STA 108 Project 1

Mahir Oza, Dylan M. Ang, In Seon Kim

1/31/2022

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:

- Fitting regression models
- Measuring linear association
- Inference about regression parameters
- Regression diagnostics
- Discussion

Part 1 Fitting Regression Models

a) Three predictors for Active Physicians

All regression functions appear in the form,

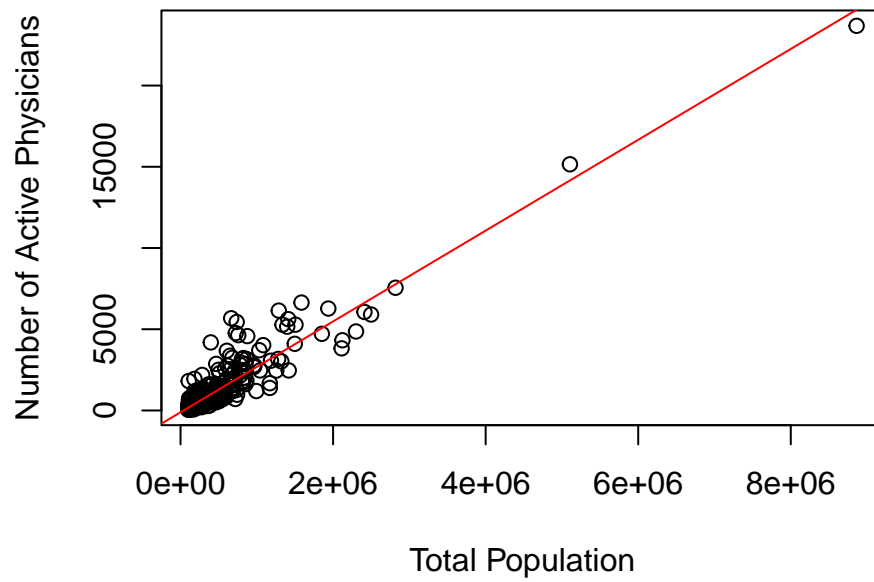$$Y_i = \hat{\beta}_0 + \hat{\beta}_1 X_i + \epsilon_i \tag{1}$$

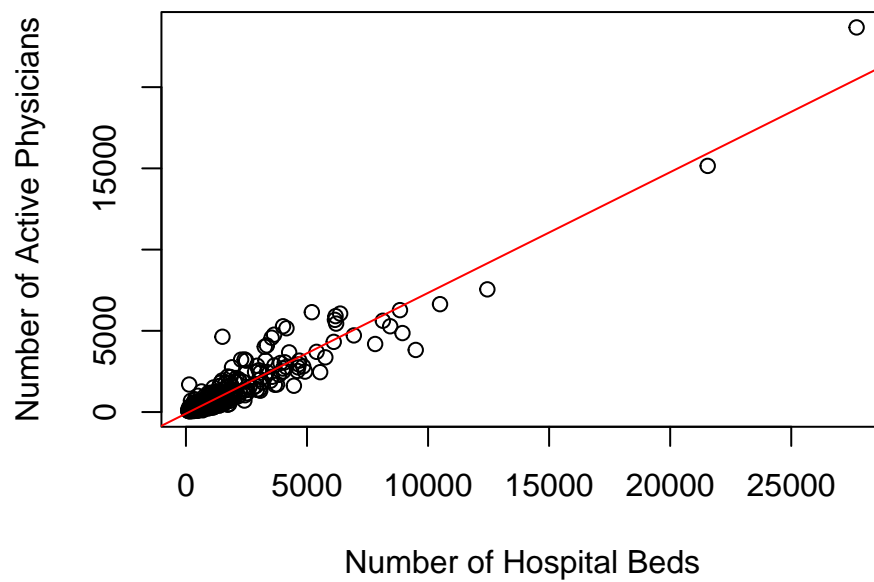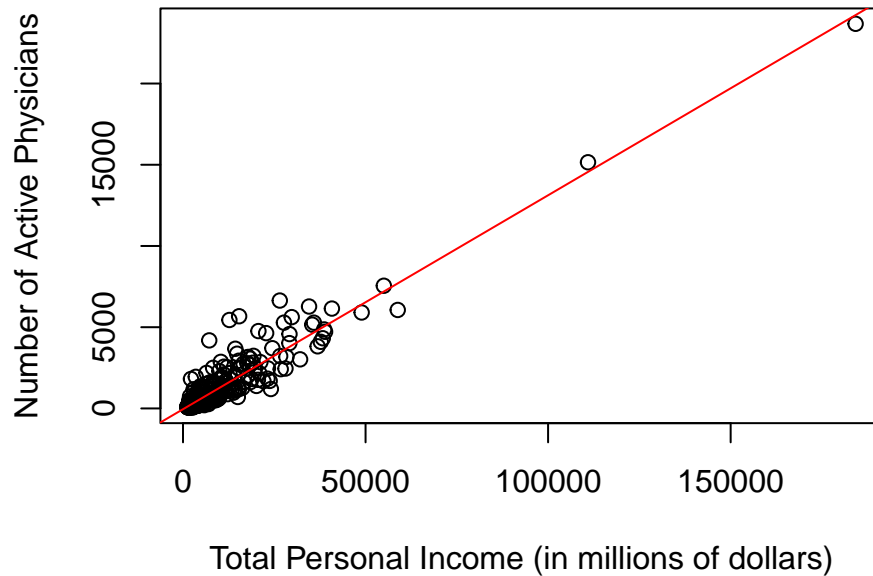|		
Active Physicians vs. Hospital Beds	$Y = -95.9321847 + 0.7431164X_i$	(2)
Active Physicians vs. Total Population	$Y = -110.6347772 + 0.0027954X_i$	(3)
Active Physicians vs. Total Personal Income	$Y = -48.3948489 + 0.1317012X_i$	(4)

Active Physicians vs. Total Population

Active Physicians vs. Number of Hospital Beds

Active Physicians vs. Total Personal Income

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.
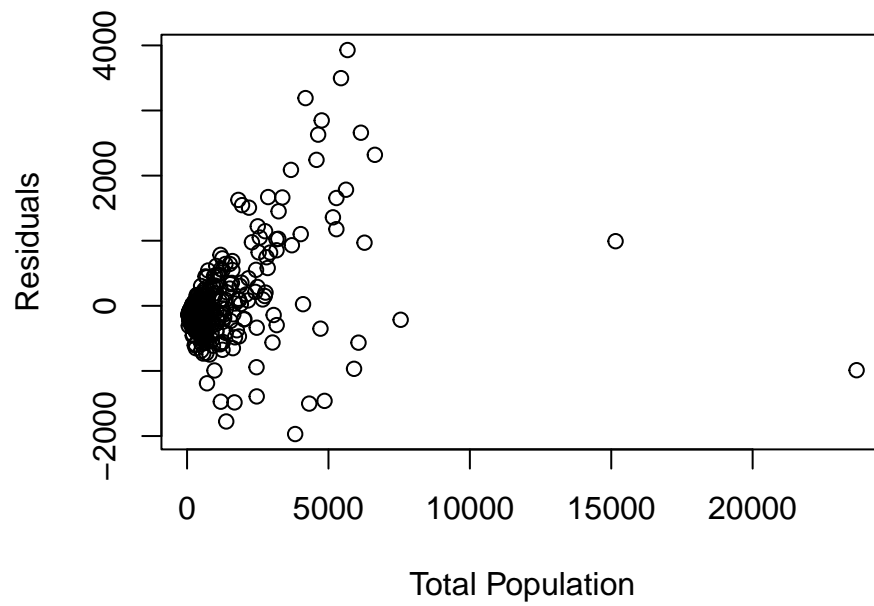
Part 3 Inference about regression parameters

```
## [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]"
```
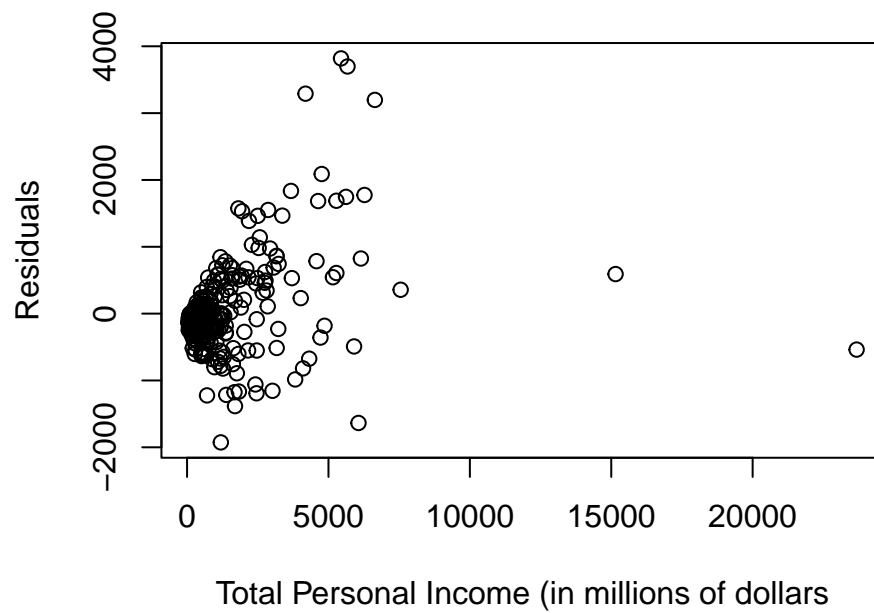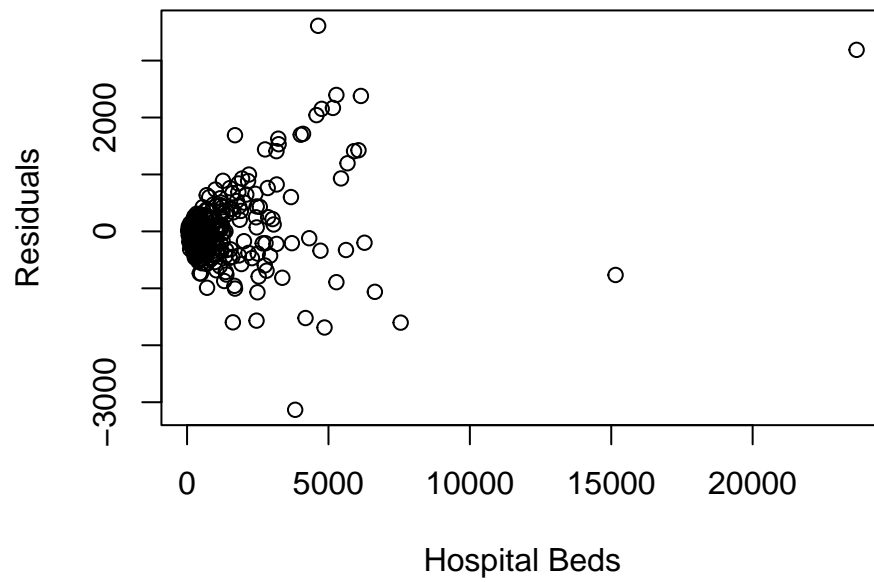
Residuals

Hospital Beds

Total Personal Income (in millions of dollars

```
## [1] "Population Sum of Residuals: 5.846914e-10"

## [1] "Hospital Beds Sum of Residuals : 9.003145e-10"

## [1] "Income Sum of Residuals : 1.212669e-09"
```

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 to these arguments for a linear regression relationship being a good fit for the data, 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.

Appendix

```r
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",
```

```r
      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)
```

```r
# 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)
# Residuals Plot
plot(x=phys,y=E_pop,
     xlab="Total Population", ylab="Residuals")
plot(x=phys,y=E_bed,
     xlab="Hospital Beds", ylab="Residuals")
plot(x=phys,y=E_inc,
     xlab="Total Personal Income (in millions of dollars", ylab="Residuals")
sprintf("Population Sum of Residuals: %e", sum(E_pop))
sprintf("Hospital Beds Sum of Residuals : %e", sum(E_bed))
sprintf("Income Sum of Residuals : %e", sum(E_inc))
```