**MATH 6357: Linear Models and Design of Experiment**

**Group Project Report**

**Fall 2023**

**Title: Data analysis on U.S. Cancer Mortality data with multivariate linear regression**

**Group members:**

* Elaheh Rajabi
* Lulu Xu
* Nabanita Gupta
* Yi Zhang

**Outline:**

* Report
* Appendix

**Introduction**

Cancer remains one of the leading causes of death worldwide. Understanding the factors that contribute to cancer mortality rates can help in formulating effective public health policies and interventions. Based on the cancer data of U.S. counties, this project aims to address the following main issues:

* Building a multivariate linear regression model based on the cancer data.
* Identifying the key factors that tightly correlated with cancer mortality rates in US counties.
* Understanding the relationships and interactions between different variables in the context of cancer mortality.
* Estimating cancer mortality rates based on these factors.

Research questions and hypotheses include:

* Which socio-economic factors have the most significant impact on cancer mortality rates?
* How do demographic factors like race, insurance coverage influence cancer mortality rates?
* What is the effect of one set of variables on cancer mortality rate after taking into account rest of the variables?
* Lower poverty associates with lower cancer mortality rates?
* Will counties with higher percentages of residents with health coverage have lower cancer mortality rates.
* How accurate can we predict cancer mortality rate for U.S. counties?

Data source:

The dataset for this project is called “cancer”, it was aggregated from three sources: American Community Survey, clinicaltrials.gov, and cancer.gov. The data was compiled by Noah Ripper on “data.World” website.

The “cancer” data includes 3047 observations which are 3047 U.S. counties. There are 34 variables in this data, with 33 independent variables and 1 target variable. These variables were related to demographics, socio-economic status, health coverage, and cancer-related clinical trials. The Target variable is “Mean per capita cancer mortalities” which is the mean number of deaths related to cancer per year for every 100,000 population in a county through year 2010-2016.

The 33 Independent variables were divided into 5 categories (detailed variable list in appendix):

1. Number of cases diagnosed, 4 variables.
2. Economic factors such as income, unemployment rate, health coverage,10 variables
3. Population, age, marriage, and other social factors, 9 variables
4. Education, 6 variables
5. Ethnicity, 4 variables

**Methodology**

We conducted a thorough examination of the dataset to ensure its completeness, addressing issues such as missing values, data errors, and outliers. Given that our dataset originates from various sources, we identified overlapping columns that could lead to multicollinearity.

* **Feature reduction:**

Several variables with categorical data and redundant information were deleted, overlapping columns were merged, more meaningful variables with reduced correlation were reacted. This step aimed to reduce multicollinearity.

* **Missing Value Handling:**

After the column combination and the creation of new variables, we scrutinized the dataset for missing values. Three of features have substantial number of missing values: 2284, 670 and 150 respectively. These 3 features were eliminated. After all we have 2895 observations in 16 independent variables, with no missing data.

* **Correlation Analysis:**

The correlation matrix and plot highlighted significant association between "Target Death Rate" and variables. The top 5 highest correlations coefficient observed are: PctUnemployed16\_Over(0.37), Employed16\_Over (-0.41), Poverty Percent (0.43), Median Income (-0.43), Mean Public Coverage (0.44).

Absolute value of correlations

**A graph with numbers and text

Description automatically generated with medium confidenceA screenshot of a computer screen

Description automatically generated**

* **Multicollinearity checking**

For the 16 variables, we checked multicollinearity with VIF values and found that all 16 variables have VIF value less than 7, the highest one is “PctWhite” with 6.98. We will pay attention to these high VIF value variables latter.

* **LASSO Regression**

We applied LASSO regression to the data to minimize predictor variables, which produced "Target Death Rate" predictions. The model was optimized using cross-validation to determine the best lambda value (0.114574). Certain variables like “MedianAge” and “PctWhite” had negligible coefficients, suggesting little impact on the model, while others like “AvgHouseholdSize” and “PctBlack” were significant, with coefficients above 0.01. The LASSO model identified 12 key predictors for predicting the "Target Death Rate." All features were standardized before building a regression model with these 12 predictors, using a dataset of 2895 observations, which adheres to the rule of thumb regarding sample size.

* **L.I.N.E assumptions**

We built an elementary multivariate linear regression model and checked that data satisfaction of the “L.I.N.E” statistics assumptions. There is linear relationship between target and variables (R square is 0.32), errors are independent, normally distributed, and have equal variance.

A collage of graphs and charts

Description automatically generated

Now, our dataset has 13 variables and 2895 observations for our next data analysis step.

**Data Analysis**

* **Category based hypothesis test:**

In this part, we addressed our research questions Hypothesis test by categories. Based on the full model (lm2.data\_normal), we categorized 12 predictors into 5 categories:

1. Race (PctBlack, PctAsian, PctOtherRace),
2. Household (AvgHouseholdSize, PctMarriedHouseholds)
3. Birth rate (BirthRate)
4. Education (PctLowEducation, PctHighEducation)
5. Economic (PctUnemployed16\_Over, povertyPercent, MeanPrivateCoverage, MeanPublicCoverage)

Instead of comparing variables within each category, each category (all variables in each category together) was tested if it is significant. We assessed whether all variables in a category had coefficients bi​ = 0 or if at least one of them was not zero.

For example, hypothesis test for race category in the model. Our full model and reduce model are:

model\_with\_race <- lm(TARGET\_deathRate~.,data = df2)  
model\_without\_race <- lm(TARGET\_deathRate ~.- PctBlack - PctAsian - PctOtherRace, data = df2)  
Null Hypothesis (H0): The coefficients of predictors in race category are equal to zero.

Alternative Hypothesis (Ha): At least one of the coefficients of predictors is NOT zero.

Decision Rule: The decision rule is based on the p-value:

If p-value < α, we reject H0,

If p-value > α, we fail to reject H0.

Conclusion: For Race category the F-statistic is 29.013. The p-value is less than 2.2e-16, which is much smaller than the significance level of 0.05, indicating a highly significant result. We reject the null hypothesis that the coefficients for "PctBlack", "PctAsian", and "PctOtherRace" are all equal to zero. This means that including these race terms significantly improves the model's predictive ability for "TARGET\_deathRate".

After that we did the same hypothesis tests for the other four categories (household, birth rate, education, and economic). They all obtain similar results. The p-value are all less than the significance level of 0.05. We reject each of the null hypothesis that the coefficients for predictors in each category are equal to zero. The addition of category to the model (lm2.data\_normal) significantly improves its ability to predict the target death rate. This is evidenced by the substantial decrease in RSS and the highly significant p-value, indicating that all these five categories are crucial in modeling the target death rate in this context. The all-category hypothesis has significantly improved its ability to predict the target death rate.

* **Stepwise regression for best subsets**

Here stepwise regression was applied to choose best model. The full model is “lm2.data\_Normal” which has 12 variables chosen by LASSO. Both forward and backward direction were tested, and they both show same results. With 10 variables included, model AIC value reached smallest -1107.08. Two variables (“PctBlack” and “AvgHouseholdSize”) were reduced.

Also, regression subsets selection procedure was applied to compare with stepwise regression results. It shows that for subsets size of 10 variables, same as step wise results, “PctBlack” and “AvgHouseholdSize” variables can be deleted.

* **Model Analysis Criterion**

Based on regression results, compared different criteria such as R2, Cp, AIC, PRESS and decided the best model.

A graph of different sizes and numbers

Description automatically generated with medium confidence

A graph of different sizes and numbers

Description automatically generated with medium confidence

It shows that: R2 reached maximum with 12 variables; adjusted R2 reached maximum at 10 variables; Cp reached minimum at 10 variables; BIC reached minimum at 9 variables. Overall, from the previous results, the model has the best results by 9 variables (exclude PctBlack).

* **Best Model Analysis**

Previously, based on the stepwise regression and model analysis criterion comparing, we change 12 predictors to 9 predictors. So, can we drop these 3 predictors from 12 predictors? We did the hypothesis test to check if we can do this. Full model is with 12 variables, and reduce model has 9 predictors. The p-value is 0.281, we fail to reject the H0, the 3 predictors could be dropped.

* **Best Model**

Thus, we get our best model with these 9 predictors and the Summary of the model expressed that R-square value of 0.3219, and adjusted R-square value of 0.3198. The intercept estimated is not significant at all (with t value of 0 and p-value =1), so the intercept term b0 of model could be 0. Other than that, all the 9 predictors are significant, and their p value is less than 0.001, The overall F statistics is 152 and p-value is also significant.

A screenshot of a computer

Description automatically generated

ANOVA table of this model shows the SSR of each single variable. It shows that “meanPublicCoverage”,”PctOtherRace”,”povertyPercent”, “PctLowEducation” and “MeanPrivateCoverage” are 5 predictors which explained most sum of squared regression amount among 9 predictors.

* **Best Model: Confidence interval for coefficients**

The 95% confidence interval for the intercept includes zero, which, along with a non-significant p-value from the model summary, suggests there is no evidence to reject the null hypothesis that the intercept (b0) is zero. Therefore, we conclude that the intercept is not statistically different from zero in our model. The confidence intervals for 9 predictors are not include zero.

* **Best Model: Outliers and Influential Points**

A graph with a line

Description automatically generatedA graph of a number of objects

Description automatically generated with medium confidence

To look for any outliers, we plotted studentized residuals plot and see a few outliers. It shows that point index with 1221, 1366,1942 are out of 4 quantile range.

To check are these outliers also influential points, Cook's distance was computed. Plot of cook’s distance shows most of the point values are less than 0.02, which is not influential. The point with value of 0.05 also not influential since the percentile value less than 2%.

Since we have checked the outliers and no influential outliers, we can simply delete those outliers for analysis data.

* **Best Model: Constant Variance Check**

group<-rep(1,length(df$AvgHouseholdSize))  
summary(lm.best9$fitted.values)

group[lm.best9$fitted.values<0]=0  
bftest(lm.best9,group,alpha=0.05)

## t.value P.Value alpha df  
## [1,] 3.067796 0.002176403 0.05 2893

It shows p-value is significant, so we can reject H0, there is equal variance, and conclude non equal variance.

We applied the “**ncvTest**(lm.best9)” function, it shows the suggested power transformation is 0.8998591. Then we did power transformation response with power value 0.8998591 and check constant variance again.

A graph of a number of values

Description automatically generated with medium confidenceA diagram of a graph

Description automatically generated

Compared to the previous plot of residuals versus fitted values, the plot of absolute standardized residuals versus fitted values includes a superimposed line of best fit.

* **Best Model: Hypothesis testing for interaction terms.**

To test the interaction terms are significant or not, model with all the possible interaction terms was build. Null Hypothesis (H0): The coefficients of interaction terms are equal to zero, which means these variables have no effect on the TARGET\_deathRate. Alternative Hypothesis (Ha): The coefficients of the variables are not all equal to zero, which means at least one term effect on target. Decision Rule: based on the p-value: If p-value < α, we reject H0.If p-value > α, we fail to reject H0.

Conclusion: Overall Model Significance: Based on the R output, the overall F-statistic (26.26) and its associated p-value (< 2.2e-16) indicate that the model with all the predictors and their interaction terms is statistically significant when compared to a model without these predictors.

Even though interaction terms have improved the model, not all of them might be necessary. We used stepwise regression to simplify the model, retaining only the terms that contribute significantly to the model. This helps to prevent overfitting. The predictors of simplified\_model was selected using stepwise, based on CIA-value. Based on the R output provided, there are 36 predictors in the simplified\_model, including 9 original predictors and 21 interaction terms. This is still too many predictors for a regression model, which could potentially lead to an overfitting problem.

* **Best model: Hypothesis testing for interaction terms.**

Except considering all the interaction term together, we also tested 4 sets of interaction terms individually, and compared with the model with no interaction. We checked interaction terms again based on 9 selected variables, and found that interaction between **PctUnemployed16\_Over** and **MeanPublicCoverage** is significant. The small p-value (1.528e-05) indicates strong evidence against the null hypothesis. we concluded that the interaction term 'PctUnemployed16\_Over \* **MeanPublicCoverage** improves the model fit.

* **Best Model: Prediction performance**

The main purpose of this project is to inference from regression model, we also To evaluate the predictive ability of the best model and compare it with the predictivity performance of the model including interactions as predictors, data was partitioned into the train and test sets by a 70 to 30 ratio. We used the train set to build our model and then implemented our models on the test set. The results show very high RMSE rate which is 0.8251. It may indicate that for prediction purpose, building regression model on this data set is not desirable.

**A screenshot of a graph

Description automatically generated**

When comparing the performance of the model with interaction terms to the model without interaction terms (the best model with 9 predictor), we found that although hypothesis testing indicated the interactions improved our best model, the prediction results showed that the model without interactions (the best model) performed better than the model with interactions. Therefore, we do not need to include the interaction terms in our model. We still choose the model with 9 predictors as our best model.

**Conclusion**

In this study, we rigorously analyzed U.S. Cancer Mortality data through multivariate linear regression to identify significant factors influencing cancer mortality rates. Our methodical approach included feature reduction to mitigate multicollinearity and the handling of missing values to ensure robustness in our model. Correlation analysis revealed key variables with a strong relationship to the target death rate, guiding our selection of predictors.

Our employment of LASSO regression refined the model, singling out 12 influential predictors, and standardization procedures enabled us to meet the statistical assumption of homogeneity. Furthermore, category-based hypothesis testing underscored the relevance of (1) socio-economic, (2) demographic, and (3) health coverage factors for predicting cancer mortality.

The iterative process of model selection, including stepwise regression and analysis of various model criteria (R^2, AIC, and PRESS), led us to a parsimonious yet effective model composed of 9 predictors. This model not only adhered to the L.I.N.E assumptions but also demonstrated the highest explanatory power without overfitting, as evidenced by the statistical insignificance of additional interaction terms.

Our analysis concluded with the validation of the model's assumptions of constant variance and the examination of outliers. While the model showed high predictive power internally, the external prediction performance indicated limitations, suggesting potential areas for further research and model refinement.

Our investigation into U.S. Cancer Mortality data via multivariate linear regression has yielded significant insights into socio-economic and demographic factors that impact cancer mortality rates. We determined that economic factors, like mean poverty percentage, health insurance coverage of a county have a profound influence on average cancer mortality rates from that respective county. Demographic factors, such as race also play a critical role. Our model indicated that counties with a higher proportion of black residents and lower insurance coverage tend to have higher cancer mortality rates.

Additionally, the effect of a single set of variables on cancer mortality rates was found to be substantial even after accounting for the rest of the variables, underscoring the importance of each factor's unique contribution. Our analysis supports the hypothesis that higher poverty is associated with higher cancer mortality rates, suggesting that economic well-being is a protective factor against cancer mortality.

Furthermore, we observed that counties with a higher percentage of residents with health coverage tend to have lower cancer mortality rates, highlighting the importance of healthcare accessibility as a determinant of health outcomes.

In conclusion, average cancer related mortality of a county can be inferred to be proportional with poverty, % population with higher employment/income, % population with high education, inversely proportional with % of population with health coverage. All indicating towards affordability of cancer treatments in a way, as that is indeed expensive. Our research addressed critical questions and provided evidence-based answers that can inform public health strategies. However, the complex nature of cancer mortality requires ongoing research to improve predictive accuracy and the need for more sophisticated or non-linear modeling techniques in future studies.

**Appendix**

* **Cancer data variable information.**

A screenshot of a computer

Description automatically generated

* **21 significantly associated interaction terms**

|  |  |
| --- | --- |
| **Interaction\_Term** | **P\_Value** |
| **AvgHouseholdSize:PctBlack** | **0.00465 (\*\*)** |
| **AvgHouseholdSize:PctUnemployed16\_Over** | **0.01584 (\*)** |
| **AvgHouseholdSize:povertyPercent** | **4.12e-05 (\*\*\*)** |
| **AvgHouseholdSize:MeanPrivateCoverage** | **0.01890 (\*)** |
| **PctBlack:PctAsian** | **1.98e-05 (\*\*\*)** |
| **PctBlack:PctLowEducation** | **3.89e-09 (\*\*\*)** |
| **PctBlack:PctHighEducation** | **0.00700 (\*\*)** |
| **PctBlack:povertyPercent** | **1.05e-05 (\*\*\*)** |
| **PctBlack:MeanPrivateCoverage** | **0.00589 (\*\*)** |
| **PctAsian:PctMarriedHouseholds** | **0.03123 (\*)** |
| **povertyPercent:MeanPublicCoverage** | **6.01e-05 (\*\*\*)** |
| **PctAsian:PctUnemployed16\_Over** | **0.00190 (\*\*)** |
| **PctAsian:MeanPublicCoverage** | **0.00619 (\*\*)** |
| **PctOtherRace:PctUnemployed16\_Over** | **0.01028 (\*)** |
| **BirthRate:PctMarriedHouseholds** | **3.38e-05 (\*\*\*)** |
| **BirthRate:povertyPercent** | **4.37e-05 (\*\*\*)** |
| **BirthRate:MeanPrivateCoverage** | **4.02e-05 (\*\*\*)** |
| **BirthRate:MeanPublicCoverage** | **0.00506 (\*\*)** |
| **PctMarriedHouseholds:PctLowEducation** | **0.00206 (\*\*)** |
| **PctUnemployed16\_Over:PctLowEducation** | **0.01450 (\*)** |
| **povertyPercent:MeanPrivateCoverage** | **0.00334 (\*\*\*)** |

* **Essential R code:**

data=read.csv("cancer\_reg.csv",header = T)  
head(data)

data$MedianAgeMaleFemale <- apply(data[, c("MedianAgeMale", "MedianAgeFemale")], 1, median)  
equality\_check <- data$MeanAgeMaleFemale == data$MedianAge  
print(data[!equality\_check, c( "MedianAgeMale", "MedianAgeFemale", "MedianAge","MedianAgeMaleFemale")])

data<- data[, !(names(data) %in% c("MedianAgeMale", "MedianAgeFemale", "studyPerCap",   
 "binnedInc", "avgDeathsPerYear", "avgAnnCount",   
 "PercentMarried","MedianAgeMaleFemale"))]  
dim(data)

#new Education category 23 variables left  
data$PctLowEducation<- rowMeans(data[, c( "PctHS18\_24", "PctHS25\_Over" ,"PctNoHS18\_24")], na.rm = TRUE)  
data$PctHighEducation<- rowMeans(data[, c( "PctBachDeg25\_Over", "PctBachDeg18\_24","PctSomeCol18\_24")], na.rm = TRUE)  
#Drop the original education columns  
data<- data[, !(names(data) %in% c("PctHS18\_24", "PctHS25\_Over", "PctSomeCol18\_24", "PctBachDeg18\_24", "PctBachDeg25\_Over","PctNoHS18\_24"))]  
dim(data)

#----------------------#manipulating health coverage categories #----------  
# Create new columns for health coverage  
data$MeanPrivateCoverage <- rowMeans(data[, c("PctPrivateCoverage", "PctPrivateCoverageAlone", "PctEmpPrivCoverage")], na.rm = TRUE)  
data$MeanPublicCoverage <- rowMeans(data[, c("PctPublicCoverage", "PctPublicCoverageAlone")], na.rm = TRUE)  
# Drop the original health coverage columns 25 left  
data<- data[, !(names(data) %in% c("PctPrivateCoverage", "PctPrivateCoverageAlone", "PctEmpPrivCoverage", "PctPublicCoverage", "PctPublicCoverageAlone"))]  
#-------  
data <- data[!is.na(data$PctEmployed16\_Over), ]  
colSums(is.na(data))

# Create a vector for each category  
demographic\_columns <- c("MedianAge", "AvgHouseholdSize", "PctWhite", "PctBlack", "PctAsian", "PctOtherRace", "PctMarriedHouseholds", "BirthRate")  
education\_columns <- c("PctLowEducation", "PctHighEducation")  
employment\_columns <- c("PctEmployed16\_Over", "PctUnemployed16\_Over")  
income\_poverty\_columns <- c("medIncome", "povertyPercent")  
health\_coverage\_columns <- c("MeanPrivateCoverage", "MeanPublicCoverage")  
# Combine all columns in the desired order  
ordered\_columns <- c( demographic\_columns,  
 employment\_columns, education\_columns,  
 income\_poverty\_columns, health\_coverage\_columns,"TARGET\_deathRate")  
# Create a new data frame with the 17 ordered columns  
ordered\_data <- data[ordered\_columns]  
colnames(ordered\_data)

r2 = round(cor(ordered\_data, use = "complete"), 2)  
ggcorrplot(r2, hc.order = TRUE, type = "full", lab = TRUE, lab\_size = 2,show.legend = TRUE, ggtheme = theme\_minimal(),title="correlation matrix between variables") # Change font size of the labels

#ABsolute value of correlation  
a = as.data.frame(sort(abs(cor(ordered\_data, use = "complete")[17,]))) # sorting the corr for easy visualization  
str(a) # 26 as we removed the categorical variable "Geography" aka County name # Investigating variables with high correlation with the target variable, true value, not abs  
cor\_17=as.data.frame(round(sort(cor(ordered\_data, use = "complete")[,17],decreasing = T),4))

lm.data=lm(TARGET\_deathRate~.,data=ordered\_data) # build a elementary model with 17 variabels to check vif  
# Calculate VIF each predictor  
vif\_values <- car::vif(lm(TARGET\_deathRate~.,data=ordered\_data))  
#generally, vif>4 or sqrt(vif)>2 indicate multicollinearity issue, 6 variabels has this issue   
needs\_investig <- names(vif\_values[vif\_values > 4])  
# Identify variables with high VIF  
high\_vif\_vars <- names(vif\_values[vif\_values > 9])  
high\_vif\_vars# is 0 means no serious multicollinearity

##residual  
residuals=resid(lm.data)  
c=hist(residuals, xlab = "Residuals", ylab="Density" ,col = "pink",main = "Histogram of residuals",freq = FALSE)

plot(lm.data)

#define response variable  
y <- ordered\_data$TARGET\_deathRate  
x <- data.matrix(ordered\_data[, -c(17)])  
# cv.glmnet() automatically performs k-fold cross validation using k=10 folds  
# Find optimal lambda value  
cv\_model2 <- cv.glmnet(x, y, alpha = 1)  
#find optimal lambda value that minimizes test MSE  
best\_lambda2 <- cv\_model2$lambda.min  
best\_lambda2 # 0.1380048

#produce plot of test MSE by lambda value  
plot(cv\_model2)

#find coefficients of best model  
best\_model2 <- glmnet(x, y, alpha = 1, lambda = best\_lambda2)  
coef(best\_model2)

# Trying to list variables with higher coeffs.  
Lasso\_coeff = as.matrix(coef(best\_model2), row.names(TRUE))  
*#———–standardize data and chose features based on LASSO REGRESION RESULTS*

df <- ordered\_data[,-c(1)]   
dim(df)

#df <- as.data.frame(DF)  
for (i in 1:length(df)){df[,i] <- (df[,i] - mean(df[,i])) / sd(df[,i])}  
# Variables considered by LASSO (variables with non-zero coeffs.).  
which(abs(Lasso\_coeff) > 0.01, arr.ind = T)

select\_var\_lasso = c(rownames(which(abs(Lasso\_coeff) > 0.01, arr.ind = T)))  
select\_var\_lasso = select\_var\_lasso[-1] # Removing the "intercept" from the list  
  
df2 = df[,c("TARGET\_deathRate",select\_var\_lasso)]  
*##this our base model on 12 variabels.*

lm2.data\_Normal <- lm(TARGET\_deathRate~.,data = df2)  
summary(lm2.data\_Normal)

par(mfrow = c(2,2))  
plot(lm2.data\_Normal)

lm.redu\_eco<-lm(TARGET\_deathRate~.-PctUnemployed16\_Over-povertyPercent-MeanPrivateCoverage-MeanPublicCoverage,data=df2) #reduce 4 related variables  
summary(lm.redu\_eco)

# Fit the model with the race terms(Full\_model)  
model\_with\_race <- lm(TARGET\_deathRate~.,data = df2)   
# Fit the model without the race terms  
model\_without\_race <- lm(TARGET\_deathRate ~ AvgHouseholdSize + PctMarriedHouseholds + BirthRate + PctUnemployed16\_Over + PctLowEducation + PctHighEducation + povertyPercent + MeanPrivateCoverage + MeanPublicCoverage, data = df2)   
# Perform an ANOVA to compare the two models  
anova\_result\_race <- anova(model\_without\_race, model\_with\_race)

*#interations*

model\_interaction <- lm(TARGET\_deathRate ~ (AvgHouseholdSize + PctBlack  
 + PctAsian + PctOtherRace + BirthRate  
 + PctMarriedHouseholds + PctUnemployed16\_Over  
 + PctLowEducation + PctHighEducation  
 + povertyPercent + MeanPrivateCoverage  
 + MeanPublicCoverage)^2, data = df2) # this bind check all the possible pairs, running takes time  
summary(model\_interaction)

stepwise\_model <- step(model\_interaction, direction = "both")

summary(stepwise\_model)

simple\_model\_formula <- formula(stepwise\_model)

print(simple\_model\_formula)

simplified\_model <- lm(simple\_model\_formula, data = df2)

summary(simplified\_model)

start.model=lm(TARGET\_deathRate~1,data = df2)  
step(lm2.data\_Normal, direction = "backward")

step(start.model, direction = "forward",scope=formula(lm2.data\_Normal))

regsub=regsubsets(TARGET\_deathRate~.,data = df2,nvmax=12)  
summary(regsub)

sumreg <- summary(regsub)   
par(mfrow = c(2,2))   
plot( sumreg$rsq, xlab = "No. of variables", ylab = "R-square", type = "l" )   
plot( sumreg$adjr2, xlab = "No. of variables", ylab = "Adjusted R-square", type = "l" )  
plot( sumreg$cp, xlab = "No. of variables", ylab = "Cp values", type = "l" )  
plot( sumreg$bic, xlab = "No. of variables", ylab = "BIC", type = "l" )

## Selected variables   
par(mfrow=c(1,1))   
plot(regsub, scale = "r2", main="Subset plot based on R2")

lm.best9=lm(formula = TARGET\_deathRate ~ MeanPublicCoverage + PctOtherRace +   
 povertyPercent + PctLowEducation + MeanPrivateCoverage +   
 PctUnemployed16\_Over + PctMarriedHouseholds + BirthRate +   
 PctHighEducation, data = df2)  
lm.best9

AICp(lm.best9)

## [1] -1105.736

pressc(lm.best9)

## [1] 1981.009

lm.full12=lm(formula = TARGET\_deathRate ~ MeanPublicCoverage + PctOtherRace +   
 povertyPercent + PctLowEducation + MeanPrivateCoverage +   
 PctUnemployed16\_Over + PctMarriedHouseholds + BirthRate +   
 PctHighEducation + PctAsian+ PctBlack+AvgHouseholdSize, data = df2) #this model has same variable with lm2.data\_Normal,but variable order is decreasing by SSR   
summary(lm.full12)

anova(lm.full12)

anova(lm.best9,lm.full12)

confint(lm.best9) # b0 =0

lm\_test\_PublicCoverage\_poverty <- lm(TARGET\_deathRate ~MeanPublicCoverage + PctOtherRace + povertyPercent +   
 PctLowEducation + MeanPrivateCoverage  
 + PctUnemployed16\_Over + PctMarriedHouseholds + BirthRate +   
   
 PctHighEducation+ MeanPublicCoverage\*povertyPercent , data = df2) #reduced\_int<-lm.best9   
# Compare models using ANOVA   
anova(lm\_test\_PublicCoverage\_poverty)

anova(lm.best9,lm\_test\_PublicCoverage\_poverty)

# Interaction term of PctUnemployed16\_Over and MeanPublicCoverage   
lm\_interation2 <- lm(TARGET\_deathRate ~MeanPublicCoverage + PctOtherRace +povertyPercent + PctLowEducation + MeanPrivateCoverage +PctUnemployed16\_Over + PctMarriedHouseholds + BirthRate + PctHighEducation + MeanPublicCoverage\*PctUnemployed16\_Over , data = df2)   
summary(lm\_interation2)

anova(lm\_interation2)

anova(lm\_interation2, lm.best9)

partial\_coefficient\_data <- data.frame(  
Variable = c('MeanPublicCoverage', 'PctOtherRace', 'PovertyPercent', 'PctLowEducation',  
'MeanPrivateCoverage', 'PctUnemployed16\_Over', 'PctMarriedHouseholds',  
'BirthRate', 'PctHighEducation'),  
R\_squared = c(1.63, 2.88, 1.97, 4.5, 1.83, 1.09, 1.15, 0.63, 0.5))  
  
partial\_coefficient\_data

plot(lm.best9)

plot(rstudent(lm.best9),type="o",xlab="case index")  
#text(rstudent(lm.best9), labels=rownames(df2), cex=0.7, font=2)   
title("(a) Studentized Delected residuals")  
  
#qqplot of studenidized residuals  
qqPlot(lm.best9,labels=row.names(df2),id.method="identify",simulate=TRUE,main = "Q-Q plot")

group<-rep(1,length(df$AvgHouseholdSize))  
summary(lm.best9$fitted.values)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## -2.21584 -0.36816 -0.02586 0.00000 0.35021 2.38944

group[lm.best9$fitted.values<0]=0  
bftest(lm.best9,group,alpha=0.05)

ncvTest(lm.best9)

## Non-constant Variance Score Test   
## Variance formula: ~ fitted.values   
## Chisquare = 40.22508, Df = 1, p = 2.2632e-10

spreadLevelPlot(lm.best9)

lamda=0.9  
y\_tr=df2$TARGET\_deathRate^lamda # I got

ncvTest(lm.best9\_transf)

## Non-constant Variance Score Test   
## Variance formula: ~ fitted.values   
## Chisquare = 71.25198, Df = 1, p = < 2.22e-16

spreadLevelPlot(lm.best9\_transf)

#cook's distance  
which.max(cooks.distance(lm.best9))

pf(cooks.distance(lm.best9)[1008],9,2895-9)

#splitting df2 into Training & test data   
random\_sample <- createDataPartition(df2 $ TARGET\_deathRate, p = 0.7, list = FALSE)  
training\_dataset <- df2[random\_sample, ]   
testing\_dataset <- df2[-random\_sample, ]  
 model <- lm(TARGET\_deathRate ~MeanPublicCoverage + PctOtherRace +   
 povertyPercent + PctLowEducation + MeanPrivateCoverage +   
 PctUnemployed16\_Over + PctMarriedHouseholds + BirthRate +  
 PctHighEducation, data = training\_dataset)   
predictions <- predict(lm.best9, testing\_dataset)  
  
# computing model performance metrics  
 data.frame(R2 = R2(predictions, testing\_dataset $ TARGET\_deathRate),  
 MSPE = RMSE(predictions, testing\_dataset $ TARGET\_deathRate)^2, #MSPE   
 RMSE = RMSE(predictions, testing\_dataset $ TARGET\_deathRate), #Root mean squared error   
 Relative\_error=( RMSE(predictions,testing\_dataset$TARGET\_deathRate))/mean(testing\_dataset$TARGET\_deathRate)) #Relative error = RMSE/ mean(Y)

## R2 MSPE RMSE Relative\_error  
## 1 0.3242369 0.6808652 0.8251456 180.2893

model2 <- lm(TARGET\_deathRate ~MeanPublicCoverage + PctOtherRace +   
 povertyPercent + PctLowEducation + MeanPrivateCoverage +   
 PctUnemployed16\_Over + PctMarriedHouseholds + BirthRate +  
 PctHighEducation + MeanPublicCoverage\*PctUnemployed16\_Over , data = training\_dataset)   
predictions2 <- predict(model2, testing\_dataset)  
anova(model, model2)

# computing model performance metrics  
 data.frame(R2 = R2(predictions2, testing\_dataset $ TARGET\_deathRate),  
 MSPE = RMSE(predictions2, testing\_dataset $ TARGET\_deathRate)^2, #MSPE   
 RMSE = RMSE(predictions2, testing\_dataset $ TARGET\_deathRate), #Root mean squared error   
 Relative\_error=( RMSE(predictions2,testing\_dataset$TARGET\_deathRate))/mean(testing\_dataset$TARGET\_deathRate)) #Relative error = RMSE/ mean(Y)