

MODELLING CANCER MORTALITY USING OLS REGRESSION

DA 546 PROJECT, JAN-APR, 2022

BUDDIKIRAN CHAITANYA 214161002 buddi.kiran@iitg.ac.in

INTRODUCTION:

-PROMLEM STATEMENT

-SCOPE AND FEASIBILITY

-WHY MORTALITY -ROADMAP & EXPECTATIONS

DATASET:

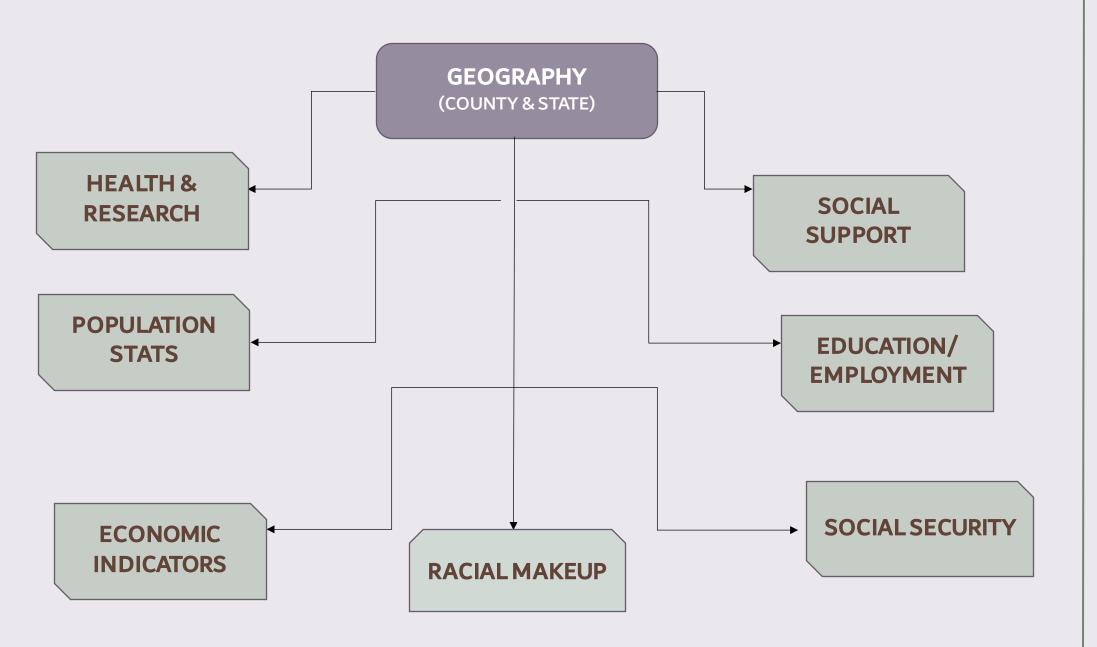
-SOURCE

-GENERAL DESCRIPTION OF THE VARIABLES

-MISSING DATA & POSSIBLE ANAMOLIES

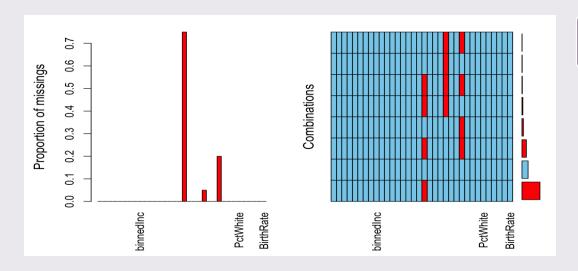
-MODIFICATIONS







MISSING DATA & MISC.:



MISSING DATA

3 ATTRIBUTES

RECOURSE

JUSTIFICATION

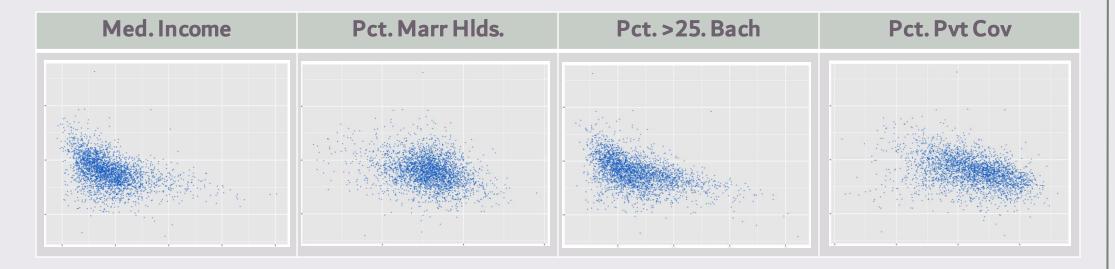
ANAMOLIES

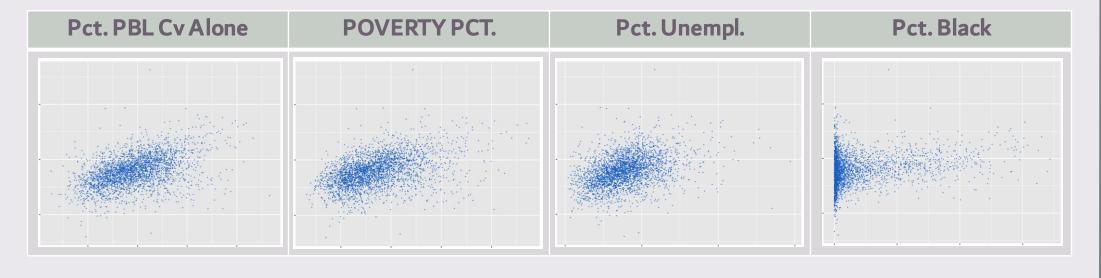
HOUSEHOLD SIZE AND AGE

CORRECTIONS



SCATTER-PLOTS:







MANUAL PUNING USING DATA CLUSTERS

- COVERAGE CLUSTER
- AGE CLUSTER
- MARRIED CLUSTER
- POVERTY CLUSTER
- POPULATION CLUSTER

1.05 1.05	0.1	0.2	0.2	0.2	0.3	-0.2									-0.2	0	-0.1	0.1	0.1	-0.1	0	-0.1	0.1	0.1	0.2	0.5	0.9	1	1
avgDeathsPerYear	0.1	0.2	0.3	0.2	0.3	-0.1	-0.1	-0.1	-0.2	-0.2	-0.2	-0.2	-0.3	-0.1	-0.1	0	-0.1	0.1	0.1	-0.1	0.1	-0.1	0.1	0.1	0.2	0.4	0.9	1	1
avgAnnCount	0.1	0.2	0.3	0.3	0.3	-0.1	-0.1	-0.1	-0.1	-0.1	-0.1	-0.2	-0.3	-0.1	-0.2	-0.1	-0.1	0	0	-0.1	0.1	0	0.1	0.1	0.2	0.4	1	0.9	0.9
PctAsian	0.2	0.3	0.3	0.4	0.4	-0.3	-0.2	-0.2	-0.3	-0.1	-0.1				-0.3	-0.2	-0.2	0	0	-0.2	0	-0.1	0.1	0.1	0.2	1	0.4	0.4	0.5
PctOtherRace	-0.2	-0.1	0	0.1	0			-0.3		-0.1	0		-0.3		-0.1	0.1	0	0	0	-0.2	-0.2	0.1	0	0.2	1	0.2	0.2	0.2	0.2
	-0.1	0	-0.1	0.1	0	-0.4	-0.4	-0.3	-0.2	-0.1	0.1	0	-0.1	0.1	-0.1	0.1	0.1	0.1	0	0	-0.1	0.1	0	1.		0.1	0.1	0.1	0.1
studyPerCap	0.1	0.1	0.1	0	0.1	0	0	0	0	0	-0.1	-0.1	-0.1	-0.1	-0.1	-0.1	-0.1	0	0	0	0.1	0	1	0	0	0.1	0.1	0.1	0.1
BirthRate	0	-0.1	-0.1	0	-0.1	-0.1	-0.1	-0.1	0	0.1	0.1	0.1	0	0.1	0	0	0	-0.1	-0.1	-0.1	-0.1	1	0	0.1	0.1	-0.1	0	-0.1	-0.1
incidenceRate	0.1	0.1	0	0	0	0	0	0	0	-0.1	-0.2	0	0.1	-0.2	0	0	0	0.1	0.1	0.4	1	-0.1	0.1	-0.1	-0.2	0	0.1	0.1	0
TARGET_deathRate							0			-0.3			0.4	0.1	0.4	0.4	0.4	0.4	0.3	1	0.4	-0.1	0	0	-0.2	-0.2	-0.1	-0.1	-0.1
PctBlack				-0.3									0	0.1	0.2	0.3	0.5	0.5	1	0.3	0.1	-0.1	0	0	0	0	0	0.1	0.1
PctUnemployed16_Over	-0.6	-0.5	-0.3	-0.5	-0.4	-0.1	-0.1	-0.1	-0.5	-0.6	-0.5	0.1	0.1	0.2	0.5	0.7	0.7	1	0.5	0.4	0.1	-0.1	0	0.1	0	0	0	0.1	0.1
povertyPercent	-0.8	-0.7	-0.4	-0.8	-0.5	-0.1	-0.2	-0.2	-0.5	-0.6	-0.6	0.1	0.2	0.3	0.7	0.8	1	0.7	0.5	0.4	0	0	-0.1	0.1	0	-0.2	-0.1	-0.1	-0.1
PctPublicCoverageAlone	-0.9	-0.7	-0.4	-0.7	-0.6	0	0	0	-0.4	-0.5	-0.5	0.2	0.3	0.3	0.9	1	0.8	0.7	0.3	0.4	0	0	-0.1	0.1	0.1	-0.2	-0.1	0	0
PctPublicCoverage	-0.7	-0.8	-0.4	-0.8	-0.6	0.5	0.4	0.4	-0.1	-0.2	-0.4	0.3	0.4	0.3	1	0.9	0.7	0.5	0.2	0.4	0	0	-0.1	-0.1	-0.1	-0.3	-0.2	-0.1	-0.2
PctNoHS18_24	-0.5	-0.4	-0.4	-0.3	-0.4	0.1	0.1	0.1	-0.2	0	0	0.1	0.2	1	0.3	0.3	0.3	0.2	0.1	0.1	-0.2	0.1	-0.1	0.1	0.1	-0.2	-0.1	-0.1	-0.1
PctHS25_Over	-0.2	-0.2	-0.4	-0.5	-0.7	0.3	0.3	0.3	0.2	0.1	0.1	0.4	1	0.2	0.4	0.3	0.2	0.1	0	0.4	0.1	0	-0.1	-0.1	-0.3	-0.4	-0.3	-0.3	-0.3
PctHS18_24	-0.3	-0.2	-0.4	-0.2	-0.4	0.2	0.2	0.2	0	0.1	0.1	1	0.4	0.1	0.3	0.2	0.1	0.1	0	0.3	0	0.1	-0.1	0	-0.1	-0.2	-0.2	-0.2	-0.2
PctMarriedHouseholds	0.4	0.3	0	0.4	0.1	0.2	0.2	0.2	0.6	0.9	1	0.1	0.1	0	-0.4	-0.5	-0.6	-0.5	-0.6	-0.3	-0.2	0.1	-0.1	0.1	0	-0.1	-0.1	-0.2	-0.1
PercentMarried	0.4	0.2	0.1	0.4	0.1	0.4	0.4	0.4	0.7	1	0.9	0.1	0.1	0	-0.2	-0.5	-0.6	-0.6	-0.6	-0.3	-0.1	0.1	0	-0.1	-0.1	-0.1	-0.1	-0.2	-0.2
PctWhite	0.4	0.3	0.1	0.2	0	0.3	0.4	0.4	1	0.7	0.6	0	0.2	-0.2	-0.1	-0.4	-0.5	-0.5	-0.8	-0.2	0	0	0	-0.2	-0.2	-0.3	-0.1	-0.2	-0.2
MedianAgeMale	0.1	-0.2	0	-0.1	-0.1	0.9	1	1	0.4	0.4	0.2	0.2	0.3	0.1	0.4	0	-0.2	-0.1	-0.2	0	0	-0.1	0	-0.3	-0.3	-0.2	-0.1	-0.1	-0.2
MedianAge	0.1	-0.2	-0.1	-0.1	-0.2	1	i	1	0.4	0.4	0.2	0.2	0.3	0.1	0.4	0	-0.2	-0.1	-0.2	0	0	-0.1	0	-0.4	-0.3	-0.2	-0.1	-0.1	-0.2
MedianAgeFemale	0	-0.3	-0.1	-0.2	-0.2	1	1	0.9	0.3	0.4	0.2	0.2	0.3	0.1	0.5	0	-0.1	-0.1	-0.2	0	0	-0.1	0	-0.4	-0.3	-0.3	-0.1	-0.1	-0.2
PctBachDeg25_Over	0.6	0.5	0.6	0.7	1	-0.2	-0.2	-0.1	0	0.1	0.1	-0.4	-0.7	-0.4	-0.6	-0.6	-0.5	-0.4	-0.1	-0.5	0	-0.1	0.1	0	0	0.4	0.3	0.3	0.3
medIncome	0.7	0.7	0.5	1	0.7	-0.2	-0.1	-0.1	0.2	0.4	0.4	-0.2	-0.5	-0.3	-0.8	-0.7	-0.8	-0.5	-0.3	-0.4	0	0	0	0.1	0.1	0.4	0.3	0.2	0.2
PctBachDeg18_24	0.5	0.5	1	0.5	0.6	-0.1	-0.1	0	0.1	0.1	0	-0.4	-0.4	-0.4	-0.4	-0.4	-0.4	-0.3	-0.1	-0.3	0	-0.1	0.1	-0.1	0	0.3	0.3	0.3	0.2
PctEmpPrivCoverage	8.0	1	0.5	0.7	0.5	-0.3	-0.2	-0.2	0.3	0.2	0.3	-0.2	-0.2	-0.4	-0.8	-0.7	-0.7	-0.5	-0.2	-0.3	0.1	-0.1	0.1	0	-0.1	0.3	0.2	0.2	0.2
PctPrivateCoverage	1	0.8	0.5	0.7	0.6	0	0.1	0.1	0.4	0.4	0.4	-0.3	-0.2	-0.5	-0.7	-0.9	-0.8	-0.6	-0.3	-0.4	0.1	0	0.1	-0.1	-0.2	0.2	0.1	0.1	0.1

FULL FIT RESULTS WITH 21 VARIABLES:

```
## 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</pre>
```

F-STATISTIC (>>1): SUFFICIENT (as n>>p) EVIDENCE AGAINST NUL HYPOTHESIS Adj. R-SQUARED: 0.51

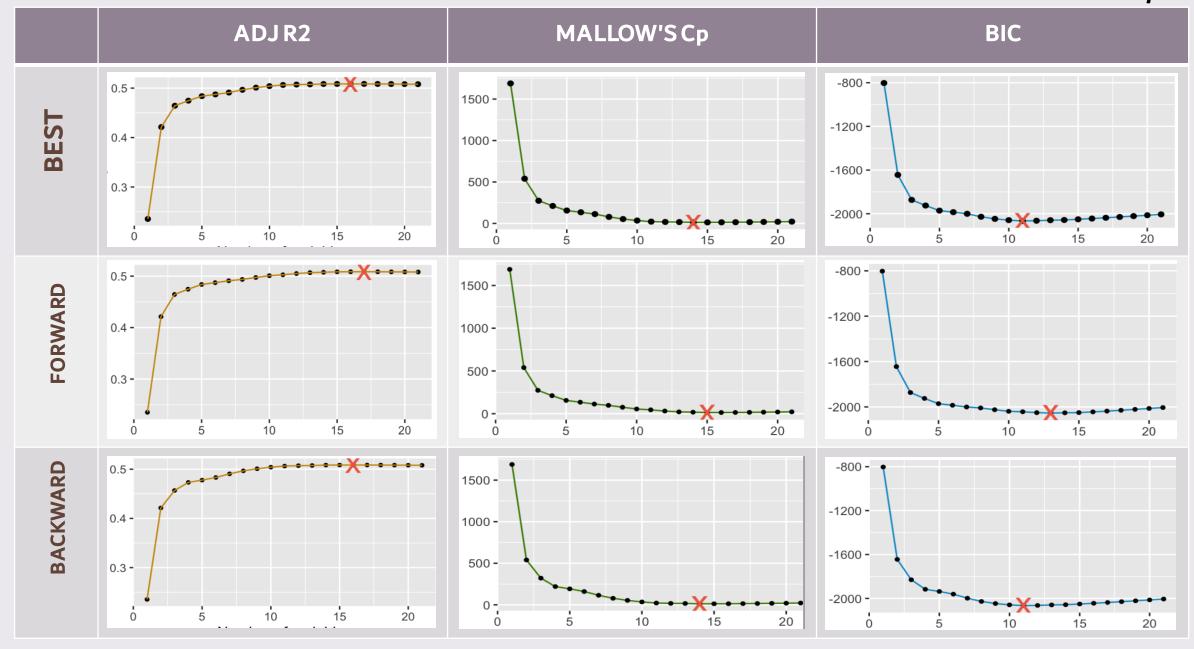
SUBSET SELECTION:

FORWARD

BACKWARD

VALIDATION SET APPROACH

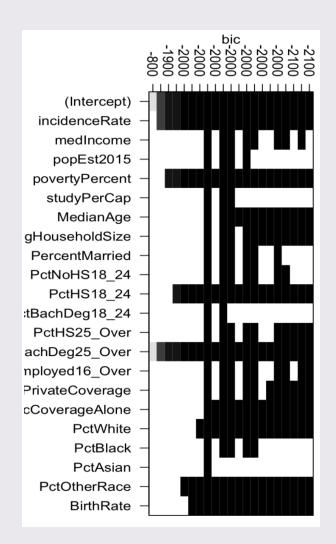




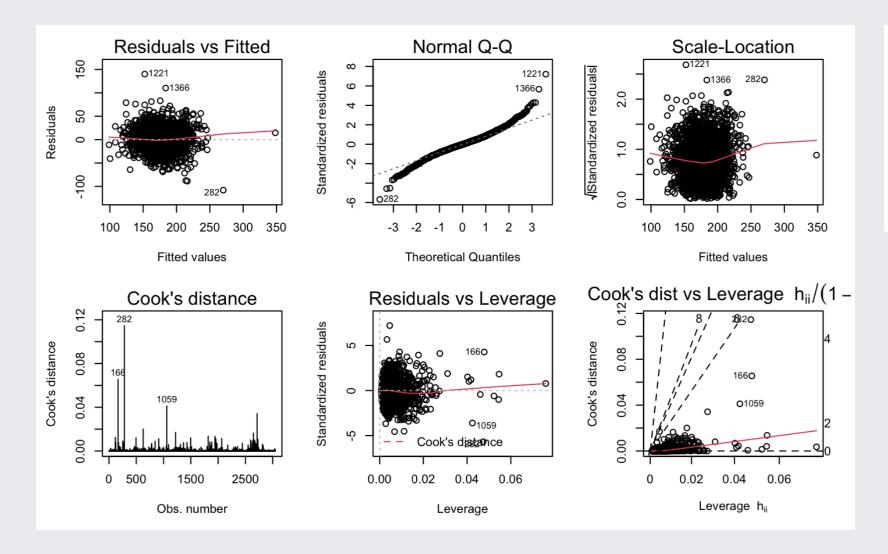
RE-FIT ON SELECTED VARIABLES, SOME OBSERVATIONS..

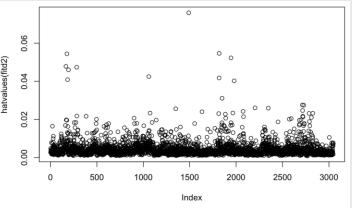
```
## Coefficients:
##
                            Estimate Std. Error t value Pr(>|t|)
   (Intercept)
                          178.902845
                                      15.240852
                                                 11.738 < 2e-16
                                       0.105323 -7.868 4.97e-15
  MedianAge
                           -0.828682
                                       2.192771
                                                 -5.971 2.63e-09
## AvgHouseholdSize
                          -13.092752
## PctHS18 24
                            0.307979
                                       0.046215
                                                  6.664 3.15e-11
                                                  4.414 1.05e-05
  PctHS25 Over
                            0.409123
                                       0.092694
  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
                           -0.102311
## PctWhite
                                       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 \quad 27.367 \quad < 2e-16
## BirthRate
                           -0.952545
                                       0.183936 -5.179 2.38e-07
```

```
## 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</pre>
```



DIAGNOSTIC PLOTS & CONCLUSIONS:







THANK YOU