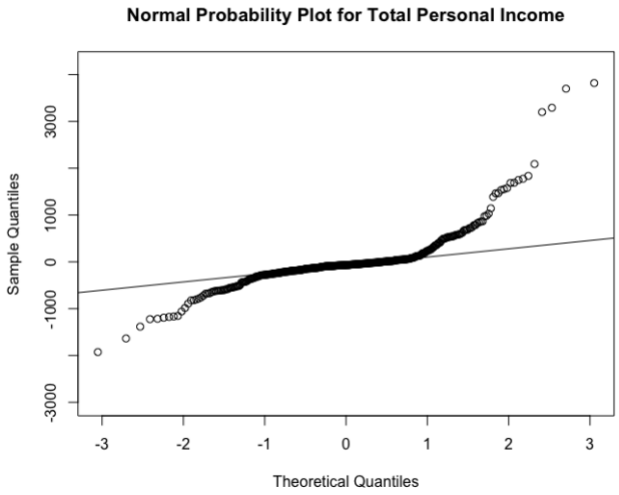

Residual Plot for Total Population
Residuals (e1)
4000
3000
2000
1000
0
-1000
-2000
0e+00
2e+06
4e+06
6e+06
8e+06
Total Population


Normal Probability Plot for Total Population
Sample Quantiles
3000
1000
0
-1000
-3000
-3
-2
-1
0
1
2
3
Theoretical Quantiles


Residual Plot for Number of Hospital beds
Residuals (e2)
3000
2000
1000
0
-1000
-3000
0
5000
10000
15000
20000
25000
Number of Hospital beds


Normal Probability Plot for Number of Hospital Beds
Sample Quantiles
3000
1000
0
-1000
-3000
-3
-2
-1
0
1
2
3
Theoretical Quantiles





Based on these plots, we can conclude that the normal distribution is not a good fit for these data. This is because all of the Normal Probability Plots have long tails. Because sample size is sufficiently large, the normality assumption is not needed as the Central Limit Theorem ensures that the distribution of the errors will be normal. The linear regression model is equally applicable to all three variables.

Conclusion

Based on the tests and calculations run in Part 1, it would appear that total population, number of hospital beds, and total personal income all have a positive correlation with number of active physicians. The number of hospital beds seemed to be the least variable predictor of number of active physicians. So, out of the three predictors, the number of hospital beds would be the best way to predict the number of active physicians. This is corroborated by the results of part 2. It was also found that per capita income was best predicted by percentage of population with bachelor's degrees in region 2. This implies that in regions 1,3, and 4, the percentage of the population with bachelor's degrees is not as reliable an indicator of per capita income. Part 3 shows a range of values for each region. We are 99% confident that the slope for each region lies between the two values listed. Based on these intervals, it does not appear that there is a similar relationship between per capita income and bachelor's degrees among each region; based on our models, region 2 has the slowest increase in per capita income as percentage of bachelor's degrees increases and region 1 has the fastest. With Part 4, we can conclude that

the models in this project are compatible with the data. In order to improve the models, we would need to increase the number of counties surveyed.