STAT 350: Final Project Report

STAT 350 D100

Tyler Oh #301320847

Simon Fraser University

Abstract Summary

This report summarizes the statistical analysis for the death rate of invasive cancer prediction. The ultimate objective is to fit an appropriate regression model and extract the factors that contribute to the death rate of the cancer. There are two large datasets of the incidence rate and death rate of invasive cancer containing up to 3141 observations each excluding the Nevada region. Also, one more dataset called US county data will be used for a better prediction of the death rate.

Introduction

There are bunch of risk factors for invasive cancer such as advancing age, family history, or stress. Among all factors, I was interested in the effects of stress on cancer. I was beginning to wonder how economic, mental, and physical differences by County affect the death rate of the cancer.

Data description

The cancer linear regression dataset has been collected by Noah Rippner based on the US census population from the year 1969 – 2014, NCI (http://seer.cancer.gov). The two individual datasets contain up to 3141 observations excluding the Nevada region.

Besides, US county data based on US population by county from year by Emil O. W. Kirkegaard (https://github.com/Deleetdk/USA.county.data) has been added for the prediction. It contains social and economic data and up to 3141 observations by County and FIPS. The dataset has no observations for the year 2008-2012 and for Alaska since it is considered as a state. I have chosen 13 variables from this dataset which may be correlated to the incidence and death of invasive cancer.

Brief look of the variables

```
$ FIPS                                                : int [1:2469] 1107 1007 1079 1115 1109 1091 1069 1003 1021 1075 ...
$ Less.Than.High.School                               : num [1:2469] 21.3 25.5 24.8 20.7 20.6 20.3 18.1 12.4 24.1 24.7 ...
$ At.Least.High.School.Diploma                        : num [1:2469] 78.7 74.5 75.2 79.3 79.4 79.7 81.9 87.6 75.9 75.3 ...
$ At.Least.Bachelor.s.Degree                          : num [1:2469] 11.5 10 10.7 14.5 23.7 17.9 19 26.8 12.2 9.2 ...
$ Graduate.Degree                                     : num [1:2469] 3.4 2.6 3.3 4.7 9.3 5.6 6.8 8.7 4.8 3.3 ...
$ School.Enrollment                                   : num [1:2469] 74.2 67.3 72 73.2 82 ...
$ Adults.65.and.Older.Living.in.Poverty               : num [1:2469] 22.1 12.4 11.7 9.8 16.2 ...
$ Poverty.Rate.below.federal.poverty.threshold        : num [1:2469] 26.6 12.2 13 10.9 28.9 ...
$ Child.Poverty.living.in.families.below.the.poverty.line   : num [1:2469] 36.1 17.9 17 14.2 37.4 ...
$ Management.professional.and.related.occupations     : num [1:2469] 20.2 20.2 22.1 27.9 26.4 ...
$ Service.occupations                                 : num [1:2469] 15.2 13.6 15.5 14 20.2 ...
$ Sales.and.office.occupations                        : num [1:2469] 25.8 22.6 21.7 26.4 24.5 ...
$ Construction.extraction.maintenance.and.repair.occupations: num [1:2469] 12.1 18.05 15.4 13.2 9.75 ...
$ Production.transportation.and.material.moving.occupations : num [1:2469] 25.3 25 24.6 18.1 17.9 ...
$ Poor.physical.health.days                           : num [1:2469] 5.2 4.7 4.6 4.2 4.7 4.8 4.5 3.3 4.6 7.1 ...
$ Poor.mental.health.days                             : num [1:2469] 3.3 5.1 5.5 4 3.8 4.5 4.3 3.8 4.9 6 ...
$ Adult.smoking                                       : num [1:2469] 0.181 0.259 0.278 0.272 0.183 0.168 0.184 0.206 0.201 0.261 ...
$ Adult.obesity                                       : num [1:2469] 0.385 0.343 0.377 0.329 0.36 0.416 0.349 0.266 0.383 0.313 ...
```

Figure1: String of selected variables

Checking correlation between each variable in the same categories

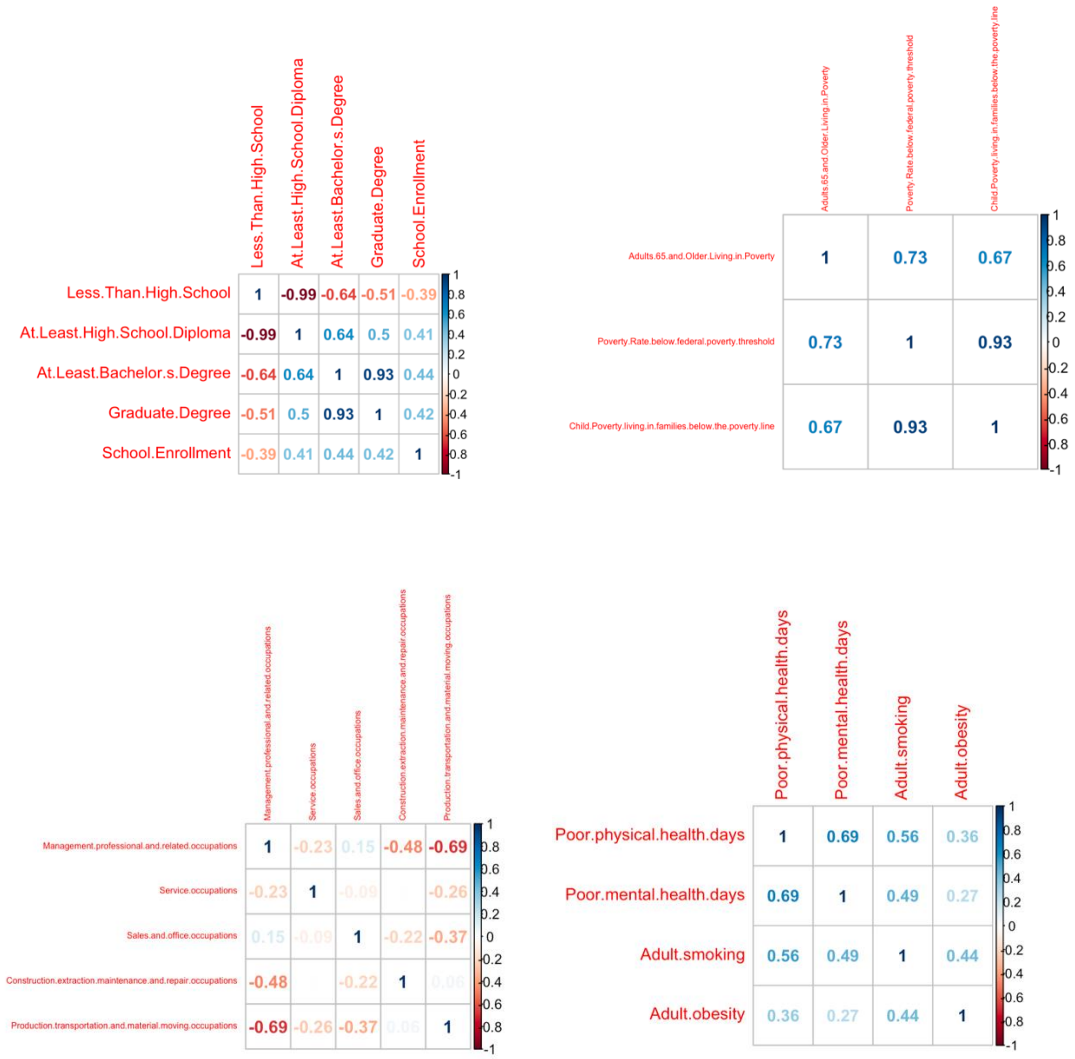


Figure2: Correlation plot of each variable within the same categories

In the education background category, It shows a high correlation between the variables "Graduate.Degree" and "At.Least.Bachelor.s.Degree" and "Less.Than.High.School" and "At.Least.High.School.Diploma". We can remove the variables "Graduate.Degree" and "At.Least.High.School.Diploma". Also, "Poverty.Rate.below.federal.poverty.threshold" and "Child.Poverty.living.in.families.below.the.poverty.line" are highly correlated. There is no reason to keep all the variables. Hence, we will drop "Poverty.Rate.below.federal.poverty.threshold".

The starting variables that the datasets provide contained Death rate(y), Incidence rate (x1), Average death per year (x2), Less than high school (x3), At least Bachelor's degree (x4), School enrollment (x5), 65 and older people living in poverty (x6), Child living in below the poverty line (x7), Management professional and related (x8), Service (x9), Sales and office (x10), Construction, maintenance and repair (x11), Production and transportation (x12), Poor physical health days (x13), Poor mental health days

(x14), Adult Smoking rate (x15), Adult obesity rate (x16), where all the variables excluding Average death per year are rates by US counties population.

Histogram and Boxplot

We will see each variables' histogram and boxplot show any outliers or special distribution which can affect the further model.

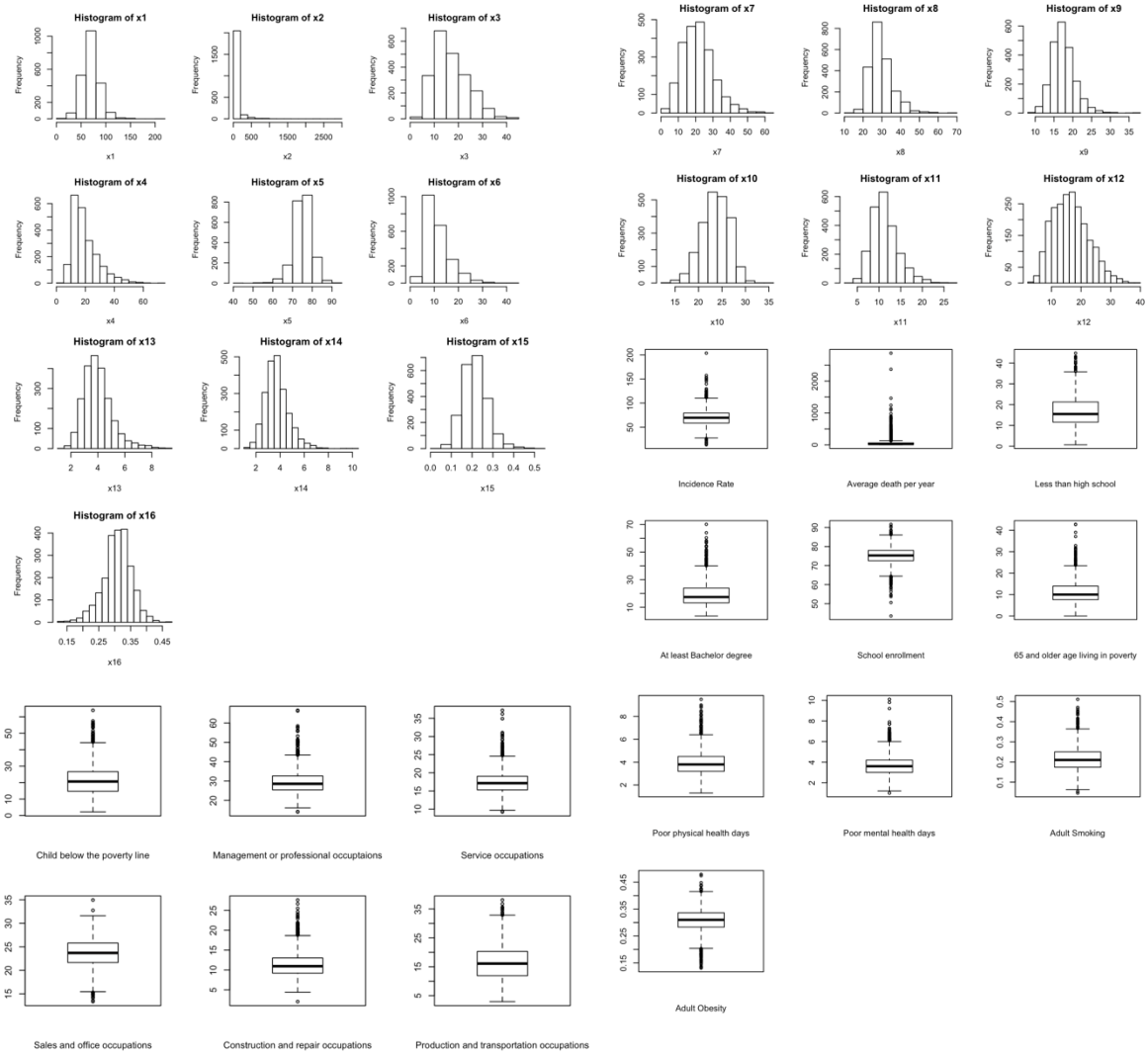


Figure 3: The histogram and the boxplot of all response variables

The histogram and the boxplot of variable x2 do not show normal distribution, and it is too skewed. However, although the histogram and the boxplot of each of the other variables have outliers, it would not affect the further model since most of the outliers are on the large-side.

Methods

Boxplot and Histogram: The boxplot and histogram are a great way to see if there is any special distribution shape or outliers which can affect our model. Thus, I would apply this method to find the unusual variable

Correlation Plot: This method can be applied to find the variables that contain similar data or meaning. It is also viewed to identify for potential multicollinearity.

Transformation: From this, we can change the distribution of usual shape or increase adjusted R-squared.

Residual Analysis: The main purpose if to check for linearity, constant variance, potential outliers, etc, for fitted model.

Multicollinearity Check: After only observing the correlations among pairs of variables, we must check if there are inflation factors which may affect our further model.

Variable Selection: There are many purposes to use stepwise selection such as reducing AIC, excluding insignificant variables, etc.

Result

Initial Model (without any changes)

```
Call:
lm(formula = y ~ x1 + x2 + x3 + x4 + x5 + x6 + x7 + x8 + x9 +
    x10 + x11 + x12 + x13 + x14 + x15 + x16)

Residuals:
    Min      1Q  Median      3Q     Max
-36.068  -3.180  -0.251   3.115  36.688

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.054115   7.651026   0.530   0.5962
x1           0.581535   0.009963  58.369  < 2e-16 ***
x2          -0.001245   0.001002  -1.242   0.2143
x3           0.040416   0.035383   1.142   0.2535
x4          -0.030217   0.036846  -0.820   0.4123
x5          -0.050271   0.032557  -1.544   0.1227
x6          -0.005669   0.038988  -0.145   0.8844
x7           0.031205   0.024004   1.300   0.1937
x8          -0.001726   0.093756  -0.018   0.9853
x9          -0.228211   0.090724  -2.515   0.0120 *
x10         -0.061679   0.081716  -0.755   0.4505
x11          0.046370   0.088944   0.521   0.6022
x12         -0.009474   0.081477  -0.116   0.9074
x13          0.863589   0.185627   4.652 3.48e-06 ***
x14          0.404900   0.178258   2.271   0.0232 *
x15         19.760797   3.105279   6.364 2.39e-10 ***
x16         25.483718   4.106789   6.205 6.51e-10 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.849 on 2182 degrees of freedom
Multiple R-squared:  0.8293,    Adjusted R-squared:  0.8281
F-statistic: 662.7 on 16 and 2182 DF,  p-value: < 2.2e-16
```

The initial model has Standard Error (5.849), Adjusted R-squared (0.8281), F-statistic (6*62.7*), and p-value (< 2.2e-16) All independent variables are not yet significant with set alpha (0.05)

According to the initial model, approximately 83% of the variability of death rate is explained through this regression.

Figure 4: Summary of lm(initial model)

Transformation

Since the histogram and the boxplot does not show any normal distribution. We will take logarithm of the variable x2 and interpret the results

```
Call:
lm(formula = y ~ x1 + log(x2) + x3 + x4 + x5 + x6 + x7 + x8 +
    x9 + x10 + x11 + x12 + x13 + x14 + x15 + x16)

Residuals:
    Min      1Q  Median      3Q     Max
-35.623  -3.286  -0.289   3.183  36.980

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 10.87632    7.82300   1.390  0.16458
x1           0.57733    0.01005  57.454  < 2e-16 ***
log(x2)      0.44611    0.15416   2.894  0.00384 **
x3           0.00649    0.03608   0.180  0.85725
x4          -0.04944    0.03719  -1.329  0.18389
x5          -0.04822    0.03251  -1.483  0.13816
x6           0.03071    0.04049   0.758  0.44831
x7           0.02483    0.02402   1.034  0.30134
x8          -0.07631    0.09526  -0.801  0.42320
x9          -0.29429    0.09212  -3.195  0.00142 **
x10         -0.19485    0.08863  -2.199  0.02801 *
x11          0.01545    0.08923   0.173  0.86259
x12         -0.07827    0.08341  -0.938  0.34811
x13          0.86830    0.18530   4.686 2.96e-06 ***
x14          0.34734    0.17860   1.945  0.05193 .
x15         20.64316    3.10385   6.651 3.67e-11 ***
x16         26.70124    4.08649   6.534 7.94e-11 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.839 on 2182 degrees of freedom
Multiple R-squared:  0.8299,    Adjusted R-squared:  0.8286
F-statistic: 665.2 on 16 and 2182 DF,  p-value: < 2.2e-16
```

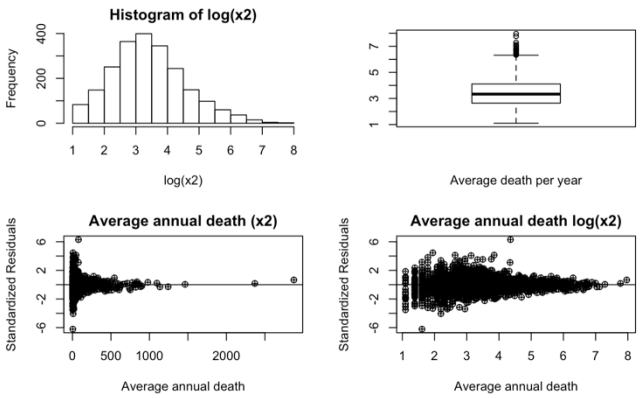


Figure 5: Summary, the histogram, the boxplot and residual plot of the variable log(x2)

We can see that F-statistic and adjusted R-squared has slightly increased. However, x2 has become significant variable by taking logarithm. Moreover, the histogram indicate normal distribution, and the dots are nicely spread compared to the plot before taking logarithm on x2. Our improved model shows many insignificant variables ($p > 0.05$).

Variable Selection

All three variable selection methods will be used to compare AIC (Akaike Information Criterion) value as it estimates the quality of each model.

Forward selection: The variables Less.than.high.school (x3), 65.and.older.living.in.poverty (x6), Service occupation (x8), Poor.physical.health.days (x12) has not been selected.

```
Step:  AIC=7773.1
y ~ x1 + x4 + x13 + x16 + x15 + x9 + x3 + x14 + x5 + x10 + log(x2) +
    x11 + x7
```

Backward Elimination: The model without x3, x6, x11 provides the AIC value of 7772.76 which is lower than the previous model.

```
Step:  AIC=7772.76
y ~ x1 + log(x2) + x4 + x5 + x7 + x8 + x9 + x10 + x12 + x13 +
    x14 + x15 + x16
```

Stepwise Selection: Stepwise gives the same result as forward selection.

```
Step:  AIC=7771.62
y ~ x1 + x4 + x13 + x16 + x15 + x9 + x14 + x5 + x10 + log(x2) +
    x11 + x7
```

```
Call:
lm(formula = y ~ x1 + x4 + x13 + x16 + x15 + x9 + x14 + x5 +
    x10 + log(x2) + x11 + x7, data = selected_data)

Residuals:
    Min      1Q  Median      3Q     Max
-35.661  -3.302  -0.315   3.168  37.096

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.853258   3.293495   1.474  0.14074
x1           0.576425   0.009905  58.194  < 2e-16 ***
x4          -0.056240   0.024389  -2.306  0.02121 *
x13          0.887160   0.181626   4.885 1.11e-06 ***
x16         26.531680   4.034532   6.576 6.02e-11 ***
x15         19.966972   3.051579   6.543 7.48e-11 ***
x9          -0.225580   0.043685  -5.164 2.64e-07 ***
x14          0.333417   0.177879   1.874  0.06101 .
x5          -0.052825   0.032020  -1.650  0.09914 .
x10         -0.131961   0.050536  -2.611  0.00908 **
log(x2)      0.398281   0.143128   2.783  0.00544 **
x11          0.085286   0.049075   1.738  0.08237 .
x7           0.040652   0.019138   2.124  0.03377 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.837 on 2186 degrees of freedom
Multiple R-squared:  0.8297,    Adjusted R-squared:  0.8288
F-statistic: 887.7 on 12 and 2186 DF,  p-value: < 2.2e-16
```

```
Call:
lm(formula = y ~ x1 + x4 + x13 + x16 + x15 + x9 + x10 + log(x2) +
    x11 + x7, data = selected_data)

Residuals:
    Min      1Q  Median      3Q     Max
-35.556  -3.287  -0.310   3.118  36.942

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.036188   2.218355   0.467  0.64048
x1           0.578070   0.009885  58.481  < 2e-16 ***
x4          -0.066409   0.023807  -2.789  0.00533 **
x13          1.049404   0.160257   6.548 7.23e-11 ***
x16         25.587036   4.016488   6.371 2.29e-10 ***
x15         20.752452   3.028572   6.852 9.42e-12 ***
x9          -0.217454   0.043426  -5.007 5.96e-07 ***
x10         -0.131608   0.050547  -2.604  0.00929 **
log(x2)      0.436894   0.141956   3.078  0.00211 **
x11          0.105859   0.047823   2.214  0.02696 *
x7           0.046536   0.018998   2.450  0.01438 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.842 on 2188 degrees of freedom
Multiple R-squared:  0.8293,    Adjusted R-squared:  0.8285
F-statistic:  1063 on 10 and 2188 DF,  p-value: < 2.2e-16
```

However, there are still insignificant variables that their p-value is bigger than set alpha (0.05). Thus, elimination is still needed. As a result, the variables x5 (School.enrollment) and x14 (Adult.smoking.rate) will be deleted. Our improved model has increased the F-statistics.

Model Adequacy Checking

We will use residuals to check if our initial regression model is adequate or not based on the following hypothesis properties:

1. The relationship between the predictor variable and response variables is linear, at least approximately
2. The error term has a zero mean
3. The error term has a constant variance.
4. The errors are correlated.
5. The errors are normally distributed.

Residual Analysis

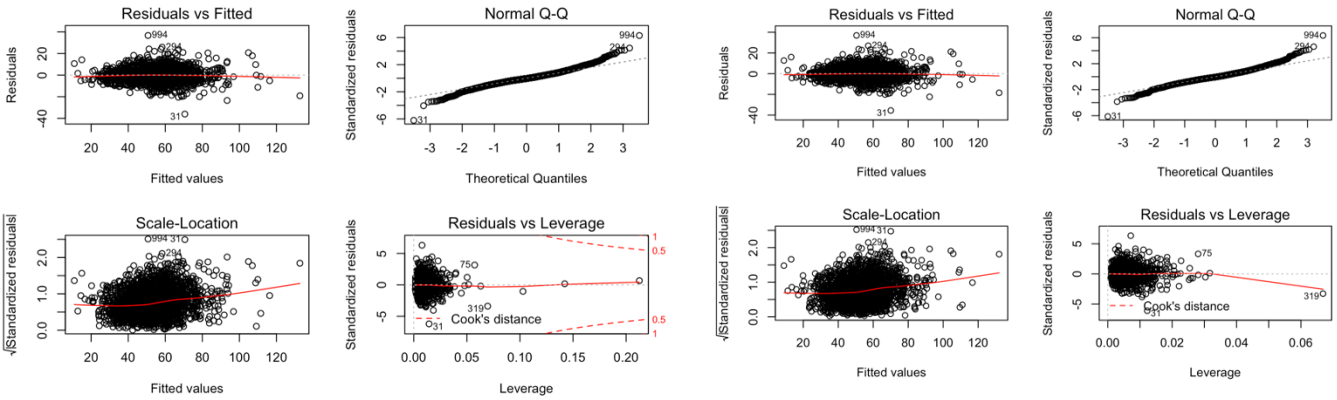


Figure 5: Residual plots for initial model and improved model

As shown in figure 5, there is no big difference between the two model in terms of residual analysis. However, we can find that few outliers in Residual vs Leverage plot has been deleted except 75th, 319th. If cook's distance of a potential outlier is bigger than 1, it should be deleted. Although the outliers in our model are too small, I will delete more accurate model. Also, the normality probability plots do not indicate any significant violations of normality. Hence, the assumption of normality of the residuals are satisfied.

Multicollinearity Check

```
      x1       x4       x13      x16      x15       x9      x14       x5      x10  log(x2)      x11       x7
1.945004 3.165927 2.491316 2.126363 2.146029 1.201725 1.995588 1.418150 1.446111 1.761739 1.540162 1.981208
      x1       x4       x13      x16      x15       x9      x10  log(x2)      x11       x7
1.933359 3.011003 1.935994 2.103483 2.109878 1.185295 1.444057 1.729791 1.459910 1.948659
```

Figure 6: VIFs of the fitted model after stepwise selection and own improved model

As seen in figure 6, Getting rid of the two variables x5 and x14 has brought down other VIFs values down from range 0.01 to 0.5. However, we could find the decrease of the F-statistics even though the change of VIFs values is too small.

Conclusion

The best fitted model is y-hat = 1.0362 + 0.5781*x1 - 0.4369*log(x2) + 0.0664*x4 0.0465*x7 - 0.2175*x9 - 0.1316*x10 + 0.105859*x11 + 1.0494*x13 +20.7525*x15+25.5870*x16 with Standard Error (5.842), Adjusted R-squared (0.8285), F-statistic (1063), and p-value (< 2.2e-16). All intendents variables are significant with set alpha (0.05). According to the final model, approximately 83% of the variability of death rate is explained through this regression.

As shown in the figure 7, the death rate is related to with factors in the order incidence rate, adult smoking rate and poor physical health days. ## Contrary to my expectations, there was no strong correlation between occupations or educational background with cancer incidence and mortality. However, we can notice that smoking and obesity contribute to invasive cancer mortality.

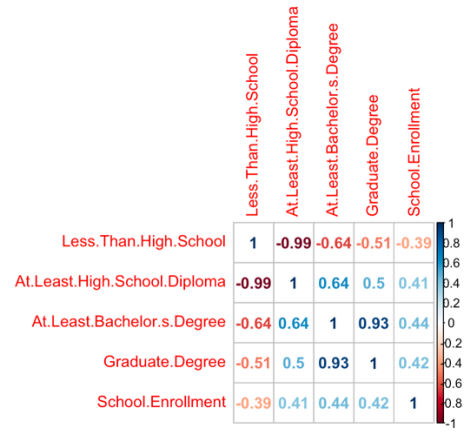
Figure 7: Correlation plot of final dataset

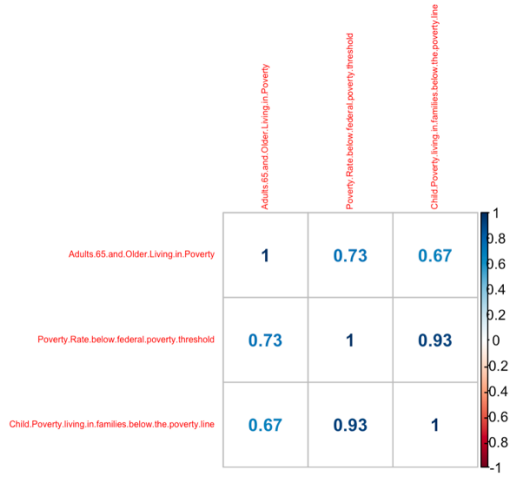
Appendix

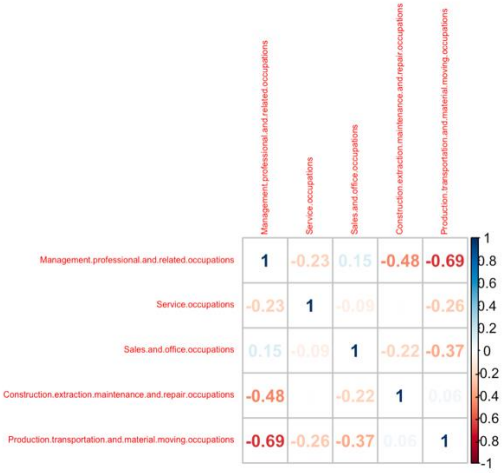
String of selected variables

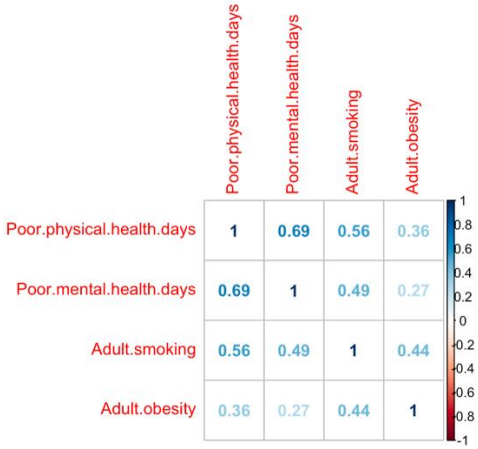
```
$ FIPS                                                  : int [1:2469] 1107 1007 1079 1115 1109 1091 1069 1003 1021 1075 ...
$ Less.Than.High.School                                 : num [1:2469] 21.3 25.5 24.8 20.7 20.6 20.3 18.1 12.4 24.1 24.7 ...
$ At.Least.High.School.Diploma                          : num [1:2469] 78.7 74.5 75.2 79.3 79.4 79.7 81.9 87.6 75.9 75.3 ...
$ At.Least.Bachelor.s.Degree                            : num [1:2469] 11.5 10 10.7 14.5 23.7 17.9 19 26.8 12.2 9.2 ...
$ Graduate.Degree                                       : num [1:2469] 3.4 2.6 3.3 4.7 9.3 5.6 6.8 8.7 4.8 3.3 ...
$ School.Enrollment                                     : num [1:2469] 74.2 67.3 72 73.2 82 ...
$ Adults.65.and.Older.Living.in.Poverty                 : num [1:2469] 22.1 12.4 11.7 9.8 16.2 ...
$ Poverty.Rate.below.federal.poverty.threshold          : num [1:2469] 26.6 12.2 13 10.9 28.9 ...
$ Child.Poverty.living.in.families.below.the.poverty.line   : num [1:2469] 36.1 17.9 17 14.2 37.4 ...
$ Management.professional.and.related.occupations       : num [1:2469] 20.2 20.2 22.1 27.9 26.4 ...
$ Service.occupations                                   : num [1:2469] 15.2 13.6 15.5 14 20.2 ...
$ Sales.and.office.occupations                          : num [1:2469] 25.8 22.6 21.7 26.4 24.5 ...
$ Construction.extraction.maintenance.and.repair.occupations: num [1:2469] 12.1 18.05 15.4 13.2 9.75 ...
$ Production.transportation.and.material.moving.occupations : num [1:2469] 25.3 25 24.6 18.1 17.9 ...
$ Poor.physical.health.days                             : num [1:2469] 5.2 4.7 4.6 4.2 4.7 4.8 4.5 3.3 4.6 7.1 ...
$ Poor.mental.health.days                               : num [1:2469] 3.3 5.1 5.5 4 3.8 4.5 4.3 3.8 4.9 6 ...
$ Adult.smoking                                         : num [1:2469] 0.181 0.259 0.278 0.272 0.183 0.168 0.184 0.206 0.201 0.261 ...
$ Adult.obesity                                         : num [1:2469] 0.385 0.343 0.377 0.329 0.36 0.416 0.349 0.266 0.383 0.313 ...
```

Correlation plot of each variable within the same categories

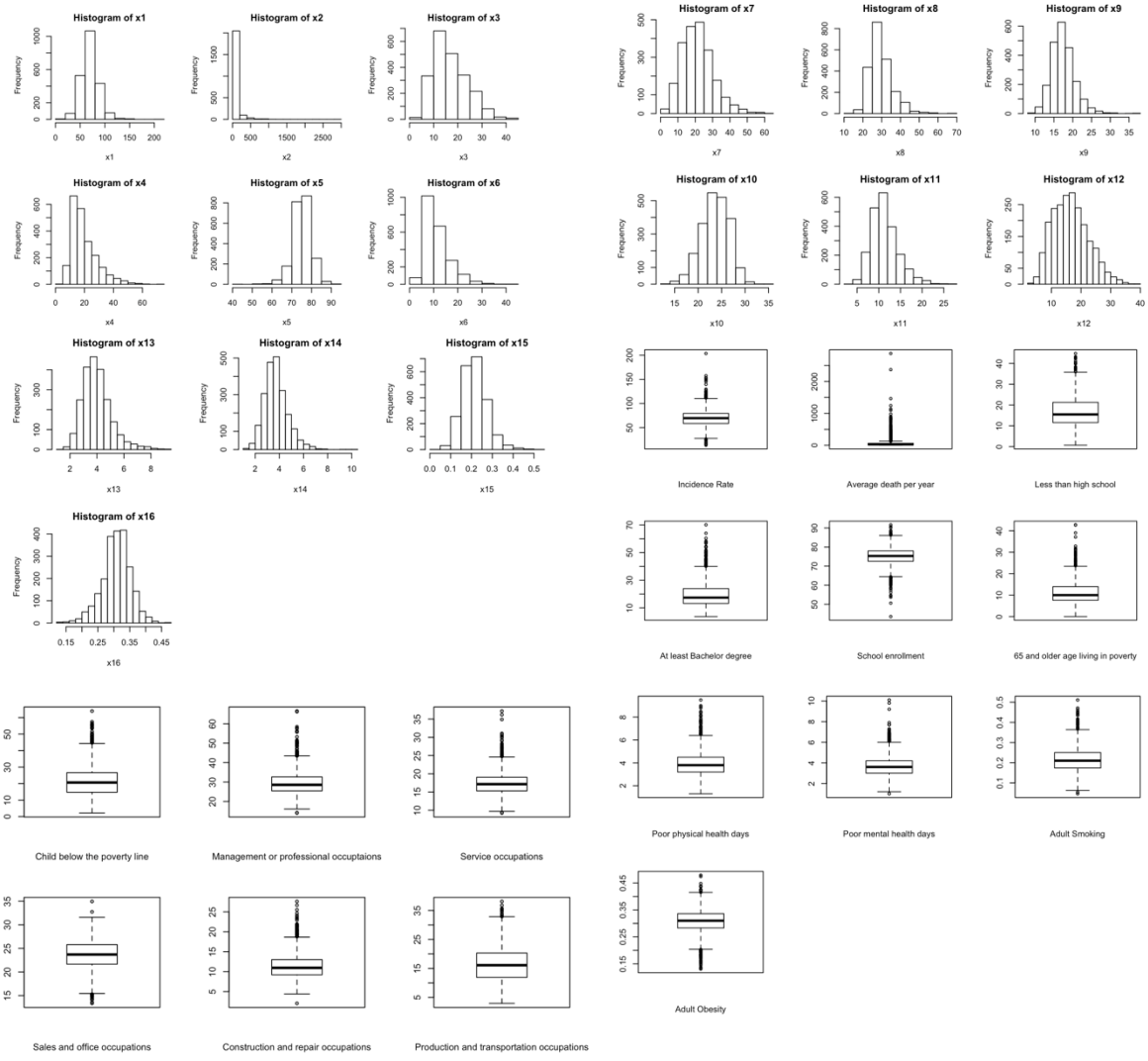








The histogram and the boxplot of all response variables



The summary of Initial Model (without any changes)

```
Call:
lm(formula = y ~ x1 + x2 + x3 + x4 + x5 + x6 + x7 + x8 + x9 +
    x10 + x11 + x12 + x13 + x14 + x15 + x16)

Residuals:
    Min      1Q  Median      3Q     Max
-36.068  -3.180  -0.251   3.115  36.688

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.054115   7.651026   0.530   0.5962
x1           0.581535   0.009963  58.369  < 2e-16 ***
x2          -0.001245   0.001002  -1.242   0.2143
x3           0.040416   0.035383   1.142   0.2535
x4          -0.030217   0.036846  -0.820   0.4123
x5          -0.050271   0.032557  -1.544   0.1227
x6          -0.005669   0.038988  -0.145   0.8844
x7           0.031205   0.024004   1.300   0.1937
x8          -0.001726   0.093756  -0.018   0.9853
x9          -0.228211   0.090724  -2.515   0.0120 *
x10         -0.061679   0.081716  -0.755   0.4505
x11          0.046370   0.088944   0.521   0.6022
x12         -0.009474   0.081477  -0.116   0.9074
x13          0.863589   0.185627   4.652 3.48e-06 ***
x14          0.404900   0.178258   2.271   0.0232 *
x15         19.760797   3.105279   6.364 2.39e-10 ***
x16         25.483718   4.106789   6.205 6.51e-10 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.849 on 2182 degrees of freedom
Multiple R-squared:  0.8293,    Adjusted R-squared:  0.8281
F-statistic: 662.7 on 16 and 2182 DF,  p-value: < 2.2e-16
```

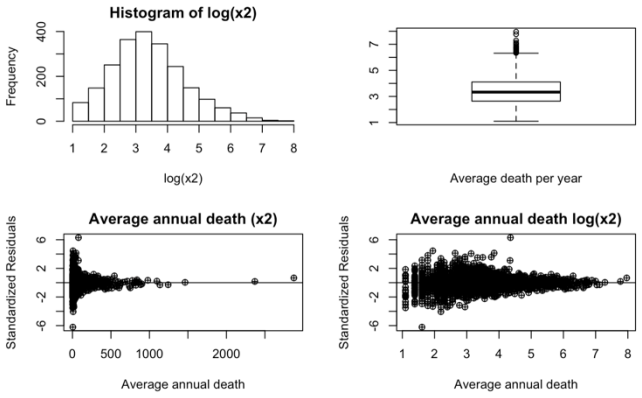
The summary of fitted model (taking logarithm on x2)

```
Call:
lm(formula = y ~ x1 + log(x2) + x3 + x4 + x5 + x6 + x7 + x8 +
    x9 + x10 + x11 + x12 + x13 + x14 + x15 + x16)

Residuals:
    Min      1Q  Median      3Q     Max
-35.623  -3.286  -0.289   3.183  36.980

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 10.87632    7.82300   1.390  0.16458
x1           0.57733    0.01005  57.454  < 2e-16 ***
log(x2)      0.44611    0.15416   2.894  0.00384 **
x3           0.00649    0.03608   0.180  0.85725
x4          -0.04944    0.03719  -1.329  0.18389
x5          -0.04822    0.03251  -1.483  0.13816
x6           0.03071    0.04049   0.758  0.44831
x7           0.02483    0.02402   1.034  0.30134
x8          -0.07631    0.09526  -0.801  0.42320
x9          -0.29429    0.09212  -3.195  0.00142 **
x10         -0.19485    0.08863  -2.199  0.02801 *
x11          0.01545    0.08923   0.173  0.86259
x12         -0.07827    0.08341  -0.938  0.34811
x13          0.86830    0.18530   4.686 2.96e-06 ***
x14          0.34734    0.17860   1.945  0.05193 .
x15         20.64316    3.10385   6.651 3.67e-11 ***
x16         26.70124    4.08649   6.534 7.94e-11 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.839 on 2182 degrees of freedom
Multiple R-squared:  0.8299,    Adjusted R-squared:  0.8286
F-statistic: 665.2 on 16 and 2182 DF,  p-value: < 2.2e-16
```



The results and summary of stepwise regression

```
Step:  AIC=7773.1
y ~ x1 + x4 + x13 + x16 + x15 + x9 + x3 + x14 + x5 + x10 + log(x2) +
    x11 + x7


Step:  AIC=7772.76
y ~ x1 + log(x2) + x4 + x5 + x7 + x8 + x9 + x10 + x12 + x13 +
    x14 + x15 + x16


Step:  AIC=7771.62
y ~ x1 + x4 + x13 + x16 + x15 + x9 + x14 + x5 + x10 + log(x2) +
    x11 + x7
```

```
Call:
lm(formula = y ~ x1 + x4 + x13 + x16 + x15 + x9 + x14 + x5 +
    x10 + log(x2) + x11 + x7, data = selected_data)

Residuals:
    Min      1Q  Median      3Q     Max
-35.661  -3.302  -0.315   3.168  37.096

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.853258   3.293495   1.474  0.14074
x1           0.576425   0.009905  58.194  < 2e-16 ***
x4          -0.056240   0.024389  -2.306  0.02121 *
x13          0.887160   0.181626   4.885 1.11e-06 ***
x16         26.531680   4.034532   6.576 6.02e-11 ***
x15         19.966972   3.051579   6.543 7.48e-11 ***
x9          -0.225580   0.043685  -5.164 2.64e-07 ***
x14          0.333417   0.177879   1.874  0.06101 .
x5          -0.052825   0.032020  -1.650  0.09914 .
x10         -0.131961   0.050536  -2.611  0.00908 **
log(x2)      0.398281   0.143128   2.783  0.00544 **
x11          0.085286   0.049075   1.738  0.08237 .
x7           0.040652   0.019138   2.124  0.03377 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.837 on 2186 degrees of freedom
Multiple R-squared:  0.8297,    Adjusted R-squared:  0.8288
F-statistic: 887.7 on 12 and 2186 DF,  p-value: < 2.2e-16
```

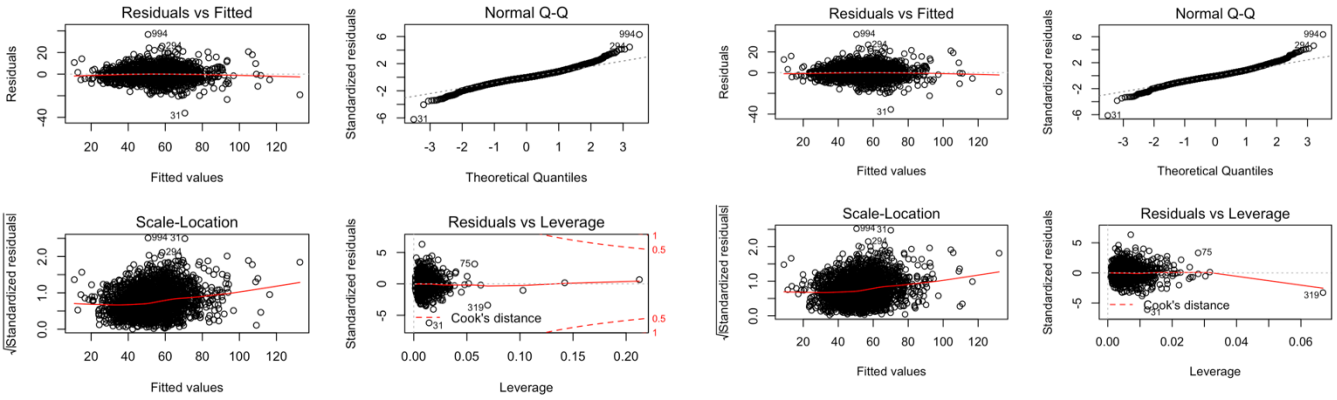
```
Call:
lm(formula = y ~ x1 + x4 + x13 + x16 + x15 + x9 + x10 + log(x2) +
    x11 + x7, data = selected_data)

Residuals:
    Min      1Q  Median      3Q     Max
-35.556  -3.287  -0.310   3.118  36.942

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.036188   2.218355   0.467  0.64048
x1           0.578070   0.009885  58.481  < 2e-16 ***
x4          -0.066409   0.023807  -2.789  0.00533 **
x13          1.049404   0.160257   6.548 7.23e-11 ***
x16         25.587036   4.016488   6.371 2.29e-10 ***
x15         20.752452   3.028572   6.852 9.42e-12 ***
x9          -0.217454   0.043426  -5.007 5.96e-07 ***
x10         -0.131608   0.050547  -2.604  0.00929 **
log(x2)      0.436894   0.141956   3.078  0.00211 **
x11          0.105859   0.047823   2.214  0.02696 *
x7           0.046536   0.018998   2.450  0.01438 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.842 on 2188 degrees of freedom
Multiple R-squared:  0.8293,    Adjusted R-squared:  0.8285
F-statistic:  1063 on 10 and 2188 DF,  p-value: < 2.2e-16
```

Residual plots of initial model and improved model



The correlation of the final dataset