Multiple Linear Regression Project Using R

By:
Gurteg Singh

Introduction:

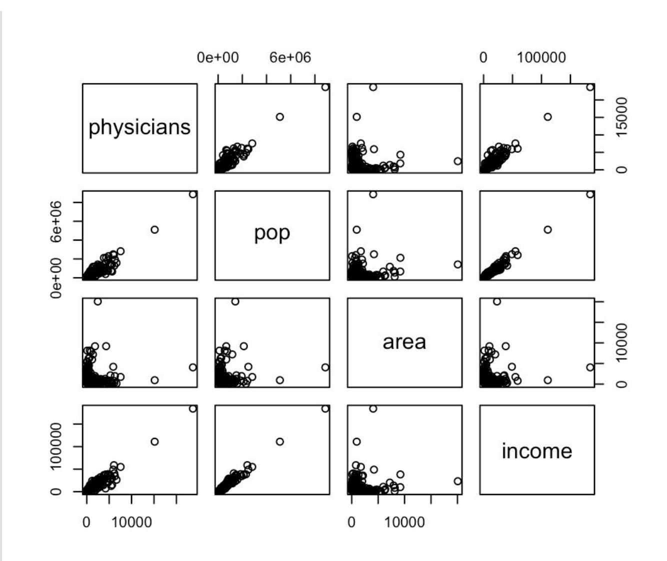
In this project we, once again, analyze multivariate CDI data which includes 16 different variables pertaining to different counties. In our previous project, we analyzed the same data using various R functions including: resid, summary, lm, confint, qqnorm, qqline, abline and anova. We analyzed relationships between the number of hospital beds, physicians and total income in general--and we also did this for four different regions. Ultimately, we uncovered strong linear relationships between presence of college degree and total income per region, as well as a positively correlated relationships between each of the three predictor variables (total income, population, and number of hospital beds) and the number of active physicians in a given county. We did, however, discover through the implementation of a normal probability plot that a different model may fit better.

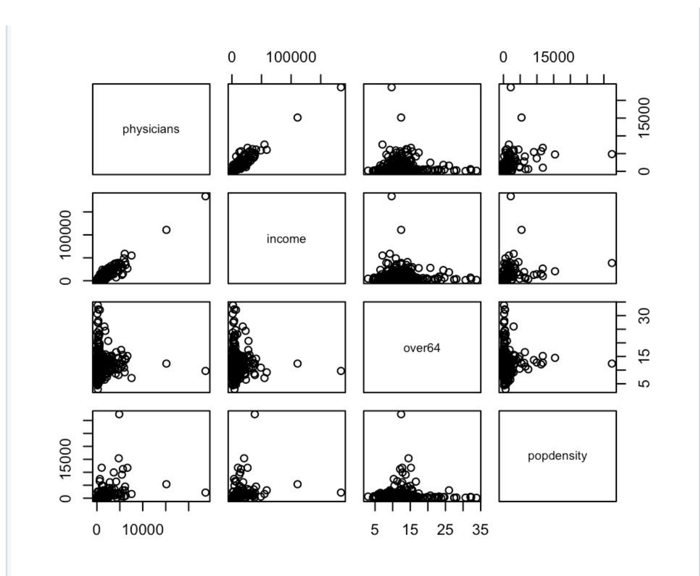
Here, we analyze the same data but at a multivariate level. In part I, we consider two models: model 1 where the predictor variables are total personal income, total land area, and total population and model 2 where the predictor variables are total personal income, proportion of people over the age of 64 and population density. The dependent variable in both cases is the total number of physicians. By analyzing normal probability residual plots, stem and leaf plots, and R-squared values (adjusted and partial), we determine whether or not one particular model fits more appropriately. In Part II, we consider the strength of 3 other predictor variables given total land area and total income. By analyzing respective coefficients of partial determination, a variety of tests for extra SS, and different pairings of the variables-- ultimately determining which variables best strengthen the existing model. We use a variety of functions in R-- among these are pairs, anova, lm, summary, and several graphing functions. Additionally appendix I will contain our computer codes, Appendix II will contain screenshots which demonstrate how our codes are run.

Part I: Multiple Linear Regression 1

6.28

A.Based on the stem and leaf plots, we can get a sense of the range of each of the variables and whether or not there are outliers or gaps. For the population variable, there is a wide range and variation in spread. For the population density variable, we see that the vast majority of counties have a smaller range in density, while there are a handful of outliers, the biggest being at 26.4. For the proportion of people over the age of 64, the spread appears to be approximately normal. For total personal income, the majority of counties are on the lower end, and there are once again some outliers on the higher end-- this is quite similar to the area spread. This information is important because it indicates that for the models which include variables such as population density, personal income, and area where there are outliers-- the model will tend to be less accurate due to the ill-fitting outliers.





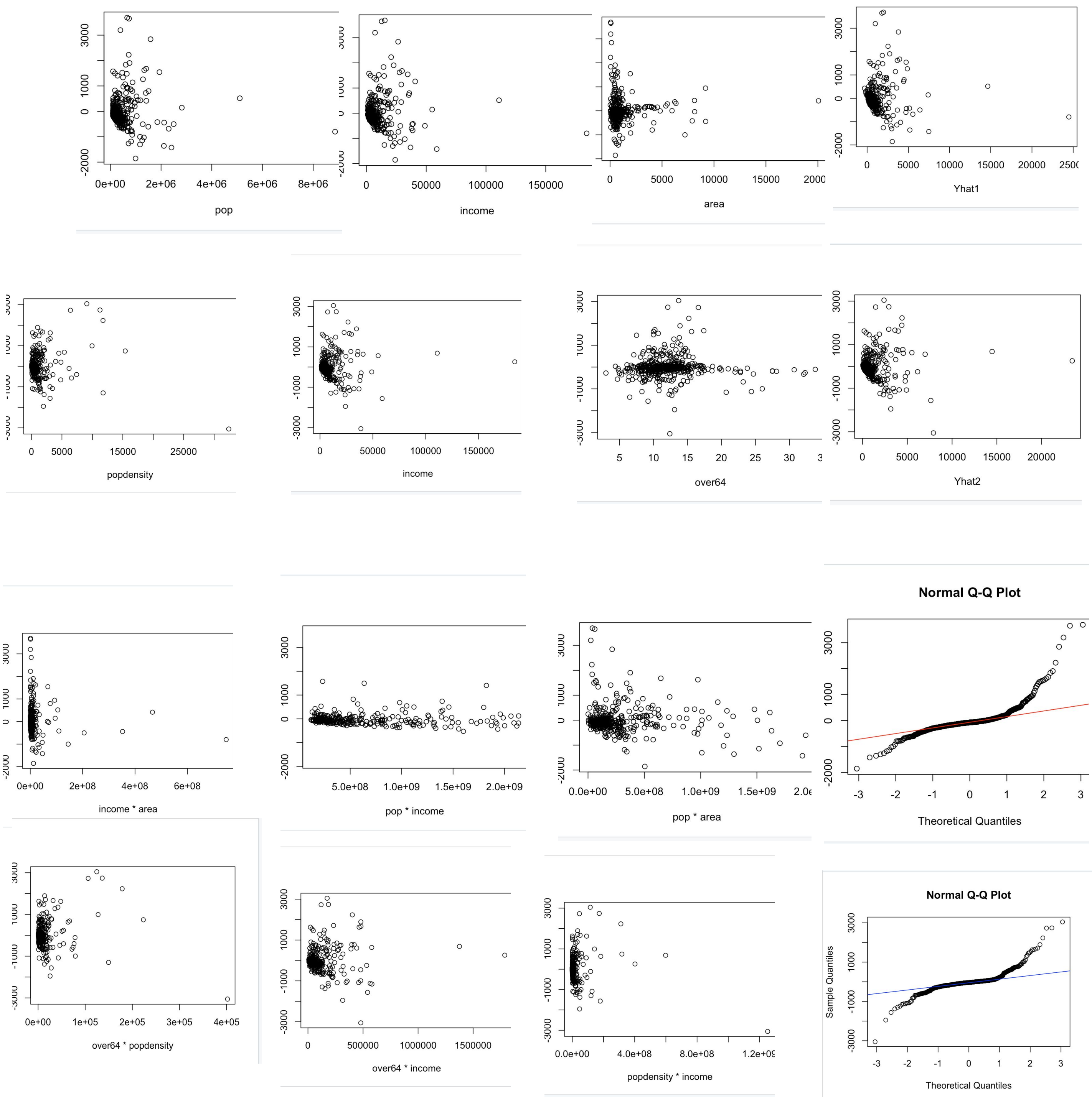
B.The scatter plots indicate that: there is a linear relation between physicians and income, there is a relationship of some sort between income and population, perhaps a relationship between area and population since much of the data is concentrated towards the bottom, relationships between over64 and income and physicians. Based on the correlation table results, there is a strong relationship (.94) between physicians and income, a weak relationship between physicians and over64, population density, and moderate relationship between income and population.

C. Model 1: # of Physicians (Y)= -1.332*10+ 8.366 * 10^-4(population)+ 9.413*10^-2(income) + -6.552 * 10^-2(area)

Model 2: # of Physicians (Y)=-1.706*10^2 +9.616*10^-2(popdensity) + 6.340(over64)+ 1.266*10^-1 (income)

D. For model 1, the R-squared value is 0.9026, for model 2, the R-squared value is 0.911. Given this, model 2 appears to be a slightly stronger model because of its higher R-squared value and R-Squared values indicate the proportion of variation that is explained by the model.

E.



E. The predictor residual plots illustrate, largely, no clear patterns. This indicates that the predictor variables for both models are generally sound and not consistently related to residuals. With the two-factor interactions, it does seem there may be some collinearity between population and area, and perhaps population density and proportion of people over 64 as well as some pattern with over 64 and income. Based on the normal probability plots, these models appear to fit the normal probability model equally poorly on low and high outlier values-- where we would expect to see a mostly straight line if the normal probability model were a very good fit, we, for both see deviation from the qqline. Finally, based on this information, it is unclear whether one model is stronger or weaker as they are mostly similar.

F.

When adding ‘total population * land area’ + ‘total population* total personal income’ + ‘total personal income * land area’ to model 1, we get an adjusted R-squared of 0.9051

When adding ‘percent of population 64 or older * population density’ +’ percent of population 64 or older * total personal income’ + ‘population density * total personal income’ to model 2, we get an adjusted R-squared of 0.922

Between the above models, the latter is preferred (model 2), due to the lower fraction of unexplained variance.

Part II: Multiple Linear Regression 2

7.37

A.

Coefficient of partial determination for ‘land area’, given total population and total personal income = (140967081 - 136903711) / 140967081 = 0.0288

Coefficient of partial determination for ‘Percent of population of 64 or older’, given total population and total personal income = (140967081 - 140425434) / 140967081 = 0.004

Coefficient of partial determination for ‘number of hospital beds’, given total population and total personal income = (140967081 - 62896949) / 140967081 = 0.554

B.

Based on the results in part A, the ‘number of hospital beds’ is the best variable predictor, giving us the highest coefficient of partial determination, 0.554. Yes, the extra sum of squares associated with this variable is the greatest since this variable has the least amount of unexplained variance compared to the other two.

C.
Ho: B3 = 0 (drop variable from model)
Ha: B3 is not = 0 (include variable in equation)
Decision Rule: If the probability of the F statistic of the reduced model is less than alpha (a = 0.01), reject Ho.
Pr(F*) = 2.2e-16 < 0.01, reject Ho.
No, the F* statistics of the other models would not be as large as this model's (541.18) since their sum of explained errors would be smaller.
D.
Coefficient of partial determination for 'land area' & 'Percent of population of 64 or older', given total population and total personal income;
(140967081 - 136295177) / 140967081 = 0.033

Coefficient of partial determination for 'land area' & 'number of hospital beds', given total population and total personal income;
(140967081 - 62614306) / 140967081 = 0.556

Coefficient of partial determination for 'Percent of population of 64 or older' & 'number of hospital beds', given total population and total personal income;
(140967081 - 61422794) / 140967081 = 0.564

For the given pairs, (land area & number of hospital beds) and (Percent of population of 64 or older & number of hospital beds) are relatively more important than (land area & Percent of population of 64 or older). Between (land area & number of hospital beds) and (Percent of population of 64 or older & number of hospital beds), they are relatively similar with the latter having a slightly higher coefficient of partial determination.

Using the F test,

Ho: B3, B4 = 0 (Use reduced model)
Ha: B3 or B4 is not = 0 (Use full model)
Decision Rule: If the probability of the F statistic of the reduced model is less than alpha, reject Ho.
Pr(F*) = 2.2e-16 < 0.01, reject Ho.
In this case, when we add the best pair (Percent of population of 64 or older & number of hospital beds) to the model with total population and total personal income, the SSR of the model substantially increases.

Part III: Discussion

From a practical standpoint, in 6.28, we take the data, and first observe for any outliers using the stem and leaf plot, this reveals that there are several outliers within the majority of the variables. Then after analyzing their residual trends on graphs to see what variables and two factor interaction terms correlate with each other and impact the graph more so than others, we create first order regression models to try and create the best line of fit. From part A, we saw the basic outliers, from part B, we saw the variable correlations with our data, from part C we created our models of best fit using the tested variables, part D had us check the effectiveness of our models, part E had us then check our residuals for any patterns that could occur between the variables themselves, and onto the models, and finally part F had us conclude which model should be chosen so as to account for the highest amount of explained error., this model was model 2 which included the variables of population density, proportion of population over the age of 64, and total personal income.

In 7.38, we first started by testing a models effectiveness when adding additional variables, and analyzing our outcomes. We then tested the best variables in the model, and checked whether or not they made a difference in the regression functions with an F test. Finally we checked with pairs of variables and analyzed whether or not the pairs would help the overall data and if they are helpful for the line of best fit with our data. Part A had us calculate the difference adding certain variables made, part B had us choose the best variable by checking for the least amounts of unexplained errors, part C had us test our most effective variable by putting it into our equation and seeing what impact it had versus our full model. Finally, part D had us calculate the difference pairs of variables could make given our original variables, and test whether or not it was helpful in adding these additional variable pairs. Most of these parts were relevant to our analysis, but calculations of our coefficients of partial determination, and single and two factor variable interactions had the biggest impact on our lines of best fit, and data. To improve our data, we would first have to. To improve our linear regression models, we could first input more data, that would let us further refine our understanding of certain variables. Second, we could try adding additional variables that were not included in the problems, to see if there could be any other factors that determine our data to be the way it is, and shape our regression models further.

Appendix II

6.27

Part A

```
> stem(popdensity)

  The decimal point is 3 digit(s) to the right of the |

   0 | 00000000000000001111111111111111111111111111111111111111111111111111111+321
   2 | 000011122334567001111 45
   4 | 05884
   6 | 2464
   8 | 19
  10 | 378
  12 |
  14 | 4
  16 |
  18 |
  20 |
  22 |
  24 |
  26 |
  28 |
  30 |
  32 | 4

>
```

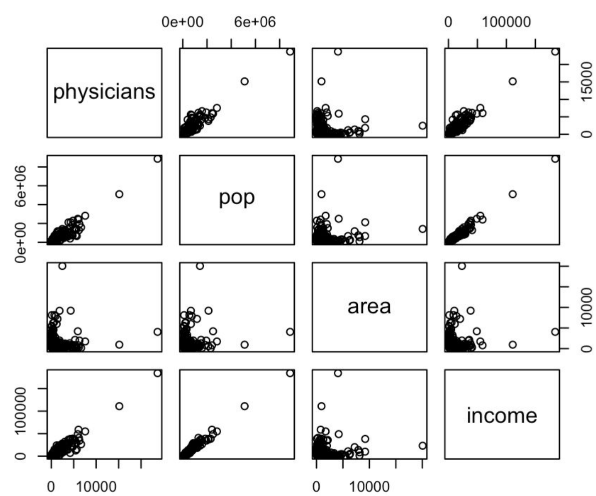
```
> stem(income)

  The decimal point is 4 digit(s) to the right of the |

   0 | 11111111111112222222222222222222222222222222222222222222222222222222+263
   1 | 00000000000011111111222223333444444555555567788888888899 9
   2 | 0011112333444777888 99
   3 | 0255678899
   4 | 19
   5 | 59
   6 |
   7 |
   8 |
   9 |
  10 |
  11 | 1
  12 |
  13 |
  14 |
  15 |
  16 |
  17 |
  18 | 4
```

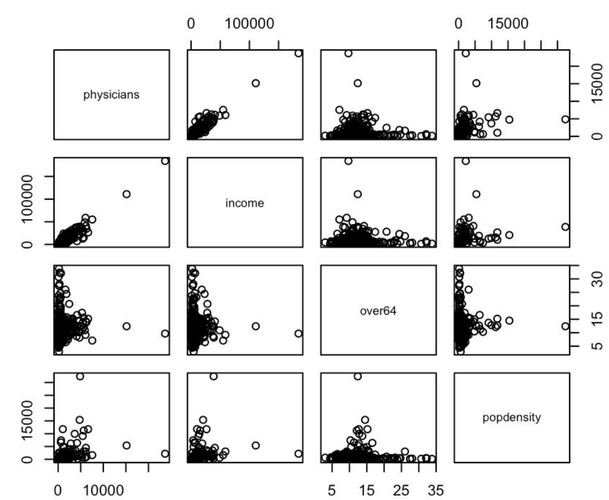
```
> stem(area)

  The decimal point is 3 digit(s) to the right of the |

   0 | 0000111111111111222222222222222222233333333333333333333333444444+252
   1 | 000000000000001111111111111122222222222223333334445566667778889999
   2 | 0001111466778
   3 | 3344688
   4 | 00122368
   5 | 45
   6 | 023
   7 | 29
   8 | 11
   9 | 22
  10 |
  11 |
  12 |
  13 |
  14 |
  15 |
  16 |
  17 |
  18 |
  19 |
  20 | 1

> stem(pop)
```

```
> stem(pop)

  The decimal point is 6 digit(s) to the right of the |

  0 | 1111111111111111111111111111111111111111111111111111111111111111+254
  0 | 55555555555555555555556666666666666667777777777777777777888888888 8
  1 | 00000122233333444
  1 | 55699
  2 | 1134
  2 | 58
  3 |
  3 |
  4 |
  4 |
  5 | 1
  5 |
  6 |
  6 |
  7 |
  7 |
  8 |
  8 | 9

>
```

Part B

```
>
> model1=[,data1:4]
Error: unexpected '[' in "model1=["
> model1=data[,1:4]
> model2=data[,c(1,4,5,6)]
> pairs(model2)
> pairs(model1)
> cor(model1)
           physicians       pop       area    income
physicians 1.00000000 0.9402486 0.07807466 0.9481106
pop        0.94024859 1.0000000 0.17308335 0.9867476
area       0.07807466 0.1730834 1.00000000 0.1270743
income     0.94811057 0.9867476 0.12707426 1.0000000
> cor(model2)
            physicians      income      over64 popdensity
physicians  1.00000000  0.94811057 -0.00312863 0.40643863
income      0.94811057  1.00000000 -0.02273315 0.31620475
over64     -0.00312863 -0.02273315  1.00000000 0.02918445
popdensity  0.40643863  0.31620475  0.02918445 1.00000000
```





Part C

```
> fit1= lm(physicians~pop+area+income)
> fit2= lm(physicians~income+over64+popdensity)
```

```
> summary(fit1)

Call:
lm(formula = physicians ~ pop + area + income)

Residuals:
    Min      1Q  Median      3Q     Max 
-2005.5  -220.3   -45.7    84.5  3711.1 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) 101.428198  76.909592   1.319  0.18793    
pop          -0.392068   0.257413  -1.523  0.12846    
area         -0.050013   0.017520  -2.855  0.00451 ** 
income        0.130267   0.002553  51.022  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 564.3 on 436 degrees of freedom
Multiple R-squared:  0.9013,	Adjusted R-squared:  0.9006 
F-statistic:  1327 on 3 and 436 DF,  p-value: < 2.2e-16

> summary(fit2)

Call:
lm(formula = physicians ~ income + over64 + popdensity)

Residuals:
    Min      1Q  Median      3Q     Max 
-1928.6  -198.7   -69.9    45.7  3801.0 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) -1.699e+02  9.024e+01  -1.882   0.0605 .  
income       1.321e-01  2.119e-03  62.325   <2e-16 ***
over64       8.653e+00  6.807e+00   1.271   0.2043    
popdensity   2.647e+01  1.922e+01   1.377   0.1691    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 568.8 on 436 degrees of freedom
Multiple R-squared:  0.8997,	Adjusted R-squared:  0.899 
F-statistic:  1304 on 3 and 436 DF,  p-value: < 2.2e-16

> 
```

Part D

Part E

```
> residuals1= fit1$residuals
> Yhat1= fitted.values(fit1)
> Yhat2=fitted.values(fit2)
> residuals2=fit2$residuals
> resid.plot
Error: object 'resid.plot' not found
> plot(x=Yhat1, y=residuals1)
> plot(x=Yhat2, y=residuals2)
> plot(x=pop,y=residuals1)
> plot(x=area, y=residuals1)
> plot(x=income, y=residuals1)
> plot(x=over64, y=residuals2)
> plot(x=income, y=residuals2)
> plot(x=popdensity, y=residuals2)
> plot(x=popdensity+over64, y=residuals2)
> plot(x=popdensity+income, y=residuals2)
> plot(x=popdensity+over64, y=residuals2)
> plot(x=popdensity+income, y=residuals2)
> plot(x=income+over64, y=residuals2)
> plot(x=pop+area, y=residuals1)
> plot(x=pop*area, y=residuals1)
> plot(x=pop*income, y=residuals1)
> plot(x= income*area, y=residuals1)
> plot(x= popdensity*income, y=residuals2)
> plot(x=popdensity*over64, y=residuals2)
> plot(x=over64*income, y=residuals2)
> qqplot1=qqnorm(residuals1)
> qqline(residuals1, col= "red")
> qqplot=qqnorm(residuals2)
> qqline(residuals2, col= "teal")
Error in int_abline(a = a, b = b, h = h, v = v, untf = untf, ...) :
  invalid color name 'teal'
> qqline(residuals2, col= "blue")
>
```

Part F

```
> fit1wi = lm(physicians ~ pop + area + income + pop*area + pop*income + income*area, data)
> summary(fit1wi)

Call:
lm(formula = physicians ~ pop + area + income + pop * area + 
    pop * income + income * area, data = data)

Residuals:
    Min      1Q  Median      3Q     Max 
-1950.2  -198.0   -61.1    76.6  3578.1 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) -5.826e+01  4.727e+01  -1.232  0.21848    
pop          7.252e-04  3.259e-04   2.225  0.02657 *  
area        -6.421e-02  3.014e-02  -2.131  0.03369 *  
income       1.087e-01  1.450e-02   7.496 3.76e-13 ***
pop:area     6.173e-07  2.058e-07   2.999  0.00287 ** 
pop:income   1.696e-09  1.041e-09   1.630  0.10392    
area:income -3.706e-05  1.152e-05  -3.217  0.00139 ** 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 551.4 on 433 degrees of freedom
Multiple R-squared:  0.9064,    Adjusted R-squared:  0.9051 
F-statistic: 698.7 on 6 and 433 DF,  p-value: < 2.2e-16

> fit2wi = lm(physicians ~ over64 + popdensity + income + over64*popdensity + over64*income + popdensity*income,data)
> summary(fit2wi)

Call:
lm(formula = physicians ~ over64 + popdensity + income + over64 * 
    popdensity + over64 * income + popdensity * income, data = data)

Residuals:
     Min       1Q   Median       3Q      Max 
-2409.57  -163.91   -12.32   103.25  2721.84 

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)    
(Intercept)       -9.367e+00  9.928e+01  -0.094    0.925    
over64            -1.106e+01  7.792e+00  -1.419    0.157    
popdensity        -4.179e-01  1.055e-01  -3.960 8.76e-05 ***
income             1.477e-01  9.739e-03  15.168  < 2e-16 ***
over64:popdensity  4.652e-02  7.925e-03   5.870 8.67e-09 ***
over64:income     -1.289e-03  8.743e-04  -1.474    0.141    
popdensity:income -3.276e-06  7.439e-07  -4.404 1.34e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 500 on 433 degrees of freedom
Multiple R-squared:  0.923,     Adjusted R-squared:  0.922 
F-statistic: 865.4 on 6 and 433 DF,  p-value: < 2.2e-16
```

7.37

Part A & B

```
> m <- lm(physicians ~ pop + income,data)
> anova(m)
Analysis of Variance Table

Response: physicians
           Df     Sum Sq    Mean Sq F value    Pr(>F)    
pop         1 1243181164 1243181164 3853.88 < 2.2e-16 ***
income      1   22058054   22058054   68.38 1.638e-15 ***
Residuals 437  140967081     322579                      
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> m <- lm(physicians ~ pop + income,data)
> m3 <- lm(physicians ~ pop + income + area, data)
> anova(m3)
Analysis of Variance Table

Response: physicians
           Df     Sum Sq    Mean Sq  F value    Pr(>F)    
pop         1 1243181164 1243181164 3959.184 < 2.2e-16 ***
income      1   22058054   22058054   70.249 7.271e-16 ***
area        1    4063370    4063370   12.941 0.0003583 ***
Residuals 436  136903711     313999                       
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> m4 <- lm(physicians~ pop + income + over64, data)
> anova(m4)
Analysis of Variance Table

Response: physicians
           Df     Sum Sq    Mean Sq   F value    Pr(>F)    
pop         1 1243181164 1243181164 3859.8919 < 2.2e-16 ***
income      1   22058054   22058054   68.4870 1.571e-15 ***
over64      1     541647     541647    1.6817    0.1954    
Residuals 436  140425434     322077                        
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> m5 <- lm(physicians ~ pop + income + hospitalbeds, data)
> anova(m5)
Analysis of Variance Table

Response: physicians
              Df     Sum Sq    Mean Sq F value    Pr(>F)    
pop            1 1243181164 1243181164 8617.70 < 2.2e-16 ***
income         1   22058054   22058054  152.91 < 2.2e-16 ***
hospitalbeds   1   78070132   78070132  541.18 < 2.2e-16 ***
Residuals    436   62896949     144259                      
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Part C

```
> anova(m, m5)
Analysis of Variance Table

Model 1: physicians ~ pop + income
Model 2: physicians ~ pop + income + hospitalbeds
  Res.Df       RSS Df Sum of Sq      F    Pr(>F)    
1    437 140967081                                  
2    436  62896949  1  78070132 541.18 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> 
```

Part D

```
> m34 <- lm(physicians ~ pop + income + area + over64, data)
> anova(m34)
Analysis of Variance Table

Response: physicians
           Df     Sum Sq    Mean Sq   F value    Pr(>F)    
pop         1 1243181164 1243181164 3967.7399 < 2.2e-16 ***
income      1   22058054   22058054   70.4005 6.842e-16 ***
area        1    4063370    4063370   12.9687 0.0003533 ***
over64      1     608535     608535    1.9422 0.1641413    
Residuals 435  136295177     313322                        
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> m35 <- lm(physicians ~ pop + income + area + hospitalbeds, data)
> anova(m35)
Analysis of Variance Table

Response: physicians
              Df     Sum Sq    Mean Sq  F value    Pr(>F)    
pop            1 1243181164 1243181164 8636.745 < 2.2e-16 ***
income         1   22058054   22058054  153.244 < 2.2e-16 ***
area           1    4063370    4063370   28.229 1.724e-07 ***
hospitalbeds   1   74289406   74289406  516.110 < 2.2e-16 ***
Residuals    435   62614306     143941                       
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

> m45 <- lm(physicians ~ pop + income + over64 + hospitalbeds, data)
> anova(m45)
Analysis of Variance Table

Response: physicians
              Df     Sum Sq    Mean Sq  F value Pr(>F)    
pop            1 1243181164 1243181164 8804.285 <2e-16 ***
income         1   22058054   22058054  156.216 <2e-16 ***
over64         1     541647     541647    3.836 0.0508 .  
hospitalbeds   1   79002640   79002640  559.502 <2e-16 ***
Residuals    435   61422794     141202                    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```