Data Analysis

Factors Affecting Lung Cancer

December 1st, 2020

Young Seok Seo(youngseok.seo@stonybrook.edu)

Table of Contents **Page**

Abstract 2

Introduction 2

Hypothesis I 4

Methodology I 4

Result I 5

Hypothesis II 6

Methodology II 6

Result II 7

Methodology III 13

Result III 13

Comparing Two Results 17

Appendix 18

Abstract

In this report, I will develop a better understanding of advantages of using multiple linear regression method in statistical analysis. The objective of this report is to design the overall problem-solving algorithm, set a hypothesis testing to check whether each independent variable is significant or not. First, I set a hypothesis testing and revised the data into categorical variables. I will check that the marriage and the residence location will affect to the lung cancer death rate or not. Multiple linear regression is extremely useful process in statistical analysis. To familiarize with this method, I will collect and clean the data to check my goal. Once I cleaned the data, I will check which variables are the most effective ones. Also, after setting up the experiment, I will fit the final model which is the most improved linear model. Ultimately, I will erase some of data in order to make some values missing.

Introduction

According to the World Health Organization, WHO, cancer is the second leading cause of death globally. 9.6 million deaths are reported in 2018 due to Cancer. There are various kinds of cancer such as lung cancer, breast cancer, kidney cancer, and lymphoma cancer. The reason that cancer is dangerous and critical to humans is it spreads along the body. Once cancer is detected, then it spreads to other organs of the body. Although medical breakthroughs have occurred over one and two decades, cancer is immedicable if it spreads to various organs. I am interested in what kinds of factors impact death rates that are caused by lung cancer. According to the data that I found on data.world, there are a total of 34 columns and 3047 rows. Rows indicate each county's data in the USA. Descriptions of each column are also from the data.world.

TARGET_deathRate (decimal)**:**

Dependent variable. Mean *per capita* (100,000) cancer mortalities

avgAnnCount (decimal):

Mean number of reported cases of cancer diagnosed annually

avgDeathsPerYear (integer):

Mean number of reported mortalities due to cancer

incidenceRate (decimal):

Mean *per capita* (100,000) cancer diagnoses

medianIncome (integer):

Median income per county

popEst2015 (integer):

Population of county

povertyPercent (decimal):

Percent of populace in poverty

studyPerCap (decimal):

Per capita number of cancer-related clinical trials per county

binnedInc (string):

Median income per capita binned by decile

MedianAge (decimal):

Median age of county residents

MedianAgeMale (decimal):

Median age of male county residents

MedianAgeFemale (decimal):

Median age of female county residents

Geography (string):

County name

AvgHouseholdSize (decimal):

Mean household size of county

PercentMarried (decimal):

Percent of county residents who are married

PctNoHS18_24 (decimal):

Percent of county residents ages 18-24 highest education attained: less than high school

PctHS18_24 (decimal):

Percent of county residents ages 18-24 highest education attained: high school diploma

PctSomeCol18_24 (decimal):

Percent of county residents ages 18-24 highest education attained: some college

PctBachDeg18_24 (decimal):

Percent of county residents ages 18-24 highest education attained: bachelor's degree

PctHS25_Over (decimal):

Percent of county residents ages 25 and over highest education attained: high school diploma

PctBachDeg25_Over (decimal):

Percent of county residents ages 25 and over highest education attained: bachelor's degree

PctEmployed16_Over (decimal):

Percent of county residents ages 16 and over employed

PctUnemployed16_Over (decimal):

Percent of county residents ages 16 and over unemployed

PctPrivateCoverage (decimal):

Percent of county residents with private health coverage

PctPrivateCoverageAlone (decimal):

Percent of county residents with private health coverage alone (no public assistance)

PctEmpPrivCoverage (decimal):

Percent of county residents with employee-provided private health coverage

PctPublicCoverage (decimal):

Percent of county residents with government-provided health coverage

PctPubliceCoverageAlone (decimal):

Percent of county residents with government-provided health coverage alone

PctWhite (decimal): Percent of county residents who identify as White

PctBlack (decimal):

Percent of county residents who identify as Black

PctAsian (decimal):

Percent of county residents who identify as Asian

PctOtherRace (decimal):

Percent of county residents who identify in a category which is not White, Black, or Asian

PctMarriedHouseholds (decimal):

Percent of married households

BirthRate (decimal) : Number of live births relative to number of women in county

In hypothesis I, I will investigate the categorical variables and their significance to the full model that includes all the independent variables. The original data had one variable that estimates the percent of married couples in one county. To make this variable a categorical one, I revised the data into two levels. The new independent categorical variable “Married” is divided into “Yes” for families with more than 50 percent married in a county, and “No” for families with less than 50 percent unmarried. Moreover, I decided to add one more categorical variable in the original data that distinguishes the location in the United States of that county. The newly added independent variable “State” is a column that was categorized by the location in the United States. 51 State code numbers in the United States were used to classify each location. I used the “Geography” variable to classify each location. Therefore, there are two categorical variables to consider in this multiple linear models, which are “State” and “Married”.

In hypothesis II, I will find the most fitted model to the data with multiple regression models. There are 34 columns in the original data. However, since the model should be

continuous and be all numerical values to analyze efficiently, I deleted all the categorical variables such as “binnedInc” and “Geography”. With the rest of the 32 columns, I will find which independent variables, columns, influence on death rate mostly.

Next, I examined the effect of missing values while testing hypothesis II. To find the effect, I randomly deleted 20% of the data, then imputed based on Predictive Mean Matching. By reexamining hypothesis 2, I found the result and compared the difference between previous steps.

Hypothesis I

a) First Hypothesis

H_0: There is no linear association between TARGET_deathRate and married.
= There is no significant difference in TARGET_deathRate of married and unmarried.
H_a: There is a linear association between TARGET_deathRate and married.
= There is a significant difference between the TARGET_deathRate of married and unmarried.

b) Second Hypothesis

H_0: There is no linear association between “TARGET_deathRate” and “State”.
= There is no significant difference in TARGET_deathRate of each location in the USA.
H_a: Here is a linear association between “TARGET_deathRate” and “State”.
= There is a significant difference between the TARGET_deathRate of each location in U.S.

Methodology I

Two hypothesis testings will be mainly used to check whether the two categorical variables are significant or not. Separated two hypothesis testing will be executed, first with the “Married” variable, and second with the “State” variable.

To begin with, I would like to test whether there is a significant difference between the TARGET_deathRate of married and unmarried. In the first hypothesis in Hypothesis I, the full model of multiple linear regression is

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \ldots + e$$

where, Y is a dependent variable, in this case, it is TARGET_deathRate, x_1= State, x_2= Married, x_3=avgAnnCount, …, x_{31}=BirthRate. If I delete the “Married” variable term (x_2), this would be the reduced model. As I revised and add the new data x_1 and x_2 are categorical variables and the rest of them are continuous numerical variables.

First, based on the “Married” variable, there comes an indicator or a dummy variable, x_i= 1 for the married and x_i= 0 for the unmarried. Here, 1 and 0 are the indicator variable values. Then, I need to fit the model with “Married” by considering dataset. R will substitute a 1 for the married couple and a 0 for the unmarried couple.

Now, by using the ANOVA function, I can easily distinguish whether there is a significant difference between the TARGET_deathRate and Married variables. If the p-value

of the ANOVA function of the two models is smaller than the significance level, I reject the null hypothesis. However, If the p-value of the ANOVA function of the two models is larger than the significance level, then I fail to reject the null hypothesis.

Next, I would like to test whether there is a significant difference between the TARGET_deathRate of each location in the United States of America. In the second hypothesis in Hypothesis I, the full model of multiple linear regression is the same as I used in the first hypothesis, which includes all the independent variables. However, in this case, I am excluding the "State" variable (x_i) from the full model.

By setting a hypothesis, I will check whether there the location in the United States matters significantly to the result of TARGET_deathRate.

There is an independent variable, "State", which has more than two levels. There are 51 states starting from Alabama to Wyoming. Same as the independent categorical variable "Married", dummy variables will be used again for "State".

Then, by using the ANOVA function, I can also see whether there is a significant difference between the full model and the reduced model. This means if I reject the null hypothesis, then the "State" variable has a significant effect on the TARGET_deathRate. Likewise, if I fail to reject the null hypothesis, then it means there is no significant effect of "State" on the full model.

Result I

For the "Married" variable, I need to check the p-value between the full model and the model that the "Married" variable has been excluded.

```
> # ANOVA testing
> anova(full.model, married.removed.model)
Analysis of Variance Table

Model 1: TARGET_deathRate ~ State + Married + avgAnnCount + avgDeathsPerYear +
    incidenceRate + medIncome + popEst2015 + povertyPercent +
    studyPerCap + MedianAge + MedianAgeMale + MedianAgeFemale +
    AvgHouseholdSize + PctNoHS18_24 + PctHS18_24 + PctSomeCol18_24 +
    PctBachDeg18_24 + PctHS25_Over + PctBachDeg25_Over + PctEmployed16_Over +
    PctUnemployed16_Over + PctPrivateCoverage + PctPrivateCoverageAlone +
    PctEmpPrivCoverage + PctPublicCoverage + PctPublicCoverageAlone +
    PctWhite + PctBlack + PctAsian + PctOtherRace + BirthRate
Model 2: TARGET_deathRate ~ State + avgAnnCount + avgDeathsPerYear + incidenceRate +
    medIncome + popEst2015 + povertyPercent + studyPerCap + MedianAge +
    MedianAgeMale + MedianAgeFemale + AvgHouseholdSize + PctNoHS18_24 +
    PctHS18_24 + PctSomeCol18_24 + PctBachDeg18_24 + PctHS25_Over +
    PctBachDeg25_Over + PctEmployed16_Over + PctUnemployed16_Over +
    PctPrivateCoverage + PctPrivateCoverageAlone + PctEmpPrivCoverage +
    PctPublicCoverage + PctPublicCoverageAlone + PctWhite + PctBlack +
    PctAsian + PctOtherRace + BirthRate
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1    513 199782
2    514 199922 -1    -139.4 0.3579 0.5499
```

Figure 1. ANOVA test for a variable "Married"

Here, I can see that the p-value is 0.5499, which is larger than the significance level alpha=0.05. Therefore, I cannot reject the null hypothesis. To interpret this, it means that the model with the "Married" variable reduced is not significantly different from the full model at the level of $\alpha = 0.05$.

For the "State" variable, I need to check the p-value between the full model that contains all the variables and the model in that the "State" variable is not included.

```
> # ANOVA testing of "State"
> anova(full.model, state.removed.model)
Analysis of Variance Table

Model 1: TARGET_deathRate ~ State + Married + avgAnnCount + avgDeathsPerYear +
    incidenceRate + medIncome + popEst2015 + povertyPercent +
    studyPerCap + MedianAge + MedianAgeMale + MedianAgeFemale +
    AvgHouseholdSize + PctNoHS18_24 + PctHS18_24 + PctSomeCol18_24 +
    PctBachDeg18_24 + PctHS25_Over + PctBachDeg25_Over + PctEmployed16_Over +
    PctUnemployed16_Over + PctPrivateCoverage + PctPrivateCoverageAlone +
    PctEmpPrivCoverage + PctPublicCoverage + PctPublicCoverageAlone +
    PctWhite + PctBlack + PctAsian + PctOtherRace + BirthRate
Model 2: TARGET_deathRate ~ Married + avgAnnCount + avgDeathsPerYear +
    incidenceRate + medIncome + popEst2015 + povertyPercent +
    studyPerCap + MedianAge + MedianAgeMale + MedianAgeFemale +
    AvgHouseholdSize + PctNoHS18_24 + PctHS18_24 + PctSomeCol18_24 +
    PctBachDeg18_24 + PctHS25_Over + PctBachDeg25_Over + PctEmployed16_Over +
    PctUnemployed16_Over + PctPrivateCoverage + PctPrivateCoverageAlone +
    PctEmpPrivCoverage + PctPublicCoverage + PctPublicCoverageAlone +
    PctWhite + PctBlack + PctAsian + PctOtherRace + BirthRate
  Res.Df    RSS  Df Sum of Sq      F    Pr(>F)
1    513 199782
2    560 242307 -47    -42525 2.3233 4.263e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 2. ANOVA testing for a variable "State"

Here, I can see that the p-value is 4.263e-06, which is smaller than the $\alpha = 0.05$. Therefore, I reject the null hypothesis. It means that there is a significant difference between the full model and the model that the "State" variable has excluded.

Hypothesis II

From the data that I found, I am interested in what factors or what independent variables relate strongly to the death rate by cancer. Multiple linear regressions are the one of linear regression analysis that are used to analyze the relationship between single dependent variables with multiple independent variables. The main objective of multiple linear regression is to select the suitable independent variables that most affects the death rate which is a dependent variable. First step is setting up the hypothesis.

$$H_0: \beta_1 = \beta_2 = \beta_3 = \beta_4 \ldots\ldots = \beta_n = 0$$

$$H_a: \text{At least one } \beta_n \neq 0$$

In this step, it represents the coefficient of each independent variable. If $\beta_1 = \beta_2 = \beta_3 = \beta_4 \ldots\ldots = \beta_n = 0$, then dependent variable has no relationship with all the environmental variables. Therefore, I will find the best fitted model which has the most influence independent variables to the data.

Methodology II

Before to start the modeling and testing the hypothesis, data should be checked whether it follows the assumptions that are required for the Multiple linear regression. There are 4 assumptions which are normality, linearity, heteroscedasticity and multicollinearity. Linearity and heteroscedasticity can be shown in the Residual vs Fitted plot. Normality can be checked with Normal Q-Q plot. Lastly, multicollinearity is important to check.

Multicollinearity occurs when the one independent variable is strongly correlated with other independent variables. It can be checked with vif() function in R.

After checking the assumptions, I should transform the multiple linear function to make it into normal shape. Although normality of the data is already substantiated, I can improve the fitted model with box-cox transformation by square to the dependent variable Y. With the transformed linear function, Stepwise Regression method helps to find the significant independent variables. There are 3 ways, forward elimination, backward elimination, and stepwise elimination, to find the most fitted model to the data. Forward elimination begins with an empty model and adds in variables one at a time. Backward elimination begins with full model and eliminate the extraneous variable in every single step until Adj-r2 increases. Stepwise elimination is combining both forward elimination and backward elimination. Predictor variables can be added or deleted in every step. I need to consider one more thing here. ols_step_both_p, ols_step_backward_p, and ols_step_forward_p functions eliminate predictor variables by p-value. Meanwhile, step functions in R eliminate independent variables by AIC value. I will consider both functions and reflect the function which has the higher Adj-R2and F value.

Result II

```
> #---------------------------------FULL MODEL-------------------------
> full_model <- lm(TARGET_deathRate~ ., data = Cancer_data)
> summary(full_model)

Call:
lm(formula = TARGET_deathRate ~ ., data = Cancer_data)

Residuals:
    Min      1Q  Median      3Q     Max
-76.293 -10.735   0.115  11.046 111.114

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)              2.419e+03  1.512e+03   1.600  0.11018
avgAnnCount             -1.984e-03  1.745e-03  -1.137  0.25600
avgDeathsPerYear         8.476e-03  9.098e-03   0.932  0.35196
incidenceRate            1.621e-01  1.707e-02   9.498  < 2e-16 ***
medIncome                2.335e-04  2.173e-04   1.074  0.28311
popEst2015              -4.019e-06  1.158e-05  -0.347  0.72881
povertyPercent          -4.084e-02  4.084e-01  -0.100  0.92038
studyPerCap             -6.553e-04  2.437e-03  -0.269  0.78815
MedianAge               -2.561e-02  1.775e-02  -1.443  0.14968
MedianAgeMale           -1.105e+00  4.898e-01  -2.257  0.02442 *
MedianAgeFemale          5.979e-01  5.217e-01   1.146  0.25227
AvgHouseholdSize         9.423e-01  2.219e+00   0.425  0.67128
PercentMarried           2.059e+00  4.077e-01   5.052 5.94e-07 ***
PctNoHS18_24            -2.277e+01  1.511e+01  -1.507  0.13243
PctHS18_24              -2.241e+01  1.512e+01  -1.483  0.13869
PctSomeCol18_24         -2.248e+01  1.512e+01  -1.487  0.13746
PctBachDeg18_24         -2.326e+01  1.512e+01  -1.538  0.12453
PctHS25_Over             5.588e-01  2.458e-01   2.273  0.02339 *
PctBachDeg25_Over       -1.054e+00  3.922e-01  -2.688  0.00740 **
PctEmployed16_Over      -7.539e-01  2.600e-01  -2.900  0.00388 **
PctUnemployed16_Over     4.579e-01  4.275e-01   1.071  0.28465
PctPrivateCoverage      -7.837e-03  7.003e-01  -0.011  0.99108
PctPrivateCoverageAlone -5.781e-01  8.229e-01  -0.702  0.48267
PctEmpPrivCoverage       5.393e-01  2.855e-01   1.889  0.05943 .
PctPublicCoverage       -1.215e+00  8.468e-01  -1.435  0.15190
PctPublicCoverageAlone   1.426e+00  9.610e-01   1.484  0.13843
PctWhite                 1.735e-01  1.394e-01   1.244  0.21408
PctBlack                 2.511e-01  1.333e-01   1.884  0.06008 .
PctAsian                 1.714e-01  6.419e-01   0.267  0.78956
PctOtherRace            -9.432e-01  3.170e-01  -2.976  0.00305 **
PctMarriedHouseholds    -2.255e+00  3.855e-01  -5.851 8.33e-09 ***
BirthRate               -3.992e-01  4.806e-01  -0.831  0.40652
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 20.2 on 559 degrees of freedom
  (2456 observations deleted due to missingness)
Multiple R-squared:   0.48,     Adjusted R-squared:  0.4512
F-statistic: 16.65 on 31 and 559 DF,  p-value: < 2.2e-16
```

Figure 3. Summary of full model

If I run full model in R as in Figure 1, Adj-R2 value is 0.4512, F = 16.65 with 31 independent variables. Adj-R2 should ameliorate for the best fitted model. The goal is finding a model fitted model to the data

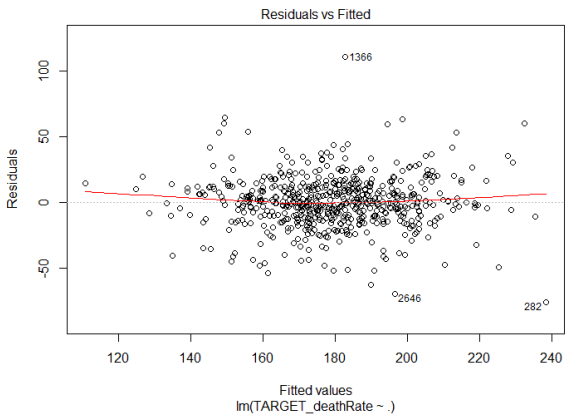

Figure 4. Residual vs. Fitted plot

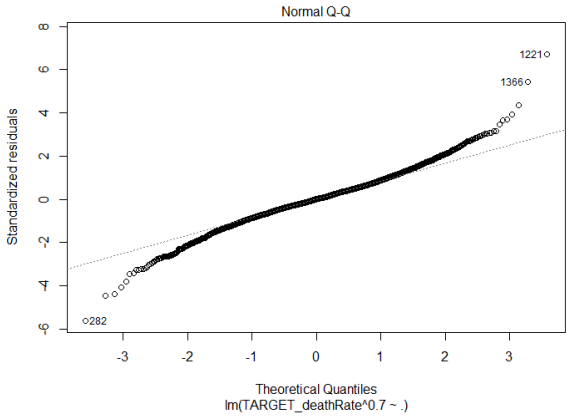

Figure 5. Normal Q-Q plot

Figure 4 shows normality, heteroscedasticity and *Figure 5* shows linearity of the data.

```
> vif(full_model)
          avgAnnCount         avgDeathsPerYear           incidenceRate                medIncome                  popEst2015
            17.207906                60.026246                 1.245404                 8.877030                  49.342208
       povertyPercent                studyPerCap                MedianAge           MedianAgeMale         MedianAgeFemale
             9.325616                 1.097954                 1.036173                 9.084853                  10.071497
    AvgHouseholdSize           PercentMarried              PctNoHS18_24              PctHS18_24           PctSomeCol18_24
             1.442801                10.046440             20358.135138             28826.423579               40208.316065
    PctBachDeg18_24             PctHS25_Over          PctBachDeg25_Over       PctEmployed16_Over     PctUnemployed16_Over
          5834.640824                 4.110425                 6.082987                 5.815717                  2.897088
PctPrivateCoverage PctPrivateCoverageAlone          PctEmpPrivCoverage        PctPublicCoverage PctPublicCoverageAlone
            77.479852                96.313652                10.733304                59.418852                 47.422347
           PctWhite             PctBlack                PctAsian             PctOtherRace       PctMarriedHouseholds
             6.709503                 4.736535                 2.309130                 1.670053                  8.363134
          BirthRate
             1.204834
```

Figure 6. vif function of full model

In *Figure 6*, vif (full_model) returns the variance inflation factor of each independent variable excluding the intercept. If the value of VIF exceeds 10, then that variable indicates near essential multicollinearity. I delete the highest VIF value one at a time till all the VIF value is less than 6 so that I can remove multicollinearity among the environmental variables.

```
> vif(full_model_dropped)
          avgAnnCount           incidenceRate                medIncome             studyPerCap               MedianAge           MedianAgeMale
             1.366823                1.232391                 5.558638                1.038582                1.023651                1.892010
    AvgHouseholdSize              PctNoHS18_24              PctHS18_24          PctBachDeg18_24             PctHS25_Over          PctBachDeg25_Over
             1.297443                1.514931                 1.541082                1.926668                3.638984                5.468823
       PctEmployed16_Over     PctUnemployed16_Over          PctEmpPrivCoverage PctPublicCoverageAlone             PctBlack                PctAsian
             3.559044                2.572251                 4.439829                3.927884                1.798821                1.660420
             PctOtherRace       PctMarriedHouseholds                BirthRate
             1.415597                2.617066                 1.140509
```

Figure 7. vif function of the dropped model

```
> summary(full_model_dropped)

Call:
lm(formula = TARGET_deathRate ~ ., data = Cancer_data_dropped)

Residuals:
     Min       1Q   Median       3Q      Max
-106.333  -11.140   -0.217   10.955  139.464

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)             1.336e+02  1.077e+01  12.407  < 2e-16 ***
avgAnnCount            -7.217e-04  2.962e-04  -2.437  0.01488 *
incidenceRate           1.845e-01  7.691e-03  23.985  < 2e-16 ***
medIncome              -3.235e-05  7.150e-05  -0.452  0.65103
studyPerCap            -3.782e-04  6.880e-04  -0.550  0.58259
MedianAge              -4.114e-03  8.308e-03  -0.495  0.62049
MedianAgeMale          -5.468e-01  9.606e-02  -5.692 1.38e-08 ***
AvgHouseholdSize        1.561e-01  9.790e-01   0.159  0.87330
PctNoHS18_24           -6.127e-02  5.551e-02  -1.104  0.26974
PctHS18_24              3.017e-01  4.984e-02   6.053 1.61e-09 ***
PctBachDeg18_24        -2.785e-02  1.113e-01  -0.250  0.80238
PctHS25_Over            3.028e-01  9.933e-02   3.049  0.00232 **
PctBachDeg25_Over      -1.235e+00  1.594e-01  -7.748 1.29e-14 ***
PctEmployed16_Over     -2.189e-01  8.290e-02  -2.640  0.00833 **
PctUnemployed16_Over    2.890e-01  1.698e-01   1.702  0.08878 .
PctEmpPrivCoverage      9.271e-02  8.156e-02   1.137  0.25578
PctPublicCoverageAlone  6.776e-01  1.184e-01   5.723 1.15e-08 ***
PctBlack                3.025e-02  3.367e-02   0.898  0.36906
PctAsian                1.499e-01  1.778e-01   0.843  0.39922
PctOtherRace           -8.050e-01  1.218e-01  -6.608 4.62e-11 ***
PctMarriedHouseholds   -3.724e-01  9.062e-02  -4.110 4.07e-05 ***
BirthRate              -8.261e-01  1.950e-01  -4.237 2.33e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 19.66 on 2873 degrees of freedom
  (152 observations deleted due to missingness)
Multiple R-squared:  0.4916,    Adjusted R-squared:  0.4878
F-statistic: 132.3 on 21 and 2873 DF,  p-value: < 2.2e-16
```

Figure 8. Summary of the dropped model

In *Figure 6*, comparing to the summary of full_model, Figure 1, Adj-R2 and the F value increases to 0.4878 from 0. 4512 and 132.3 from 16.65. I check all assumptions which are normality, linearity, heteroscedasticity and multicollinearity.

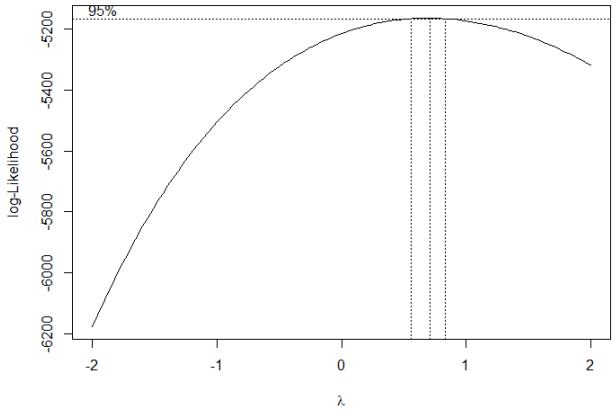

Figure 9. Box-cox transformation

Although my data already follows normality, I try to find a better fit regression model with box-cox transformation.

$$\lambda \approx 0.7$$

I set the linear function as

$$Y^{0.7} = \beta_0 + 1x_1 + 2x_2 + 3x_3 + \ldots + n\,x_n$$

```
> full_model_dropped_transformed <- lm(TARGET_deathRate**0.7~ ., data = Cancer_data_dropped)
> summary(full_model_dropped_transformed)

Call:
lm(formula = TARGET_deathRate^0.7 ~ ., data = Cancer_data_dropped)

Residuals:
     Min       1Q   Median       3Q      Max
-15.7855  -1.6250   0.0253   1.6347  19.4518

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)             3.087e+01  1.592e+00  19.392  < 2e-16 ***
avgAnnCount            -9.959e-05  4.378e-05  -2.275  0.02299 *
incidenceRate           2.746e-02  1.137e-03  24.151  < 2e-16 ***
medIncome              -4.918e-06  1.057e-05  -0.465  0.64175
studyPerCap            -4.814e-05  1.017e-04  -0.473  0.63602
MedianAge              -6.159e-04  1.228e-03  -0.501  0.61607
MedianAgeMale          -8.010e-02  1.420e-02  -5.641 1.86e-08 ***
AvgHouseholdSize        3.682e-02  1.447e-01   0.254  0.79916
PctNoHS18_24           -1.077e-02  8.205e-03  -1.313  0.18927
PctHS18_24              4.285e-02  7.368e-03   5.815 6.72e-09 ***
PctBachDeg18_24        -6.657e-03  1.645e-02  -0.405  0.68571
PctHS25_Over            4.760e-02  1.468e-02   3.242  0.00120 **
PctBachDeg25_Over      -1.864e-01  2.356e-02  -7.911 3.62e-15 ***
PctEmployed16_Over     -3.207e-02  1.225e-02  -2.617  0.00891 **
PctUnemployed16_Over    4.353e-02  2.510e-02   1.734  0.08297 .
PctEmpPrivCoverage      1.578e-02  1.206e-02   1.309  0.19065
PctPublicCoverageAlone  9.636e-02  1.750e-02   5.506 4.00e-08 ***
PctBlack                5.041e-03  4.977e-03   1.013  0.31119
PctAsian                2.226e-02  2.628e-02   0.847  0.39699
PctOtherRace           -1.172e-01  1.801e-02  -6.507 9.00e-11 ***
PctMarriedHouseholds   -5.546e-02  1.340e-02  -4.140 3.57e-05 ***
BirthRate              -1.266e-01  2.882e-02  -4.393 1.16e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.906 on 2873 degrees of freedom
  (152 observations deleted due to missingness)
Multiple R-squared:  0.4931,    Adjusted R-squared:  0.4894
F-statistic: 133.1 on 21 and 2873 DF,  p-value: < 2.2e-16
```

Figure 10. Summary of dropped and transformed of full model

Comparing to previous model,*Figure 6*, value of Adj-R2 and F slightly increase again. Adj-R2 and the F value of the first model, *Figure 1*, is 0.4512 and 16.65. Although I did not run the stepwise method yet, those two values increase significantly.

Here, I will consider a stepwise regression method with two different ways. Firstly, I eliminate the independent variable by Akaike Information Criterion(AIC) value with step function in R. A good model is the one that has minimum AIC among all the backward, forward and stepwise elimination.(CITE - sciencedirect). Backward Elimination(AIC = 6185.55) and both direction(AIC = 6185.55) has lower AIC value, *Figure 9*,then the forward elimination(AIC = 6197.6). Results from backward elimination and both direction elimination are the same.

```
> summary(step_both_2_transformed)

Call:
lm(formula = TARGET_deathRate^0.7 ~ avgAnnCount + incidenceRate +
    MedianAgeMale + PctHS18_24 + PctHS25_Over + PctBachDeg25_Over +
    PctEmployed16_Over + PctUnemployed16_Over + PctEmpPrivCoverage +
    PctPublicCoverageAlone + PctOtherRace + PctMarriedHouseholds +
    BirthRate, data = Cancer_data_dropped)

Residuals:
     Min       1Q   Median       3Q      Max
-15.7581  -1.6253   0.0176   1.6327  19.5391

Coefficients:
                        Estimate Std. Error t value Pr(>|t|)
(Intercept)            3.130e+01  1.410e+00  22.195  < 2e-16 ***
avgAnnCount           -9.148e-05  4.177e-05  -2.190  0.02858 *
incidenceRate          2.751e-02  1.119e-03  24.591  < 2e-16 ***
MedianAgeMale         -8.708e-02  1.291e-02  -6.746 1.83e-11 ***
PctHS18_24             4.468e-02  6.900e-03   6.475 1.11e-10 ***
PctHS25_Over           4.643e-02  1.438e-02   3.230  0.00125 **
PctBachDeg25_Over     -1.866e-01  2.166e-02  -8.615  < 2e-16 ***
PctEmployed16_Over    -3.463e-02  1.192e-02  -2.906  0.00369 **
PctUnemployed16_Over   4.804e-02  2.417e-02   1.988  0.04694 *
PctEmpPrivCoverage     1.612e-02  1.045e-02   1.542  0.12312
PctPublicCoverageAlone 9.562e-02  1.715e-02   5.576 2.70e-08 ***
PctOtherRace          -1.211e-01  1.753e-02  -6.910 5.96e-12 ***
PctMarriedHouseholds  -6.338e-02  1.067e-02  -5.940 3.20e-09 ***
BirthRate             -1.307e-01  2.861e-02  -4.569 5.11e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.903 on 2881 degrees of freedom
  (152 observations deleted due to missingness)
Multiple R-squared:  0.4924,    Adjusted R-squared:  0.4901
F-statistic:   215 on 13 and 2881 DF,  p-value: < 2.2e-16
```

Figure 11. Summary of Stepwise Regression

With step function with both directions, 8 independent variables are eliminated. Total 13 environmental variables left with estimated coefficient as indicated above. According to the *Figure 9*, avgAnnCount, IncidenceRate, MedianAgeMale, PctHS18_24, PctHS25_Over,PctBachDeg25_Over,PctEmployed16_Over,PctUnemployed16_Over,PctEmp PrivCoverage, PctPublicCoverageAlone, PctOtherRate, PctMarriedHouseholds, and Birthrate are significant variables to the Target_Deathrate. Adj-R2 = 0.4901 and F = 215. From the full_model,*Figure 1*, Adj-R2 and F value increased considerably from 0.4512 and 16.65.

Lastly, I will consider stepwise elimination that removed independent variable by P-value with olsrr package in R. Backward and both direction elimination’s result are identical to each other. I set p-remove value as 0.05. The function removes the independent variable which has the highest p-value in every single step till there is no independent variable that has higher p-value than 0.05.

```
> backwardEliminationModel_2_transformed


                        Elimination Summary
-------------------------------------------------------------------------------------
        Variable                      Adj.
Step     Removed          R-Square    R-Square      C(p)        AIC          RMSE
-------------------------------------------------------------------------------------
   1    AvgHouseholdSize     0.4931      0.4896    20.0647    14413.3242    2.9050
   2    PctBachDeg18_24      0.4931      0.4897    18.2297    14411.4904    2.9046
   3    studyPerCap           0.493      0.4898    16.4396    14409.7019    2.9042
   4    medIncome             0.493       0.490    14.6549    14407.9188    2.9038
   5    MedianAge            0.4929      0.4901    12.8929    14406.1586    2.9034
   6    PctAsian             0.4928      0.4902    11.4298    14404.6994    2.9032
   7    PctBlack             0.4927      0.4902    10.3340    14403.6099    2.9032
   8    PctNoHS18_24         0.4924      0.4901     9.9129    14403.1992    2.9035
   9    PctEmpPrivCoverage    0.492      0.4899    10.2880    14403.5883    2.9042
-------------------------------------------------------------------------------------
```

Figure 12. Number of steps in elimination

```
No more variables satisfy the condition of p value = 0.05


Variables Removed:

x AvgHouseholdSize
x PctBachDeg18_24
x studyPerCap
x medIncome
x MedianAge
x PctAsian
x PctBlack
x PctNoHS18_24
x PctEmpPrivCoverage
```

Figure 13. Eliminated variables by backward elimination

With a backward elimination method with ols_step_backward_p function, There are a total 9 steps to find the most fitted model to the cancer data. I can see that AvgHouseholdSize, PctBachDeg18_24, studyPerCap, medIncome, MedianAge, PctAsian, PctBlack, PctNoHS18_24, and PctEmpPrivCoverage variables are eliminated as in *Figure 11* because of greater P-value than 0.05.

```
Final Model Output
------------------

                    Model Summary
---------------------------------------------------------------
R                       0.701       RMSE                 2.904
R-Squared               0.492       Coef. Var            7.722
Adj. R-Squared          0.490       MSE                  8.434
Pred R-Squared          0.485       MAE                  2.155
---------------------------------------------------------------
 RMSE: Root Mean Square Error
 MSE: Mean Square Error
 MAE: Mean Absolute Error

                              ANOVA
-------------------------------------------------------------------------
                Sum of
               Squares          DF    Mean Square       F          Sig.
-------------------------------------------------------------------------
Regression    23539.750          12       1961.646    232.585    0.0000
Residual      24307.111        2882          8.434
Total         47846.861        2894
-------------------------------------------------------------------------

                                  Parameter Estimates
-----------------------------------------------------------------------------------------------------
                model      Beta    Std. Error    Std. Beta      t        Sig      lower      upper
-----------------------------------------------------------------------------------------------------
          (Intercept)    31.606         1.396                22.639    0.000    28.869     34.343
          avgAnnCount     0.000         0.000       -0.030    -2.016    0.044     0.000      0.000
       incidenceRate      0.028         0.001        0.362    25.531    0.000     0.026      0.030
       MedianAgeMale     -0.094         0.012       -0.120    -7.670    0.000    -0.118     -0.070
          PctHS18_24      0.044         0.007        0.099     6.401    0.000     0.031      0.058
          PctHS25_Over    0.052         0.014        0.089     3.710    0.000     0.024      0.079
    PctBachDeg25_Over    -0.181         0.021       -0.238    -8.474    0.000    -0.222     -0.139
   PctEmployed16_Over    -0.029         0.011       -0.059    -2.557    0.011    -0.051     -0.007
 PctUnemployed16_Over     0.052         0.024        0.044     2.144    0.032     0.004      0.099
PctPublicCoverageAlone    0.085         0.016        0.127     5.429    0.000     0.054      0.115
          PctOtherRace    -0.122         0.018       -0.107    -6.965    0.000    -0.156     -0.088
 PctMarriedHouseholds     -0.061         0.011       -0.098    -5.774    0.000    -0.082     -0.040
            BirthRate     -0.139         0.028       -0.068    -4.953    0.000    -0.194     -0.084
-----------------------------------------------------------------------------------------------------
```

Figure 14. Final model output

According to the observation as in *Figure 12*, there are 12 columns in the model. Adj-R2 and F values are higher than the model that is eliminated by AIC value, Figure 9. In the final model ANOVA table, Figure 12, F = 232.585 and Adj-R2 = 0.490. Compare to the step function, Adj-R2 has the same value but value of F increases from 215 to 232.585 while number of columns, independent variable decreases. Therefore, my final model would be DeathRate =

0+ 1avgAnnCount + 2incidenceRate + 3MedianAgeMale + 4PctHS18_24 + 5PctHS25_Over + 6PctBachDeg25_Over + 7PctEmployed16_Over + 8PctUnemployed16_Over + 9PctPublicCoverageAlone + 10PctOtherRace + 11PctMarriedHouseholds + 12BirthRate.

Methodology III

To compare the regression model when missing values exist, I used the "simFrame" R package to set the missing rate and randomly delete the values from the cleansed data. The "MICE" R package is used to perform the imputation. Three iterations using the Predictive Mean Matching method, known as "pmm", I obtained the imputed data. (possibly explain why PMM?) I perform regression analysis following previous steps, resulting in different conclusions.

Result III

Using the imputed data, I first checked the existence of high multicollinearity. One by one, I removed the independent variable with the highest multicollinearity in order until multicollinearity lower than 6 remains. Then, I removed following variables: "PctPrivateCoverageAlone","PctSomeCol18_24","PctPublicCoverage","PctPrivateCoverage", "PercentMarried", "popEst2015", "MedianAgeFemale", "povertyPercent", and "PctWhite". Obtaining the full model, I checked assumptions required for the regression model. As seen in *Figure 13*, the constant variance and normality assumptions are evaluated through residual plot and Q-Q Normal plot respectively. Two plots show that the assumptions are satisfied.

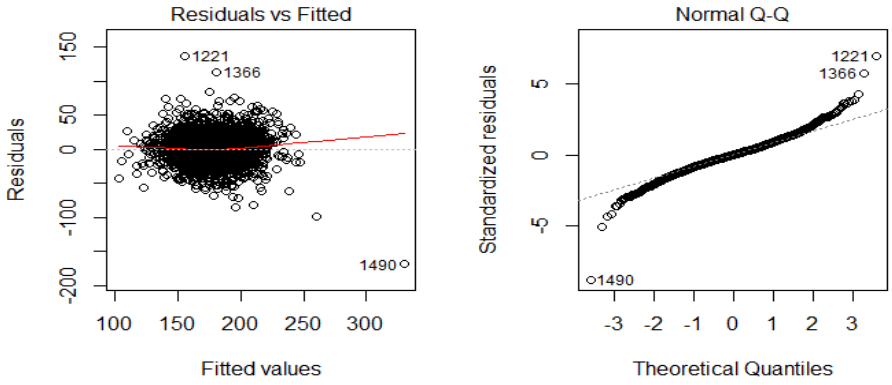

Figure 15. Residual plot & Q-Q plot

To be more precise, I checked the normality assumption again by conducting Box-Cox function. The Box-Cox plot in *Figure 14* indicates the lambda is approximately 1 so that I had concluded that no transformation on the dependent variable is needed.

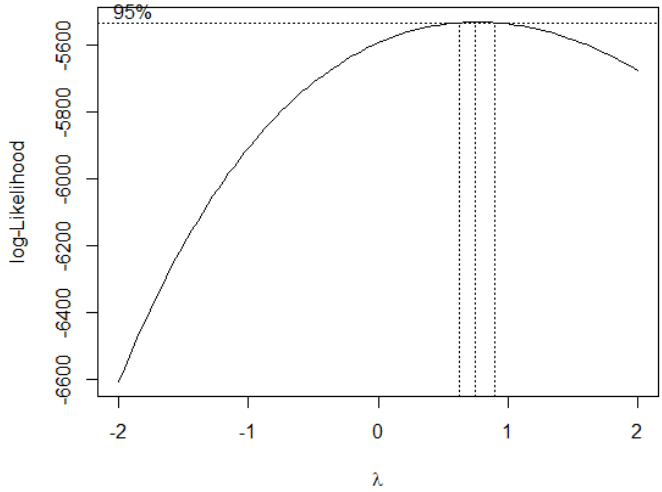

Figure 16. Box-Cox plot

I performed a regression analysis under Akaike's Information Criterion, known as AIC, using step() function in R. Diversifying direction, three tests on "forward", "backward", and "both" directions are performed. The result shows each test conclude 22, 18, and 15 independent variables are required to obtain adjusted R-squared 0.49.

Noting this result, I performed stepwise regression using functions included in the "olsrr" R package for stepwise regression. Again three tests varying in direction are performed in "both", "backward", and "forward" order. In the stepwise regression analysis, regardless of the direction, all three analyses yield identical results. Obtaining the identical regression model, it contains 12 independent variables with R-squared 0.491, and adjusted R-squared 0.489 as seen in Figure 3. Team obtained total 4 models; Three models using step() function and one model using stepwise regression. Comparing between these models, I selected the model with the highest F statistics.

The final model includes "avgAnnCount", "incidenceRate", "MedianAgeMal", "PctHS18_24","PctHS25_Over","PctBachDeg25_Over","PctEmployed16_Over", "PctPublicCoverageAlone", "PctOtherRace", "PctMarriedHouseholds", "BirthRate", and "avgDeathsPerYear".

Models	DF	R-squared	Adj R-squared	F statistics
Model 1	22	0.4928	0.4891	133.6
Model 2	15	0.4925	0.49	196.1
Model 3	15	0.4925	0.49	196.1
Model 4,5,6	12	0.4909	0.4889	243.8

Table 1. Model Comparison

Final Model:

TARGET_deathRate = 140.169 - 0.002 ***avgAnnCount*** *+ 0.182* ***incidenceRate*** *- 0.741* ***MedianAgeMale*** *+ 0.357* ***PctHS18_24*** *+ 0.538* ***PctHS25_Over*** *- 0.754* ***PctBachDeg25_Over*** *- 0.422* ***PctEmployed16_Over*** *+ 0.712* ***PctPublicCoverageAlone*** *- 0.711* ***PctOtherRace*** *- 0.388* ***PctMarriedHouseholds*** *- 0.741* ***BirthRate*** *+ 0.004* ***avgDeathsPerYear***

Below are the individual t-test results of each independent variables given other variables are in the Imputed Final Model.

```
Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)             1.402e+02  8.470e+00  16.549  < 2e-16 ***
avgAnnCount            -1.751e-03  5.455e-04  -3.210  0.00134 **
incidenceRate           1.816e-01  6.954e-03  26.111  < 2e-16 ***
MedianAgeMale          -7.408e-01  7.896e-02  -9.382  < 2e-16 ***
PctHS18_24              3.569e-01  4.538e-02   7.863 5.15e-15 ***
PctHS25_Over            5.381e-01  9.053e-02   5.944 3.11e-09 ***
PctBachDeg25_Over      -7.541e-01  1.416e-01  -5.326 1.08e-07 ***
PctEmployed16_Over     -4.222e-01  6.997e-02  -6.034 1.80e-09 ***
PctPublicCoverageAlone  7.120e-01  1.009e-01   7.056 2.12e-12 ***
PctOtherRace           -7.114e-01  1.150e-01  -6.185 7.03e-10 ***
PctMarriedHouseholds   -3.878e-01  6.836e-02  -5.673 1.53e-08 ***
BirthRate              -7.413e-01  1.851e-01  -4.005 6.34e-05 ***
avgDeathsPerYear        4.456e-03  1.768e-03   2.521  0.01177 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 19.76 on 3034 degrees of freedom
Multiple R-squared:  0.4909,    Adjusted R-squared:  0.4889
F-statistic: 243.8 on 12 and 3034 DF,  p-value: < 2.2e-16
```

Figure 17. Individual t-test

Non-ignorable Missing Values and their Effects

Non-ignorable missing values indicate missing values are not ignorable. This means that there is a close relationship between tendency to be missed and missed values of variables. This phenomenon occurs when people are asked personal information such that it can be seen as against social norms or unpopular about themselves. For example, if people are asked if they smoked during childhood, it is very highly likely that people will say no even if they smoked, under the fear that they may be punished. The other example of missing values may be a question about low income. If people are asked what their income is, for people with satisfying income they may honestly reply to questions. However, if they think they are getting unsatisfyingly low income, they may feel ashamed of their income leading to nonresponse which results in missing values. Regardless what actual reasons are, the existence of non-ignorable missing values creates bias in its analysis. If bias exists in the analysis, bias will also exist in the result. Thus, it is crucial to be careful in wording when making survey questions to avoid nonresponse as much as possible.

To handle randomly deleted missing values, assuming they are actually non-ignorable missing values, I performed an imputation on numerically missing values using the Predictive Mean Matching method.

Predictive Mean Matching method forms a small subset of dataset of complete observations that have closest non-missing values to that of missing observations. By randomly

drawing one from selected dataset, the missing value is imputed. This method helps to reduce the bias because imputed values are reasonably imputed.

To use the Predictive Mean Matching, I used the “MICE” R package. It checks the locations and types of missing values. As non-ignorable missing values include only numerical values, I fixed the method to be Predictive Mean Matching. Setting “maxit” to three, the program iterates the same imputation three times. Then, it pools the result into the final complete data.

Comparing two results

	Full Data	Missing Valued Data
Adj-R2(Full Model)	0.4512	0.4891
F (Full Model)	16.65	133.6
Removed variables by Multicollinearity	• *PctSomeCol18_24* • *PctPrivateCoverageAlone* • *avgDeathsPerYear* • *PctPublicCoverage* • *PctPrivateCoverage* • *PercentMarried* • *MedianAgeFemale* • *popEst2015* • *povertyPercent* • *PctWhite"*	• *PctSomeCol18_24* • *PctPrivateCoverageAlone* • *PctPublicCoverage* • *PctPrivateCoverage* • *PercentMarried* • *popEst2015* • *MedianAgeFemale* • *povertyPercent* • *PctWhite*
Number of Eliminated by stepwise regression	9	10
Number of Independent variable (Final Model)	12	12
Name of Independent Variables (Final Model)	• *avgAnnCount* • *incidenceRate* • *MedianAgeMale* • *PctHS18_24* • *PctHS25_Over* • *PctBachDeg25_Over* • *PctEmployed16_Over* • *PctUnemployed16_Over* • *PctPublicCoverageAlone* • *PctOtherRace* • *PctMarriedHouseholds* • *BirthRate*	• *avgAnnCount* • *incidenceRate* • *MedianAgeMale* • *PctHS18_24* • *PctHS25_Over* • *PctBachDeg25_Over* • *PctEmployed16_Over* • *PctPublicCoverageAlone* • *PctOtherRace* • *PctMarriedHouseholds* • *BirthRate* • *avgDeathsPerYear*
Adj-R2(Final Model)	0.490	0.489
F (Final Model)	232.585	243.825

When I compared two final regression models, using the original data, I learned the model weighted more importance on the effect of including the percent of unemployed over 16. On the other hand, the regression model that was obtained from the dataset which had gone through imputation after random deletion put more importance on the effect of number of average deaths per year.

Appendix Part 1.

```
install.packages("tidyverse")
library(tidyverse)
# Importing Data
Cancer_data <-
read.csv("C:/Users/SEC/Desktop/cancer_reg_categorical.csv")
View(Cancer_data)
colnames(Cancer_data)
# Data Processing
head(Cancer_data)
# Converting Variables as factor
Cancer_data$State=as.factor(Cancer_data$State)
Cancer_data$Married=as.factor(Cancer_data$Married)
table(Cancer_data$State)
table(Cancer_data$Married)
# Fitting the Linear Regression Model
full.model <- lm(TARGET_deathRate~ ., data =
Cancer_data)
full.model
summary(full.model)
# Setting all the linear models
married.removed.model <- lm(TARGET_deathRate ~ State
+ avgAnnCount + avgDeathsPerYear + incidenceRate +
medIncome + popEst2015 + povertyPercent +
studyPerCap + MedianAge + MedianAgeMale +
MedianAgeFemale + AvgHouseholdSize + PctNoHS18_24 +
PctHS18_24 + PctSomeCol18_24 + PctBachDeg18_24 +
PctHS25_Over + PctBachDeg25_Over +
PctEmployed16_Over + PctUnemployed16_Over +
PctPrivateCoverage + PctPrivateCoverageAlone +
PctEmpPrivCoverage + PctPublicCoverage +
PctPublicCoverageAlone + PctWhite + PctBlack + PctAsian
+ PctOtherRace + BirthRate,data = Cancer_data)
state.removed.model <- lm(TARGET_deathRate ~ Married
+ avgAnnCount + avgDeathsPerYear + incidenceRate +
medIncome + popEst2015 + povertyPercent +
studyPerCap + MedianAge + MedianAgeMale +
MedianAgeFemale + AvgHouseholdSize + PctNoHS18_24 +
PctHS18_24 + PctSomeCol18_24 + PctBachDeg18_24 +
PctHS25_Over + PctBachDeg25_Over +
PctEmployed16_Over + PctUnemployed16_Over +
PctPrivateCoverage + PctPrivateCoverageAlone +
PctEmpPrivCoverage + PctPublicCoverage +
PctPublicCoverageAlone + PctWhite + PctBlack + PctAsian
+ PctOtherRace + BirthRate,data = Cancer_data)
# ANOVA testing
anova(full.model, married.removed.model)
anova(full.model, state.removed.model)
```

Appendix Part 2.

```
install.packages("dplyr")
library(dplyr)
Cancer_data <-
read.csv("C:/Users/ryans/Desktop/archive/cancer_reg.csv")
Cancer_data
colnames(Cancer_data)
cancer_data_linear <- lm((avgAnnCount)~ ., data =
Cancer_data)
summary(cancer_data_linear)
install.packages("boxcoxmix")
install.packages("MASS")
library(MASS)
boxcox(cancer_data_linear)
plot(cancer_data_linear)
plot(Cancer_data)

#----------P-VALUE----------#
install.packages("olsrr")
library(olsrr)
install.packages("car")
library(car)
#FULL MODEL
all_model <- lm(TARGET_deathRate~ ., data = Cancer_data)
summary(all_model)
vif(all_model)
#STEPWISE REGRESSION
stepWiseRegressionModel <-
ols_step_both_p(all_model,pent = 0.05,prem = 0.05, detail
= TRUE)
stepWiseRegressionModel
stepWiseRegressionModel$model
stepWiseRegressionModel$steps
stepWiseRegressionModel$rsquare
stepWiseRegressionModel$orders
plot(stepWiseRegressionModel)
#BACKWARD ELIMINATION
backwardEliminationModel <-
ols_step_backward_p(all_model,prem=0.05, detail= TRUE)
plot(backwardEliminationModel)
backwardEliminationModel$model
backwardEliminationModel$indvar
backwardEliminationModel
#FORWARD ELIMINATION
forWardEliminationModel <-
ols_step_forward_p(all_model,pent = 0.05,detail = TRUE)
forWardEliminationModel
#----------AKAIKI----------#
step(all_model, direction = "forward")
step_back <- step(all_model, direction = "backward")
summary(step_back)
vif(step_back)
step_both <-step(all_model, direction = "both")
summary(step_both)
```

Appendix Part 3.

```
Cancer_data2 <- read.csv("cancer_reg.csv", header=TRUE)
library(car)
# Random Deletion
library(simFrame)
dataFrame <-as.data.frame(Cancer_data2)
nac<-NAControl(NArate=0.2) #set missing rate
Newdata<-setNA(dataFrame,nac) #get new data(with missing data)

# Imputation
library(mice)
data_incomplete = Newdata
imp = mice(data_incomplete, printFlag = FALSE)
meth = imp$meth
meth = "pmm"
Imp = mice(data_incomplete, maxit = 3, printFlag = FALSE, meth = meth)
Cancer_data22 = complete(Imp)
all_model_2 <- lm(TARGET_deathRate~ ., data = Cancer_data22)
vif(all_model_2)
drop <- c("PctPrivateCoverageAlone", "PctSomeCol18_24", "PctPublicCoverage", "PctPrivateCoverage", "PercentMarried", "popEst2015", "MedianAgeFemale", "povertyPercent", "PctWhite")
Cancer_data32 <- Cancer_data22[!(names(Cancer_data22) %in% drop)]
all_model_2 <- lm( TARGET_deathRate ~ ., data= Cancer_data32)
vif(all_model_2)
library(MASS) # For BoxCox Normality Check
boxcox(all_model_2)

# Regression using AIC

step_model1 <- step(all_model_2, direction = "forward")
summary(step_model1)

step_model2 <- step(all_model_2, direction = "backward")
summary(step_model2)

step_model3 <-step(all_model_2, direction = "both")
summary(step_model3)


library(olsrr) # Stepwise Regression

#Both Direction
step_model4 <- ols_step_both_p(all_model_2, pent = 0.05, prem = 0.05, detail = TRUE)
step_model4
step_model4$model

#Backward
step_model5 <- ols_step_backward_p(all_model_2, prem=0.05, detail= TRUE)
step_model5
step_model5$model

#Forward
step_model6 <-ols_step_forward_p(all_model_2,pent = 0.05,detail = TRUE)
step_model6
step_model6$model

Final_imp_model <- lm( TARGET_deathRate ~
                       avgAnnCount+
                       incidenceRate+
                       MedianAgeMale+
                       PctHS18_24+
                       PctHS25_Over+
                       PctBachDeg25_Over+
                       PctEmployed16_Over+
                       PctPublicCoverageAlone+
                       PctOtherRace+
                       PctMarriedHouseholds +
                       BirthRate+
                       avgDeathsPerYear, data = Cancer_data32)

summary( Final_imp_model )
```