Project 1

2024-09-12

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.

```r
CDI <- read.table("http://www.cnachtsheim-text.csom.umn.edu/Kutner/Appendix%20C%20Data%20Sets/APPENC02.t


head(CDI)
```

```
##   V1          V2 V3   V4      V5   V6   V7    V8    V9    V10  V11  V12  V13
## 1  1 Los_Angeles CA 4060 8863164 32.1  9.7 23677 27700 688936 70.0 22.3 11.6
## 2  2        Cook IL  946 5105067 29.2 12.4 15153 21550 436936 73.4 22.8 11.1
## 3  3      Harris TX 1729 2818199 31.3  7.1  7553 12449 253526 74.9 25.4 12.5
## 4  4   San_Diego CA 4205 2498016 33.5 10.9  5905  6179 173821 81.9 25.3  8.1
## 5  5      Orange CA  790 2410556 32.6  9.2  6062  6369 144524 81.2 27.8  5.2
## 6  6       Kings NY   71 2300664 28.3 12.4  4861  8942 680966 63.7 16.6 19.5
##   V14   V15    V16 V17
## 1 8.0 20786 184230   4
## 2 7.2 21729 110928   2
## 3 5.7 19517  55003   3
## 4 6.1 19588  48931   4
## 5 4.8 24400  58818   4
## 6 9.5 16803  38658   1
```

```r
colnames(CDI) <- c("identification_number", "county", "state", "land_area", "total_pop", "pop18_34", "po

head(CDI)
```

```
##   identification_number      county state land_area total_pop pop18_34 pop_65+
## 1                     1 Los_Angeles    CA      4060   8863164     32.1     9.7
## 2                     2        Cook    IL       946   5105067     29.2    12.4
## 3                     3      Harris    TX      1729   2818199     31.3     7.1
## 4                     4   San_Diego    CA      4205   2498016     33.5    10.9
## 5                     5      Orange    CA       790   2410556     32.6     9.2
## 6                     6       Kings    NY        71   2300664     28.3    12.4
##   physicians  beds serious_crimes high_school bachelors below_poverty
## 1      23677 27700         688936        70.0      22.3          11.6
## 2      15153 21550         436936        73.4      22.8          11.1
## 3       7553 12449         253526        74.9      25.4          12.5
## 4       5905  6179         173821        81.9      25.3           8.1
## 5       6062  6369         144524        81.2      27.8           5.2
```

```
## 6       4861  8942         680966        63.7       16.6            19.5
##   unemployment per_capita_income personal_income region
## 1          8.0             20786          184230      4
## 2          7.2             21729          110928      2
## 3          5.7             19517           55003      3
## 4          6.1             19588           48931      4
## 5          4.8             24400           58818      4
## 6          9.5             16803           38658      1
```

```
n= nrow(CDI)
```

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

Regressing the number of Active Physicians using Total Population as the Predictor Variable

```
population_model <- lm(physicians ~ total_pop, data = CDI)

population_summary <- summary(population_model)

population_summary
```

```
##
## Call:
## lm(formula = physicians ~ total_pop, data = CDI)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1969.4  -209.2   -88.0    27.9  3928.7
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.106e+02  3.475e+01  -3.184  0.00156 **
## total_pop    2.795e-03  4.837e-05  57.793  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 610.1 on 438 degrees of freedom
## Multiple R-squared:  0.8841, Adjusted R-squared:  0.8838
## F-statistic:  3340 on 1 and 438 DF,  p-value: < 2.2e-16
```

The estimated regression function using Total Population to predict the Number of Active Physicians is

yhat = -1.106e + 2.795e-03x

Regressing the Number of Active Physicians using the Number of Beds as the Predictor Variable

```
beds_model <- lm(physicians ~ beds, data = CDI)

beds_summary <- summary(beds_model)
```

```r
beds_summary
```

```
##
## Call:
## lm(formula = physicians ~ beds, data = CDI)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3133.2  -216.8   -32.0    96.2  3611.1
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -95.93218   31.49396  -3.046  0.00246 **
## beds          0.74312    0.01161  63.995  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 556.9 on 438 degrees of freedom
## Multiple R-squared:  0.9034, Adjusted R-squared:  0.9032
## F-statistic:  4095 on 1 and 438 DF,  p-value: < 2.2e-16
```

The estimated regression function when using Number of Beds to predict Number of Active Physicians is

yhat = -95.93218 + 0.74312x

Regressing the Number of Active Physicians using Personal Income as the Predictor Variable

```r
income_model <- lm(physicians ~ personal_income, data = CDI)

income_summary <- summary(income_model)

income_summary
```

```
##
## Call:
## lm(formula = physicians ~ personal_income, data = CDI)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1926.6  -194.5   -66.6    44.2  3819.0
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)     -48.39485   31.83333   -1.52    0.129
## personal_income   0.13170    0.00211   62.41   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 569.7 on 438 degrees of freedom
## Multiple R-squared:  0.8989, Adjusted R-squared:  0.8987
## F-statistic:  3895 on 1 and 438 DF,  p-value: < 2.2e-16
```

The estimated regression function using Personal Income to predict the Number of Active Physicians is

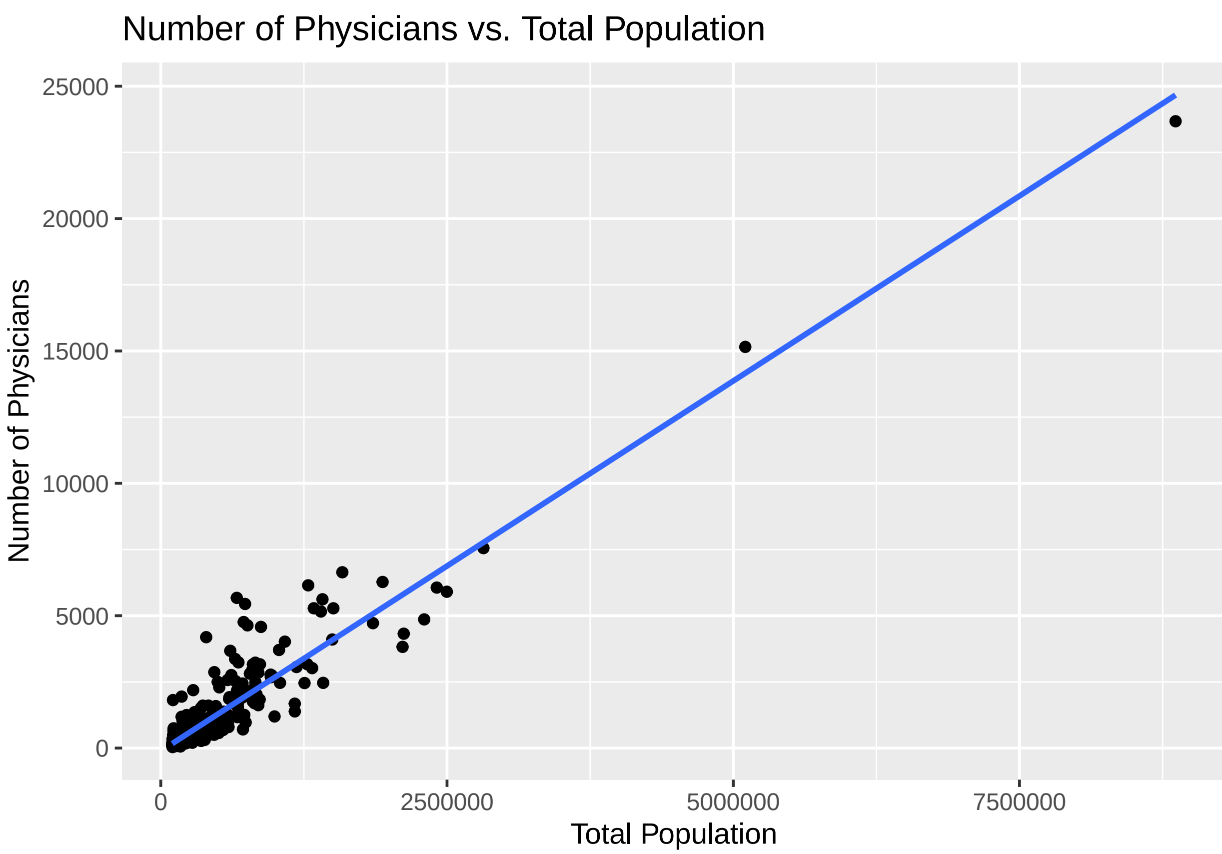
yhat = -48.39485 + 0.13170x

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

Population and Physician Regression Function Plotted

```r
library(ggplot2)

# Plot for Population
ggplot(CDI, aes(x = total_pop, y = physicians)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE)+
  labs(title = "Number of Physicians vs. Total Population",
       x = "Total Population",
       y = "Number of Physicians")
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



Upon examining the summary of the linear regression model assessing the relationship between Total Population and the Number of Active Physicians, it is observed that the model has an R-squared value of 0.8841. This high R-squared value indicates a strong relationship between the predictor (Total Population) and the response variable (Number of Active Physicians), suggesting that the model explains a significant portion of the variance in the number of physicians based on the total population. The linear regression plot of Total Population and the Number of Active Physicians appears to generally support this conclusion but the current visualization is difficult to thoroughly read and determine the fit of the linear regression relation Looking closely, it appears this difficulty in visualization could be due to some outliers in the Total Population. To better look at how well the linear regression model fits the data, I'm going to remove the outliers in the Total Population and re-plot the data.

Finding outliers in the Total Population

```
total_pop_IQR <- IQR(CDI$total_pop) # Calculate IQR

total_pop_Q1 <- quantile(CDI$total_pop, 0.25) #Calculate Q1 for total population
total_pop_Q3 <- quantile(CDI$total_pop, 0.75) #Calculate Q3 for total population

total_pop_Q1 # Print total population Q1
```

```
##      25%
## 139027.2
```

```
total_pop_Q3 #Print total population Q3
```

```
##      75%
## 436064.5
```

```
#Calculate and print lower outliers

total_pop_lower_threshold <- total_pop_Q1 - 1.5 * total_pop_IQR

total_pop_lower_threshold
```

```
##       25%
## -306528.6
```

```
# Calculate and print upper outliers

total_pop_upper_threshold <- total_pop_Q3 + 1.5 * total_pop_IQR

total_pop_upper_threshold
```

```
##      75%
## 881620.4
```

```
total_pop_outliers <- CDI$total_pop > total_pop_upper_threshold | CDI$total_pop < total_pop_lower_thres

CDI_total_pop_outliers_removed <- CDI[!total_pop_outliers,] #Create a new data frame with the CDI dataf

total_pop_outliers_removed_model <- lm(physicians ~ total_pop, data = CDI_total_pop_outliers_removed)
summary(total_pop_outliers_removed_model)
```

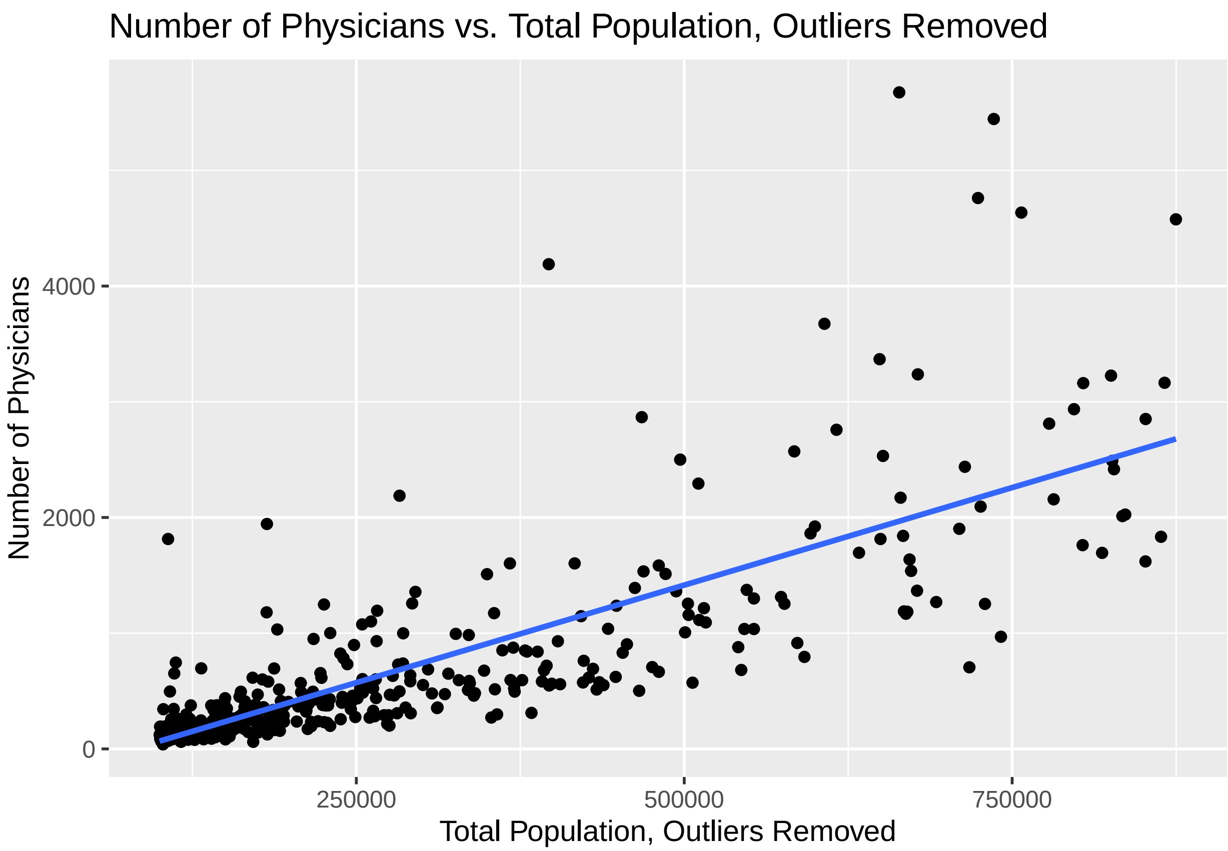
```
##
## Call:
## lm(formula = physicians ~ total_pop, data = CDI_total_pop_outliers_removed)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1443.3  -190.9   -22.0    74.7  3706.0
```

```
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2.701e+02  4.478e+01  -6.032 3.63e-09 ***
## total_pop    3.371e-03  1.302e-04  25.889  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 521 on 407 degrees of freedom
## Multiple R-squared:  0.6222, Adjusted R-squared:  0.6213
## F-statistic: 670.2 on 1 and 407 DF,  p-value: < 2.2e-16
```

Total Population and Physicians Regression Function Plotted With Outliers Removed

```r
# Plot for Population
ggplot(CDI_total_pop_outliers_removed, aes(x = total_pop, y = physicians)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE) +
  labs(title = "Number of Physicians vs. Total Population, Outliers Removed",
       x = "Total Population, Outliers Removed",
       y = "Number of Physicians")
```

```
## `geom_smooth()` using formula = 'y ~ x'
```

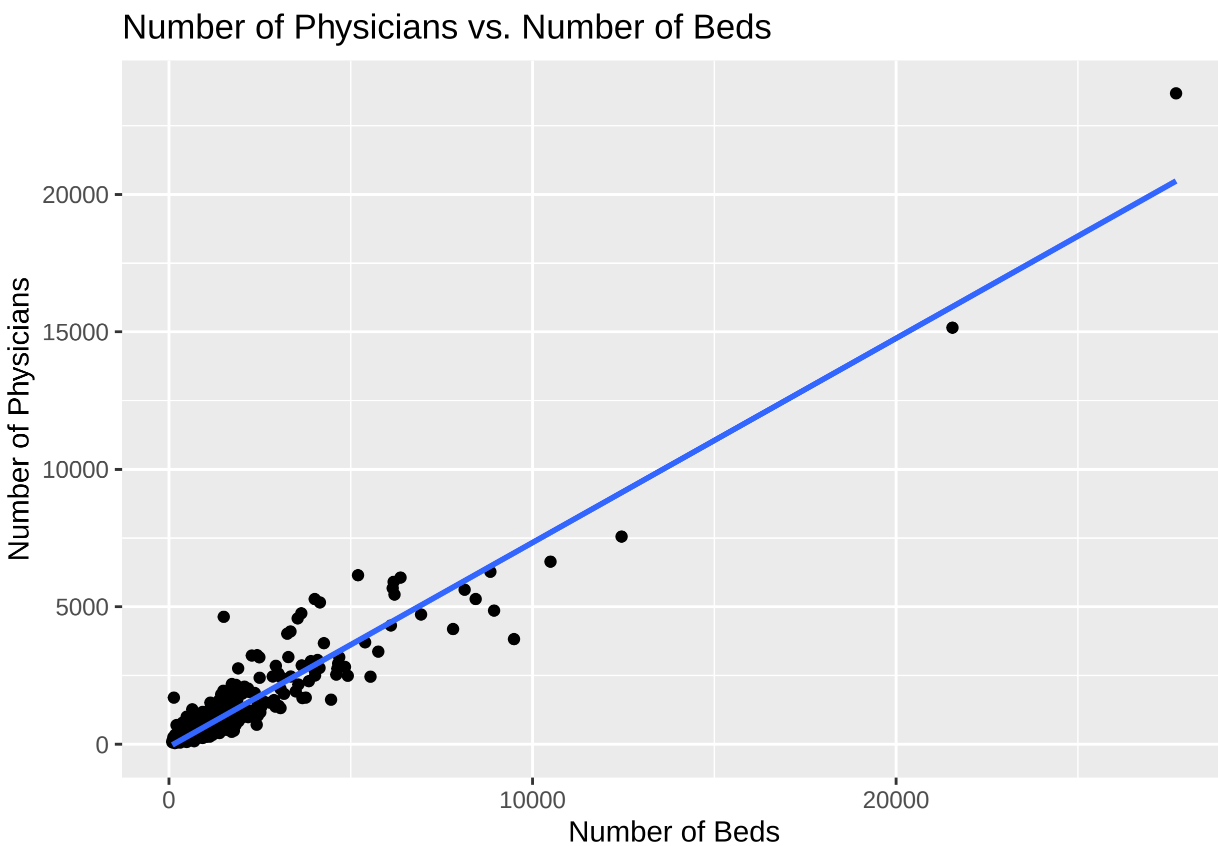


With this modified linear regression plot, it is much easier to see that the linear regression plot of the Number of Active Physicians against the Total Population demonstrates a moderately good fit as the regression line moderately-closely follows the trend of the data points. Looking at the summary of the linear regression model, however, it does appear we lose some of the strength of the relationship between the predictor variable (Total Population) and the response variable (Number of Active Physicians) as the R-squared value goes from 0.8841 to 0.6222. This suggests that the outliers of the first plot likely skewed the fit slightly, inflating it. While this is a slightly less strong relationship however an R-squared value of 0.6222 still indicates a moderate relationship, as supported by the plot as well.

Number of Beds and Physicians Regression Function Plotted

```
# Plot for Beds
ggplot(CDI, aes(x = beds, y = physicians)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE) +
  labs(title = "Number of Physicians vs. Number of Beds",
       x = "Number of Beds",
       y = "Number of Physicians")
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



Upon examining the summary of the linear regression model assessing the relationship between the Number of beds and the Number of Active Physicians, it is observed that the model has an R-squared value of 0.9034. This high R-squared value indicates a strong relationship between the predictor (Number of Beds) and the response variable (Number of Active Physicians), suggesting that the model explains a significant portion of the variance in the number of physicians based on the total population. The linear regression plot of the Number of Beds and the Number of Active Physicians appears to generally support this conclusion but the current visualization is difficult to thoroughly read and determine the fit of the linear regression relation Looking closely, it appears this difficulty in visualization could be due to some outliers in the Total Population. To better look at how well the linear regression model fits the data, I'm going to remove the outliers in the

Total Population and re-plot the data.

Finding outliers in the Number of Beds

```
beds_IQR <- IQR(CDI$beds) # Calculate IQR for number of beds

beds_Q1 <- quantile(CDI$beds, 0.25) #Calculate Q1 for number of beds
beds_Q3 <- quantile(CDI$beds, 0.75) #Calculate Q3 for number of beds

beds_Q1 # Print number of beds Q1
```

```
##    25%
## 390.75
```

```
beds_Q3 #Print number of beds Q3
```

```
##     75%
## 1575.75
```

```
#Calculate and print lower outliers

beds_lower_threshold <- beds_Q1 - 1.5 * beds_IQR

beds_lower_threshold
```

```
##      25%
## -1386.75
```

```
# Calculate and print upper outliers

beds_upper_threshold <- beds_Q3 + 1.5 * beds_IQR

beds_upper_threshold
```

```
##     75%
## 3353.25
```

```
beds_outliers <- CDI$beds > beds_upper_threshold | CDI$beds < beds_lower_threshold #Create a vector for

CDI_beds_outliers_removed <- CDI[!beds_outliers,] #Create a new data frame with the CDI dataframe itsel

beds_outliers_removed_model <- lm(physicians ~ beds, data = CDI_beds_outliers_removed)
summary(beds_outliers_removed_model)
```

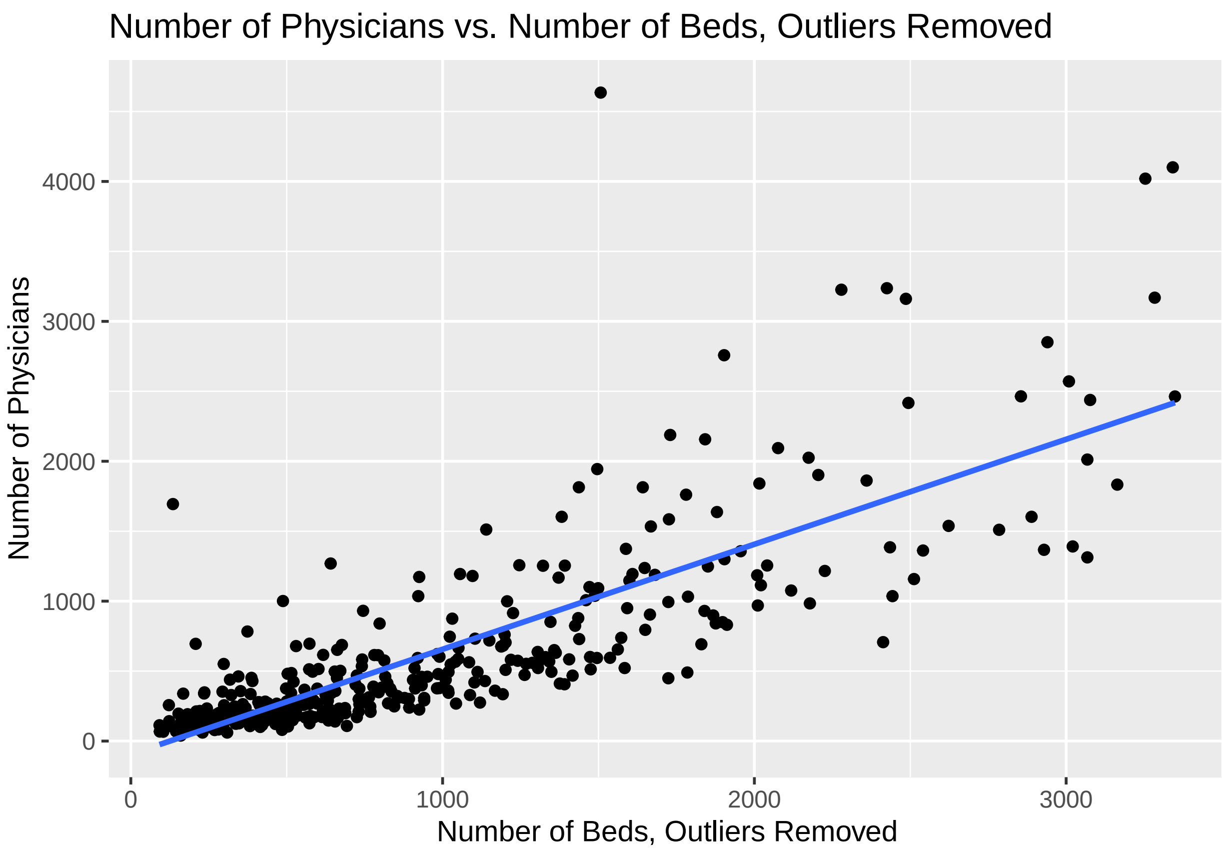
```
##
## Call:
## lm(formula = physicians ~ beds, data = CDI_beds_outliers_removed)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1009.5  -192.9   -30.7    88.8  3598.3
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -94.17720   32.17328  -2.927  0.00362 **
## beds          0.75039    0.02759  27.199  < 2e-16 ***
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 396.9 on 395 degrees of freedom
## Multiple R-squared:  0.6519, Adjusted R-squared:  0.651
## F-statistic: 739.8 on 1 and 395 DF,  p-value: < 2.2e-16
```

Number of Beds and Physicians Regression Function Plotted with Outliers excluded

```r
# Plot for Beds
ggplot(CDI_beds_outliers_removed, aes(x = beds, y = physicians)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE) +
  labs(title = "Number of Physicians vs. Number of Beds, Outliers Removed",
       x = "Number of Beds, Outliers Removed",
       y = "Number of Physicians")
```

```
## `geom_smooth()` using formula = 'y ~ x'
```

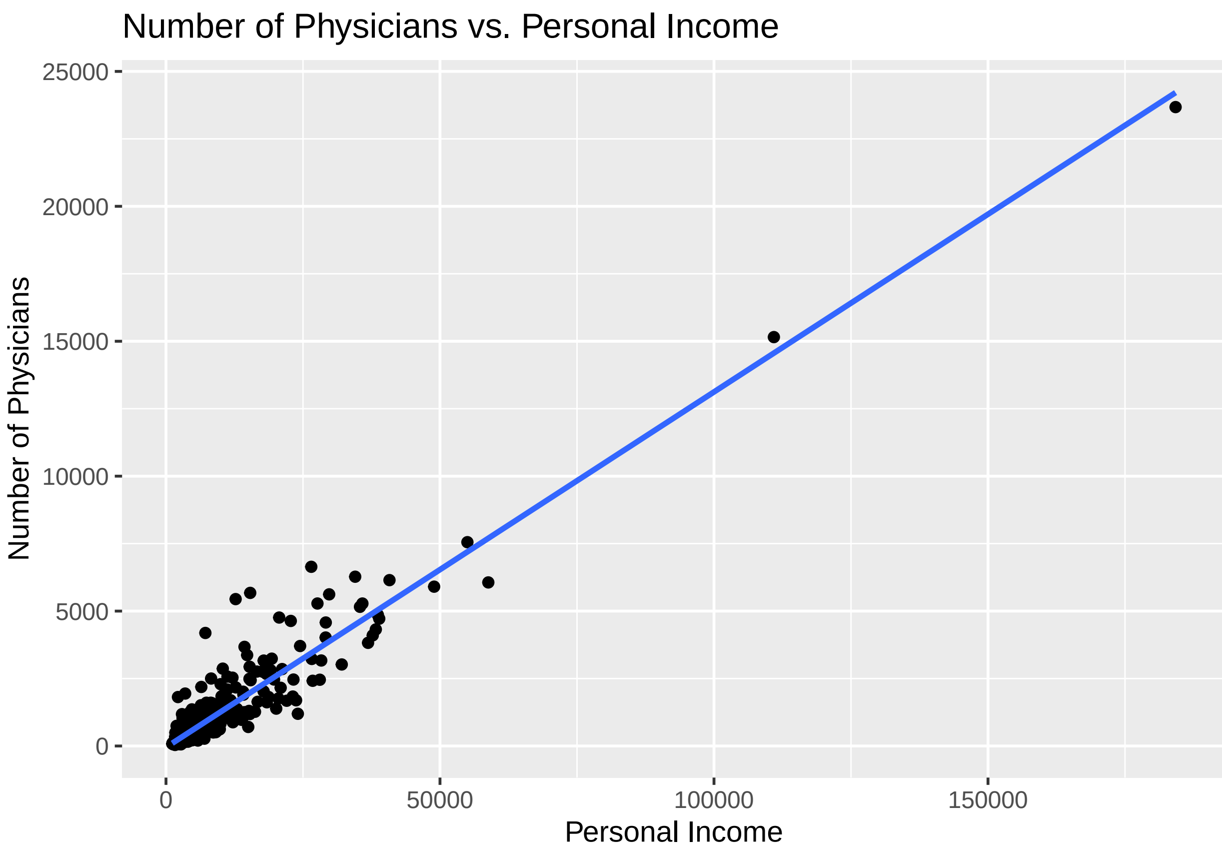


With this modified linear regression plot, it is much easier to see that the linear regression plot of the Number of Active Physicians against the Number of Beds demonstrates a moderately good fit as the regression line moderately-closely follows the trend of the data points. Looking at the summary of the linear regression model, however, it does appear we lose some of the strength of the relationship between the predictor variable (Total Population) and the response variable (Number of Active Physicians) as the R-squared value goes from 0.9034 to 0.6519. This suggests that the outliers of the first plot likely skewed the fit slightly, inflating it. While this is a slightly less strong relationship however an R-squared value of 0.6519 still indicates a moderate relationship, a conclusion also supported by the plot.

Personal Income and Number of Physicians Regression Function Plotted

```
# Plot for Income
ggplot(CDI, aes(x = personal_income, y = physicians)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE) +
  labs(title = "Number of Physicians vs. Personal Income",
       x = "Personal Income",
       y = "Number of Physicians")
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



Upon examining the summary of the linear regression model assessing the relationship between Personal Income and the Number of Active Physicians, it is observed that the model has an R-squared value of 0.8989. This high R-squared value indicates a strong relationship between the predictor (Personal Income) and the response variable (Number of Active Physicians), suggesting that the model explains a significant portion of the variance in the number of physicians based on the total population. The linear regression plot of the Number of Beds and the Number of Active Physicians appears to generally support this conclusion but the current visualization is difficult to thoroughly read and determine the fit of the linear regression relation Looking closely, it appears this difficulty in visualization could be due to some outliers in the Total Population. To better look at how well the linear regression model fits the data, I'm going to remove the outliers in the Total Population and re-plot the data.

Finding outliers in Personal Income

```
income_IQR <- IQR(CDI$personal_income) # Calculate IQR for personal income
```

```r
income_Q1 <- quantile(CDI$personal_income, 0.25) #Calculate Q1 for personal income
income_Q3 <- quantile(CDI$personal_income, 0.75) #Calculate Q3 for personal income

income_Q1 # Print personal income Q1
```

```
##  25%
## 2311
```

```r
income_Q3 #Print personal income Q3
```

```
##     75%
## 8654.25
```

```r
#Calculate and print lower outliers

income_lower_threshold <- income_Q1 - 1.5 * income_IQR

income_lower_threshold
```

```
##       25%
## -7203.875
```

```r
# Calculate and print upper outliers

income_upper_threshold <- income_Q3 + 1.5 * income_IQR

income_upper_threshold
```

```
##      75%
## 18169.12
```

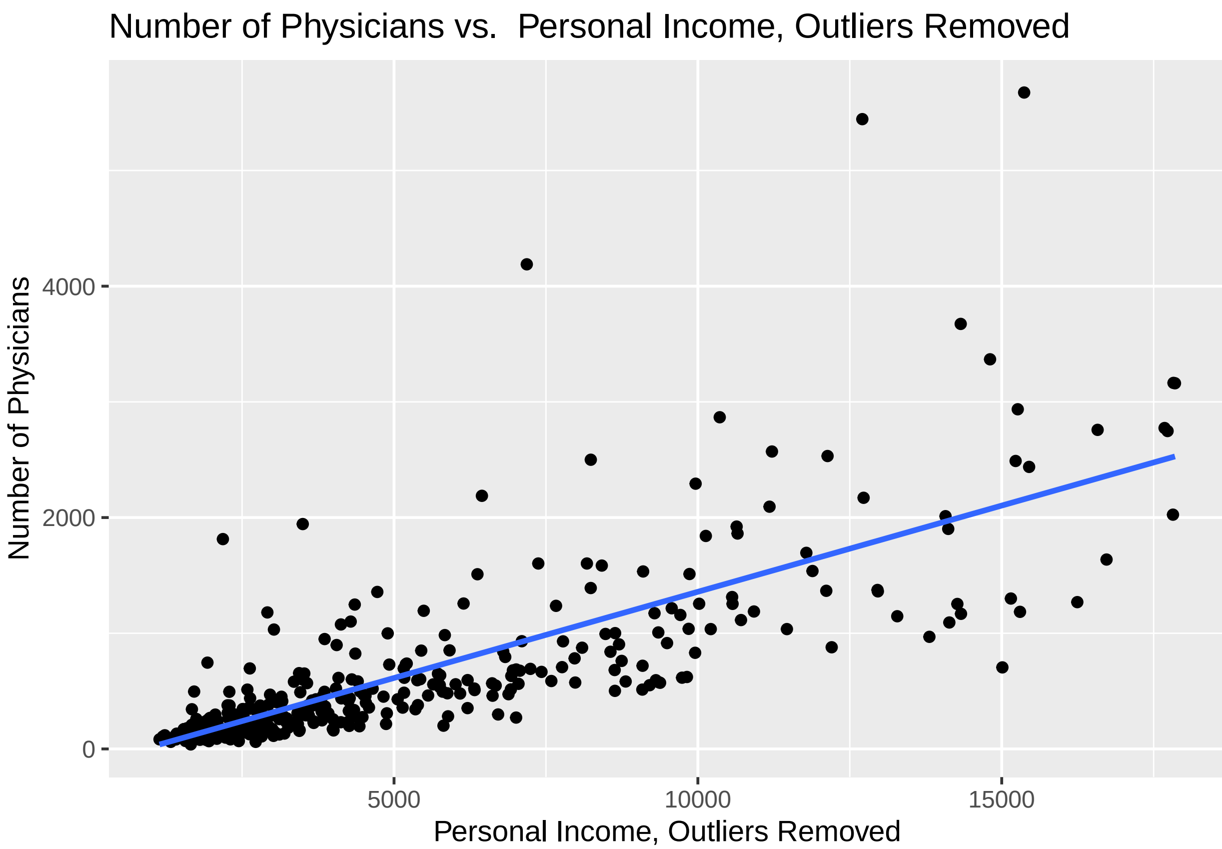
```r
income_outliers <- CDI$personal_income > income_upper_threshold | CDI$personal_income < income_lower_thr

CDI_income_outliers_removed <- CDI[!income_outliers,] #Create a new data frame with the CDI dataframe i

income_outliers_removed_model <- lm(physicians ~ personal_income, data = CDI_income_outliers_removed)

summary(income_outliers_removed_model)
```

```
##
## Call:
## lm(formula = physicians ~ personal_income, data = CDI_income_outliers_removed)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1400.2  -173.5   -34.0    72.0  3682.3
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)     -1.318e+02  3.883e+01  -3.394  0.00076 ***
## personal_income  1.490e-01  6.154e-03  24.215  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 472 on 395 degrees of freedom
## Multiple R-squared:  0.5975, Adjusted R-squared:  0.5965
## F-statistic: 586.4 on 1 and 395 DF,  p-value: < 2.2e-16
```

Personal Income and Number of Physicians Regression Function Plotted, Outliers Removed

```
# Plot for Income
ggplot(CDI_income_outliers_removed, aes(x = personal_income, y = physicians)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE) +
  labs(title = "Number of Physicians vs.  Personal Income, Outliers Removed",
       x = "Personal Income, Outliers Removed",
       y = "Number of Physicians")
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



With this modified linear regression plot, it is much easier to see that the linear regression plot of the Number of Active Physicians against Personal Income demonstrates a moderately good fit as the regression line moderately-closely follows the trend of the data points. Looking at the summary of the linear regression model, however, it does appear we lose some of the strength of the relationship between the predictor variable (Total Population) and the response variable (Number of Active Physicians) as the R-squared value goes from 0.8989 to 0.5975 This suggests that the outliers of the first plot likely skewed the fit slightly, inflating it. While this is a slightly less strong relationship however an R-squared value of 0.5975 still indicates, while less strong than the other two variables, a moderate relationship, a conclusion also supported by the plot.

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

MSE For Total Population

```r
eitotalpop <- population_summary$residuals

ssetotalpop = sum(eitotalpop^2)

msetotalpop = ssetotalpop/(n - 2)
msetotalpop
```

```
## [1] 372203.5
```

The MSE for Total Population is 372203.5 number of active physicians squared

MSE for Number of Beds

```r
eibeds <- beds_summary$residuals

ssebeds = sum(eibeds^2)

msebeds = ssebeds/(n - 2)
msebeds
```

```
## [1] 310191.9
```

The MSE for Number of Beds is 310191.9 number of active physicians squared

MSE For Personal Income

```r
eiincome <- income_summary$residuals

sseincome = sum(eiincome^2)

mseincome = sseincome/(n - 2)
mseincome
```

```
## [1] 324539.4
```

The MSE for Personal Income is 324539.4 number of active physicians squared

With MSE providing us an estimate for variability, it can be seen that using the number of beds to estimate the number of active physicians leads to the smallest variability around the fitted regression line at MSE = 310191.9 active physicians sqared as compared to MSE = 324539.4 active physicians squared when using personal income to estimate the number of active physicians and MSE = 372203.5 active physicians squared when using the total population to estimate the number of active physicians.