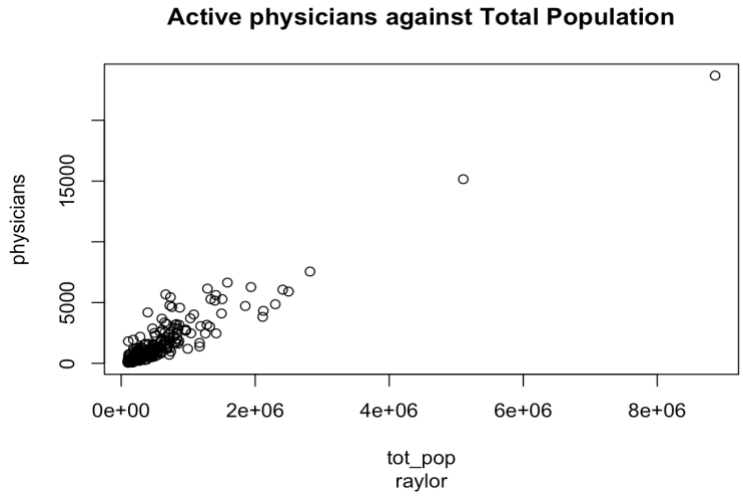
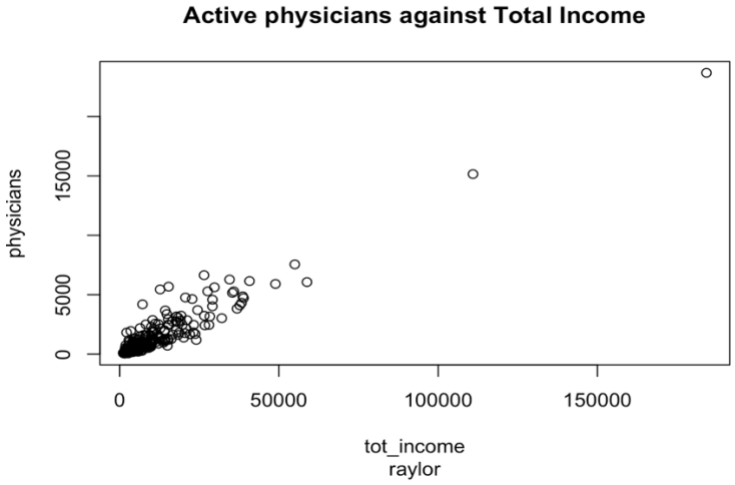
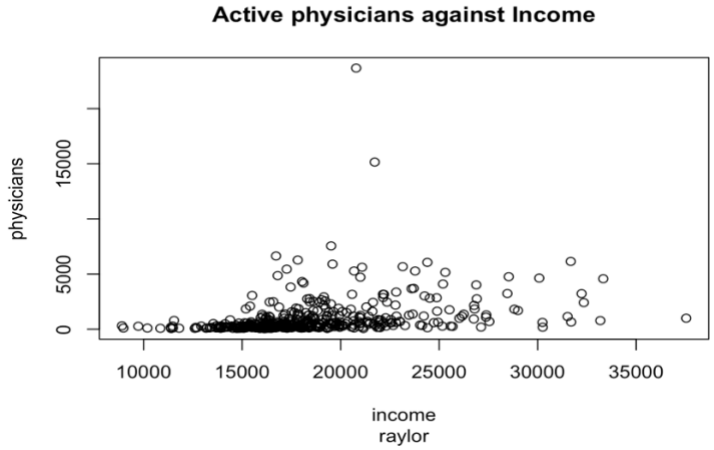
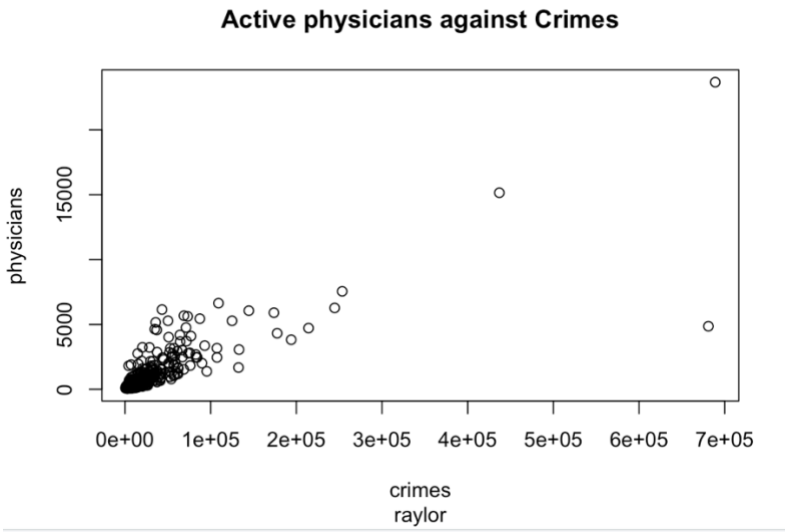
The goal is to model the number of physicians per 1000 inhabitants, using the other demographic variables.

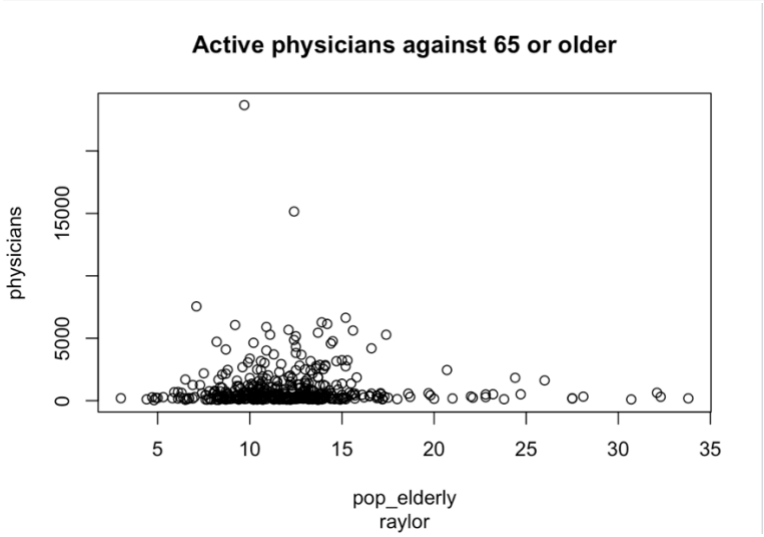
Plot Number of active physicians against each of Total Population, Total personal income, per capita income, Total serious crimes and pop65plus.











Also plot **Number of active physicians** against the others (Total Population, Total personal income, per capita income , Total serious crimes and pop65plus).

Same with last one

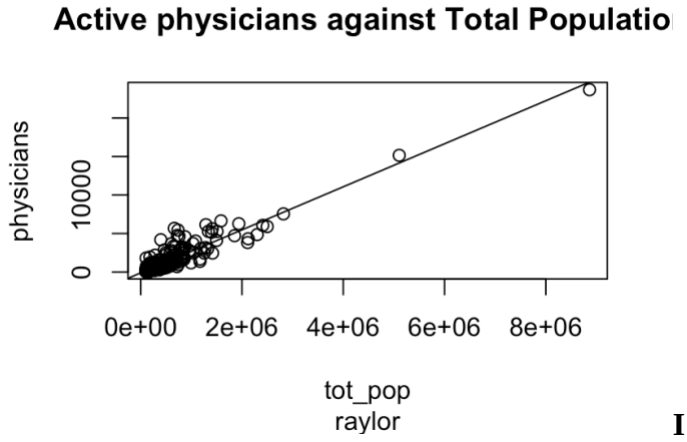
Part I

a. Regress the number of active physicians in turn on (SLR)each of the three predictor variables

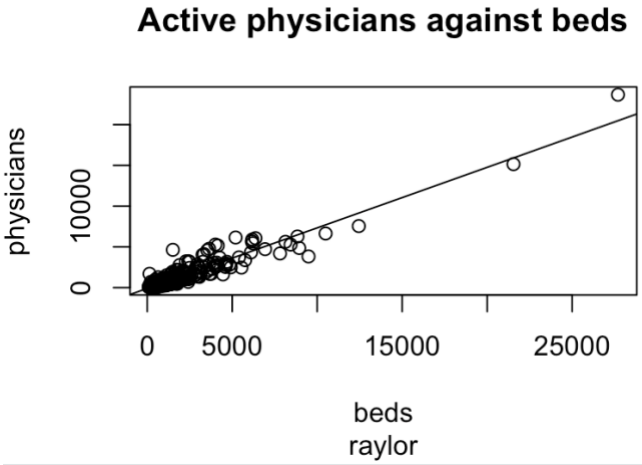
Y=-1.106e-2 + 2.795e-03 total pop Y=-95.93218+0.74312 beds

Y=-48.39485+0.1317total income

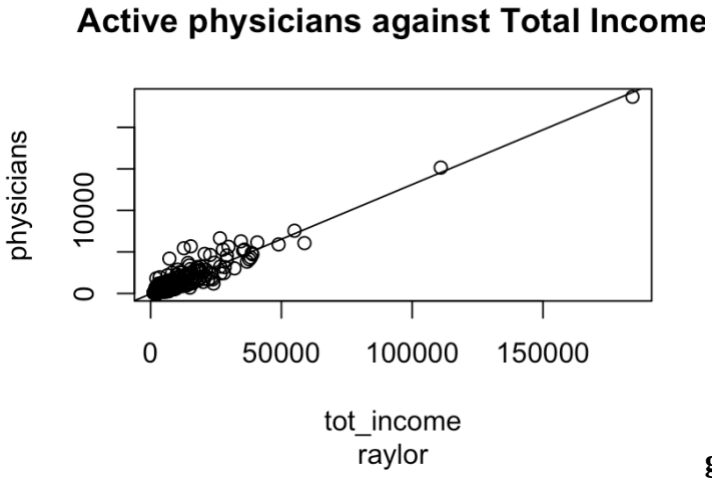
b. Plot the three estimated regression functions and data on separate graphs. Does a linear regression relation appear to provide a good fit for each of the three predictor variables?



I think it is a almost fit.



Good fit and one out linear



good fit.

They are all most fit and their square is quite high.

c. Calculate s($\sqrt{MSE}$) for each of the three predictor variables. Which predictor variable leads to the smallest variability around the fitted regression line?

Total population $\sqrt{MSE}$=610.1

Beds $\sqrt{MSE}$=556.9

Total income $\sqrt{MSE}$=569.7

d. Obtain Bonferroni joint confidence intervals for β0 and β1 using a 95 percent family confidence coefficient and interpret the interval for all the models.

```
> confint(a,level=(1-0.05/2))
                    1.25 %       98.75 %
(Intercept) -1.887833e+02 -32.486285498
tot_pop      2.686636e-03   0.002904214
> confint(b,level=(1-0.05/2))
                   1.25 %     98.75 %
(Intercept) -166.7663435 -25.0980260
beds           0.7169992   0.7692337
> confint(c,level=(1-0.05/2))
                   1.25 %    98.75 %
(Intercept) -119.9922981 23.2026003
tot_income     0.1269549  0.1364475
```

e. An investigator has suggested that for **model with total population** β0 should be -100 and β1 should be .0028. Do the joint confidence intervals in part (d) support this view?
Yes, because this β0 and β1 are fall into the confidence interval in part d.

f. Estimate the expected number of active physicians for counties with total population of X = 500, 1000, 5000 thousand with Bonferroni family confidence coefficient 0.90.

```
>   predict (a, newdata = new.data, interval= 'con
fidence', level=1- 0.1/ 2)
        fit       lwr       upr
1  1287.078  1229.017  1345.138
2  2684.790  2603.566  2766.014
3 13866.490 13424.812 14308.168
>
```

Part II

a. For geographic region**, regress per capital income in a CDI (Y)** against the percentage of individuals in a county having at least a bachelor's degree (X) Assume that first-order regression model is appropriate for each region. State the estimated regression functions.

```
Call:
lm(formula = income ~ bsgrad + as.factor(region))

Residuals:
     Min       1Q   Median       3Q      Max
-10253.0  -1438.4     72.2   1315.7  11174.5

Coefficients:
                   Estimate Std. Error t value
(Intercept)        12616.68     457.40  27.583
bsgrad               366.41      17.03  21.515
as.factor(region)2 -1561.38     375.49  -4.158
as.factor(region)3 -2840.25     346.73  -8.191
as.factor(region)4 -2372.29     409.05  -5.800
                   Pr(>|t|)
(Intercept)         < 2e-16 ***
bsgrad              < 2e-16 ***
as.factor(region)2 3.86e-05 ***
as.factor(region)3 2.90e-15 ***
as.factor(region)4 1.28e-08 ***
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2715 on 435 degrees of freedom
Multiple R-squared:  0.5567,    Adjusted R-squared:  0.5526
F-statistic: 136.6 on 4 and 435 DF,  p-value: < 2.2e-16
```

y= 12616.68+ 366.41* bsgrad- 11561.38* region2- 2840.25* region3–2372.29* region4

b. **Ho: β21 = β22 = β23 = β24 (Ha: At least one of β2i does not equal to others, i=1, 2, 3, 4.**

```
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = income ~ factor(region))

$`factor(region)`
          diff       lwr       upr     p adj
2-1 -2297.67440 -3681.5586  -913.7902 0.0001337
3-1 -3111.78015 -4394.1659 -1829.3944 0.0000000
4-1 -2276.18257 -3789.9574  -762.4077 0.0007004
3-2  -814.10575 -2078.6732   450.4617 0.3461252
4-2    21.49182 -1477.2183  1520.2019 0.9999818
4-3   835.59757  -569.9307  2241.1259 0.4184819
```

because of 2-1，3-1，4-1’s p-value<0.05, so they are not similar and 3-2,4-2,4-2 >0.05, so they are equal.

Part III

- The number of active physicians(Y) is to be regressed against total population (X1), total personal income (X2), and geographic region (X3, X4, X5). Fit a first-order regression model. Let X3 =1 if NE and 0 otherwise, X4 = 1 if Midwest and 0 otherwise, and X5 = 1 if S and 0 otherwise.

```
Call:
lm(formula = physicians ~ tot_pop + tot_income + as.factor(r
egion))

Residuals:
    Min      1Q  Median      3Q     Max
-1866.8  -207.7   -81.5    72.4  3721.7

Coefficients:
                     Estimate Std. Error t value
(Intercept)        -5.848e+01  5.882e+01  -0.994
tot_pop             5.515e-04  2.835e-04   1.945
tot_income          1.070e-01  1.325e-02   8.073
as.factor(region)2 -3.493e+00  7.881e+01  -0.044
as.factor(region)3  4.220e+01  7.402e+01   0.570
as.factor(region)4 -1.490e+02  8.683e+01  -1.716
                   Pr(>|t|)
(Intercept)          0.3207
tot_pop              0.0524 .
tot_income          6.8e-15 ***
as.factor(region)2   0.9647
as.factor(region)3   0.5689
as.factor(region)4   0.0868 .
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 566.1 on 434 degrees of freedom
Multiple R-squared:  0.9011,    Adjusted R-squared:  0.8999
F-statistic: 790.7 on 5 and 434 DF,  p-value: < 2.2e-16
```

Y= -58.48+ 0.0005515* totpop+ 0.0005515* tot_pop +0.107* totincome -3.493* Midwest+42.2* South- 149* West

We find northeast is default value,

- Examine whether the effect for the northeastern region on number of active physicians differs from the effect for the midwest region by constructing an appropriate 90 percent confidence interval. Interpret your interval estimate.

```
> TukeyHSD (aov.e, conf.level = 0.9)
  Tukey multiple comparisons of means
    90% family-wise confidence level

Fit: aov(formula = physicians ~ factor(region))

$`factor(region)`
          diff        lwr       upr     p adj
2-1 -234.04935 -798.94247  330.8438 0.7765657
3-1 -273.33444 -796.79653  250.1276 0.6271356
4-1  259.45606 -358.45764  877.3698 0.7694215
3-2  -39.28509 -555.47385  476.9037 0.9980969
4-2  493.50541 -118.25895 1105.2698 0.2496929
4-3  532.79050  -40.93761 1106.5186 0.1439341
```

90 confidence interval is (-330.8438, 798.94247). We are 90% predict that the mean effect for the northeastern region on physicians differs from the effect for the Midwest region is between -330.8438 and 798.94247

- Test whether any geographic effects are present; use α =.10. State the alternatives, decision rule, and conclusion. What is the P-value of the test?

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

Response: physicians
                   Df     Sum Sq    Mean Sq   F value
tot_pop             1 1243181164 1243181164 3878.9792
tot_income          1   22058054   22058054   68.8256
as.factor(region)   3    1873626     624542    1.9487
Residuals         434  139093455     320492
                     Pr(>F)
tot_pop           < 2.2e-16 ***
tot_income        1.369e-15 ***
as.factor(region)     0.121
Residuals
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Ho: β1 = β2 = β3= β4

Ha: At least one of the β does not equal to the other Test statistic is F = 1.9487, p-value = 0.121 is large than α, we fail to reject null hypothesis.

Part IV.

What is the best model for predicting the # of active physicians in a county? Possible predictors include: Write the final expression of the estimated regression model, based on Adj R^2 and PRESS statistics. (Use Forward, Backward and Both stepwise selection procedure). Start with your full model as complete second order model.

Forward

```
Step:  AIC=5178.41
y ~ x5 + x2 + x1 + I(x4^2) + I(x1^2) + I(x5^2) + x3 + x8 + x5:x1 +
    x2:x8 + x1:x8

         Df Sum of Sq      RSS    AIC
<none>                53830559 5178.4
+ x5:x8    1    147006 53683552 5179.2
+ I(x7^2)  1    139434 53691125 5179.3
+ I(x8^2)  1    103365 53727194 5179.6
+ x7       1     83186 53747373 5179.7
+ x2:x5    1     51286 53779273 5180.0
+ I(x2^2)  1     45243 53785316 5180.0
+ x1:x3    1     21649 53808910 5180.2
+ x2:x3    1     13267 53817292 5180.3
+ x3:x8    1      4492 53826067 5180.4
+ x6       1      4428 53826131 5180.4
+ I(x3^2)  1      2348 53828211 5180.4
+ x1:x2    1      1395 53829164 5180.4
+ x3:x5    1      1360 53829199 5180.4
+ I(x6^2)  1       272 53830287 5180.4
+ x4       1         8 53830550 5180.4
```

```
Call:
lm(formula = y ~ x5 + x2 + x1 + I(x4^2) + I(x1^2) + I(x5^2) +
    x3 + x8 + x5:x1 + x2:x8 + x1:x8)

Residuals:
     Min       1Q   Median       3Q      Max
-1444.83  -116.84   -14.68    73.31  1943.73

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.149e+01  7.475e+01   0.287  0.77389
x5           5.512e-01  3.460e-02  15.931  < 2e-16 ***
x2           2.011e-01  2.799e-02   7.185 3.00e-12 ***
x1          -3.167e-03  6.262e-04  -5.057 6.32e-07 ***
I(x4^2)     -3.965e-01  1.341e-01  -2.956  0.00329 **
I(x1^2)      5.539e-10  1.085e-10   5.107 4.94e-07 ***
I(x5^2)      2.605e-05  8.634e-06   3.018  0.00270 **
x3           2.185e-02  1.241e-02   1.761  0.07901 .
x8          -5.757e+00  1.081e+01  -0.533  0.59454
x5:x1       -2.440e-07  5.993e-08  -4.072 5.56e-05 ***
x2:x8       -1.042e-02  4.363e-03  -2.388  0.01737 *
x1:x8        1.567e-04  8.641e-05   1.814  0.07045 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 354.6 on 428 degrees of freedom
Multiple R-squared:  0.9617,    Adjusted R-squared:  0.9607
F-statistic: 977.5 on 11 and 428 DF,  p-value: < 2.2e-16
```

```
Call:
lm(formula = y ~ x5 + x2 + x1 + I(x4^2) + I(x1^2) + I(x5^2) + 
    x8 + x5:x1 + x2:x8)

Residuals:
     Min       1Q   Median       3Q      Max 
-1401.98  -119.26   -14.40    74.91  2098.76 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) -3.055e+01  7.003e+01  -0.436 0.662869    
x5           5.322e-01  3.379e-02  15.750  < 2e-16 ***
x2           1.492e-01  9.769e-03  15.277  < 2e-16 ***
x1          -1.963e-03  2.414e-04  -8.134 4.50e-15 ***
I(x4^2)     -4.121e-01  1.338e-01  -3.079 0.002209 ** 
I(x1^2)      5.877e-10  1.081e-10   5.436 9.17e-08 ***
I(x5^2)      2.871e-05  8.609e-06   3.335 0.000925 ***
x8           4.480e+00  9.810e+00   0.457 0.648142    
x5:x1       -2.651e-07  5.961e-08  -4.447 1.11e-05 ***
x2:x8       -2.916e-03  1.253e-03  -2.328 0.020359 *  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 356.4 on 430 degrees of freedom
Multiple R-squared:  0.9612,    Adjusted R-squared:  0.9604 
F-statistic:  1183 on 9 and 430 DF,  p-value: < 2.2e-16
```

Keep $p<0.05$.

backward

```
Step:  AIC=5073.09
y ~ x1 + x2 + x3 + x4 + x5 + x6 + x7 + x8 + I(x1^2) + I(x2^2) + 
    I(x5^2) + I(x7^2) + I(x8^2) + x1:x2 + x1:x3 + x1:x4 + x1:x5 + 
    x1:x6 + x2:x5 + x2:x6 + x2:x7 + x3:x5 + x3:x6 + x4:x5 + x5:x6 + 
    x5:x8 + x6:x7 + x6:x8

          Df Sum of Sq      RSS    AIC
<none>                 39220369 5073.1
- I(x7^2)  1    196842 39417211 5073.3
- I(x2^2)  1    214790 39435159 5073.5
- x3:x6    1    240876 39461245 5073.8
- x1:x6    1    243320 39463689 5073.8
- I(x8^2)  1    397338 39617707 5075.5
- I(x5^2)  1    579882 39800251 5077.5
- x1:x5    1    615600 39835969 5077.9
- x1:x3    1    633225 39853594 5078.1
- x1:x2    1    687520 39907889 5078.7
- x2:x6    1    741737 39962106 5079.3
- I(x1^2)  1    835648 40056017 5080.4
- x3:x5    1   1132103 40352472 5083.6
- x2:x5    1   1478466 40698835 5087.4
- x5:x6    1   2097729 41318098 5094.0
- x1:x4    1   2261543 41481912 5095.8
- x5:x8    1   2460982 41681351 5097.9
- x4:x5    1   3240252 42460621 5106.0
- x6:x8    1   5685418 44905787 5130.7
- x2:x7    1   6014154 45234523 5133.9
- x6:x7    1   8521585 47741954 5157.6
```

```
Call:
lm(formula = y ~ x1 + x2 + x3 + x4 + x5 + x6 + x7 + x8 + I(x1^2) +
    I(x2^2) + I(x5^2) + I(x7^2) + I(x8^2) + x1:x2 + x1:x3 + x1:x4 +
    x1:x5 + x1:x6 + x2:x5 + x2:x6 + x2:x7 + x3:x5 + x3:x6 + x4:x5 +
    x5:x6 + x5:x8 + x6:x7 + x6:x8)

Residuals:
    Min      1Q  Median      3Q     Max
-935.74 -138.56   -8.44   97.78 1558.22

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -2.569e+03  1.381e+03  -1.860 0.063552 .
x1          -3.904e-03  6.405e-04  -6.095 2.52e-09 ***
x2          -5.831e-01  8.921e-02  -6.537 1.87e-10 ***
x3           2.671e-02  1.968e-02   1.357 0.175500
x4          -6.478e+00  6.156e+00  -1.052 0.293278
x5           7.361e-01  1.174e-01   6.268 9.25e-10 ***
x6           2.591e-01  2.599e-02   9.969  < 2e-16 ***
x7           6.375e+01  3.798e+01   1.678 0.094042 .
x8          -4.731e+01  3.058e+01  -1.547 0.122637
I(x1^2)      4.975e-09  1.681e-09   2.959 0.003263 **
I(x2^2)      3.244e-06  2.162e-06   1.500 0.134310
I(x5^2)      3.220e-05  1.306e-05   2.465 0.014105 *
I(x7^2)     -3.665e-01  2.552e-01  -1.436 0.151697
I(x8^2)      3.173e+00  1.555e+00   2.041 0.041935 *
x1:x2       -3.185e-07  1.187e-07  -2.684 0.007565 **
x1:x3        1.787e-07  6.938e-08   2.576 0.010344 *
x1:x4        1.877e-04  3.855e-05   4.868 1.61e-06 ***
x1:x5       -5.005e-07  1.971e-07  -2.540 0.011456 *
x1:x6       -1.478e-08  9.255e-09  -1.597 0.111076
x2:x5        3.029e-05  7.696e-06   3.936 9.72e-05 ***
x2:x6        1.272e-06  4.562e-07   2.788 0.005550 **
x2:x7        8.526e-03  1.074e-03   7.939 1.97e-14 ***
x3:x5       -1.180e-04  3.425e-05  -3.444 0.000631 ***
```

Keep p<0.05,

```
Residuals:
    Min      1Q  Median      3Q     Max
-898.82 -127.86   -9.49   87.35 1638.95

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -3.335e+01  3.786e+01  -0.881 0.378869
x1          -4.264e-03  5.698e-04  -7.483 4.31e-13 ***
x2          -5.166e-01  6.883e-02  -7.505 3.73e-13 ***
x5           6.870e-01  1.110e-01   6.191 1.43e-09 ***
x6           2.299e-01  2.111e-02  10.892  < 2e-16 ***
I(x1^2)      2.470e-09  6.872e-10   3.594 0.000364 ***
I(x5^2)      3.924e-05  1.212e-05   3.238 0.001300 **
I(x8^2)      8.617e-02  4.630e-01   0.186 0.852463
x1:x2       -1.315e-07  3.234e-08  -4.066 5.71e-05 ***
x1:x3        2.137e-07  6.837e-08   3.125 0.001902 **
x1:x4        1.725e-04  3.451e-05   4.997 8.56e-07 ***
x1:x5       -4.428e-07  1.935e-07  -2.288 0.022609 *
x2:x5        2.921e-05  7.548e-06   3.870 0.000126 ***
x2:x6        5.348e-07  2.155e-07   2.482 0.013472 *
x2:x7        7.913e-03  7.844e-04  10.087  < 2e-16 ***
x5:x3       -8.026e-05  2.697e-05  -2.976 0.003092 **
x5:x4       -4.991e-02  8.379e-03  -5.956 5.47e-09 ***
x5:x6       -5.876e-06  1.027e-06  -5.721 2.02e-08 ***
x5:x8        6.496e-02  1.117e-02   5.814 1.21e-08 ***
x6:x7       -2.385e-03  2.228e-04 -10.705  < 2e-16 ***
x6:x8       -5.037e-03  6.692e-04  -7.527 3.20e-13 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 314.3 on 419 degrees of freedom
Multiple R-squared:  0.9706,    Adjusted R-squared:  0.9692
F-statistic:   691 on 20 and 419 DF,  p-value: < 2.2e-16
```

Both

```
Step:  AIC=5073.09
y ~ x1 + x2 + x3 + x4 + x5 + x6 + x7 + x8 + I(x1^2) + I(x2^2) +
    I(x5^2) + I(x7^2) + I(x8^2) + x1:x2 + x1:x3 + x1:x4 + x1:x5 +
    x1:x6 + x2:x5 + x2:x6 + x2:x7 + x3:x5 + x3:x6 + x4:x5 + x5:x6 +
    x5:x8 + x6:x7 + x6:x8

          Df Sum of Sq      RSS    AIC
<none>                  39220369 5073.1
- I(x7^2)  1    196842 39417211 5073.3
+ x2:x8    1    143902 39076467 5073.5
- I(x2^2)  1    214790 39435159 5073.5
- x3:x6    1    240876 39461245 5073.8
- x1:x6    1    243320 39463689 5073.8
+ x3:x4    1     93082 39127287 5074.0
+ x4:x7    1     83294 39137075 5074.2
+ x1:x8    1     78355 39142014 5074.2
+ x4:x8    1     41033 39179336 5074.6
+ I(x3^2)  1     36586 39183783 5074.7
+ x3:x8    1     34234 39186135 5074.7
+ x2:x4    1     31775 39188594 5074.7
+ x5:x7    1     29478 39190891 5074.8
+ x1:x7    1     20654 39199715 5074.9
+ x4:x6    1     10683 39209686 5075.0
+ x2:x3    1      9652 39210717 5075.0
+ I(x4^2)  1      3484 39216885 5075.1
+ I(x6^2)  1       499 39219870 5075.1
+ x3:x7    1       364 39220005 5075.1
+ x7:x8    1        75 39220294 5075.1
- I(x8^2)  1    397338 39617707 5075.5
- I(x5^2)  1    579882 39800251 5077.5
- x1:x5    1    615600 39835969 5077.9
- x1:x3    1    633225 39853594 5078.1
- x1:x5    1    615600 39835969 5077.9
- x1:x3    1    633225 39853594 5078.1
- x1:x2    1    687520 39907889 5078.7
- x2:x6    1    741737 39962106 5079.3
- I(x1^2)  1    835648 40056017 5080.4
- x3:x5    1   1132103 40352472 5083.6
- x2:x5    1   1478466 40698835 5087.4
- x5:x6    1   2097729 41318098 5094.0
- x1:x4    1   2261543 41481912 5095.8
- x5:x8    1   2460982 41681351 5097.9
- x4:x5    1   3240252 42460621 5106.0
- x6:x8    1   5685418 44905787 5130.7
- x2:x7    1   6014154 45234523 5133.9
- x6:x7    1   8521585 47741954 5157.6
```

```
Call:
lm(formula = y ~ x1 + x2 + x3 + x4 + x5 + x6 + x7 + x8 + I(x1^2) +
    I(x2^2) + I(x5^2) + I(x7^2) + I(x8^2) + x1:x2 + x1:x3 + x1:x4 +
    x1:x5 + x1:x6 + x2:x5 + x2:x6 + x2:x7 + x3:x5 + x3:x6 + x4:x5 +
    x5:x6 + x5:x8 + x6:x7 + x6:x8)

Residuals:
    Min      1Q  Median      3Q     Max
-935.74 -138.56   -8.44   97.78 1558.22

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -2.569e+03  1.381e+03  -1.860 0.063552 .
x1          -3.904e-03  6.405e-04  -6.095 2.52e-09 ***
x2          -5.831e-01  8.921e-02  -6.537 1.87e-10 ***
x3           2.671e-02  1.968e-02   1.357 0.175500
x4          -6.478e+00  6.156e+00  -1.052 0.293278
x5           7.361e-01  1.174e-01   6.268 9.25e-10 ***
x6           2.591e-01  2.599e-02   9.969  < 2e-16 ***
x7           6.375e+01  3.798e+01   1.678 0.094042 .
x8          -4.731e+01  3.058e+01  -1.547 0.122637
I(x1^2)      4.975e-09  1.681e-09   2.959 0.003263 **
I(x2^2)      3.244e-06  2.162e-06   1.500 0.134310
I(x5^2)      3.220e-05  1.306e-05   2.465 0.014105 *
I(x7^2)     -3.665e-01  2.552e-01  -1.436 0.151697
I(x8^2)      3.173e+00  1.555e+00   2.041 0.041935 *
x1:x2       -3.185e-07  1.187e-07  -2.684 0.007565 **
x1:x3        1.787e-07  6.938e-08   2.576 0.010344 *
x1:x4        1.877e-04  3.855e-05   4.868 1.61e-06 ***
x1:x5       -5.005e-07  1.971e-07  -2.540 0.011456 *
x1:x6       -1.478e-08  9.255e-09  -1.597 0.111076
x2:x5        3.029e-05  7.696e-06   3.936 9.72e-05 ***
x2:x6        1.272e-06  4.562e-07   2.788 0.005550 **
x2:x7        8.526e-03  1.074e-03   7.939 1.97e-14 ***
x3:x5       -1.180e-04  3.425e-05  -3.444 0.000631 ***
x3:x6        1.518e-06  9.553e-07   1.589 0.112881
x4:x5       -5.054e-02  8.672e-03  -5.827 1.14e-08 ***
x5:x6       -5.363e-06  1.144e-06  -4.689 3.75e-06 ***
x5:x8        6.808e-02  1.341e-02   5.078 5.79e-07 ***
x6:x7       -2.767e-03  2.928e-04  -9.450  < 2e-16 ***
x6:x8       -5.590e-03  7.242e-04  -7.719 9.02e-14 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 308.9 on 411 degrees of freedom
Multiple R-squared:  0.9721,    Adjusted R-squared:  0.9702
F-statistic: 511.6 on 28 and 411 DF,  p-value: < 2.2e-16

>
```

Keep p<0.05

```
lm(formula = y ~ x1 + x2 + x5 + x6 + I(x1^2) + I(x5^2) + I(x8^2) +
    x1:x2 + x1:x3 + x1:x4 + x1:x5 + x2:x5 + x2:x6 + x2:x7 + x3:x5 +
    x4:x5 + x5:x6 + x5:x8 + x6:x7 + x6:x8)

Residuals:
    Min      1Q  Median      3Q     Max
-898.82 -127.86   -9.49   87.35 1638.95

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -3.335e+01  3.786e+01  -0.881 0.378869
x1          -4.264e-03  5.698e-04  -7.483 4.31e-13 ***
x2          -5.166e-01  6.883e-02  -7.505 3.73e-13 ***
x5           6.870e-01  1.110e-01   6.191 1.43e-09 ***
x6           2.299e-01  2.111e-02  10.892  < 2e-16 ***
I(x1^2)      2.470e-09  6.872e-10   3.594 0.000364 ***
I(x5^2)      3.924e-05  1.212e-05   3.238 0.001300 **
I(x8^2)      8.617e-02  4.630e-01   0.186 0.852463
x1:x2       -1.315e-07  3.234e-08  -4.066 5.71e-05 ***
x1:x3        2.137e-07  6.837e-08   3.125 0.001902 **
x1:x4        1.725e-04  3.451e-05   4.997 8.56e-07 ***
x1:x5       -4.428e-07  1.935e-07  -2.288 0.022609 *
x2:x5        2.921e-05  7.548e-06   3.870 0.000126 ***
x2:x6        5.348e-07  2.155e-07   2.482 0.013472 *
x2:x7        7.913e-03  7.844e-04  10.087  < 2e-16 ***
x5:x3       -8.026e-05  2.697e-05  -2.976 0.003092 **
x5:x4       -4.991e-02  8.379e-03  -5.956 5.47e-09 ***
x5:x6       -5.876e-06  1.027e-06  -5.721 2.02e-08 ***
x5:x8        6.496e-02  1.117e-02   5.814 1.21e-08 ***
x6:x7       -2.385e-03  2.228e-04 -10.705  < 2e-16 ***
x6:x8       -5.037e-03  6.692e-04  -7.527 3.20e-13 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 314.3 on 419 degrees of freedom
Multiple R-squared:  0.9706,    Adjusted R-squared:  0.9692
F-statistic:   691 on 20 and 419 DF,  p-value: < 2.2e-16
```

```
Call:
lm(formula = y ~ x1 + x2 + x5 + x6 + I(x1^2) + I(x5^2) + x1:x2 +
    x1:x3 + x1:x4 + x1:x5 + x2:x5 + x2:x6 + x2:x7 + x3:x5 + x4:x5 +
    x5:x6 + x5:x8 + x6:x7 + x6:x8)

Residuals:
    Min      1Q  Median      3Q     Max
-898.48 -128.53   -9.54   88.91 1641.05

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -2.961e+01  3.204e+01  -0.924 0.355904
x1          -4.255e-03  5.674e-04  -7.500 3.83e-13 ***
x2          -5.168e-01  6.874e-02  -7.518 3.40e-13 ***
x5           6.874e-01  1.108e-01   6.203 1.33e-09 ***
x6           2.297e-01  2.106e-02  10.908  < 2e-16 ***
I(x1^2)      2.464e-09  6.858e-10   3.593 0.000365 ***
I(x5^2)      3.957e-05  1.197e-05   3.305 0.001032 **
x1:x2       -1.314e-07  3.230e-08  -4.069 5.64e-05 ***
x1:x3        2.114e-07  6.721e-08   3.145 0.001776 **
x1:x4        1.730e-04  3.434e-05   5.038 7.00e-07 ***
x1:x5       -4.446e-07  1.931e-07  -2.303 0.021761 *
x2:x5        2.928e-05  7.530e-06   3.889 0.000117 ***
x2:x6        5.397e-07  2.136e-07   2.527 0.011871 *
x2:x7        7.909e-03  7.833e-04  10.097  < 2e-16 ***
x5:x3       -7.931e-05  2.646e-05  -2.998 0.002881 **
x5:x4       -5.012e-02  8.291e-03  -6.045 3.30e-09 ***
x5:x6       -5.907e-06  1.013e-06  -5.830 1.11e-08 ***
x5:x8        6.516e-02  1.111e-02   5.865 9.11e-09 ***
x6:x7       -2.385e-03  2.225e-04 -10.717  < 2e-16 ***
x6:x8       -5.011e-03  6.535e-04  -7.667 1.23e-13 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 313.9 on 420 degrees of freedom
Multiple R-squared:  0.9706,    Adjusted R-squared:  0.9692
F-statistic:   729 on 19 and 420 DF,  p-value: < 2.2e-16
```

Press

```
> PRESS.statistic <- sum((resid(forward_red)/(1-hatvalues(forward_re
d))) ^2)
>  print (paste ("PRESS statistic= ", PRESS.statistic))
[1] "PRESS statistic=  73023898.5749624"
> PRESS.statistic <- sum((resid(backward_red)/(1-hatvalues(backward_re
d))) ^2)
> print (paste ("PRESS statistic= ", PRESS.statistic))
[1] "PRESS statistic=  97445809.2663577"
> PRESS.statistic <- sum((resid(both_red2)/(1-hatvalues(both_red2)))^2)
> print (paste ("PRESS statistic= ", PRESS.statistic))
[1] "PRESS statistic=  96482848.7014274"
>
```

By press st and ad r square, the final equation should be Physicians= -3.055 + 0.53*beds + 0.1492*tot_income - 0.001963*totpop-0.4121* pop_elderly^2 + 0.0000000005877*totpop^2+ 0.00002871* beds^2+4.48unemploy-0.0000002651* beds: totpop – 0.002916* tot_income: unemploy

a. Interpret each of the regression coefficients for the final model.
   β0 means that the expect value of physicians is -3.055 when all independent variables are 0.
   β1 increases the expect value of physicians by 0.53 for each unit increase in bed, with other variables held fixed.
   β2 increases the expect value of physicians by 0.1492 for each unit increase in total income, with other variables held fixed.

β3 decreases the expect value of physicians by 0.001963 for each unit increase in total population, with other variables held fixed.
B7 increase the expect value of physicians by 4.48 for each unit increase in unemploy. population, with other variables held fixed.
β4 is -0.4121, β5 is 0.0000000005877 and β6 is 0.00002871 for the type of surface and rates of curvature.
B8 is -0.0000002651 and β9 is 0.002916 for the rate of twist in the ruled surface.

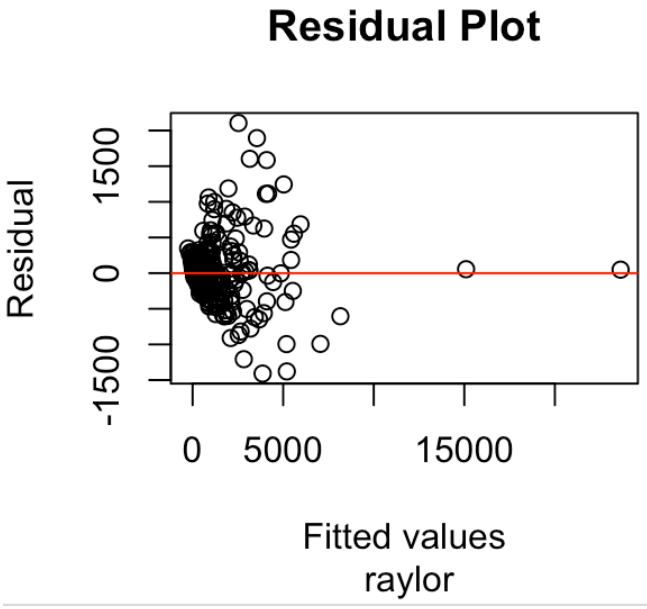
b. Discuss the coefficient of determination, R-squared and Adj R-squared for the final model.

r-2=0.9611 ad=0.9604 96.11 percentage of the variability in the physicians are explained by the eight predictor variables.

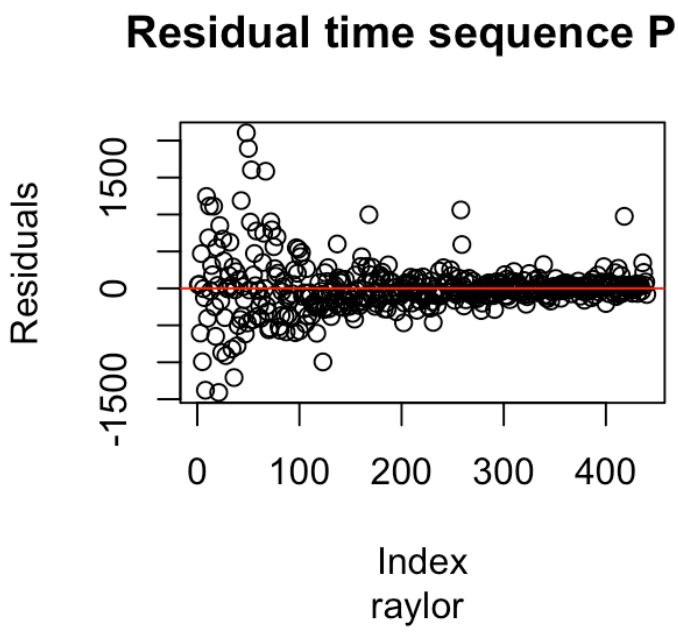
c. Test overall F-test for regression for the final model. Make sure to write null and alternate hypothesis, test statistic, p-value and conclusion.

Ho: β1 = β2 = β3 = β4= β5= β6= β7= β8 = 0 Ha: At least one βi is not 0, i= 1~8 F = 1332.8 degrees of freedom is 8 and 431 p-value = 0.00000000000000022, is less than α, so we reject null hypothesis and conclude that at least one βi is not 0.

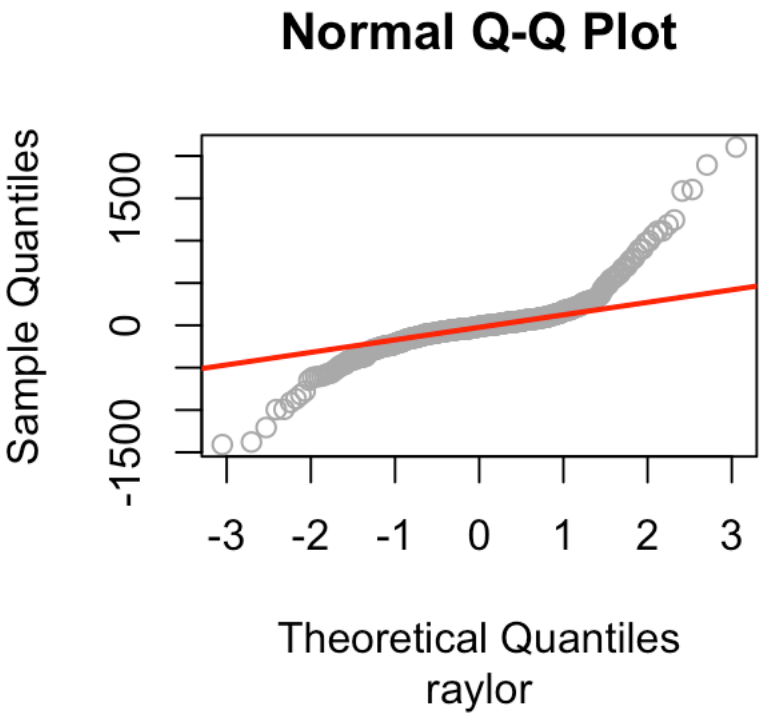
d. Discuss the model assumptions. (Both graphical and hypothesis tests, Make sure to write null and alternate hypothesis, test statistic, p-value and conclusion.



it is not constant.



it is not independent



not normal

```
> dwtest(forward_red)

        Durbin-Watson test

data:  forward_red
DW = 2.1803, p-value = 0.9655
alternative hypothesis: true autocorrelation is greater than 0
```

Ho: The error variance is constant Ha: The error variance is not constant BP = 642.79, degrees of freedom is 1 p-value = 0.00000000000000022, is less than α, so. we reject null hypothesis and state that the variance is not equal or constant.

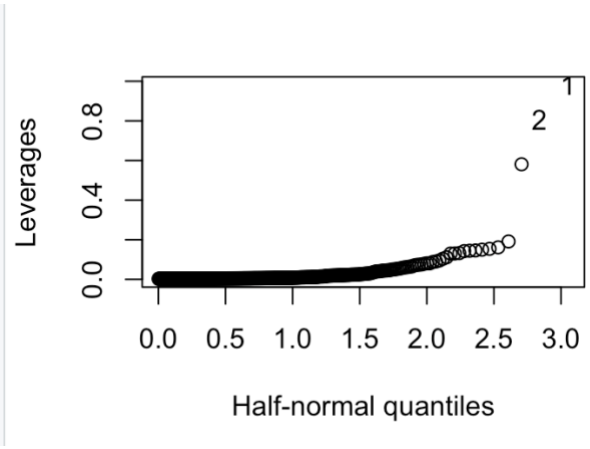
```
>  shapiro.test(resid(forward_red))

        Shapiro-Wilk normality test

data:  resid(forward_red)
W = 0.82356, p-value < 2.2e-16
```

Ho: Random error comes from Normal distribution Ha: Random error does not come from Normal distribution. W = 0.82356 p-value = 0 is less than α, we reject null hypothesis and that random error does not come from Normal distribution.

```
> dwtest(forward_red)

        Durbin-Watson test

data:  forward_red
DW = 2.1803, p-value = 0.9655
alternative hypothesis: true autocorrelation is greater than 0
```

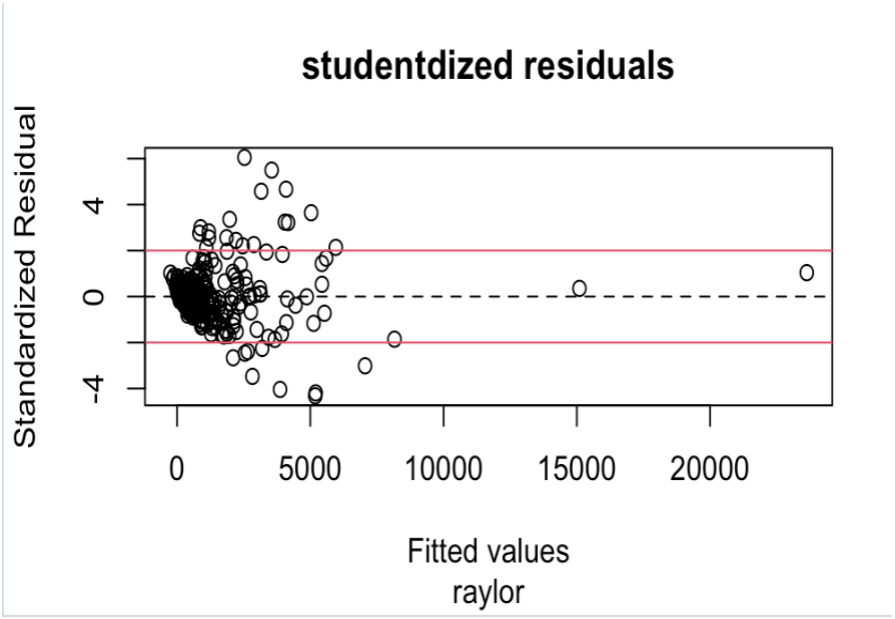
Ho: The errors are uncorrelated over time. Ha: The errors are positively correlated. DW = 2.1803 p-value = 0.9655, is larger than α, we fail to reject null hypothesis ad state that the errors are not positively correlated, and they are independent.

e. Use the best model and plot the leverage. Is there any observation with a dangerously large leverage? Try to identify in which variable the problem lies. Find out which county this is?.



there are two observations with a dangerously large leverage. Variable: country / la and cook

f. Using the studentized residuals identify the outliers. Which county produced the largest residuals?

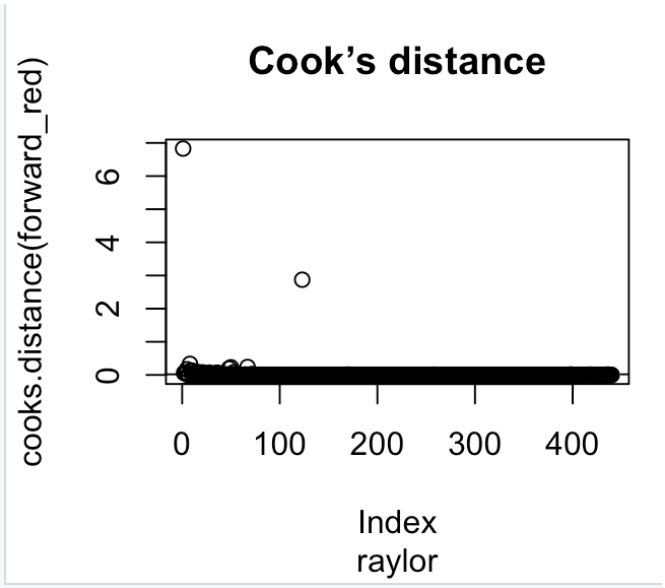


There are outlinears

```
> (1: length(rstandard(forward_new)))[rstandard(forward_new)>2]
 [1]  16  20  28  41  44  63  65  67  68 102 126 133 222 223
> (1: length(rstandard(forward_new)))[rstandard(forward_new)< -2]
[1]  19  24  26  39  46  53  66 166 195
```

Montgomery produced the largest residuals.

g. Calculate Cook's distance, Di and plot it. Any problems? What about the problematic counties identified from (e) and (f)?




```
 [1] "1"   "2"   "3"   "4"   "5"   "8"   "9"   "11"  "12"  "16"  "18"  "19"  "21"
[14] "22"  "24"  "25"  "28"  "32"  "34"  "36"  "39"  "42"  "43"  "47"  "48"  "50"
[27] "52"  "53"  "58"  "65"  "67"  "70"  "72"  "73"  "123" "171" "398" "418" "436"
```

The problematic counties identified from (e) and (f) is similar in the cook's distance.

h. Leave out the problematic county from (g) and re-fit the largest model. Then redo (c)–(g) and see what happens.

```
Call:
lm(formula = y ~ x5 + x2 + x1 + I(x4^2) + I(x1^2) + I(x5^2) +
    x8 + x5:x1 + x2:x8)

Residuals:
    Min      1Q  Median      3Q     Max
-768.67  -87.46  -10.12   60.31 2208.74

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -5.848e+01  5.593e+01  -1.046 0.296343
x5           4.154e-01  3.715e-02  11.179  < 2e-16 ***
x2           1.217e-01  9.276e-03  13.121  < 2e-16 ***
x1          -9.941e-04  2.839e-04  -3.502 0.000515 ***
I(x4^2)     -2.718e-01  1.182e-01  -2.299 0.022011 *
I(x1^2)      1.264e-09  2.438e-10   5.183 3.49e-07 ***
I(x5^2)      1.002e-04  1.420e-05   7.053 7.90e-12 ***
x8           2.731e+00  7.218e+00   0.378 0.705364
x5:x1       -7.314e-07  1.143e-07  -6.399 4.44e-10 ***
x2:x8       -3.438e-03  1.166e-03  -2.948 0.003388 **
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 243.8 on 394 degrees of freedom
Multiple R-squared:  0.9315,    Adjusted R-squared:   0.93
F-statistic: 595.6 on 9 and 394 DF,  p-value: < 2.2e-16
```

This is the most fit model.

```
>  bptest(forward_red, studentize = FALSE)

        Breusch-Pagan test

data:  forward_red
BP = 656.56, df = 9, p-value < 2.2e-16
```

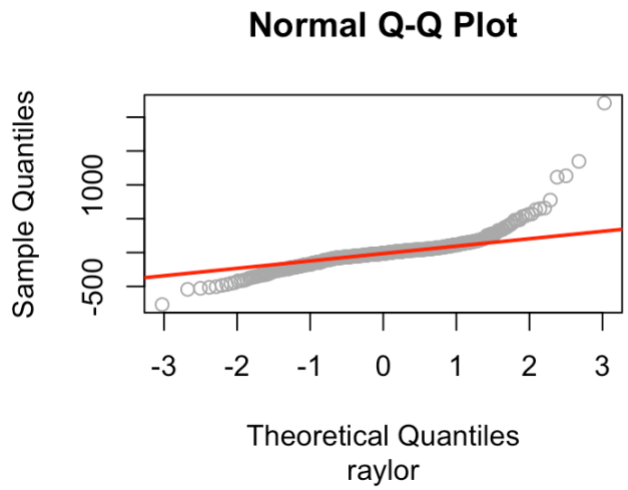
Ho: The error variance is constant Ha: The error variance is not constant BP = 656.56, degrees of freedom is 9 p-value = 0.00000000000000022, is less than α, so. we reject null hypothesis and state that the variance is not equal or constant.

```
>  shapiro.test(resid(forward_red))

        Shapiro-Wilk normality test

data:  resid(forward_red)
W = 0.78112, p-value < 2.2e-16
```

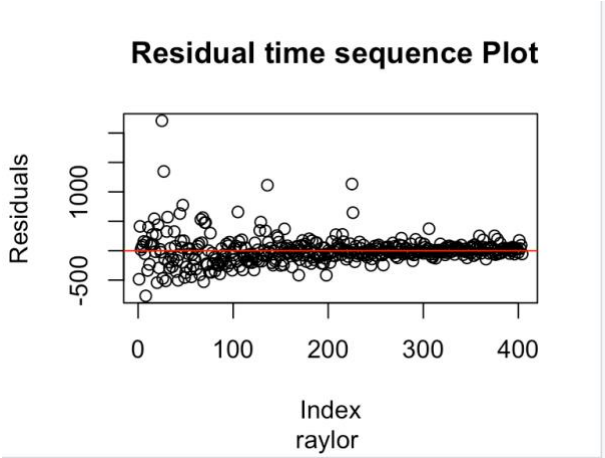
Ho: Random error comes from Normal distribution Ha: Random error does not come from Normal distribution. W = 0.78112 p-value = 0.000000000000000022 is less than α, we reject null hypothesis and that random error does not come from Normal distribution.

```
>  dwtest(forward_red)

        Durbin-Watson test

data:  forward_red
DW = 2.1032, p-value = 0.8277
alternative hypothesis: true autocorrelation is greater than 0
```

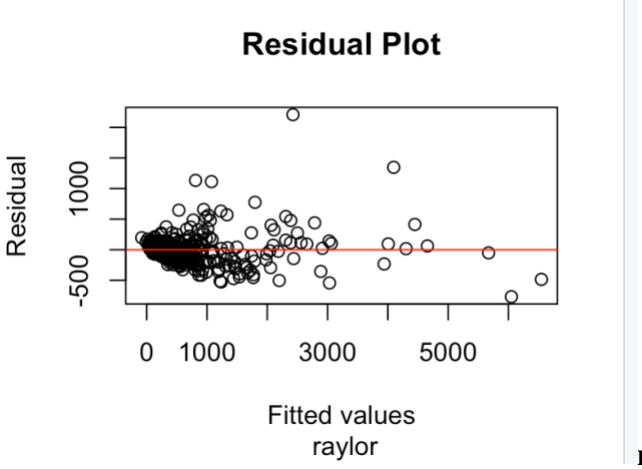
Ho: The errors are uncorrelated over time. Ha: The errors are positively correlated. DW = 2.1032 p-value = 0.8277 is larger than α, we fail to reject null hypothesis ad state that the errors are not positively correlated, and they are independent.



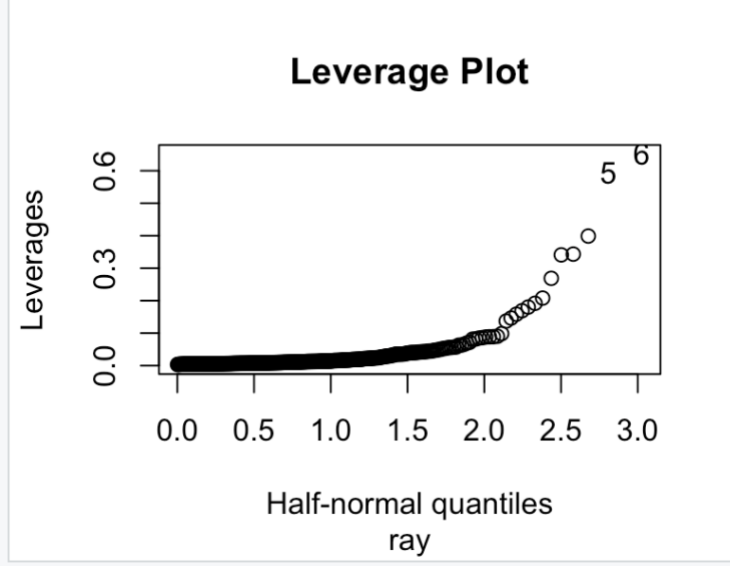
not normal



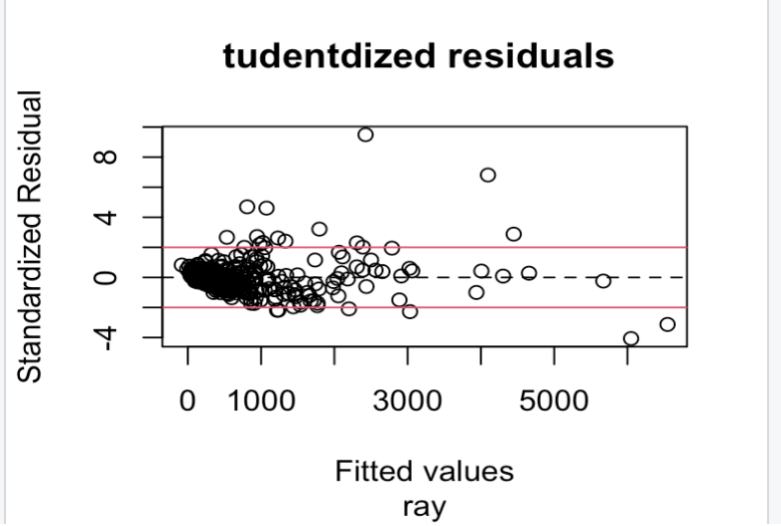
independent



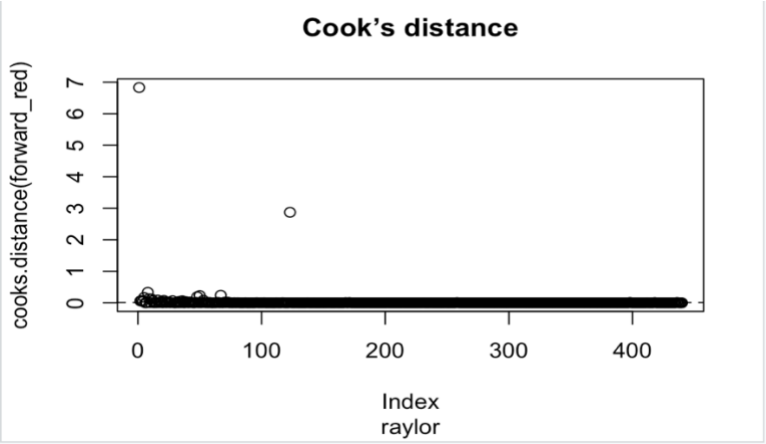
not constant



there are two outlinears.



country Kent



the country found in previous also found in this.

```
cdi <- read.table(file.choose(),header=TRUE,col.names = c ('county', 'state', 'land_area',
'tot_pop', 'pop_young','pop_elderly', 'physicians', 'beds', 'crimes', 'hsgrad', 'bsgrad','pov',
'unemploy', 'income', 'tot_income', 'region'))
> View(cdi)
attach (cdi)
plot (physicians~ tot_pop, main="Active physicians against Total Population", sub="raylor")
plot (physicians~ tot_income, main="Active physicians against Total Income", sub="raylor")
plot (physicians~ income, main="Active physicians against Income", sub="raylor")
plot (physicians~ crimes, main="Active physicians against Crimes", sub="raylor")
plot (physicians~pop_elderly, main="Active physicians against 65 or older", sub="raylor")
a= lm(physicians~tot_pop)
 summary (a)
 b= lm(physicians~beds)
 summary (b)
 c= lm(physicians~tot_income)
 summary (c)
  plot(physicians~ tot_pop, main=" Active physicians against Total Population", sub= "raylor")
  abline( lm( physicians~ tot_pop))
 plot(physicians~beds, main="Active physicians against beds", sub= "raylor")
 abline(lm(physicians~beds))
 plot (physicians~ tot_income, main="Active physicians against Total Income", sub= "raylor")
 abline( lm( physicians~ tot_income))
  new.data= data.frame( tot_pop= c (500000, 1000000, 5000000))
  predict (a, newdata = new.data, interval= 'confidence', level=1- 0.1/ 2)
d= lm(income~ bsgrad+ as.factor(region))
  summary (d)
   anova(d)
   aov.d <- aov (physicians ~ factor(region))
```

```
  TukeyHSD (aov.d)
> y = physicians
> x1 = tot_pop
> x2 = tot_income
> x3 = land_area
> x4 = pop_elderly
> x5 = beds
> x6 = crimes
> x7 = hsgrad
> x8 = unemploy
> null <- lm(y~1)
> full = lm(y~(x1+x2+x3+x4+x5+x6+x7+x8) ^2
        +I(x1^2) +I(x2^2) +I(x3^2) +I(x4^2) +I(x5^2) +I(x6^2) +I(x7^2) +I(x8^2))

> forward.step = step (null, data=CDI, list(upper=full), direction="forward")
> forward = lm(y ~ x5 + x2 + x1 + I(x4^2) + I(x1^2) + I(x5^2) + x3 + x8 + x5:x1 +
            x2:x8 + x1:x8)
> summary(forward)
> forward_red <- lm(y ~ x5 + x2 + x1 + I(x4^2) + I(x1^2) + I(x5^2) + x5:x1 + x2:x8)
> summary(forward_red)

> backward.step = step (full, direction="backward")
> backward = lm(y ~ x1 + x2 + x3 + x4 + x5 + x6 + x7 + x8 + I(x1^2) + I(x2^2) +
I(x5^2) + I(x7^2) + I(x8^2) + x1:x2 + x1:x3 + x1:x4 + x1:x5 +
x1:x6 + x2:x5 + x2:x6 + x2:x7 + x3:x5 + x3:x6 + x4:x5 + x5:x6 +
x5:x8 + x6:x7 + x6:x8)

> summary(backward)
```

```
backward_red <- lm(y ~ x1 + x2 + x5 + x6 + I(x1^2) +
                   I(x5^2) + I(x8^2) + x1:x2 + x1:x3 + x1:x4 + x1:x5 +
                   x2:x5 + x2:x6 + x2:x7 + x3:x5 + x4:x5 + x5:x6 +
                   x5:x8 + x6:x7 + x6:x8)
summary (backward_red)


 full.step.both=step (full, direction="both")
 both = lm(y ~ x1 + x2 + x3 + x4 + x5 + x6 + x7 + x8 + I(x1^2) + I(x2^2) +
           I(x5^2) + I(x7^2) + I(x8^2) + x1:x2 + x1:x3 + x1:x4 + x1:x5 +
           x1:x6 + x2:x5 + x2:x6 + x2:x7 + x3:x5 + x3:x6 + x4:x5 + x5:x6 +
           x5:x8 + x6:x7 + x6:x8)
 summary(both)


 both_red <- lm(y ~ x1 + x2+ x5 + x6 + I(x1^2) + I(x5^2) + I(x8^2) + x1:x2 + x1:x3 + x1:x4 +
x1:x5 + x2:x5 + x2:x6 + x2:x7 + x3:x5+ x4:x5 + x5:x6 +
            x5:x8 + x6:x7 + x6:x8)
summary(both_red)


both_red2= lm(formula = y ~ x1 + x2 + x5 + x6 + I(x1^2) + I(x5^2) +
              x1:x2 + x1:x3 + x1:x4 + x1:x5 + x2:x5 + x2:x6 + x2:x7 + x3:x5 +
              x4:x5 + x5:x6 + x5:x8 + x6:x7 + x6:x8)
summary(both_red2)
summary(forward)
> a=lm(y ~ x5 + x2 + x1 + I(x4^2) + I(x1^2) + I(x5^2) + x3 + x8 + x5:x1 +  x2:x8 + x1:x8)
> summary(a)
PRESS.statistic <- sum((resid(forward_red)/(1-hatvalues(forward_red))) ^2)
 print (paste ("PRESS statistic= ", PRESS.statistic))
PRESS.statistic <- sum((resid(backward_red)/(1-hatvalues(backward_red))) ^2)
print (paste ("PRESS statistic= ", PRESS.statistic))
```

```
PRESS.statistic <- sum((resid(both_red2)/(1-hatvalues(both_red2)))^2)

print (paste ("PRESS statistic= ", PRESS.statistic))

  > b= lm(y~1)

>  anova(b, forward_red)

plot (forward_red$fitted.values, forward_red$residuals, main="Residual Plot", sub= "
raylor",xlab="Fitted values",ylab="Residual")

abline (h=0, col="red")

plot(forward_red$residuals, ylab="Residuals",main=" Residual time sequence Plot", sub=
"raylor ")

 abline(h=0, col="red")

 qqnorm(resid(forward_red), main = "Normal Q-Q Plot", col = "darkgrey", sub="raylor")

qqline(resid(forward_red), col = "red", lwd = 2)

 bptest(forward_red, studentize = FALSE)

 dwtest(forward_red)

 shapiro.test(resid(forward_red))

e <- hatvalues(forward_red)

 f<- rstandard(forward_red)

 id <- row.names(cdi)

  library(faraway)

 halfnorm(e, labs=id, ylab="Leverages")

standard_res <- rstandard(forward_red)

 plot(forward_red$fitted.values,standard_res,main="studentdized residuals", xlab="Fitted
values",ylab="Standardized Residual", sub="raylor")

abline(h=c (-2,0,2), col=c (2,1,2),lty=c (1,2,1))

identify(forward_red$fitted.values, rstandard(forward_red), labels=row.names(CDI))

 (1: length(rstandard(forward_red)))[rstandard(forward_red)>2]

 (1: length(rstandard(forward_red)))[rstandard(forward_red)< -2]

plot(cooks.distance(forward_red), main="Cook’s distance", sub="raylor")

cutoff <- with(forward_red, 8/df.residual)
```

```
> cdinew <- read.table(file.choose(),header=TRUE,col.names = c ('county', 'state', 'land_area',
'tot_pop', 'pop_young','pop_elderly', 'physicians', 'beds', 'crimes', 'hsgrad', 'bsgrad','pov',
'unemploy', 'income', 'tot_income', 'region'))
Warning message:
In read.table(file.choose(), header = TRUE, col.names = c("county",  :
  header and 'col.names' are of different lengths
> attach (cdinew)
> y = physicians
> x1 = tot_pop
> x2 = tot_income
> x3 = land_area
> x4 = pop_elderly
> x5 = beds
> x7 = hsgrad
> x6 = crimes
> x8 = unemploy
> null <- lm(y~1)
> full = lm(y~(x1+x2+x3+x4+x5+x6+x7+x8) ^2
        +I(x1^2) +I(x2^2) +I(x3^2) +I(x4^2) +I(x5^2) +I(x6^2) +I(x7^2) +I(x8^2))

> forward.step = step (null, data=CDI, list(upper=full), direction="forward")
> forward = lm(y ~ x5 + x2 + x1 + I(x4^2) + I(x1^2) + I(x5^2) + x3 + x8 + x5:x1 +
            x2:x8 + x1:x8)
> summary(forward)
> forward_red <- lm(y ~ x5 + x2 + x1 + I(x4^2) + I(x1^2) + I(x5^2) + x5:x1 + x2:x8)
> summary(forward_red)

> backward.step = step (full, direction="backward")
```

```
> backward = lm(y ~ x1 + x2 + x3 + x4 + x5 + x6 + x7 + x8 + I(x1^2) + I(x2^2) +
I(x5^2) + I(x7^2) + I(x8^2) + x1:x2 + x1:x3 + x1:x4 + x1:x5 +
x1:x6 + x2:x5 + x2:x6 + x2:x7 + x3:x5 + x3:x6 + x4:x5 + x5:x6 +
x5:x8 + x6:x7 + x6:x8)

> summary(backward)
backward_red <- lm(y ~ x1 + x2 + x5 + x6 + I(x1^2) +
                   I(x5^2) + I(x8^2) + x1:x2 + x1:x3 + x1:x4 + x1:x5 +
                   x2:x5 + x2:x6 + x2:x7 + x3:x5 + x4:x5 + x5:x6 +
                   x5:x8 + x6:x7 + x6:x8)
summary (backward_red)

 full.step.both=step (full, direction="both")
 both = lm(y ~ x1 + x2 + x3 + x4 + x5 + x6 + x7 + x8 + I(x1^2) + I(x2^2) +
           I(x5^2) + I(x7^2) + I(x8^2) + x1:x2 + x1:x3 + x1:x4 + x1:x5 +
           x1:x6 + x2:x5 + x2:x6 + x2:x7 + x3:x5 + x3:x6 + x4:x5 + x5:x6 +
           x5:x8 + x6:x7 + x6:x8)
 summary(both)

 both_red <- lm(y ~ x1 + x2+ x5 + x6 + I(x1^2) + I(x5^2) + I(x8^2) + x1:x2 + x1:x3 + x1:x4 +
x1:x5 + x2:x5 + x2:x6 + x2:x7 + x3:x5+ x4:x5 + x5:x6 +
                 x5:x8 + x6:x7 + x6:x8)
summary(both_red)

both_red2= lm(formula = y ~ x1 + x2 + x5 + x6 + I(x1^2) + I(x5^2) +
                x1:x2 + x1:x3 + x1:x4 + x1:x5 + x2:x5 + x2:x6 + x2:x7 + x3:x5 +
                x4:x5 + x5:x6 + x5:x8 + x6:x7 + x6:x8)
summary(both_red2)
summary(forward)
```

```
> a= lm(y~1)

> anova(a, forward_red)

> plot (forward_red$fitted.values, forward_red$residuals, main="Residual Plot", sub= "
raylor",xlab="Fitted values",ylab="Residual")

> abline (h=0, col="red")

> plot(forward_red$residuals, ylab="Residuals",main=" Residual time sequence Plot", sub=
"raylor ")

>  abline(h=0, col="red")

>  qqnorm(resid(forward_red), main = "Normal Q-Q Plot", col = "darkgrey", sub="raylor")

> qqline(resid(forward_red), col = "red", lwd = 2)

>  bptest(forward_red, studentize = FALSE)

        Breusch-Pagan test

data:  forward_red

BP = 656.56, df = 9, p-value < 2.2e-16

>  dwtest(forward_red)

        Durbin-Watson test

data:  forward_red

DW = 2.1032, p-value = 0.8277

alternative hypothesis: true autocorrelation is greater than 0

>  shapiro.test(resid(forward_red))

        Shapiro-Wilk normality test
```

data: resid(forward_red)

W = 0.78112, p-value < 2.2e-16

e <- hatvalues(forward_red)

f<- rstandard(forward_red)

id <- row.names(cdi)

library(faraway)

halfnorm(e, labs=id, ylab="Leverages")

standard_res <- rstandard(forward_red)

plot(forward_red$fitted.values,standard_res,main="studentdized residuals", xlab="Fitted values",ylab="Standardized Residual", sub="raylor")

abline(h=c (-2,0,2), col=c (2,1,2),lty=c (1,2,1))

identify(forward_red$fitted.values, rstandard(forward_red), labels=row.names(CDI))

(1: length(rstandard(forward_red)))[rstandard(forward_red)>2]

(1: length(rstandard(forward_red)))[rstandard(forward_red)< -2]