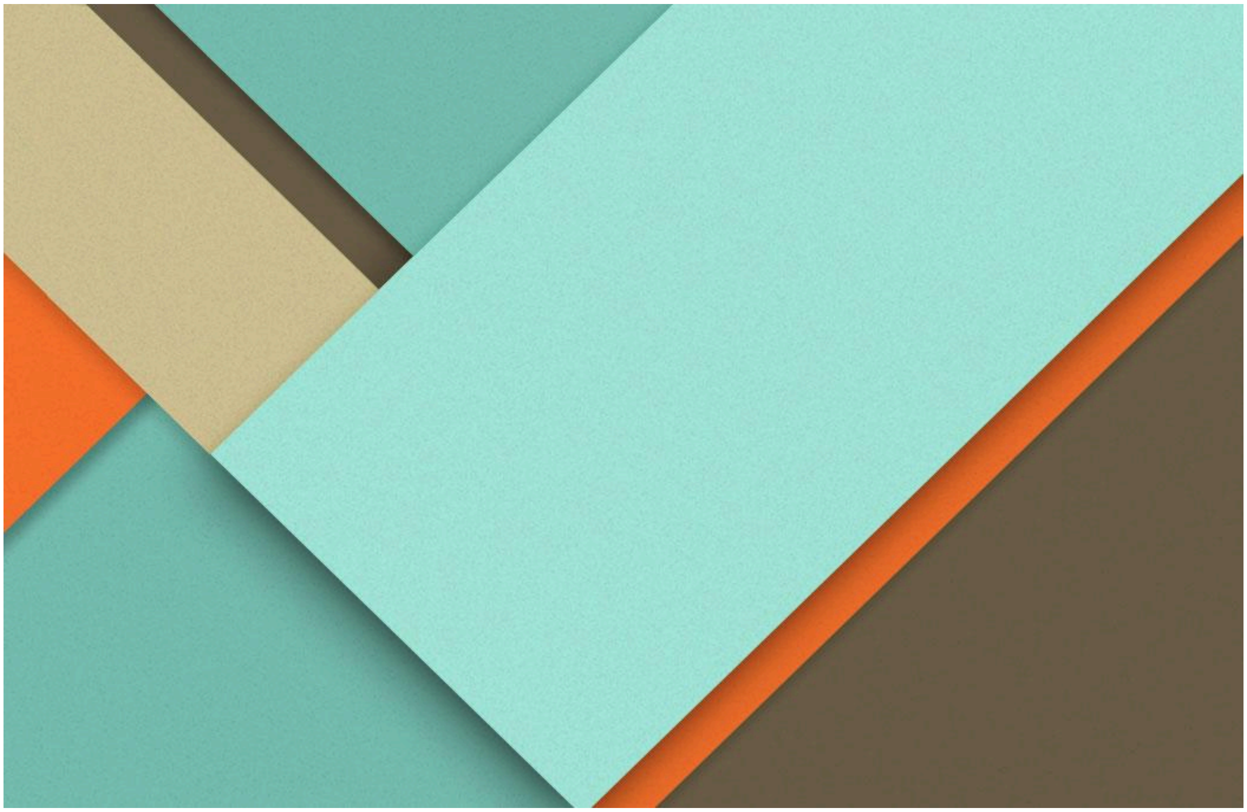


County Demographic Information: Case Study

02.10.2020

—

Radhika Kulkarni & Sameerah Helal

STA 108 A01

Professor Jiming Jiang

Introduction

In this project we would like to determine:

1) the relationships, if any, between three predictor variables and the number of active physicians in a US county, and
2) whether the per capita income has any relationship with the percentage of the population with bachelor's degrees in the county.

The variables in 1) are the total population, number of hospital beds, and the total personal income. We hypothesized that each of these variables would have no relationship with the number of physicians in a county. We will determine whether or not this statement is true for each of the predictor variables.

We plan to test our hypothesis by comparing regression models, MSE calculations, residual plots, and normal probability plots. We will also use R squared calculations to determine how much of the variability in the number of active physicians in the county is explained by the regression functions involving, respectively, total population, number of hospital beds, and total personal income. We realize, however, that we cannot definitively prove whether or not our hypothesis is true because we will not have performed the standard method of hypothesis testing to use as sufficient evidence to support or reject our hypothesis.

Our second hypothesis deals with per capita income and the percent of population who have bachelor's degrees. We hypothesize that these two variables also share no linear relationship. By separating our data into four geographical regions, which may have a different per capita income, we aim to be more cognizant of any error we could make in that respect; hence we would like to generate estimated regression functions that are more accurate. To test our hypothesis we will compare the regression function data and use the F-test technique. We will use this to determine whether or not there is a relationship between per capita income and the proportion of a population who have bachelor's degrees in each region.

We believe that this project may be valuable to the medical community because it can prove to hospitals whether or not more beds in a hospital will mean that more physicians will be working there, which could save them money and help hospitals invest in their future. Additionally, urban planners and city officials can use the relationship between total population and number of physicians to build more medical centers in populous areas, since there could be a larger number of active physicians working there, and vice versa. Economists could also find out whether a county is fiscally thriving because the number of active physicians can change based on the total personal income. Similarly, the per capita income of a region may be able determine what percentage of the population has a bachelor's degree, which can be used as an advertisement for cities looking to attract college graduates to their workforce.

The main tool we will use to generate our graphs and output is R-Studio. Screenshots of code and output are given at the end of the document.

Part I: Fitting Regression Models

1.43

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.

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

REGRESSION FUNCTION 1:

For X = total population

$\hat{Y}$ = -110.6348 + 0.002795425X

Estimated Number of Active Physicians = -110.6348 + 0.002795425 (Total Population)

REGRESSION FUNCTION 2:

For X = number of hospital beds

$\hat{Y}$= -95.93218 + 0.7431164X

Estimated Number of Active Physicians = -95.93218 + 0.7431164 (Number of Hospital Beds)

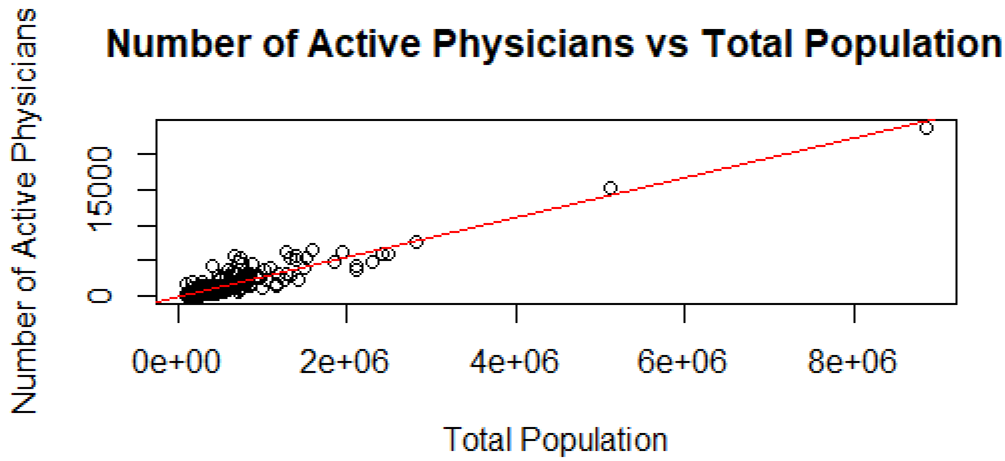
REGRESSION FUNCTION 3:

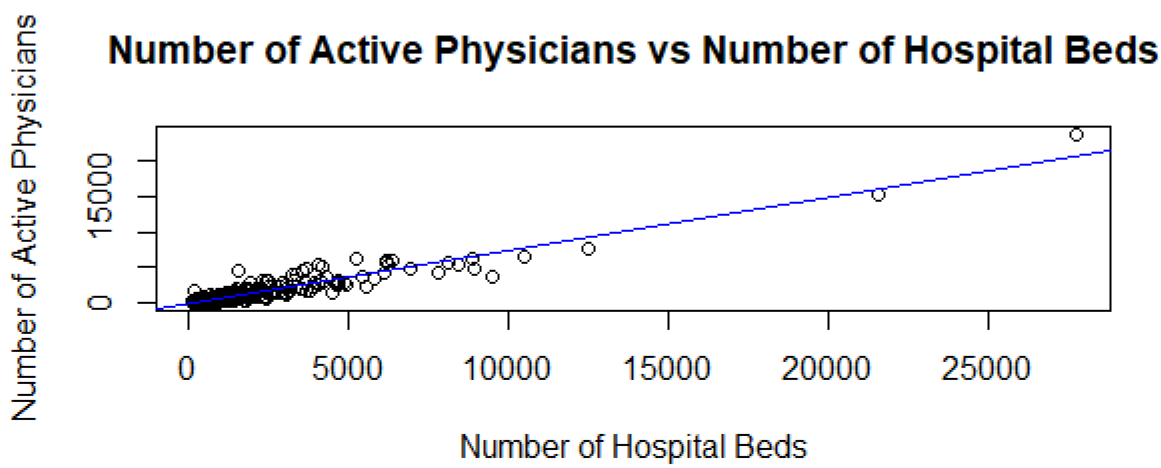
For X = total personal income

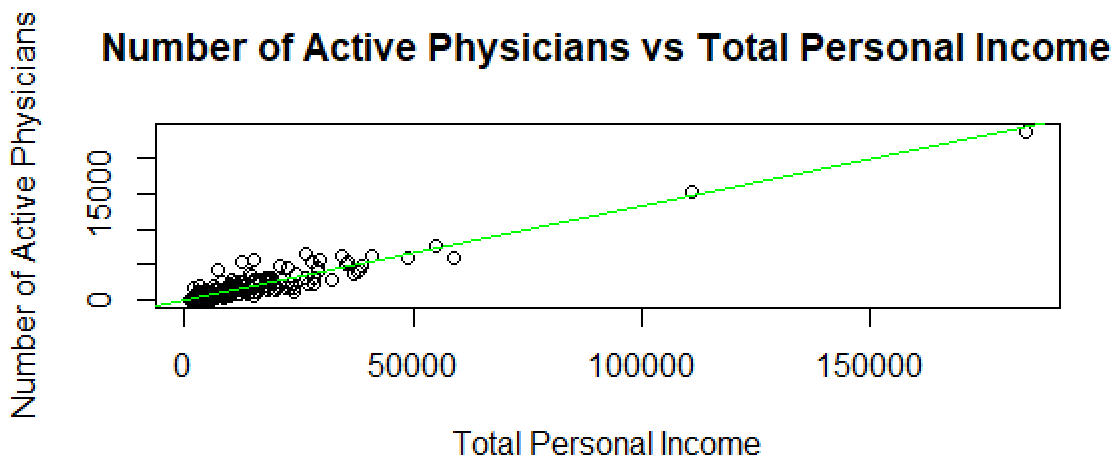
$\hat{Y} = -48.39485 + 0.1317012X$

Estimated Number of Active Physicians = -48.39485 + 0.1317012 (Total Personal Income)

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?







Yes, the respective linear regression relations appear to be a good fit for each of the three predictor variables, the line fits the data and its trajectory, hitting almost every point.

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

MSE = 372203.5
MSE2 = 310191.9
MSE3 = 324539.4
The predictor variable which leads to the smallest variability around the regression line is the number of hospital beds.

1.44

a) For each geographic region, regress per capita income in a CD (Y) against the percentage of individuals in a county having at least a bachelor's degree (X). State the estimated regression functions. b. Are the estimated regression functions similar for the four regions? Discuss.

REGRESSION FUNCTIONS:

Region 1:

For X = percentage of individuals with bachelor's degrees in region 1

$\hat{Y} = 9223.8156 + 422.1588X$

Region 2:

For X = percentage of individuals with bachelor's degrees in region 1

$\hat{Y}$ = 13581.4052 + 238.6694X

Region 3:

For X = percentage of individuals with bachelor's degrees in region 1

$\hat{Y}$ = 10529.7851 + 330.6117X

Region 4:

For X = percentage of individuals with bachelor's degrees in region 1

$\hat{Y}$ = 8615.0527 + 440.3157X

b. Are the estimated regression functions similar for the four regions? Discuss.

The estimated regression functions for the 4 regions are somewhat similar. The intercepts range from 9000 to 13000, and the slopes are all positive and in the range of 230 to 530.

c. Calculate MSE for each region. Is the variability·around the fitted regression line approximately the same for the four regions? Discuss.

The MSE was extracted from the ANOVA, which can be found below in part 3.

MSE for region 1 is 733,5008.

MSE for region 2 is 441,1341.

MSE for region 3 is 747,4349.

MSE for region 4 is 821,4318.

While the MSE for regions 1, 3, and 4 are fairly close, that of region 2 deviates noticeably. The variability around the fitted line cannot be said to be approximately the same for all 4 of the regions.

Part II: Measuring linear associations

2.62

Using R^2 as the criterion, which predictor variable accounts for the largest reduction in the variability in the number of active physicians?

Total Population:

R^2 is 0.8840674,

Hospital Beds:

R^2 is 0.9033826,

Total Personal Income:

R^2 is 0.8989137.

The predictor variable that accounts for the largest reduction in the variability in the number of physicians is the number of hospital beds, because 90.338% of the variability in number of active physicians can be explained by the regression function with X = number of hospital beds.

Part III. Inference about regression parameters

2.63

Obtain a separate interval estimate of β_1 for each region. Use a 90 percent confidence coefficient in each case. Do the regression lines for the different regions appear to have similar slopes? Also carry out the analysis of variance (ANOVA) for each regression model and state the results of the F-tests. What do you conclude in each case?

The interval estimate of β_1 for region 1 is [460.5177,583.80].
The interval estimate of β_1 for region 2 is [193.4858,283.853].
The interval estimate of β_1 for region 3 is [285.7076,375.5158].
The interval estimate of β_1 for region 4 is [364.7585,515.8729].

Comparing respectively the F* and F values for the 4 regions:

$197.75 > 2.755868$
$76.826 > 2.753462$
$148.49 > 2.739275$
$94.195 > 2.773642$.

For all four regions, the value of F* is greater than the quantity of F for the level 0.9 and the particular degrees of freedom. So we reject the null hypothesis that the slopes of any of the regression lines are 0, and accept the alternative hypothesis that all of the predictor variables have a relationship with the response variable.

ANOVA Tables:

```
> anova(regFit1)
Analysis of Variance Table

Response: perCapInc1
           Df     Sum Sq    Mean Sq F value    Pr(>F)    
bDeg1       1 1450517671 1450517671  197.75 < 2.2e-16 ***
Residuals 101  740835765    7335008                      
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> anova(regFit2)
Analysis of Variance Table

Response: perCapInc2
           Df    Sum Sq   Mean Sq F value    Pr(>F)    
bDeg2       1 338907694 338907694  76.826 3.344e-14 ***
Residuals 106 467602149   4411341                      
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> anova(regFit3)
Analysis of Variance Table

Response: perCapInc3
           Df     Sum Sq    Mean Sq F value    Pr(>F)    
bDeg3       1 1109873245 1109873245  148.49 < 2.2e-16 ***
Residuals 150 1121152411    7474349                      
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> anova(regFit4)
Analysis of Variance Table

Response: perCapInc4
          Df    Sum Sq   Mean Sq F value    Pr(>F)    
bDeg4      1 773745787 773745787  94.195 6.856e-15 ***
Residuals 75 616073841   8214318                      
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

F-intervals

```
> confint(regFit1, level = 0.9)
                 5 %     95 %
(Intercept) 7809.8077 10637.82
bDeg1        460.5177   583.80
> confint(regFit2, level = 0.9)
                  5 %      95 %
(Intercept) 12627.0363 14535.774
bDeg2         193.4858   283.853
> confint(regFit3, level = 0.9)
                 5 %       95 %
(Intercept) 9516.0773 11543.4929
bDeg3        285.7076   375.5158
> confint(regFit4, level = 0.9)
                 5 %       95 %
(Intercept) 6862.6967 10367.4086
bDeg4        364.7585   515.8729
```

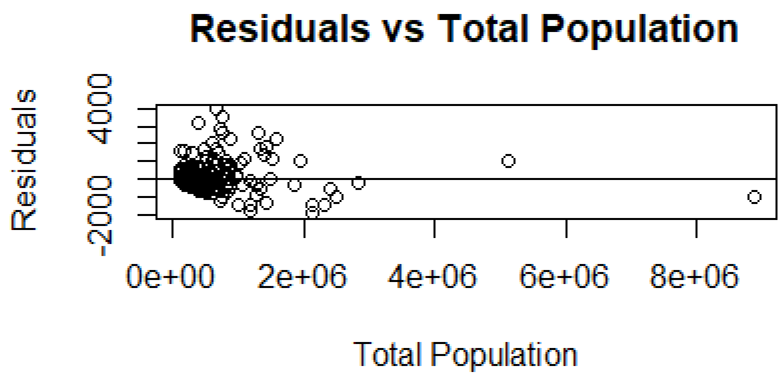
```
> regions <- CDI$V17
> perCapInc <- CDI$V15
> bDeg <- CDI$V12
>
> reg1 <- regions==1
> reg2 <- regions==2
> reg3 <- regions==3
> reg4 <- regions==4
>
> perCapInc1 <- perCapInc[reg1]
> perCapInc2 <- perCapInc[reg2]
> perCapInc3 <- perCapInc[reg3]
> perCapInc4 <- perCapInc[reg4]
>
> bDeg1 <- bDeg[reg1]
> bDeg2 <- bDeg[reg2]
> bDeg3 <- bDeg[reg3]
> bDeg4 <- bDeg[reg4]
>
> regFit1 <- lm(perCapInc1~bDeg1)
> regFit2 <- lm(perCapInc2~bDeg2)
> regFit3 <- lm(perCapInc3~bDeg3)
> regFit4 <- lm(perCapInc4~bDeg4)
>
> coef1 <- coef(summary(regFit1))
> coef2 <- coef(summary(regFit2))
> coef3 <- coef(summary(regFit3))
> coef4 <- coef(summary(regFit4))
>
> # Region n: y = coefn[1] + coefn[2]x
> # Region 1: y = 9223.8156 + 522.1588x
> # Region 2: y = 13581.4052 + 238.6694x
> # Region 3: y = 10529.7851 + 330.6117x
> # Region 4: y = 8615.0527 + 440.3157x
```

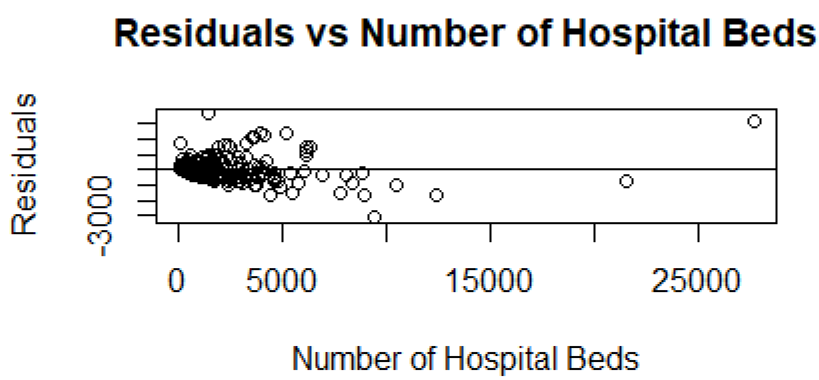
```
> qf(.9,1,101)
[1] 2.755868
> qf(.9,1,106)
[1] 2.753462
> qf(.9,1,150)
[1] 2.739275
> qf(.9,1,75)
[1] 2.773642
> qf(.9,1,8)
[1] 3.457919
```

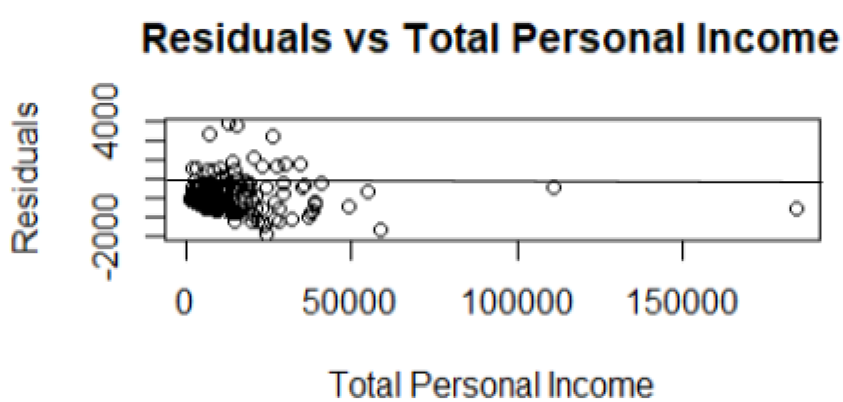
Part IV: Regression diagnostics

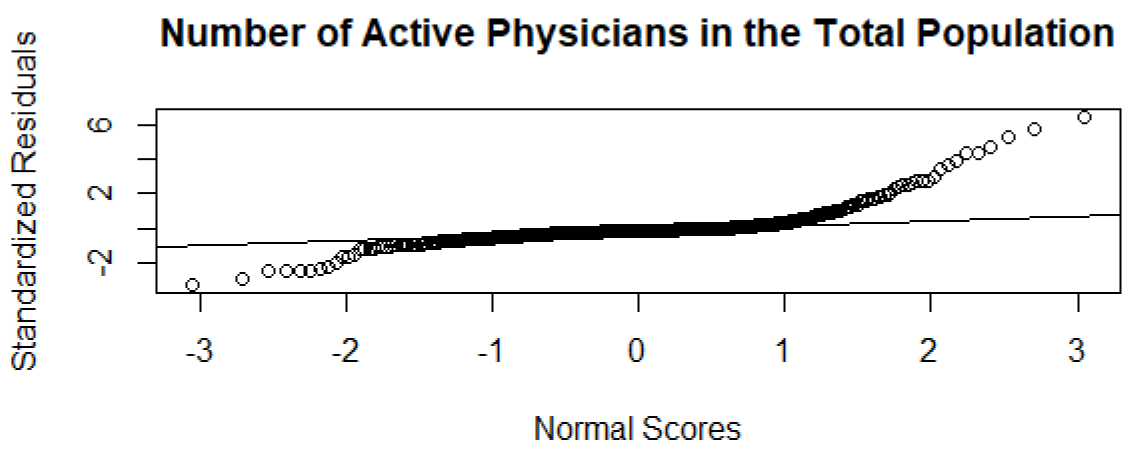
3.25

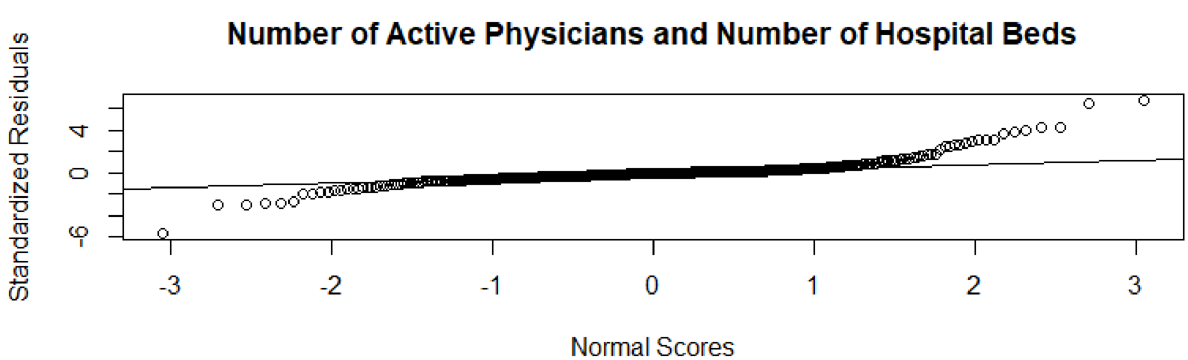
For each of the three fitted regression models, obtain the residuals and prepare a residual plot against X and a normal probability plot. Summarize your conclusions. Is linear regression model (2.1) more appropriate in one case than in the others?

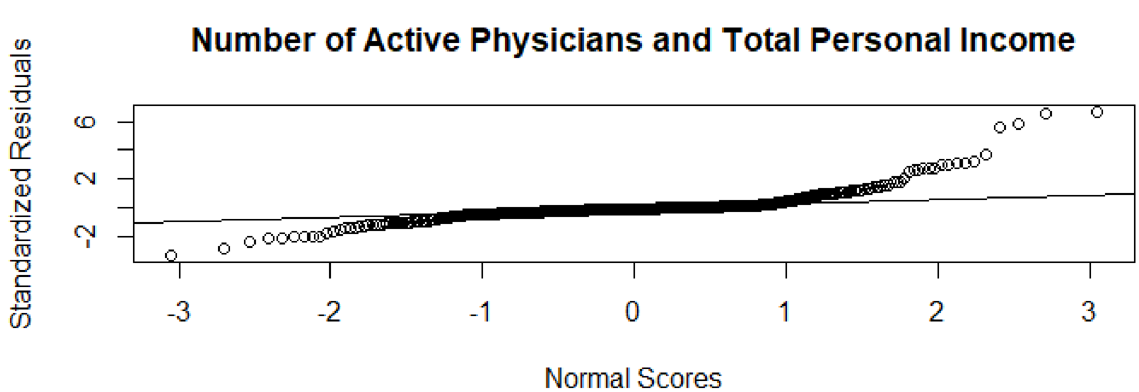

Residuals vs Total Population
Residuals
4000
-2000
0e+00
2e+06
4e+06
6e+06
8e+06
Total Population


Residuals vs Number of Hospital Beds
Residuals
-3000
0
5000
15000
25000
Number of Hospital Beds


Residuals vs Total Personal Income
Residuals
4000
-2000
0
50000
100000
150000
Total Personal Income







Yes, for us to obtain the best fitting linear regression model $Y_i = \beta_0 + \beta_1 X + \varepsilon_i$, we must have the most accurate predictor. In this case, the number of hospital beds is the most accurate predictor variable for the number of active physicians in the county.

Part V: Discussion

In the first and second parts, we analyzed the relationship of the predictor variables (total population, number of hospital beds, and total personal income) with the number of active physicians in a county. Regressing the fourth variables in turn against the first three, i.e. trying out whether any of the three predictor variables had a relationship to our response variable- number of physicians- told us that, while there was a decent line through the data formed by plotting each pair, the line was most strongly connected to the data when the predictor was the number of hospital beds and weakest when it was the value of the total population. The number of hospital beds was also that variable that accounted for the largest reduction in the variability of the number of physicians. This means that the number of hospital beds had a stronger relationship than the other two variables with the number of physicians; the number of hospital beds makes it so that there is less chance that the reason for the number of physicians having a certain value will be unexplained by the predictor. This isn't enough evidence still to reject or accept our hypothesis that there is no relationship with each of the variables to the number of active physicians in the county.

In parts I and III, the number of people with bachelor's degrees in the four different regions all seemed to have a positive linear relationship with the per capita income in the region. So the more people there were with bachelor's degrees, the higher we would expect the per capita income to be. We found 90 percent confidence intervals for the slope of the line relating the two variables for each region, meaning that we are 90% confident that each slope value is in the interval we gave. We used the F-test to check that none of the slopes were 0 to confirm that there was a significant linear relationship present. As a result of our test, we have rejected our hypothesis and accepted that there is in fact a relationship between the per capita income of a county and the percent of the population who have bachelor's degrees in each region.

We might improve the linear regression models by testing every variable, not just a few, as a predictor to see which one gives us the optimized estimated regression function.

Outputs

```
> #PART II
> #R^2 and r info:
>
> summary(fit1)$r.squared
[1] 0.8840674
> #summary(fit1)$adj.r.squared
>
> summary(fit2)$r.squared
[1] 0.9033826
> #summary(fit2)$adj.r.squared
>
> summary(fit3)$r.squared
[1] 0.8989137
> #summary(fit3)$adj.r.squared
> #***********************************
>
```

```
> #MSE totalPop - 1:
> residualStandardError1 = summary(fit1)$sigma
> MSE1 = residualStandardError1^2
> #MSE1 = 372203.5
>
> #MSE numHospBeds - 2:
> residualStandardError2 = summary(fit2)$sigma
> MSE2 = residualStandardError2^2
> #MSE2 = 310191.9
>
> #MSE totalPersonalIncome - 3:
> residualStandardError3 = summary(fit3)$sigma
> MSE3 = residualStandardError3^2
> #MSE3 = 324539.4
>
```

```
Call:
lm(formula = numActivePhys ~ totalPop, data = CDI_dataset)

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 ** 
totalPop     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
```

```
Call:
lm(formula = numActivePhys ~ numHospBeds, data = CDI_dataset)

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 ** 
numHospBeds   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
```

```
Call:
lm(formula = numActivePhys ~ totalPersonalIncome, data = CDI_dataset)

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    
totalPersonalIncome   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
```

```
> regions
  [1] 4 2 3 4 4 1 4 2 3 3 1 4 4 4 2 1 1 1 1 4 3 3 4 3 2 4 2 2 1 2 2 1 2 3 1 3 4 3 1 3 1 3 1
 [44] 4 2 2 1 3 4 3 3 4 4 2 1 1 3 1 3 3 1 1 4 4 4 3 1 3 4 3 2 1 3 3 1 3 4 4 3 2 1 1 1 3 4 1
 [87] 2 2 3 3 1 3 1 2 3 1 2 4 4 2 1 4 4 2 1 1 3 3 1 4 1 1 2 3 3 1 1 2 3 2 3 4 2 1 3 4 4 3 3
[130] 3 1 3 4 1 4 2 2 4 2 4 2 3 3 3 4 1 1 1 3 3 1 1 3 2 4 4 1 2 2 3 1 2 3 1 3 3 3 2 4 2 3 4
[173] 3 1 2 3 3 2 3 1 4 2 3 4 2 3 3 3 1 1 4 3 2 1 2 3 1 2 1 1 1 3 3 4 1 4 1 4 4 2 3 3 3 4 1
[216] 2 3 1 1 3 4 3 3 2 2 2 1 2 3 3 2 3 4 1 2 1 3 2 3 3 3 4 3 3 4 4 2 3 1 4 1 2 2 3 3 4 3 3
[259] 3 1 3 4 2 3 2 3 3 1 2 4 2 3 3 3 3 3 1 2 3 1 1 2 2 4 1 2 2 4 2 1 2 1 3 3 1 3 3 3 3 2 3
[302] 3 3 2 1 2 2 4 1 1 4 2 2 1 3 3 3 3 3 1 2 3 3 1 4 2 2 3 2 1 1 2 2 4 3 2 3 4 1 3 3 1 2 2
[345] 1 2 2 1 2 2 4 3 3 2 4 2 1 2 3 4 3 2 3 1 1 1 3 4 3 1 2 3 4 3 2 1 3 3 2 4 3 1 2 2 3 3 1
[388] 4 3 2 4 2 1 2 3 3 1 3 1 4 3 1 2 4 2 2 2 3 3 1 3 4 2 2 4 3 3 2 2 3 1 3 3 3 1 2 2 3 2 3
[431] 2 1 4 2 3 3 3 3 4 3
```

```
> perCapInc
  [1] 20786 21729 19517 19588 24400 16803 18042 17461 17823 21001 16721 23779 25193 16399
 [15] 21086 25312 20681 24262 31679 22148 22355 15508 17185 18825 26884 18934 23705 24219
 [29] 18305 19040 18431 33330 20580 26798 24875 21610 21307 16876 32342 18430 32230 28999
 [43] 22197 25523 19148 26772 24523 30081 18625 17263 19568 15399 28532 20924 21641 19895
 [57] 23470 28462 17879 17662 24896 22834 21420 16365 15191 19140 23150 18624 28819 22819
 [71] 18611 26909 23603 17741 17866 11545 16194 19215 18340 18410 27391 18463 23658 21005
 [85] 15881 22548 27378 18583 20942 19505 18521 19295 19930 18674 16578 26248 20303 15453
 [99] 17518 16327 19401 22156 18545 17815 19073 21973 17101 21933 22284 20997 21500 20974
[113] 16829 22797 20658 18878 31520 19629 14835 19276 17668 16807 18113 23008 17697 22507
[127] 22055  8899 17881 14389 24732 15648 15238 17069 21902 16898 20087 16365 18787 19465
[141] 26156 19861 18225 20349 17268 19502 19655 22581 17382 18877 16405 26026 17874 21684
[155] 14710 19932 19788 23004 19123 16015 21003 16750 15124 19785 17885 17137 18242 22782
[169] 15701 17458 13944 19942 24948 16331 21123 12923 17801 16006 20645 26757 16116 16256
[183] 22303 11467 16190 15392 16412  9728 22173 20259 21327 16215 18376 16477 17980 16337
[197] 18336 17211 21770 21362 33180 17418 18990 16790 18348 37541 18523 22025 16022 16144
[211] 15776 18301 19320 21421 24035 18288 15443 16647 16963 17744 17221 17776 20543 19692
[225] 17816 18753 18058 16904 17997 17469 16630 17192 18786 16625 15419 19254 13802 18490
[239] 16422 17951 13536 17009 15941 14925 15374 13394 18360 27546 23267 17140 15162 21855
[253] 18342 17084 20941 15051 16171 19238 16058 18857 15505 13961 19601 16319 18426 16934
[267] 14443 25161 16957 20168 15896 30242 15327 14968 18126 13691 18824 18093 16868 17908
[281] 14473 14134 16232 17312 20086 19558 14767 15301 16770 17774 18395 15853 17496 25589
[295] 17251 16924 17511 15113 19954 16231 14137 17548 10190 15750 20679 17818 16676 16277
[309] 15521 17853 15582 20682 17480 14051 14205 17129 14693 15803 15747 24132 16031 13869
[323] 16935 15197 19727 17182 17645 14934 16742 20068 16819 18161 15944 11379 14743 17278
[337]  8973 15874 19940 14615 16713 24405 16018 15847 14779 18961 17566 21944 16412 17338
[351] 16002 14814 15079 16191 19250 18526 15476 18008 22002 14197 17119 18892 12641 14834
[365] 16281 15177 17898 16728 17119 20600 15697 16021 16138 14766 14757 15778 15501 17396
[379] 18021 11396 13776 17131 21153 16305 13475 14961 16500 17272 14736 17522 17332 17175
[393] 12704 16499 13228 31699 14946 16362 15205 22668 15691 19449 16542 14523 14266 25681
[407] 12597 17306 15852 30255 16451 13681 16655 16119 11490 19345 14721 20515 15036 16029
[421] 16154 10849 16775 13350 17182 18061 16342 16514 16275 11803 16137 18070 13907 16464
[435] 19317 13919 27125 13169 18504 16458
```

```
> bDeg
  [1] 22.3 22.8 25.4 25.3 27.8 16.6 22.1 13.7 18.8 26.3 15.2 32.8 32.6 14.9 20.1 35.4 22.6
 [18] 23.0 30.0 28.8 18.8 19.7 14.6 24.0 30.2 23.0 31.6 29.2 20.0 26.6 19.3 35.3 23.7 22.1
 [35] 25.8 18.5 24.6 20.2 34.2 20.8 31.7 49.0 24.2 31.6 21.4 36.0 24.0 49.9 13.8 15.5 25.5
 [52] 23.8 35.0 13.5 26.3 22.2 25.0 32.1 21.2 18.4 26.5 25.9 23.0 16.9 23.3 19.3 27.7 19.9
 [69] 31.3 31.6 20.0 34.4 33.3 22.6 18.3 15.2 17.5 23.7 34.7 20.0 28.4 19.7 24.8 32.7 13.3
 [86] 24.8 32.0 19.7 28.3 24.4 15.9 23.7 21.0 20.7 22.4 25.0 28.8 13.2 26.7 12.8 24.4 29.0
[103] 19.3 17.0 17.6 18.7 18.8 33.0 25.2 30.7 22.2 15.3 12.8 24.6 35.3 16.7 36.7 24.9 12.9
[120] 22.2 20.4 25.8 15.3 23.6 25.5 35.2 24.5 11.5 27.5 15.5 34.7 14.8 13.0 15.4 26.6 14.3
[137] 34.2 20.6 18.0 21.5 40.5 29.6 23.5 24.8 18.7 13.9 15.1 26.4 23.9 16.4 13.1 29.5 21.0
[154] 21.4 11.8 29.8 19.5 27.1 19.0 22.4 28.3 18.7 17.0 19.6 26.3 28.0 19.7 41.9 22.2 29.2
[171]  9.1 23.6 21.9 16.2 27.6 16.6 32.3 12.3 24.1 33.0 13.0 14.0 39.1 26.2 14.7 18.2 16.8
[188] 12.0 24.8 21.8 20.7 26.4 16.7 16.7 14.4 18.2 16.7 19.2 25.9 27.6 38.3 15.5 30.1 16.8
[205] 18.6 44.0 18.1 29.7 17.5 11.4 14.3 30.2 30.6 42.1 16.4 27.1 23.4 13.6 14.8 19.3 22.9
[222] 18.6 27.6 17.5 27.6 21.2 20.7 13.0 15.5 24.2 20.7 17.6 24.9 13.6 17.3 22.9 11.5 17.7
[239] 37.1 15.1 17.2 19.8 17.3 16.6 13.7 23.5 18.7 46.9 28.1 32.3 11.9 21.0 19.5 19.4 21.5
[256] 19.5 14.7 33.4 34.6 25.2 16.6 12.0 22.4 10.8 16.5 19.1 25.9 25.0 34.1 22.7  9.0 52.3
[273] 18.4 14.7 21.0 16.3 21.6 16.0 22.5 19.0 10.8 10.3 16.7 24.7 20.5 22.3 11.1 18.0 14.2
[290] 19.5 20.0  8.1 12.7 22.3 15.6 16.9 19.8 20.0 22.0 14.5 13.1 17.0 13.4 12.9 23.0 15.0
[307] 12.2 13.7 17.7 31.9 17.6 26.2 10.7  9.3 12.9 23.1 16.0 21.0 15.6 28.2 17.6 18.9 17.0
[324] 14.2 30.3 16.7 18.2 24.6 13.3 20.9 10.8 26.0 13.8 21.9 19.1 10.5 11.1 18.4 34.0 14.6
[341] 16.9 24.9 11.7 30.7 10.5 29.0 18.5 29.2 13.0 12.7 22.0 15.7 10.0 11.6 20.8 21.3 32.3
[358] 13.6 19.6 14.0 16.5 18.0 35.8 12.9 12.4 13.6 17.2 18.5 21.2 25.4 16.5 20.7 20.0 11.4
[375] 17.5 12.3 18.7 14.2 14.8  8.2 14.2 18.1 19.6 13.5 14.4 14.8 11.8 21.5 20.0 21.9 23.3
[392] 36.5 15.1 11.0 18.4 48.5 15.0 13.4 13.6 22.3 11.7 29.1 11.4  9.7 32.9 36.2  8.5 14.6
[409] 21.9 34.6 11.2 17.7 11.7 10.5 12.7 26.4  9.1 29.5 18.1 17.9 12.6 11.0 17.7 12.7 21.7
[426] 13.8 15.5 14.4 26.5 15.0 25.4 16.8  9.0 13.9 16.2  9.7 20.3 16.5 17.8 15.5
```

```
> perCapInc1
  [1] 16803 16721 25312 20681 24262 31679 18305 33330 24875 32342 32230 22197 24523 21641
 [15] 19895 28462 24896 22834 23150 26909 17866 27391 18463 23658 22548 18521 19930 26248
 [29] 19401 19073 21973 22284 21500 20974 18878 31520 23008 24732 17069 19502 19655 22581
 [43] 16405 26026 19788 21003 19785 16331 26757 22173 20259 16477 18336 21770 21362 33180
 [57] 18348 18523 24035 16647 16963 18058 16625 19254 23267 15162 18857 25161 18824 17908
 [71] 14473 20086 17774 15853 17251 20679 15521 17853 14051 24132 15197 20068 16819 19940
 [85] 24405 14779 21944 15476 14834 16281 15177 20600 15778 17131 16500 12704 14946 15205
 [99] 19449 30255 16154 17182 18070
> perCapInc2
  [1] 21729 17461 21086 26884 23705 24219 19040 18431 20580 19148 26772 20924 18611 18410
 [15] 27378 18583 18674 20303 16327 17815 16829 19629 19276 18113 16898 20087 18787 26156
 [29] 21684 23004 19123 16750 22782 17458 21123 16006 16256 16190 18376 17980 17211 16144
 [43] 18288 19692 17816 18753 16904 16630 15419 18490 18360 21855 18342 19601 18426 16957
 [57] 15896 18093 14134 16232 19558 14767 16770 18395 16231 15750 17818 16676 20682 17480
 [71] 16031 17182 17645 16742 18161 15944 17278 16018 15847 18961 17566 16412 17338 16191
 [85] 18526 18008 18892 15697 14757 18021 21153 16305 17522 17175 16499 16542 14266 25681
 [99] 12597 16655 16119 20515 15036 18061 16342 16275 16137 16464
> perCapInc3
  [1] 19517 17823 21001 22355 15508 18825 26798 21610 16876 18430 28999 30081 17263 19568
 [15] 23470 17879 17662 19140 18624 22819 23603 17741 11545 18340 21005 20942 19505 19295
 [29] 16578 17101 21933 22797 20658 14835 17668 17697  8899 17881 14389 15648 19861 18225
 [43] 20349 17382 18877 17874 16015 15124 17885 17137 18242 13944 24948 12923 17801 20645
 [57] 22303 15392 16412  9728 16215 16337 17418 18990 15776 18301 19320 15443 17744 17776
 [71] 20543 17997 17469 17192 13802 16422 17951 13536 15941 14925 27546 17084 20941 16171
 [85] 19238 16058 15505 16319 16934 14443 30242 15327 14968 18126 13691 16868 17496 25589
 [99] 16924 17511 15113 19954 14137 17548 10190 14205 17129 14693 15803 15747 13869 16935
[113] 14934 14743  8973 14615 16713 14814 15079 22002 17119 12641 17898 17119 16021 14766
[127] 15501 17396 13776 13475 14961 14736 13228 31699 16362 15691 17306 15852 16451 19345
[141] 14721 16029 10849 16775 13350 16514 11803 19317 13919 27125 13169 16458
> perCapInc4
 [1] 20786 19588 24400 18042 23779 25193 16399 22148 17185 18934 21307 25523 18625 15399
[15] 28532 21420 16365 15191 28819 16194 19215 15881 15453 17518 22156 18545 20997 16807
[29] 22507 22055 15238 21902 16365 19465 17268 14710 19932 15701 19942 16116 11467 21327
[43] 16790 37541 22025 16022 21421 17221 18786 17009 15374 13394 17140 15051 13961 20168
[57] 17312 15301 16277 15582 19727 11379 15874 16002 19250 14197 16728 16138 11396 17272
[71] 17332 22668 14523 13681 11490 13907 18504
```

```
> bDeg1
  [1] 16.6 15.2 35.4 22.6 23.0 30.0 20.0 35.3 25.8 34.2 31.7 24.2 24.0 26.3 22.2 32.1 26.5
 [18] 25.9 27.7 34.4 18.3 28.4 19.7 24.8 24.8 15.9 21.0 25.0 24.4 17.6 18.7 25.2 22.2 15.3
 [35] 16.7 36.7 23.6 34.7 15.4 13.9 15.1 26.4 13.1 29.5 19.5 28.3 19.6 16.2 33.0 24.8 21.8
 [52] 16.7 16.7 25.9 27.6 38.3 18.6 18.1 16.4 13.6 14.8 20.7 13.6 22.9 28.1 11.9 25.2 25.0
 [69] 21.6 19.0 10.8 20.5 19.5  8.1 15.6 23.0 17.7 31.9  9.3 28.2 14.2 20.9 10.8 34.0 24.9
 [86] 10.5 29.2 32.3 12.9 12.4 13.6 25.4 12.3 18.1 11.8 15.1 15.0 13.6 29.1 34.6 12.6 21.7
[103] 16.8
> bDeg2
  [1] 22.8 13.7 20.1 30.2 31.6 29.2 26.6 19.3 23.7 21.4 36.0 13.5 20.0 20.0 32.0 19.7 20.7
 [18] 28.8 12.8 17.0 12.8 24.9 22.2 15.3 14.3 34.2 18.0 40.5 21.4 27.1 19.0 18.7 41.9 29.2
 [35] 27.6 12.3 14.0 14.7 16.7 14.4 19.2 11.4 27.1 17.5 27.6 21.2 13.0 20.7 17.3 17.7 18.7
 [52] 21.0 19.5 22.4 16.5 34.1  9.0 16.0 10.3 16.7 22.3 11.1 14.2 20.0 14.5 12.9 15.0 12.2
 [69] 26.2 10.7 17.6 16.7 18.2 13.3 26.0 13.8 10.5 11.7 30.7 29.0 18.5 13.0 12.7 11.6 21.3
 [86] 13.6 18.0 16.5 17.5 14.8 19.6 13.5 21.9 36.5 11.0 11.4 32.9 36.2  8.5 11.7 10.5 29.5
[103] 18.1 13.8 15.5 26.5 25.4 13.9
> bDeg3
  [1] 25.4 18.8 26.3 18.8 19.7 24.0 22.1 18.5 20.2 20.8 49.0 49.9 15.5 25.5 25.0 21.2 18.4
 [18] 19.3 19.9 31.6 33.3 22.6 15.2 34.7 32.7 28.3 24.4 23.7 22.4 18.8 33.0 24.6 35.3 12.9
 [35] 20.4 25.5 11.5 27.5 15.5 14.8 29.6 23.5 24.8 23.9 16.4 21.0 22.4 17.0 26.3 28.0 19.7
 [52]  9.1 21.9 16.6 32.3 24.1 39.1 18.2 16.8 12.0 26.4 18.2 15.5 30.1 14.3 30.2 30.6 23.4
 [69] 19.3 18.6 27.6 15.5 24.2 17.6 11.5 37.1 15.1 17.2 17.3 16.6 46.9 19.4 21.5 14.7 33.4
 [86] 34.6 16.6 10.8 19.1 25.9 52.3 18.4 14.7 21.0 16.3 22.5 12.7 22.3 16.9 19.8 20.0 22.0
[103] 13.1 17.0 13.4 12.9 23.1 16.0 21.0 15.6 18.9 17.0 24.6 19.1 11.1 14.6 16.9 15.7 10.0
[120] 19.6 16.5 35.8 17.2 21.2 20.7 11.4 18.7 14.2 14.2 14.4 14.8 20.0 18.4 48.5 13.4 11.7
[137] 14.6 21.9 11.2 26.4  9.1 17.9 11.0 17.7 12.7 14.4 15.0 16.2  9.7 20.3 16.5 15.5
> bDeg4
 [1] 22.3 25.3 27.8 22.1 32.8 32.6 14.9 28.8 14.6 23.0 24.6 31.6 13.8 23.8 35.0 23.0 16.9
[18] 23.3 31.3 17.5 23.7 13.3 13.2 26.7 29.0 19.3 30.7 25.8 35.2 24.5 13.0 26.6 20.6 21.5
[35] 18.7 11.8 29.8 22.2 23.6 13.0 26.2 20.7 16.8 44.0 29.7 17.5 42.1 22.9 24.9 19.8 13.7
[52] 23.5 32.3 19.5 12.0 22.7 24.7 18.0 13.7 17.6 30.3 21.9 18.4 22.0 20.8 14.0 18.5 20.0
[69]  8.2 21.5 23.3 22.3  9.7 17.7 12.7  9.0 17.8
```

```
> coef1
             Estimate Std. Error  t value     Pr(>|t|)
(Intercept) 9223.8156  851.77065 10.82899 1.347038e-18
bDeg1        522.1588   37.13141 14.06246 1.589101e-25
> coef2
              Estimate Std. Error  t value     Pr(>|t|)
(Intercept) 13581.4052   575.1441 23.61392 5.068448e-44
bDeg2         238.6694    27.2296  8.76507 3.344138e-14
> coef3
              Estimate Std. Error  t value     Pr(>|t|)
(Intercept) 10529.7851  612.48431 17.19193 2.751512e-37
bDeg3         330.6117   27.13115 12.18569 3.539586e-24
> coef4
             Estimate Std. Error  t value     Pr(>|t|)
(Intercept) 8615.0527 1052.19722 8.187679 5.242270e-12
bDeg4        440.3157   45.36812 9.705399 6.855606e-15
```

Code

```
#PROJECT BEGINS: PART I

#Project 1.43:

# first predictor variable
 totalPop <- CDI_dataset$V5
# second predictor variable
 numHospBeds <- CDI_dataset$V9
```

```
# third predictor variable
 totalPersonalIncome <- CDI_dataset$V16
# response variable
 numActivePhys <- CDI_dataset$V8
#
#  1: totalPop vs numActivePhys
fit1 = lm(numActivePhys~totalPop, data = CDI_dataset)
summary(fit1)

beta1hat1 = coef(summary(fit1))[2]
beta0hat1 = coef(summary(fit1))[1]

#a) REGRESSION FUNCTION 1:
#Yhat = -110.6348 + 0.002795425X
#Estimated Number of Active Physicians = -110.6348 + 0.002795425
(Total Population)

# 2: numHospBeds vs numActivePhys
fit2 = lm(numActivePhys~numHospBeds, data = CDI_dataset)
summary(fit2)
beta1hat2 = coef(summary(fit2))[2]
beta0hat2 = coef(summary(fit2))[1]

#a) REGRESSION FUNCTION 2:
#Yhat = -95.93218 + 0.7431164X
#Estimated Number of Active Physicians = -95.93218 + 0.7431164(Number
of Hospital Beds)

# 3: totalPersonalIncome vs numActivePhys
fit3 = lm(numActivePhys~totalPersonalIncome, data = CDI_dataset)
summary(fit3)
beta1hat3 = coef(summary(fit3))[2]
beta0hat3 = coef(summary(fit3))[1]

#a) REGRESSION FUNCTION 3:
#Yhat = -48.39485 + 0.1317012X
#Estimated Number of Active Physicians = -48.39485 + 0.1317012(Total
Personal Income)

#1.43 b)
#1st plot and regression line
```

```
plot(totalPop,numActivePhys, xlab = 'Total Population', ylab =
'Number of Active Physicians', main = 'Number of Active Physicians vs
Total Population')
abline(fit1, col = 'red')

#2nd plot and regression line
plot(numHospBeds,numActivePhys, xlab = 'Number of Hospital Beds',
ylab = 'Number of Active Physicians', main = 'Number of Active
Physicians vs Number of Hospital Beds')
abline(fit2, col = 'blue')

#3rd plot and regression line
plot(totalPersonalIncome, numActivePhys, xlab = 'Total Personal
Income', ylab = 'Number of Active Physicians', main = 'Number of
Active Physicians vs Total Personal Income')
abline(fit3, col = 'green')
#Yes, the linear regression relations appear to be a good fit for the
data.

#c)
#The predictor variable which leads to the smallest variability
around the regression line is number of hospital beds.
#MSE totalPop - 1:
residualStandardError1 = summary(fit1)$sigma
MSE1 = residualStandardError1^2
#MSE1 = 372203.5

#MSE numHospBeds - 2:
residualStandardError2 = summary(fit2)$sigma
MSE2 = residualStandardError2^2
#MSE2 = 310191.9

#MSE totalPersonalIncome - 3:
residualStandardError3 = summary(fit3)$sigma
MSE3 = residualStandardError3^2
#MSE3 = 324539.4



# # #_______________________________________________
#Project 1.44:
```

```
#data stored in regions
regions <- CDI_dataset$V17

#per capita income
perCapInc <- CDI_dataset$V15

#bachelor degree
bDeg <- CDI_dataset$V12



#region 1: NE
reg1 <- regions==1
perCapInc1 <- perCapInc[reg1]
bDeg1 <- bDeg[reg1]
coef1 <- coef(summary(lm(perCapInc1~bDeg1)))


#region 2: NC
reg2 <- regions==2
bDeg2 <- bDeg[reg2]
perCapInc2 <- perCapInc[reg2]
coef2 <- coef(summary(lm(perCapInc2~bDeg2)))


#region 3: S
reg3 <- regions==3
bDeg3 <- bDeg[reg3]
perCapInc3 <- perCapInc[reg3]
coef3 <- coef(summary(lm(perCapInc3~bDeg3)))

#region 4: W
reg4 <- regions==4
bDeg4 <- bDeg[reg4]
perCapInc4 <- perCapInc[reg4]
coef4 <- coef(summary(lm(perCapInc4~bDeg4)))


#a) REGRESSION FUNCTIONS:
# Region 1: y = 9223.8156 + 422.1588x
```

```
# Region 2: y = 13581.4052 + 238.6694x
# Region 3: y = 10529.7851 + 330.6117x
# Region 4: y = 8615.0527 + 440.3157x
#*********************************
#PART II
#R^2 and r info:

summary(fit1)$r.squared
#summary(fit1)$adj.r.squared

summary(fit2)$r.squared
#summary(fit2)$adj.r.squared

summary(fit3)$r.squared
#summary(fit3)$adj.r.squared
#**********************************

#PART III:
#2.63:
anova(lm(perCapInc1~bDeg1))
anova(lm(perCapInc2~bDeg2))
anova(lm(perCapInc3~bDeg3))
anova(lm(perCapInc4~bDeg4))

#************************************

# #________________________________
#PART IV:
#residuals 3.25:
residuals1 = fit1$residuals
residuals2 = fit2$residuals
residuals3 = fit3$residuals

#residual plots against x

resPlot1.lm = lm(numActivePhys ~ totalPop, data = CDI_dataset)
numActivePhys.resid = resid(resPlot1.lm)
plot(totalPop,numActivePhys.resid , ylab="Residuals", xlab="Total
Population", main="Residuals vs Total Population")
abline(0, 0)
```

```
resPlot2.lm = lm(numActivePhys ~ numHospBeds, data = CDI_dataset)
numActivePhys.resid = resid(resPlot2.lm)
plot(numHospBeds,numActivePhys.resid , ylab="Residuals", xlab="Number
of Hospital Beds", main="Residuals vs Number of Hospital Beds")
abline(0, 0)

resPlot3.lm = lm(numActivePhys ~ totalPersonalIncome, data =
CDI_dataset)
numActivePhys.resid = resid(resPlot3.lm)
plot(totalPersonalIncome,numActivePhys.resid , ylab="Residuals",
xlab="Total Personal Income", main="Residuals vs Total Personal
Income")
abline(0, 0)

#qq plot 1
resplot1.lm = lm(numActivePhys ~ totalPop, data=CDI_dataset)
numActivePhys.stdres1 = rstandard(resplot1.lm)
qqnorm(numActivePhys.stdres1, ylab="Standardized Residuals",
xlab="Normal Scores", main="Number of Active Physicians in the Total
Population")
qqline(numActivePhys.stdres1)

#qq plot 2
resplot2.lm = lm(numActivePhys ~ numHospBeds, data=CDI_dataset)
numActivePhys.stdres2 = rstandard(resplot2.lm)
qqnorm(numActivePhys.stdres2, ylab="Standardized Residuals",
xlab="Normal Scores", main="Number of Active Physicians and Number of
Hospital Beds")
qqline(numActivePhys.stdres2)

#qq plot 3
resplot3.lm = lm(numActivePhys ~ totalPersonalIncome,
data=CDI_dataset)
numActivePhys.stdres3 = rstandard(resplot3.lm)
qqnorm(numActivePhys.stdres3, ylab="Standardized Residuals",
xlab="Normal Scores", main="Number of Active Physicians and Total
Personal Income")
qqline(numActivePhys.stdres3)
```