STA108 Project 1

Tomer Fidelman, Meghana Vemula, Quilvio Hernandez

4/29/2019

Background and Introduction

The CDI is a data set that provides county demographic information for 440 of the most populous counties in the US. The primary objective of the project was to identify if there was a linear relationship between several significant attributes of US continental counties. Initially, the project team was interested in determining if there were linear relationships present between the number of active physicians in a county and 3 specific predictor variables: total population, number of hospital beds, and total personal income. For this part, the team formed linear regression models for each predictor variable. Later, the R2 goodness of fit values were compared between the three predictor values to determine which factor accounted for the largest variability reduction of the number of physicians. Next, the research team analyzed per capita income of the counties against the predictor variable containing the percentage of bachelor degree holding county residents. The researchers analyzed regression models between the four continental regions of the US for similarity. They cross compared error values for variance and regression coefficients. To further analyze the effects of bachelor degree holding residents of counties on per capita income, confidence intervals and F-tests were utilized to identify similarity and presence of linear relationships between these variables. Lastly, the researchers used residual and normal probability plots as regression diagnostics to determine if a linear model was an appropriate fit for the original 3 predictor variables (total population, number of hospital beds, and total personal income) when related to the number of active physicians. Overall, the research team used advanced linear regression analysis tools such as confidence intervals, normal probability plots, goodness of fit statistics, residual plots, and regression modeling to determine if there were linear relationships between several unique variables of interest.

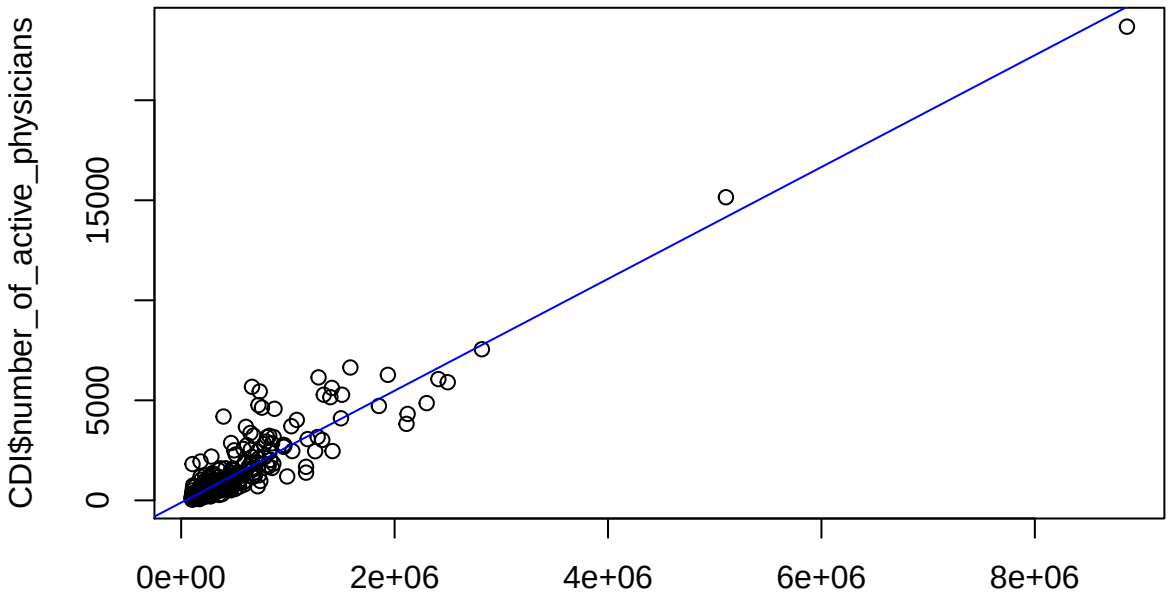
Part I. Fitting Regression Models

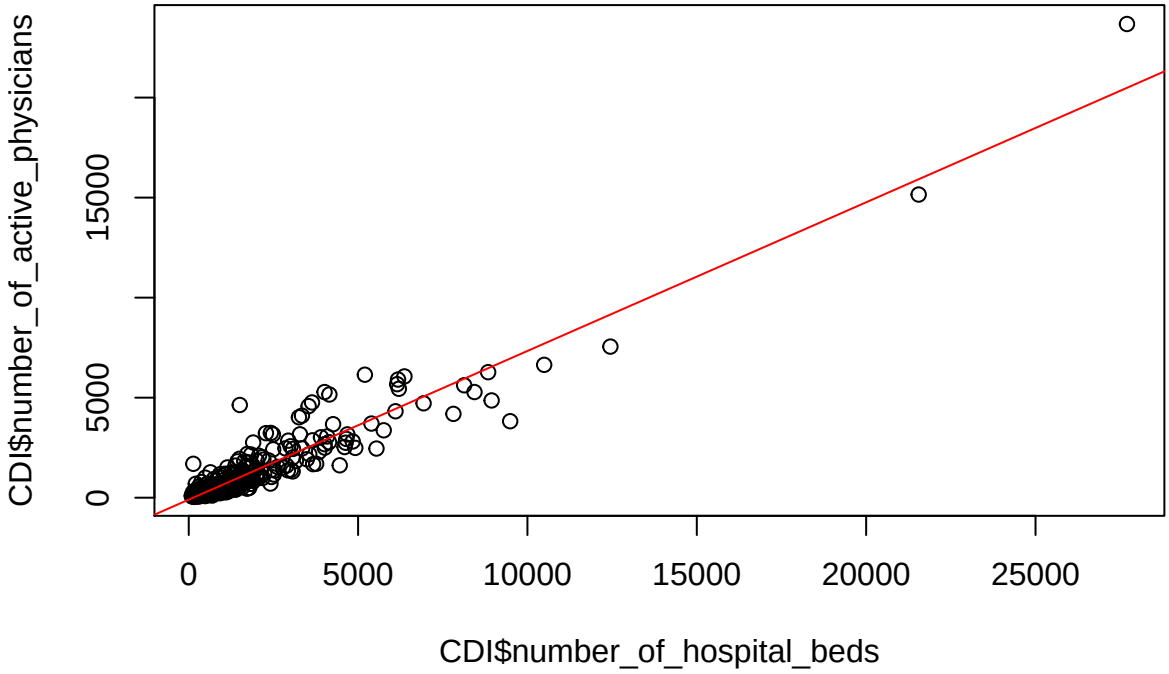
1.43 Refer to the CDI data set in Appendix C.2. The number of active physicians in a CDI(Y) is expected to be related to total population, number of hospital beds, and total personal income. Assume that first-order regression model (1.1) is appropriate for each of the three predictor variables.

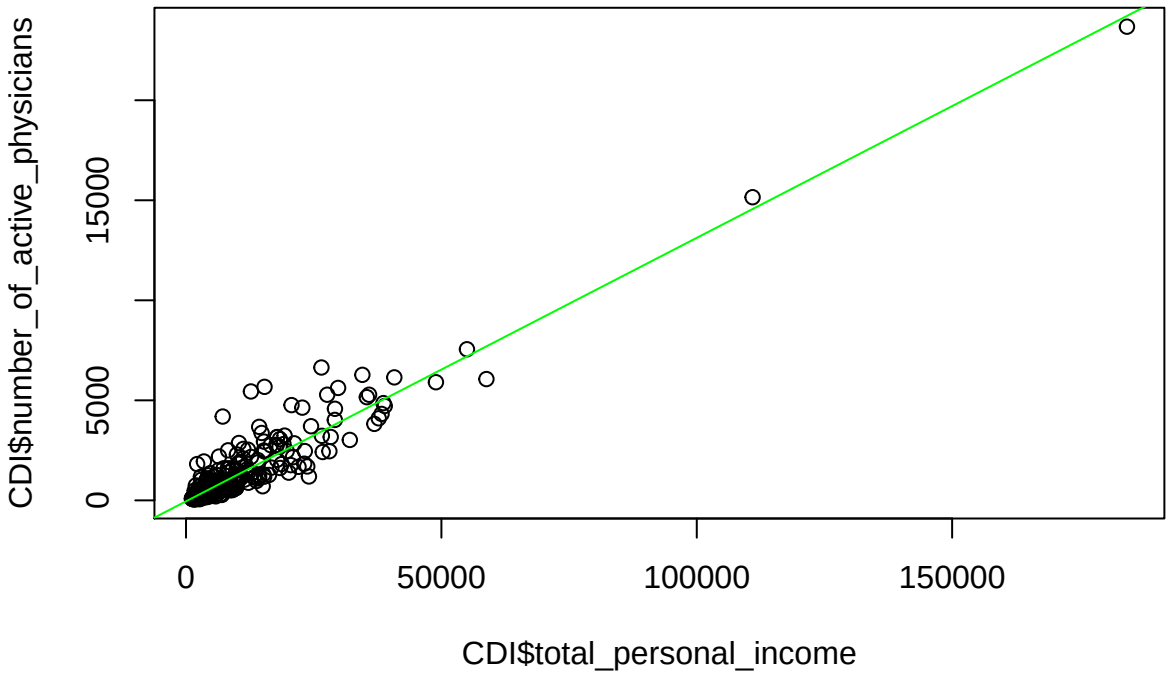
1a. Regress the number of active physicians in turn on each of the three predictor variables. State the estimated regression functions.

i. The regression function for the number of active physicians using the total population as a predictor is $-110.63 + .00279\hat{x}_i$.
ii. The regression function for the number of active physicians using the number of hospital beds as a predictor is $-95.93 + .7431\hat{x}_i$.
iii. The regression function for the number of active physicians using total personal income as a predictor is $-48.39 + .1317\hat{x}_i$.

1b. Plot the three estimated regression function and data on separate graphs. Does a linear regression relation appear to provide a good fit for each of the three predictor variables


CDI$number_of_active_physicians
0
5000
15000
0e+00
2e+06
4e+06
6e+06
8e+06
CDI$Total_population


CDI$number_of_active_physicians
0
5000
15000
0
5000
10000
15000
20000
25000
CDI$number_of_hospital_beds



Yes, a linear regression relation initially appear to provide a good fit for each of the three predictor variables.

1c. Calculate MSE for each of the three predictor variables. Which predictor variable leads to the smallest variability around the fitted regression line?

i. MSE for population is 3.722035×10^5.
ii. MSE for number of beds is 3.1019188×10^5.
iii. MSE for personal income is 3.2453939×10^5.

MSE for number of beds is the smallest and as a result leads to the smallest variability around the fitted regression line.

1.44 Refer to the CDI data set in Appendix C.2.

1.44a. For each geographic regions, regress per capita income in a CDI (Y) agaisnt the percentage of individuals in a county having at least a bachelor's degree (X). Assume that first-order regression model is appropriate for each region. State the estimated regression functions.

i. The regression function for the Northeast region is $\hat{per\ capita\ income} = 9223.82 + 522.16(\hat{bachelors})$, where $\hat{bachelors}$ denotes a percent point increase in the proportion of the population that hold a bachelor's degree.
ii. The regression function for the North Central region is $\hat{per\ capita\ income} = 13581.41 + 238.67(\hat{bachelors})$, where $\hat{bachelors}$ denotes a percent point increase in the proportion of the population that hold a bachelor's degree.
iii. The regression function for the South region is $\hat{per\ capita\ income} = 10529.79 + 330.61(\hat{bachelors})$, where $\hat{bachelors}$ denotes a percent point increase in the proportion of the population that hold a bachelor's degree.
iv. The regression function for the West region is $8615.05 + 440.32(\hat{bachelors})$, where $\hat{bachelors}$ denotes a percent point increase in the proportion of the population that hold a bachelor's degree.

1.44b. Are the estimated regression function similar for the four regions. Discuss.

The estimated regression are similar in the fact that they all show a positive relationship between per capita income and the percent of bachelor's degree in the population. However, the degree to which the a percentage point increase affect the regions differs. β_1 in the regression model for the Northeast region is more than double that of the North Central region and over 1.5 times more than the South. The West and Northeast

having higher β_1 parameters can possibly be characterized by the urban centers located in the regions such as New York in the Northeast and San Francisco in the West incentivising bachelor's degree holders to flock to the city. While the North Central region has the lowest slope, it has the highest intercept by a margin of approximately 3000. The intercept and slope of this model could indicate the workforce in the region is not as dependent on a bachelor's degree.

1.44c. Calculate MSE for each of the regions. Is the variability around the fitted regression line approximately the same for the four regions.

The variability around the fitted regression line is approximately the same for the Northeast and South regions (7.3350076×10^6 and 7.4743494×10^6, respectively), while the variability for the North Central is lower (4.411341×10^6) and the West is larger (8.2143179×10^6).

Part II - Measuring Linear Associations

Refer to the CDI data set in Appendix C.2 and Project 1.43. Using R^2 as the criterion, which predictor variable accounts for the largest reduction in the variability in the number of physicians?

The largest reduction in the variability in the number of physicians is attributed to number of hospital beds because it has the highest R^2 value at 0.9033826 compared to the the R^2 values of total population and total personal income, 0.8840674 and 0.8989137 respectively.

Part III - Inference About Regression Parameters

2.63 Refer to the CDI data set in Appendix C.2 and Project 1.44. Obtain a seperate interval estimate of β_1 for each region. Use a 90 percent confidence coefficient in each case. Do the regression lines for the different regions appear to have similar slopes?

i. We are 90% confident the true value of β_1 for the Northeast region is in the interval (460.52, 583.80).
ii. We are 90% confident the true value of β_1 for the North Central region is in the interval (193.49, 283.85).
iii. We are 90% confident the true value of β_1 for the South region is in the interval (285.71, 375.52).
iv. We are 90% confident the true value of β_1 for the West region is in the interval (364.76, 515.87).

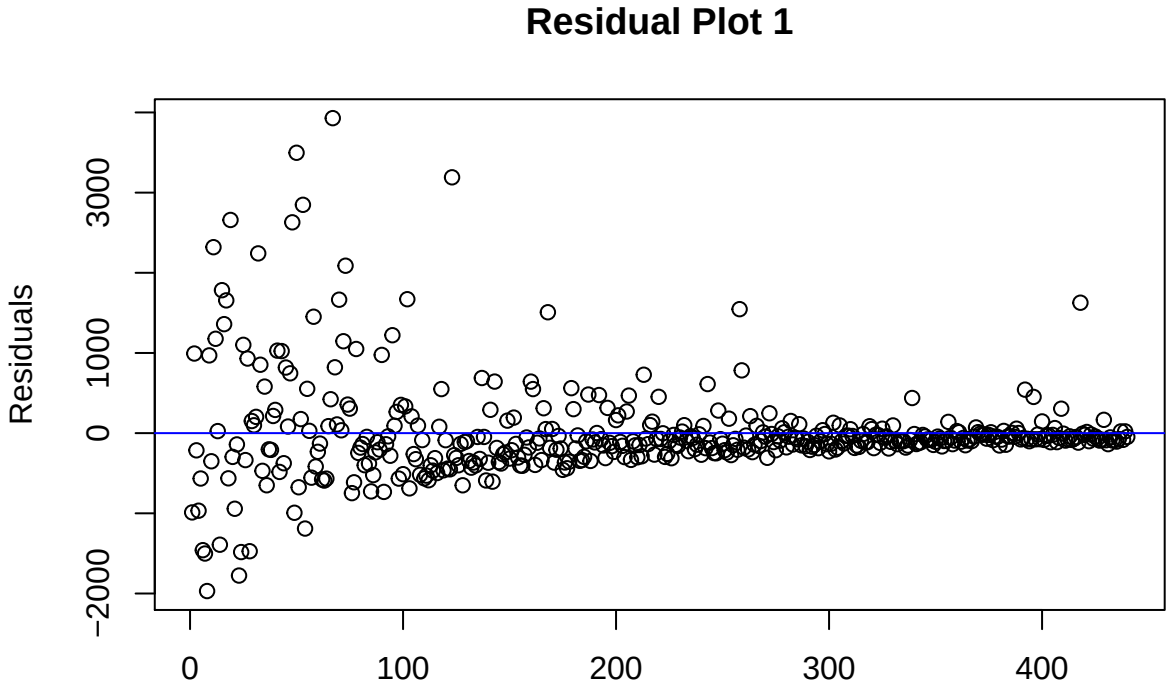
The regression lines for the different regions appear to have different slopes because the intervals do not all overlap.

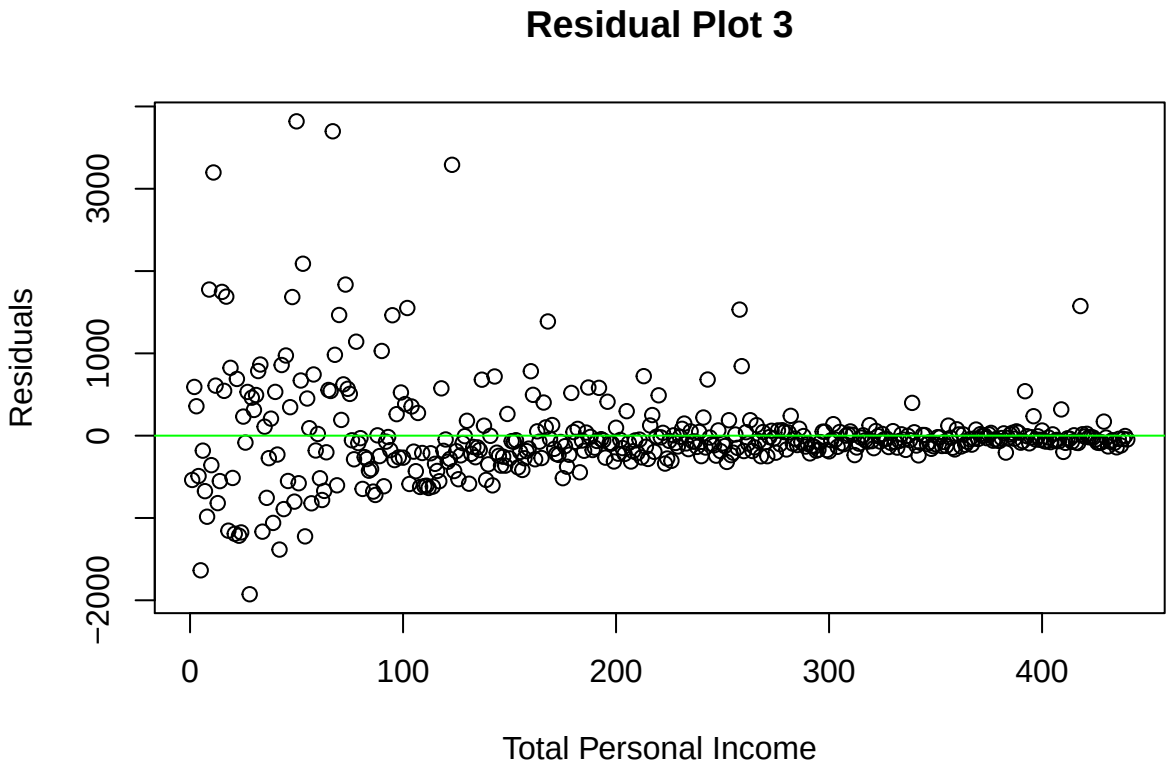
i. For the Northeast Region, with a value of $F^* = 197.7527162$ which is greater than our critical F value of 2.7558677, we reject H_0 at a 90% confidence level ($\alpha = .1$) and conclude there is a linear relationship between per capita income and the percent of the population with a bachelor's degree.
ii. For the North Central Region, $F^* = 76.8264551$ which is greater than our critical F value of 2.7534624, we reject H_0 at a 90% confidence level ($\alpha = .1$) and conclude there is a linear relationship between per capita income and the percent of the population with a bachelor's degree.
iii. For the South, $F^* = 148.4909503$ which is greater than our critical F value of 2.7392749, we reject H_0 at a 90% confidence level ($\alpha = .1$) and conclude there is a linear relationship between per capita income and the percent of the population with a bachelor's degree.
iv. For the West, $F^* = 94.1947705$ which is greater than our critical F value of 2.7736417, we reject H_0 at a 90% confidence level ($\alpha = .1$) and conclude there is a linear relationship between per capita income and the percent of the population with a bachelor's degree.

Part IV - Regression Diagnostics

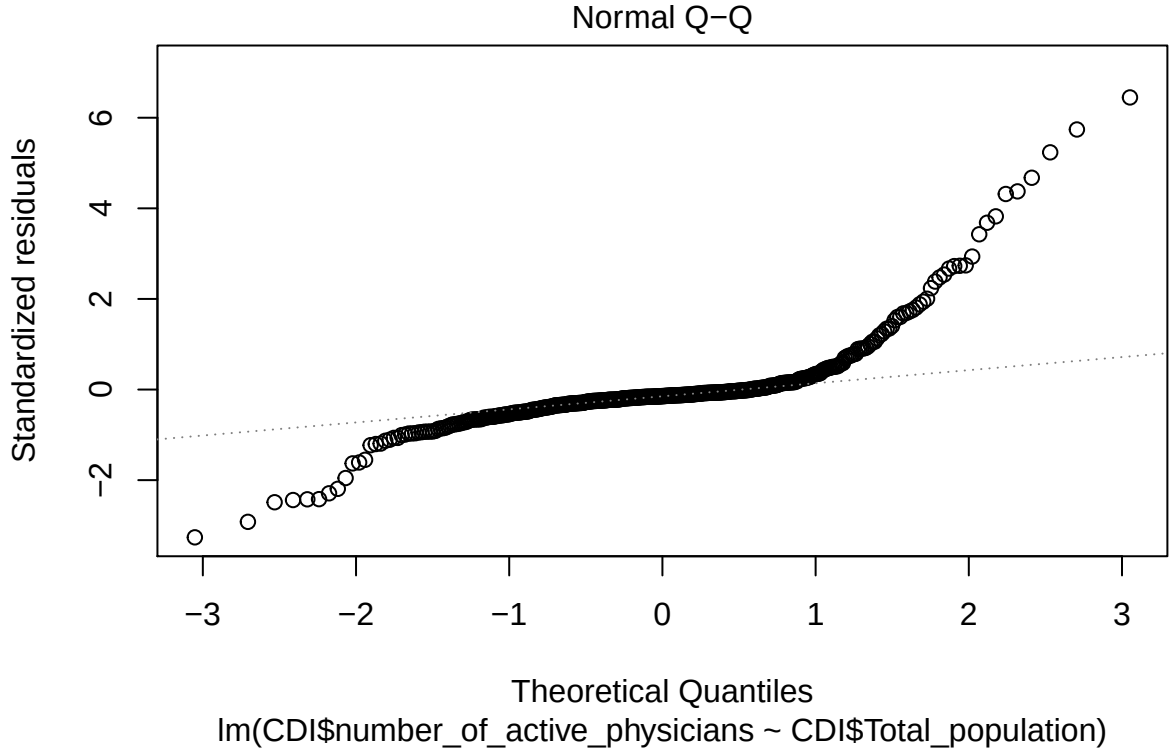
3.25 Refer to the CDI dataset in Appendix C.2 and Project1.43. For each of the three fitted regression models, obtain the residuals and prepare a residual plot against X and a normal probability plot. Summarize your

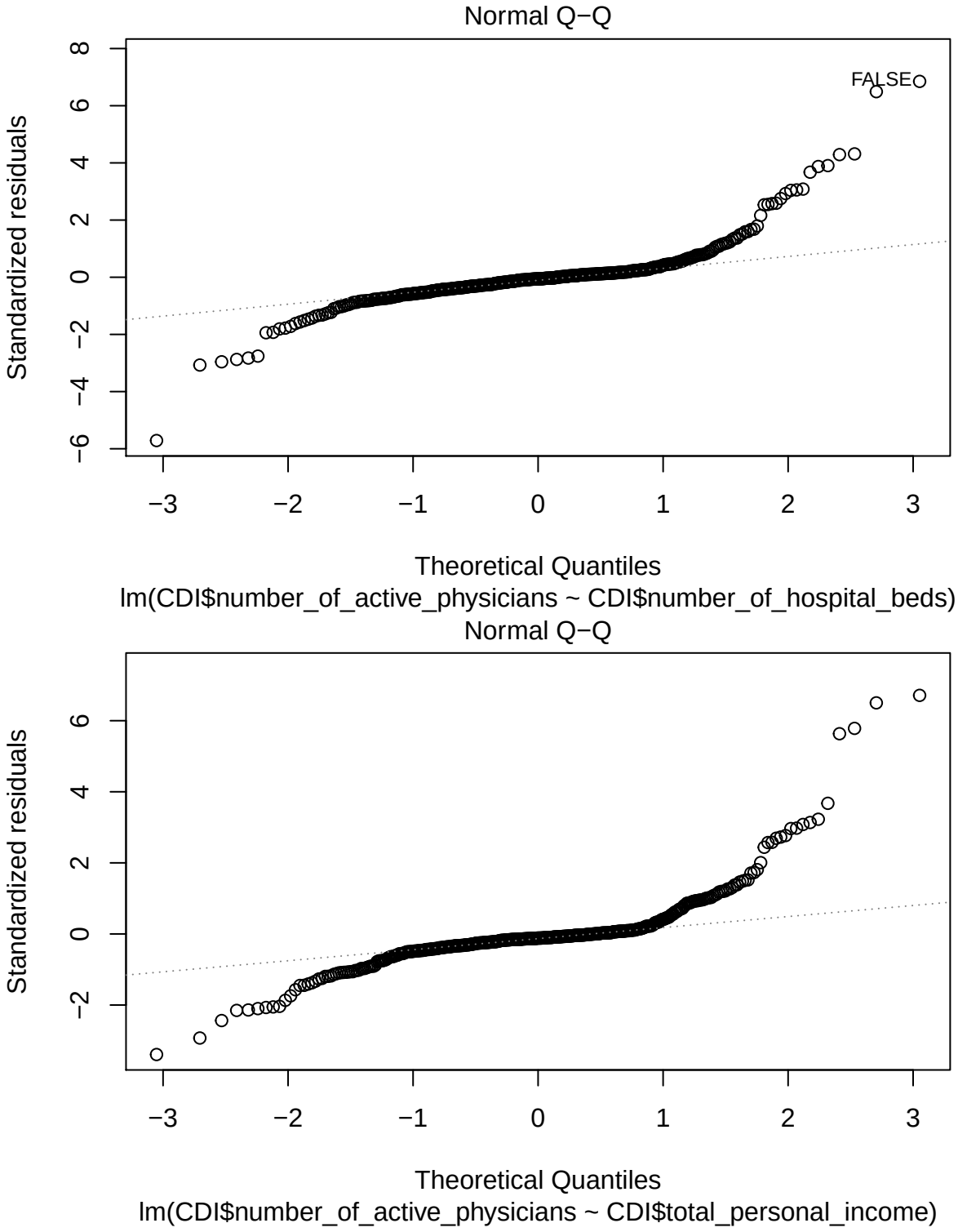
conclusions. Is linear regression model (2.1) more appropriate in one case than in the others?





As total population, number of hospital beds, and total personal income increase, there is less variation in the number of active physicians.





Linear regression doesn't seem to be appropriate for any of the models because the normal probability shows points towards both extreme ends straying from the normal qqline.

Part V - Discussion

When the researchers looked to see whether total population, number of hospital beds, and total personal income were related to the number of active physicians in a country, they found there to be positive linear relationship. Generally, as total population, number of hospital beds, and total personal income increase, the number of active physicians also increase. Number of hospital beds seemed to be the most direct predictor of number of physicians, but the other predictors (total population and total personal income) can also be used to predict the number of active physicians with relative accuracy. The wide range of data appeared to skew the fit of a linear model. While all three variables originally appeared to be good candidates for a linear model, other diagnostic tests showed the variability in the data decreased as the predictor variables increased and the extremes strayed from our normal assumption indicating a nonlinear model would be a better approximation for number of active physician for all three variables.

The research also showed a positive linear relationship between the percent of population holding a bachelor's degree and per capita income in all four US geographic regions. The degree to which an increase in the percentage of the population holding a bachelor's degree affect per capita income differs between the four regions with Northeast having the highest increase (~$522 per increase in bachelor's percentage point), while the North Central region had the smallest increase (~$238 per increase in bachelor's percentage point). While the North Central region has the lowest slope, it has the highest intercept by a margin of approximately 3000. The intercept and slope of this model could indicate the workforce in the region is not as dependent on a bachelor's degree.

The linear regression models mentioned above could potentially be improved for the given data in the CDI data set by using transformations. As noted above, the wide range of data resulted in a significant skew of the linear fit of the models. This conclusion was reached after the use of multiple regression diagnostic tools. This wide range of data could have been altered using a log transformation or other manipulation to reduce the magnitude of the values while still maintaining the integrity of the distribution. By using an appropriate transformation, the researcher might be able to more accurately fit a linear model to the data analyses.

Code Appendix

```
knitr::opts_chunk$set(echo = FALSE, results = 'hide')
#import the data
CDI <- read.table("~/Desktop/Spring 2019/STA 108 /Projects/CDI.txt", quote="\"", comment.char="")
#give proper variable names to the data
names(CDI) <- c("ID", "City", "State", "Land_Area", "Total_population", "percent_of_population_aged_18-3

#Problem 1
#part a
#create linear regression models for all the predictor variables
physiciansxPopulation = lm(CDI$number_of_active_physicians ~ CDI$Total_population)
physiciansxBed = lm(CDI$number_of_active_physicians ~ CDI$number_of_hospital_beds)
physiciansxIncome = lm(CDI$number_of_active_physicians ~ CDI$total_personal_income)

#print the coefficients of the linear regression models
physiciansxPopulation$coefficients
physiciansxBed$coefficients
physiciansxIncome$coefficients

#part b
#create scatterplots for each one

{plot(x = CDI$Total_population,y =  CDI$number_of_active_physicians)
```

```
abline(physiciansxPopulation, col = "blue")}

{plot(x = CDI$number_of_hospital_beds,y =  CDI$number_of_active_physicians)

abline(physiciansxBed, col = "red")}


{plot(x = CDI$total_personal_income,y =  CDI$number_of_active_physicians)

abline(physiciansxIncome, col = "green")
}
#part c
#calculate mse by taking the mean of the squared residuals
mse_Population = summary(physiciansxPopulation)$sigma^2
mse_Bed = summary(physiciansxBed)$sigma^2
mse_Income = summary(physiciansxIncome)$sigma^2
#subset data into geographic regions
geo_NE <- subset(CDI, geographic_region == 1)
geo_NC <- subset(CDI, geographic_region == 2)
geo_S <- subset(CDI, geographic_region == 3)
geo_W <- subset(CDI, geographic_region == 4)

#create regression for each region
per_capita_incomexBachelors_NE = lm(geo_NE$per_capita_income ~ geo_NE$`percent_bachelor's_degrees`)
per_capita_incomexBachelors_NC = lm(geo_NC$per_capita_income ~ geo_NC$`percent_bachelor's_degrees`)
per_capita_incomexBachelors_S = lm(geo_S$per_capita_income ~ geo_S$`percent_bachelor's_degrees`)
per_capita_incomexBachelors_W = lm(geo_W$per_capita_income ~ geo_W$`percent_bachelor's_degrees`)

#print the coefficients for each one to find the regression model and make sure we didn't actually calc
per_capita_incomexBachelors_NE$coefficients
per_capita_incomexBachelors_NC$coefficients
per_capita_incomexBachelors_S$coefficients
per_capita_incomexBachelors_W$coefficients
#calculate MSE for the different regions
mse_NE = summary(per_capita_incomexBachelors_NE)$sigma^2
mse_NC = summary(per_capita_incomexBachelors_NC)$sigma^2
mse_S = summary(per_capita_incomexBachelors_S)$sigma^2
mse_W = summary(per_capita_incomexBachelors_W)$sigma^2
#Problem 2
# extract coeffiecient of determination R2 from linear models
tp.r2 = summary(physiciansxPopulation)$r.squared
hosbeds.r2 = summary(physiciansxBed)$r.squared
totalpers.r2 = summary(physiciansxIncome)$r.squared

# store all values in vector
r2.values = c(tp.r2, hosbeds.r2, totalpers.r2)

# find max correlation coefficient (largest reduction in variability)
solution = max(r2.values)
#Problem 3
#calculating 90% confidence interval for beta 1 for each geographic region
beta1_NE <- confint(per_capita_incomexBachelors_NE, level = .90)
beta1_NC <- confint(per_capita_incomexBachelors_NC, level = .90)
```

```r
beta1_S <- confint(per_capita_incomexBachelors_S, level = .90)
beta1_W <- confint(per_capita_incomexBachelors_W, level = .90)
#running anova test for all regions
anova_NE = aov(per_capita_incomexBachelors_NE, data = geo_NE)
anova_NC = aov(per_capita_incomexBachelors_NC, data = geo_NC)
anova_S = aov(per_capita_incomexBachelors_S, data = geo_S)
anova_W = aov(per_capita_incomexBachelors_W, data = geo_W)

#obtain F* from the summary table and calculate critical F to compare the two
#if F* > critF, reject H_0 and conclude there is a linear relationship
summary(anova_NE)
critf_NE = qf(1-.1, 1, 101)

summary(anova_NC)
critf_NC = qf(1-.1, 1, 106)

summary(anova_S)
critf_S = qf(1-.1, 1, 150)

summary(anova_W)
critf_W = qf(.9, 1, 75)
#obtain residuals

plot(resid(physiciansxPopulation), ylab ="Residuals", xlab ="Total Population", main ="Residual Plot 1"
abline(0,0, col = "blue")
plot(resid(physiciansxBed), ylab ="Residuals", xlab ="Number Hospital Beds", main ="Residual Plot 2")
abline(0,0, col = "red")
plot(resid(physiciansxIncome), ylab ="Residuals", xlab ="Total Personal Income", main ="Residual Plot 3
abline(0,0, col = "green")
#normal probability plots
plot(physiciansxPopulation, which = 2, labels.id = FALSE)
plot(physiciansxBed, which = 2, labels.id = FALSE)
plot(physiciansxIncome, which = 2, labels.id = FALSE)
```