STA 108 Project 1

Noe Velasquez, ______, and _________

2023-01-27

Introduction

1

```
# 1
# a
cdi = read.table("CDI.txt")
colnames(cdi) <- c("id_num", "county",
                   "state", "land_area",
                   "pop_total", "pop_18_34",
                   "pop_65_old", "active_physicians",
                   "hospital_beds", "serious_crimes",
                   "pct_hsgrad", "pct_bachelors",
                   "pct_poverty", "pct_unemp",
                   "income_percap", "income_total",
                   "region")
model_1 = lm(active_physicians ~ pop_total, data = cdi)
model_2 = lm(active_physicians ~ hospital_beds, data = cdi)
model_3 = lm(active_physicians ~ income_total, data = cdi)
```

a

The estimated regression functions are:

1. The number of active physicians in relation to total population is estimated by $\hat{Y}$ =-110.63478+0.0028X.
2. The number of active physicians in relation to number of hospital beds is estimated by $\hat{Y}$ =-95.93218+0.74312X.
3. The number of active physicians in relation to total personal income is estimated by $\hat{Y}$ =-48.39485+0.1317X.

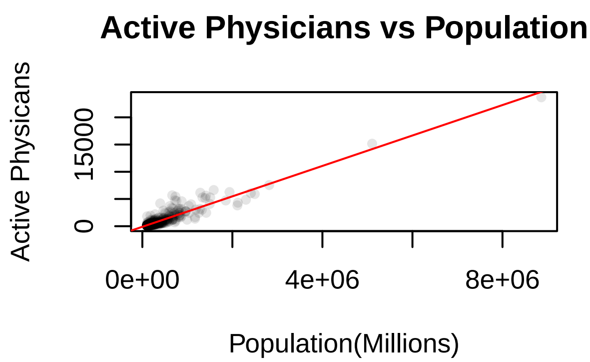
b

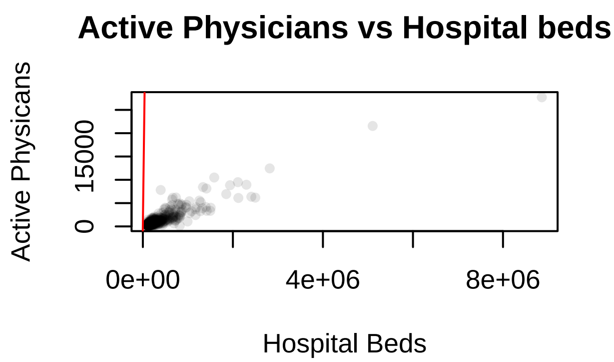
```
#b
par(mfrow = c(2,2))
plot(cdi$pop_total, cdi$active_physicians,
    main = "Active Physicians vs Population",
```

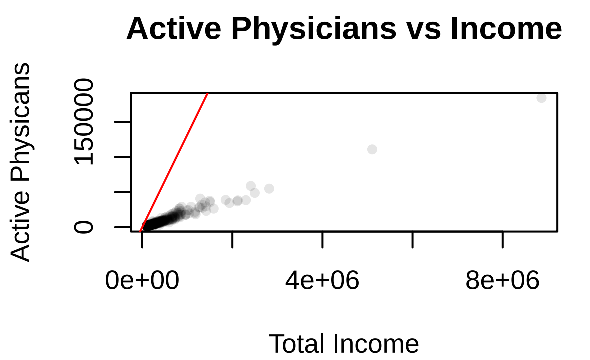
```
    xlab = "Population(Millions)",
    ylab = "Active Physicans",
    pch = 20,
    col = rgb(red=0, green = 0, blue = 0, alpha = 0.1))
abline(model_1, col="red")


plot(cdi$pop_total, cdi$hospital_beds,
    main = "Active Physicians vs Hospital beds",
    xlab = "Hospital Beds",
    ylab = "Active Physicans",
    pch = 20,
    col = rgb(red=0, green = 0, blue = 0, alpha = 0.1))
abline(model_2, col="red")


plot(cdi$pop_total, cdi$income_total,
    main = "Active Physicians vs Income",
    xlab = "Total Income",
    ylab = "Active Physicans",
    pch = 20,
    col = rgb(red=0, green = 0, blue = 0, alpha = 0.1))
abline(model_3, col="red")
```







Based on the three graphs a linear fit would only provide a good fit for the graph of when Population is the predictor variable since for that one the data points are all near the line where for the other graphs the linear regression line is way off from the data points.

```
# c
n = nrow(cdi)
MSE_1 = sum(residuals(model_1)^2) / (n-2)
MSE_2 = sum(residuals(model_2)^2) / (n-2)
MSE_3 = sum(residuals(model_3)^2) / (n-2)
```

The MSE for model 1 is 3.722035×10^5, MSE for model 2 is 3.1019188×10^5, and MSE for model 3 is 3.2453939×10^5.

c