STA 108 Project 1

Mahir Oza, Dylan M. Ang, In Seon Kim

1/31/2022

Part 1 1.43

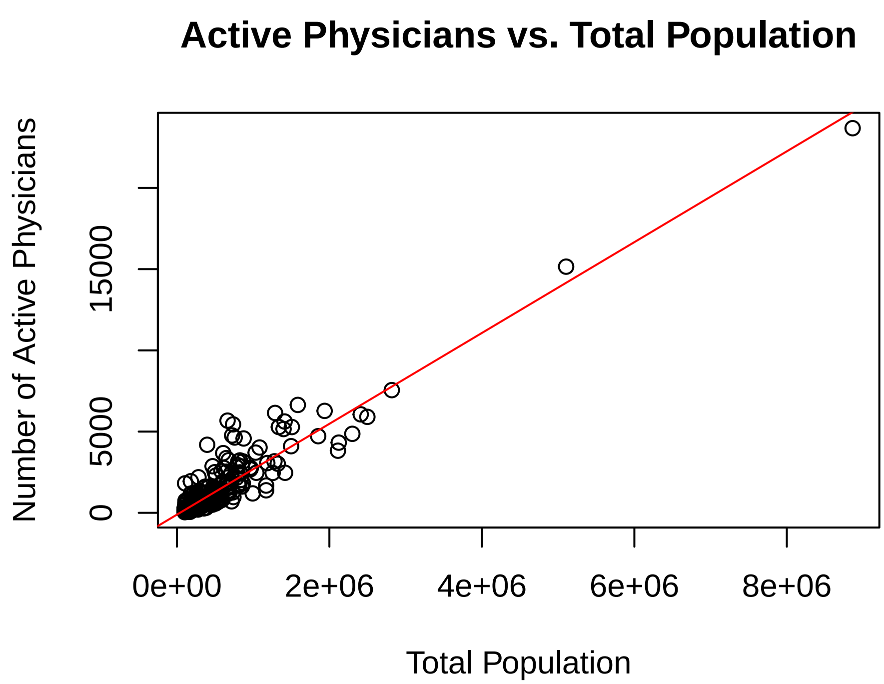
a) Three predictors for Active Physicians

All regression functions appear in the form,

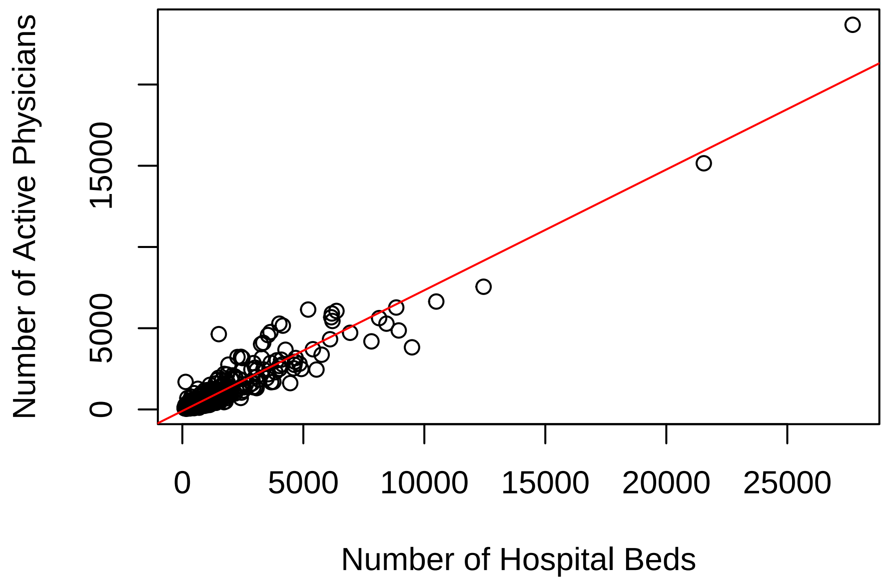
$$Y_i = \hat{\beta}_0 + \hat{\beta}_1 X_i + \epsilon_i \tag{1}$$

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)

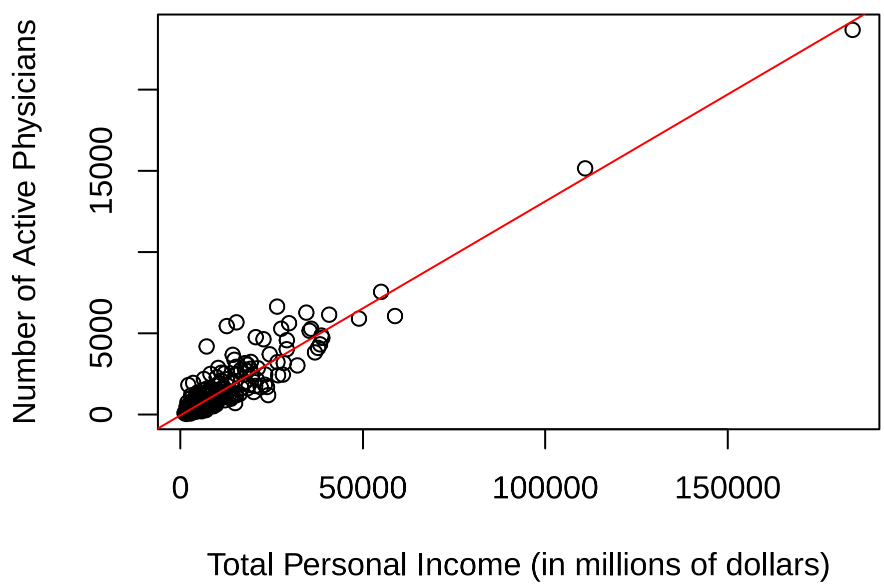
b) Plots



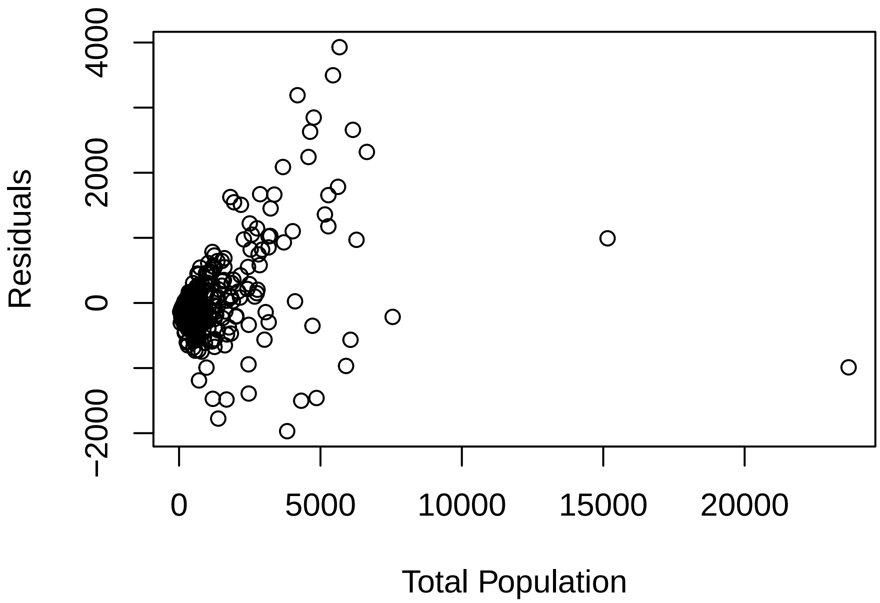
Active Physicians vs. Number of Hospital Beds



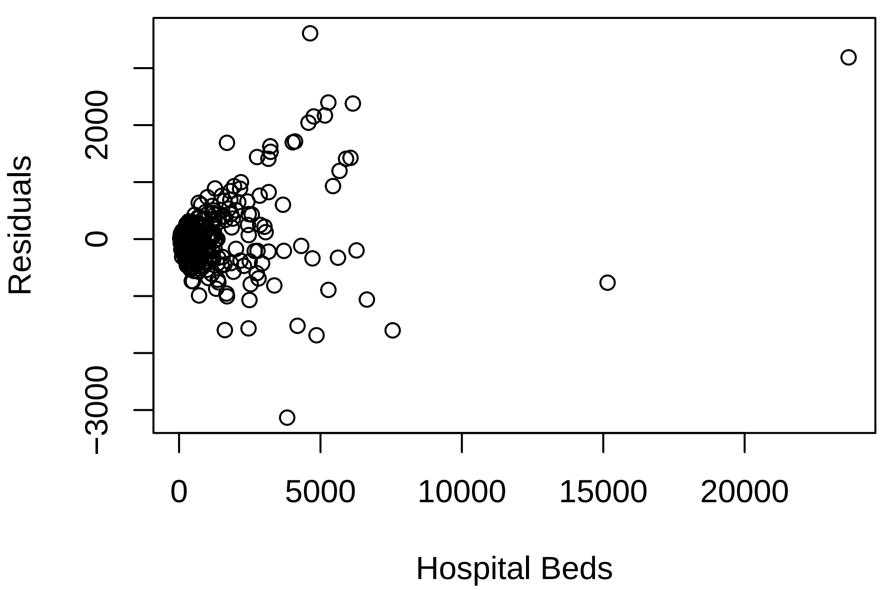
Active Physicians vs. Total Personal Income



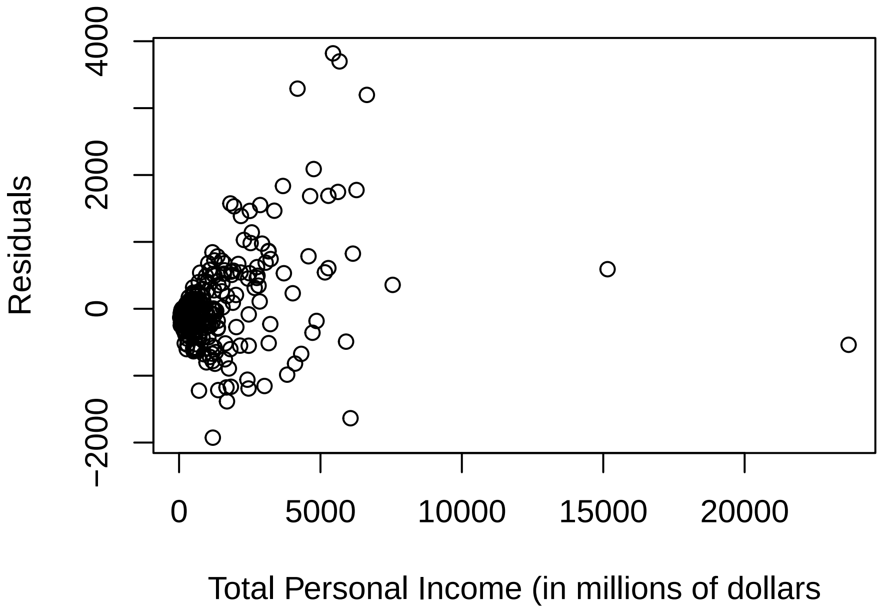
Residuals



Residuals
4000
2000
0
−2000
0
5000
10000
15000
20000
Total Population



Residuals
2000
0
−3000
0
5000
10000
15000
20000
Hospital Beds




```
## [1] "Population SSE: 5.846914e-10"

## [1] "Hospital Beds SSE : 9.003145e-10"

## [1] "Income SSE : 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.

c. MSE

The mean of squared errors of the number of hospital beds as the predictor for number of active physicians in a county, is 310,191.88. This is the lower MSE value compared to predictors total personal income, which had an MSE of 324,539.39, and total population of the county, which had the highest MSE of 372,203.50. 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

Rˆ2

For predicting the number of active physicians in a county, the rˆ2 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.

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))
```

```r
# 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 SSE: %e", sum(E_pop))
sprintf("Hospital Beds SSE : %e", sum(E_bed))
sprintf("Income SSE : %e", sum(E_inc))
npop=length(pop)
nbed=length(bed)
ninc=length(inc)
SSEp=sum((E_pop)^2)
SSEb=sum((E_bed)^2)
SSEi=sum((E_inc)^2)
MSEp=SSEp/(npop-2)
MSEb=SSEb/(nbed-2)
MSEi=SSEi/(ninc-2)
popFit2=summary(fit_pop)
bedFit2=summary(fit_bed)
incFit2=summary(fit_inc)
r2p=popFit2$r.squared
r2b=bedFit2$r.squared
r2i=incFit2$r.squared
```