Adam Hetherwick
Raees Qadir
Russell Su

STA 108 Project 2

Brief Intro:

In Project 1, we were given a data set that represents selected county demographic information (CDI) for 440 of the most populous counties in the United States. Each county has its state, an identification number, and 14 different variables associated with it. We used this data to analyze the relationship between the number of active physicians and three specific variables: the total population, number of hospital beds and total personal income. Not only were we able to fit a linear regression model to each relationship, but we also expanded further on their regression parameters and performed F-tests. Lastly, we went ahead and created residual and normal plots for each of the three variable's relationship to the number of active physicians.

In Project 2, we started by using the same data to evaluate two different models for predicting the number of active physicians in a CDI. The first model included the predictor variables total population, land area, and total personal income, while the second included population density, percentage of population older than 64 years, and total personal income. The analysis included all two-factor interactions as well. We then started a new model with total population and total personal income, and evaluated which of four different predictor variables would be the best to complete the model.

Adam Hetherwick
Raees Qadir
Russell Su

Part I: Multiple linear regression I

This part consists of Project 6.28 in the book:

a. Prepare a stem-and-Leaf plot for each of the predictor variables. What noteworthy information is provided by your plots?

Stem and leaf plot for Total Population (Model I, X_1)

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

0 | 111111111111111111111111111111111111111111111111111111111111111111111+254
0 | 555555555555555555555555566666666666666777777777777777777778888888888
1 | 000000122233333444
1 | 55699
2 | 1134
2 | 58
3 |
3 |
4 |
4 |
5 | 1
5 |
6 |
6 |
7 |
7 |
8 |
8 | 9
```

Adam Hetherwick
Raees Qadir
Russell Su

Stem and leaf plot for Land Area (Model I, X_2)

```
Column: Land Area

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

   0 | 0000111111111112222222222222222222222333333333333333333333333344444+252
   1 | 00000000000000111111111111122222222222223333334444555666677778889999
   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 and leaf plot for Total Personal Income (Model I, X_3):

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

  0 | 11111111111112222222222222222222222222222222222222222222222222222+263
  1 | 0000000000011111111122222333334444444555555556778888888899
  2 | 00111123334447778899
  3 | 0255678899
  4 | 19
  5 | 59
  6 |
  7 |
  8 |
  9 |
 10 |
 11 | 1
 12 |
 13 |
 14 |
 15 |
 16 |
 17 |
 18 | 4
```

Adam Hetherwick
Raees Qadir
Russell Su

Stem and leaf plot for Population Density (Model II, X_1)

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

   0 | 0000000000000000111111111111111111111111111111111111111111111111111+321
   2 | 00001112233456700111145
   4 | 05884
   6 | 2464
   8 | 19
  10 | 378
  12 |
  14 | 4
  16 |
  18 |
  20 |
  22 |
  24 |
  26 |
  28 |
  30 |
  32 | 4
```

Stem and leaf plot for percent of people older than 64 years old (Model II, X_2):

```
The decimal point is at the |

   2 | 0
   4 | 47890389
   6 | 1123455677990134566678899
   8 | 0011222223333444455566677777888889999000222233333344444444555666677
  10 | 0001111122222222223333334444445555555666666677777778888888888899999+36
  12 | 0000000011111222233333333334444555555566666677777777788889990000000+36
  14 | 0000111111223334444455567788900000011112222345566 7778
  16 | 12556699901122345
  18 | 06778
  20 | 070
  22 | 018828
  24 | 47
  26 | 055
  28 | 1
  30 | 7
  32 | 138
```

Adam Hetherwick
Raees Qadir
Russell Su

<u>Stem and leaf plot for total personal income (Model II, X_3):</u>

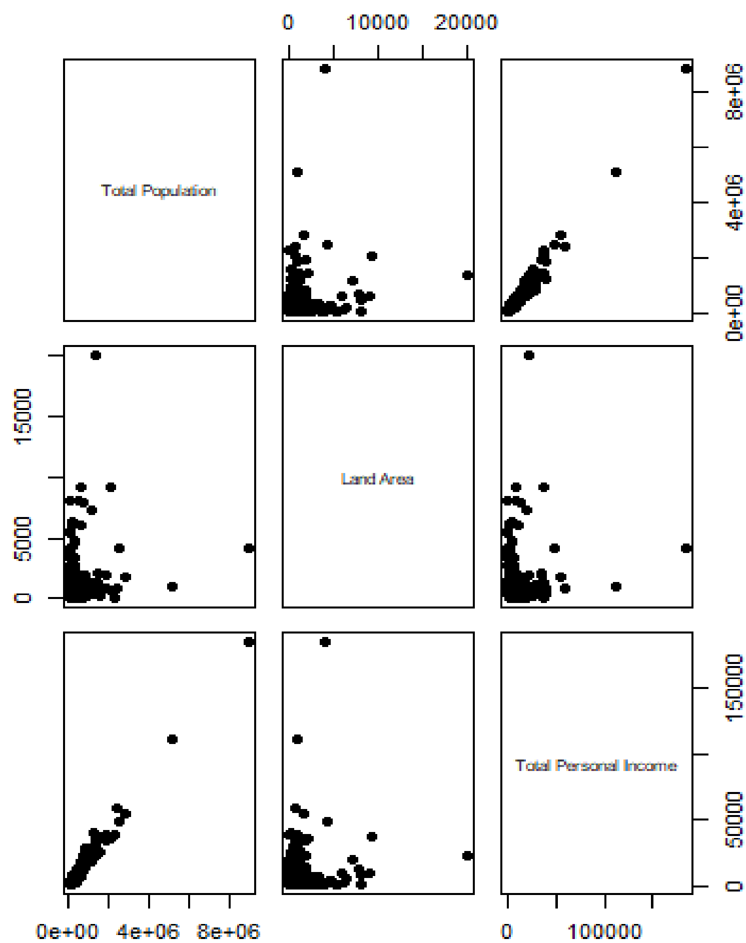
```
 The decimal point is 4 digit(s) to the right of the |

   0 | 11111111111112222222222222222222222222222222222222222222222222222222+263
   1 | 000000000000111111111222223333344444445555555567788888888999
   2 | 001111233344477788899
   3 | 0255678899
   4 | 19
   5 | 59
   6 |
   7 |
   8 |
   9 |
  10 |
  11 | 1
  12 |
  13 |
  14 |
  15 |
  16 |
  17 |
  18 | 4
```

The stem and leaf plots for each predictor variable all appear to follow a skewed distribution, with the lower end having most of the data and the higher numbers have less frequency. This indicates that a majority of the 440 counties are similar to one another in terms of said predictor variables, with a few outlier counties.

Adam Hetherwick
Raees Qadir
Russell Su

b. Obtain the scatter plot matrix and the correlation matrix for each proposed model. Summarize the information provided.

Scatter plot matrix of Model I

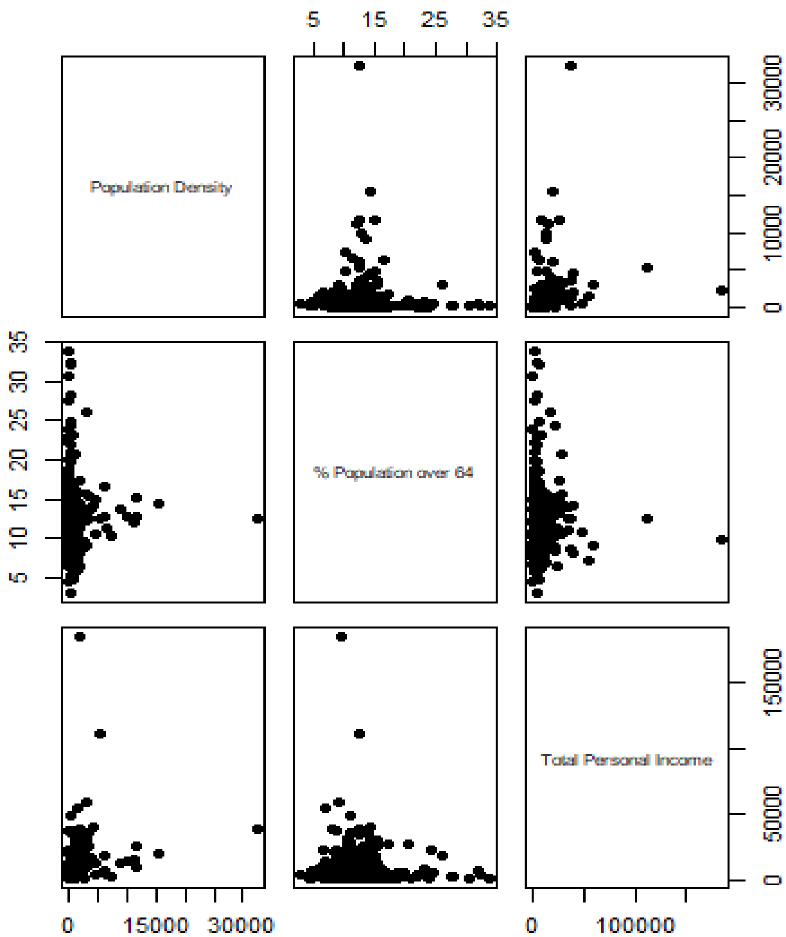


Correlation matrix of Model I

	total_pop	land_area	total_personal_income
total_pop	1.0000000	0.1730834	0.9867476
land_area	0.1730834	1.0000000	0.1270743
total_personal_income	0.9867476	0.1270743	1.0000000

Adam Hetherwick
Raees Qadir
Russell Su

Scatter plot matrix of Model II



Correlation matrix of Model II

	pop_density	perc_pop_65_older	total_personal_income
pop_density	1.00000000	0.02918445	0.31620475
perc_pop_65_older	0.02918445	1.00000000	-0.02273315
total_personal_income	0.31620475	-0.02273315	1.00000000

Adam Hetherwick
Raees Qadir
Russell Su

c. For each proposed model, fit the first-order regression model (6.5) with three predictor variables.

Model I

```
Call:
lm(formula = physicians ~ total.pop + land.area + total.income,
    data = CDI)

Coefficients:
 (Intercept)     total.pop     land.area  total.income
  -1.332e+01     8.366e-04    -6.552e-02     9.413e-02
```

Model II

```
Call:
lm(formula = physicians ~ pop.density + pop.over65 + total.income,
    data = CDI)

Coefficients:
 (Intercept)   pop.density    pop.over65  total.income
  -170.57422       0.09616       6.33984       0.12657
```

Adam Hetherwick
Raees Qadir
Russell Su

d. Calculate R^2 for each model. Is one model clearly preferable in terms of this measure?

<u>Model I</u>

```
Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)  -1.332e+01  3.537e+01  -0.377 0.706719
total.pop     8.366e-04  2.867e-04   2.918 0.003701 **
land.area    -6.552e-02  1.821e-02  -3.597 0.000358 ***
total.income  9.413e-02  1.330e-02   7.078 5.89e-12 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 560.4 on 436 degrees of freedom
Multiple R-squared:  0.9026,	Adjusted R-squared:  0.902
F-statistic:  1347 on 3 and 436 DF,  p-value: < 2.2e-16
```

R^2= 0.9026

<u>Model II</u>

```
Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)  -1.706e+02  8.353e+01  -2.042   0.0418 *
pop.density   9.616e-02  1.224e-02   7.857  3.1e-14 ***
pop.over65    6.340e+00  6.384e+00   0.993   0.3212
total.income  1.266e-01  2.084e-03  60.723  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 533.5 on 436 degrees of freedom
Multiple R-squared:  0.9117,	Adjusted R-squared:  0.9111
F-statistic:  1501 on 3 and 436 DF,  p-value: < 2.2e-16
```

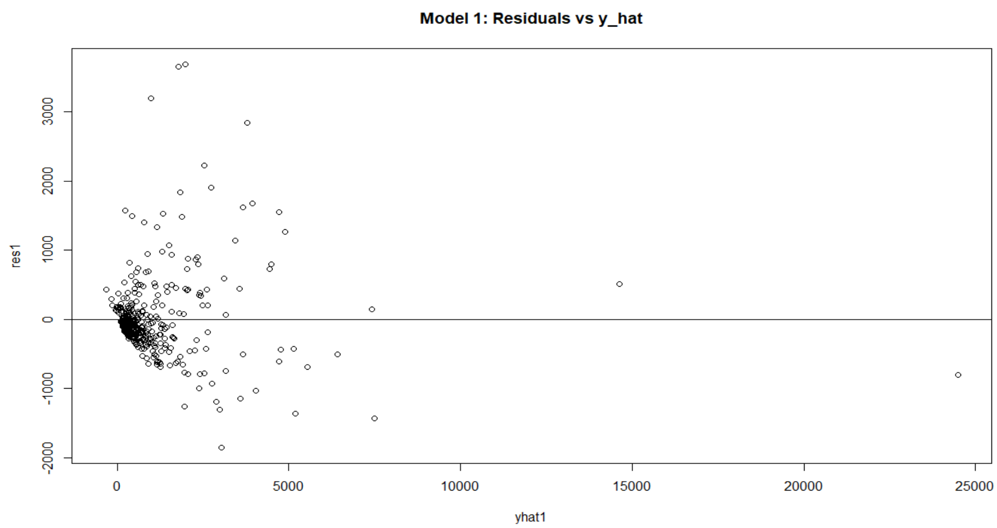
R^2= 0.9117

Based on R^2 alone, neither model is clearly preferable, as the R^2 are very close (less than 0.01 apart)

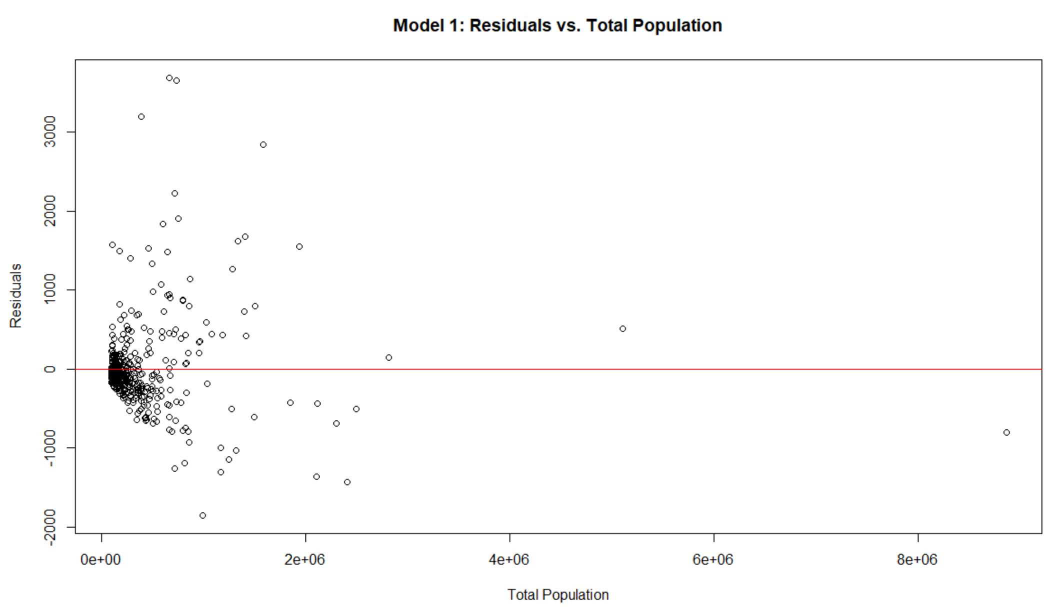
Adam Hetherwick
Raees Qadir
Russell Su

e. For each model, obtain the residuals and plot them against Y, each of the three predictor variables, and each of the two-factor interaction terms. Also prepare a normal probability plot for each of the two fitted models. Interpret your plots and state your findings. Is one model clearly preferable in terms of appropriateness?

<u>Plots for Model 1:</u>

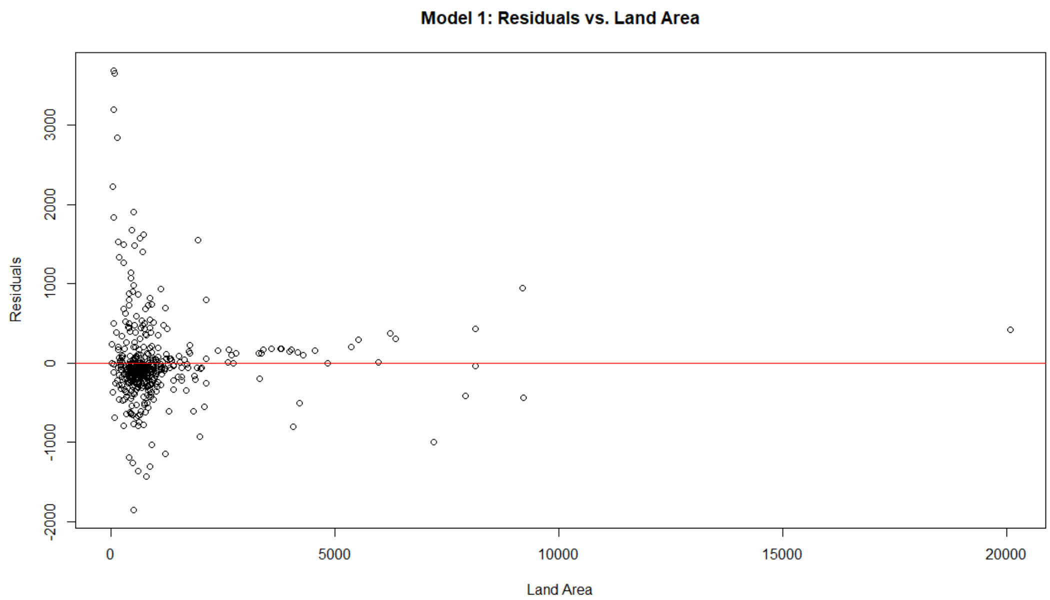


The residuals plotted over the expected values of number of active physicians shows a cluster around x = 0 though it clumps up as $\hat{Y}$ approaches 0. The cluster is a good sign of further analysis. .

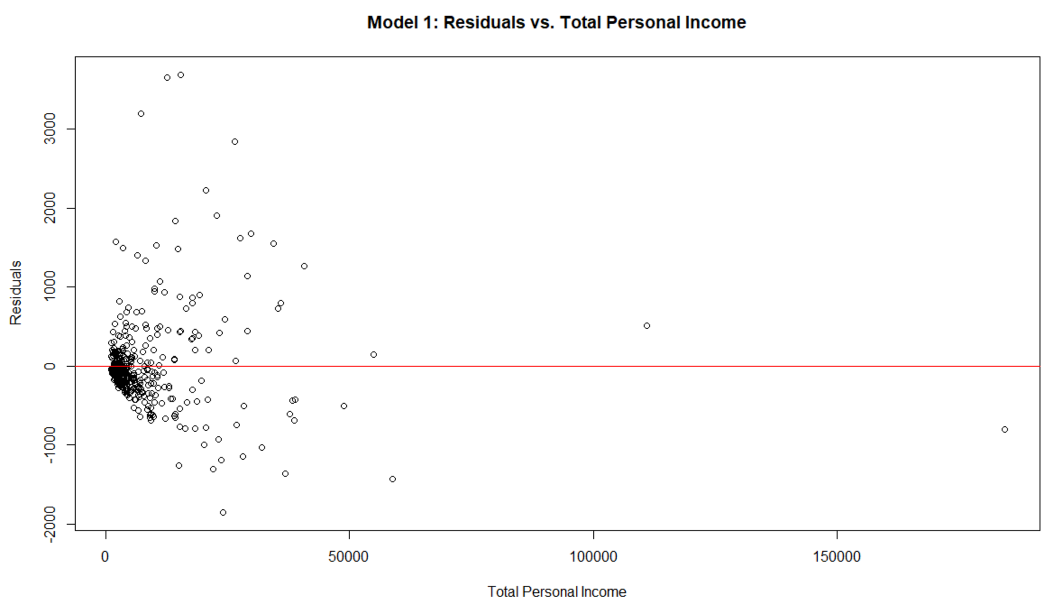


The residuals plotted over total population shows no relative pattern which is good, and also clusters around where $\hat{Y}$ would equal 0.

Adam Hetherwick
Raees Qadir
Russell Su

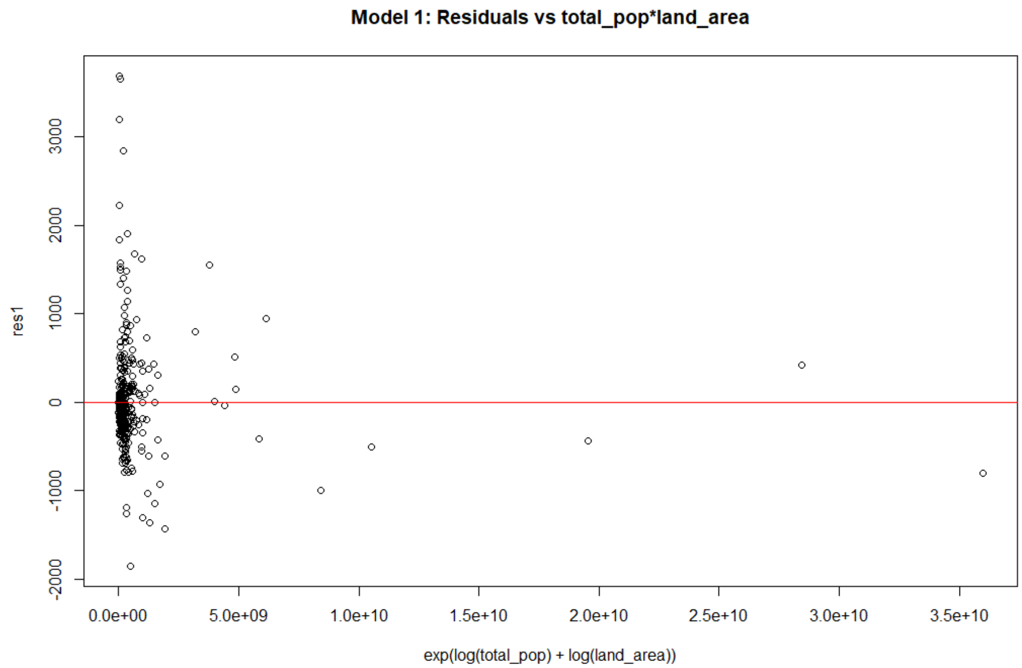


The residuals plotted over land area shows a clump around where land area approaches 0, then after that it appears to thin out. This suggests that as land area decreases, the variation in active physicians may also decrease.

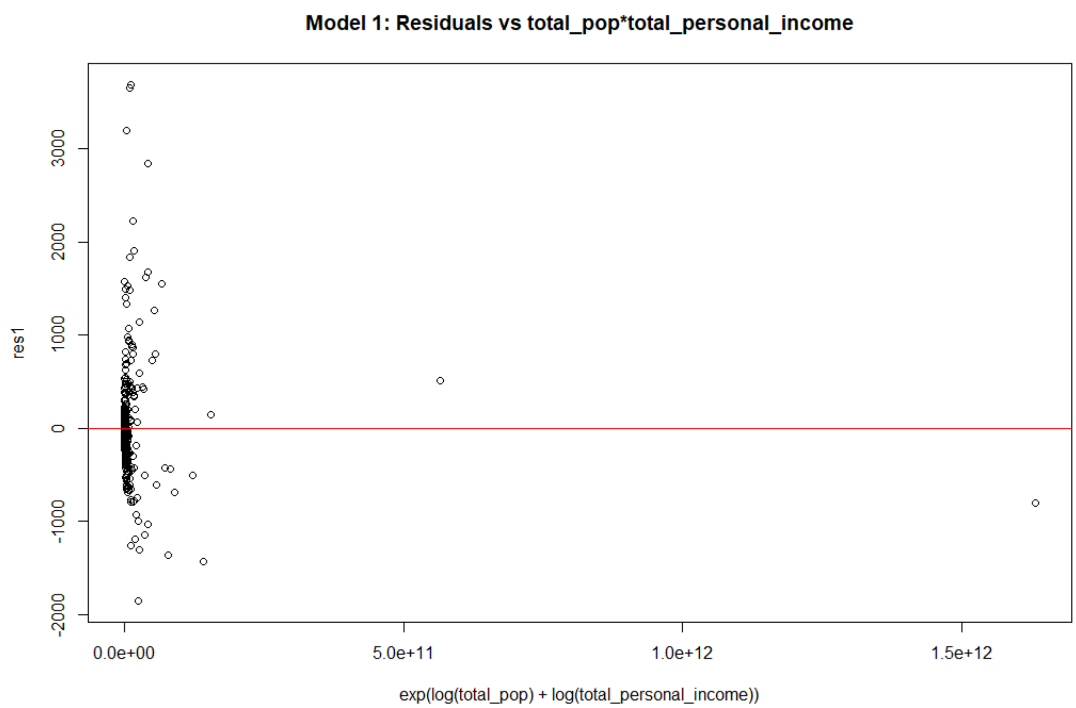


The residuals plotted over total personal income shows a clump around where total personal income approaches 0, then scatters out as total personal income increases.

Adam Hetherwick
Raees Qadir
Russell Su

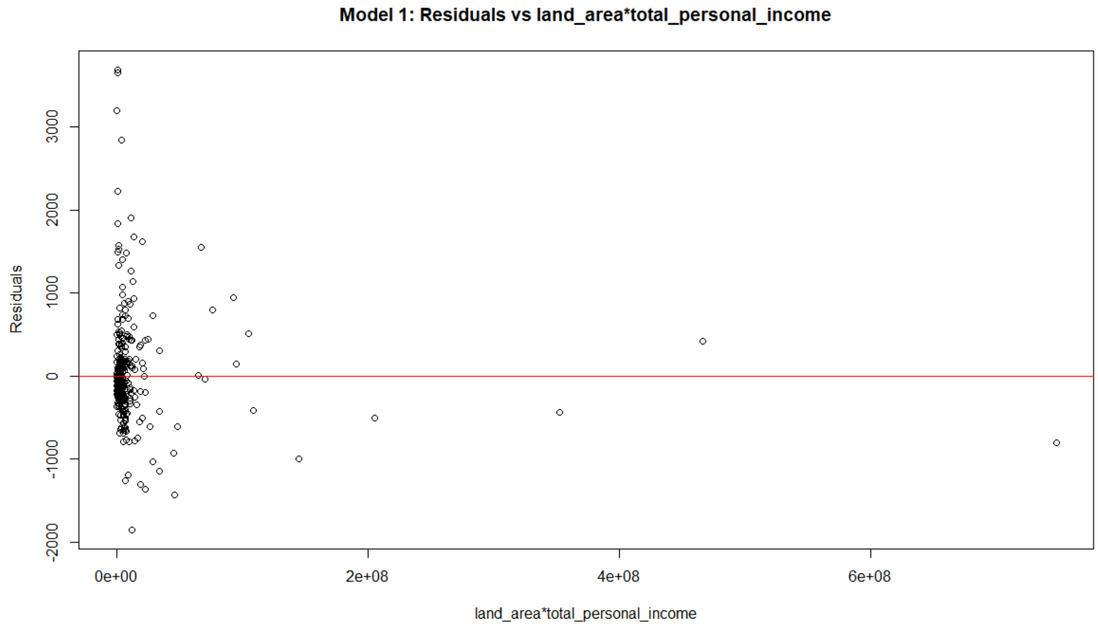


When examining the residuals over the total population and land area, we notice a cluster around where the total population and land area is nearly 0, with high variability.

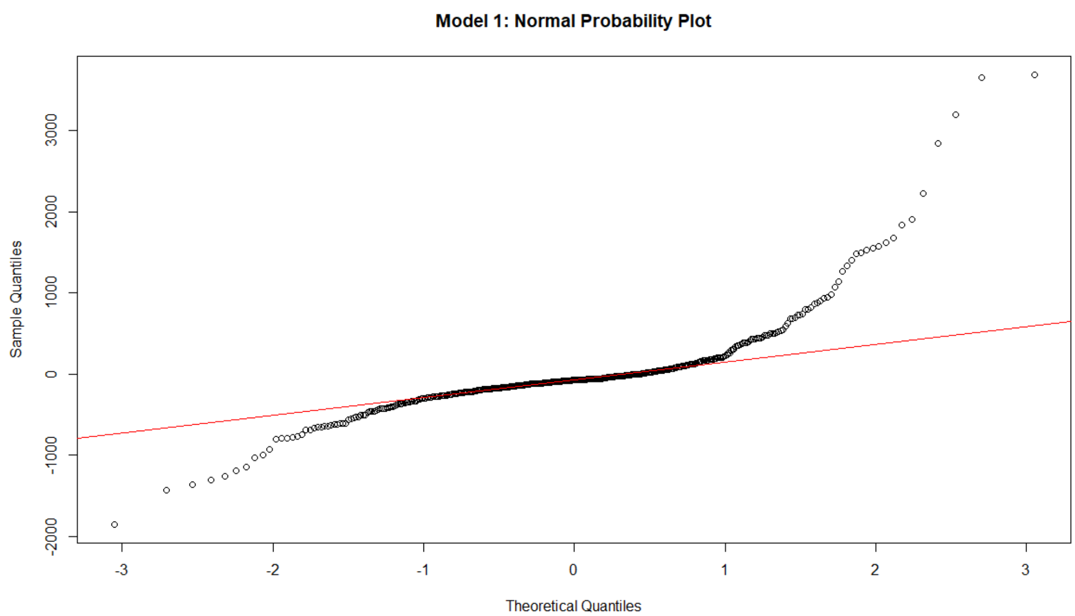


When plotting our residuals over total population and total personal income, a large cluster around where total population and total personal income is 0.

Adam Hetherwick
Raees Qadir
Russell Su



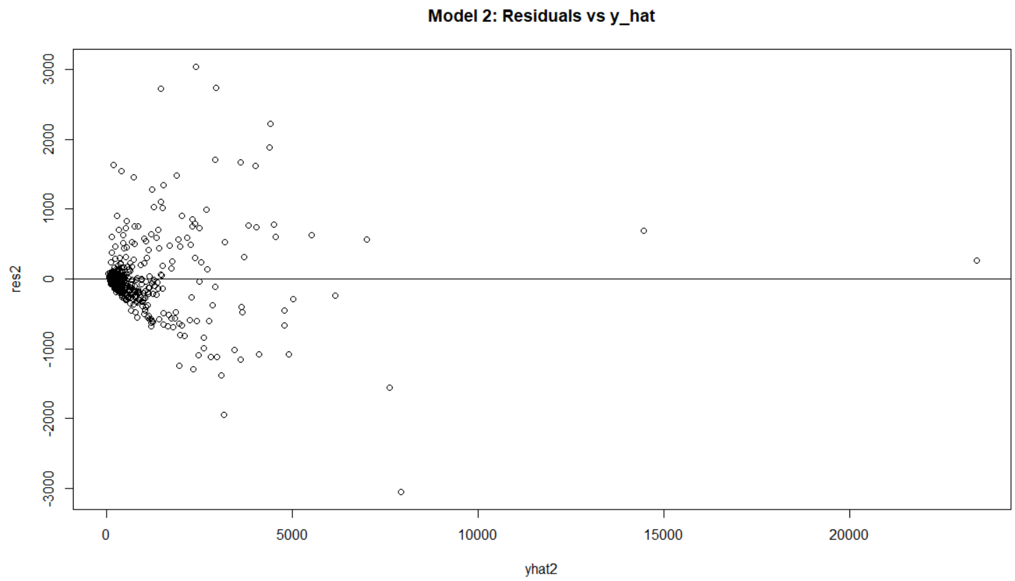
Looking at the residuals over land area and total personal income, we see that there is a massive cluster with high variability around where land area and total personal income approach 0, and a thin cluster moving forward. This suggests that land area may not be that predictive of a variable.



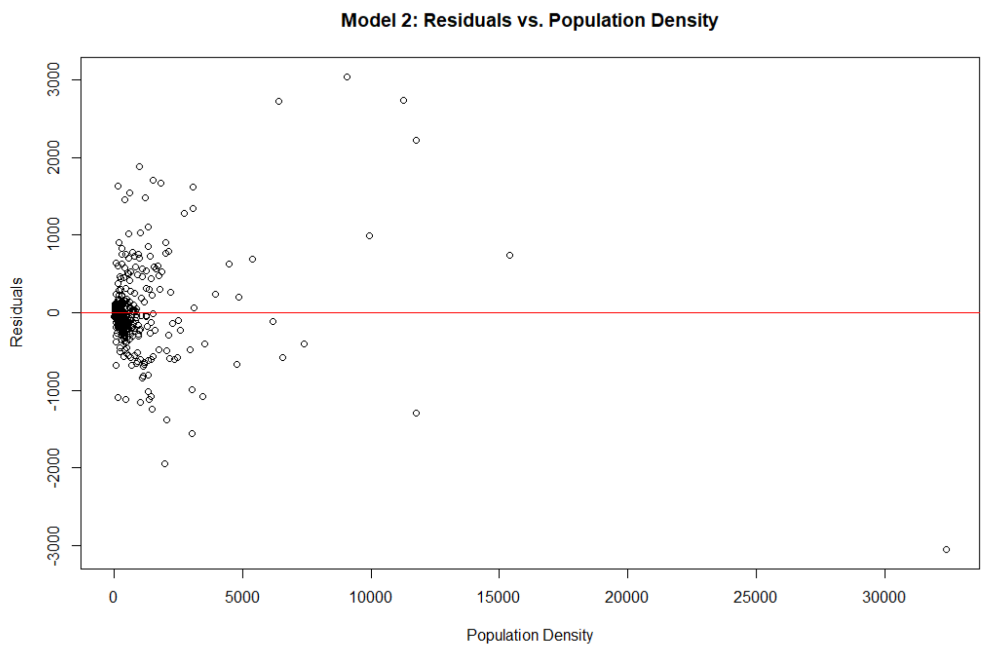
The normal probability plot for Model 1 signifies that there is a strong linear relationship around quantiles [-2, 2] though after that the residuals begin to tail off and are thus more difficult to predict.

Adam Hetherwick
Raees Qadir
Russell Su

<u>Plots for Model 2:</u>

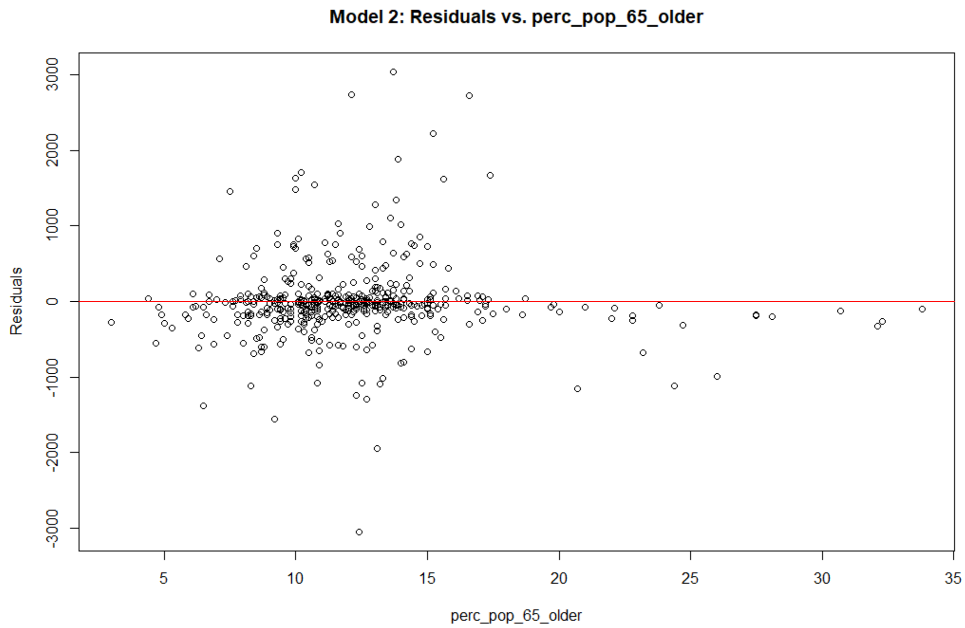


When examining the residuals over the expected values of active physicians, we notice a large cluster around where $\hat{Y}$ equals 0, then a highly variable cluster as $\hat{Y}$ increases.

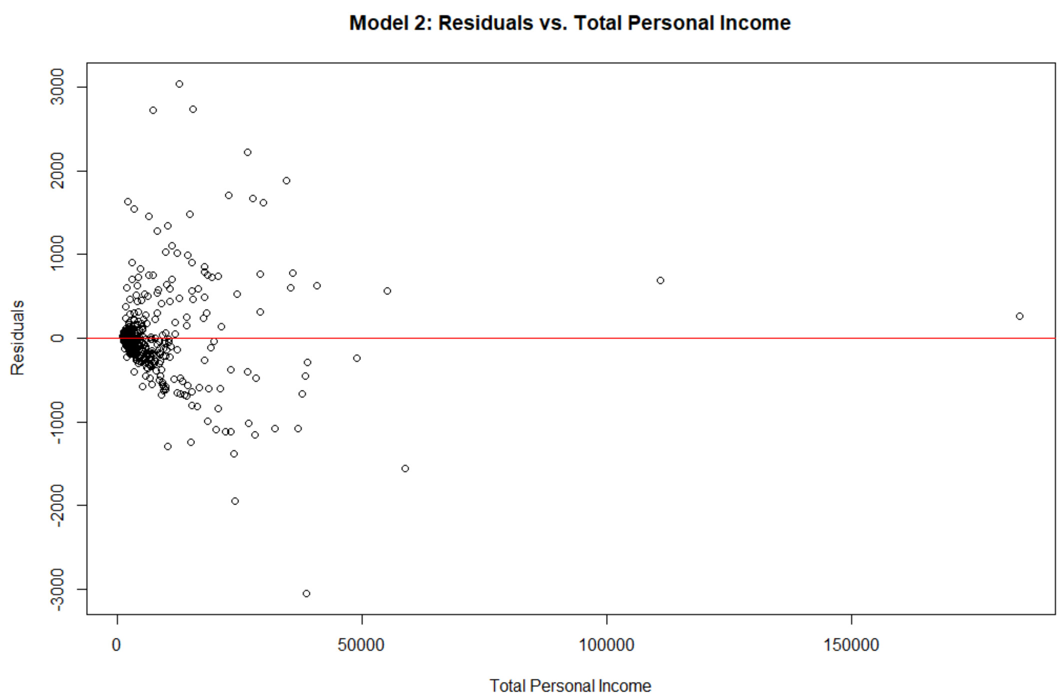


When examining the residuals over the population density, we notice a large cluster near where population density is equal to 0, then a smaller scatter as density increases.

Adam Hetherwick
Raees Qadir
Russell Su

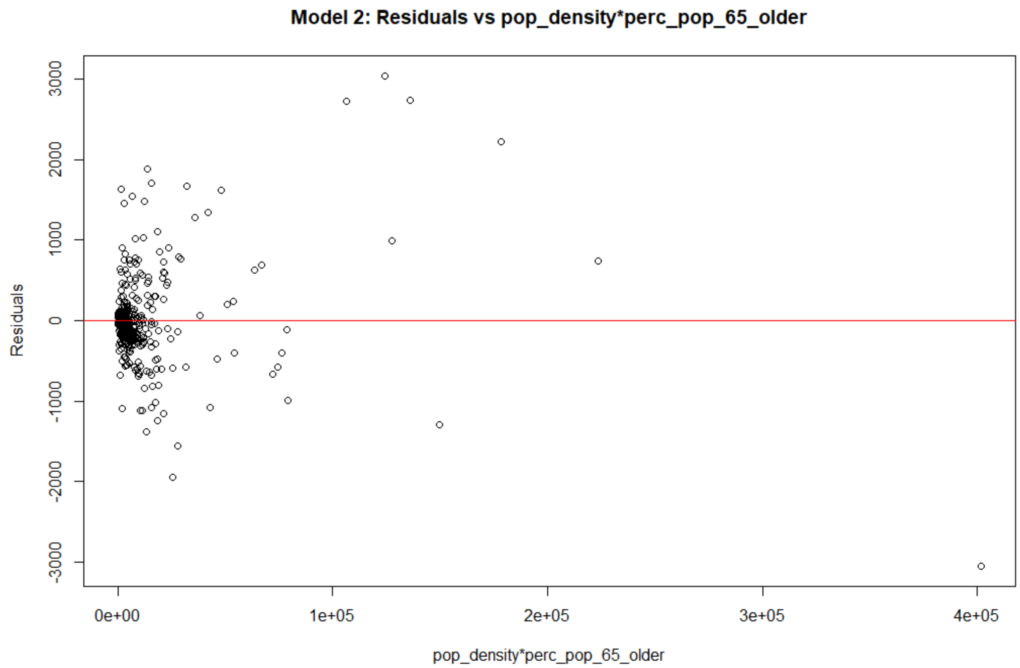


When examining the residuals over the percent of the population that is age 65 or older, we notice a clump near 10-15% then a cluster from 15% onward that has lower variance than compared to the clump.

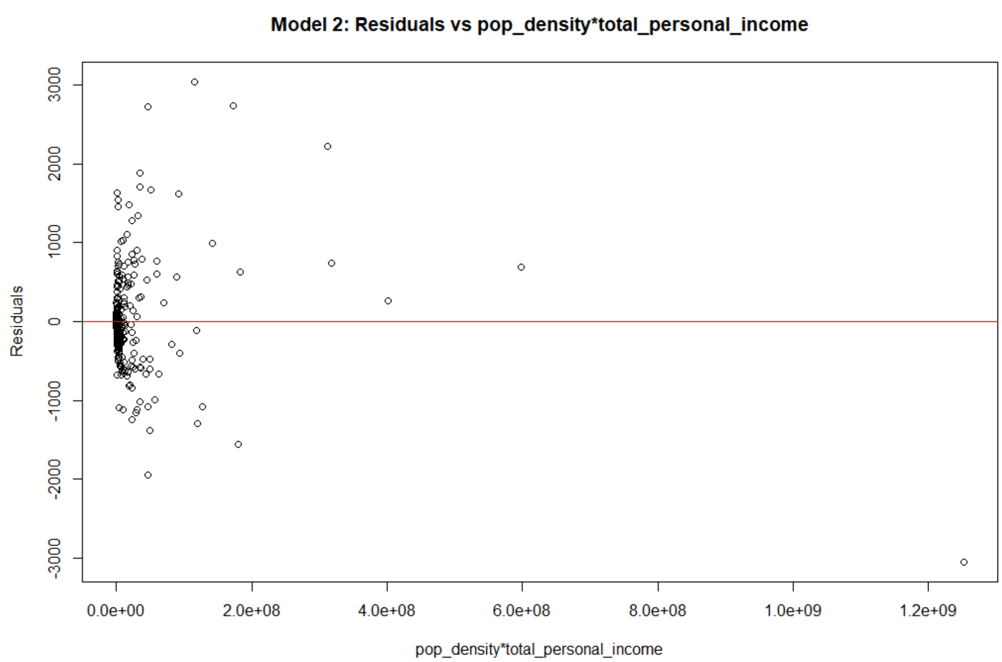


When looking at the residuals over the total personal income, we notice that there is a large clump around where total personal income approaches 0, then it tapers off as income increases.

Adam Hetherwick
Raees Qadir
Russell Su

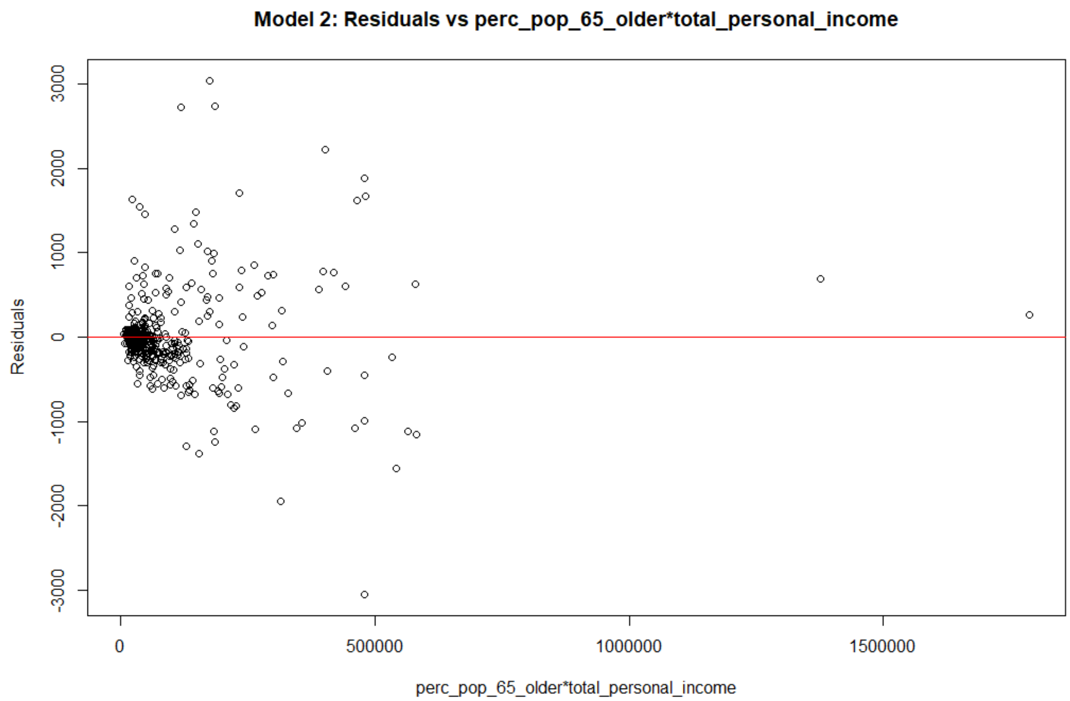


When examining the residuals for model 2 over the population density and percent population over 65 years old, we see a large clump around where population density and percent of the population that is ages 65 or older equals 0, then small cluster as that increases.

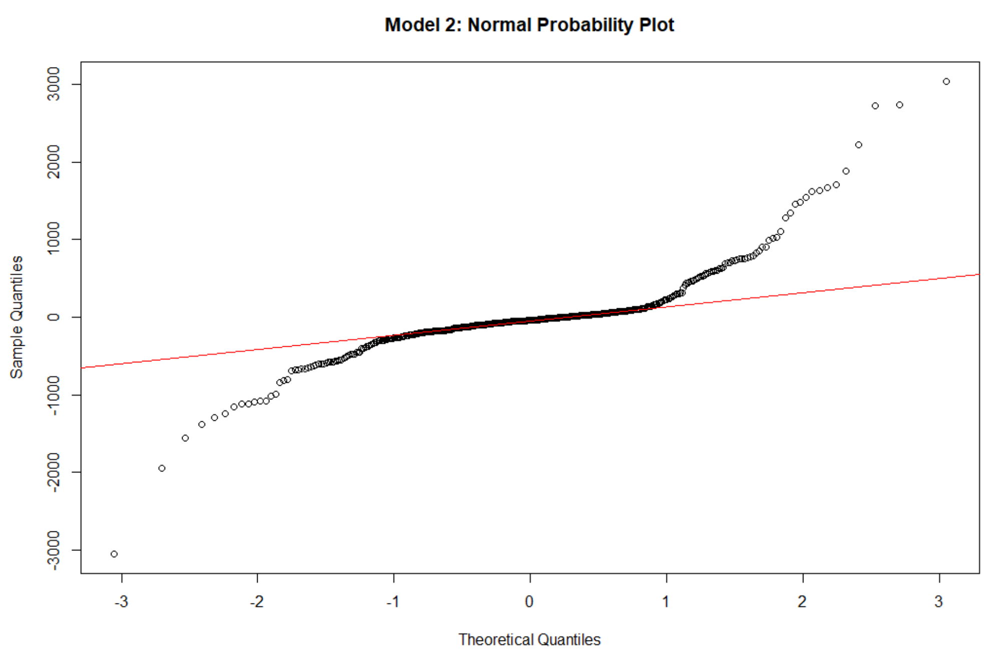


The model 2 residuals over population density and total personal income show a large cluster with high variance around where total personal income and population density equal zero, then a highly variable scatter as that increases.

Adam Hetherwick
Raees Qadir
Russell Su



When examining the residuals over the percent of the population over 65 years old and the total personal income, we see a dense cluster around where x is close to 0.



When examining the normal probability plot for model 2 residuals, we notice a strong linear relationship in between theoretical quantiles [-2, 2], then after that we see the residuals taper off thus suggesting more variability.

Adam Hetherwick
Raees Qadir
Russell Su

We infer that model 2 is slightly more predictive than model 1, as shown through the less variability in the residual plots. Model 2's normal probability plot seems to have a stronger correlation for longer than compared to model 1. Model 2 does not appear to have a stronger and definitive correlation than that of Model 1, though it does appear to be slightly more predictive. That also aligns with our summary statistics, as we found that Model 2 has a slightly higher R^2 value, although they are very similar.

f. Now expand both models proposed above by adding all possible two-factor interactions. Note that, for a model with X1, X2, X3 as the predictors, the two-factor interactions are X1X2, X1X3, X2X3. Repeat part d for the two expanded models.

<u>Model I</u>

```
Call:
lm(formula = physicians ~ total.pop + land.area + total.income +
    total.pop:land.area + total.pop:total.income + land.area:total.income,
    data = CDI)

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
total.pop               7.252e-04  3.259e-04   2.225  0.02657 *
land.area              -6.421e-02  3.014e-02  -2.131  0.03369 *
total.income            1.087e-01  1.450e-02   7.496 3.76e-13 ***
total.pop:land.area     6.173e-07  2.058e-07   2.999  0.00287 **
total.pop:total.income  1.696e-09  1.041e-09   1.630  0.10392
land.area:total.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
```

```
> summary(model1_a)$r.squared
[1] 0.9063789
```

R^2 = 0.9064

Adam Hetherwick
Raees Qadir
Russell Su

Model II

```
Call:
lm(formula = physicians ~ pop.density + pop.over65 + total.income +
    pop.density:pop.over65 + pop.density:total.income + pop.over65:total.income,
    data = CDI)

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
pop.density              -4.179e-01  1.055e-01  -3.960 8.76e-05 ***
pop.over65               -1.106e+01  7.792e+00  -1.419    0.157
total.income              1.477e-01  9.739e-03  15.168  < 2e-16 ***
pop.density:pop.over65    4.652e-02  7.925e-03   5.870 8.67e-09 ***
pop.density:total.income -3.276e-06  7.439e-07  -4.404 1.34e-05 ***
pop.over65:total.income  -1.289e-03  8.743e-04  -1.474    0.141
---
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

> summary(model2_a)$r.squared
[1] 0.9230238
```

R^2 = 0.9230

There is still only a 0.02 difference between the R^2 of each expanded model, so it cannot be said that there is a clear preference based on this alone.

Adam Hetherwick
Raees Qadir
Russell Su

Part II: Multiple linear regression II.

This part consists of Project 7.37 in the book.

a. For each of the following variables, calculate the coefficient of partial determination given that X1 (total pop) and X2 (personal income) are included in the model: land area (X3), percent of population 65 or older (X4), and number of hospital beds (X5)

Land Area

```
Analysis of Variance Table

Response: num_active_phys
                      Df     Sum Sq    Mean Sq  F value    Pr(>F)    
total_pop              1 1243181164 1243181164 3959.184 < 2.2e-16 ***
total_personal_income  1   22058054   22058054   70.249 7.271e-16 ***
land_area              1    4063370    4063370   12.941 0.0003583 ***
Residuals            436  136903711     313999                       
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

SSR(X3|X1, X2)= **4063370**
SSE(X1,X2)= 4063370 + 136903711= **140967081**
Coefficient of partial determination for land area= **0.02882496**

Population over 65

```
Analysis of Variance Table

Response: physicians
              Df     Sum Sq    Mean Sq   F value    Pr(>F)    
total.pop      1 1243181164 1243181164 3859.8919 < 2.2e-16 ***
total.income   1   22058054   22058054   68.4870 1.571e-15 ***
pop.over65     1     541647     541647    1.6817    0.1954    
Residuals    436  140425434     322077                        
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

SSR (X4|X1, X2)= **541647**
SSE(X1,X2)= 541647+140425434= **140967081**
Coefficient of partial determination for population over 65= **0.003842**

Adam Hetherwick
Raees Qadir
Russell Su

Hospital Beds

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

SSR (X5|X1, X2)= **78070132**
SSE(X1,X2)= 78070132+ 62896949 = **140967081**
Coefficient of partial determination for Hospital Beds= **0.553818**

b. On the basis of the results in part (a), which of the three additional predictor variables is best? Is the extra sum of squares associated with this variable larger than those for the other two variables?

The best predictor variable is Hospital Beds. Yes, the sum of squares for hospital beds is significantly larger than the sum of squares for the other two variables. This means that when adding hospital beds to the linear model, a large percent of the error can be explained by the number of hospital beds. This implies that hospital beds are a strong variable for analysis.

c. Using the F test statistic, test whether or not the variable determined to be best in part (b) is helpful in the regression model when X1 and X2 are included in the model; use a = .01. State the alternatives, decision rule, and conclusion. Would the F* test statistics for the other three potential predictor variables be as large as the one here? Discuss.*

H_0: β5 = 0
H_A: β5 ≠ 0

SSE(R) = SSE(X1, X2) = 78070132+ 62896949= 140967081
SSE(F) = SSE(X1, X2, X5) = 78070132
df(R) = n - 3 = 440 - 3 = 437
df(F) = n - 4 = 440 - 4 = 436

F* = ((SSE(R) - SSE(F)) / (df(R) - df(F))) / (SSE(F) / df(F))
= ((140967081 - 78070132) / (437 - 436)) / (78070132 / 436)
= 351.262

Adam Hetherwick
Raees Qadir
Russell Su

For $\alpha = 0.01$, we require $F(.99; 1,\ 436) = 6.63$. Since F* = 351.262 ≥ 6.63, we reject the null hypothesis that X5 can be removed from the regression model that already contains X1 and X2, and thus conclude that it is a significant variable. No, for the other prediction variables, the F* would not be as high because their slopes are less significant and thus have a higher likelihood of being dropped from the regression model.

d. Compute three additional coefficients of partial determination: R2Y,X3,X4|X1,X2, R2Y,X3,X5|X1,X2, and R2Y,X4,X5|X1,X2. Which pair of predictors is relatively more im- portant than other pairs? Use the F test to find out whether adding the best pair to the model is helpful given that X1, X2 are already included.

$R^2_{Y,\ X3,\ X4|X1,\ X2}$:

```
Analysis of Variance Table

Response: physicians
              Df     Sum Sq    Mean Sq   F value    Pr(>F)    
total.pop      1 1243181164 1243181164 3967.7399 < 2.2e-16 ***
total.income   1   22058054   22058054   70.4005 6.842e-16 ***
land.area      1    4063370    4063370   12.9687 0.0003533 ***
pop.over65     1     608535     608535    1.9422 0.1641413    
Residuals    435  136295177     313322                      
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
>
```

SSR(X3, X4|X1, X2) = 4063370+608535 = **4671905**
SSE(X1,X2) = 4671905 + 136295177 = **140967082**
Coefficient of partial determination = **0.0331**

Adam Hetherwick
Raees Qadir
Russell Su

$R^2_{Y, X3, X5|X1, X2}$:

```
Analysis of Variance Table

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

SSR(X3, X5|X1, X2) = 4063370+74289406 = **78352776**
SSE(X1,X2)= 78352776 + 62614306 = **140967082**
Coefficient of partial determination = **0.5558232**

$R^2_{Y, X4, X5|X1, X2}$:

```
Analysis of Variance Table

Response: physicians
              Df     Sum Sq    Mean Sq  F value Pr(>F)
total.pop      1 1243181164 1243181164 8804.285 <2e-16 ***
total.income   1   22058054   22058054  156.216 <2e-16 ***
pop.over65     1     541647     541647    3.836 0.0508 .
beds           1   79002640   79002640  559.502 <2e-16 ***
Residuals    435   61422794     141202
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

SSR(X4, X5|X1, X2)= 541647+79002640= **79544287**
SSE(X1,X2)= 79544287+61422794= **140967082**
Coefficient of partial determination= **0.56427562**

X4 (percent of population over 65) and X5 (hospital beds) are the most important pair.

H_0: $\beta 4 = \beta 5 = 0$
H_A: Not both β4 and β5 are equal to 0

SSE(X1, X2) = 79544287+61422794= 140967082
SSR(X4, X5 | X1, X2) = 541647+79002640= 79544287

F* = (SSR(X4, X5 | X1, X2)/2) / (SSE(X1, X2) / (n - 4)))
= (79544287 / 2) / (140967082 / (440 - 4))
= 123.0121

For $\alpha = 0.01$, we require $F(.99; 1, 436) = 6.63$. Since F* = 123.0121 ≥ 6.63, we reject the null hypothesis that X5 and X4 can be removed from the regression model that already contains X1 and X2, and thus conclude that they are significant variables.

Part III: Discussion.

In problem 1, we compared two models for predicting the number of active physicians in a CDI. Based on our plots as well as our tests for R^2, we could not conclude that either model is clearly preferable, as the plots and values for R^2 of each model are relatively similar. In problem 2, we looked at various predictors and how one may be best for the model over the other three. In relation to their coefficients of partial determination, we found the number of hospital beds to be the factor with the greatest coefficient of partial determination, and the factor to cause a greater coefficient among the three additional coefficients calculated. As such, we can conclude that the number of hospital beds is the best additional predictor variable to be added to the model of problem 2. For this project, the aspect of the class that was used the most was the ANOVA table, all in the second problem as it required so. Lastly, it is possible to further improve the regression models by considering all of the untested predictor variables for a possible stronger correlation with the number of hospital beds.

Adam Hetherwick
Raees Qadir
Russell Su

R Code Appendix

```r
CDI <- read.table("C:/Users/cheif/RProjects/STA108/CDI.txt", quote="\"", comment.char="")
View(CDI)

#1a
attach(CDI)
stem(total_pop)
colnames(CDI) <- c('id', 'county', 'state', 'land_area', 'total_pop', 'perc_pop_18-34',
                   'perc_pop_65_older', 'num_active_phys', 'num_hospital_num_hospital_beds',
                   'serious_crimes', 'perc_hs_grads', 'perc_b_degree', 'perc_below_pov',
                   'perc_unemployed', 'per_capita_income', 'total_personal_income',
                   'geographic_region')
stem(land_area)
stem(total_personal_income)
pop_density = total_pop / land_area
pop_density
stem(pop_density)
stem(perc_pop_65_older)
stem(total_personal_income)

#1b
model1_response <- data.frame(total_pop, land_area, total_personal_income)
pairs(model1_response, pch=19, cex.lab = 0.8)
cor(model1_response)

model2_response <- data.frame(pop_density, perc_pop_65_older, total_personal_income)
pairs(model2_response, pch=19, cex.lab = 0.8)
cor(model2_response)

#1c
model1 <- lm(num_active_phys ~ total_pop + land_area + total_personal_income)
model1
model2 <- lm(num_active_phys ~ pop_density + perc_pop_65_older + total_personal_income)
model2

#1d
summary(model1)
summary(model1)$r.squared
```

Adam Hetherwick
Raees Qadir
Russell Su

```
summary(model2)$r.squared

#1e
#model1
res1 <- residuals(model1)
yhat1 <- fitted(model1)
plot(res1 ~ yhat1, main = "Model 1: Residuals vs y_hat")
abline(h=0)
plot(res1 ~ total_pop, main = "Model 1: Residuals vs. Total Population",
     xlab = "Total Population",
     ylab = "Residuals")
abline(h=0, col = 'red')
plot(res1 ~ land_area, main = "Model 1: Residuals vs. Land Area",
     xlab = "Land Area",
     ylab = "Residuals")
abline(h=0, col = 'red')
plot(res1 ~ total_personal_income,
     main = "Model 1: Residuals vs. Total Personal Income",
     xlab = "Total Personal Income",
     ylab = "Residuals")
abline(h=0, col = 'red')
plot(exp(log(total_pop)+log(land_area)), res1,
     main = "Model 1: Residuals vs total_pop*land_area")
abline(h=0, col = "red")
plot(land_area*total_personal_income, res1,
     main = "Model 1: Residuals vs land_area*total_personal_income",
     ylab = "Residuals",
     xlab = "land_area*total_personal_income")
abline(h=0, col = "red")
plot(exp(log(total_pop)+log(total_personal_income)), res1,
     main = "Model 1: Residuals vs total_pop*total_personal_income")
abline(h=0, col = "red")
qqnorm(res1, main = "Model 1: Normal Probability Plot")
qqline(res1, col='red')

#model2
#model2 <- lm(num_active_phys ~ pop_density + perc_pop_65_older + total_personal_income)
```

Adam Hetherwick
Raees Qadir
Russell Su

```r
res2 <- residuals(model2)
yhat2 <- fitted(model2)
plot(res2 ~ yhat2, main = "Model 2: Residuals vs y_hat")
abline(h=0)
plot(res2 ~ pop_density, main = "Model 2: Residuals vs. Population Density",
     xlab = "Population Density",
     ylab = "Residuals")
abline(h=0, col = 'red')
plot(res2 ~ perc_pop_65_older, main = "Model 2: Residuals vs. perc_pop_65_older",
     xlab = "perc_pop_65_older",
     ylab = "Residuals")
abline(h=0, col = 'red')
plot(res2 ~ total_personal_income,
     main = "Model 2: Residuals vs. Total Personal Income",
     xlab = "Total Personal Income",
     ylab = "Residuals")
abline(h=0, col = 'red')
plot(pop_density*perc_pop_65_older, res2,
     main = "Model 2: Residuals vs pop_density*perc_pop_65_older",
     ylab = "Residuals",
     xlab = "pop_density*perc_pop_65_older")
abline(h=0, col = "red")
plot(pop_density*total_personal_income, res2,
     main = "Model 2: Residuals vs pop_density*total_personal_income",
     ylab = "Residuals",
     xlab = "pop_density*total_personal_income")
abline(h=0, col = "red")
plot(perc_pop_65_older*total_personal_income, res2,
     main = "Model 2: Residuals vs perc_pop_65_older*total_personal_income",
     ylab = "Residuals",
     xlab = "perc_pop_65_older*total_personal_income")
abline(h=0, col = "red")
qqnorm(res2, main = "Model 2: Normal Probability Plot")
qqline(res2, col='red')

#1f
model1_a <- lm(num_active_phys ~ pop_density + perc_pop_65_older + total_personal_income +
```

Adam Hetherwick
Raees Qadir
Russell Su

```r
                 pop:perc_pop_65_older + pop.density:total_personal_income +
                 perc_pop_65_older:total_personal_income, data = CDI)
summary(model1_a)
model2_a <- lm(num_active_phys ~ pop_density + perc_pop_65_older + total_personal_income +
                 pop.density:perc_pop_65_older + pop.density:total_personal_income +
                 perc_pop_65_older:total_personal_income, data = CDI)
summary(model2_a)
summary(model2_a)$r.squared

#2a
anova(model1)
anova(lm(num_active_phys ~ total_pop + total_personal_income + land_area))
4063370 + 136903711
4063370 / 140967081
anova(lm(num_active_phys ~ total_pop + total_personal_income + perc_pop_65_older))
anova(lm(num_active_phys ~ total_pop + total_personal_income + num_hospital_num_hospital_beds))
541647 / (541647+140425434)
78070132 / (78070132 + 62896949)


#2c
MSR =
F_stat=MSR/MSE
F_stat
qf(1-0.01, p-1,n-p) #critical value
1-pf(F_stat, p-1,n-p) #p-value
78070132 / 140967081
length(CDI$id)
# ((SSE(R) - SSE(F)) / (df(R) - df(F))) / (SSE(F) / df(F))
((140967081 - 78070132) / (437 - 436)) / (78070132 / 436)
78352776 + 62614306
78352776 / 140967082
4671905 + 136295177
4063370+74289406 + 62614306
541647+79002640 + 61422794
79544287 / 2
140967082 / 2
```

```r
(79544287 / 2) / (140967082 / (440 - 4))

#2d
anova(lm(num_active_phys ~ total_pop + total_personal_income + land_area +
           perc_pop_65_older, data = CDI))
anova(lm(num_active_phys ~ total_pop + total_personal_income + land_area + num_hospital_beds,
         data = CDI))
anova(lm(num_active_phys ~ total_pop + total_personal_income + perc_pop_65_older +
           num_hospital_beds, data = CDI))

```