Project 1

Arya Gupta

1/28/2022

Introduction

- This project consists of 5 parts: fitting regression models, measuring linear associations, inference about regression parameters, regression diagnostics, and a conclusion on the project. We are given a data set which consists of county demographic information for 440 of the most populous counties in the United States. The data set provides 14 different variables for a single county. In this project, I worked with 6 of these variables: Total Population, Number of Active Physicians, Number of Hospital Beds, Percent of Bachelor's Degrees, Total Personal Income, and Geographic Location.
- For Parts 1, 2, and 4, the three predictor variables were Number of Hospital Beds, Total Population, and Total Personal Income. The response variable was Number of Active Physicians.
- For Part 3, for each geographic region, the predictor variable was Percent of Bachelor's Degrees and the response variable was Per Capita Income.
- There will be a few tools I use, mainly "qf" which is used for finding the critical value of the F-distribution, and "qt" which is used for finding the critical value of the t-distribution. I will use the "plot" function to create our graphs. Lastly, to create a normal probability plot, I will need to use the "qqplot" function.
- In this report, I have detailed the 5 different parts of the project, ranging from part I to part V. I used the "echo = FALSE" command to hide the code from the project so it only shows the results. I have attached the code at the end of the project which is called the Appendix. In the Appendix, I used the "results = 'hide' " command to hide the results.

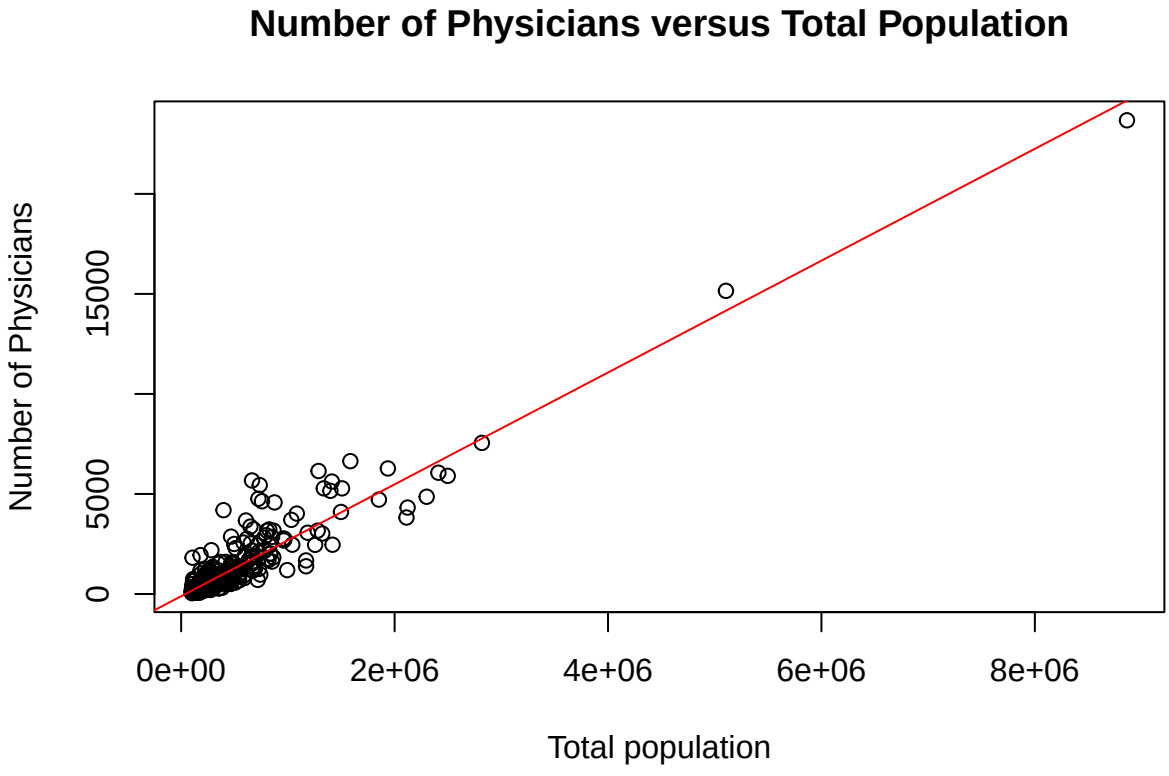
Part I

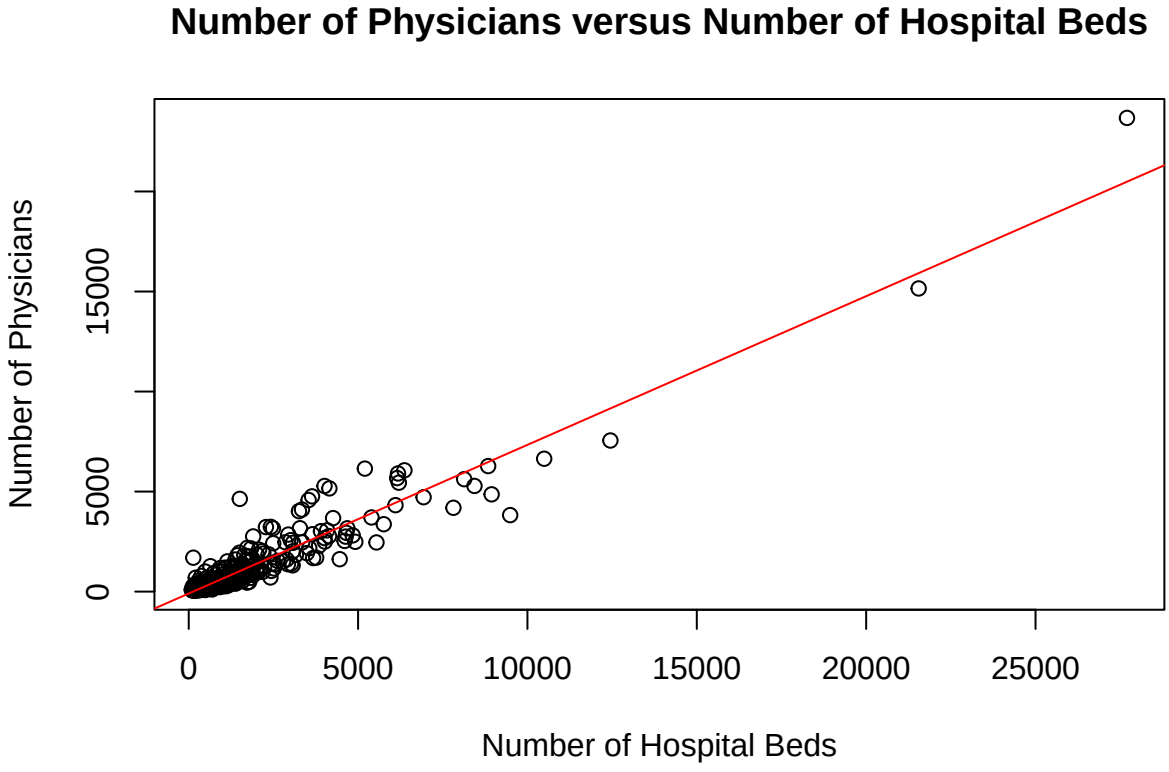
a)

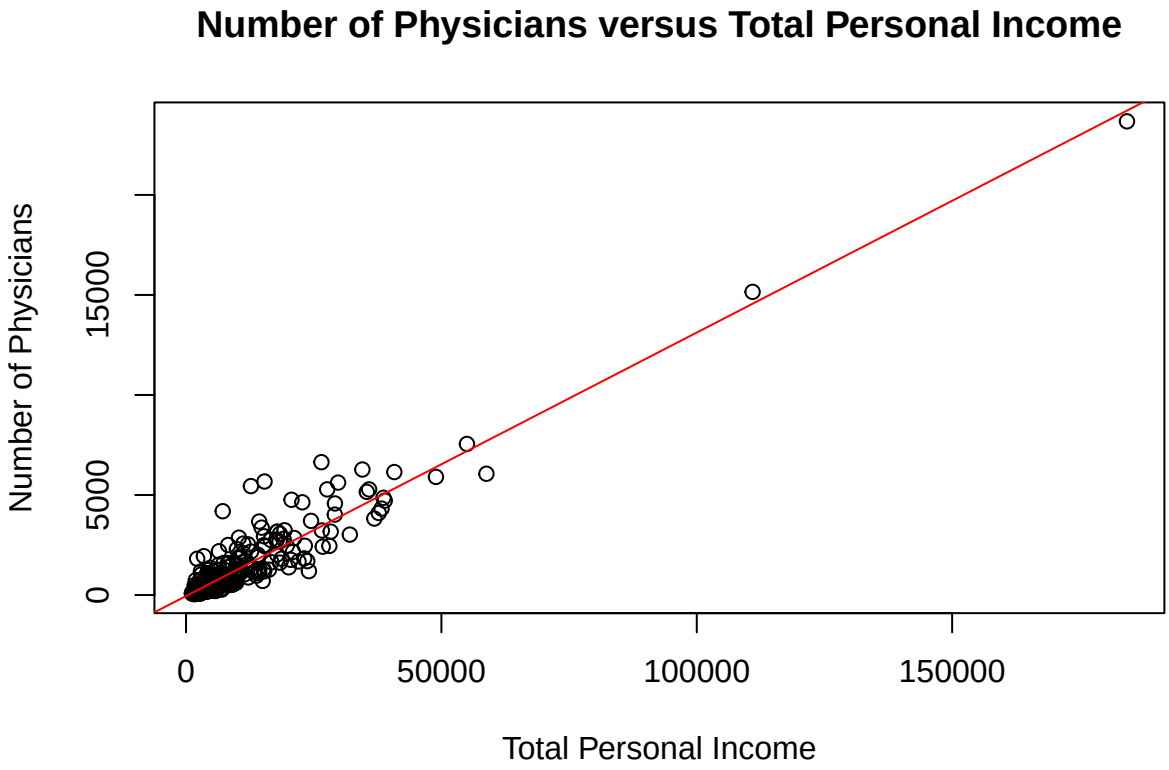
Let Y = Number of Active Physicians, Let XPop = Total Population, Let XBed = Number of Hospital Beds, Let XInc = Total Personal Income

- The Regression function for the relationship between Total Population and Number of Active Physicians is: Yhat = -110.6348 + 0.002795425(X).
- The Regression function for the relationship between Number of Hospital Beds and Number of Active Physicians is: Yhat = -95.93218 + 0.7431164(X).
- The Regression function for the relationship between Total Personal Income and Number of Active Physicians is: Yhat = -48.39485 + 0.1317012(X).

b)


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


Number of Physicians versus Number of Hospital Beds
Number of Physicians
0
5000
15000
0
5000
10000
15000
20000
25000
Number of Hospital Beds



For all three scatter plots, a linear regression relation appears to provide a good fit for each of the three predictor variables.

c)

- The MSE for Total Population is 372203.5
- The MSE for Number of Hospital Beds is 310191.9
- The MSE for Total Personal Income is 324539.4
- The Number of Hospital Beds leads to the smallest variability around the fitted regression line as it has the lowest MSE.

Part II

- R^2 value for Total Population versus Number of Active Physicians is 0.8840674
- R^2 value for Number of Hospital Beds versus Number of Active Physicians is 0.9033826
- R^2 value for Total Personal Income versus Number of Active Physicians is 0.8989137
- R^2 in this case represents the percentage of variation in active physicians that can be explained by the predictor variables. Since the value of R^2 is the highest for the relationship between Number of Hospital Beds and Number of Active Physicians, I can say that the predictor variable which accounts for the largest reduction in the variability in the Number of Active Physicians is the Number of Hospital Beds.

Part III

Region 1

H0: $B_1 = 0$

HA: B_1 != 0

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
XRegion1	1	1450517671	1450517671	197.7527	0
Residuals	101	740835765	7335008	NA	NA

- 90% Confidence interval for region 1 is (460.5177 , 583.8).
- Since the F* statistic(197.7527) is larger than the critical value(2.755868), I reject the null hypothesis.

Region 2

H0: $B_1 = 0$

HA: B_1 != 0

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
XRegion2	1	338907694	338907694	76.82646	0
Residuals	106	467602149	4411341	NA	NA

- 90% Confidence interval for region 2 is (193.4858 , 283.853)
- Since the F* statistic(76.82646) is larger than the critical value(2.753462), I reject the null hypothesis.

Region 3

H0: $B_1 = 0$

HA: B_1 != 0

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
XRegion3	1	1109873245	1109873245	148.491	0
Residuals	150	1121152411	7474349	NA	NA

- 90% Confidence interval for region 3 is (285.7076 , 375.5158)
- Since the F* statistic(148.491) is larger than the critical value(2.739275), I reject the null hypothesis.

Region 4

H0: $B_1 = 0$

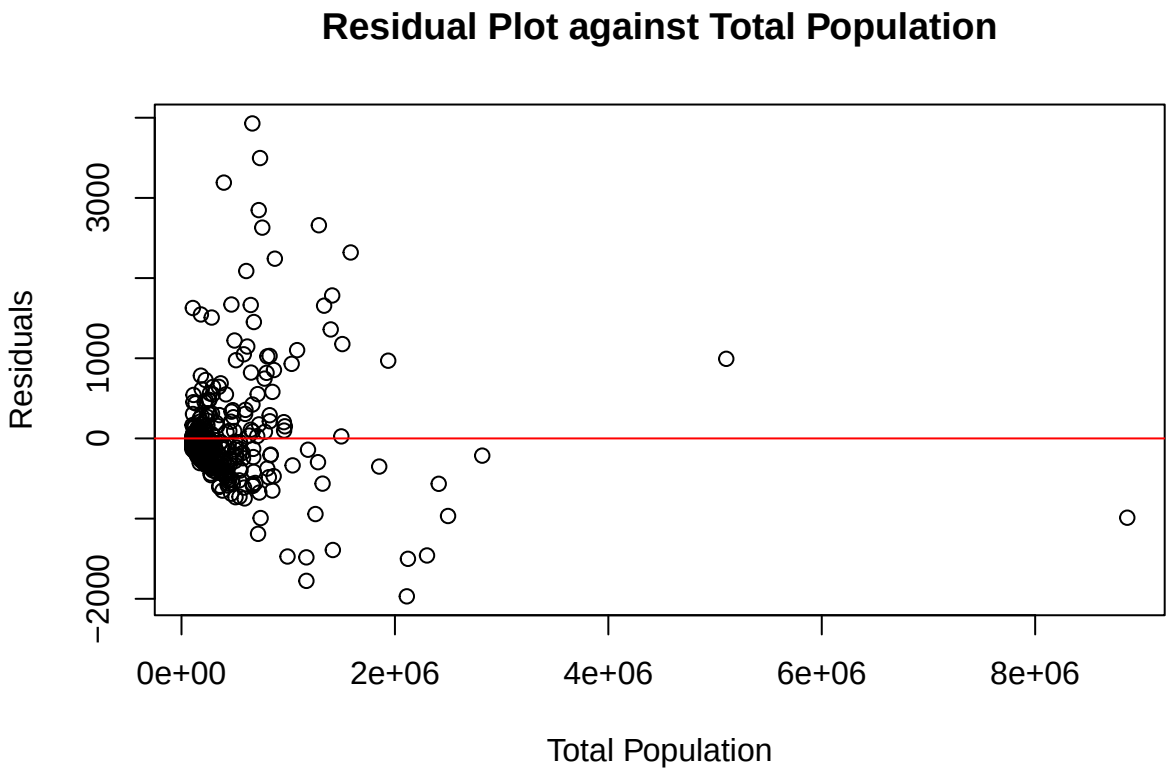
HA: B_1 != 0

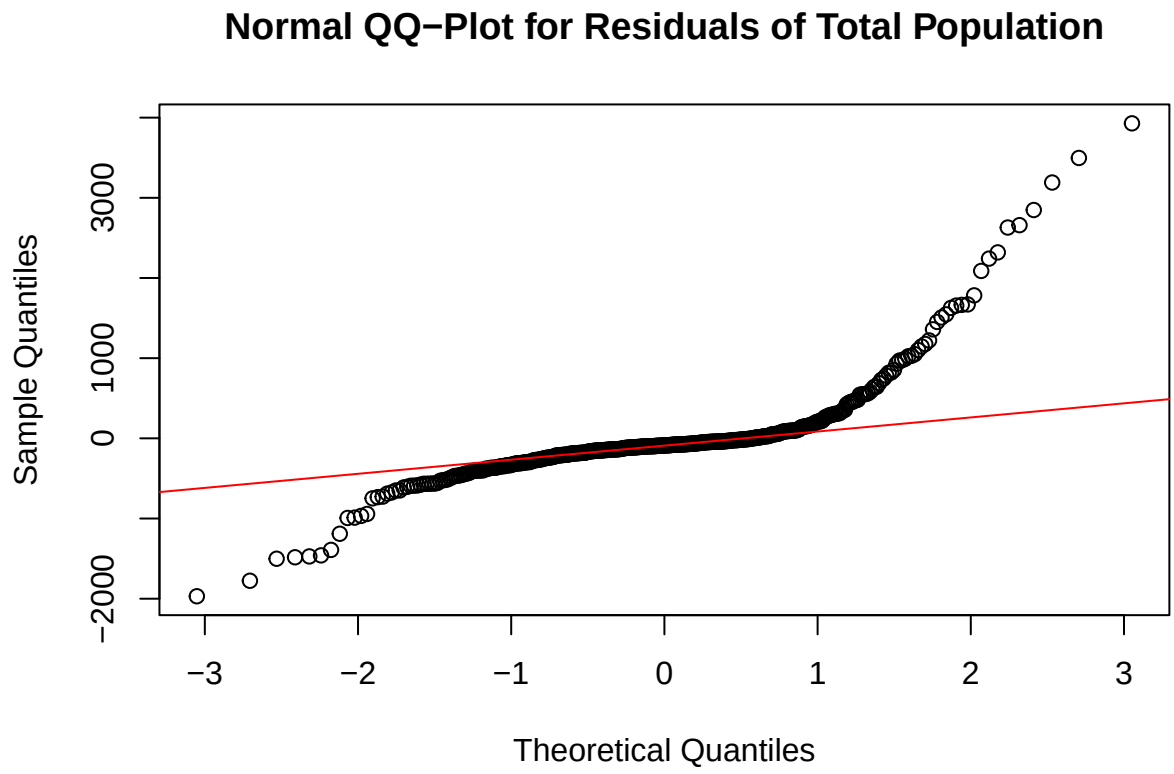
	Df	Sum Sq	Mean Sq	F value	Pr(>F)
XRegion4	1	773745787	773745787	94.19477	0
Residuals	75	616073841	8214318	NA	NA

- 90% Confidence interval for region 4 is (364.7585 , 515.8729)
- Since the F* statistic(94.19477) is larger than the critical value(2.773642), I reject the null hypothesis.
- The regression lines for the different regions do not appear to have similar slopes. It is apparent that Region 1 has the largest slope and Region 2 has the smallest slope. The slopes are vastly different.

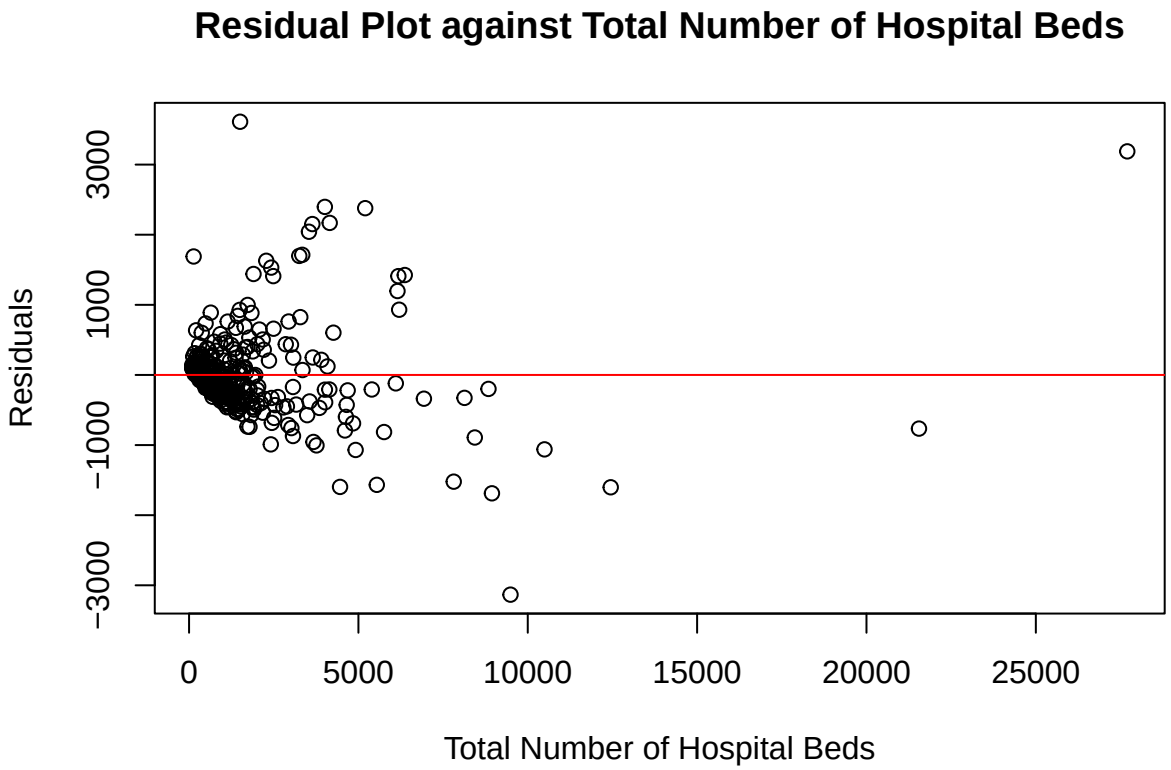
Part IV

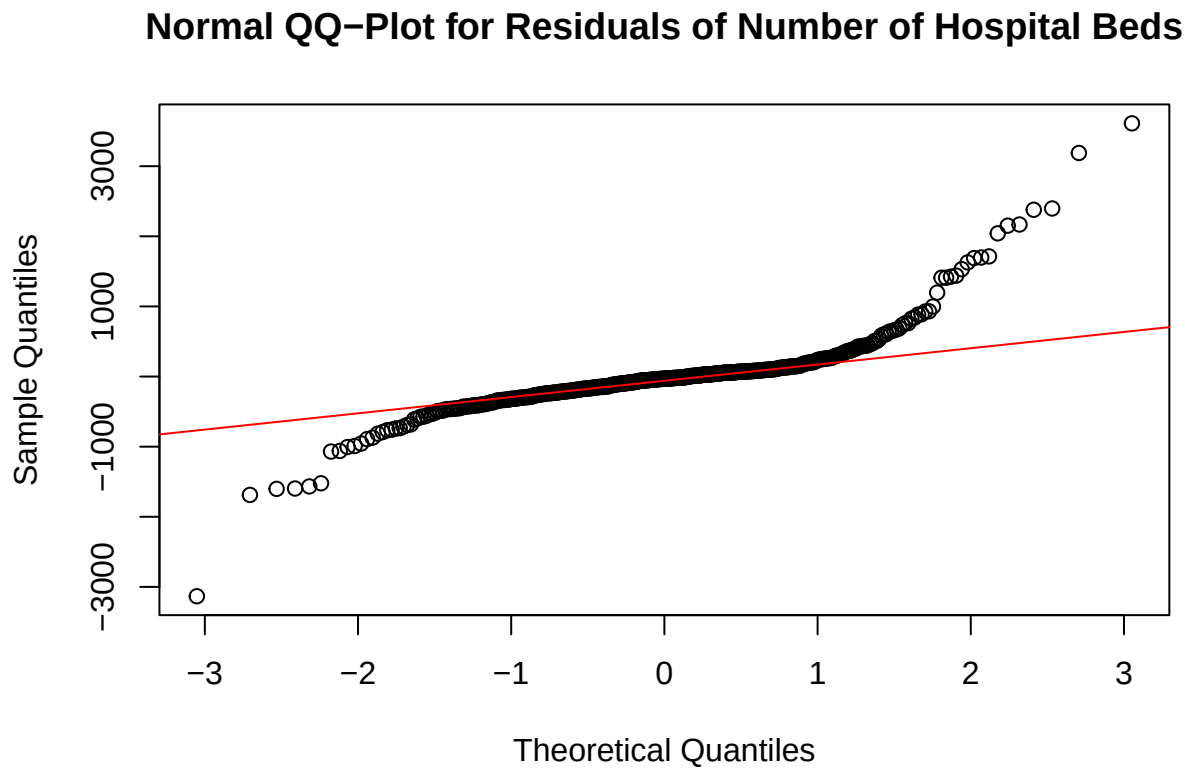
a)



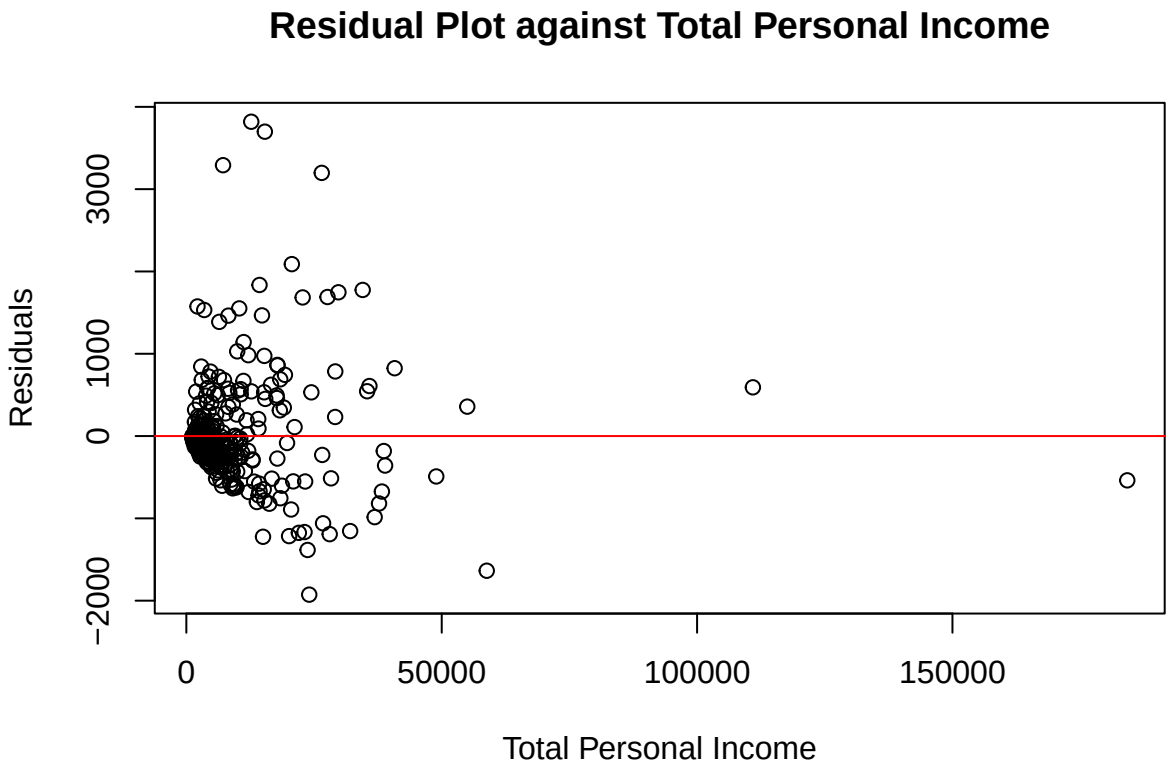

Normal QQ−Plot for Residuals of Total Population
Sample Quantiles
Theoretical Quantiles
−2000
0
1000
3000
−3
−2
−1
0
1
2
3

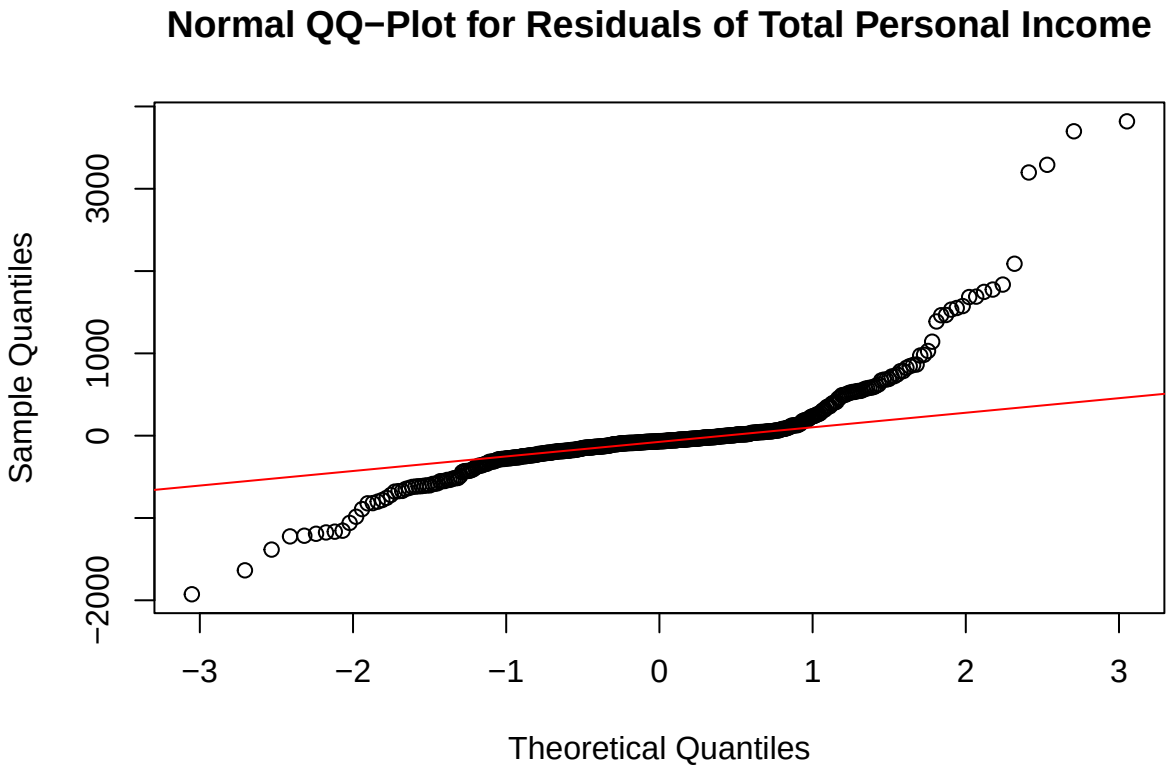
b)


Residual Plot against Total Number of Hospital Beds
Residuals
Total Number of Hospital Beds
−3000
−1000
1000
3000
0
5000
10000
15000
20000
25000


Normal QQ−Plot for Residuals of Number of Hospital Beds
Sample Quantiles
Theoretical Quantiles
−3000
−1000
1000
3000
−3
−2
−1
0
1
2
3

c)


Residual Plot against Total Personal Income
Residuals
Total Personal Income
−2000
0
1000
3000
0
50000
100000
150000



- After calculating the correlation test for normality, I found that the correlation coefficient for total population vs number of active physicians is equal to 0.8520252.
- After calculating the correlation test for normality, I found that the correlation coefficient for number of hospital beds vs number of active physicians is equal to 0.8803917.
- After calculating the correlation test for normality, I found that the correlation coefficient for number of hospital beds vs number of active physicians is equal to 0.8596853.
- Linear regression model (2.1) is more appropriate for the number of hospital beds versus number of active physicians, since its correlation coefficient for normality is the largest out of all the predictor variables.

Part V

- In Part I calculated and plotted the linear regression models for each of the predictor variables (Number of Hospital Beds, Total Population, Total Personal Income) against the response variable (Number of Active Physicians). I found that all three regression functions had negative Y-intercepts, but a positive slope. The MSE (mean squared error) for each model was also calculated, with the MSE of the Number of Hospital Beds being the lowest.
- In Part II the Coefficients of Determination were calculated for the three models. I found that since the Coefficient of Determination (R^2) was the highest for the relationship between Number of Hospital Beds and Number of Active Physicians, I can say that 90.33% of the variability observed in the Number of Active Physicians can be attributed to Number of Hospital Beds.
- In Part III, I found the slope of Per Capita Income versus Percent Bachelor's degrees in each of the four geographic regions. I then created 90% confidence intervals for $\hat{B}1$ in each of the four regions. Finally, I carried out the ANOVA of each regression model and their respective F-tests. Through the

F-tests, I concluded that I must reject the null hypothesis for all four regression models. I found that the slope for Region 1 was the largest and the slope for Region 2 was the smallest.

- Observational data impacts our results because it creates a larger margin of variance that could be attributed to other parameters besides the one that is being tested against active physicians. Without treatments or controls it is difficult to identify how much of the data is attributed to the observation. There is no randomization of treatment. In Part IV I identified that the predictor variables exhibited high residual distributions and correlation tests that appeared normal (both in shape and in values). Overall, in the calculations for correlation test for normality, I found that the numbers of hospital beds versus number of active physicians had the highest value (0.8803917) indicating the regression model was the best fit. The linear regression model can be improved for this particular subset by, for example, increasing the amount of hospital beds in a given county and monitoring the effects it has on active physicians. Throughout this project, it has been clear that a higher number of hospital beds leads to a greater number of active physicians compared to the other two predictor variables. Although both total population and total personal income gave us a slightly smaller correlation coefficient, they both still showed a high correlation against the number of active physicians, with total personal income being slightly higher than total population in relation to the number of active physicians.

Appendix

Project 1.43 code

a)

```
data1 = read.table("CDI.txt")
Y = data1[,8]
XPop = data1[,5]
n = length(XPop)

b1Pop = sum((XPop-mean(XPop))*(Y-mean(Y)))/sum((XPop-mean(XPop))^2)
b0Pop = mean(Y)-b1Pop*mean(XPop)
YPop = b0Pop + b1Pop*(XPop)
fitPop = lm(Y~XPop)
print(b1Pop)
print(b0Pop)
```

```
XBed = data1[,9]
b1Bed = sum((XBed-mean(XBed))*(Y-mean(Y)))/sum((XBed-mean(XBed))^2)
b0Bed = mean(Y)-b1Bed*mean(XBed)
YBed = b0Bed + b1Bed*(XBed)
fitBed = lm(Y~XBed)
print(b1Bed)
print(b0Bed)
```

```
XInc = data1[,16]
b1Inc = sum((XInc-mean(XInc))*(Y-mean(Y)))/sum((XInc-mean(XInc))^2)
b0Inc = mean(Y)-b1Inc*mean(XInc)
YInc = b0Inc + b1Inc*(XInc)
fitInc = lm(Y~XInc)
print(b1Inc)
print(b0Inc)
```

b)

```
plot(XPop, Y, main = 'Plot of number of physicians versus total population',
     xlab = 'Total population', ylab = 'Number of Physicians')
abline(b0Pop, b1Pop, col='red')
```

```
plot(XBed, Y, main = 'Plot of number of physicians versus Number of Hospital Beds',xlab = 'Number of Hos
abline(b0Bed, b1Bed, col='red')
```

```
plot(XInc, Y, main = 'Plot of number of physicians versus Total Population Income',xlab = 'Total Popula
abline(b0Inc, b1Inc, col='red')
```

c)

```
MSEPop = (sum((Y-YPop)^2))/(n-2)
print(MSEPop)
```

```
MSEBed = (sum((Y-YBed)^2))/(n-2)
print(MSEBed)
```

```
MSEInc = (sum((Y-YInc)^2))/(n-2)
print(MSEInc)
```

Project 2.62 code

```
SSR_Pop = sum((YPop - mean(Y))^2)
SSE_Pop = sum((Y - YPop)^2)
SSTO_Pop = SSR_Pop + SSE_Pop

RSquared_Pop = SSR_Pop/SSTO_Pop

print(RSquared_Pop)
```

```
SSR_Bed = sum((YBed - mean(Y))^2)
SSE_Bed = sum((Y - YBed)^2)
SSTO_Bed = SSR_Bed + SSE_Bed

RSquared_Bed = SSR_Bed/SSTO_Bed
print(RSquared_Bed)
```

```
SSR_Inc = sum((YInc - mean(Y))^2)
SSE_Inc = sum((Y - YInc)^2)
SSTO_Inc = SSR_Inc + SSE_Inc

RSquared_Inc = SSR_Inc/SSTO_Inc
print(RSquared_Inc)
```

Project 2.63 code

Region 1

```
ind1<-seq(1,440)[data1[,17]==1]
YRegion1 = data1[ind1,][,15]
XRegion1 = data1[ind1,][,12]
b1_R1 = sum((XRegion1-mean(XRegion1))*(YRegion1-mean(YRegion1)))/sum((XRegion1-mean(XRegion1))^2)
b0_R1 = mean(YRegion1)-b1_R1*mean(XRegion1)
YFittedR1 = b0_R1 + b1_R1*XRegion1
MSE_R1 = sum((YRegion1-YFittedR1)^2)/(101)
fit_R1 = lm(YRegion1~XRegion1)
t_R1 = qt(0.95,101)

SE_R1 = (sqrt((MSE_R1)/sum((XRegion1-mean(XRegion1))^2)))
PM_R1 = t_R1*SE_R1

Pinterval_R1 = b1_R1 + PM_R1
Ninterval_R1 = b1_R1 - PM_R1
print(Ninterval_R1)
print(Pinterval_R1)

SSR_R1 = sum((YFittedR1 - mean(YRegion1))^2)
SSE_R1 = sum((YRegion1 - YFittedR1)^2)
SSTO_R1 = SSR_R1 + SSE_R1
MSR_R1 = SSR_R1/1

F_R1 = MSR_R1/MSE_R1
print(F_R1)
qf(0.90,1,101)
```

```
library(knitr)
kable(anova(fit_R1))
```

Region 2

```
ind2<-seq(1,440)[data1[,17]==2]
YRegion2 = data1[ind2,][,15]
XRegion2 = data1[ind2,][,12]
b1_R2 = sum((XRegion2-mean(XRegion2))*(YRegion2-mean(YRegion2)))/sum((XRegion2-mean(XRegion2))^2)
b0_R2 = mean(YRegion2)-b1_R2*mean(XRegion2)
YFittedR2 = b0_R2 + b1_R2*XRegion2
MSE_R2 = sum((YRegion2-YFittedR2)^2)/(106)
fit_R2 = lm(YRegion2~XRegion2)
t_R2 = qt(0.95,106)
SE_R2 = (sqrt((MSE_R2)/sum((XRegion2-mean(XRegion2))^2)))
PM_R2 = t_R2*SE_R2

Pinterval_R2 = b1_R2 + PM_R2
Ninterval_R2 = b1_R2 - PM_R2
```

```
print(Ninterval_R2)
print(Pinterval_R2)

SSR_R2 = sum((YFittedR2 - mean(YRegion2))^2)
SSE_R2 = sum((YRegion2 - YFittedR2)^2)
SSTO_R2 = SSR_R2 + SSE_R2
MSR_R2 = SSR_R2/1

F_R2 = MSR_R2/MSE_R2
print(F_R2)
qf(0.90,1,106)
```

```
library(knitr)
kable(anova(fit_R2))
```

Region 3

```
ind3<-seq(1,440)[data1[,17]==3]
YRegion3 = data1[ind3,][,15]
XRegion3 = data1[ind3,][,12]

b1_R3 = sum((XRegion3-mean(XRegion3))*(YRegion3-mean(YRegion3)))/sum((XRegion3-mean(XRegion3))^2)
b0_R3 = mean(YRegion3)-b1_R3*mean(XRegion3)
YFittedR3 = b0_R3 + b1_R3*XRegion3
MSE_R3 = sum((YRegion3-YFittedR3)^2)/(150)
fit_R3 = lm(YRegion3~XRegion3)
t_R3 = qt(0.95,150)
SE_R3 = (sqrt((MSE_R3)/sum((XRegion3-mean(XRegion3))^2)))
PM_R3 = t_R3*SE_R3

Pinterval_R3 = b1_R3 + PM_R3
Ninterval_R3 = b1_R3 - PM_R3
print(Ninterval_R3)
print(Pinterval_R3)

SSR_R3 = sum((YFittedR3 - mean(YRegion3))^2)
SSE_R3 = sum((YRegion3 - YFittedR3)^2)
SSTO_R3 = SSR_R3 + SSE_R3
MSR_R3 = SSR_R3/1

F_R3 = MSR_R3/MSE_R3
print(F_R3)
qf(0.90,1,150)
```

```
library(knitr)
kable(anova(fit_R3))
```

Region 4

```
ind4<-seq(1,440)[data1[,17]==4]
YRegion4 = data1[ind4,][,15]
XRegion4 = data1[ind4,][,12]


b1_R4 = sum((XRegion4-mean(XRegion4))*(YRegion4-mean(YRegion4)))/sum((XRegion4-mean(XRegion4))^2)
b0_R4 = mean(YRegion4)-b1_R4*mean(XRegion4)
YFittedR4 = b0_R4 + b1_R4*XRegion4
MSE_R4 = sum((YRegion4-YFittedR4)^2)/(75)
fit_R4 = lm(YRegion4~XRegion4)
t_R4 = qt(0.95,75)
SE_R4 = (sqrt((MSE_R4)/sum((XRegion4-mean(XRegion4))^2)))
PM_R4 = t_R4*SE_R4

Pinterval_R4 = b1_R4 + PM_R4
Ninterval_R4 = b1_R4 - PM_R4
print(Ninterval_R4)
print(Pinterval_R4)

SSR_R4 = sum((YFittedR4 - mean(YRegion4))^2)
SSE_R4 = sum((YRegion4 - YFittedR4)^2)
SSTO_R4 = SSR_R4 + SSE_R4
MSR_R4 = SSR_R4/1

F_R4 = MSR_R4/MSE_R4
print(F_R4)
qf(0.90,1,75)
```

```
library(knitr)
kable(anova(fit_R4))
```

Project 3.25 code

```
Residuals_Pop = Y - YPop
plot(XPop, Residuals_Pop, main='Residual Plot against Total Population', xlab = 'Total Population', yla
abline(h=0, col='red')
```

```
res_pop <- qqnorm(Residuals_Pop, plot.it = FALSE)
r_pop=cor(res_pop$x, res_pop$y)
```

```
qqnorm(Residuals_Pop, main='Normal QQ-Plot for Residuals of Total Population')
qqline(Residuals_Pop, col='red')
```

```
r_pop
```

```
Residuals_Bed = Y - YBed
plot(XBed, Residuals_Bed, main='Residual Plot against Total Number of Hospital Beds', xlab = 'Total Numb
abline(h=0, col='red')
```

```
res_bed <- qqnorm(Residuals_Bed, plot.it = FALSE)
r_bed = cor(res_bed$x, res_bed$y)
```

```
qqnorm(Residuals_Bed, main = 'Normal QQ-Plot for Residuals of Number of Hospital Beds')
qqline(Residuals_Bed, col='red')
```

```
r_bed
```

```
Residuals_Inc = Y - YInc
plot(XInc, Residuals_Inc, main='Residual Plot against Total Personal Income', xlab = 'Total Personal In
abline(h=0, col='red')
```

```
res_inc <- qqnorm(Residuals_Inc, plot.it = FALSE)
r_inc = cor(res_inc$x, res_inc$y)
```

```
qqnorm(Residuals_Inc, main = 'Normal QQ-Plot for Residuals of Total Personal Income')
qqline(Residuals_Inc, col='red')
```

```
r_inc
```