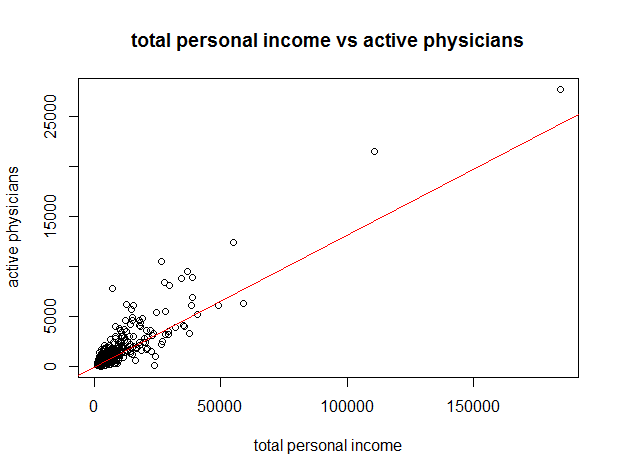
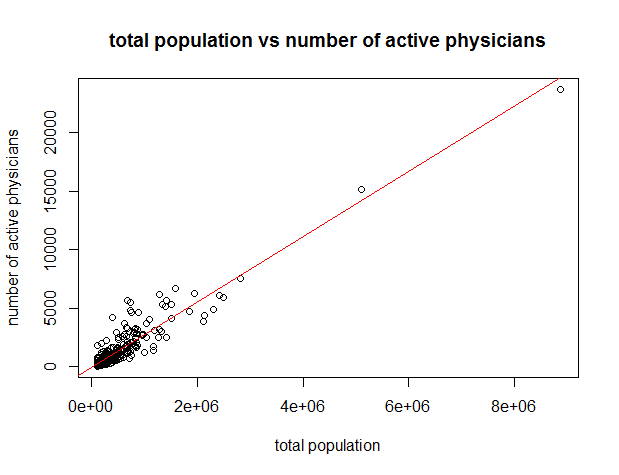
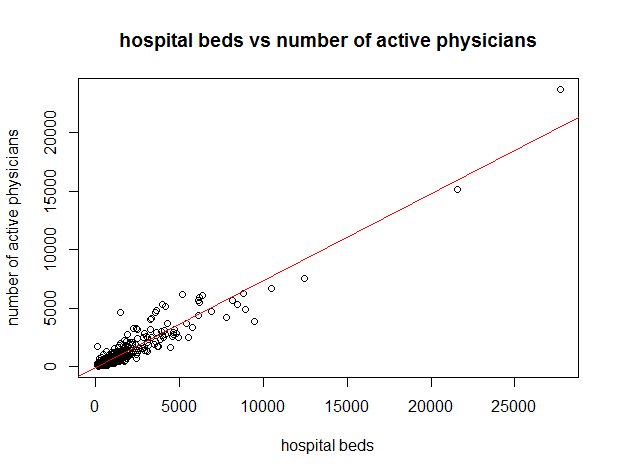
1.43

|  |
| --- |
| Predictor: Total population: |
| Hospital beds: |
| Total personal income: |

**a.**



**b**.



Based on the graphs, the linear regression models appear to provide a good fit for each of the three predictor variables. In the meanwhile, all of them have obvious outliers.

**c.**

Since 310192<324539<372204, number of

|  |
| --- |
| MSE |
| Total population: 372204 |
| Hospital beds: 310192 |
| Total personal income: 324539 |

hospital beds leads to the smallest variability

around the fitted regression line.

1.44

**a.**

|  |
| --- |
| Predictor: Bachelor Degree Fitted value: per capita income |
| Region 1: |
| Region 2: |
| Region 3: |
| Region 4: |

**b.**

The regression models for the four regions are roughly similar that all of them are upward sloping with similar intercepts. However, region 2 has higher per capita income with a lower percentage of bachelor’s degrees, compared with other regions. Therefore, the intercepts and slopes are suggesting the different level significance of the percentage of bachelor’s degrees.

**c.**

|  |
| --- |
| MSE |
| Region 1: 7335008 |
| Region 2: 4411341 |
| Region 3: 7474349 |
| Region 4: 8214318 |

Region1 and region 3’ s variabilities are approximately the same. Region 2’s variability is much lower than the others. Region 4’s variability is slightly higher than region 1 and region 3.

2.62

|  |
| --- |
| Coefficient of determination (R^2) |
| Predictor: Total population: 0.8841 |
| Hospital beds: 0.9034 |
| Total personal income: 0.8989 |

Since 0.9034 > 0.8989 > 0.8841, the number of hospital beds accounts for the largest reduction in the variability in the number of active physicians.

2.63

|  |
| --- |
| 90% confidence interval |
| Region 1: (460.52, 583.80) |
| Region 2: (193.47, 283.87) |
| Region 3: (285.57,375.65) |
| Region 4: (365.002, 515.64) |

|  |
| --- |
| s.e() |
| Standard error |
| Region 1: 37.13 |
| Region 2: 27.23 |
| Region 3: 27.13 |
| Region 4: 45.37 |

Region 2 and region 3 have overlapping intervals that they might have very similar slopes. Region1and region 4 also have overlapping intervals. Region 3 and 4 have a small overlapping interval. However, Region 1 has much higher slope interval than region 2 and 3, so they will not be similar.

ANOVA TABLE

SOURCE SS df MS F test

SSR MSR

1. 1450517671 1 1450517671 197.75

Regression 2. 338907694 1 338907694 76.826

3. 1109873245 1 1109873245 148.49

4. 773745787 1 7743745787 94.195

SSE MSE

1. 740835765 101 7335008

Error 2. 467602149 106 4411341

3. 1121152411 150 7474349

4. 616073841 75 8214318

SSTO

1. 2191353436 102

Total 2. 806509843 107

3. 2231025656 151

4. 3004771443 76

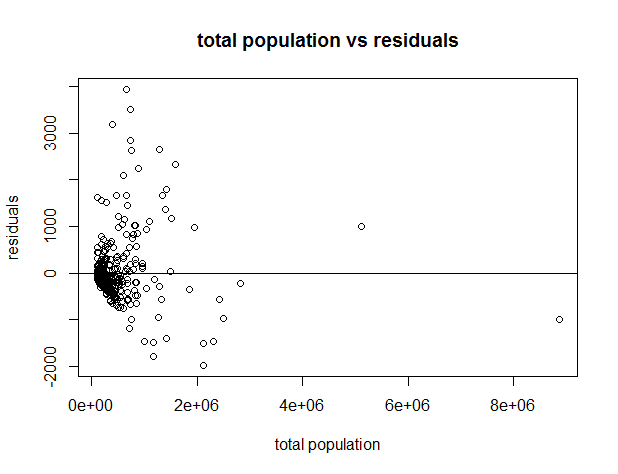
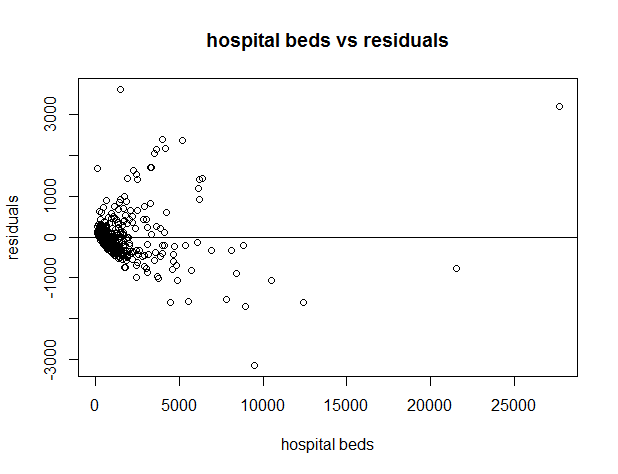
Null Hypothesis: 0 Alternate Hypothesis:

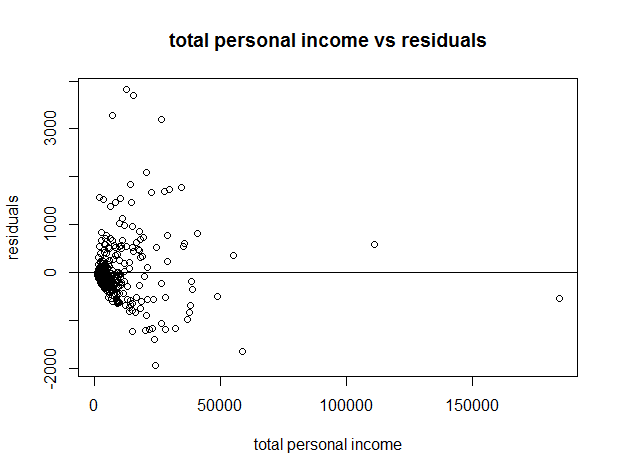
F1,101(0.9) = 2.155868<197.75 F1,106(0.9) = 0.7535<76.865

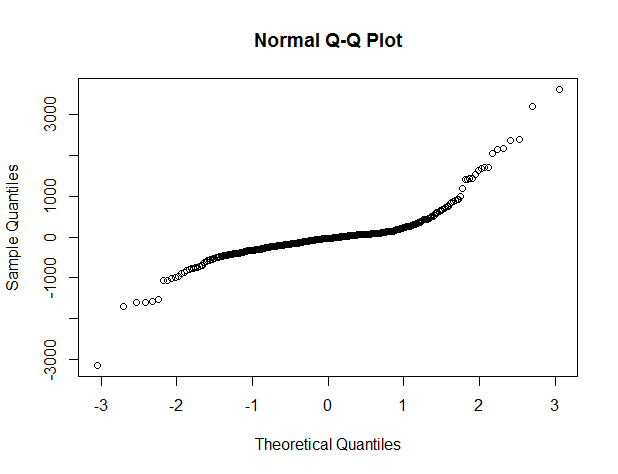
F1,150(0.9) = 2.74<148.49 F1,75(0.9) = 2.77<94.195

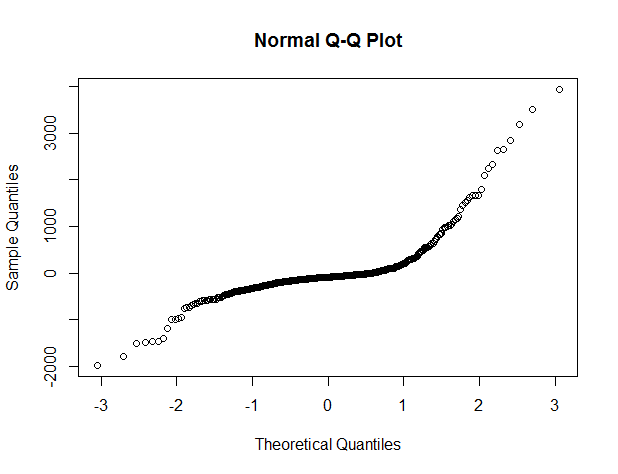
Conclusion: Accept null hypothesis, since all the F scores are all significantly large. Therefore, there are linear relation between percent bachelor’s degrees and per capita income for four regions.

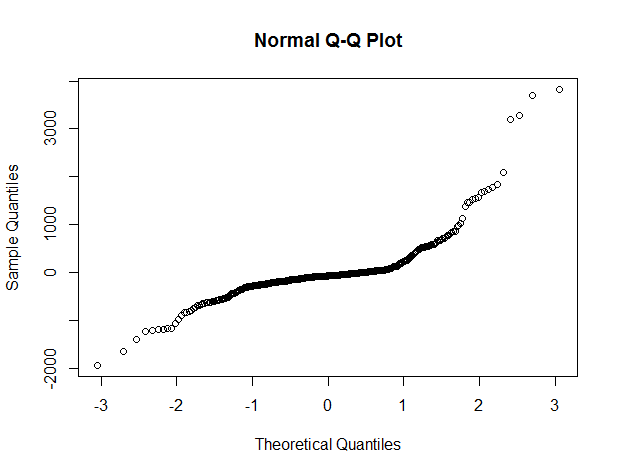
3.25





 Total population Hospital beds



 Total personal income

**Conclusion:**

1. Regarding the residuals against predictor plots, predictor total population appears to have more appropriate linear regression model than the others. On the other hand, residuals are asymmetrically lying around zero line, so they might not meet the normality assumption. All of them have constant variance and some obvious outliers.
2. Base on the normal Q-Q plot, none of them meet the normality assumption. Since none of them roughly follow a straight line.

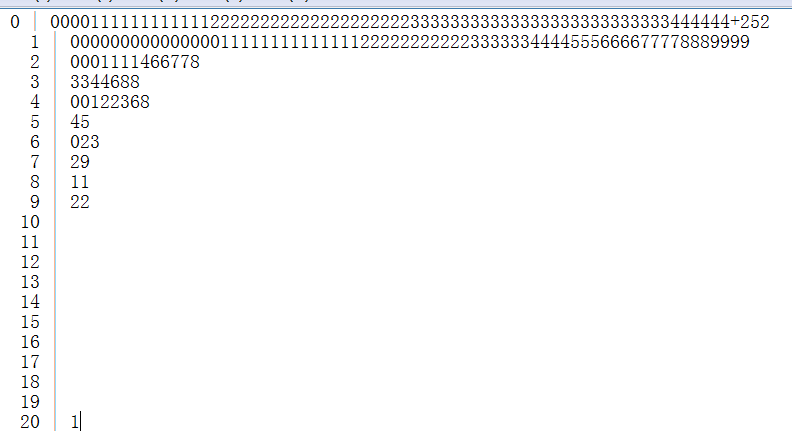
General Interpretation

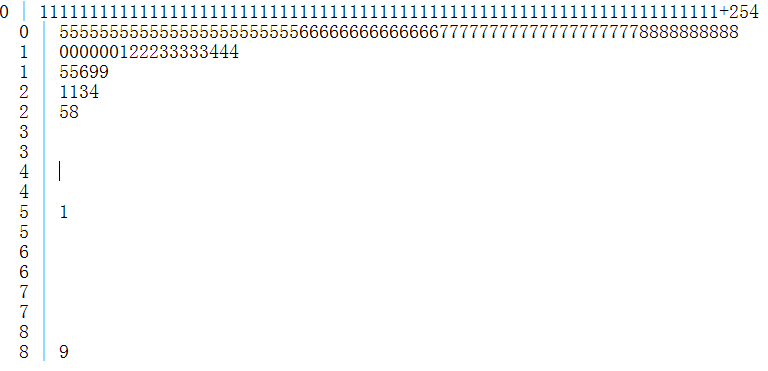
In this report, the relationships among the number of active physicians and the predictors (total population, number of hospital beds, and total personal income) are analyzed. It was found that the number of active physicians has positive relations with the predictors such that the number of active physicians increases as the total population gets larger. However, it was found that total population has the strongest relation with the number of active physicians compared with other two predictors.

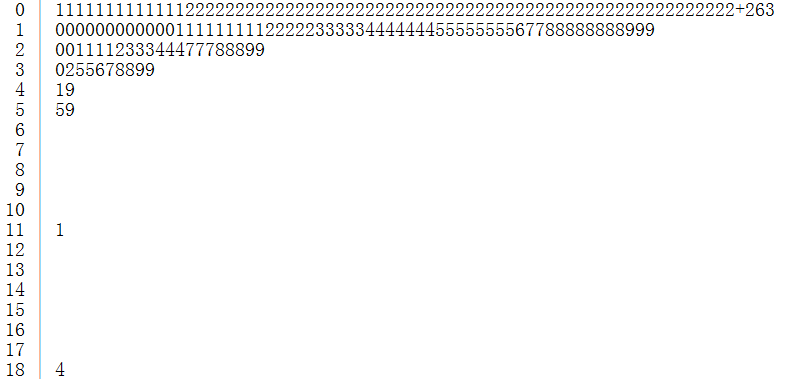
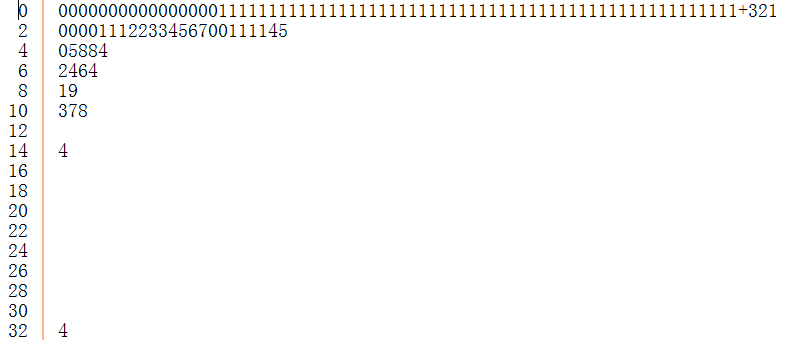
The report also finds that region 2 has much higher per capita income than other regions even when its percent of bachelor’s degrees is relatively low. However, the percent of bachelor’s degrees attributes more to the per capita income in region 1 and 4.

Part II

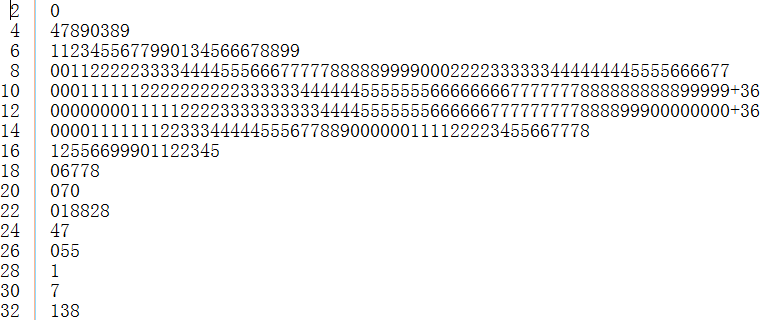
6.28

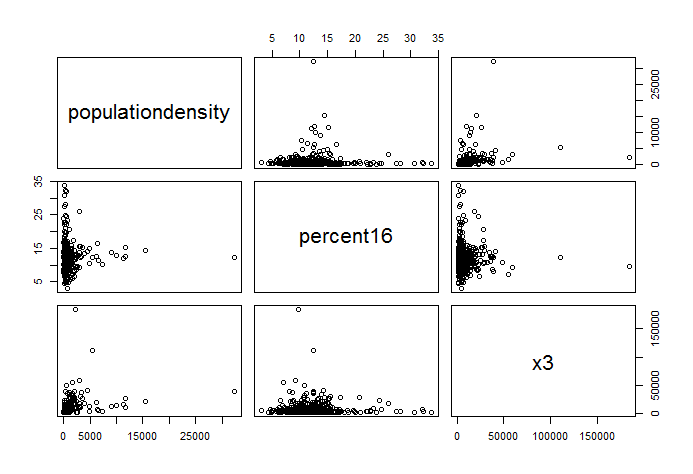
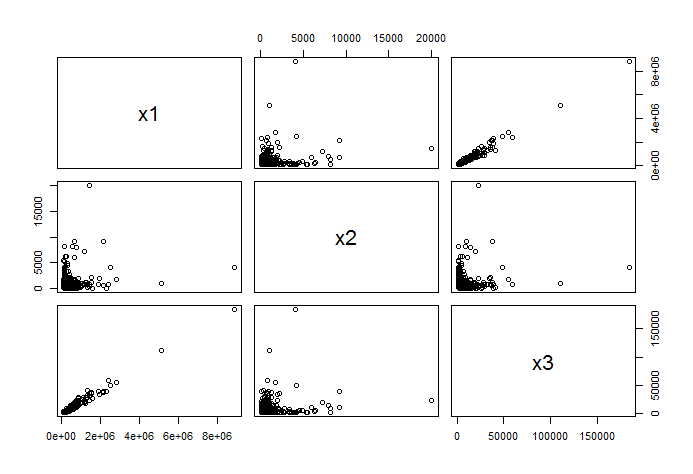
a. Total Population Land Area

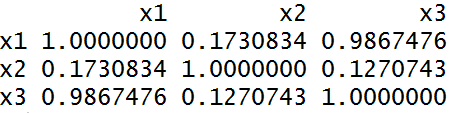
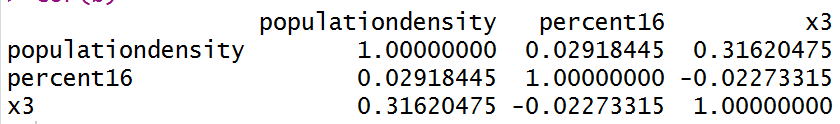


 Total personal income Population density

Percent of population greater than 64 years old

Except for Percent of population greater than 64 years old, all other predictors have outstanding outliers. Most of the data in different predictors concentrate in the first half of the plots.

b. model 1 model 2



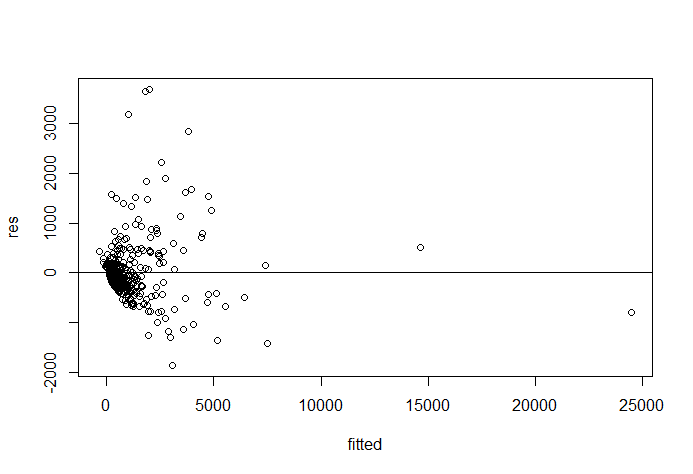
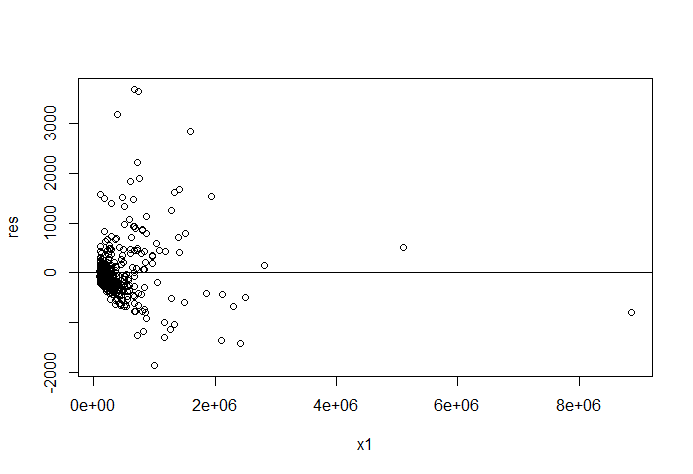
In model I, x1 and x3 are strongly correlated. However, x1 and x2, x2 and x3 have little correlation. In model II, all the predictors do not have strong relation with each other.

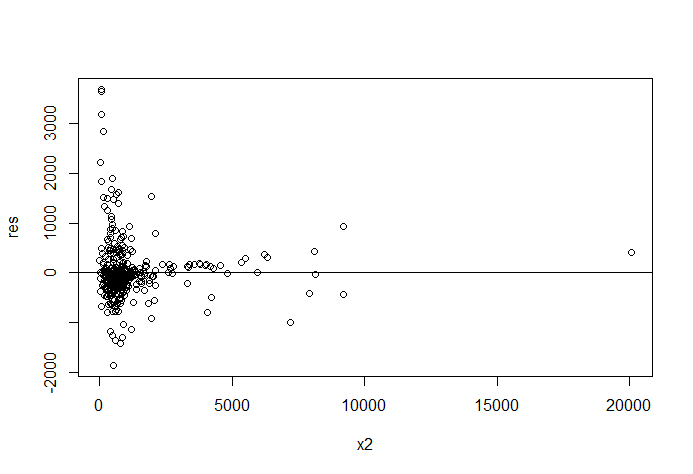
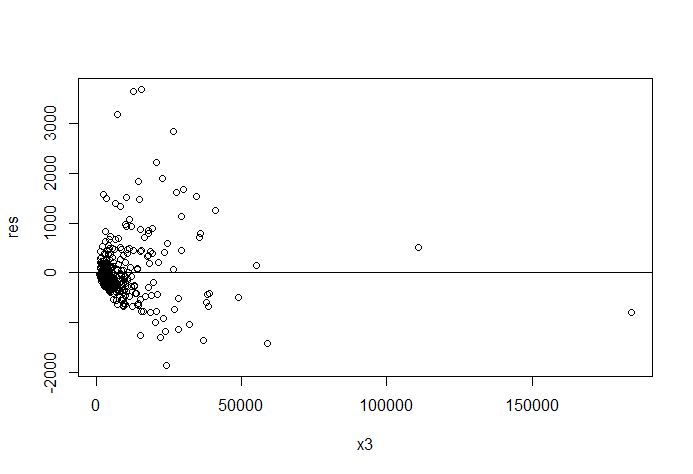
c. model I:

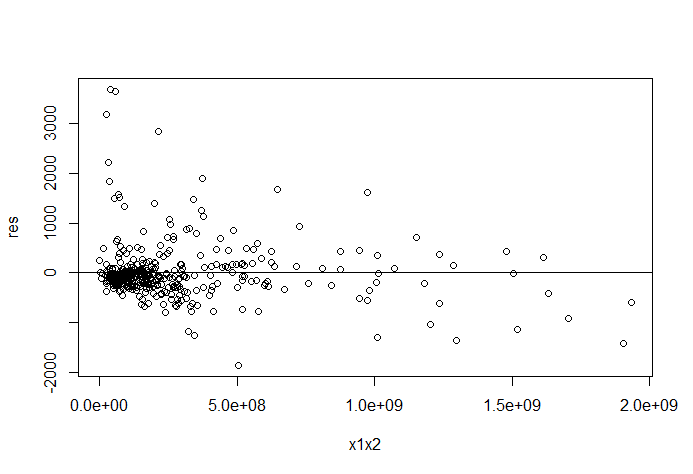
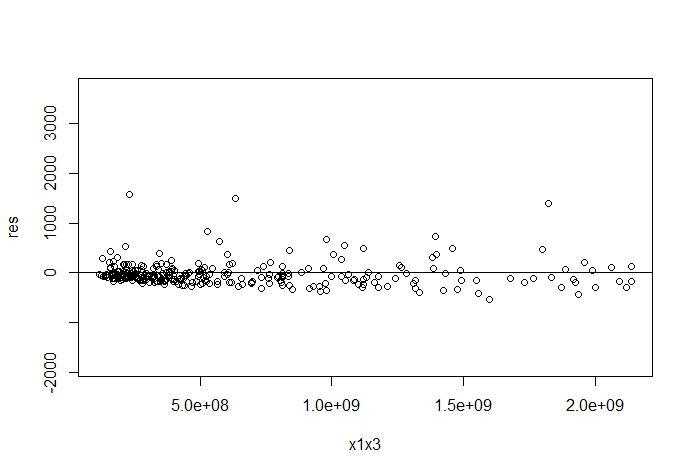
model II:

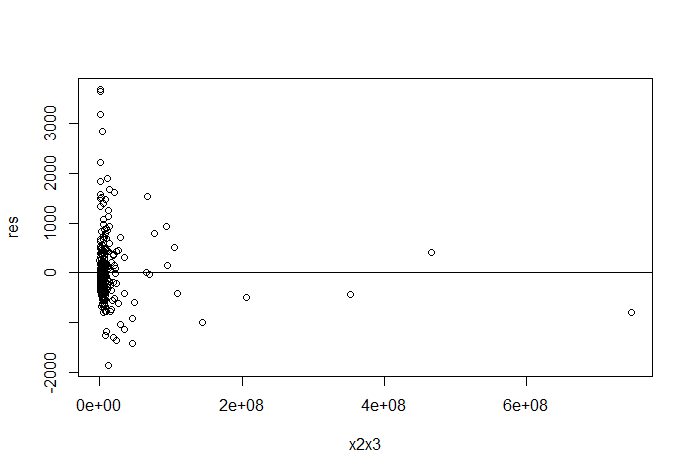
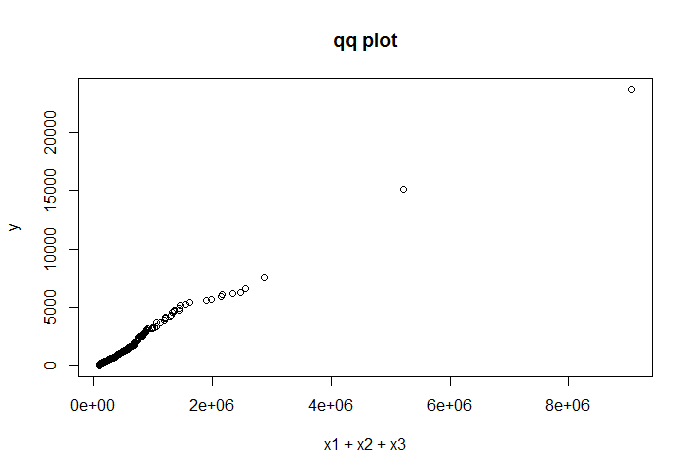
d. model I: R^2=0.903 model II: R^2=0.9117

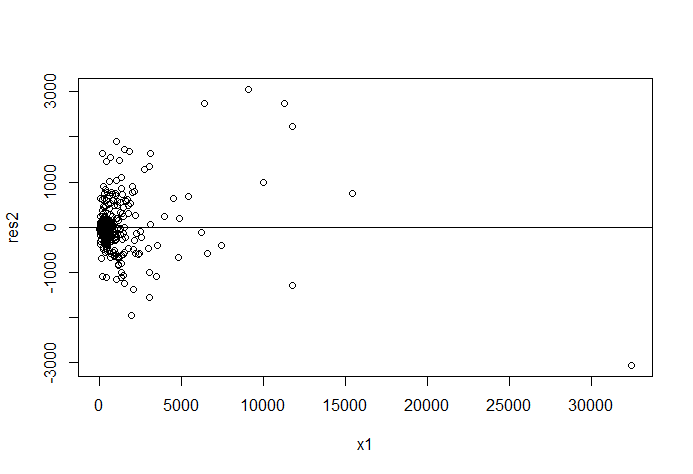
Model I and model II have very close coefficient of determination, but model II is more preferable.

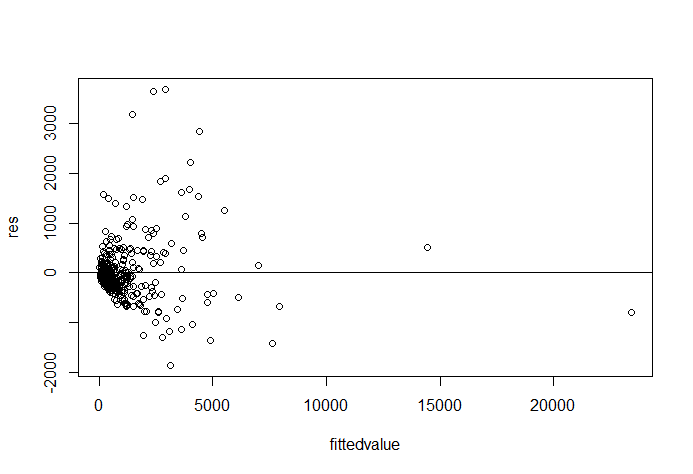
e. Model I

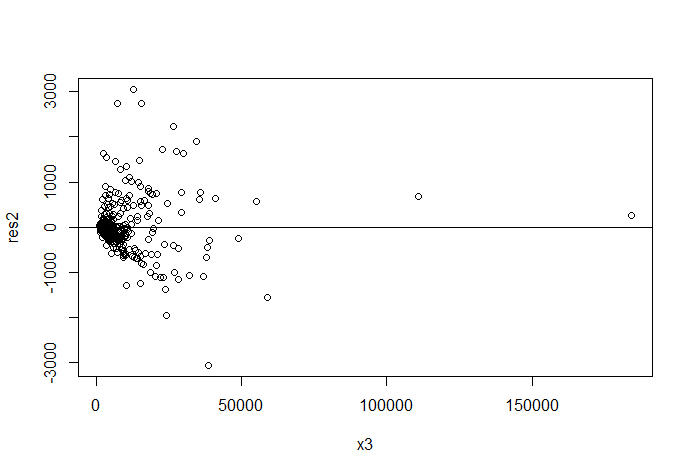
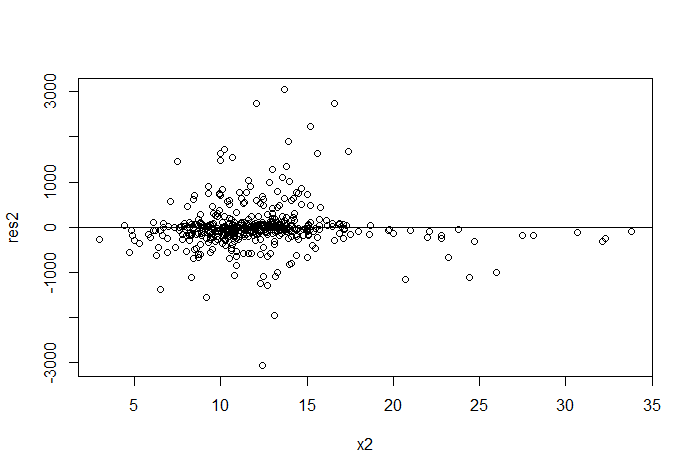


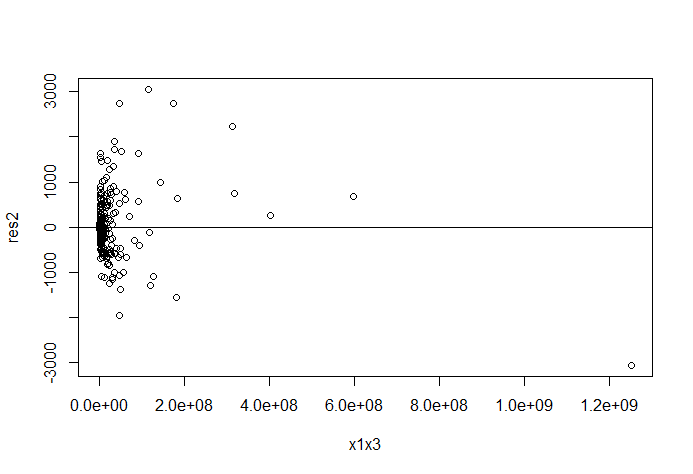


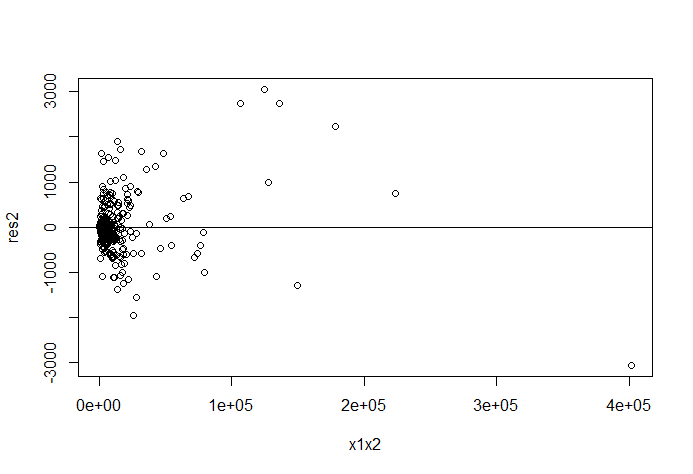


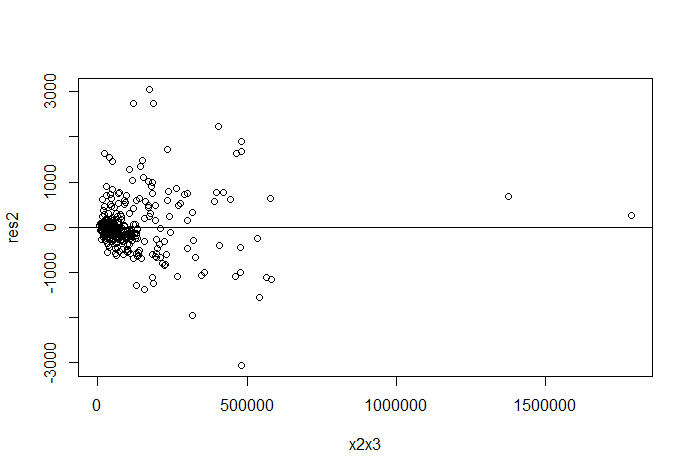
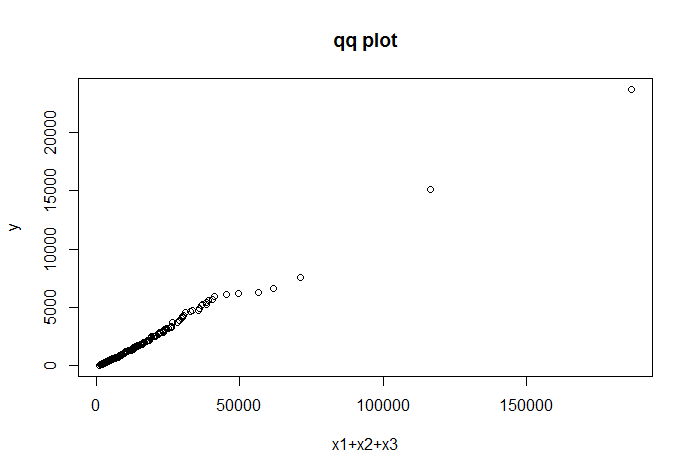


Model II:









In both models, there are some obvious outliers; they both have constant variance. Regarding the plots, both models meet the normality assumption. However, model I’s residual against fitted value plot indicates that the data do not lie evenly around the 0 line. It’s slightly asymmetric. On the other hand, model II’s plot has better pattern and its qqplot follow a straight line strictly despite some outliers. Therefore, model II is more preferable.

f.

model I expand: = 0.357

model II expand:

Part II

7.37 a

Coefficient of partial determination

b. Predictor number of hospital beds (X5) is the best because it has the largest coefficient of partial determination. It means given X1 and X2, around 55% of the variation can be explained by X5. Meanwhile, other two predictors have very small coefficients of partial determination.

c. H0: β5=0 H1: Not H0

F\*=(SSR(X5|X1,X2)/1)/MSE(X1,X2,X5)

=541.18

F1,436(0.99) = 6.69

Decision rule: Reject null hypothesis if F\* > F

Since 541.18 > 6.69, reject null hypothesis. X5 is helpful and should not be dropped.

Discussion: Since (SSR(X3|X1,X2) and (SSR(X4|X1,X2) are much smaller than (SSR(X5|X1,X2) and they all need to be divided by the same SSE, the F\* of them would be much smaller.

d.

0.564

Pair (X4, X5) is relatively more important than the others.

Null Hypothesis: β4 = β5 = 0 Alternate Hypothesis: Not null hypothesis

F\* = 281.6688

F2,435(0.99) = 4.654

Decision rule: Reject null hypothesis if F\* > F

Since 281.6688 > 6.654, we reject null hypothesis. Therefore, the pair is helpful to the model given X1 and X2.

DISCUSSION

In this report, the relationships among the number of active physicians(y) and the predictors (quantitative and qualitative) are analyzed. It was found that interactive predictor like population density can provide more and better information of the relation among y and predictors.

Practically, the hospital beds, population density, percent of population 65 or older, and total personal income has large contribution to the number of active physicians.

Residual plots against different factors and Anova table are the most relevant course material in the analysis. Because most of the time, we are analyzing the data by analyzing MSE, MSR, coefficient of (partial) determination, etc. And residual plots provide more information about the distribution of the data.

By eliminating outliers, enlarging the sample size, adding more interactive predictors (avoid multicollinearity), the linear regression model should be improved.