Project 2

Lillian Quynni (913323065), Marco Oviedo (919345322), Colin Lauffer (918125242)

3/8/2022

Library and Load

Introduction

In this project, we were given a data set with 440 points and 17 different variables. All the data and graphs were calculated and made in R. For the first part of this project we created stem-and-leaf plots for each of the predictor variables in each model. The predictor variables for model 1 are active physicians (Y), total population (X1), land area (X2), and total personal income (X3). For model 2 the predictor variables are active physicians (Y), population density (X1), percent of population greater than 64 years old (X2), and total personal income (X3). We then created a scatter plot and a correlation matrix for each of the two models. Then we created a first order regression equation and calculated the R2 value for each model. Lastly for part one we obtained the residuals and plotted them against Y and each predictor variables. For the second part of the project we calculated the R2 values given X1 and X2. The given variables for the calculations are active physicians (Y), total population (X1) and total personal income (X2). The predictor variables are land area (X3), percent of population 65 or older (X4), and the number of hospital beds (X5). We then preformed an F-test on the best predictor variable to see if it was helpful to the regression model. Lastly we calculated the R2 values for the paired predictor variables and preformed another F-test.

6.28

Total Population

```
##
##   The decimal point is 6 digit(s) to the right of the |
##
##   0 | 1111111111111111111111111111111111111111111111111111111111111111111+254
##   0 | 5555555555555555555555556666666666666777777777777777777778888888888
##   1 | 000000122233333444
##   1 | 55699
##   2 | 1134
##   2 | 58
##   3 |
##   3 |
##   4 |
##   4 |
##   5 | 1
##   5 |
##   6 |
##   6 |
##   7 |
##   7 |
##   8 |
##   8 | 9
```

Land Area

```
##
##   The decimal point is 3 digit(s) to the right of the |
##
##    0 | 0000111111111111222222222222222222223333333333333333333333333333444444+252
##    1 | 00000000000000011111111111111222222222223333334444555666677778889999
##    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
```

Total Personal Income

```
##
##   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
```

Population Density

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

```
##
##     0 | 00000000000000001111111111111111111111111111111111111111111111111111+321
##     2 | 00001112233456700111145
##     4 | 05884
##     6 | 2464
##     8 | 19
##    10 | 378
##    12 |
##    14 | 4
##    16 |
##    18 |
##    20 |
##    22 |
##    24 |
##    26 |
##    28 |
##    30 |
##    32 | 4
```

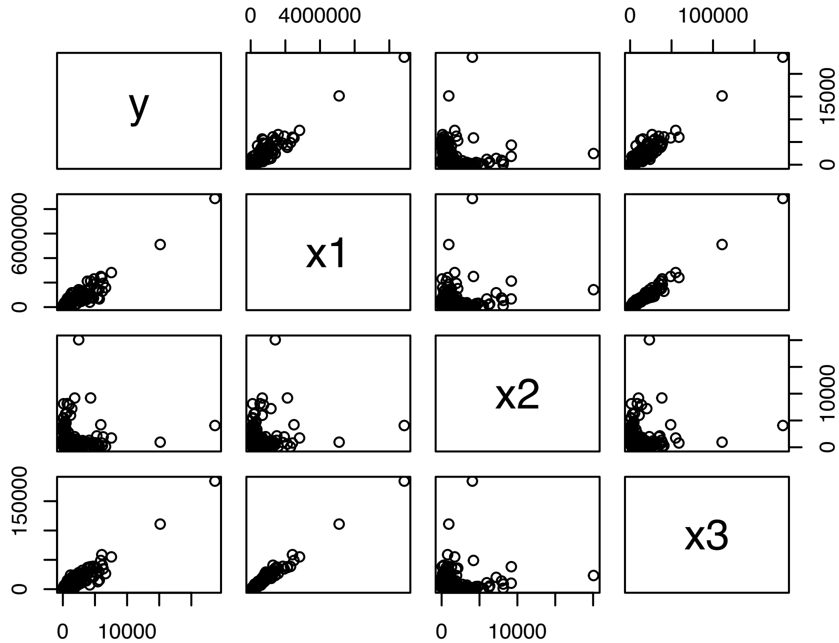
Percent of Population Older than 64

```
##
##    The decimal point is at the |
##
##     2 | 0
##     4 | 47890389
##     6 | 1123455677990134566678899
##     8 | 0011222223333444455566677777888889999000222233333344444445555666677
##    10 | 00011111122222222222333333444444555555566666666677777778888888888899999+36
##    12 | 00000000111112222333333333334444555555566666667777777778888999000000000+36
##    14 | 000011111111223334444445555677889000000111122223455667778
##    16 | 12556699901122345
##    18 | 06778
##    20 | 070
##    22 | 018828
##    24 | 47
##    26 | 055
##    28 | 1
##    30 | 7
##    32 | 138
```

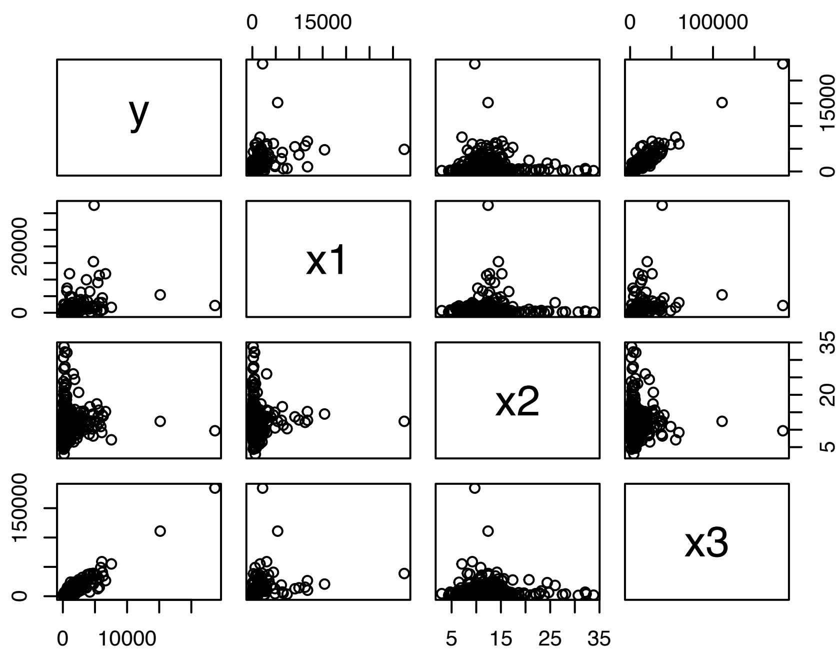
These plots provide us useful information about where the concentration of the data lies and we have a good visual comparison of how the data is distributed for our predictor variables.

b)

Model I: number of active physicians (y), total population (x1, land area (x2), and total personal income (x3)




```
##             y        x1         x2        x3
## y  1.00000000 0.9402486 0.07807466 0.9481106
## x1 0.94024859 1.0000000 0.17308335 0.9867476
## x2 0.07807466 0.1730834 1.00000000 0.1270743
## x3 0.94811057 0.9867476 0.12707426 1.0000000
```




```
##              y          x1          x2          x3
## y   1.00000000 0.40643863 -0.00312863  0.94811057
## x1  0.40643863 1.00000000  0.02918445  0.31620475
## x2 -0.00312863 0.02918445  1.00000000 -0.02273315
## x3  0.94811057 0.31620475 -0.02273315  1.00000000
```

Scatter plots provide us with information about how strong the relationship is between our predictor variable and the response variable, as well as outliers and gaps. By organizing them together in a matrix we are able to compare all 3 of our predictor variables and our response variable at the same time, and detect outliers, gaps, and the nature of the relationships simultaneously.

c)

```
##
## Call:
## lm(formula = y ~ x1 + x2 + x3, data = CDI2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1855.6  -215.2   -74.6    79.0  3689.0
##
## Coefficients:
##               Estimate  Std. Error t value         Pr(>|t|)
## (Intercept) -13.3161522  35.3670835  -0.377         0.706719
## x1            0.0008366   0.0002867   2.918         0.003701 **
## x2           -0.0655230   0.0182144  -3.597         0.000358 ***
## x3            0.0941320   0.0132996   7.078 0.00000000000589 ***
## ---
## 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: < 0.00000000000000022
```

Model I: $Y = -13.3 + 0.0008366(x_1) - 0.0655230(x_2) + 0.0941320(x_3)$

```
##
## Call:
## lm(formula = y ~ x1 + x2 + x3, data = CDI2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -3055.75  -175.30   -38.05    72.88  3045.81
##
## Coefficients:
##               Estimate  Std. Error t value             Pr(>|t|)
## (Intercept) -170.574223   83.532885  -2.042               0.0418 *
## x1             0.096159    0.012238   7.857    0.000000000000031 ***
## x2             6.339841    6.383772   0.993               0.3212
## x3             0.126566    0.002084  60.723 < 0.0000000000000002 ***
## ---
## 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: < 0.00000000000000022
```

Model II: $Y = -170.574223 + 0.096159(x_1) + 6.339841(x_2) + 0.126566(x_3)$

d)

```
## [1] 0.9026432
```

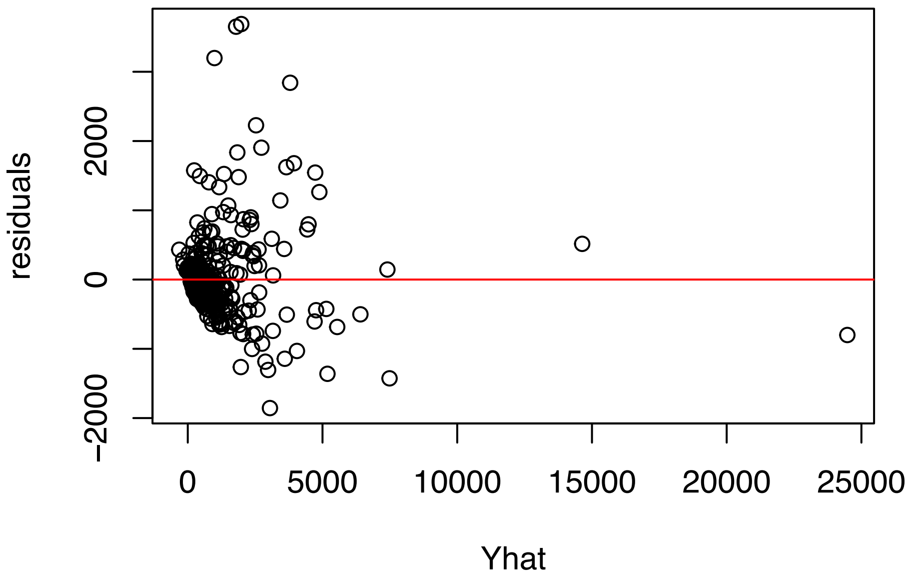
```
## [1] 0.9117491
```

Model I: $R^2 = 0.9026432$ Model II: $R^2 = 0.9117491$

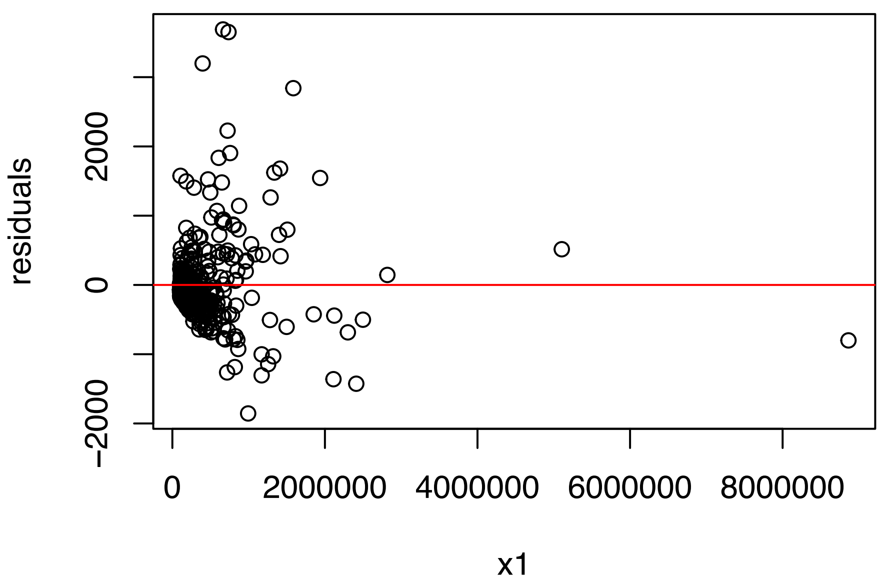
Model II is a better fit as it is closer to 1/is bigger.

e)

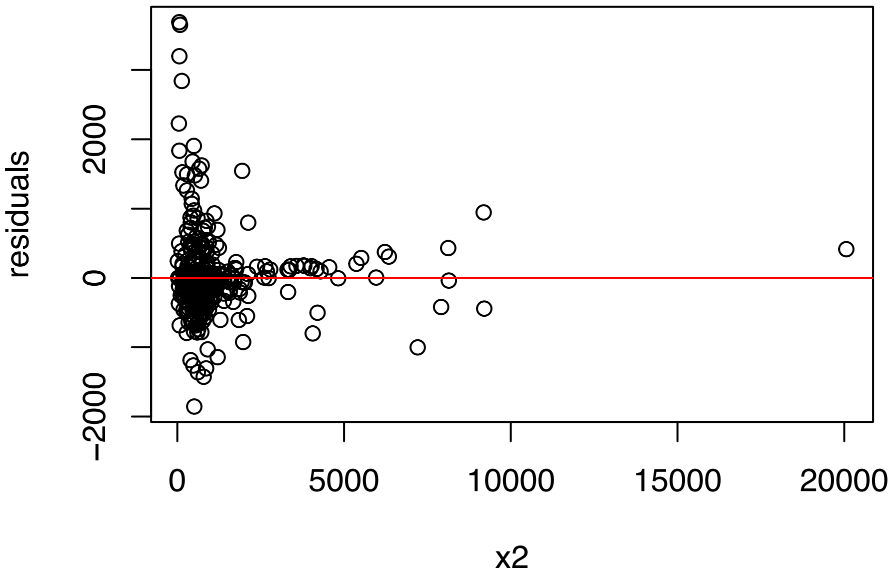
Model 1 Residuals Against Y__hat



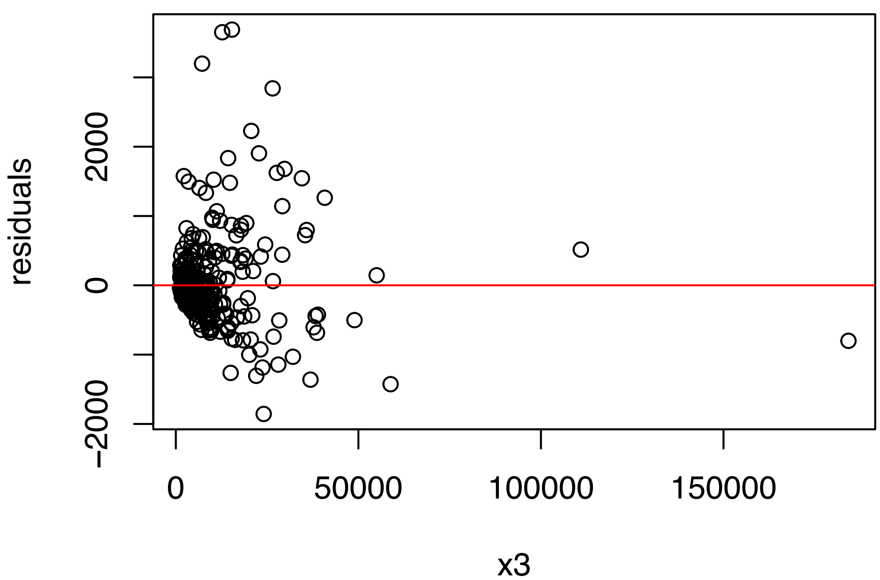
Model 1 Against Total population



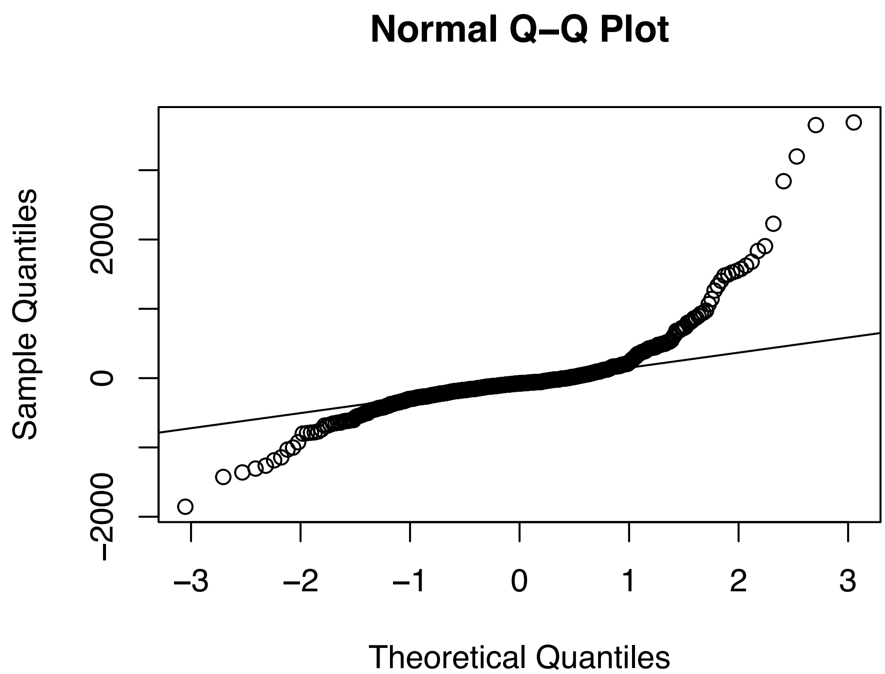
Model 1 Against Land Area



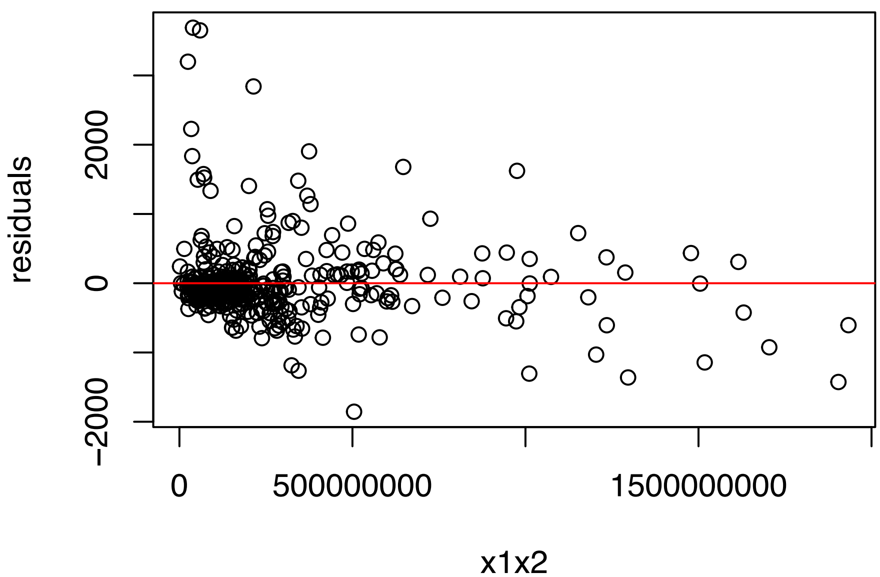
Model 1 Against Total Personal Income



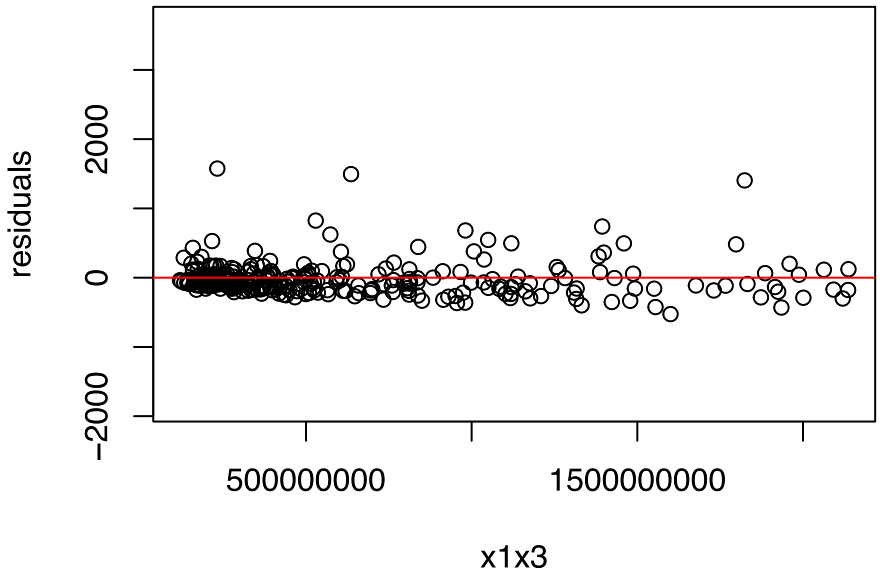
Normal Probability Plot for Model 1



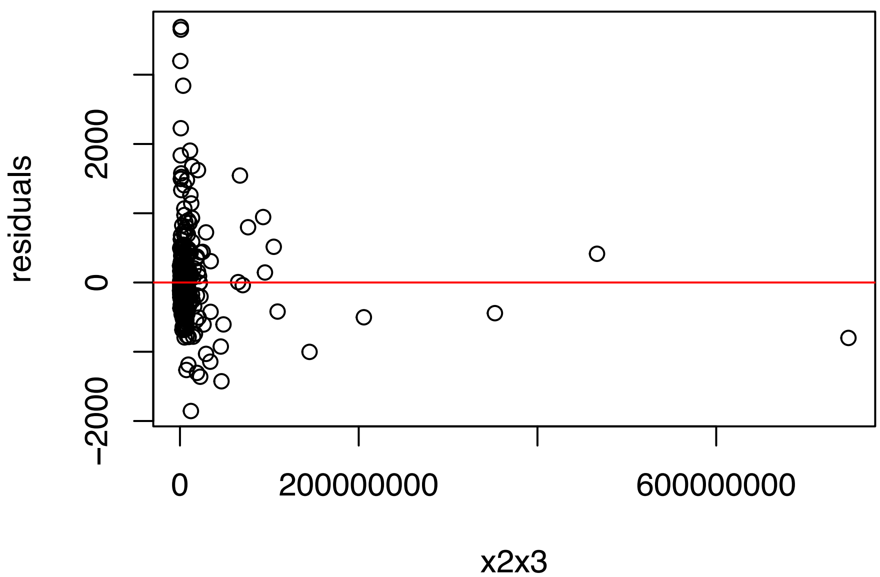
Residual Values Against X1X2



Residual Values Against X1X3



Residuals Against X2X3

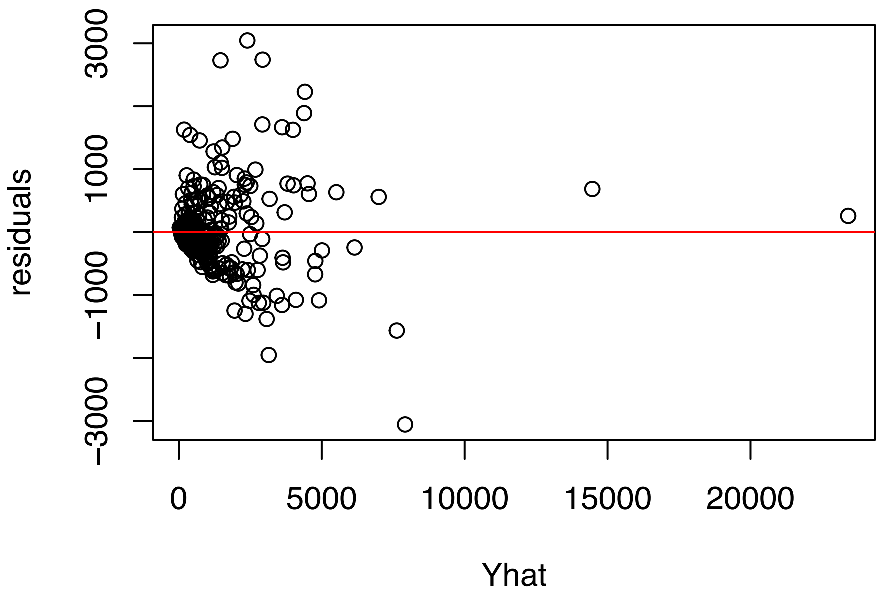


All of the plots look good for Model 1. They show no clear pattern, which is a good thing. Our residuals are split about evenly above the line 0 and they appear to be normally distributed. The Normal Probability Plot has heavy tails with both ends, but that could be due to large outliers we can see in the other residual plots.

Model 2 Against Y_hat

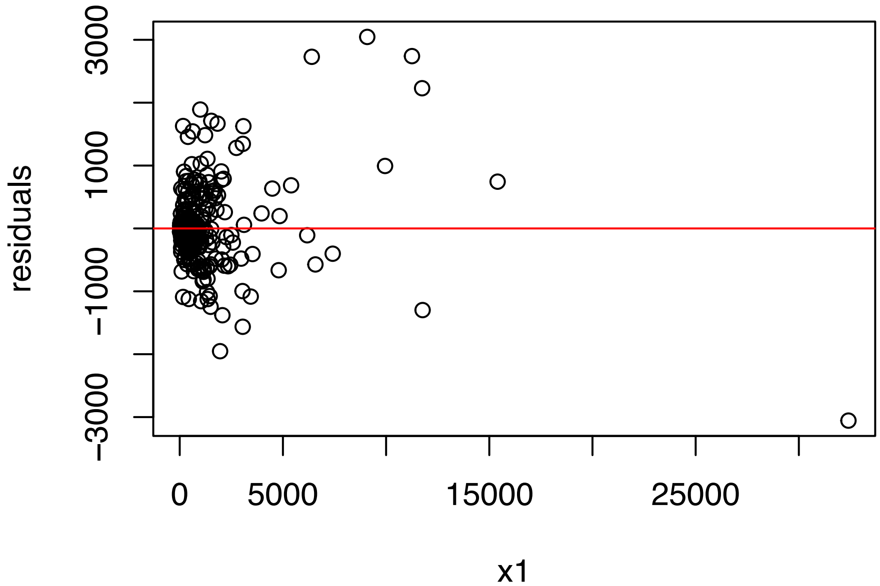
```
##
## Call:
## lm(formula = y ~ x1 + x2 + x3, data = CDI2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -3055.75  -175.30   -38.05    72.88  3045.81
##
```

```
## Coefficients:
##                Estimate  Std. Error t value             Pr(>|t|)
## (Intercept) -170.574223   83.532885  -2.042               0.0418 *
## x1             0.096159    0.012238   7.857    0.000000000000031 ***
## x2             6.339841    6.383772   0.993               0.3212
## x3             0.126566    0.002084  60.723 < 0.0000000000000002 ***
## ---
## 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: < 0.00000000000000022
```



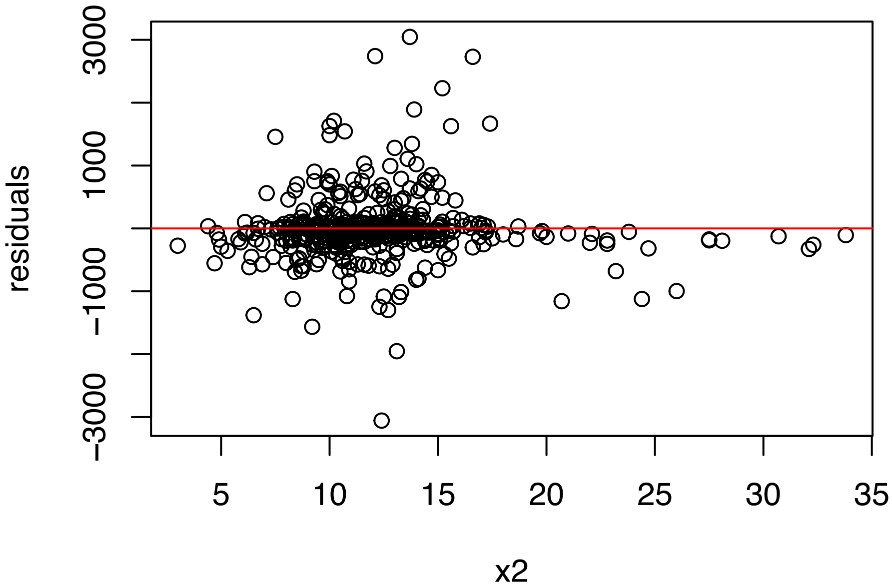
Model 2 Against Total population Divided By land

```
##
## Call:
## lm(formula = y ~ x1 + x2 + x3, data = CDI2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -3055.75  -175.30   -38.05    72.88  3045.81
##
## Coefficients:
##                Estimate  Std. Error t value             Pr(>|t|)
## (Intercept) -170.574223   83.532885  -2.042               0.0418 *
## x1             0.096159    0.012238   7.857    0.000000000000031 ***
## x2             6.339841    6.383772   0.993               0.3212
## x3             0.126566    0.002084  60.723 < 0.0000000000000002 ***
## ---
## 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: < 0.00000000000000022
```



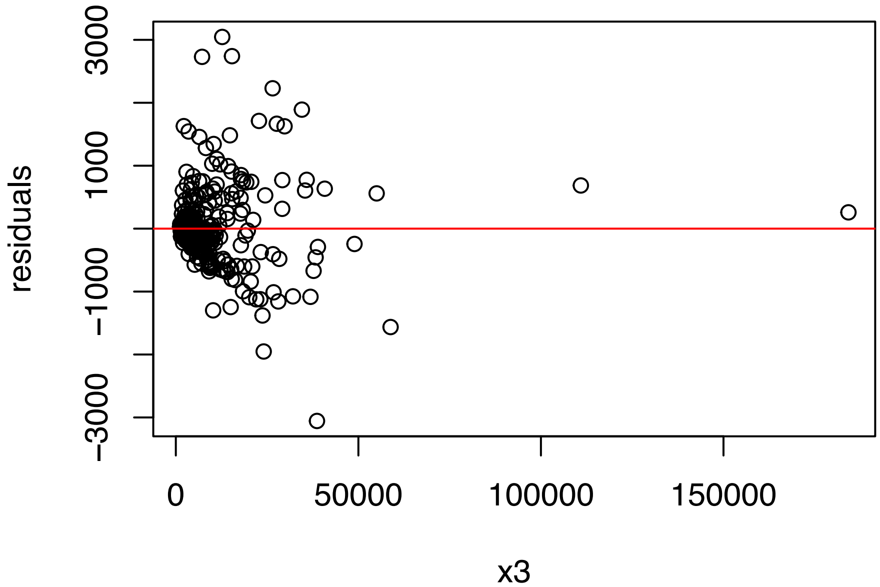
Model 2 Against Population Older than 64

```
##
## Call:
## lm(formula = y ~ x1 + x2 + x3, data = CDI2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -3055.75  -175.30   -38.05    72.88  3045.81
##
## Coefficients:
##                Estimate  Std. Error t value             Pr(>|t|)
## (Intercept) -170.574223   83.532885  -2.042               0.0418 *
## x1             0.096159    0.012238   7.857    0.000000000000031 ***
## x2             6.339841    6.383772   0.993               0.3212
## x3             0.126566    0.002084  60.723 < 0.0000000000000002 ***
## ---
## 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: < 0.00000000000000022
```



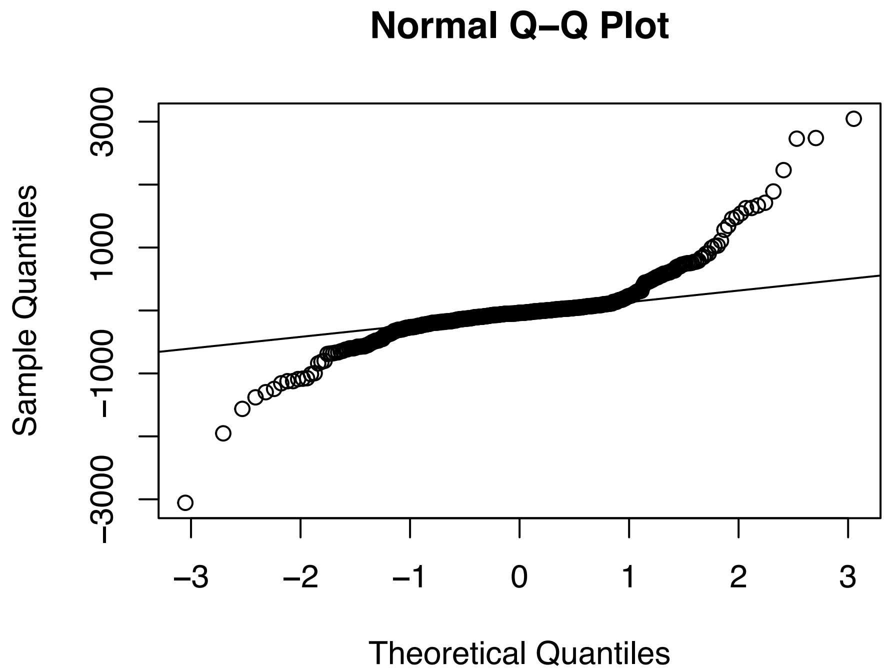
Model 2 Against Total Personal Income

```
##
## Call:
## lm(formula = y ~ x1 + x2 + x3, data = CDI2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -3055.75  -175.30   -38.05    72.88  3045.81
##
## Coefficients:
##                Estimate  Std. Error t value             Pr(>|t|)
## (Intercept) -170.574223   83.532885  -2.042               0.0418 *
## x1             0.096159    0.012238   7.857    0.000000000000031 ***
## x2             6.339841    6.383772   0.993               0.3212
## x3             0.126566    0.002084  60.723 < 0.0000000000000002 ***
## ---
## 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: < 0.00000000000000022
```



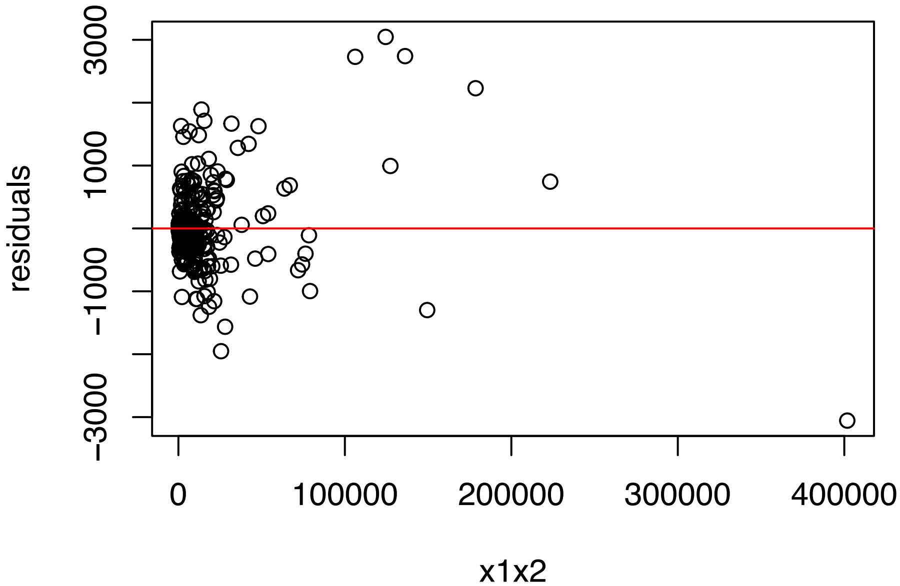
Normal Probability for Model 2

```
##
## Call:
## lm(formula = y ~ x1 + x2 + x3, data = CDI2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -3055.75  -175.30   -38.05    72.88  3045.81
##
## Coefficients:
##                Estimate  Std. Error t value             Pr(>|t|)
## (Intercept) -170.574223   83.532885  -2.042               0.0418 *
## x1             0.096159    0.012238   7.857    0.000000000000031 ***
## x2             6.339841    6.383772   0.993               0.3212
## x3             0.126566    0.002084  60.723 < 0.0000000000000002 ***
## ---
## 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: < 0.00000000000000022
```



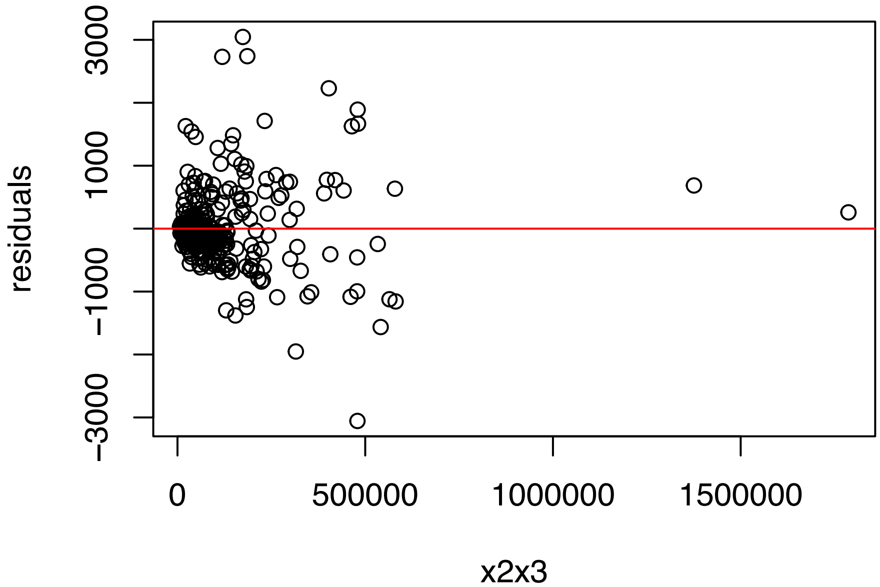
Model 2 Against X1X2

```
##
## Call:
## lm(formula = y ~ x1 + x2 + x3, data = CDI2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -3055.75  -175.30   -38.05    72.88  3045.81
##
## Coefficients:
##                Estimate  Std. Error t value             Pr(>|t|)
## (Intercept) -170.574223   83.532885  -2.042               0.0418 *
## x1             0.096159    0.012238   7.857    0.000000000000031 ***
## x2             6.339841    6.383772   0.993               0.3212
## x3             0.126566    0.002084  60.723 < 0.0000000000000002 ***
## ---
## 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: < 0.00000000000000022
```



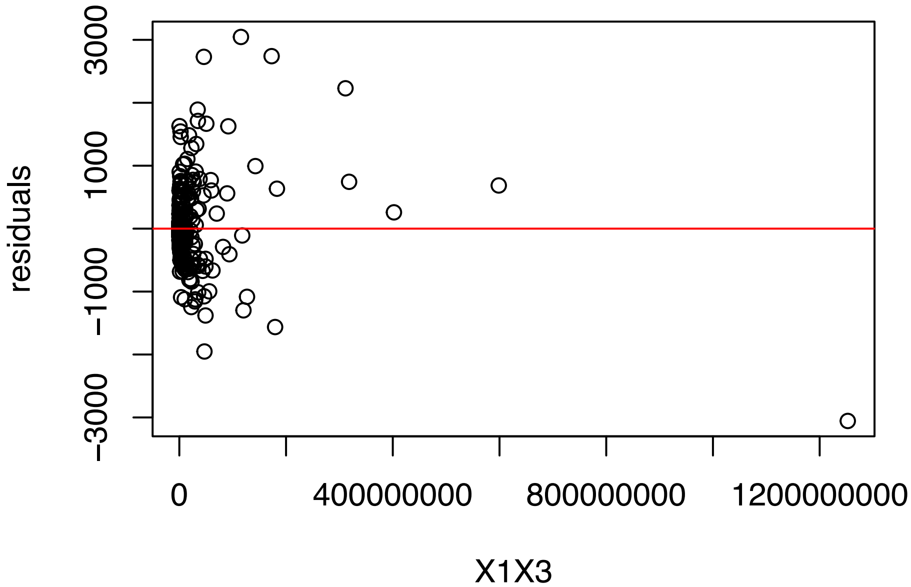
Residual Values Against X2X3

```
##
## Call:
## lm(formula = y ~ x1 + x2 + x3, data = CDI2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -3055.75  -175.30   -38.05    72.88  3045.81
##
## Coefficients:
##                Estimate  Std. Error t value             Pr(>|t|)
## (Intercept) -170.574223   83.532885  -2.042               0.0418 *
## x1             0.096159    0.012238   7.857    0.000000000000031 ***
## x2             6.339841    6.383772   0.993               0.3212
## x3             0.126566    0.002084  60.723 < 0.0000000000000002 ***
## ---
## 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: < 0.00000000000000022
```



Residual Values Against X1X3

```
##
## Call:
## lm(formula = y ~ x1 + x2 + x3, data = CDI2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -3055.75  -175.30   -38.05    72.88  3045.81
##
## Coefficients:
##                Estimate  Std. Error t value             Pr(>|t|)
## (Intercept) -170.574223   83.532885  -2.042               0.0418 *
## x1             0.096159    0.012238   7.857    0.000000000000031 ***
## x2             6.339841    6.383772   0.993               0.3212
## x3             0.126566    0.002084  60.723 < 0.0000000000000002 ***
## ---
## 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: < 0.00000000000000022
```



Model II appears to also be a good fit. The distribution of data form regression plots appears to be normally distributed and has no pattern, which is good. Most of the data is around the line 0, split evenly above and below too. The normal probability plot seems to be symmetrical with heavy tails, which can be why we see some outliers in our plots.

I don't think we can really say one model is better than the other here as all the plots are pretty similar between the two, which is backed up by the fact that they have such close correlation coefficients.

f)

```
##
## Call:
## lm(formula = y ~ x1 + x2 + x3 + x1 * x2 + x1 * x3 + x2 * x3,
##     data = CDI2)
##
## Coefficients:
##      (Intercept)                x1                x2                x3
## -58.257142564791    0.000725238876   -0.064213594715    0.108695122071
##            x1:x2             x1:x3             x2:x3
##   0.000000617305    0.000000001696   -0.000037062252
```

Model 1: $Y = -58.257142564791 + 0.000725238876(x_1) - 0.064213594715(x_2) + 0.108695122071(x_3) + 0.000000617305(x1x2) + 0.000000001696(x1x3) - 0.000037062252(x2x3)$

```
## [1] 0.9063789
```

Model 1: $R^2 = 0.9063789$

```
##
## Call:
## lm(formula = y ~ x1 + x2 + x3 + x1 * x2 + x1 * x3 + x2 * x3,
##     data = CDI2)
##
## Coefficients:
##   (Intercept)            x1            x2            x3         x1:x2
##   -9.367001586   -0.417949211  -11.058565175    0.147717007    0.046522448
##         x1:x3         x2:x3
##   -0.000003276   -0.001288565
```

$Y = -9.367001586 - 0.417949211(x_1) - 11.058565175(x_2) + 0.147717007(x_3) + 0.046522448(x1x2) - 0.000003276(x1x3) - 0.001288565(x2x3)$

```
## [1] 0.9230238
```

Model II: $R^2 = 0.9117491$

Model II appears to be a better fit again due to it having a higher R^2 that is closer to 1.

7.37

a)

```
## Analysis of Variance Table
##
## Response: Y
##            Df     Sum Sq    Mean Sq F value                Pr(>F)
## X1          1 1243181164 1243181164 3853.88 < 0.00000000000000022 ***
## X2          1   22058054   22058054   68.38  0.000000000000001638 ***
## Residuals 437  140967081     322579
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
## Analysis of Variance Table
##
## Response: Y
##            Df     Sum Sq    Mean Sq  F value                Pr(>F)
## X1          1 1243181164 1243181164 3959.184 < 0.00000000000000022 ***
## X2          1   22058054   22058054   70.249 0.0000000000000007271 ***
## X3          1    4063370    4063370   12.941             0.0003583 ***
## Residuals 436  136903711     313999
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
## Analysis of Variance Table
##
## Response: Y
##            Df     Sum Sq    Mean Sq   F value                Pr(>F)
## X1          1 1243181164 1243181164 3859.8919 < 0.00000000000000022 ***
## X2          1   22058054   22058054   68.4870  0.000000000000001571 ***
## X4          1     541647     541647    1.6817                0.1954
## Residuals 436  140425434     322077
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
## Analysis of Variance Table
##
## Response: Y
##            Df     Sum Sq    Mean Sq F value                Pr(>F)
## X1          1 1243181164 1243181164 8617.70 < 0.00000000000000022 ***
## X2          1   22058054   22058054  152.91 < 0.00000000000000022 ***
## X5          1   78070132   78070132  541.18 < 0.00000000000000022 ***
## Residuals 436   62896949     144259
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

$R^2_{Y,(X3|X1,X2)} = 0.028$

$R^2_{Y,(X4|X1,X2)} = 0.0038$

$R^2_{Y,(X5|X1,X2)} = 0.5538$

b)

Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
437	140967081				
436	136903711	1	4063370	12.94069	0.0003583

Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
437	140967081				
436	140425434	1	541647.3	1.681734	0.1953801

Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
437	140967081				
436	62896949	1	78070132	541.1801	0

The predictor variable X5 would be best because it has the largest R2 value. Also, it has the largest extra sum of squares associated to it.

c)

```
## Analysis of Variance Table
##
## Response: Y
##            Df     Sum Sq    Mean Sq F value                Pr(>F)
## X1          1 1243181164 1243181164 8617.70 < 0.00000000000000022 ***
## X2          1   22058054   22058054  152.91 < 0.00000000000000022 ***
## X5          1   78070132   78070132  541.18 < 0.00000000000000022 ***
## Residuals 436   62896949     144259
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
## [1] 6.693358
```

$H0 : B_k = B_5 = 0$ $H1 : B_k \neq b_5 \neq 0$

F-value = 541.18 F-Critical value = 6.693

Since the F-value is greater than the critical value we can conclude that X5 is helpful in the regression model.

d)

```
## Analysis of Variance Table
##
## Response: Y
##            Df     Sum Sq    Mean Sq   F value                Pr(>F)
## X1          1 1243181164 1243181164 3967.7399 < 0.00000000000000022 ***
## X2          1   22058054   22058054   70.4005 0.0000000000000006842 ***
## X3          1    4063370    4063370   12.9687             0.0003533 ***
## X4          1     608535     608535    1.9422             0.1641413
## Residuals 435  136295177     313322
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

## Analysis of Variance Table
##
## Response: Y
##            Df     Sum Sq    Mean Sq  F value                Pr(>F)
## X1          1 1243181164 1243181164 8636.745 < 0.00000000000000022 ***
## X2          1   22058054   22058054  153.244 < 0.00000000000000022 ***
## X3          1    4063370    4063370   28.229          0.0000001724 ***
## X5          1   74289406   74289406  516.110 < 0.00000000000000022 ***
## Residuals 435   62614306     143941
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

## Analysis of Variance Table
##
## Response: Y
##            Df     Sum Sq    Mean Sq  F value              Pr(>F)
## X1          1 1243181164 1243181164 8804.285 <0.0000000000000002 ***
## X2          1   22058054   22058054  156.216 <0.0000000000000002 ***
## X4          1     541647     541647    3.836              0.0508 .
## X5          1   79002640   79002640  559.502 <0.0000000000000002 ***
## Residuals 435   61422794     141202
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

## [1] 4.654269
```

$R^2_{Y,(X3,X4|X1,X2)} = 0.0331$

$R^2_{Y,(X3,X5|X1,X2)} = 0.5558$

$R^2_{Y,(X4,X5|X1,X2)} = 0.5642$

X_4 and X_5 is the best pair of predictors because it has the largest R^2 value compared to the other pairs

$H_0 : B_4 = B_5 = 0$ $H_1 : B_4 \neq B_5 \neq 0$ F-value = 281.66 F-Critical Value = 4.654

Since the F-value is greater than the critical value we can conclude that the pair X_4 , X_5 is helpful in the regression model.

Part 3 Disciussion

We found Model 1 and Model 2 to be very similar in correlation and we had to do a lot of work using ANOVA table for part 2 of our project.

Multiple linear regression was most relevant to our analyses of out data, not much information from before Midterm 1 was used, expect for interpreting our Residual/Normal Probability Plots. However, we did need to know how to interpret out results and also how to calculate coefficients of partial determination. Ways we could improve our regression model is by having more sample data to fit our model better. This could help us maybe explain the outliers and other important information we may not be able to see currently.

```
knitr::opts_chunk$set(
    error = FALSE,
    message = FALSE,
    warning = FALSE,
    echo = FALSE, # hide all R codes!!
    fig.width=5, fig.height=4,#set figure size
```

```r
    fig.align='center',#center plot
    options(knitr.kable.NA = ''), #do not print NA in knitr table
    tidy = FALSE #add line breaks in R codes
)
library(tidyverse)
CDI <- read.table("CDI.txt")

options(scipen=999)
CDI_new=CDI%>%mutate(new=V5/V4)
X=CDI_new$V5
stem(X)
X=CDI_new$V4
stem(X)
X=CDI_new$V16
stem(X)
X=CDI_new$new
stem(X)
X=CDI_new$V7
stem(X)
CDI_new=CDI%>%mutate(new=V5/V4)
CDI2=CDI_new[ , c(8, 5, 4, 16)]
colnames(CDI2) = c("y","x1","x2","x3")
pairs(CDI2)
cor(CDI2)
#Model 2:
CDI_new=CDI%>%mutate(new=V5/V4)
CDI2=CDI_new[ , c(8, 18, 7, 16)]
colnames(CDI2) = c("y","x1","x2","x3")
pairs(CDI2)
cor(CDI2)

CDI_new=CDI%>%mutate(new=V5/V4)
CDI2=CDI_new[ , c(8, 5, 4, 16)]
colnames(CDI2) = c("y","x1","x2","x3")
fit = lm(y~x1+x2+x3, data=CDI2)
summary(fit)
CDI_new=CDI%>%mutate(new=V5/V4)
CDI2=CDI_new[ , c(8, 18, 7, 16)]
colnames(CDI2) = c("y","x1","x2","x3")
fit = lm(y~x1+x2+x3, data=CDI2)
summary(fit)#for model II
CDI_new=CDI%>%mutate(new=V5/V4)
CDI2=CDI_new[ , c(8, 5, 4, 16)]
colnames(CDI2) = c("y","x1","x2","x3")
fit = lm(y~x1+x2+x3, data=CDI2)
summary(fit)$r.squared
#Model II:
CDI_new=CDI%>%mutate(new=V5/V4)
CDI2=CDI_new[ , c(8, 18, 7, 16)]
colnames(CDI2) = c("y","x1","x2","x3")
fit = lm(y~x1+x2+x3, data=CDI2)
summary(fit)$r.squared#for model II
CDI_new=CDI%>%mutate(new=V5/V4)
```

```
CDI2=CDI_new[ , c(8, 5, 4, 16)]
colnames(CDI2) = c("y","x1","x2","x3")
fit = lm(y~x1+x2+x3, data=CDI2)
residuals = fit$residuals
y_hat = fit$fitted.values
plot(x=y_hat, y=residuals, xlab="Yhat", ylab="residuals")
abline(h=0, col="red")
CDI_new=CDI%>%mutate(new=V5/V4)
CDI2=CDI_new[ , c(8, 5, 4, 16)]
colnames(CDI2) = c("y","x1","x2","x3")
fit = lm(y~x1+x2+x3, data=CDI2)
residuals = fit$residuals
y_hat = fit$fitted.values
plot(x=CDI2$x1, y=residuals, xlab="x1", ylab="residuals")
abline(h=0, col="red")
CDI_new=CDI%>%mutate(new=V5/V4)
CDI2=CDI_new[ , c(8, 5, 4, 16)]
colnames(CDI2) = c("y","x1","x2","x3")
fit = lm(y~x1+x2+x3, data=CDI2)
residuals = fit$residuals
y_hat = fit$fitted.values
plot(x=CDI2$x2, y=residuals, xlab="x2", ylab="residuals")
abline(h=0, col="red")
CDI_new=CDI%>%mutate(new=V5/V4)
CDI2=CDI_new[ , c(8, 5, 4, 16)]
colnames(CDI2) = c("y","x1","x2","x3")
fit = lm(y~x1+x2+x3, data=CDI2)
residuals = fit$residuals
y_hat = fit$fitted.values
plot(x=CDI2$x3, y=residuals, xlab="x3", ylab="residuals")
abline(h=0, col="red")
CDI_new=CDI%>%mutate(new=V5/V4)
CDI2=CDI_new[ , c(8, 5, 4, 16)]
colnames(CDI2) = c("y","x1","x2","x3")
fit = lm(y~x1+x2+x3, data=CDI2)
residuals = fit$residuals
qqnorm(residuals)
qqline(residuals)
CDI_new=CDI%>%mutate(new=V5/V4)
CDI2=CDI_new[ , c(8, 5, 4, 16)]
colnames(CDI2) = c("y","x1","x2","x3")
fit = lm(y~x1+x2+x3, data=CDI2)
residuals = fit$residuals
y_hat = fit$fitted.values
plot(x=CDI2$x1*CDI2$x2, y=residuals, xlab="x1x2", ylab="residuals")
abline(h=0, col="red")
CDI_new=CDI%>%mutate(new=V5/V4)
CDI2=CDI_new[ , c(8, 5, 4, 16)]
colnames(CDI2) = c("y","x1","x2","x3")
fit = lm(y~x1+x2+x3, data=CDI2)
residuals = fit$residuals
y_hat = fit$fitted.values
plot(x=CDI2$x1*CDI2$x3, y=residuals, xlab="x1x3", ylab="residuals")
```

```r
abline(h=0, col="red")
CDI_new=CDI%>%mutate(new=V5/V4)
CDI2=CDI_new[ , c(8, 5, 4, 16)]
colnames(CDI2) = c("y","x1","x2","x3")
fit = lm(y~x1+x2+x3, data=CDI2)
residuals = fit$residuals
y_hat = fit$fitted.values
plot(x=CDI2$x2*CDI2$x3, y=residuals, xlab="x2x3", ylab="residuals")
abline(h=0, col="red")
CDI_new=CDI%>%mutate(new=V5/V4)
CDI2=CDI_new[ , c(8, 18, 7, 16)]
colnames(CDI2) = c("y","x1","x2","x3")
fit = lm(y~x1+x2+x3, data=CDI2)
summary(fit)#for model II
residuals = fit$residuals
y_hat = fit$fitted.values
plot(x=y_hat, y=residuals, xlab="Yhat", ylab="residuals")
abline(h=0, col="red")
CDI_new=CDI%>%mutate(new=V5/V4)
CDI2=CDI_new[ , c(8, 18, 7, 16)]
colnames(CDI2) = c("y","x1","x2","x3")
fit = lm(y~x1+x2+x3, data=CDI2)
summary(fit)#for model II
residuals = fit$residuals
y_hat = fit$fitted.values
plot(x=CDI2$x1, y=residuals, xlab="x1", ylab="residuals")
abline(h=0, col="red")
CDI_new=CDI%>%mutate(new=V5/V4)
CDI2=CDI_new[ , c(8, 18, 7, 16)]
colnames(CDI2) = c("y","x1","x2","x3")
fit = lm(y~x1+x2+x3, data=CDI2)
summary(fit)#for model II
residuals = fit$residuals
y_hat = fit$fitted.values
plot(x=CDI2$x2, y=residuals, xlab="x2", ylab="residuals")
abline(h=0, col="red")
CDI_new=CDI%>%mutate(new=V5/V4)
CDI2=CDI_new[ , c(8, 18, 7, 16)]
colnames(CDI2) = c("y","x1","x2","x3")
fit = lm(y~x1+x2+x3, data=CDI2)
summary(fit)#for model II
residuals = fit$residuals
y_hat = fit$fitted.values
plot(x=CDI2$x3, y=residuals, xlab="x3", ylab="residuals")
abline(h=0, col="red")
CDI_new=CDI%>%mutate(new=V5/V4)
CDI2=CDI_new[ , c(8, 18, 7, 16)]
colnames(CDI2) = c("y","x1","x2","x3")
fit = lm(y~x1+x2+x3, data=CDI2)
summary(fit)#for model II
residuals = fit$residuals
qqnorm(residuals)
qqline(residuals)
```

```r
CDI_new=CDI%>%mutate(new=V5/V4)
CDI2=CDI_new[ , c(8, 18, 7, 16)]
colnames(CDI2) = c("y","x1","x2","x3")
fit = lm(y~x1+x2+x3, data=CDI2)
summary(fit)#for model II
residuals = fit$residuals
y_hat = fit$fitted.values
plot(x=CDI2$x2*CDI2$x1, y=residuals, xlab="x1x2", ylab="residuals")
abline(h=0, col="red")
CDI_new=CDI%>%mutate(new=V5/V4)
CDI2=CDI_new[ , c(8, 18, 7, 16)]
colnames(CDI2) = c("y","x1","x2","x3")
fit = lm(y~x1+x2+x3, data=CDI2)
summary(fit)#for model II
residuals = fit$residuals
y_hat = fit$fitted.values
plot(x=CDI2$x2*CDI2$x3, y=residuals, xlab="x2x3", ylab="residuals")
abline(h=0, col="red")
CDI_new=CDI%>%mutate(new=V5/V4)
CDI2=CDI_new[ , c(8, 18, 7, 16)]
colnames(CDI2) = c("y","x1","x2","x3")
fit = lm(y~x1+x2+x3, data=CDI2)
summary(fit)#for model II
residuals = fit$residuals
y_hat = fit$fitted.values
plot(x=CDI2$x1*CDI2$x3, y=residuals, xlab="X1X3", ylab="residuals")
abline(h=0, col="red")
CDI_new=CDI%>%mutate(new=V5/V4)
CDI2=CDI_new[ , c(8, 5, 4, 16)]
colnames(CDI2) = c("y","x1","x2","x3")
fit = lm(y~x1+x2+x3+x1*x2+x1*x3+x2*x3, data=CDI2)
fit
summary(fit)$r.squared
CDI_new=CDI%>%mutate(new=V5/V4)
CDI2=CDI_new[ , c(8, 18, 7, 16)]
colnames(CDI2) = c("y","x1","x2","x3")
fit = lm(y~x1+x2+x3+x1*x2+x1*x3+x2*x3, data=CDI2)
fit#for model II
summary(fit)$r.squared
Y = CDI[,8]
X1 = CDI[,5]
X2 = CDI[,16]
X3 = CDI[,4]
X4 = CDI[,7]
X5 = CDI[,9]
fit = lm(Y~X1 + X2)
anova(fit)
fit = lm(Y~X1 + X2 + X3)
anova(fit)
fit = lm(Y~X1 + X2 + X4)
anova(fit)
fit = lm(Y~X1 + X2 + X5)
anova(fit)
```

```r
reduced=lm(Y~X1+X2)
full= lm(Y~X1+X2+X3)
library(knitr)
kable(anova(reduced,full))

full= lm(Y~X1+X2+X4)
kable(anova(reduced,full))

full= lm(Y~X1+X2+X5)
kable(anova(reduced,full))
fit = lm(Y~X1 + X2 + X5)
anova(fit)
alpha = 0.01
qf(1-alpha, 1, 436)
fit = lm(Y~X1 + X2 + X3 + X4)
anova(fit)

fit = lm(Y~X1 + X2 + X3 + X5)
anova(fit)

fit = lm(Y~X1 + X2 + X4 + X5)
anova(fit)

qf(1-alpha, 2, 435)
```

7.37 A

$$R^2_{(X_3|X_1,X_2)} = \frac{SSR(X_3|X_1,X_2)}{SSE(X_1,X_2)} = \frac{4063370}{140967081} = 0.0288$$

$$R^2_{(X_4|X_1,X_2)} = \frac{SSR(X_4|X_1,X_2)}{SSE(X_1,X_2)} = \frac{541647}{140967081} = 0.0038$$

$$R^2_{(X_5|X_1,X_2)} = \frac{SSR(X_5|X_1,X_2)}{SSE(X_1,X_2)} = \frac{78070132}{140967081} = 0.5538$$

D. $$R^2_{(X_3,X_4|X_1,X_2)} = \frac{SSR(X_3|X_1,X_2)+SSR(X_4|X_1,X_2,X_3)}{SSE(X_1,X_2)} = \frac{4063370+608535}{140967081}$$

$$= 0.033$$

$$R^2_{(X_3,X_5|X_1,X_2)} = \frac{SSR(X_3|X_1,X_2)+SSR(X_5|X_1,X_2,X_3)}{SSE(X_1,X_2)} = 0.5558$$

$$R^2_{(X_4,X_5|X_1,X_2)} = \frac{SSR(X_4|X_1,X_2)+SSR(X_5|X_1,X_2,X_4)}{SSE(X_1,X_2)} = 0.5642$$

$$F = \frac{SSR(X_4|X_1,X_2)+SSR(X_5|X_1,X_2,X_4)}{2} \div MSE(X_1,X_2,X_4,X_5)$$

$$F = 281.66$$