DA 546 PROJECT: Analysis/Report, Jan-Apr, 2022

Title: OLS Regression modelling of cancer mortality rates

Name: Buddi Kiran Chaitanya

Roll No: 214161002

Statement/Objective:

To explore a linear model built upon county level socio-economic & demographic statistics that can possibly help us to zero in on the potential social determinants of the cancer mortality rates.

Incidence might be uniform or random, the mortality that ensues from it is a form of social disability, in the sense that groups/individuals who are socially & economically disadvantaged and disenfranchised face a greater risk of transitioning from incidence to mortality.

While the correlation between socio-economic indicators of disadvantage and the cancer mortality rates is a well established consensus, the sticking point is in the form of the existence of myriad forms of social disadvantages, which often interact in a non-linear fashion.
We wish to identify those disadvantages that have high predictive ability in relation to mortality compared to others, and what sort of impact the intersection of disabilities has on the mortality rates. Such knowledge has implications for public health policy, distribution/re-distribution of public resources and can also be of aid to those advocating for better health outcomes in low income & socially disadvantaged groups/populations.

Description of the data:

The dataset that is being utilised in this project has been aggregated, by Noah Rippner, using table joins, from data made available openly by several State institutions such as the American Community Survey (census.gov), clinicaltrials.gov, National Institutes of Health (NIH), and cancer.gov. The data corresponds to several socio-economic variables aggregated at the level of counties of the USA.

The dataset has 3047 entries and 34 columns. Each one of the 3047 samples/rows correspond to data of one of the counties, whose name and home state details are provided by the '**Geography**' attribute, a categorical variable. Thus, **'Geography'** serves as the unique identifier/key for an observation/row.

Apart from Geography, '**binnedInc**' is the only other categorical variable, which assigns all the counties to decile bins based upon the median county income. The remaining 32 variables are of the numeric-continuous variety, with the variable/attribute '**TARGET_deathRate**', the county level lung cancer mortality rate, being the one that we seek to model/predict using some combination of the rest of the variables.

The attributes available in the dataset are detailed below:

Variable label (data type)	Sample data	Unique %	NA%	description
Geography (string)	Kitsap County, Washington Madison County, Tennessee Blaine County, Idaho	100	0	County name & State in which the county is situated.

Target								
Variable label (data type)	NA %	Quantiles & spread						
		Min	1st Q	Median	Mean	3rd Q	Max	Std. dev
'**TARGET_deathRate**' (integer)	mean per capita (100,000) **lung cancer** mortalities in a given county over the period of 2010-2016							
	0	59.7	161.2	178.1	178.7	195.2	362.8	27.75

Five variables can be grouped into the **medical health/research statistics** category, these are mean annual county level figures for over the period 2010-16, their description is as follows:

Medical Health/Research Statistics								
Variable label (data type)	NA %	Quantiles & spread						
		Min	1st Q	Median	Mean	3rd Q	Max	Std. dev
avgAnnCount (integer)	Mean number of reported cases of cancer diagnosed annually							
	0	6	76	171	606.34	518	38150	1416.36
incidenceRate (decimal)	Mean per-capita (100,000) cancer diagnoses							
	0	201.3	420.3	453.5	448.3	480.9	1206.9	54.56
avgDeathsPerYear (integer)	Mean number of reported mortalities due to lung cancer							
	0	3	28	61	186	149	14010	504.13
TARGET_deathRate (decimal)	The **targeted Dependent variable**. Mean per capita (100,000) lung cancer mortalities							
	0	59.7	161.2	178..1	178.7	195.2	362.8	27.75
studyPerCap (decimal)	Per-capita number of cancer-related clinical trials per county							
	0	0	0	0	155.4	83.65	9762.3	529.6

Five variables deal with aspects relating to **population statistics**, sourced from the 2013 census estimates. Their description follows:

Population Statistics								
Variable label (data type)	NA %	Quantiles & spread						
		Min	1st Q	Median	Mean	3rd Q	Max	Std. dev
popEst2015 (integer)	Population of the concerning county							
	0	827	11684	26643	102637	68671	10170292	329059
BirthRate (decimal)	Number of live births relative to number of women in the county							
	0	0	4.52	5.38	5.64	6.49	21.33	1.97
MedianAge	Median age of county residents							
	0	22.3	37.7	41	45.3	44	624	45.3
MedianAgeMale	Median age of male county residents.							
	0	22.4	36.4	39.6	39.57	42.5	64.7	5.27
MedianAgeFemale	Median age of female county residents							
	0	22.3	39.1	42.4	42.15	45.3	65.7	5.29

Four of the variables, together, describe the **racial makeup** of the county as per 2013 census estimates:

Racial makeup statistics								
Variable label (data type)	NA %	Quantiles & spread						
		Min	1st Q	Median	Mean	3rd Q	Max	Std. dev
PctWhite (decimal)	percent of county residents who identify as **White**							
	0	10.2	77.3	90.1	83.7	95.5	100	16.4
PctBlack (decimal)	percent of county residents who identify as **Black**							
	0	0	0.62	2.25	9.11	10.51	85.95	14.53
PctAsian (decimal)	percent of county residents who identify as **Asian**							
	0	0	0.25	0.55	1.25	1.22	42.62	2.61
PctOtherRace (decimal)	percent of residents who identify in a category which is not White/ Black/ Asian							
	0	0	0.3	0.83	1.98	2.18	41.93	3.52

Eight of the variables, together, can be categorised as '**educational & employment**' statistics. These details, as per 2013 census estimates, are described below.

Educational & Employment Statistics								
Variable label (data type)	NA %	Quantiles & spread						
		Min	1st Q	Median	Mean	3rd Q	Max	Std. dev
PctNoHS18_24 (decimal)	Percent of county residents ages 18-24 highest education attained: less than high school							
	0	0	12.8	17.1	18.2	22.7	64.1	8.1
PctHS18_24 (decimal)	Percent of county residents ages 18-24 highest education attained: high school diploma							
	0	0	29.2	34.7	35	40.7	72.5	9.07
PctSomeCol18_24 (decimal)	Percent of county residents ages 18-24 highest education attained: some college							
	75	7.1	34	40.4	41	46.4	79	11.12
PctBachDeg18_24 (decimal)	Percent of county residents ages 18-24 highest education attained: bachelor's degree							
	0	0	3.1	5.4	6.16	8.2	51.8	4.53
PctHS25_Over (decimal)	Percent of county residents ages 25 & over highest education attained: high school diploma							
	0	7.5	30.4	35.3	34.8	39.7	54.8	7.03
PctBachDeg25_Over (decimal)	Percent of county residents ages 25 & over highest education attained: bachelor's degree							
	0	2.5	9.4	12.3	13.3	16.1	42.2	5.4
PctEmployed16_Over (decimal)	Percent of county residents ages 16 & over employed							
	5	17.6	48.6	54.5	54.2	60.3	80.1	8.32
PctUnemployed16_Over (decimal)	Percent of county residents ages 16 & over unemployed							
	0	0.4	5.5	7.6	7.85	9.7	29.4	3.45

Three variables, together, are used to gauge the '**social/domestic support makeup**.

Social Support Statistics								
Variable label (data type)	NA %	Quantiles & spread						
		Min	1st Q	Median	Mean	3rd Q	Max	Std. dev
PercentMarried (decimal)	Percent of county residents who are married							
	0	23.1	47.8	52.4	51.8	56.4	72.5	6.9
PctMarriedHouseholds (decimal)	Percent of married households							
	0	23	47.8	51.7	51.2	55.4	78	6.57
AvgHouseholdSize (decimal)	Mean household size of county							
	0	0.022	2.37	2.5	2.48	2.63	3.97	0.43

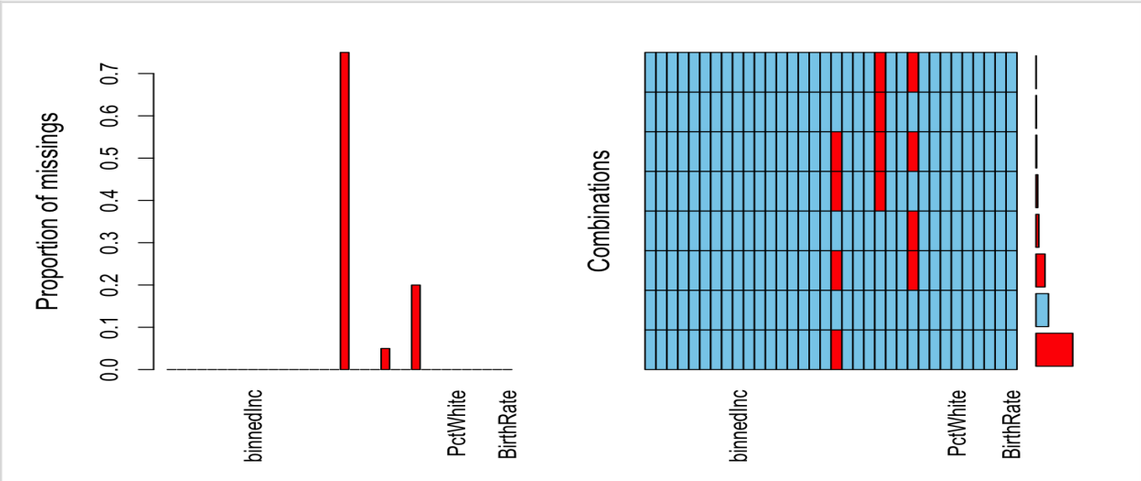
Five variables, together, describe the aggregate '**social security coverage**' status of the county residents, as per 2013 census estimates.

Social Security Coverage								
Variable label (data type)	NA %	Quantiles & spread						
		Min	1st Q	Median	Mean	3rd Q	Max	Std. dev
PctPrivateCoverage (decimal)	Percent of county residents with private health coverage							
	0	22.3	57.2	65.1	64.4	72.1	92.3	10.65
PctPublicCoverage (decimal)	Percent of county residents with private health coverage.							
	0	11.2	30.9	36.3	36.3	41.6	65.1	7.84
PctPrivateCoverageAlone (decimal)	Percent of county residents with private health coverage alone (no public assistance)							
	20	15.7	41	48.7	48.5	55.6	78.9	10.1
PctPublicCoverageAlone (decimal)	Percent of county residents with government-provided health coverage alone							
	0	2.6	14.9	18.8	19.2	23.1	46.6	6.11
PctEmpPrivCoverage (decimal)	Percent of county residents with employee-provided private health coverage							
	0	13.5	34.5	41.1	41.2	47.7	70.7	9.45

The remaining **Three** variables, together, are used to capture the '**economic standing**' of the county residents, as per 2013 census estimates. These include:

Economic Indicators								
Variable label (data type)	NA %	Quantiles & spread						
		Min	1st Q	Median	Mean	3rd Q	Max	Std. dev
medIncome (integer)	Median income per county							
	0	22640	38882	45207	47060	52492	125635	12040
povertyPercent (decimal)	Percent of the populace in poverty							
	0	3.2	12.15	15.9	16.9	20.4	47.4	6.41
binnedInc (string/categorical)	Median income per capita, binned by decile, each decile thus has ~300 counties grouped into one of income bin (bin intervals below)							
'[22640, 34218]' '(34218, 37414]' '(37414, 40363]' '(40363, 42724]' '(42724, 45201]' '(45201, 48022]' '(48022, 51046]' '(51046, 54546]' '(54546, 61495]' '(61495, 125635]'								

Missing Data :



Observations & Manual modifications to the Data:

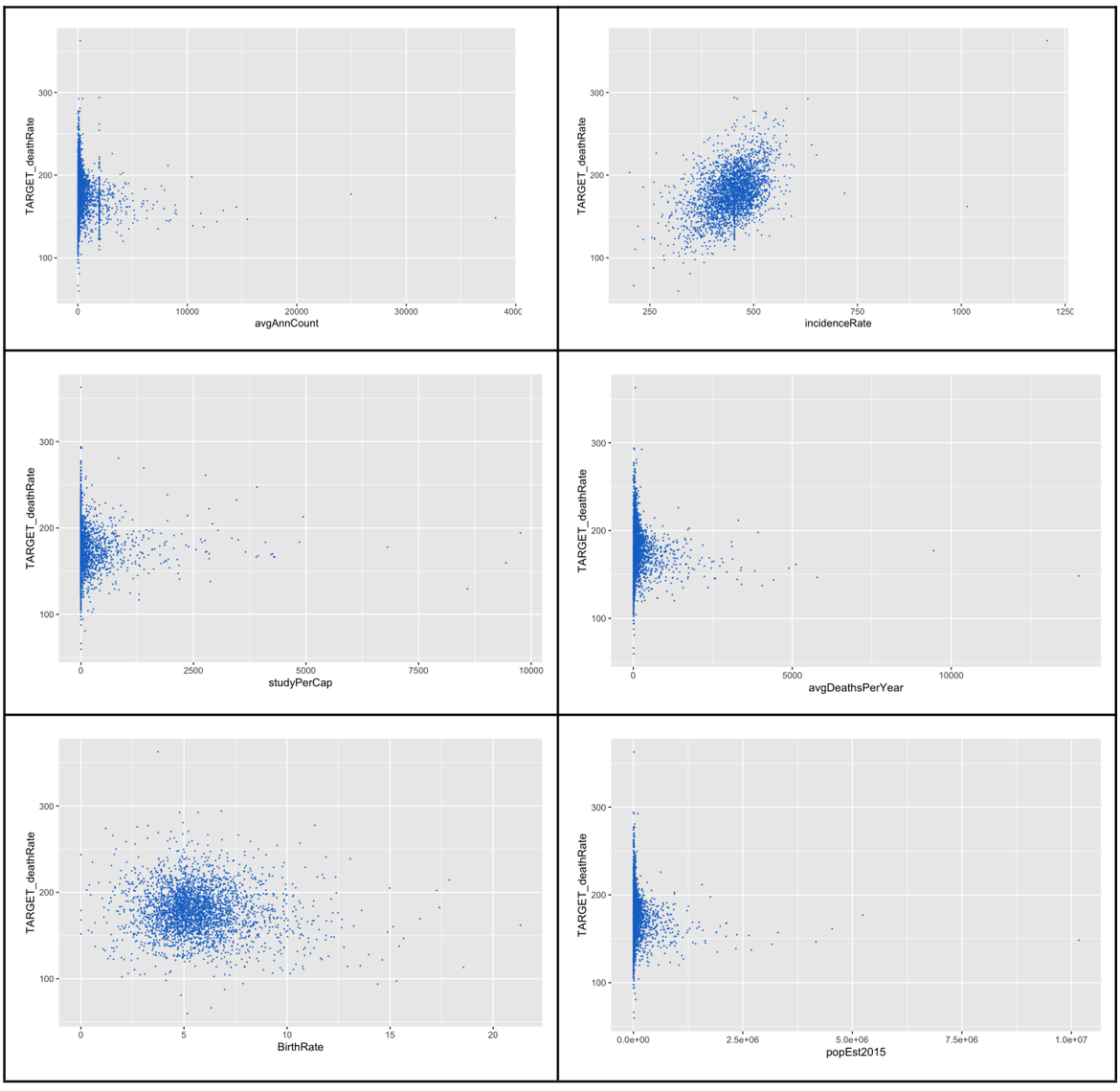
-3 attributes, namely **PctPrivateCoverageAlone, PctEmployed16_Over & PctSomeCol18_24** have missing data for some counties. Those variables are dropped, with the justification being that the relatively high level of correlation observed between these and other 'complete' attributes will be sufficient to capture the variation that these variables might have produced in the output.

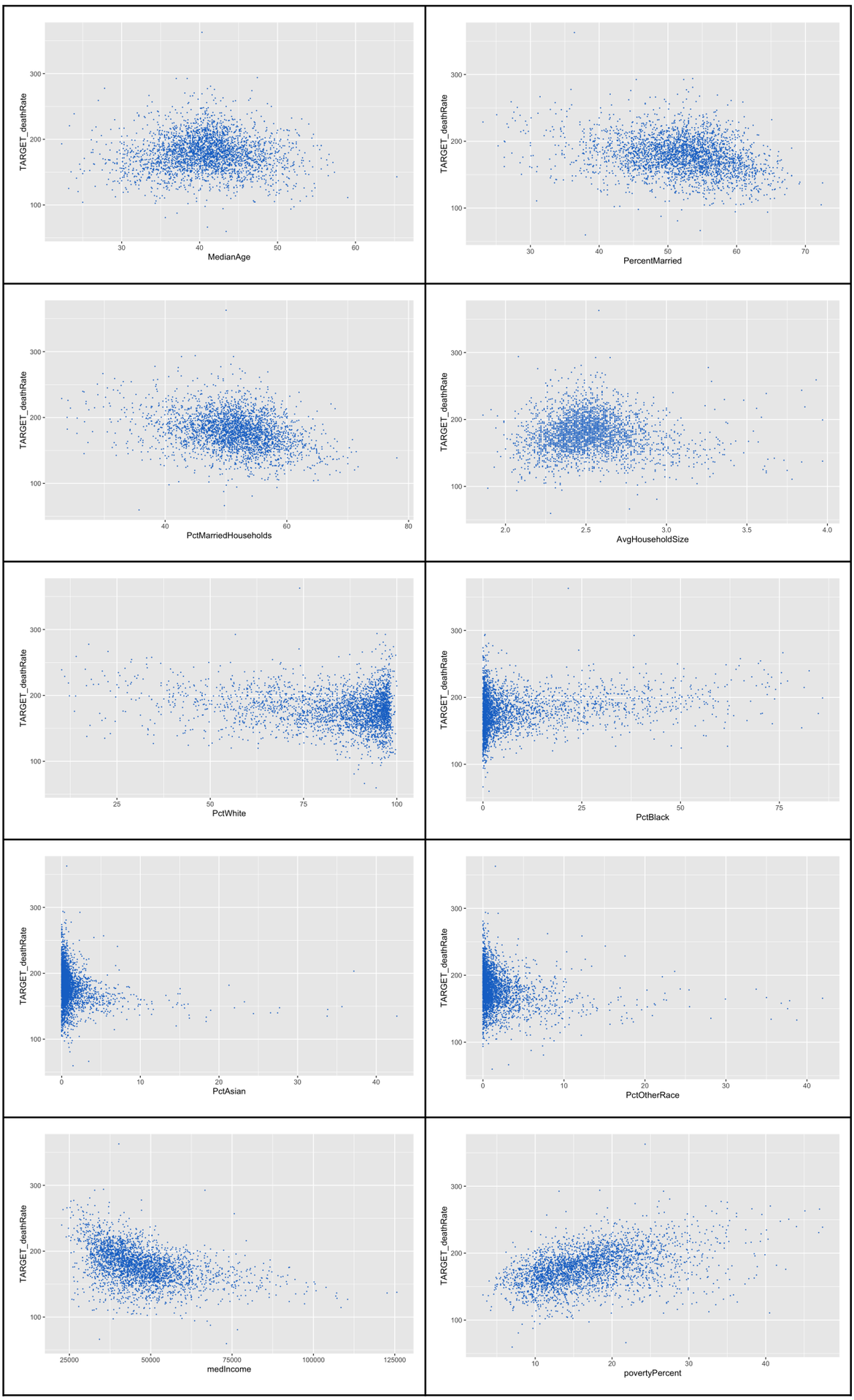
-the '**binnedInc'** attribute has been dropped, as the income data is already captured in the 'Median Income' attribute. While the decile bin is a crucial aspect that has implications for the model, with it being highly likely that a model that holds for one decile bin would not hold for other, this problem is not explored in this project and we only seek a composite model that might generalise over all the income decile bins.

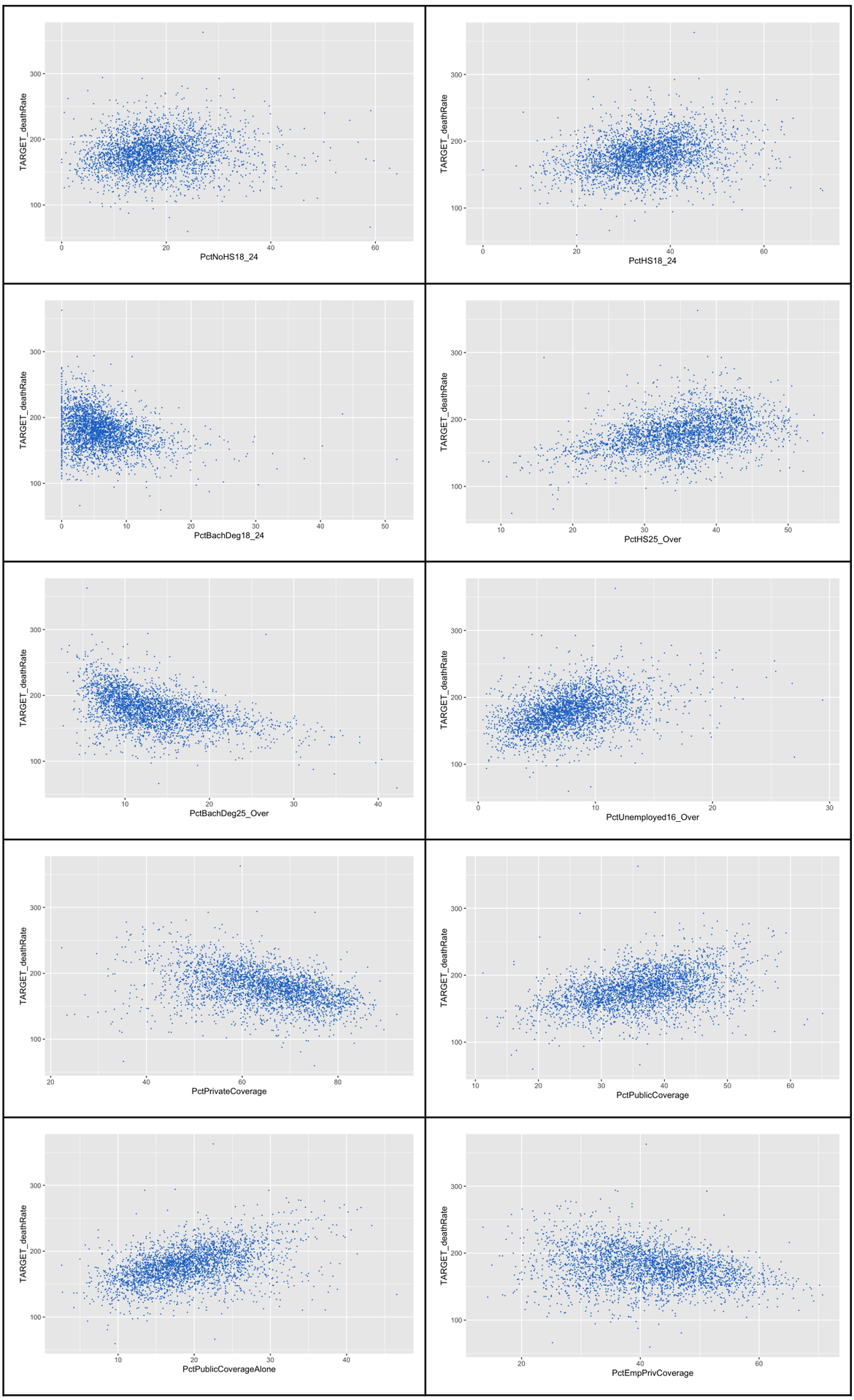
-The variable **'AvgHouseholdSize'** has around 61 entries/observations less than 0.1, which is an impossibility as the household size has to start at 1, suggesting some error in the process of data aggregation. It's likely that these anomalous values might have to be multiplied by 100, and a correction has been applied likewise.

-The **'MedianAge'** attribute has around entries for 30 counties listed with abnormally high (>300) values. These errors have been corrected using the average of **MedianAgeMale** & **MedianAgeFemale** values.

To get a sense of how the variables might be related to the target, the scatterplots of target ('**TARGET_deathRate**') against all other variables are visualised below:



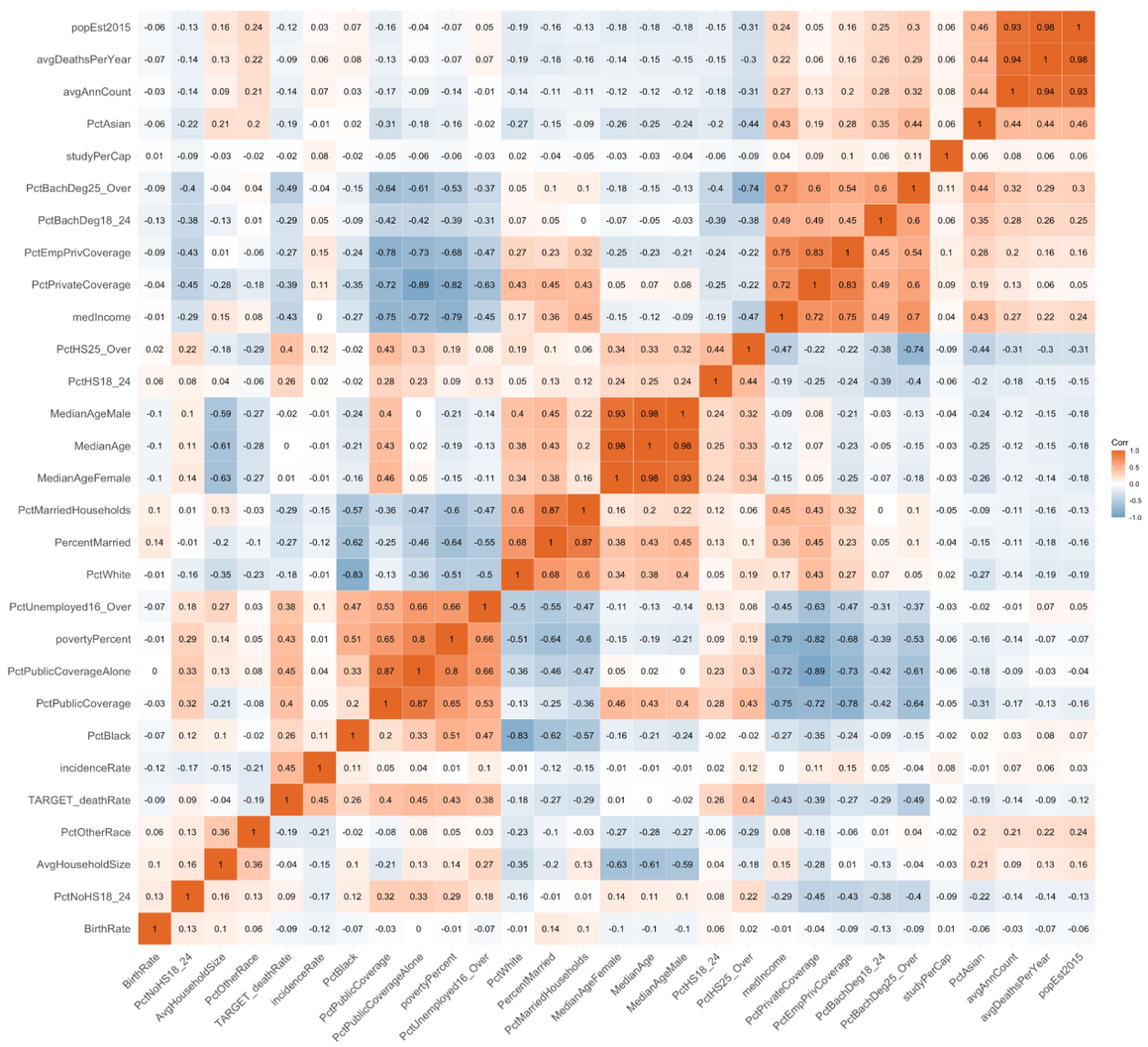

TARGET_deathRate
MedianAge
PercentMarried
PctMarriedHouseholds
AvgHouseholdSize
PctWhite
PctBlack
PctAsian
PctOtherRace
medIncome
povertyPercent


TARGET_deathRate
PctNoHS18_24
PctHS18_24
PctBachDeg18_24
PctHS25_Over
PctBachDeg25_Over
PctUnemployed16_Over
PctPrivateCoverage
PctPublicCoverage
PctPublicCoverageAlone
PctEmpPrivCoverage

-In line with common perceptions, it is visually apparent that the target death rate is positively correlated with the usual suspects such as the incidence rate, poverty percentage, percent unemployed, and percent of public social security coverage, and negatively correlated with factors such as median income & percent population.

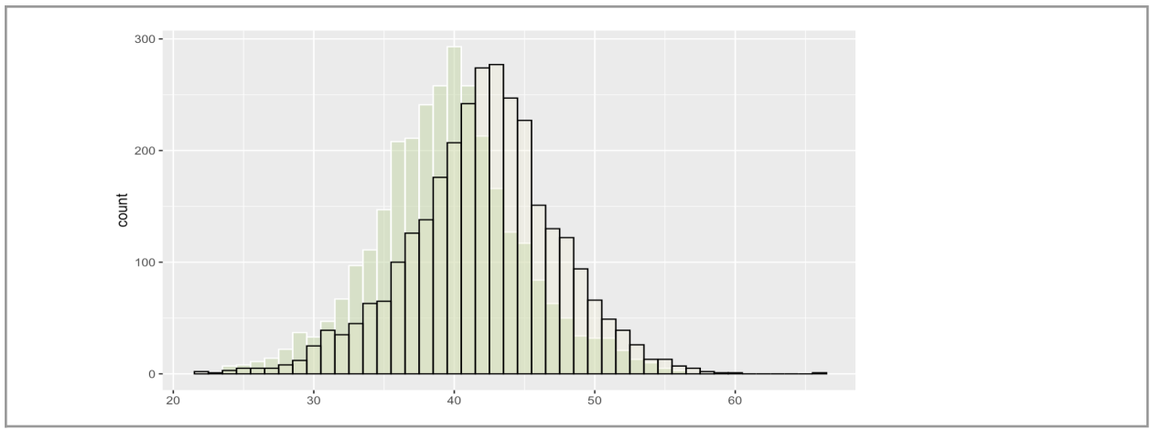
While the bachelor's degree percentage seems to have some negative correlation with death rates, the relation between lower levels of educational attainment, such as high school & college, and mortality rate, is, at the outset, lacking in any visible trend, linear or otherwise.

A Hierarchically **Clustered Correlation heatmap** of the 29 numeric variables, after the above corrections/modifications, with the resulting correlation values rounded, is displayed below :

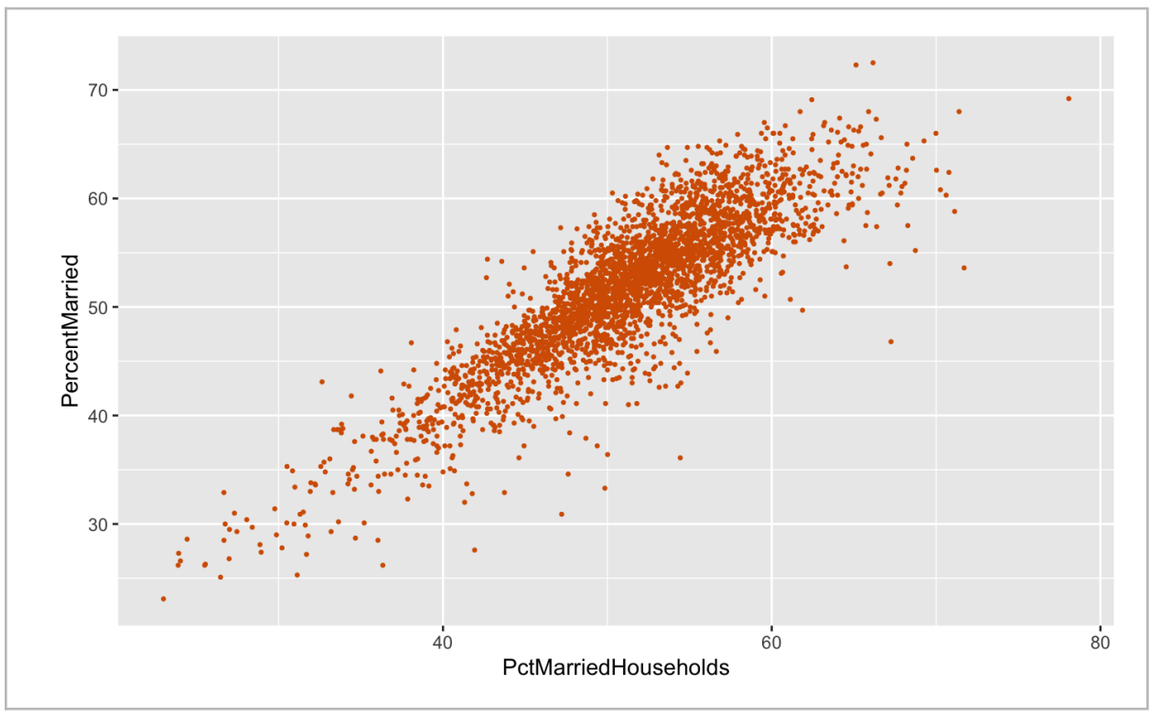


Manual Attribute/Variable Pruning:

Trivially, the Median Age is highly correlated with **MedianAge** is correlated with **MedianAgeFemale** and **MedianAgeMale**.
Hence it doesn't make much sense in carrying all the 3 variables, and **MedianAge** can be used as a proxy for all the three variables. This is also justified by the near similar distribution of the **MedianAgeFemale** attribute & the **MedianAgeMale** attribute, as can be seen below**:**



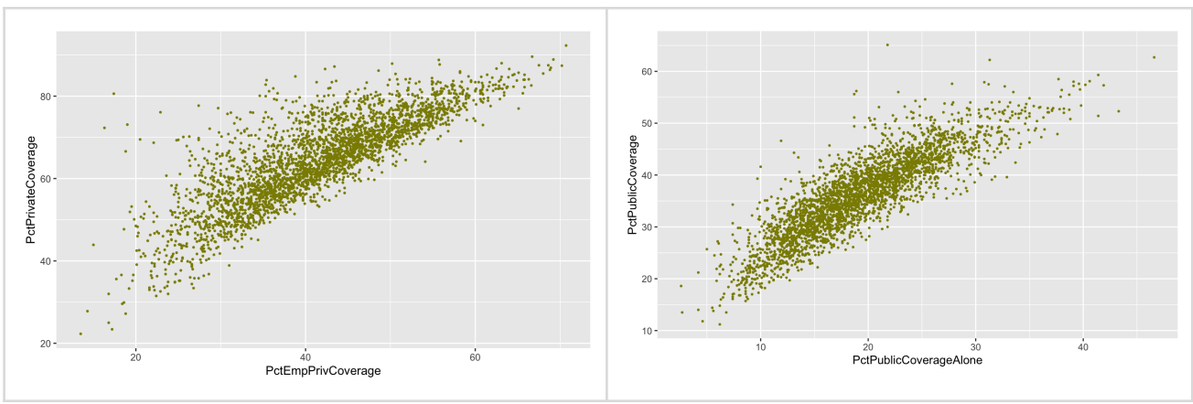
Another trivial correlation, is that between the **PctMarriedHouseholds** and the **PercentMarried**. Thus, it would suffice to include only one of these variables, say, **PercentMarried**, and drop the other one, **PctMarriedHouseholds,** for further analysis.



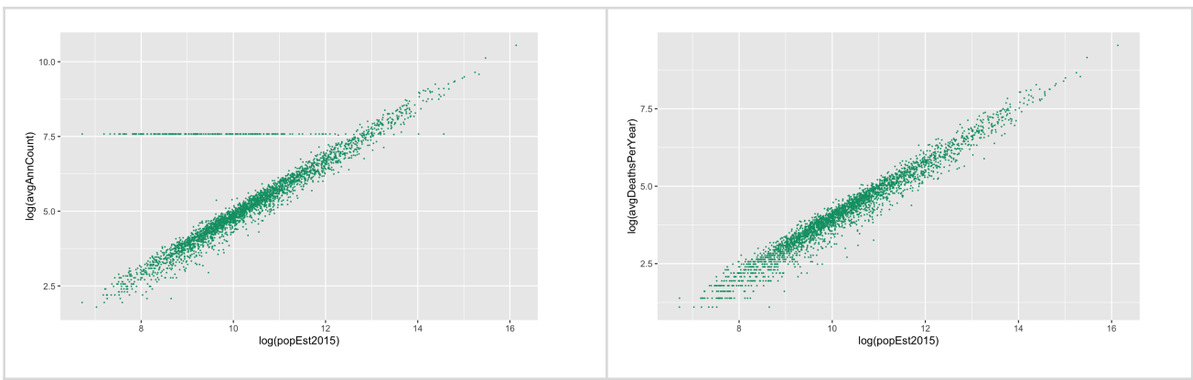
Also intuitive is the rather high (0.83) correlation between the **PctEmpPrivCoverage** and **PctPrivateCoverage,** which inturn has a correlation of 0.75 with **medIncome.** This suggests the use of medIncome as a proxy for the three variables together, but for now we retain both **PctPrivateCoverage** & **medIncome** for further analysis, while dropping **PctEmpPrivCoverage** alone.

Public Social security coverage (**PctPublicCoverage**) is unsurprisingly highly (0.87) correlated with public coverage alone (**PctPublicCoverageAlone**), as one might expect on as public coverage is, for the most part, the last resort for many individuals on account of the public products being poor on service and high on inane formalities that one has to fulfil at access.

The above descriptions are captured in the plots below:



As can be expected, the gross aggregate incidence and mortality counts (**avgAnnCount** & **avgDeathsPerYear**) are correlated with the total county population (**popEst2015**). And this is explored in the plots below:



The predominantly linear trend between **avgAnnCount** and **popEst2015** is disrupted by a series of observations related to 206 counties, for whom the mean number of cases recorded (incidence count as given by **avgAnnCount**) has been, inexplicably, recorded as the same number (1962.668), indicating an error in data processing. Despite that, the **popEst2015** is correlated to an extent of 0.93, and thus we can hope that by dropping both **avgAnnCount** & **avgDeathsPerYear**, while retaining only **popEst2015**, so that we can, in one shot, do away with the errors in data while also capturing the variance that these three variables might have brought to the output.

Multiple Regression Model:

At the outset, a linear model that is also linear in all the variables is fit using all of the remaining 21 variables and the summary is displayed below.

```
## Call:
## lm(formula = TARGET_deathRate ~ ., data = cancer_df_mod)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -105.052  -11.162   -0.306   10.689  140.393
##
## Coefficients:
##                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)              1.850e+02  1.630e+01  11.349  < 2e-16 ***
## incidenceRate            1.942e-01  7.183e-03  27.032  < 2e-16 ***
## medIncome                1.955e-04  7.439e-05   2.628  0.00863 **
## popEst2015              -9.342e-07  1.270e-06  -0.735  0.46213
## povertyPercent           3.946e-01  1.522e-01   2.593  0.00956 **
## studyPerCap              1.352e-04  6.788e-04   0.199  0.84208
## MedianAge               -8.199e-01  1.134e-01  -7.227 6.21e-13 ***
## AvgHouseholdSize        -1.569e+01  2.430e+00  -6.456 1.25e-10 ***
## PercentMarried           1.159e-01  9.325e-02   1.243  0.21411
## PctNoHS18_24            -1.438e-01  5.589e-02  -2.572  0.01015 *
## PctHS18_24               2.608e-01  4.921e-02   5.300 1.24e-07 ***
## PctBachDeg18_24         -2.143e-02  1.071e-01  -0.200  0.84141
## PctHS25_Over             4.603e-01  9.457e-02   4.867 1.19e-06 ***
## PctBachDeg25_Over       -1.243e+00  1.519e-01  -8.186 3.95e-16 ***
## PctUnemployed16_Over     5.283e-01  1.598e-01   3.305  0.00096 ***
## PctPrivateCoverage      -5.257e-01  1.045e-01  -5.031 5.17e-07 ***
## PctPublicCoverageAlone   1.431e-01  1.476e-01   0.970  0.33237
## PctWhite                -1.573e-01  5.622e-02  -2.798  0.00518 **
## PctBlack                -5.581e-02  5.492e-02  -1.016  0.30954
## PctAsian                 4.664e-03  1.890e-01   0.025  0.98031
## PctOtherRace            -8.929e-01  1.248e-01  -7.152 1.07e-12 ***
## BirthRate               -9.782e-01  1.899e-01  -5.151 2.76e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 19.47 on 3025 degrees of freedom
## Multiple R-squared:  0.5114, Adjusted R-squared:  0.508
## F-statistic: 150.7 on 21 and 3025 DF,  p-value: < 2.2e-16
```

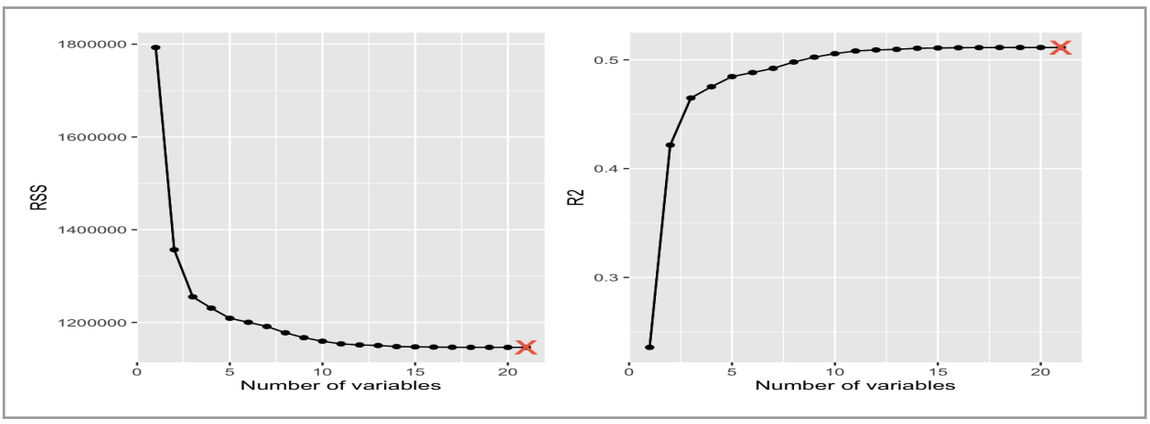
An F-statistic value of ~150 (>>1) suggests that the possibility of all the regression coefficients being zero is very unlikely and that there exists a linear relationship with at least one of the variables.

But, the Adjusted R-squared value, a proxy for test error, being a mere 0.508 suggests that a model that is only linear in the variables is inadequate in explaining the variance in the dependent variable and that we need to incorporate additional variables, either those beyond the ones dataset or that we need to seek out ‘better’ variables from among the available ones, and this includes the possibility of considering interaction terms.

Subset Selection:

Within the scope of a linear model that is also linear in the variables, subset selection is attempted, using best, forward and backward, to gauge what the optimal number of attributes might be and what they might be.

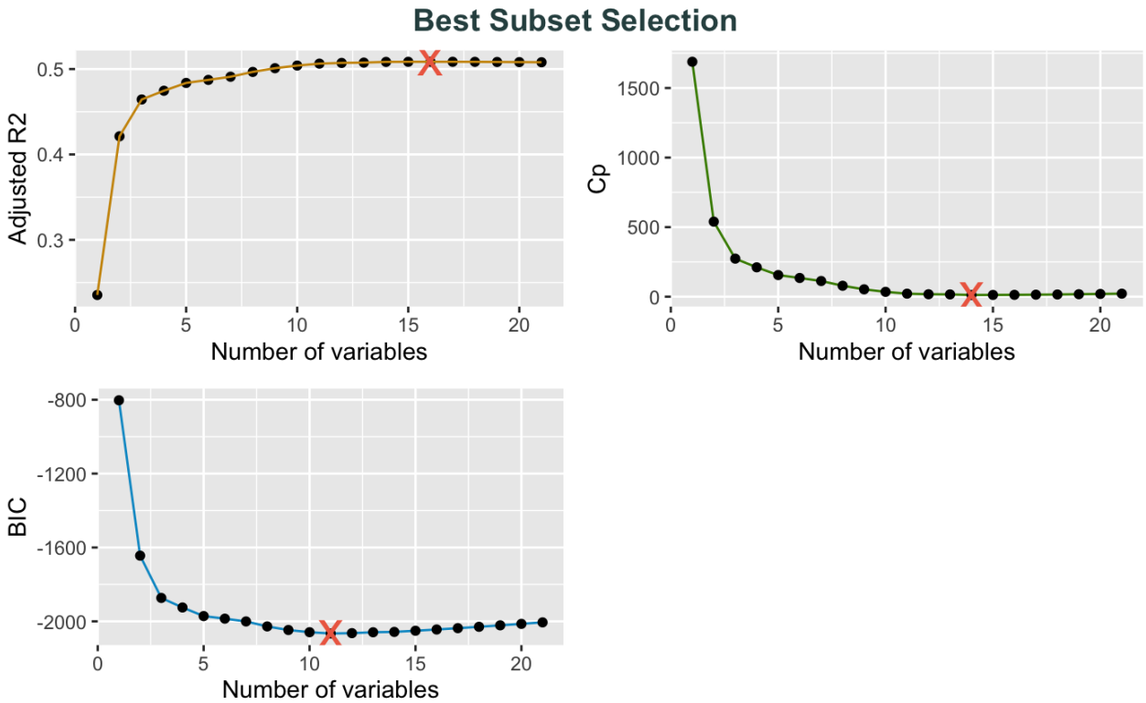
Best Subset Selection:



As the model flexibility/complexity increases, there is overfitting and this is responsible for the monotonic increase in the R2 statistic & a monotonic decrease in RSS statistic.
If these train error estimates are chosen to select acmong from models of different numbers of predictors, we would be underestimating test error and would always end up choosing the model with max predictors (21) allowed, as can be seen in the above plots.

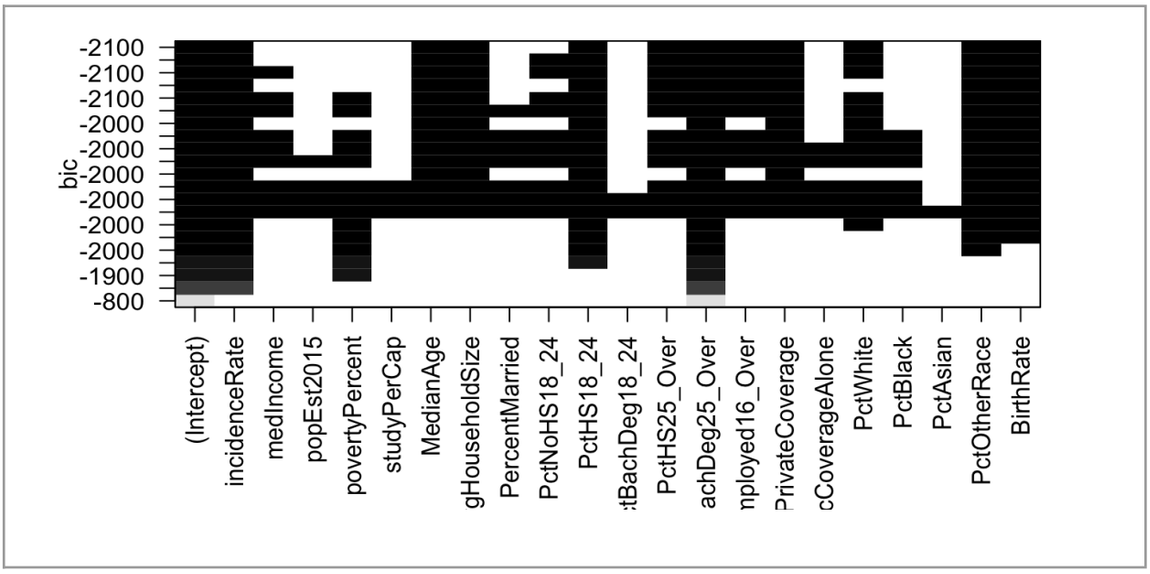
While the R2 and RSS are good criteria to choose from within different models of a given flexibility (given number of predictors), they are not ideal to compare models of different flexibility.

To estimate test error, we have the option of choosing from Mallows's (Cp), BIC or adjusted R squared (adjr2) statistics, which are arrived at by including a flexibility penalty correction to the train error estimate.
The model which has the highest Adjusted R2 or the lowest BIC or the lowest Cp can be chosen as the 'best' from among models of different numbers of predictors.



As can be seen from the above plots, after the penalty for over estimation is applied, the model with 21 predictors is no longer the optimal one.
Adjusted R2, the Cp & BIC criterion suggest the 16-predictor, 14-predictor model & the 11-predictor model respectively, are the optimals ones according to the test error estimates.

The BIC results are explored to see what are the attributes that have produced the minimum test error estimate:



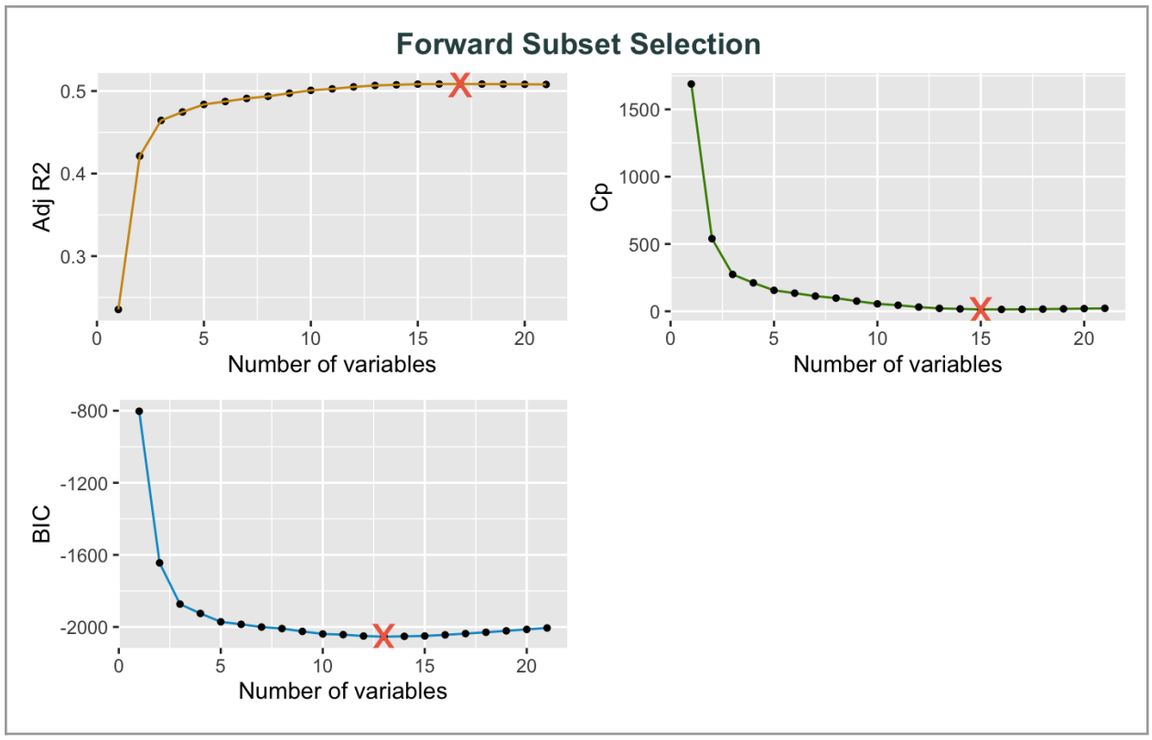
The variables that have made the cut via the best subset selection are: the bias term, **incidenceRate**, **MedianAge**, **AvgHouseholdSize**, **PctHS18_24**, **PctHS25_Over**, **PctBachDeg25_Over**, **PctUnemployed16_Over**, **PctPrivateCoverage**, **PctWhite**, **PctOtherRace**, & **BirthRate.**

A model is now fit upon only these selected variables:

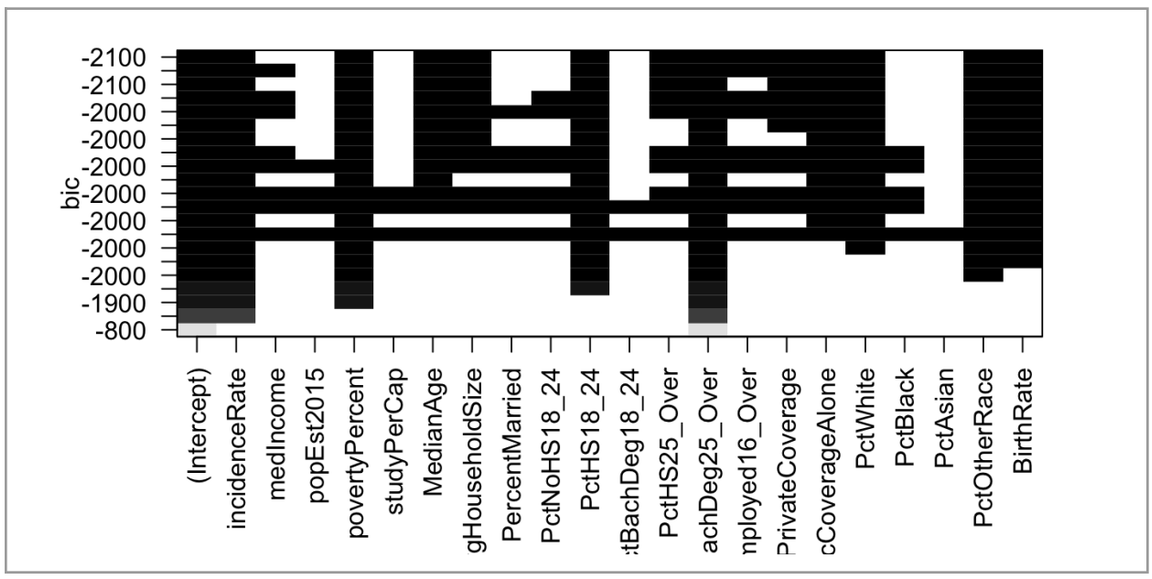
```
## Call:
## lm(formula = TARGET_deathRate ~ ., data = cancer_df_mod2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -108.736  -11.323   -0.198   10.677  140.582
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)           200.779495  10.304600  19.484  < 2e-16 ***
## incidenceRate           0.197039   0.006977  28.239  < 2e-16 ***
## MedianAge              -0.874192   0.096694  -9.041  < 2e-16 ***
## AvgHouseholdSize      -14.654067   2.007534  -7.300 3.67e-13 ***
## PctHS18_24              0.295388   0.045710   6.462 1.20e-10 ***
## PctHS25_Over            0.422683   0.092271   4.581 4.82e-06 ***
## PctBachDeg25_Over      -1.142970   0.135361  -8.444  < 2e-16 ***
## PctUnemployed16_Over    0.633140   0.148165   4.273 1.99e-05 ***
## PctPrivateCoverage     -0.560941   0.056526  -9.924  < 2e-16 ***
## PctWhite               -0.109429   0.028260  -3.872  0.00011 ***
## PctOtherRace           -0.894554   0.116643  -7.669 2.32e-14 ***
## BirthRate              -0.973972   0.183438  -5.310 1.18e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 19.5 on 3035 degrees of freedom
## Multiple R-squared:  0.5081, Adjusted R-squared:  0.5063
## F-statistic:   285 on 11 and 3035 DF,  p-value: < 2.2e-16
```

Even with half the variables (11/21), the Adjusted R-square has almost remained the same, while the F-statistic has gone higher, indicating the redundancy among the attributes.
Also apparent, is the low p-values of all these coefficients, an indication that the null hypothesis that these coefficients are zero is very unlikely.

Forward subset selection:



When we use forward selection algorithm, Adjusted R2 criterion of test error estimate suggests that a 17-predictor model is the optimal one, while the Cp favours a 15-variable model and BIC a 13-variable one.



In the forward selection, using BIC criterion, 2 additional variables, namely **povertyPercent** & **PctPublicCoverageAlone** have been included in the optimal model, apart from all the 11 variables that best subset selection has shortlisted.

A model is fit using these 13 variables:

```
## Call:
## lm(formula = TARGET_deathRate ~ ., data = cancer_df_mod2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -108.036  -11.257   -0.195   10.632  140.111
##
## Coefficients:
##                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)             178.902845  15.240852  11.738  < 2e-16 ***
## MedianAge                -0.828682   0.105323  -7.868 4.97e-15 ***
## AvgHouseholdSize        -13.092752   2.192771  -5.971 2.63e-09 ***
## PctHS18_24                0.307979   0.046215   6.664 3.15e-11 ***
## PctHS25_Over              0.409123   0.092694   4.414 1.05e-05 ***
## PctBachDeg25_Over        -1.112476   0.137121  -8.113 7.09e-16 ***
## PctUnemployed16_Over      0.530894   0.156159   3.400 0.000683 ***
## PctPrivateCoverage       -0.403922   0.094881  -4.257 2.13e-05 ***
## PctWhite                 -0.102311   0.029314  -3.490 0.000489 ***
## PctOtherRace             -0.879265   0.118489  -7.421 1.51e-13 ***
## povertyPercent            0.149029   0.126172   1.181 0.237632
## PctPublicCoverageAlone    0.209718   0.143807   1.458 0.144853
## incidenceRate             0.195023   0.007126  27.367  < 2e-16 ***
## BirthRate                -0.952545   0.183936  -5.179 2.38e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 19.49 on 3033 degrees of freedom
## Multiple R-squared:  0.5088, Adjusted R-squared:  0.5067
## F-statistic: 241.7 on 13 and 3033 DF,  p-value: < 2.2e-16
```

The estimates for slopes give an indication of how significant an impact the attributes have upon the dependent variable (mortality rates).

The coefficient associated with **BirthRate** (no. of live births) is '-0.95', which is not surprising as a higher BirthRate is indicative of access to better medical services in the given county and hence the lower mortality rate.

Intuitively, the attribute corresponding to the '**Avg household size**' is a bit confusing as to its impact on mortality rates, and the regression model gives us a possible clarification on this. On the one hand a bigger household implies a better social support system to cope with illness and pull through, but that is only when there are surplus resources at disposal. In resource scarce settings, a bigger household has been observed to have the effect of rationing of resources, especially towards its female members, and this could result in higher mortality rates.
A coefficient of '-0.82' suggests that the support effect of a bigger household dominates the rationing effect and those in bigger households have a better chance of survival than otherwise.

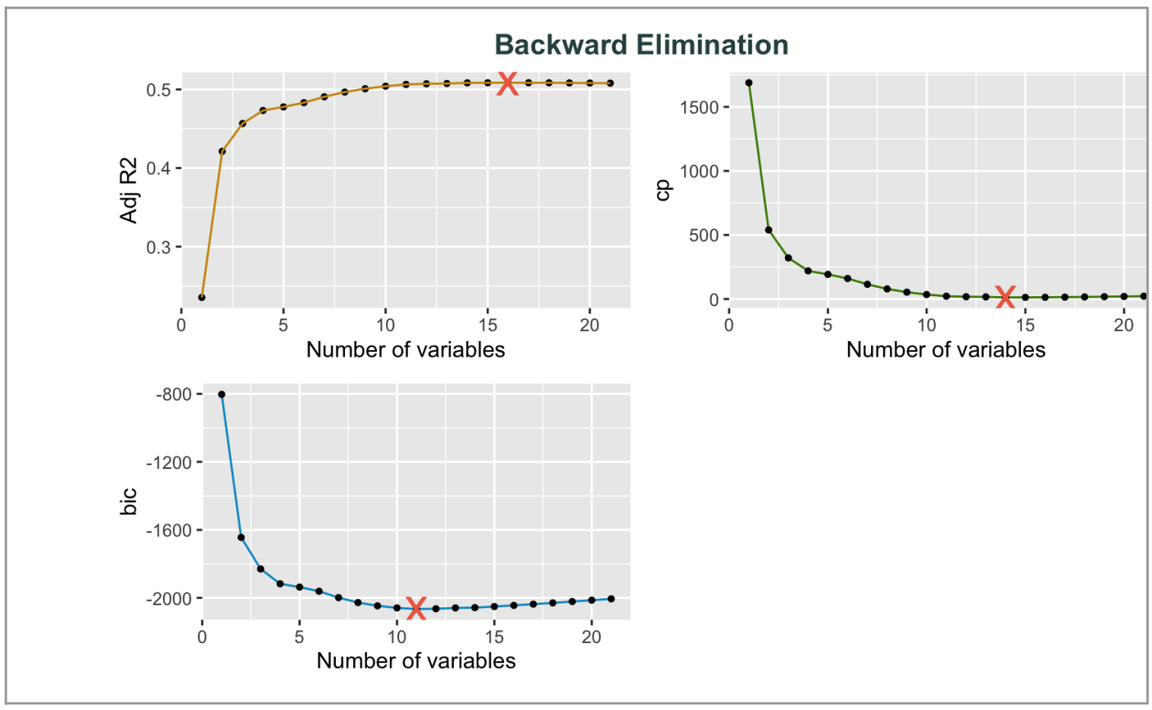
The negative coefficients associated with Percent White & Percent Other Race, together, capture the added disability that black communities face when dealing with life threatening diseases such as cancer. It might be most likely that those counties that have historically had high percent of black population might have suffered from systemic racial neglect in terms of allocation of resources towards public health infrastructure and thus the greater observed risk of an incidence leading to mortality in such counties.

The regression coefficients also suggest the importance of improving the quality of public health coverage services/products. Expectedly, the increase in access to private coverage has an effect of decrease in mortality rates, but a slightly surprising though not entirely unexpected aspect is that of a positive coefficient associated with the attribute 'Public Coverage alone'. This makes sense though, because the percent of population with access to public coverage alone is a great proxy for people who might not have been captured by the poverty percentage, yet who

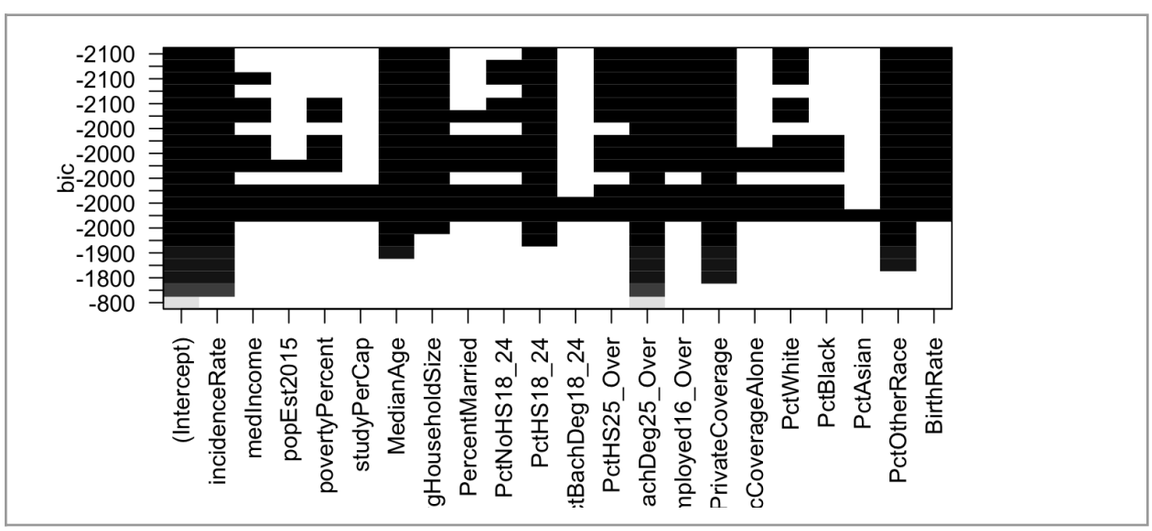
are disadvantaged in their access to quality medical services. The people who have public coverage as their only resort are more likely to be deprived than those who can access private coverage too, thus making them vulnerable to incidence and mortality.

MedianAge is another attribute that sends mixed signals when common logic is applied. On the one hand, a higher median age would imply longer exposure to risk factors and hence higher levels of incidence and mortality, but on the other hand a higher median age would also imply more social support/connections, probably a more stable source of income and reduced risk taking behaviour which might translate to less incidence & mortality.
If we go by the regression coefficient the latter effect of age seems to be more prominent than the former, in that the death rates seem to come down as the median age of the county rises.

Backward subset Selection:



When we use backward elimination algorithm, Adjusted BIC criterion of test error estimate suggests that a 11-predictor model is the optimal one, while Cp & Adjusted R2 endorse a 14 and 16-variable model respectively.
The models arrived at using backward elimination seem to have in general a bias towards higher number of predictors, compared to best subset and forward selection.

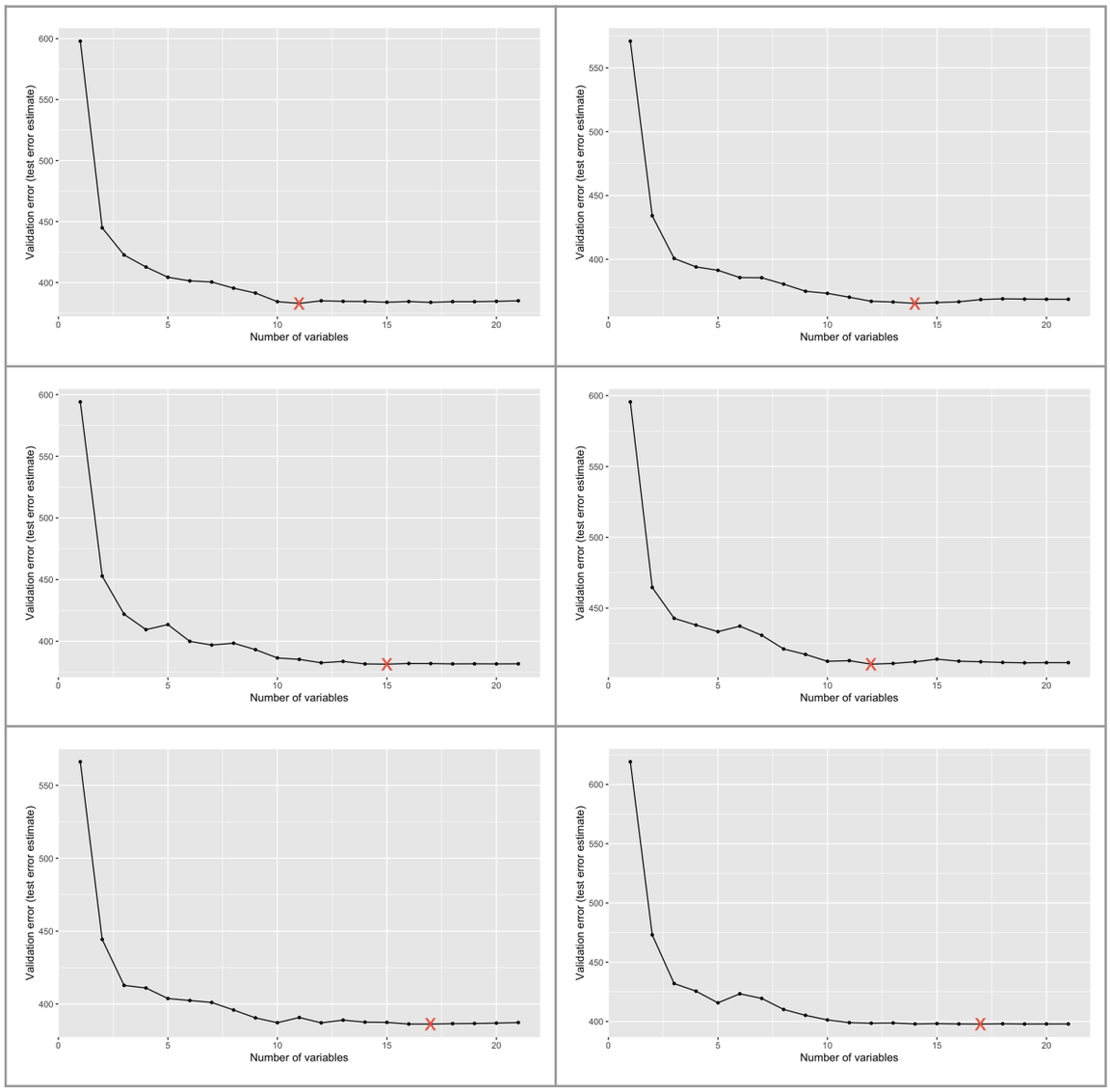


As BIC statistic places a heavier penalty on models with many variables (n>7), and the models chosen by that criterion have expectedly smaller number of features/variables in them compared to the models endorsed by Cp.

The backward selection, using the BIC criterion, has fetched the same 11 features as those that have been picked up using best subset selection.

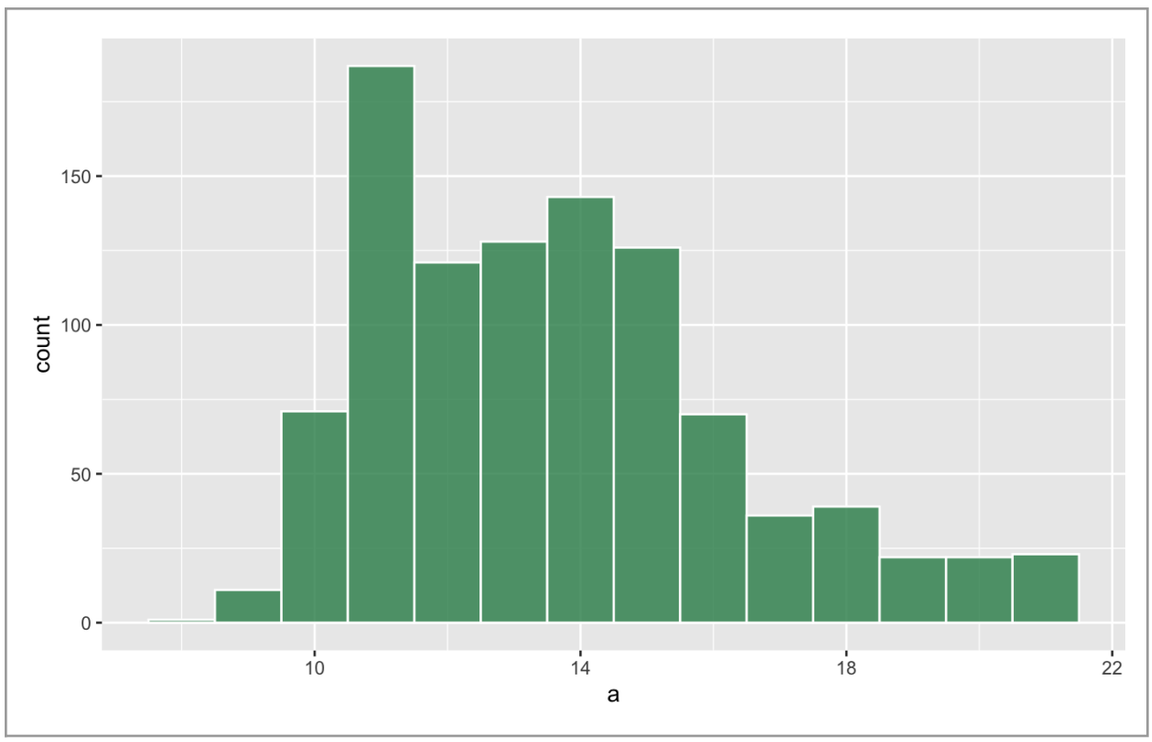
Validation set approach to identifying optimal number of variables:

A certain portion of data is split randomly and set aside for validation of the model and the following plots emerge from different random splits.



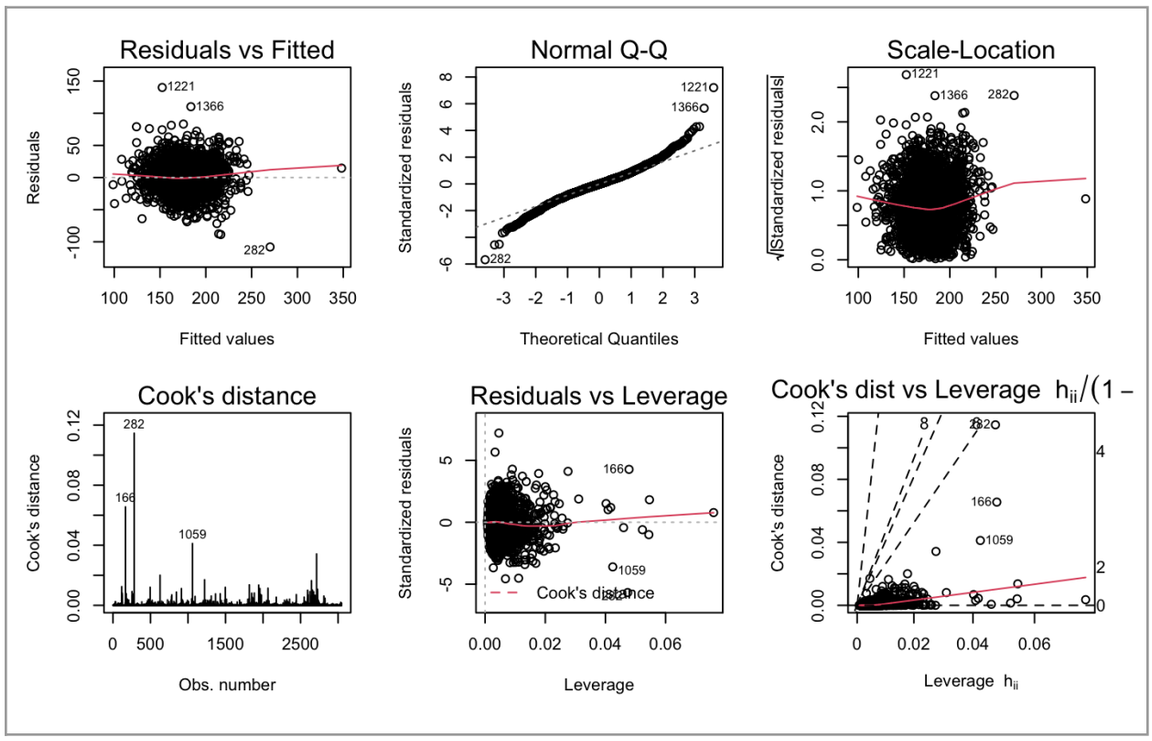
As can be seen above, the model that is predicted as the 'best' by the validation set approach seems to be highly sensitive to the random seed used for test-train split.
Depending upon the split, the model with least test error estimate is as small as one with 8 predictors to one with 21 variables.

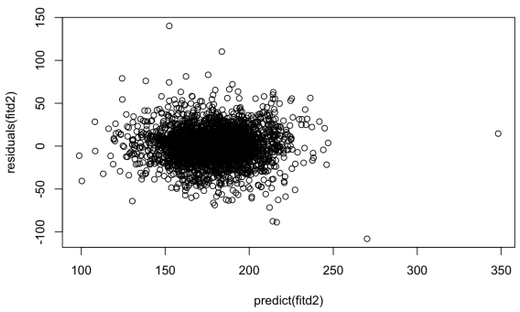
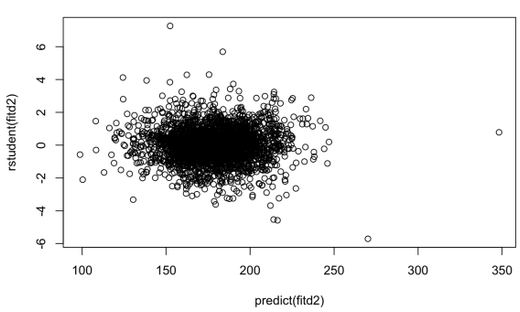
The following is a histogram of the validation set results of around 1000 iterations, and the model with 11 variables stands out as the mode, giving an indication as to the optimal number of variables.



Diagnostic plots:

Regression diagnostic analysis is performed on a best subset of 13 variables obtained through forward selection.



Residuals	Standardised Residuals
	

Residuals vs Fitted plot:

This plot gives an indication about the presence of non-linear patterns in the residuals. Residuals that are relatively equally/randomly spread around the horizontal (red) line without any distinct pattern is an indication that there might not be any non-linear relationships, which appears to be the case here.

Absence of any strong trend/pattern in the residuals plot, as evident from the horizontal red line that has been obtained from smoothing the data, might be suggestive of the absence of non-linearity in the relationships between variables.

Normal Q-Q plot:

The uncorrelatedness of the error term, that has been made the catch all for all the elements that are outside our ability to model, is a crucial assumption in regression estimates. If the residuals are lined up relatively well along a straight line, as the case seems to be in this instance, we can conclude that our assumption of error terms being uncorrelated has in fact held up well.

Deviation from line, suggestive of correlation between error terms will result in the underestimation of standard error, thus causing the prediction and confidence intervals to be narrower than they ought to be and p-values lesser than they ought to be.
We could end up where the stated confidence interval (95%) contains the true parameter with a probability less than 0.95. i.e. we would run the risk of having a misplaced sense of confidence in our model.

Correlation between the error terms can be observed when a factor such as geographical location & proximity, predominant climate of a county, etc. that could in fact be modelled has not been accounted for or when care has not been taken to control those factors.
For instance, the climate & pollution levels in counties have an impact on lung cancer mortalities and counties with greater geographical proximity are likely to have similar mortality rates on account of this, and since this has not been modelled, it might show up as a correlation in the error term.

Scale-Location or Spread-Location Plot:

This plot gives an indication as to the spread of residuals along the range of the predictor. This can help us check the validity of our assumption of equal/constant variance (homoscedasticity) in the error terms. The validity of this assumption has implications for standard error, confidence intervals and hypothesis tests.

Unequal variance or heteroscedasticity is borne out via funnel shaped spread in the residual plot. Our current spread plot seems to be free of this funnelling effect, thus suggesting that our assumption of homoscedasticity might be a valid one.

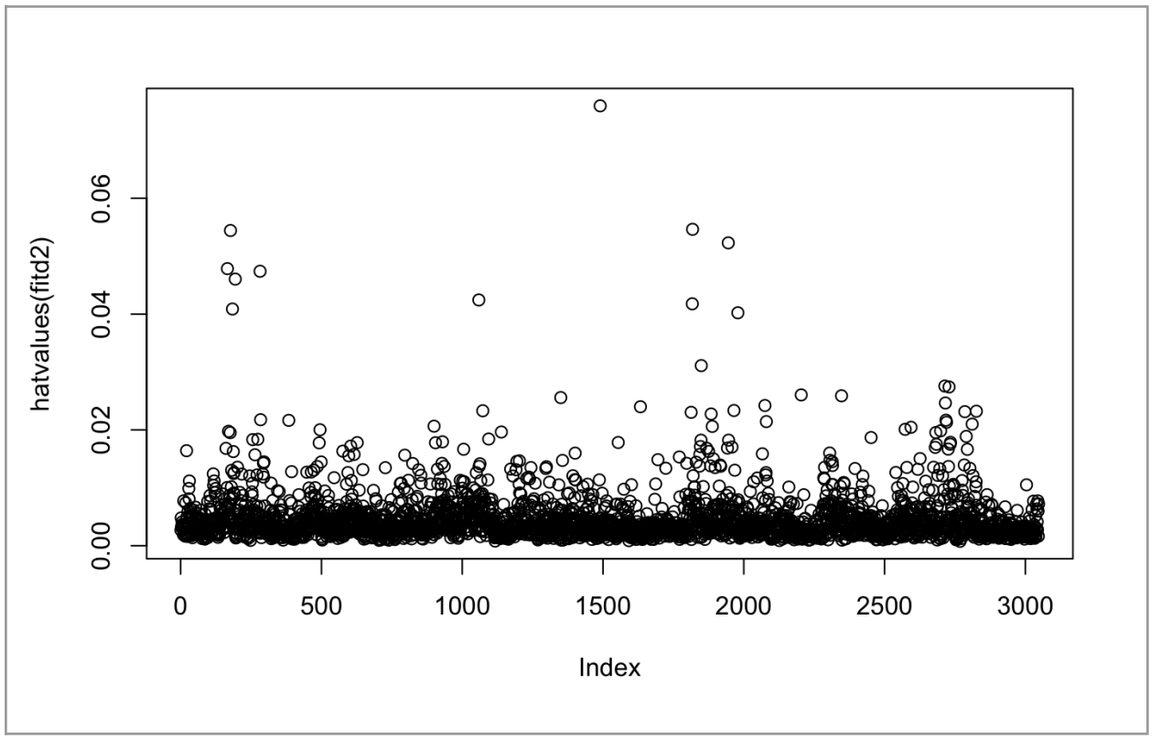
Cook's distance & influence:

This plot gives an indication of the influence of outliers. Whatever the definition we may choose for an outlier, not all of them might be influential to impact the regression fit, suggesting that their exclusion/inclusion would not substantially alter the results of the regression analysis. When they don't get along with the trend reasonably, they become influential, in the sense that their exclusion will impact the results of regression.
Even when the fit is not significantly altered, outliers could still have implications for standard errors, p-values & confidence intervals.

To quantify this impact & detect influential outliers with less ambiguity, we go for studentized/standardised residuals, instead of just relying on residuals. These are obtained by dividing the residuals by the estimated standard error.
Cases with studentized residual values of above 3 are possible outliers.
Upper right corner and the lower right corner are the spaces in the plot where we can find subjects that are influential against regression. Cases with high cook's distance scores (lie outside cook's distance) are influential to regression results.
In our case all the cases are within the cook's distance line (red dashed line) and the upper cook's line is not even visible, thus suggesting that there are no influential outliers.

While outliers have unusual output/response, unusual predictor values could have high leverage. While identifying a high leverage point in simple linear regression is pretty straightforward of picking out points for which the predictor value is beyond the common range, in case of multiple predictors, visual aids are either deceiving or not available.

To overcome this, we can quantify an observation's leverage through leverage statistic (h). In simple regression this statistic takes the value between 1/n and 1, where n is the number of observations and sum of the leverages is the number of parameters (including intercept). This makes the average as (params+1)/n, so, when the h value of an observation exceeds this average by significant margin, we can conclude that the observation point has high leverage.



The function **hatvalues()** in R can be used to compute the leverage statistics for the predictor and the observation number **1490** seems to have the highest leverage.

Conclusions:

An attempt to link socio-economic indicators and demographic variables to cancer mortality rates has been taken up as a part of this project. After applying various subset selection procedures, an optimal model with 11-23 variables has been found to be the optimal one with an adjusted R^2 Value of ~0.51, suggesting that a linear model that is also linear in the variables can only account for half of the variability in the dependent variable.
While the assumptions related to normality of error terms or their constant variance have not been violated, the model's middling performance does not render it with much utility. Though some insights could be drawn from the coefficients, the model as a whole has not been able to 'explain' the target attribute in a satisfactory manner. Several other alternate models incorporating many combinations of interaction terms have been tried with little to no improvement to the adjusted R^2 statistic, unless an obscene amount of attributes have been incorporated.

This suggests that either that there were additional flaws in data aggregation apart from those identified or that many other variables with greater predictive power have not been modelled and thus the greater value of unexplained variability. Another flaw in the model could be the choice of county as a unit of analysis. Many counties have such a large diversity that many of the crucial determinants of mortality could have been averaged/cancelled out when data has been aggregated at the level of a county.