Regression Analysis of US County Premature Mortality Rate

Roy Subhadeep
Bothara Shubham
Illinois Institute of Technology

Abstract

The County Health Rankings measure the health of all US counties and rank them within states. These ranks are estimated using composite scores from variety of health measures like quality of life, socio-economic factors, clinical care etc. The major goal of the rankings is to raise awareness about many factors that influence health and subsequently take corrective actions. A good indicator of the mortality measure for each county is YPLL (years of potential life lost) which indicates the cumulative years lost (typically calculated using predefined standards) due to premature death. Contribution to YPLL is the difference of premature deaths that are below 75 years of age as defined by the CDC. For example, if a person dies at the age of 72, then there is a contribution of 3 years towards the county YPLL. Any age of death above 75 is not counted towards YPLL as the criterion of expected age of 75 has supposedly been met. Can we identify years of preventable life loss?

Introduction

YPLL data for each US county is available under general public license but without any supporting data. But, YPLL can be regressed if combined with external demographic data which well may act as a supporting evidence to a preventable action.

The YPLL for each 100,000 people, averaged across counties in the United States is between 8000 and 9000 depending on the year. The data file ypll.csv contains per-county YPLL's for the United States in 2011. The attached file additional_measures.csv contains demographic measures for each US counties.

As a part of this project, we analyzed which of the additional measures correlate strongly with our mortality measure, and fit/analyze regression models for YPLL.

Development Environment

The regression analysis was done in R 3.4.1 with rms, ggplot as helper libraries. The iPython notebook for the complete implementations can be found in the below link.
https://github.com/shubhambothara/AppliedStats/YPLL.ipynb

Data Sources

Data for this project is available on:
http://www.countyhealthrankings.org/ranking-methods/exploring-data
http://opengovdata.pbworks.com/w/page/27141180/County%20Health%20Rankings.

Data Cleaning and Preprocessing

Below are the data from two sources.

Measures Data:

```
head(measures)
```

FIPS	State	County	Population	X..18	X65.and.over	African.American	Female	Rural	X.Diabetes	HIV.rate	Physical.Inactivity	mental.health.provider.rate	m
1000	Alabama		4708708	23.9	13.8	26.1	51.6	44.6	12	NA	31	20	
1001	Alabama	Autauga	50756	27.8	11.6	18.4	51.4	44.8	11	170	33	2	
1003	Alabama	Baldwin	179878	23.1	17	10	51	54.2	10	176	25	17	
1005	Alabama	Barbour	29737	22.3	13.8	46.6	46.8	71.5	14	331	35	7	
1007	Alabama	Bibb	21587	23.3	13.5	22.3	48	81.5	11	90	37	0	
1009	Alabama	Blount	58345	24.2	14.7	2.1	50.2	91	11	66	35	2	

YPLL Data:

```
head(ypll)
```

FIPS	State	County	Unreliable	YPLL.Rate
1000	Alabama			10189
1001	Alabama	Autauga		9967
1003	Alabama	Baldwin		8322
1005	Alabama	Barbour		9559
1007	Alabama	Bibb		13283
1009	Alabama	Blount		8475

To bring the data in a single data frame for analysis, the two sources of data (County YPLL and demography measures) were merged based on the unique identifier FIPS. The below transformations were made to the data as a part of data cleaning process.

- Records marked as unreliable were omitted so that unreliable data do not affect the model.
- Records depicting the cumulative data for a state were removed as the focus of the project is analysis of county related data.
- Records with missing YPLL value were removed as it is the target variable and training will not be possible.
- Records with missing values for Child Illiteracy, Rural and free lunch variables as they are very few in number (< 1%) and will not affect the model.
- Unique ID values (FIPS) were then removed from the data frame as they would not contribute much information to the model.
- Predictors were then assigned proper data types to build the models.
- After the initial cleaning, there still existed 22% missing values in the independent variable HIV Rate. As HIV Rate is an important predictor, it couldn't be ignored. HIV Rate was

bucketed into 4 quartile levels – Very Low, Low, High, Very High.

```
> range(HIVRate)
[1]   13 4534
> quantile(HIVRate)
  0%  25%  50%  75% 100%
  13   57   98  194 4534
>
```

Dummy Variables were created for these HIV factors. An extra dummy variable was created as an indicator of missing HIV values.

dfHigh	dfLow	dfNotAvail	dfVeryHigh	dfVeryLow
1	0	0	0	0
1	0	0	0	0
0	0	0	1	0
0	1	0	0	0
0	1	0	0	0
0	0	0	1	0
0	0	0	1	0
1	0	0	0	0

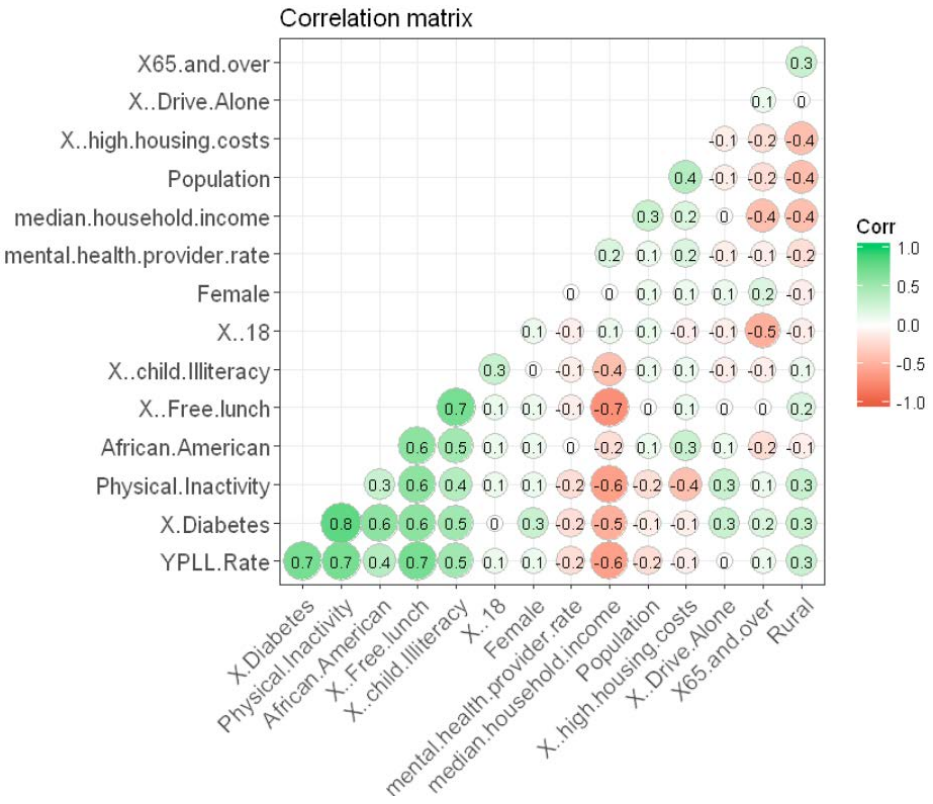
The below is the snapshot of the final merged and cleaned data frame.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T
1	Populatio	X..18	X65.and.o	African.Ar	Female	Rural	X.Diabete	Physical.Ir	mental.he	median.hc	X..high.ho	X..Free.lur	X..child.Illi	X..Drive.A	dfHigh	dfLow	dfNotAvai	dfVeryHig	dfVeryLov	YPLL..Rate
2	50756	27.8	11.6	18.4	51.4	44.8	11	33	2	51622	25	29	12.7	86	1	0	0	0	0	9967
3	179878	23.1	17	10	51	54.2	10	25	17	51957	29	29	10.6	83	1	0	0	0	0	8322
4	29737	22.3	13.8	46.6	46.8	71.5	14	35	7	30896	36	65	23.2	82	0	0	0	1	0	9559
5	21587	23.3	13.5	22.3	48	81.5	11	37	0	41076	18	48	17.5	83	0	1	0	0	0	13283
6	58345	24.2	14.7	2.1	50.2	91	11	35	2	46086	21	37	13.9	80	0	1	0	0	0	8475
7	10985	24.6	10.8	68.2	44.5	64.7	15	32	0	26980	36	92	34.2	83	0	0	0	1	0	15433
8	19964	24.8	16.2	41.9	52.9	74.8	15	35	5	31449	31	64	20.6	87	0	0	0	1	0	12652
9	114081	23.4	15	20	52.1	31	15	32	11	39997	28	45	14.6	86	1	0	0	0	0	11720
10	34320	22.5	16.8	37.8	52.6	49.8	16	36	3	35614	29	60	20.2	85	0	0	0	1	0	11319
11	24448	20.9	18.6	5.6	51.4	100	14	37	0	38028	24	43	16.4	77	0	1	0	0	0	11753
12	42971	24.5	13.4	11.4	50.3	88	10	32	2	40292	30	42	15.5	79	0	1	0	0	0	12243
13	13990	23.3	17.5	43.6	53.1	100	15	34	0	30728	21	66	22.5	79	0	0	0	1	0	10116
14	26042	25.4	15.9	43.6	52.3	74.6	14	36	4	34101	27	56	20.8	89	0	1	0	0	0	9455
15	13640	21.3	18.3	15.8	50.6	100	13	36	7	33032	24	47	18.2	77	0	1	0	0	0	9201
16	14759	23	15.2	4.5	49.6	100	11	32	0	37742	25	45	16.7	79	1	0	0	0	0	9305
17	48635	24	14.3	17.9	51.4	55.9	12	30	2	43145	25	36	13.4	84	1	0	0	0	0	9527
18	54639	22	16.8	16.6	52	46.9	14	32	9	39947	26	40	14.6	86	0	1	0	0	0	10660
19	12931	23.1	17.8	43.6	52.8	100	15	36	8	27068	26	83	22.8	86	0	0	0	1	0	12759
20	10556	19.9	17.4	30.4	50.2	97.4	13	35	0	36050	26	60	21	79	1	0	0	0	0	12965
21	36678	22.6	18.8	13.1	52.1	72	13	37	0	33773	27	45	15.5	81	1	0	0	0	0	11183
22	13781	24.2	16.2	25.2	52.3	100	13	31	7	34402	24	46	18.8	80	1	0	0	0	0	10062
23	81778	23.3	15.6	1.7	50.4	75.7	12	32	5	39276	27	37	13.4	79	0	0	0	0	1	9565
24	48147	24.9	13	20.6	51.3	55.5	14	29	6	42867	22	49	12.9	80	0	0	0	1	0	9477
25	41925	27	14.8	67.8	54.3	46.6	16	35	12	27992	39	76	23.9	83	0	0	0	1	0	14265

Data Analysis

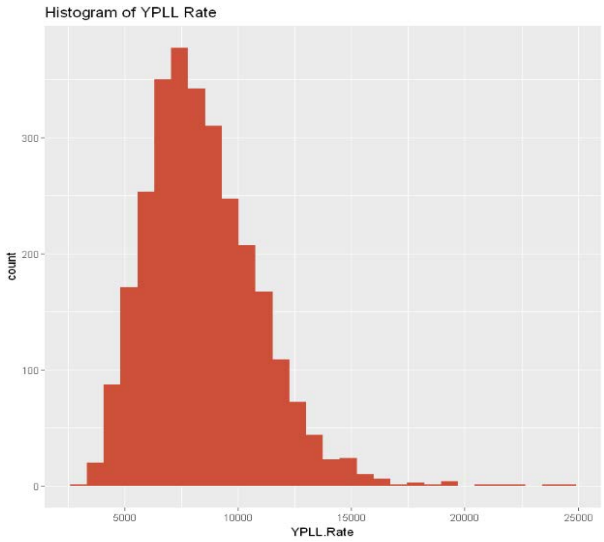
Multicollinearity & distribution:

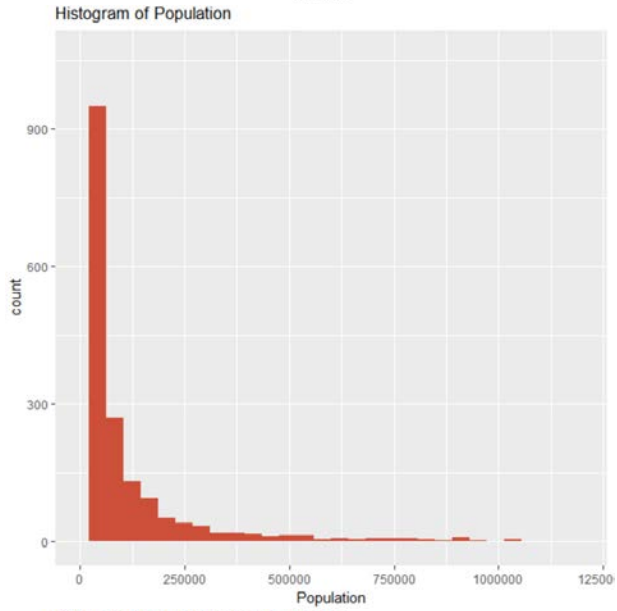
A correlation matrix was plotted between continuous variables to understand collinearity between independent variables. The highest correlation was found between %Diabetes and Physical Inactivity which seems obvious. There were no any issues with collinearity.

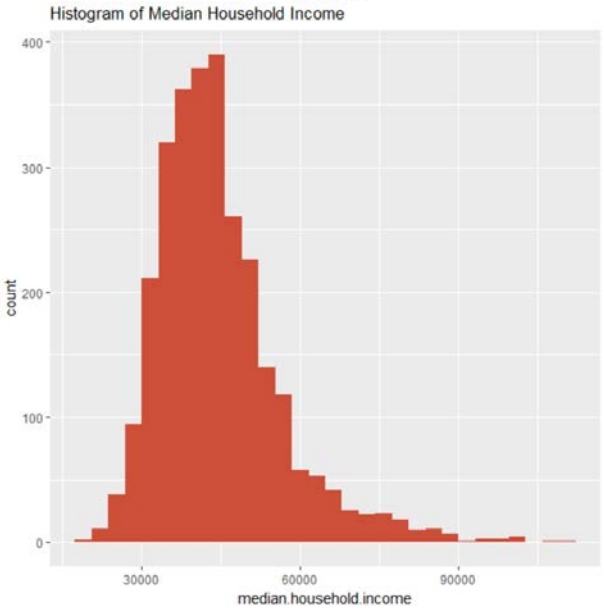


Below is the distribution of continuous variables with high range values (YPLL Rate, population, median household distribution. The response variable is slightly left skewed which indicates it will have potential effect of scaling. The population is highly skewed to the left.

```
'data.frame':   2834 obs. of  16 variables:
 $ YPLL.Rate                 : int  9967 8322 9559 13283 8475 15433 12652 11720 11319 11753 ...
 $ Population                : int  50756 179878 29737 21587 58345 10985 19964 114081 34320 24448 ...
 $ X..18                     : num  27.8 23.1 22.3 23.3 24.2 24.6 24.8 23.4 22.5 20.9 ...
 $ X65.and.over              : num  11.6 17 13.8 13.5 14.7 10.8 16.2 15 16.8 18.6 ...
 $ African.American          : num  18.4 10 46.6 22.3 2.1 68.2 41.9 20 37.8 5.6 ...
 $ Female                    : num  51.4 51 46.8 48 50.2 44.5 52.9 52.1 52.6 51.4 ...
 $ Rural                     : num  44.8 54.2 71.5 81.5 91 64.7 74.8 31 49.8 100 ...
 $ X.Diabetes                : int  11 10 14 11 11 15 15 15 16 14 ...
 $ Physical.Inactivity       : int  33 25 35 37 35 32 35 32 36 37 ...
 $ mental.health.provider.rate: int  2 17 7 0 2 0 5 11 3 0 ...
 $ median.household.income   : int  51622 51957 30896 41076 46086 26980 31449 39997 35614 38028 ...
 $ X..high.housing.costs     : int  25 29 36 18 21 36 31 28 29 24 ...
 $ X..Free.lunch             : int  29 29 65 48 37 92 64 45 60 43 ...
 $ X..child.Illiteracy       : num  12.7 10.6 23.2 17.5 13.9 34.2 20.6 14.6 20.2 16.4 ...
 $ X..Drive.Alone            : int  86 83 82 83 80 83 87 86 85 77 ...
 $ HIV                       : Factor w/ 5 levels "High","Low","NotAvail",..: 1 1 4 2 2 4 4 1 4 2 ...
```


Histogram of YPLL Rate
count
0
100
200
300
5000
10000
15000
20000
25000
YPLL.Rate

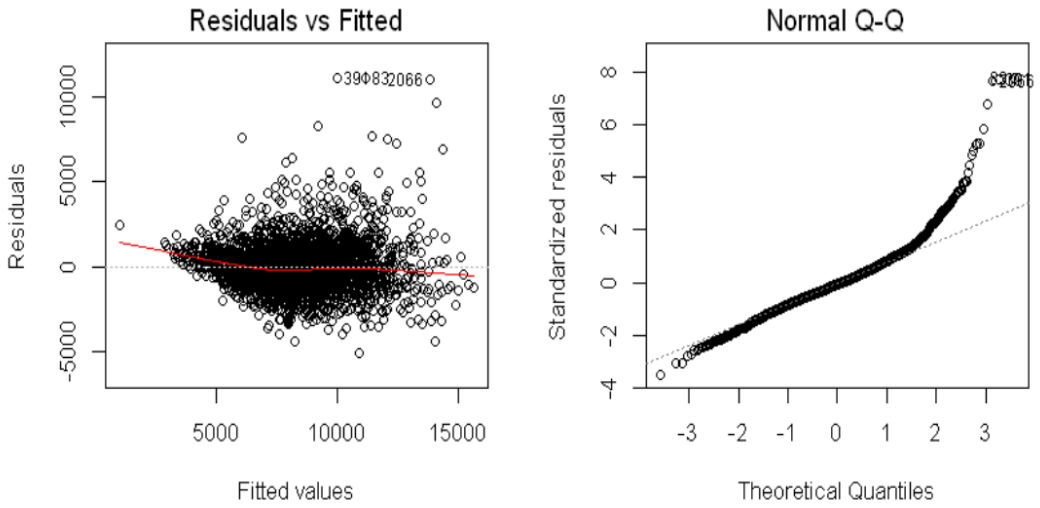

Histogram of Population
count
0
300
600
900
0
250000
500000
750000
1000000
12500
Population


Histogram of Median Household Income
count
0
100
200
300
400
30000
60000
90000
median.household.income

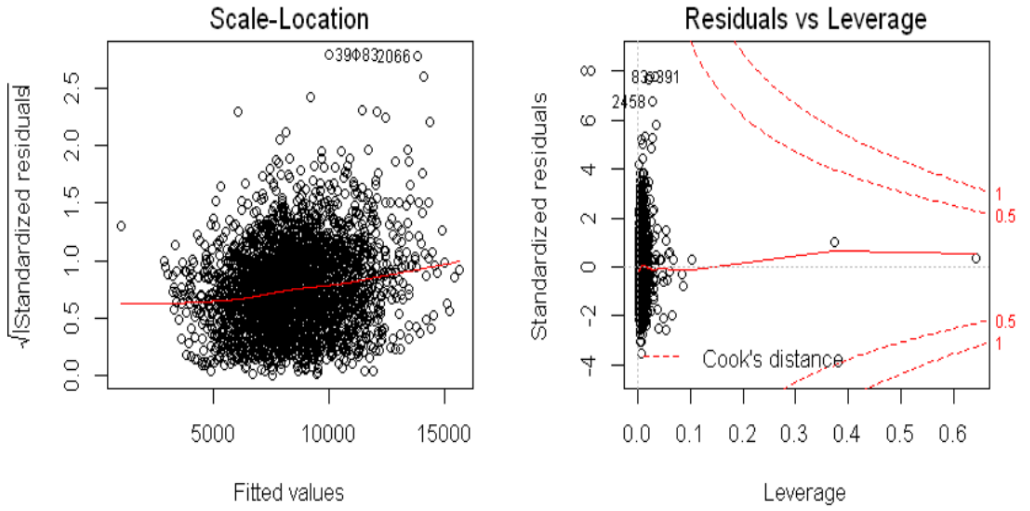
Regression Model

A basic regression model was fitted to understand the initial results and data distribution. Below are the residual plots.

```
Residual standard error: 1443 on 2815 degrees of freedom
Multiple R-squared:  0.9734,    Adjusted R-squared:  0.9732
F-statistic:  5414 on 19 and 2815 DF,  p-value: < 0.00000000000000022
```



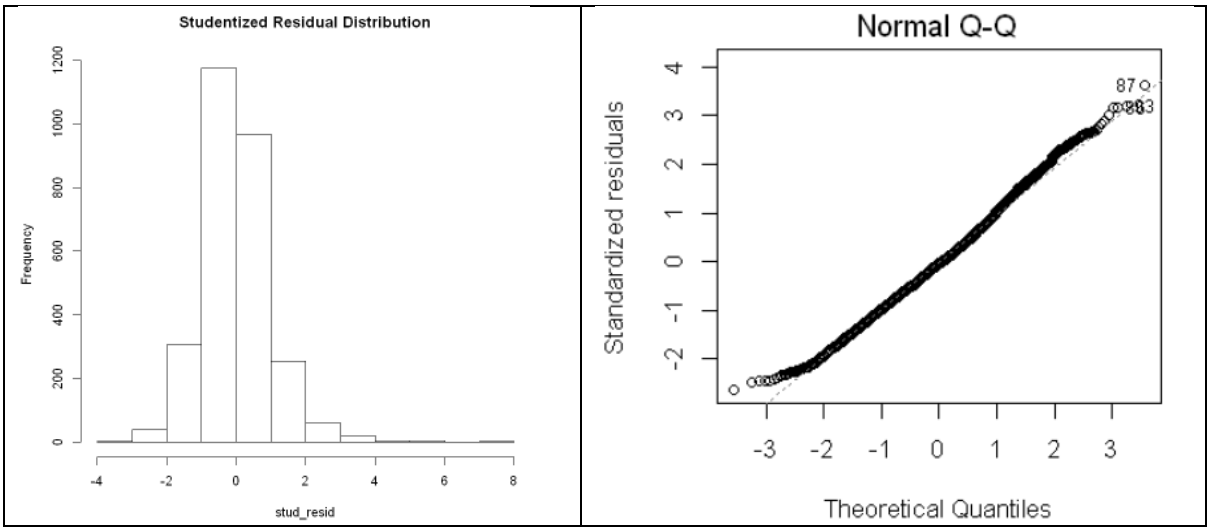
The residual plot vs fitted values has a near horizontal distribution indicating some amount constant variance. The normal plot does not look like a straight line indicating that the normality assumption has been violated.



The above scale location plot indicates the spread is almost equal along the range of predictors. The residuals vs leverage plot indicates couple of isolated points away from the main cluster but will not make much difference to the regression model as they are not way beyond cook's distance cutoff.

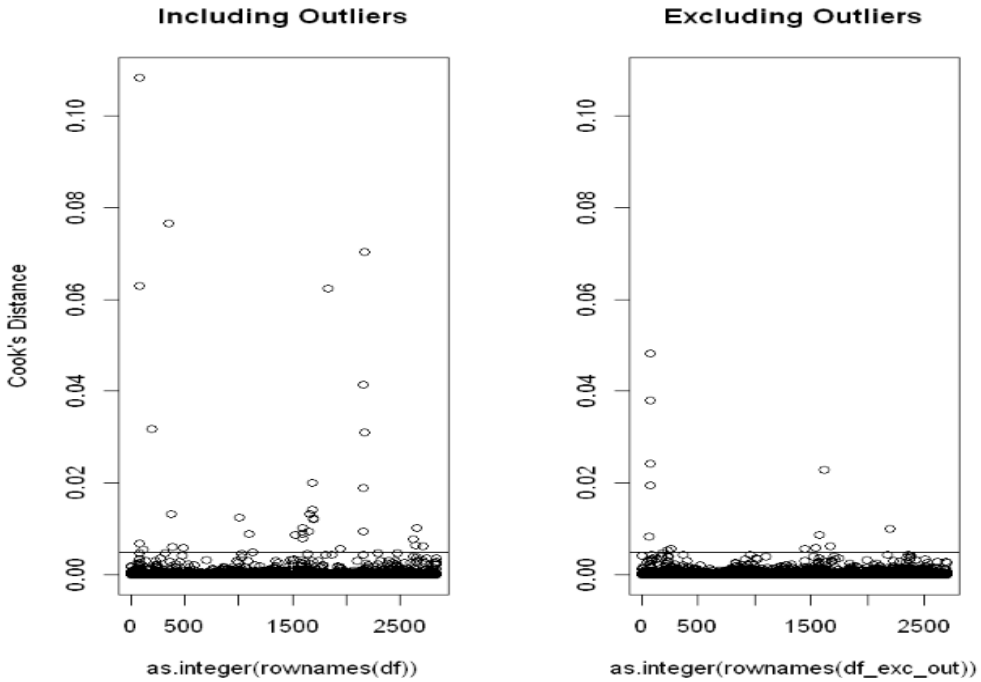
Removing outliers using studentized residual:

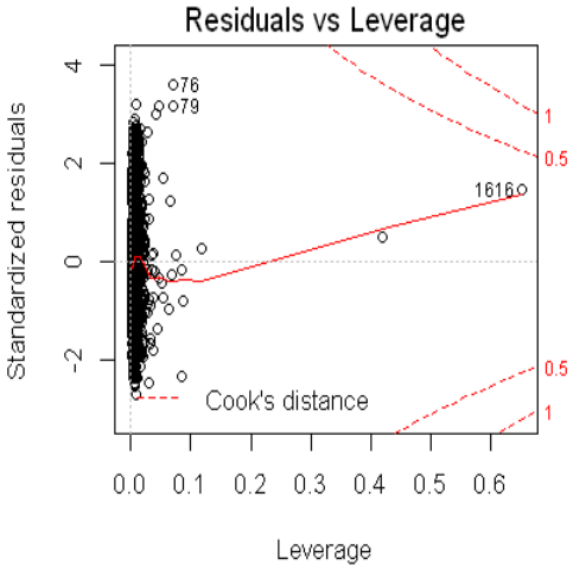
The studentized residual distribution is given below indicating few outliers. The model is rerun by excluding the outliers whose absolute value of studentized residual is greater than 2. We notice the normal plot of residuals has a reduced curvature than the base model.



Removing influential observations using Cook's distance:

Influential observations are identified below using Cook's distance. The Cook's distance is identified by a cutoff formula (4/n-k-1) where k is the number of independent variables. We notice in the leverage plot after removing influential observations is that the dotted lines are now more cornered towards the right side indicating the few isolated points are well far off from the cutoff.

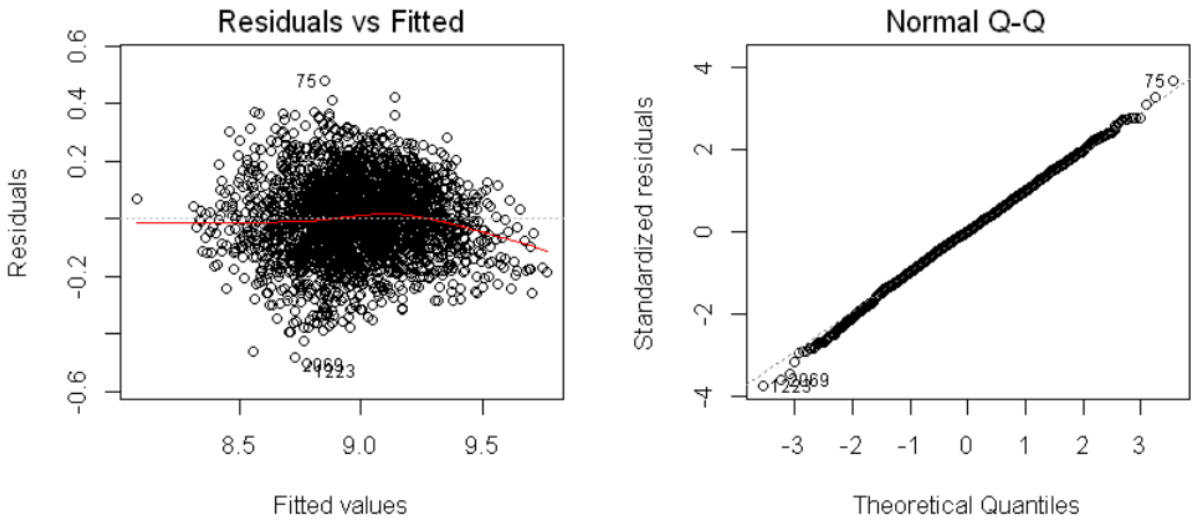




Reducing Heteroskedasticity (scaling):

Some of the predictor variables (Population and median.household.income) and response variable (YPLL.Rate) was log scaled as their range differed a lot from other variables causing heteroskedasticity. The model was rerun using individual scale of response/predictor variable as well as the combination of both.

The scaling of just the response variable achieved the best output considering all model adequacies. The normal residual plot was now a perfect straight line. The final reference model in this case is now log scaled to y with all outliers and leverage values removed.



Finalizing regression model (stepwise):

Since we now have a good model for regressing, it is imperative to check if the same model performance can be achieved with less number of variables. Bidirectional stepwise regression was performed using minimum AIC method and important features were selected. It is noticed that the same performance can be achieved using 4 lesser number of features with an adjusted R squared of 0.9998.

Reference Model:

```
YPLL.Rate ~ Population + X..18 + X65.and.over + African.American +
    Female + Rural + X.Diabetes + Physical.Inactivity + mental.health.provider.rate +
    median.household.income + X..high.housing.costs + X..Free.lunch +
    X..child.Illiteracy + X..Drive.Alone + NAHigh + NALow + NANotAvail +
    NAVeryHigh + NAVeryLow - 1
<environment: 0x000000001afccff8>


Coefficients:
                              Estimate    Std. Error t value             Pr(>|t|)
Population                -0.00021906    0.00009676  -2.264               0.0237 *
X..18                     87.86276811   11.97182726   7.339  0.00000000000027989 ***
X65.and.over             -20.18209330   10.94826598  -1.843               0.0654 .
African.American          -2.03624212    3.45369035  -0.590               0.5555
Female                    -5.59661428   15.40356529  -0.363               0.7164
Rural                      5.39630494    1.21619807   4.437  0.00000947121182251 ***
X.Diabetes               307.54548665   26.33322345  11.679 < 0.0000000000000002 ***
Physical.Inactivity       83.59794555    9.53463728   8.768 < 0.0000000000000002 ***
mental.health.provider.rate -0.99053873  0.44923676  -2.205               0.0275 *
median.household.income   -0.04992777    0.00426563 -11.705 < 0.0000000000000002 ***
X..high.housing.costs      3.55465583    5.78602292   0.614               0.5390
X..Free.lunch             46.95881427    3.30803546  14.195 < 0.0000000000000002 ***
X..child.Illiteracy      -55.25836847    7.01888792  -7.873  0.00000000000000491 ***
X..Drive.Alone           -38.14721949    4.74009814  -8.048  0.00000000000000123 ***
dfHigh                  5641.13331657  861.62189238   6.547  0.00000000006946332 ***
dfLow                   5449.67325888  862.45739885   6.319  0.00000000030546362 ***
dfNotAvail              5060.68439483  850.52594974   5.950  0.00000000301174339 ***
dfVeryHigh              5919.26078870  853.81754207   6.933  0.00000000000509994 ***
dfVeryLow               5040.67560934  861.59357942   5.850  0.00000000546931041 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1443 on 2815 degrees of freedom
Multiple R-squared:  0.9734,    Adjusted R-squared:  0.9732
F-statistic:  5414 on 19 and 2815 DF,  p-value: < 0.00000000000000022
```

Stepwise Model:

```
Call:
lm(formula = YPLL.Rate ~ X..18 + African.American + Rural + X.Diabetes +
    Physical.Inactivity + mental.health.provider.rate + median.household.income +
    X..Free.lunch + X..child.Illiteracy + X..Drive.Alone + NAHigh +
    NALow + NANotAvail + NAVeryHigh + NAVeryLow - 1, data = df_exc_scaley)


Coefficients:
                              Estimate    Std. Error t value             Pr(>|t|)
X..18                     0.0073447309  0.0009729181   7.549   0.0000000000000604 ***
African.American         -0.0005568442  0.0003286073  -1.695             0.090280 .
Rural                     0.0005710608  0.0001119566   5.101   0.0000003629566670 ***
X.Diabetes                0.0301097917  0.0024389582  12.345 < 0.0000000000000002 ***
Physical.Inactivity       0.0105785387  0.0008826982  11.984 < 0.0000000000000002 ***
mental.health.provider.rate -0.0001973440 0.0000418156 -4.719   0.0000024920530126 ***
median.household.income  -0.0000066828  0.0000003978 -16.801 < 0.0000000000000002 ***
X..Free.lunch             0.0049587220  0.0003231683  15.344 < 0.0000000000000002 ***
X..child.Illiteracy      -0.0043818542  0.0006758664  -6.483   0.0000000001071790 ***
X..Drive.Alone           -0.0017238583  0.0005053146  -3.411             0.000656 ***
dfHigh                    8.5442595168  0.0514377003 166.109 < 0.0000000000000002 ***
dfLow                     8.5222753937  0.0512637717 166.244 < 0.0000000000000002 ***
dfNotAvail                8.4478371931  0.0494648512 170.785 < 0.0000000000000002 ***
dfVeryHigh                8.5737956915  0.0506956560 169.123 < 0.0000000000000002 ***
dfVeryLow                 8.4669611299  0.0509501941 166.181 < 0.0000000000000002 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1343 on 2575 degrees of freedom
Multiple R-squared:  0.9998,    Adjusted R-squared:  0.9998
F-statistic: 7.736e+05 on 15 and 2575 DF,  p-value: < 0.00000000000000022
```

Conclusion

To statistically verify the model performance, Chi-squared test is performed which compares the reduction in the residual sum of squares between two models.

```
anova(lm.5,lm.7, test="Chisq")
```

Res.Df	RSS	Df	Sum of Sq	Pr(>Chi)
2571	46.36974	NA	NA	NA
2575	46.44811	-4	-0.07836842	0.3613009

From the output, it is noticed that the RSS for model 5 (reference model) is slightly better than model 7 (stepwise) but this minute difference seems insignificant. It is much better to have smaller model parameters than factoring in the little RSS difference between the two models.

Appendix

Summary Statistics:

```
> summary(df)
   YPLL.Rate       Population          X..18         X65.and.over   African.American      Female          Rural          X.Diabetes
 Min.   : 3275   Min.   :   1485   Min.   :11.30   Min.   : 2.60   Min.   : 0.000   Min.   :25.10   Min.   :  0.00   Min.   : 3.00
 1st Qu.: 6696   1st Qu.:  14371   1st Qu.:21.60   1st Qu.:12.70   1st Qu.: 0.900   1st Qu.:49.80   1st Qu.: 33.50   1st Qu.: 9.00
 Median : 8146   Median :  29840   Median :23.50   Median :15.00   Median : 2.900   Median :50.60   Median : 56.85   Median :10.00
 Mean   : 8461   Mean   : 107552   Mean   :23.64   Mean   :15.18   Mean   : 9.867   Mean   :50.23   Mean   : 56.09   Mean   :10.01
 3rd Qu.: 9934   3rd Qu.:  75279   3rd Qu.:25.30   3rd Qu.:17.30   3rd Qu.:11.800   3rd Qu.:51.30   3rd Qu.: 78.97   3rd Qu.:11.00
 Max.   :24829   Max.   :9848011   Max.   :40.50   Max.   :34.30   Max.   :86.300   Max.   :58.00   Max.   :100.00   Max.   :18.00
 Physical.Inactivity mental.health.provider.rate median.household.income X..high.housing.costs X..Free.lunch    X..child.Illiteracy
 Min.   :10.00       Min.   :   0.00             Min.   : 19829          Min.   :11.00         Min.   :  0.00   Min.   : 3.90
 1st Qu.:24.00       1st Qu.:   0.00             1st Qu.: 36635          1st Qu.:25.00         1st Qu.: 26.00   1st Qu.: 8.50
 Median :27.00       Median :   7.00             Median : 42714          Median :29.00         Median : 37.00   Median :11.90
 Mean   :27.11       Mean   :  19.18             Mean   : 44542          Mean   :29.31         Mean   : 38.44   Mean   :13.07
 3rd Qu.:30.00       3rd Qu.:  22.00             3rd Qu.: 49749          3rd Qu.:33.00         3rd Qu.: 48.00   3rd Qu.:16.10
 Max.   :45.00       Max.   :2618.00             Max.   :111582          Max.   :56.00         Max.   :100.00   Max.   :64.60
 X..Drive.Alone      dfHigh           dfLow          dfNotAvail      dfVeryHigh       dfVeryLow
 Min.   : 7.00   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000
 1st Qu.:76.00   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000
 Median :79.00   Median :0.0000   Median :0.0000   Median :0.0000   Median :0.0000   Median :0.0000
 Mean   :78.32   Mean   :0.1941   Mean   :0.1958   Mean   :0.2205   Mean   :0.1944   Mean   :0.1951
 3rd Qu.:82.00   3rd Qu.:0.0000   3rd Qu.:0.0000   3rd Qu.:0.0000   3rd Qu.:0.0000   3rd Qu.:0.0000
 Max.   :92.00   Max.   :1.0000   Max.   :1.0000   Max.   :1.0000   Max.   :1.0000   Max.   :1.0000
```

List of attributes:

FIPS, State, County, Population, < 18,65 and over, African American, Female, Rural, %Diabetes, HIV rate, Physical Inactivity, mental health provider rate, median household income, % high housing costs, % Free lunch, % child Illiteracy, % Drive Alone, YPLL rate