STA 108 Project 1

Mahir Oza, Dylan M. Ang, In Seon Kim

2/6/2022

Background

The United States Census Bureau is a government organization that collects and produces data about the American people and economy. Its mission is to display quality data about the population in all regions of America. Every 10 years, they conduct a population and housing census that includes all residents living in the United States. They not only count the population, but also gather information about different factors in order to analyze the people and economy.

Throughout our project, we analyze the County Demographic Information (CDI) provided by the United States Census Bureau. The data set displays information about 440 populous counties in the United States with 14 different variables to describe the county's economy. The counties range from all throughout the US, from Orange County in California to Wayne County in North Carolina. Various data variables such as land area, total population, number of active physicians, number of hospital beds, total serious crime, percent high school graduates, percent bachelor's degree, etc were gathered for a single county. Some counties with missing data were deleted from the data set. The data reflects the years 1990 and 1992.

We will be using R studio, a known professional software used in statistics, in our research to analyze and graph our data. The purpose of this is to demonstrate and analyze a simple linear regression model for our random variables and display inference methods that can be applied to our model. Data is provided from the textbook "Applied Linear Statistical Models" and our project will contain five parts:

1. Fitting regression models
2. Measuring linear association
3. Inference about regression parameters
4. Regression diagnostics
5. Discussion

Part 1. Fitting Regression Models

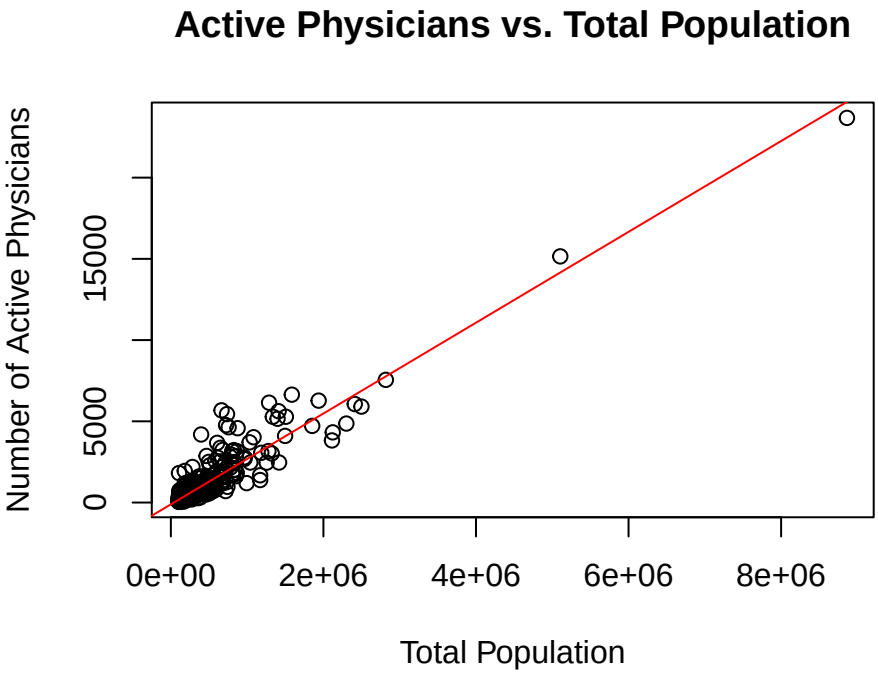
a) Three Predictors for Active Physicians

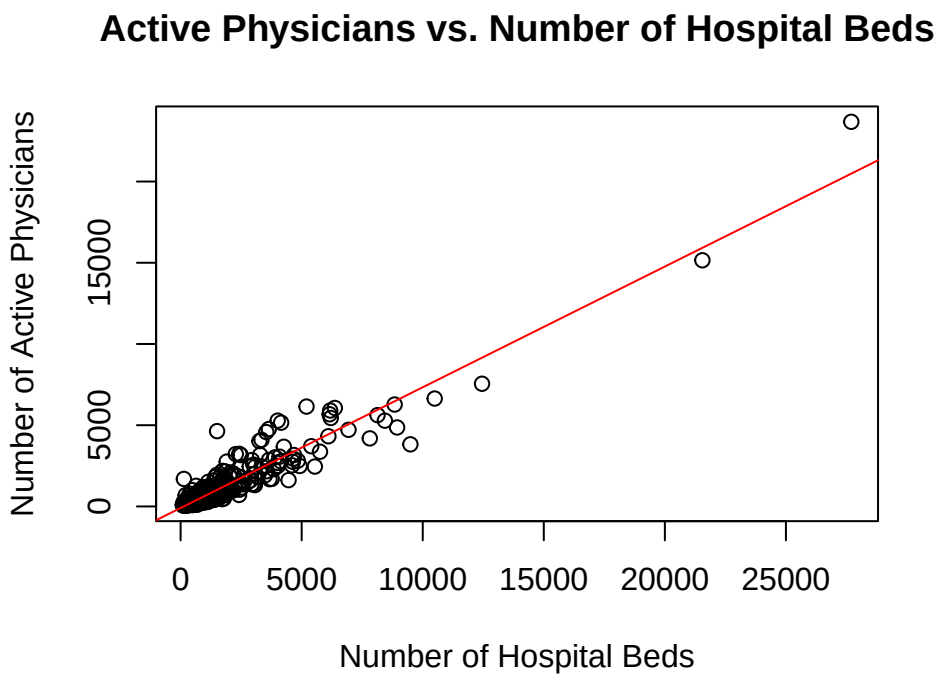
All regression functions appear in the form,

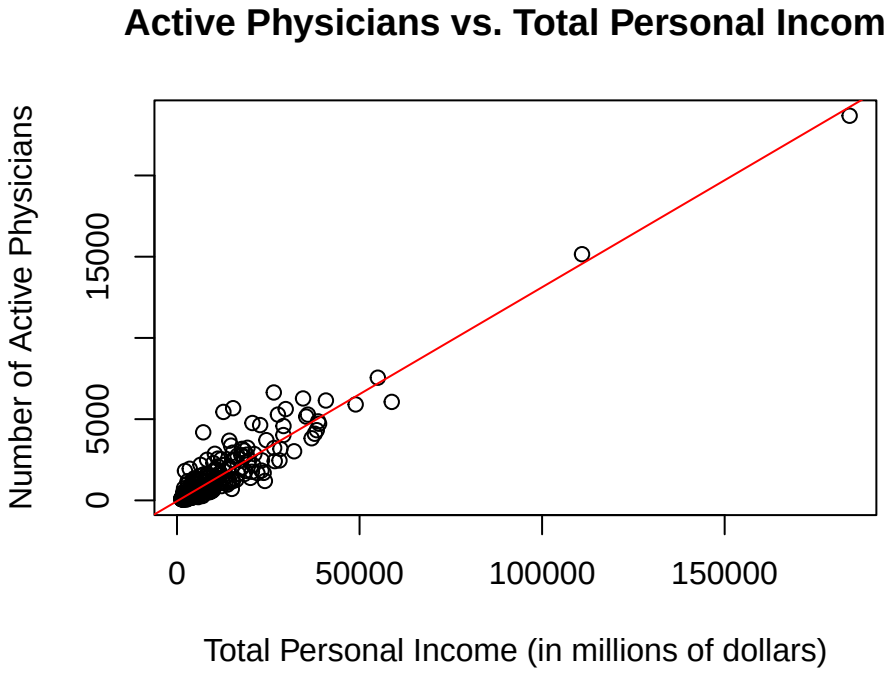
$$Y_i = \hat{\beta}_0 + \hat{\beta}_1 X_i + \epsilon_i \tag{1}$$

Active Physicians vs. Hospital Beds	$Y = -95.9321847 + 0.7431164 X_i$	(2)
Active Physicians vs. Total Population	$Y = -110.6347772 + 0.0027954 X_i$	(3)
Active Physicians vs. Total Personal Income	$Y = -48.3948489 + 0.1317012 X_i$	(4)

b) Plots







We can see that the data itself appears to demonstrate a linear relationship, showing some evidence that the linear regression relation appears to show a good fit for the data. Adding the regression function to the data further shows that the data of all 3 predictor variables follows the behavior of this regression function and further supports a linear regression relationship for all 3 predictor variables for the number of active physicians.

c) MSE Values

```
## [1] "Population: 372203.504917"
```

```
## [1] "Beds: 310191.883540"
```

```
## [1] "Income: 324539.393676"
```

The mean of squared errors of the number of hospital beds as the predictor for number of active physicians in a county is 310,191.883. This is the lower MSE value compared to predictors total personal income, which had an MSE of 324,539.394, and total population of the county, which had the highest MSE of 372,203.505. This shows the mean squared errors in which the number of hospital beds in a county predicting the number of active physicians is lowest and therefore has the lowest variability around the fitted regression line.

Part 2. Measuring Linear Associations

R^2 values

```
## [1] "Population: 0.884067"
```

```
## [1] "Beds: 0.903383"
```

```
## [1] "Income: 0.898914"
```

For predicting the number of active physicians in a county, the R^2 value for total population as the predictor is 0.8841, for number of hospital beds is 0.9034, and for total personal income is 0.8989. Based on these values, the number of hospital beds as the predictor has the largest coefficient of determination meaning that 90.34% of the variation in the number of active physicians is due to introducing number of hospital beds into the regression model. This shows that the number of hospital beds accounts for the largest reduction in variability in the number of active physicians.

Part 3. Inference About Regression Parameters

Confidence Intervals for per capita income (y) vs percent of individuals with atleast bachelor's degree (x) for each region

```
## [1] "NE(1): [460.518, 583.800]"
## [1] "NC(2): [193.486, 283.853]"
## [1] "S (3): [285.708, 375.516]"
## [1] "W (4): [364.758, 515.873]"
```

The regression lines for all 4 regions seem to have pretty large, positive slopes based on their 90% confidence intervals where the lower limit of each interval seems to be greater that 190. However the magnitudes of these large, positive slopes and even the size of the confidence intervals differ slightly based on region. The NC region has an interval of [193.486, 283.853], the S region has an interval of [285.708, 375.516], the W region has an interval of [364.758,515,873], and the NE region has an interval of [460.518,583.800]. These intervals represent that we are 90% confident the true slope of per capita income vs percent of individuals with at least a bachelor's degree lies in the interval for each respective region.

NE F-test

per capita income (y) vs percent of individuals with at least bachelor's degree (x)

$$H_0 : \beta_1 = 0$$
$$H_1 : \beta_1 \neq 0$$

$$F = 197.7527162 \tag{5}$$
$$\text{p-value} = 0 \tag{6}$$

For significance level $\alpha = 0.01$ the F-statistic for the NE region, 197.753, shows an incredibly small p-value where p-value $= 0 < 0.01 = \alpha$, meaning that H_0 is rejected. This conclusion as well as the fact that the F-statistic is very large, shows that regression for per capita income vs percent of individuals with a bachelor's degree in the NE region is useful.

Source	df	ss	ms	f_val	p_val
Regression	1	1450517671	1450517671	197.7527	0
Error	101	740835765	7335008		
Total	102	2191353436			

NC F-test per capita income (y) vs percent of individuals with at least bachelor's degree (x)

$$H_0 : \beta_1 = 0$$
$$H_1 : \beta_1 \neq 0$$

$$F = 76.8264551 \tag{7}$$
$$\text{p-value} = 3.3417713 \times 10^{-14} \tag{8}$$

For significance level $\alpha = 0.01$ the F-statistic for the NC region, 76.826, shows a small p-value such that p-value $= 3.3417713 \times 10^{-14} < 0.01 = \alpha$, meaning that H_0 is rejected. This conclusion as well as the fact

that the F-statistic is fairly large, shows that regression for per capita income vs percent of individuals with a bachelor's degree in the NC region is useful.

Source	df	ss	ms	f_val	p_val
Regression	1	338907694	338907694	76.82646	0
Error	106	467602149	4411341		
Total	107	806509843			

S F-test

per capita income (y) vs percent of individuals with at least bachelor's degree (x)

$$H_0 : \beta_1 = 0$$
$$H_1 : \beta_1 \neq 0$$

$$F = 148.4909503 \tag{9}$$
$$\text{p-value} = 0 \tag{10}$$

For significance level $\alpha = 0.01$ the F-statistic for the S region, 148.491, shows a very small p-value where p-value $= 0 < 0.01 = \alpha$, meaning that H_0 is rejected. This conclusion as well as the fact that the F-statistic is very large, shows that regression for per capita income vs percent of individuals with a bachelor's degree in the S region is useful.

Source	df	ss	ms	f_val	p_val
Regression	1	1109873245	1109873245	148.491	0
Error	150	1121152411	7474349		
Total	151	2231025656			

W F-test

per capita income (y) vs percent of individuals with at least bachelor's degree (x)

$$H_0 : \beta_1 = 0$$
$$H_1 : \beta_1 \neq 0$$

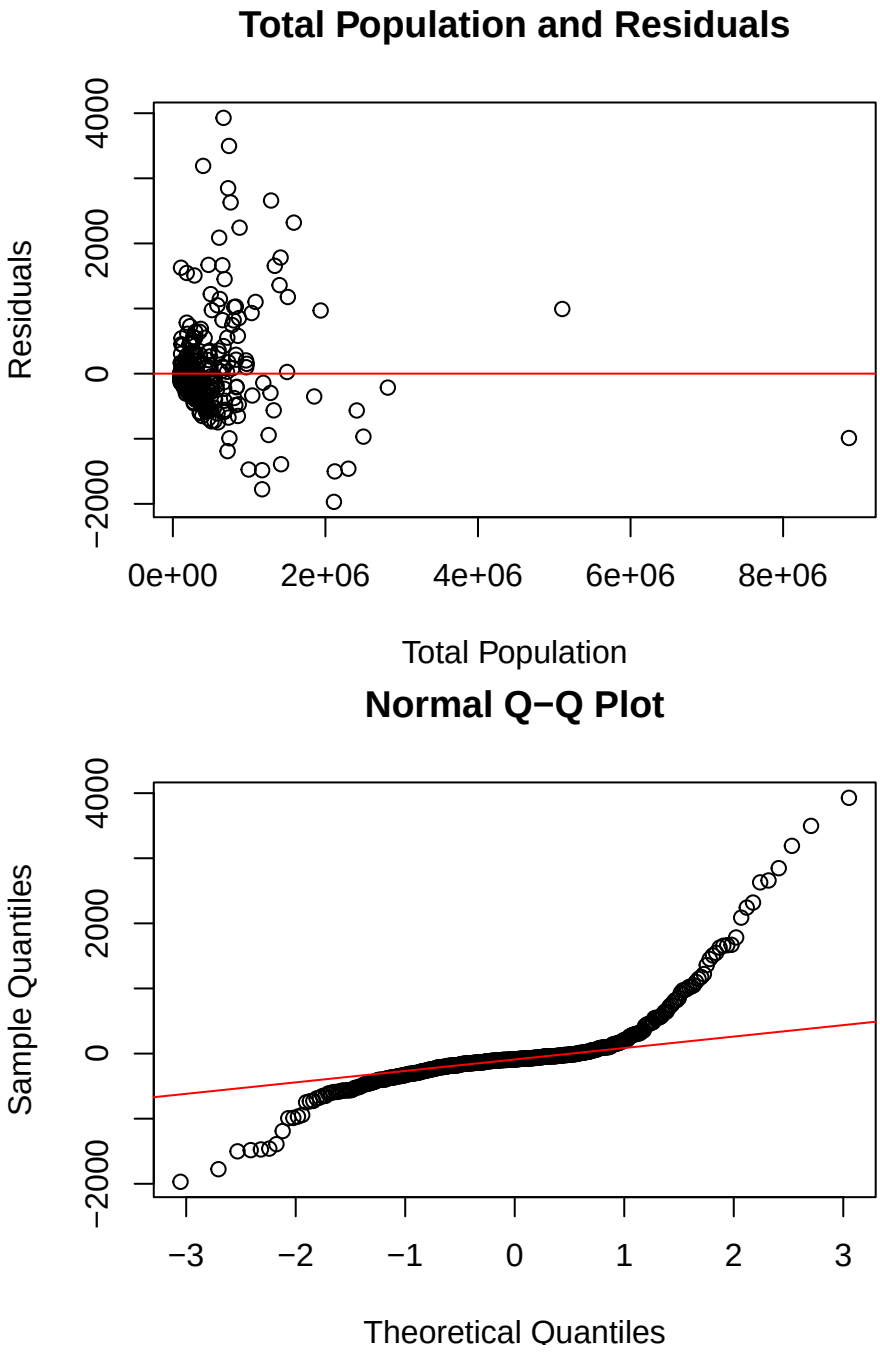
$$F = 94.1947705 \tag{11}$$
$$\text{p-value} = 6.8833828 \times 10^{-15} \tag{12}$$

For significance level $\alpha = 0.01$ the F-statistic for the W region, 94.195, shows a small p-value such that p-value $= 6.8833828 \times 10^{-15} < 0.01 = \alpha$, meaning that H_0 is rejected. This conclusion as well as the fact that the F-statistic is fairly large, shows that regression for per capita income vs percent of individuals with a bachelor's degree in the W region is useful.

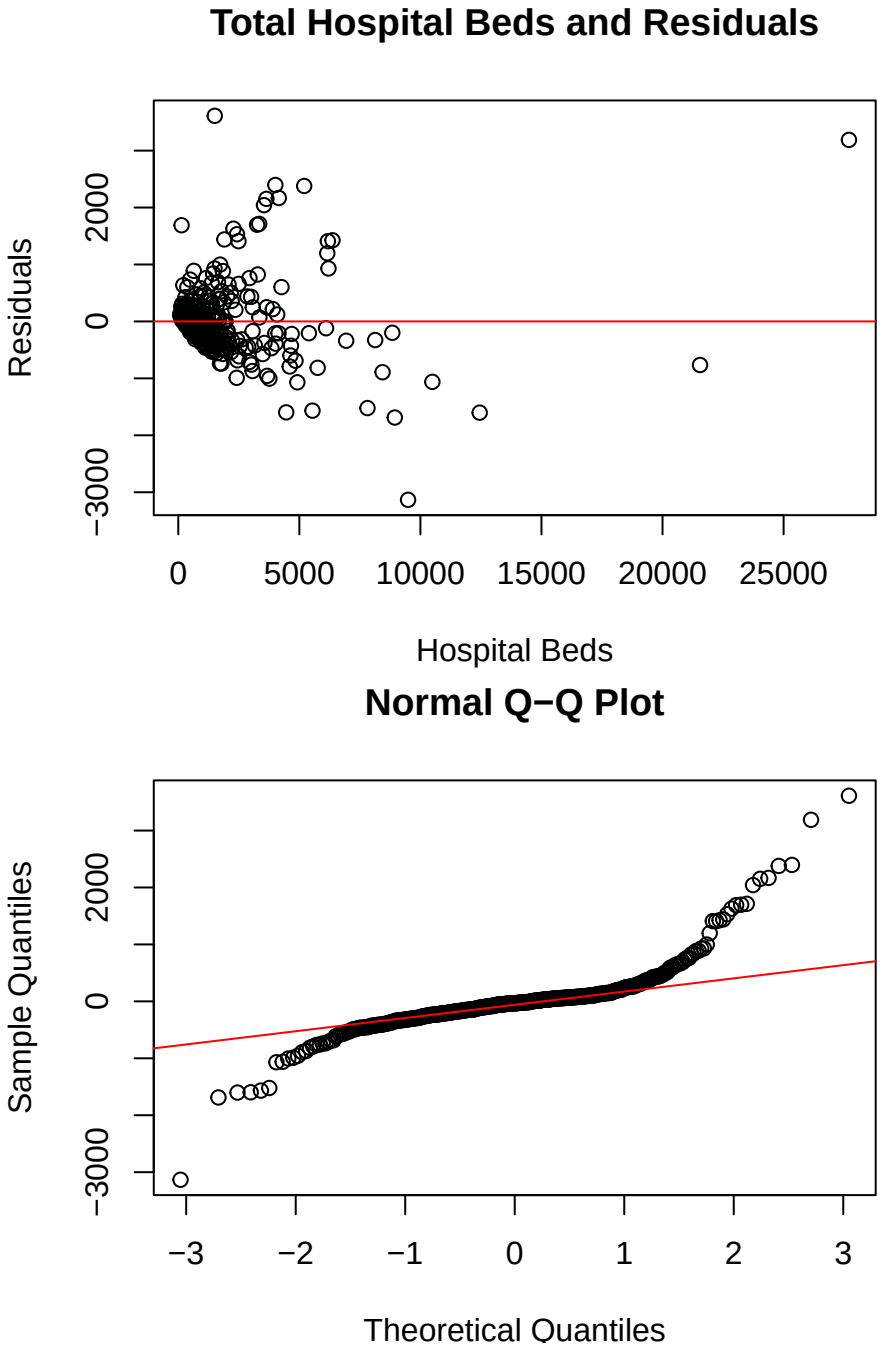
Source	df	ss	ms	f_val	p_val
Regression	1	773745787	773745787	94.19477	0
Error	75	616073841	8214318		
Total	76	1389819629			

Part 4. Regression Diagnostics

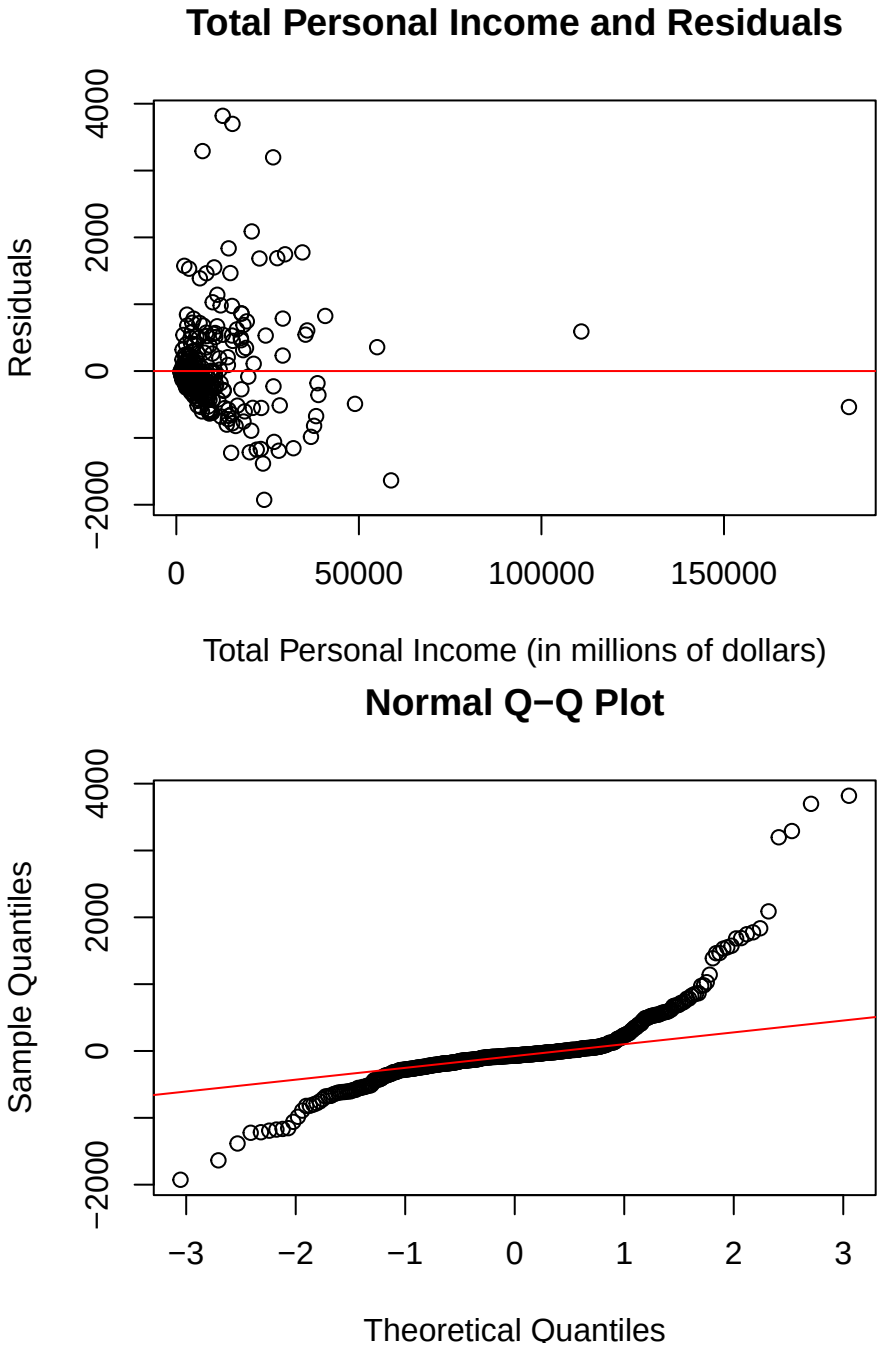
Plot 1: Number of Active Physicians (Y) vs Total Population (Xp)



Plot 2: Number of Active Physicians (Y) vs Number of Hospital Beds (Xh)



Plot 3: Number of Active Physicians (Y) vs Total Personal Income (Xi)



The residual plot for all 3 predictor variables as a function of the number of active physicians shows no distinguishable mathematical trend nor any pattern. Also the sum of all the residuals for the 3 predictor variables are extremely small values, close to 0. These show that a linear regression model is appropriate in showing the relationship of the data. In addition, we can see that the data itself appears to demonstrate a linear relationship. Adding the regression function to the data further shows that the data of all 3 predictor variables follows the behavior of this regression function and further supports a linear regression relationship for all 3 predictor variables for the number of active physicians. While all 3 QQ plots demonstrate a similar form, it seems that the QQ plot for the number of active physicians vs number of hospital beds, shows the best linear trend in the plot compared to the other 2 predictor variables. This helps shows that this predictor variable is slightly more appropriate in determining a regression relationship for number of active physicians.

Part 5. Discussion

We discussed three possible sources of error that may have been present in our observation data. First, there are many counties who are missing from the data set because they may have lacked one or more of the values. In addition, the data set is described as consisting of 440 of the most populous counties. Therefore, the data set may not be representative of the entire United States, especially less populous counties. Finally, we only compared the total number of doctors to three predictor variables, but there could be other better variables that we didn't consider.

As for improving our models, we can make the model more representative of the United States, by taking random samples of counties rather than only picking the most populous, reducing bias. To make the bachelors degrees to per capita income model more accurate, we could split up the divisions by state rather than region. This would be more precise and give us more opportunities to see patterns.

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
)
# import data
CDI = read.table("./CDI.txt")

# define active physicians columns
phys = CDI$V8

# define predictor variables total pop,
#         hospital beds, and personal income
pop = CDI$V5
bed = CDI$V9
inc = CDI$V16
# Physicians vs Population
fit_pop = lm(phys ~ pop)
b0_pop = fit_pop$coefficients[1]
b1_pop = fit_pop$coefficients[2]

# Physicians vs Beds
fit_bed = lm(phys ~ bed)
b0_bed = fit_bed$coefficients[1]
b1_bed = fit_bed$coefficients[2]

# Physicians vs Income
fit_inc = lm(phys ~ inc)
b0_inc = fit_inc$coefficients[1]
b1_inc = fit_inc$coefficients[2]
plot(x=pop,y=phys,
     xlab="Total Population", ylab="Number of Active Physicians",
     main="Active Physicians vs. Total Population")
abline(fit_pop,col="red")

plot(x=bed,y=phys,
     xlab="Number of Hospital Beds", ylab="Number of Active Physicians",
     main="Active Physicians vs. Number of Hospital Beds")
abline(fit_bed,col="red")

plot(x=inc,y=phys,
     xlab="Total Personal Income (in millions of dollars)", ylab="Number of Active Physicians",
     main="Active Physicians vs. Total Personal Income")
abline(fit_inc,col="red")
E_pop = phys - (b0_pop + (b1_pop * pop))
E_bed = phys - (b0_bed + (b1_bed * bed))
E_inc = phys - (b0_inc + (b1_inc * inc))
```

```
n_pop = length(pop)
n_bed = length(bed)
n_inc = length(inc)
SSE_p = sum((E_pop)^2)
SSE_b = sum((E_bed)^2)
SSE_i = sum((E_inc)^2)
MSE_p = SSE_p/(n_pop - 2)
MSE_b = SSE_b/(n_bed - 2)
MSE_i = SSE_i/(n_inc - 2)
sprintf("Population: %f", MSE_p)
sprintf("Beds: %f", MSE_b)
sprintf("Income: %f", MSE_i)
fit_pop_sum = summary(fit_pop)
fit_bed_sum = summary(fit_bed)
fit_inc_sum = summary(fit_inc)
r2p = fit_pop_sum$r.squared
r2b = fit_bed_sum$r.squared
r2i = fit_inc_sum$r.squared
sprintf("Population: %f", r2p)
sprintf("Beds: %f", r2b)
sprintf("Income: %f", r2i)

# split region
# V12: % Bachelors and V15: per capita income
region = CDI$V17 == "1"
NE = CDI[region, c("V12", "V15")]

region = CDI$V17 == "2"
NC = CDI[region, c("V12", "V15")]

region = CDI$V17 == "3"
S = CDI[region, c("V12", "V15")]

region = CDI$V17 == "4"
W = CDI[region, c("V12", "V15")]

# linear models
fit_NE = lm(NE$V15 ~ NE$V12)
fit_NC = lm(NC$V15 ~ NC$V12)
fit_S = lm(S$V15 ~ S$V12)
fit_W = lm(W$V15 ~ W$V12)

# betas
b0_NE = fit_NE$coefficients[1]
b0_NC = fit_NC$coefficients[1]
b0_S = fit_S$coefficients[1]
b0_W = fit_W$coefficients[1]

b1_NE = fit_NE$coefficients[2]
b1_NC = fit_NC$coefficients[2]
b1_S = fit_S$coefficients[2]
b1_W = fit_W$coefficients[2]
```

```
# SSEs
SSE_NE = sum( ( NE$V15 - (b0_NE + b1_NE * NE$V12) )^2 )
SSE_NC = sum( ( NC$V15 - (b0_NC + b1_NC * NC$V12) )^2 )
SSE_S = sum( ( S$V15 - (b0_S + b1_S * S$V12) )^2 )
SSE_W = sum( ( W$V15 - (b0_W + b1_W * W$V12) )^2 )

# MSEs
MSE_NE = SSE_NE / (nrow(NE) - 2)
MSE_NC = SSE_NC / (nrow(NC) - 2)
MSE_S = SSE_S / (nrow(S) - 2)
MSE_W = SSE_W / (nrow(W) - 2)

# t vals
alpha = 0.1
t_NE = qt(1 - alpha/2, nrow(NE) - 2)
t_NC = qt(1 - alpha/2, nrow(NC) - 2)
t_S = qt(1 - alpha/2, nrow(S) - 2)
t_W = qt(1 - alpha/2, nrow(W) - 2)

# SEs
SE_NE = sqrt( MSE_NE / sum( (NE$V12 - mean(NE$V12) )^2) )
SE_NC = sqrt( MSE_NC / sum( (NC$V12 - mean(NC$V12) )^2) )
SE_S = sqrt( MSE_S / sum( (S$V12 - mean(S$V12) )^2) )
SE_W = sqrt( MSE_W / sum( (W$V12 - mean(W$V12) )^2) )

#  CIs
NE_up = b1_NE + t_NE * SE_NE
NE_lo = b1_NE - t_NE * SE_NE

NC_up = b1_NC + t_NC * SE_NC
NC_lo = b1_NC - t_NC * SE_NC

S_up = b1_S + t_S * SE_S
S_lo = b1_S - t_S * SE_S

W_up = b1_W + t_W * SE_W
W_lo = b1_W - t_W * SE_W

# print CIs
sprintf("NE(1): [%.3f, %.3f]", NE_lo, NE_up)
sprintf("NC(2): [%.3f, %.3f]", NC_lo, NC_up)
sprintf("S (3): [%.3f, %.3f]", S_lo, S_up)
sprintf("W (4): [%.3f, %.3f]", W_lo, W_up)
# SSRs
SSR_NE = sum( ( (b0_NE + b1_NE * NE$V12) - mean(NE$V15) )^2 )
SSR_NC = sum( ( (b0_NC + b1_NC * NC$V12) - mean(NC$V15) )^2 )
SSR_S = sum( ( (b0_S + b1_S * S$V12) - mean(S$V15) )^2 )
SSR_W = sum( ( (b0_W + b1_W * W$V12) - mean(W$V15) )^2 )

# MSRs
MSR_NE = SSR_NE
MSR_NC = SSR_NC
MSR_S = SSR_S
```

```
MSR_W = SSR_W

# F-statistics
F_NE = MSR_NE/MSE_NE
F_NC = MSR_NC/MSE_NC
F_S = MSR_S/MSE_S
F_W = MSR_W/MSE_W
anova_table <- function(region, ssr, sse, msr, mse, f, p) {
  antable = data.frame(Source = c("Regression", "Error", "Total"),
                       df = c(1, nrow(region) - 2, nrow(region) - 1),
                       ss = c(ssr, sse, ssr + sse),
                       ms = c(msr, mse, NA),
                       f_val = c(f, NA, NA),
                       p_val = c(p, NA, NA))
  return(antable)
}

p_NE = 1-pf(F_NE,1,nrow(NE)-2)
library(knitr)
kable(anova_table(NE, SSR_NE, SSE_NE, MSR_NE, MSE_NE, F_NE, p_NE))
p_NC = 1-pf(F_NC,1,nrow(NC)-2)
kable(anova_table(NC, SSR_NC, SSE_NC, MSR_NC, MSE_NC, F_NC, p_NC))
p_S = 1-pf(F_S,1,nrow(S)-2)
kable(anova_table(S, SSR_S, SSE_S, MSR_S, MSE_S, F_S, p_S))
p_W = 1-pf(F_W,1,nrow(W)-2)
kable(anova_table(W, SSR_W, SSE_W, MSR_W, MSE_W, F_W, p_W))
#Residual Plot
plot(CDI$V5, fit_pop$residuals,
     xlab="Total Population", ylab="Residuals",
     main="Total Population and Residuals")
abline(h=0, col = 'red')

#Normal Probability Plot
qqplot = qqnorm(fit_pop$residuals)
qqline(fit_pop$residuals, col='red')
#Residual Plot
plot(CDI$V9, fit_bed$residuals,
     xlab="Hospital Beds", ylab="Residuals",
     main="Total Hospital Beds and Residuals")
abline(h=0, col='red')

#Normal Probability Plot
qqplot = qqnorm(fit_bed$residuals)
qqline(fit_bed$residuals, col='red')
plot(CDI$V16, fit_inc$residuals,
     xlab="Total Personal Income (in millions of dollars)", ylab = "Residuals",
     main="Total Personal Income and Residuals")
abline(h=0, col='red')

#Normal Probability Plot
qqplot = qqnorm(fit_inc$residuals)
qqline(fit_inc$residuals, col='red')
```