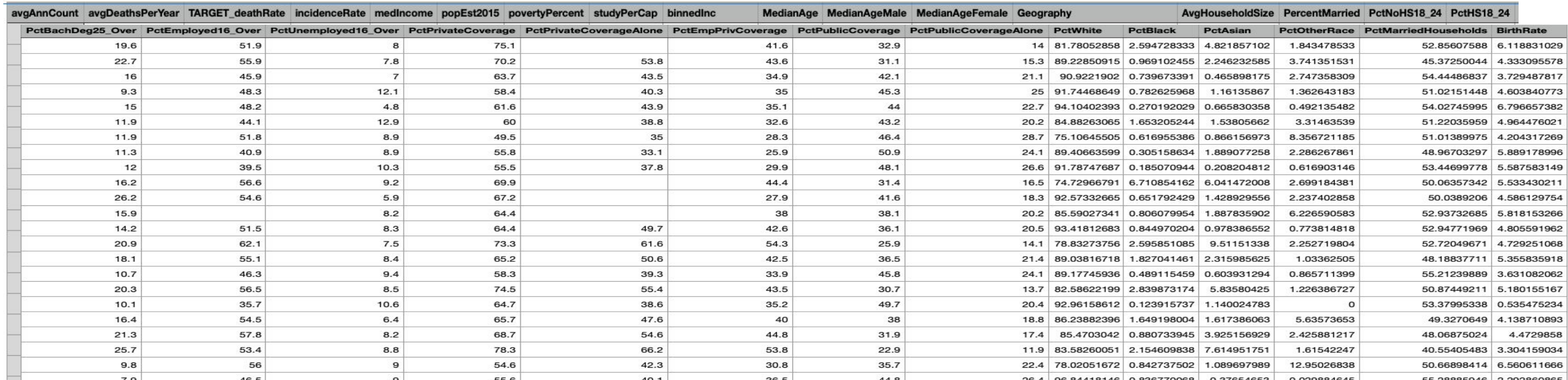
# Analyzing Socioeconomic Factors on USA Cancer Mortality

Jasen Zhang, Alexander Zhu, Christian Pascual

# Background

- ? leading cause of death in the U.S.
- Estimations of cancer mortality statistics
- A comprehensive investigatigation on socioeconomic factors (Focus)
- Aggregated dataset
  2010-2016, over 3,000 counties nationwide



| avgAnnCount | avgDeathsPerYear | TARGET_deathRate | incidenceRate | medIncome | popEst2015 | povertyPercent | studyPerCap | binnedInc | MedianAge | MedianAgeMale | MedianAgeFemale | Geography | AvgHouseholdSize | PercentMarried | PctNoHS18_24 | PctHS18_24 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|

| PctBachDeg25_Over | PctEmployed16_Over | PctUnemployed16_Over | PctPrivateCoverage | PctPrivateCoverageAlone | PctEmpPrivCoverage | PctPublicCoverage | PctPublicCoverageAlone | PctWhite | PctBlack | PctAsian | PctOtherRace | PctMarriedHouseholds | BirthRate |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 19.6 | 51.9 | 8 | 75.1 |  | 41.6 | 32.9 | 14 | 81.78052858 | 2.594728333 | 4.821857102 | 1.843478533 | 52.85607588 | 6.118831029 |
| 22.7 | 55.9 | 7.8 | 70.2 | 53.8 | 43.6 | 31.1 | 15.3 | 89.22850915 | 0.969102455 | 2.246232585 | 3.741351531 | 45.37250044 | 4.333095578 |
| 16 | 45.9 | 7 | 63.7 | 43.5 | 34.9 | 42.1 | 21.1 | 90.9221902 | 0.739673391 | 0.465898175 | 2.747358309 | 54.44486837 | 3.729487817 |
| 9.3 | 48.3 | 12.1 | 58.4 | 40.3 | 35 | 45.3 | 25 | 91.74468649 | 0.782625968 | 1.16135867 | 1.362643183 | 51.02151448 | 4.603840773 |
| 15 | 48.2 | 4.8 | 61.6 | 43.9 | 35.1 | 44 | 22.7 | 94.10402393 | 0.270192029 | 0.665830358 | 0.492135482 | 54.02745995 | 6.796657382 |
| 11.9 | 44.1 | 12.9 | 60 | 38.8 | 32.6 | 43.2 | 20.2 | 84.88263065 | 1.653205244 | 1.53805662 | 3.31463539 | 51.22035959 | 4.964476021 |
| 11.9 | 51.8 | 8.9 | 49.5 | 35 | 28.3 | 46.4 | 28.7 | 75.10645505 | 0.616955386 | 0.866156973 | 8.356721185 | 51.01389975 | 4.204317269 |
| 11.3 | 40.9 | 8.9 | 55.8 | 33.1 | 25.9 | 50.9 | 24.1 | 89.40663599 | 0.305158634 | 1.889077258 | 2.286267861 | 48.96703297 | 5.889178996 |
| 12 | 39.5 | 10.3 | 55.5 | 37.8 | 29.9 | 48.1 | 26.6 | 91.78747687 | 0.185070944 | 0.208204812 | 0.616903146 | 53.44699778 | 5.587583149 |
| 16.2 | 56.6 | 9.2 | 69.9 |  | 44.4 | 31.4 | 16.5 | 74.72966791 | 6.710854162 | 6.041472008 | 2.699184381 | 50.06357342 | 5.533430211 |
| 26.2 | 54.6 | 5.9 | 67.2 |  | 27.9 | 41.6 | 18.3 | 92.57332665 | 0.651792429 | 1.428929556 | 2.237402858 | 50.0389206 | 4.586129754 |
| 15.9 |  | 8.2 | 64.4 |  | 38 | 38.1 | 20.2 | 85.59027341 | 0.806079954 | 1.887835902 | 6.226590583 | 52.93732685 | 5.818153266 |
| 14.2 | 51.5 | 8.3 | 64.4 | 49.7 | 42.6 | 36.1 | 20.5 | 93.41812683 | 0.844970204 | 0.978386552 | 0.773814818 | 52.94771969 | 4.805591962 |
| 20.9 | 62.1 | 7.5 | 73.3 | 61.6 | 54.3 | 25.9 | 14.1 | 78.83273756 | 2.595851085 | 9.51151338 | 2.252719804 | 52.72049671 | 4.729251068 |
| 18.1 | 55.1 | 8.4 | 65.2 | 50.6 | 42.5 | 36.5 | 21.4 | 89.03816718 | 1.827041461 | 2.315985625 | 1.03362505 | 48.18837711 | 5.355835918 |
| 10.7 | 46.3 | 9.4 | 58.3 | 39.3 | 33.9 | 45.8 | 24.1 | 89.17745936 | 0.489115459 | 0.603931294 | 0.865711399 | 55.21239889 | 3.631082062 |
| 20.3 | 56.5 | 8.5 | 74.5 | 55.4 | 43.5 | 30.7 | 13.7 | 82.58622199 | 2.839873174 | 5.83580425 | 1.226386727 | 50.87449211 | 5.180155167 |
| 10.1 | 35.7 | 10.6 | 64.7 | 38.6 | 35.2 | 49.7 | 20.4 | 92.96158612 | 0.123915737 | 1.140024783 | 0 | 53.37995338 | 0.535475234 |
| 16.4 | 54.5 | 6.4 | 65.7 | 47.6 | 40 | 38 | 18.8 | 86.23882396 | 1.649198004 | 1.617386063 | 5.63573653 | 49.3270649 | 4.138710893 |
| 21.3 | 57.8 | 8.2 | 68.7 | 54.6 | 44.8 | 31.9 | 17.4 | 85.4703042 | 0.880733945 | 3.925156929 | 2.425881217 | 48.06875024 | 4.4729858 |
| 25.7 | 53.4 | 8.8 | 78.3 | 66.2 | 53.8 | 22.9 | 11.9 | 83.58260051 | 2.154609838 | 7.614951751 | 1.61542247 | 40.55405483 | 3.304159034 |
| 9.8 | 56 | 9 | 54.6 | 42.3 | 30.8 | 35.7 | 22.4 | 78.02051672 | 0.842737502 | 1.089697989 | 12.95026838 | 50.66898414 | 6.560611666 |

# Data Overview

1. **Cancer Mortality**
   cancer mortalities, reported cases of cancer diagnosed annually, reported mortalities due to cancer, cancer diagnoses rate

2. **Economics**
   median income, percentage of population in poverty, employment rate

3. **Demographics**
   population, percentage of population in races, median age(male/female/general),household size, marriage percentage, birth rate

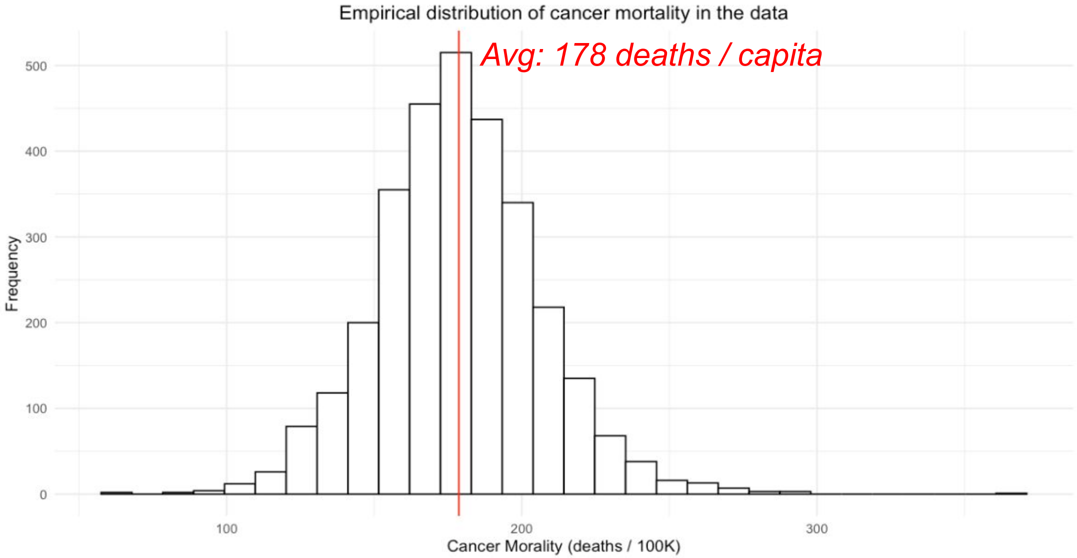
4. **Education**
   residents ages (18-24/25 above) with education level (less than high school/high school diploma/college/bachelor’s degree)

5. **Medical Coverage**
   percent of county residents with private alone/employee-provided/government-provided health coverage

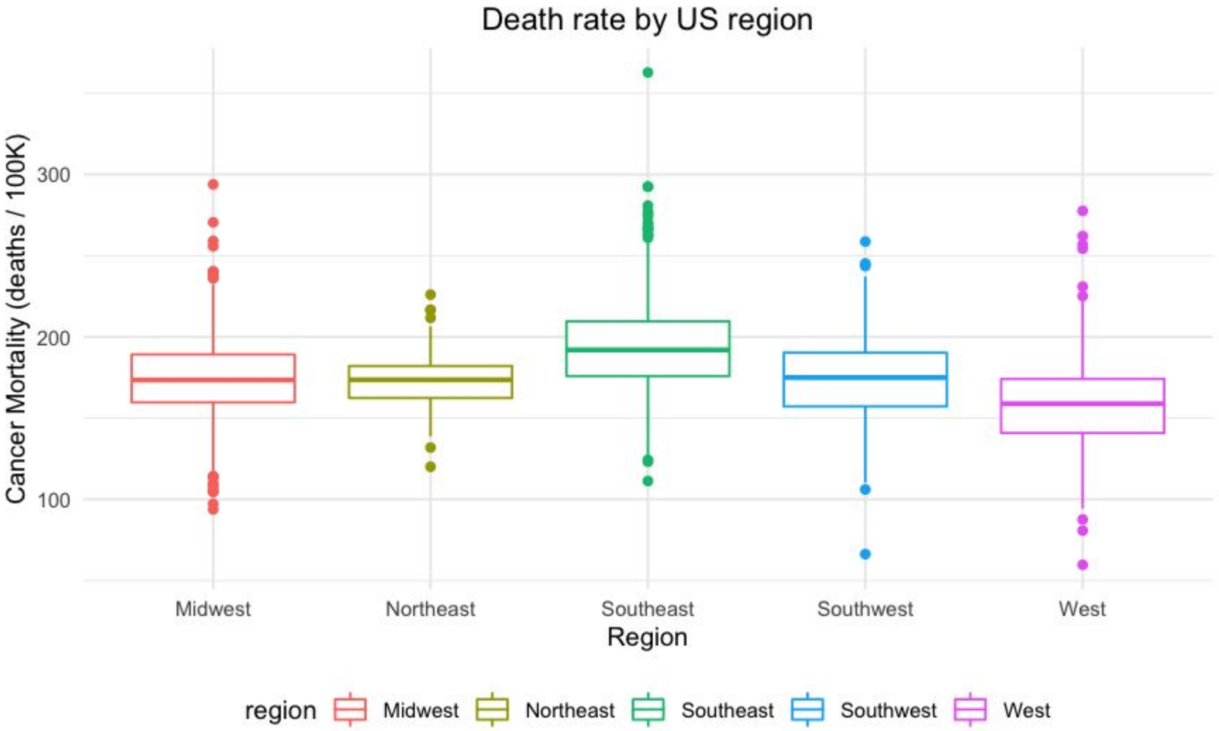
## Variable Types & Models

**Continuous:**cancer mortality,income,population,employment rate,birth rate…
**Categorical:**education**(ordinal),**medical coverage,ethnicity...
**Models:** Multiple regression, ANOVA, Bootstrap,PCA...

# Exploratory Data Analysis(EDA)

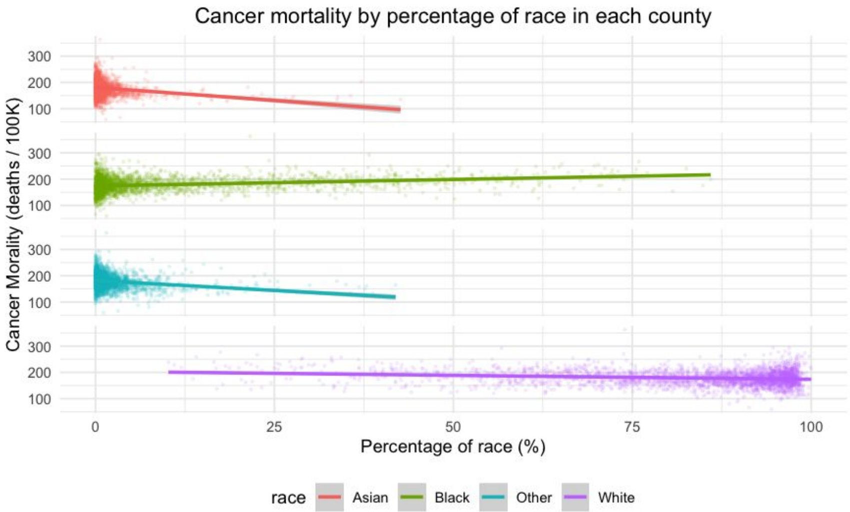


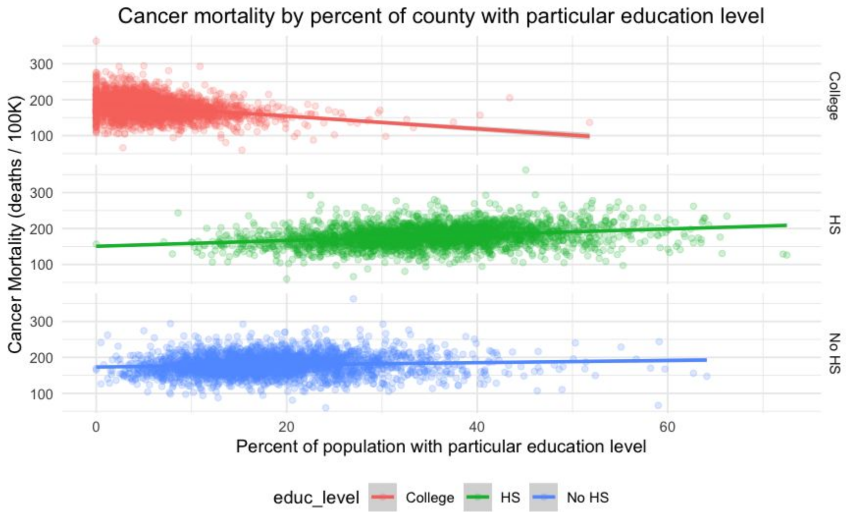
# EDA



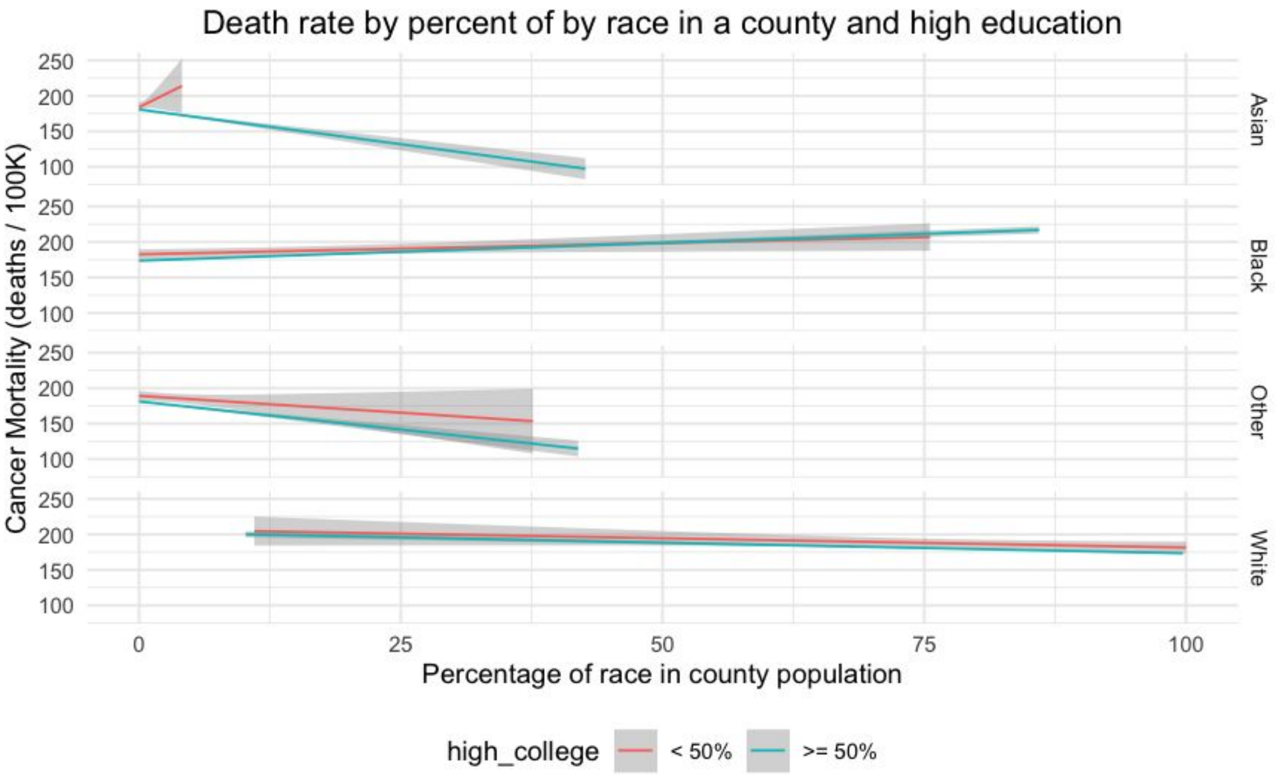
Death rate by US region
Cancer Mortality (deaths / 100K)
300
200
100
Midwest
Northeast
Southeast
Southwest
West
Region
region
Midwest
Northeast
Southeast
Southwest
West

# EDA


Cancer mortality by percentage of race in each county
Cancer Morality (deaths / 100K)
300
200
100
0
25
50
75
100
Percentage of race (%)
race
Asian
Black
Other
White


Cancer mortality by percent of county with particular education level
Cancer Morality (deaths / 100K)
300
200
100
College
HS
No HS
0
20
40
60
Percent of population with particular education level
educ_level
College
HS
No HS

# EDA


Death rate by percent of by race in a county and high education
Cancer Mortality (deaths / 100K)
250
200
150
100
Asian
Black
Other
White
0
25
50
75
100
Percentage of race in county population
high_college
< 50%
>= 50%

# Questions Of Interest

1. Is there a significant interaction between any demographic and economic factors that contributes to increased cancer mortality in a county?

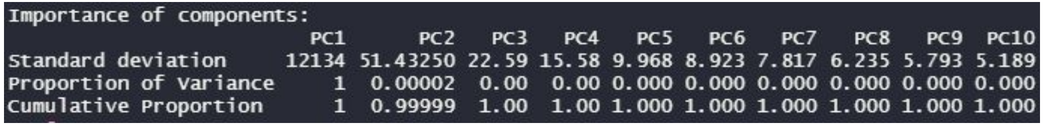
2. Can we use any significant findings from (1) that can help predict future cancer mortality in the counties?

# Proposed Inference Model

$$\overbrace{Y}^{\substack{\text{Cancer}\\ \text{Mortality}}} = \beta_0 + \overbrace{\beta_1 X_1}^{\substack{\text{Education}\\ \text{Rates}}} + \overbrace{\beta_2 X_2}^{\substack{\text{Demographic}\\ \text{Rates}}} + \underbrace{\beta_3}_{\substack{\text{Are these}\\ \text{interactions}\\ \text{significant?}}} \overbrace{X_1 X_2}^{\text{Interactions}} + \sum \overbrace{\beta_k \underbrace{X_k}_{\substack{\bullet\ \text{Cancer Incidence}\\ \bullet\ \text{Age}\\ \bullet\ \text{County Size}\\ \bullet\ \text{etc}}}}^{\text{Confounders}} + \epsilon$$

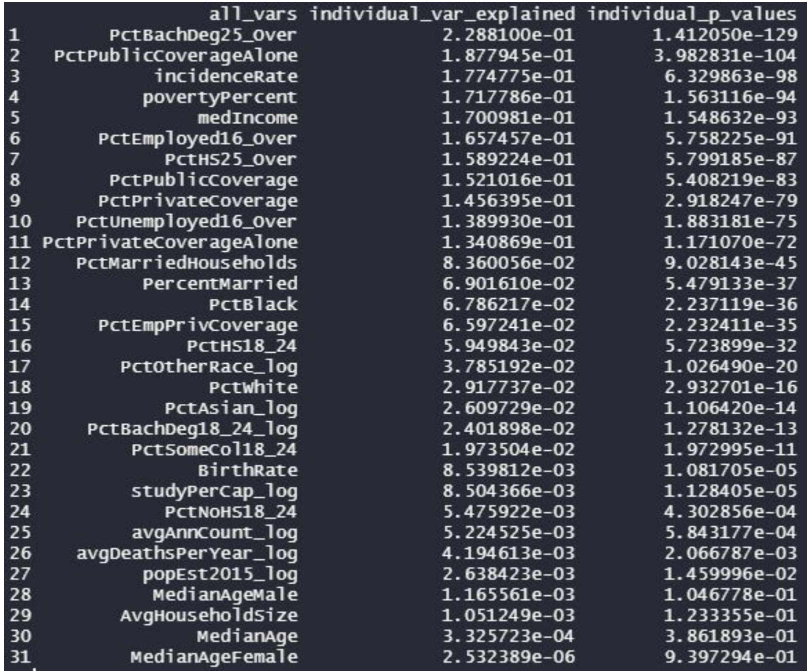
Null Hypothesis: All the coefficients for interaction are zero. (Wald Test)

# PCA



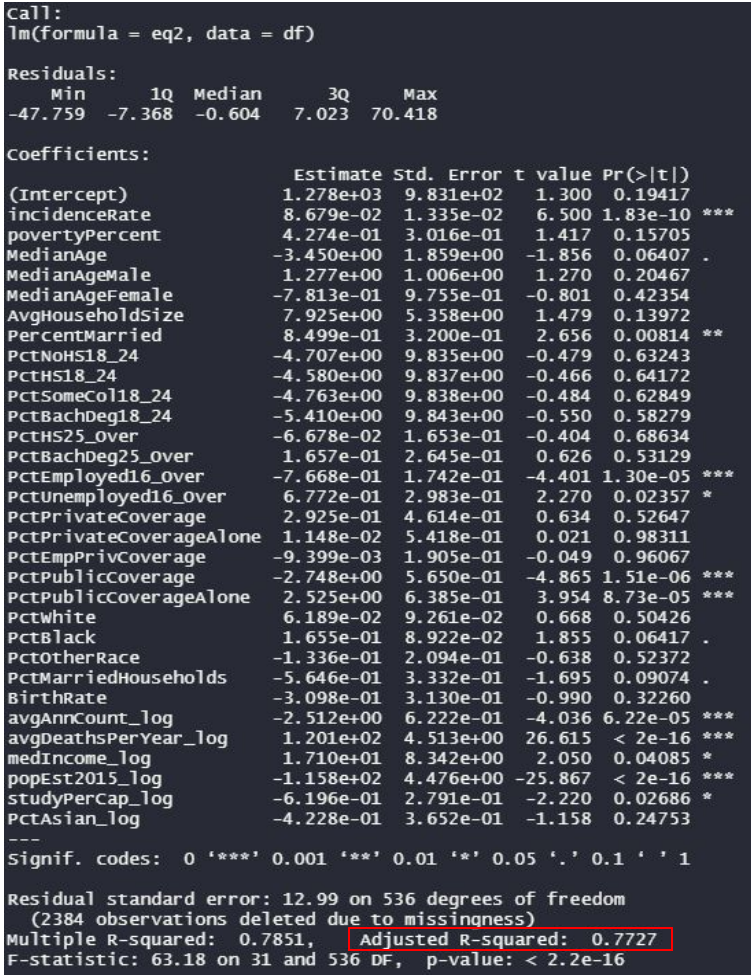
```
Importance of components:
                          PC1      PC2   PC3   PC4   PC5   PC6   PC7   PC8   PC9  PC10
Standard deviation      12134 51.43250 22.59 15.58 9.968 8.923 7.817 6.235 5.793 5.189
Proportion of Variance      1  0.00002  0.00  0.00 0.000 0.000 0.000 0.000 0.000 0.000
Cumulative Proportion       1  0.99999  1.00  1.00 1.000 1.000 1.000 1.000 1.000 1.000
```

# Individual Models



```
                 all_vars individual_var_explained individual_p_values
1        PctBachDeg25_Over             2.288100e-01        1.412050e-129
2   PctPublicCoverageAlone             1.877945e-01        3.982831e-104
3            incidenceRate             1.774775e-01         6.329863e-98
4           povertyPercent             1.717786e-01         1.563116e-94
5                medIncome             1.700981e-01         1.548632e-93
6       PctEmployed16_Over             1.657457e-01         5.758225e-91
7             PctHS25_Over             1.589224e-01         5.799185e-87
8        PctPublicCoverage             1.521016e-01         5.408219e-83
9       PctPrivateCoverage             1.456395e-01         2.918247e-79
10    PctUnemployed16_Over             1.389930e-01         1.883181e-75
11 PctPrivateCoverageAlone             1.340869e-01         1.171070e-72
12    PctMarriedHouseholds             8.360056e-02         9.028143e-45
13          PercentMarried             6.901610e-02         5.479133e-37
14                PctBlack             6.786217e-02         2.237119e-36
15      PctEmpPrivCoverage             6.597241e-02         2.232411e-35
16              PctHS18_24             5.949843e-02         5.723899e-32
17       PctOtherRace_log             3.785192e-02         1.026490e-20
18                PctWhite             2.917737e-02         2.932701e-16
19            PctAsian_log             2.609729e-02         1.106420e-14
20     PctBachDeg18_24_log             2.401898e-02         1.278132e-13
21         PctSomeCol18_24             1.973504e-02         1.972995e-11
22               BirthRate             8.539812e-03         1.081705e-05
23        studyPerCap_log             8.504366e-03         1.128405e-05
24            PctNoHS18_24             5.475922e-03         4.302856e-04
25        avgAnnCount_log             5.224525e-03         5.843177e-04
26    avgDeathsPerYear_log             4.194613e-03         2.066787e-03
27         popEst2015_log             2.638423e-03         1.459996e-02
28           MedianAgeMale             1.165561e-03         1.046778e-01
29        AvgHouseholdSize             1.051249e-03         1.233355e-01
30               MedianAge             3.325723e-04         3.861893e-01
31         MedianAgeFemale             2.532389e-06         9.397294e-01
```

# Full Model



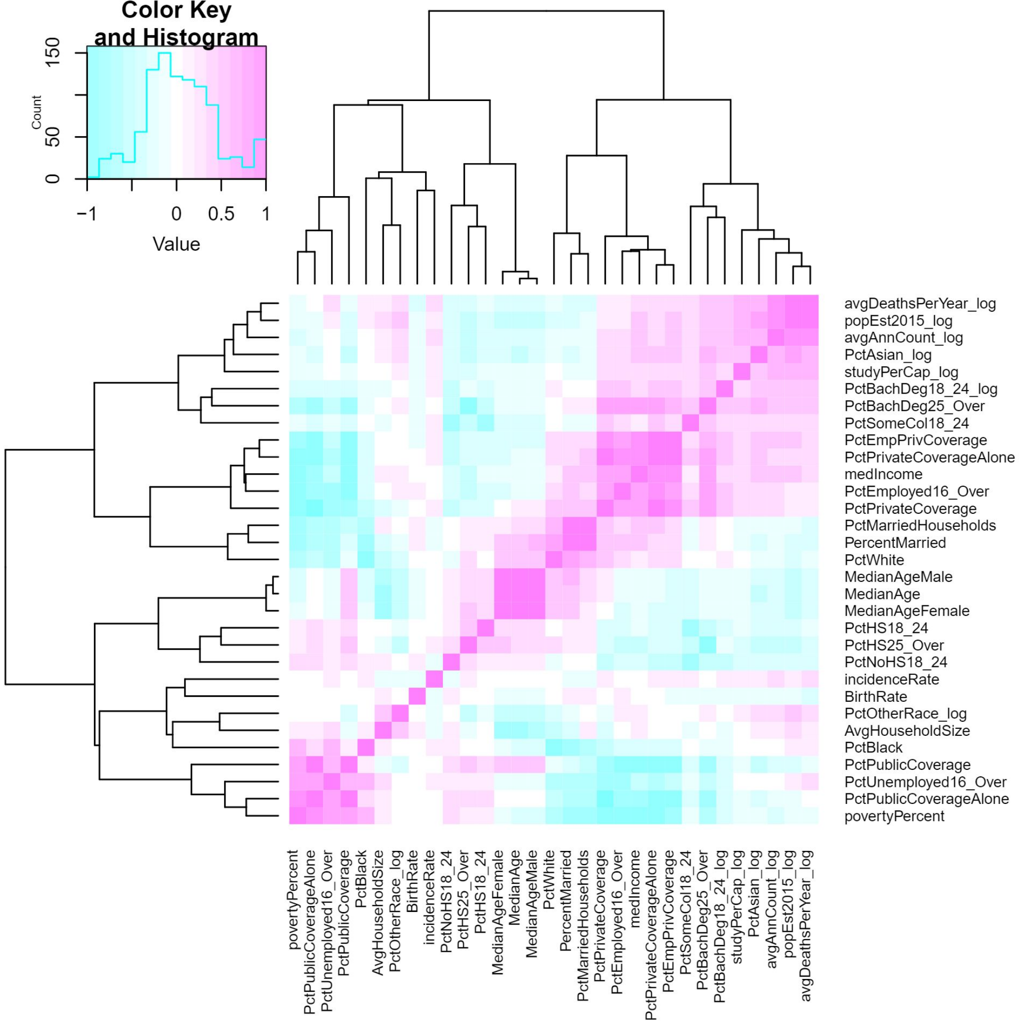
```
Call:
lm(formula = eq2, data = df)

Residuals:
    Min      1Q  Median      3Q     Max 
-47.759  -7.368  -0.604   7.023  70.418 

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)    
(Intercept)              1.278e+03  9.831e+02   1.300  0.19417    
incidenceRate            8.679e-02  1.335e-02   6.500 1.83e-10 ***
povertyPercent           4.274e-01  3.016e-01   1.417  0.15705    
MedianAge               -3.450e+00  1.859e+00  -1.856  0.06407 .  
MedianAgeMale            1.277e+00  1.006e+00   1.270  0.20467    
MedianAgeFemale         -7.813e-01  9.755e-01  -0.801  0.42354    
AvgHouseholdSize         7.925e+00  5.358e+00   1.479  0.13972    
PercentMarried           8.499e-01  3.200e-01   2.656  0.00814 ** 
PctNoHS18_24            -4.707e+00  9.835e+00  -0.479  0.63243    
PctHS18_24              -4.580e+00  9.837e+00  -0.466  0.64172    
PctSomeCol18_24         -4.763e+00  9.838e+00  -0.484  0.62849    
PctBachDeg18_24         -5.410e+00  9.843e+00  -0.550  0.58279    
PctHS25_Over            -6.678e-02  1.653e-01  -0.404  0.68634    
PctBachDeg25_Over        1.657e-01  2.645e-01   0.626  0.53129    
PctEmployed16_Over      -7.668e-01  1.742e-01  -4.401 1.30e-05 ***
PctUnemployed16_Over     6.772e-01  2.983e-01   2.270  0.02357 *  
PctPrivateCoverage       2.925e-01  4.614e-01   0.634  0.52647    
PctPrivateCoverageAlone  1.148e-02  5.418e-01   0.021  0.98311    
PctEmpPrivCoverage      -9.399e-03  1.905e-01  -0.049  0.96067    
PctPublicCoverage       -2.748e+00  5.650e-01  -4.865 1.51e-06 ***
PctPublicCoverageAlone   2.525e+00  6.385e-01   3.954 8.73e-05 ***
PctWhite                 6.189e-02  9.261e-02   0.668  0.50426    
PctBlack                 1.655e-01  8.922e-02   1.855  0.06417 .  
PctOtherRace            -1.336e-01  2.094e-01  -0.638  0.52372    
PctMarriedHouseholds    -5.646e-01  3.332e-01  -1.695  0.09074 .  
BirthRate               -3.098e-01  3.130e-01  -0.990  0.32260    
avgAnnCount_log         -2.512e+00  6.222e-01  -4.036 6.22e-05 ***
avgDeathsPerYear_log     1.201e+02  4.513e+00  26.615  < 2e-16 ***
medIncome_log            1.710e+01  8.342e+00   2.050  0.04085 *  
popEst2015_log          -1.158e+02  4.476e+00 -25.867  < 2e-16 ***
studyPerCap_log         -6.196e-01  2.791e-01  -2.220  0.02686 *  
PctAsian_log            -4.228e-01  3.652e-01  -1.158  0.24753    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 12.99 on 536 degrees of freedom
  (2384 observations deleted due to missingness)
Multiple R-squared:  0.7851,	Adjusted R-squared:  0.7727 
F-statistic: 63.18 on 31 and 536 DF,  p-value: < 2.2e-16
```

# Correlation Heatmap



# Reference

1. Tiwari RC, Ghosh K, Jemal A et al.. A new method of predicting US and state- level cancer mortality counts for the current calendar year. CA Cancer J Clin 2004; 54: 30–40.

2. Jemal A, Siegel R, Ward E et al.. Cancer statistics, 2009. CA Cancer J Clin 2009; 59: 225–249.

3. Chen HS, Portier K, Ghosh K et al.. Predicting US- and state-level cancer counts for the current calendar year: part I: evaluation of temporal projection methods for mortality. Cancer 2012; 118: 1091–1099.