Assignment 1: A Linear Regression Theory

Shashwat Patel
Metallurgical and Materials Engineering
Indian Institute of Technology,Madras
mm19b053@smail.iitm.ac.in

***Abstract*—Linear regression is a mathematical technique of fitting the given data to a function. This technique is best known for fitting straight lines. In this paper, we explain the theory behind Linear regression and implement this technique on a data set. Using linear regression we examine whether low income groups are at greater risk for being diagnosed and dying from cancer.**

***Index Terms*—Cancer, Linear regression, Ordinary Least Square(OLS), Cost function,residuals**

I. Introduction

"Cancer is referred to as an ailment characterised by an unrestrained growth of abnormal cells which if untreated and unchecked eventually kills the patient". There are near about 100 types of cancers affecting human body.In most people's minds there is no scarier diagnosis than that of cancer. It is the second leading cause of death as per the world health organization. The cancer burden continues to grow globally, exerting tremendous physical, emotional and financial strain on individuals, families, communities and health systems. Cancer is often thought of as an untreatable, unbearably painful disease with no cure but in reality it is not so. According to a recent study by American Cancer Society epidemiologists, it is estimated than in 2021, there will be 1.9 million new cancer cases diagnosed and 608,570 cancer deaths in the United States. [1]

Linear regression is a very useful and widely used statistical learning method. It is a process to estimate the relationship between variables, where the focus is towards identifying a linear relationship between a dependent variable(response) and one or more independent variables (predictors). Specifically, linear regression helps to understand how the dependent variable changes linearly when any one of the independent variables is changed, while other independent variables are kept fixed. The general formula for linear regression is:

$$Y = \beta X + \epsilon$$

Our main goal in linear regression is to estimate the coefficients β, which helps in identifying the relationship between the dependent and independent variable.Ordinary least square method is a very known technique used to estimate the coefficients.

Cancer, still remains a very problematic disease in the developed countries like USA. The goal of this paper is to study whether low income groups are affected more as compared to other people due to cancer. Our other goal is to study whether socioeconomic factors like Income, Poverty, Population density, education etc. effect the mortality rate due to cancer. This study will help in identifying groups that can be benefited more from fundraising and Cancer research.

This paper majorly deals with the theory of Linear regression and its technical aspects about how straight lines are fit on the data and the mathematics behind it. This paper also helps in studying the effect of different social-economic factors on cancer incidence rate and mortality rate in different states of US by implementation of Linear regression technique.

II. Linear Regression

Linear regression is an approach for modeling the relationship between a scalar dependent variable Y and one or more predictors (or independent variables) denoted X. The case of one predictor only is called simple linear regression. For more than one predictor, the process is called multiple linear regression. Polynomial regression is generalisation of multiple linear regression which includes higher power of the predictors. In linear regression, the relationships are modeled using linear predictor functions whose unknown model parameters are estimated from the data.

The general formula for linear regression is:

$$Y = \beta X + \epsilon \tag{1}$$

where,
Y is the dependent variable
X is the independent variable
β is the regression coefficient
ϵ is a mean-zero random error term

For multiple linear regression, the formula is expressed in this way:

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{1,1} & x_{1,2} & \dots & x_{1,p-1} \\ 1 & x_{2,1} & x_{2,2} & \dots & x_{2,p-1} \\ 1 & x_{3,1} & x_{3,2} & \dots & x_{3,p-1} \\ 1 & x_{4,1} & x_{4,2} & \dots & x_{4,p-1} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n,1} & x_{n,2} & \dots & x_{n,p-1} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \\ \vdots \\ \beta_{p-1} \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \\ \epsilon_4 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

The first step in linear regression model is to estimate the coefficient β using the training data. The main goal is to minimize the value of cost function [2]:

$$J(\beta) = \frac{1}{2N} \sum_{i=1}^{N} (\hat{y_i} - y_i)^2 \tag{2}$$

where,
N is the total number of sample
$\hat{y_i}$ is the predicted output
y_i is the given output

The objective is to find β that minimizes the cost function J. There are several methods to minimize the cost function J like Ordinary Least Square(OLS), Gradient Descent, Grid Search, maximum likelihood approach etc. The ordinary least square is the most commonly used method in general.

A. Methods to estimate the coefficients

1) Ordinary Least Square: In OLS, given a regression line through the data we calculate the distance from each data point to the regression line, square it, and sum all of the squared errors together. This is the quantity that ordinary least squares seeks to minimize.This quantity is commonly known as residual sum of squares(RSS). This approach treats the data as a matrix and uses linear algebra operations to estimate the optimal values for the coefficients.

Firstly, residual(e) is defined as the difference between observed output and predicted output. The estimates of the coefficients are referred as $\hat{\beta}$.Then vector of residuals e is given by:

$$e = Y - X\hat{\beta} \tag{3}$$

The sum of squared residual is given by $e^T e$. The RSS can be written as:

$$e^T e = (Y - X\hat{\beta})^T (Y - X\hat{\beta})$$
$$e^T e = Y^T Y - 2\hat{\beta}^T X^T Y + \hat{\beta}^T X^T X \hat{\beta} \tag{4}$$

To find $\hat{\beta}$, derivative of Eq. 4 is taken with respect to $\hat{\beta}$,

$$\frac{\partial e^T e}{\partial \hat{\beta}} = -2X^T Y + 2X^T X\hat{\beta} \tag{5}$$

From Eq. 5, the normal equation is obtained.

$$(X^T X)\hat{\beta} = X^T Y \tag{6}$$

Here, $X^T X$ is a square matrix, and always a symmetric matrix. So the inverse of $X^T X$ exists. Pre-multiplying both sides by the inverse gives us

$$(X^T X)^{-1}(X^T X)\hat{\beta} = (X^T X)^{-1} X^T Y \tag{7}$$

Here,

$$(X^T X)^{-1}(X^T X) = I$$

where, I is the identity matrix.This gives us:

$$I\hat{\beta} = (X^T X)^{-1} X^T Y$$
$$\hat{\beta} = (X^T X)^{-1} X^T Y \tag{8}$$

One must have enough memory to fit the data and perform the above matrix operations. This procedure is very fast to calculate computationally. [3]

2) Gradient Descent: This method works by starting with random values for each coefficient. The sum of the squared errors are calculated for each pair of input and output values. A learning rate is used as a scale factor and the coefficients are updated in the direction towards minimizing the error. The process is repeated until a minimum sum squared error is achieved or no further improvement is possible.In practice, it is useful when you have a very large dataset either in the number of rows or the number of columns that may not fit into memory.

B. Assessing accuracy of coefficient estimates

After estimating the coefficients, it is important to assess the accuracy of coefficient estimates. Statistical analysis like Hypothesis testing, t-statistics and p-values are done to assess the accuracy of coefficient estimates [4]. In hypothesis testing, if the null hypothesis H_0 is rejected, we conclude that at least one of the regression coefficients is non-zero; hence at least one of the X variables is useful in predicting Y. If H_0 is not rejected, then we cannot conclude that any of the X variables is useful in predicting Y.

C. Assessing accuracy of model

After assessing accuracy of coefficient estimates, the assessing quality of a linear regression fit is necessary. This is assessed using two quantities mainly: the residual standard error(RSE) and the R^2 statistics. Here,

$$RSE = \sqrt{\frac{1}{N-2}\sum_{i=1}^{N}(y_i - \hat{y_i})^2} \tag{9}$$

The RSE is considered as a measure of the lack of fit of the model to the data. The R^2 statistics gives the proportion of variation in the response variable explained by the explanatory variables. It takes values between 0 to 1 [4]. An R^2 statistics that is close to 1 indicates that a large proportion of the variability in the response has been explained by the regression and R^2 statistics close to 0 indicates regression did not explain much of the variability in the response.

There are certain assumptions made while preparing the Linear regression model [4]. These assumptions should be kept in mind:
1.Linearity: The relationship between X and the mean of Y is linear.
2.Homoscedasticity: The variance of residual is the same for any value of X.
3.Independence: Observations are independent of each other.
4.Normality: For any fixed value of X, Y is normally distributed.

III. The problem

In our problem we have fitted a linear regression model to identify how socioeconomic factors affect the mortality rate of cancer in the states of the USA.

The dataset consists of features like, total number of people below poverty line, Female/Male population below poverty line, median income in general as well as median income by ethnicity, total number of people having as well as not having a medical insurance, Female/Male population having as well as not having an insurance, Incidence rate, recent trend in incidence and mortality rate.

*Data Cleaning:*The columns with Median Income based on ethnicity have a lot of missing data, so those columns have been removed to prevent noise in the model. The missing data in recent trend column has been replaced with ”stable” trend, as those rows had ”*” as the value which means less than 16 reported cases. For other numerical missing data, MICE algorithm has been used for imputation [5]. MICE assumes that the missing data are Missing at Random, which means that the probability that a value is missing depends only on observed value and can be predicted using them. By default, linear regression is used to predict continuous missing values.

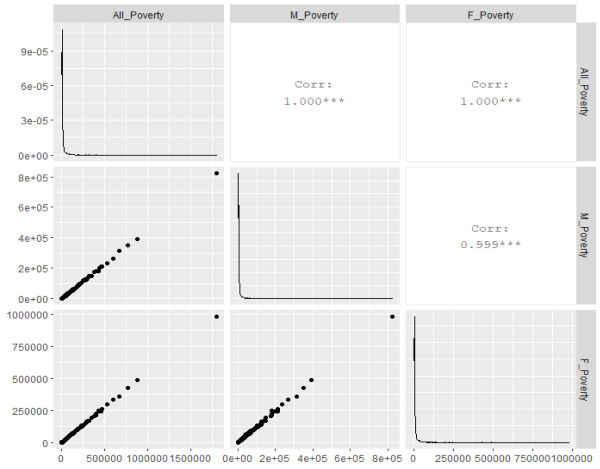

Fig. 1. Correlation

From Fig 1, It is seen that there is a very strong correlation ship between Total Poverty and Male/Female Poverty, so Male poverty and female poverty columns are removed as features as these variables are redundant. Similar case is for All_With Insurance, Male/Female Insurance and All_Without Insurance, Male/Female without Insurance. This is due to the fact that Male Poverty+Female Poverty=All Poverty, i.e. the variables are linearly dependent so the variables become redundant.

Finally, the columns used to make the model are All Poverty,MedIncome,All With Insurance, All Without Insurance,Incidence Rate,Avg annual incidence,Avg annual death, recent trend.One-hot encoding has been done to create dummy variable of recent trend column.

*Visualizations and insights:*From Fig. 2, the mortality rate values looks like they are normally distributed but Fig 3. in reality shows that the the distribution is right skewed.Fig 3 is a quantile-quantile plot which is a tool to explore how a batch of numbers deviates from a theoretical distribution.

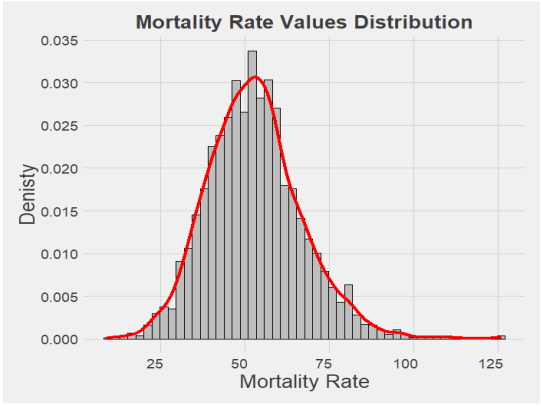

Fig. 2. Distribution of Mortality Rate

One peculiar case found is the case of Los Angeles, California area which reported the maximum of male and female population below poverty line. The mortality rate reported there is of 31.1 which is below the average mortality rate of the USA(Avg Mortality Rate is 53). The incidence rate is also in falling trend in this area.

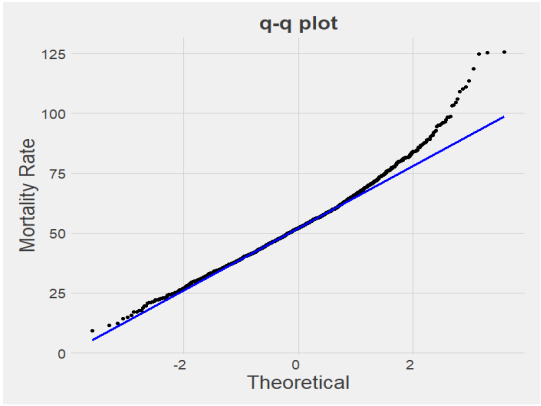

Fig. 3. q-q plot of Mortality Rate

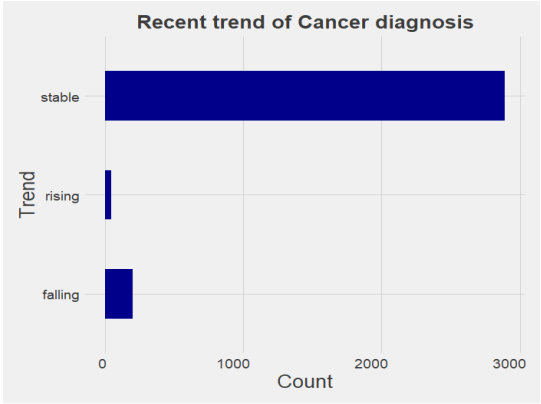

Fig. 4. Recent Trends

From Fig 4, we see most of the states of the USA have a stable trend of Cancer diagnosis but Iowa(IA) and Missouri(MO) are the two states having most of the areas where there is rising trend of Cancer Diagnosis.

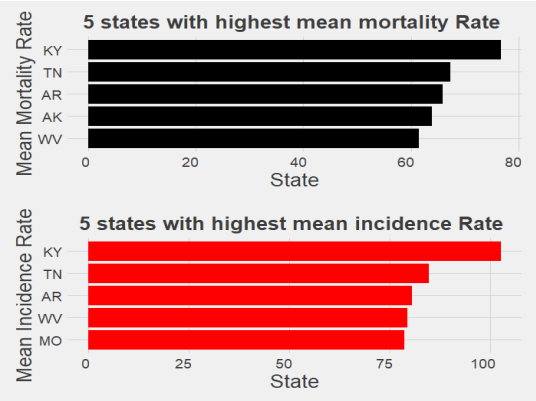

Fig. 5. Top 5 states

Fig. 5 shows the top 5 states with highest mortality rate and incidence rate respectively. Kentucky(KY), Tennessee(TN), Arkanas(AR) are the states where mortality rate as well as Incidence rate is quite high. On further analysis it is found that Kentucky and Tennessee have the state-averaged median income less than the averaged median income of USA but Arkanas has the state-averaged income above the averaged median income of USA.

In Fig 6. the plot between the mortality rate and total poverty, most of the values of total poverty is lesser than 2,50,000 suggesting other values as outliers, similar case is seen in the plot of Mortality Rate and People with Insurance(Fig 6.), most of the values of amount of people having insurance is lesser than 20,00,000 suggesting other values as outliers. These outliers can cause noise in the model and should be checked accordingly.

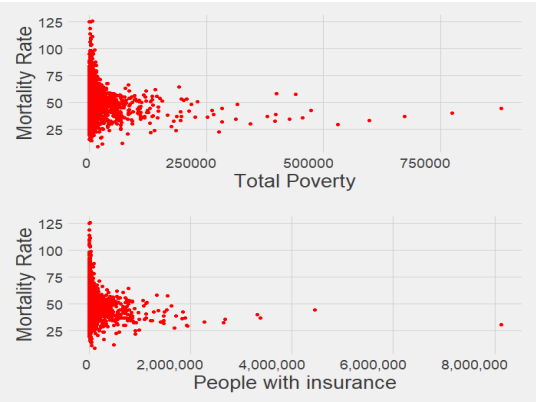

Fig. 6.

In Fig.7, the plot between Mortality rate and Income shows a general trend that mortality rate is lesser for those people having high income. This can be due to the fact that higher income people can afford treatments of cancer like chemotherapy, which is generally quite expensive. The income variable is one of the most important factor in checking whether cancer mortality rate is affected due to socio-economic factors.

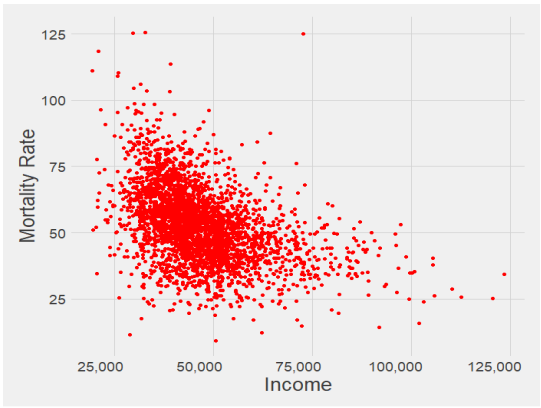

Fig. 7. Mortality Rate vs Income

On fitting the linear regression model using ordinary least square(OLS) method we obtain the following result as shown in Fig 8. The p-value for People with Insurance as well as People without insurance is quite high, showing that these variable are statistically insignificant for fitting the model. The Median Income, Incidence rate, Average annual incidence, Average annual death, Incidence rate, All poverty columns are quite statistically significant.

```
Residuals:
    Min      1Q  Median      3Q     Max
-59.993  -3.338  -0.634   3.215  29.874

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)       1.403e+01  8.729e-01  16.076  < 2e-16 ***
All_Poverty      -3.565e-05  1.134e-05  -3.143  0.00169 **
Med_Income       -1.429e-04  1.199e-05 -11.917  < 2e-16 ***
All_With          4.164e-06  2.606e-06   1.597  0.11026
All_Without      -4.823e-06  9.280e-06  -0.520  0.60334
Incidence_Rate    6.539e-01  7.283e-03  89.786  < 2e-16 ***
Avg_Ann_Incidence -1.493e-01  7.790e-03 -19.168  < 2e-16 ***
Avg_Ann_Deaths    2.103e-01  1.096e-02  19.188  < 2e-16 ***
falling           1.087e+00  5.075e-01   2.141  0.03233 *
rising           -7.671e-01  1.004e+00  -0.764  0.44477
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6.529 on 3124 degrees of freedom
Multiple R-squared:  0.7853,    Adjusted R-squared:  0.7847
F-statistic:  1270 on 9 and 3124 DF,  p-value: < 2.2e-16
```

Fig. 8. Linear Regression Model

```
Residuals:
    Min      1Q  Median      3Q     Max
-59.967  -3.368  -0.604   3.293  29.951

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)       1.372e+01  8.490e-01  16.155  < 2e-16 ***
All_Poverty      -2.878e-05  5.496e-06  -5.236 1.75e-07 ***
Med_Income       -1.356e-04  1.102e-05 -12.307  < 2e-16 ***
Incidence_Rate    6.533e-01  7.266e-03  89.911  < 2e-16 ***
Avg_Ann_Incidence -1.466e-01  7.479e-03 -19.603  < 2e-16 ***
Avg_Ann_Deaths    2.106e-01  1.088e-02  19.356  < 2e-16 ***
falling           1.141e+00  5.065e-01   2.252   0.0244 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6.529 on 3127 degrees of freedom
Multiple R-squared:  0.7851,    Adjusted R-squared:  0.7847
F-statistic:  1904 on 6 and 3127 DF,  p-value: < 2.2e-16
```

Fig. 9. 2nd Linear Regression Model

The adjusted R^2 is 0.7847(Fig 8.), meaning that the model accounts for 78% of the total observed variance in the target variable.The root mean squared error is 42.533. On fitting a 2nd linear regression model(Fig 9.) without insurance data, the adjusted R^2 value remains the same, suggesting that the insurance data doesn't effect the mortality rate. This is quite sup rising, as it is expected that people with insurance can easily fund the cancer treatment. The Pearson R correlation

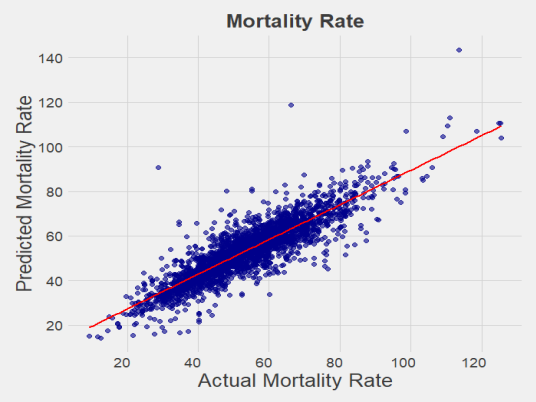

Fig. 10. Predicted Vs Actual Mortality Rate

coefficient value between the predicted mortality rate and actual mortality rate is 0.8860, which is quite consistent with our reported R^2 value. Fig 10. shows that a good linear fit was obtained.

The plot of residuals versus fitted values(Fig 11.) shows that

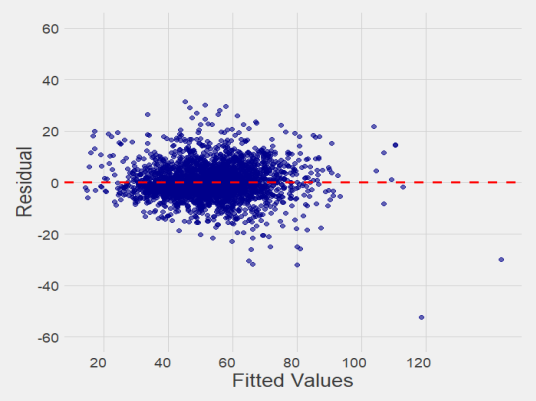

Fig. 11. Residual Vs Fitted Values

the residuals are fairly symmetric relative to the fitted values.

Finally, the model suggests that Cancer mortality actually depends on parameters like Income, Total Poverty, Incidence Rate, Average annual death, average annual incidence. Insurance data doesn't seem to be much important for cancer mortality rate.

IV. Conclusion

From the model, it is learnt that socio-economic factors like income, poverty actually effect the cancer mortality rate. It is learnt that with more income, mortality rate tends to decrease and it is due to the fact that people can afford the expensive cancer treatment. The states and area where number of people below poverty line is quite high reported high mortality rate. States where there high incidence rate of cancer, the mortality rate is also quite high in those states. It is safe to say that low income groups are at greater risk for being diagnosed and dying from cancer.

Further improvements in the model is highly possible. Using the data of population in different areas of US as variable would likely improve the model. The mortality rate data given is based on "per capita", so normalizing our other data based on "per capita" would really be helpful and improve our model and that's why using population data is highly suggested. While imputing for missing values, MICE algorithm was directly used instead of that different methods of imputation can be used based on further research on the data. In the present model, outliers were not accounted for, further work can done on addressing the outliers present in the data. The data on "Income on the basis of ethnicity" was removed due to lot of missing values because imputation might have led to noise in the data but further work is possible to check whether mortality rate is dependent on "income based on ethnicity". Segregation of male and female data can be done to check who is more affected due to cancer. Other methods of fitting a linear regression model like grid search, gradient descent can be done as well.

References

[1] https://www.cancer.org/content/dam/cancer-org/research/cancer-facts-and-statistics/annual-cancer-facts-and-figures/2021/cancer-facts-and-figures-2021.pdf

[2] Hettiarachchi, Changa. (2021). Machine Learning : Model and Cost Function. 10.13140/RG.2.2.16338.27844.

[3] https://web.stanford.edu/ mrosenfe/soc_meth_proj3/matrix_OLS_NYU_notes.pdf

[4] Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani. An Introduction to Statistical Learning : with Applications in R. New York :Springer, 2013.

[5] https://cran.r-project.org/web/packages/mice/mice.pdf

Code Documentation

Load the required Libraries

library(readxl)

library(tidyverse)

library(stringr)

library(mltools)

library(mice)

library(ggthemes)

library(GGally)

library(scales)

library(gridExtra)

library(Metrics)

Reading the data

all_data <- read_excel("merged_data.xlsx",na=c(" ","**"))

The FIPS, fips_x fips_y is the same (used for identification of area), so those columns are removed.

all_data<-all_data[,c(-1,-20,-24)]

Checking basic structure of dataset

summary(all_data)

str(all_data)

head(all_data)

Data Cleaning and processing

- Some incidence rate data had "#" written at the end with the number, below code extracts out the number.

check_1<-str_extract_all(all_data$Incidence_Rate,pattern = “\d+.?\d+”)

all_data$Incidence_Rate<-as.numeric(all_data$Incidence_Rate)

all_data$Incidence_Rate<-check_1

- Function to check number of NAs column-wise

sapply(all_data,function(x) sum(is.na(x)))

- Replacing "_“,”*" in the columns with NAs, where these are present

all_data$Incidence_Rate<-str_replace_all(all_data$Incidence_Rate,pattern = " *",replacement=NA_character*)

all_data$Avg_Ann_Incidence<-str_replace_all(all_data$Avg_Ann_Incidence,pattern=" *",replacement=NA_character*)

all_data$recent_trend<-str_replace_all(all_data$recent_trend,pattern = " *",replacement=NA_character*)

all_data$Mortality_Rate<-str_replace_all(all_data$Mortality_Rate,pattern = "*",replacement=NA_character_)

all_data$Avg_Ann_Deaths<-str_replace_all(all_data$Avg_Ann_Deaths,pattern="*",replacement=NA_character_)

- The "*" and NAs present in recent_trend column are replace with “stable” as , those values mean less than 16 reported cases.

all_data$recent_trend<-str_replace_all(all_data$recent_trend,"*“,”stable")

all_data$recent_trend<-replace_na(all_data$recent_trend,“stable”)

One hot encoding on recent trend column to create dummy variable

recent_trend<-as.data.frame(all_data$recent_trend)

recent<-data.table::data.table(recent_trend)

encoder<-one_hot(recent)

colnames(encoder)<-c(“falling”,“rising”,“stable”)

all_data<-cbind(all_data,encoder)

remove the recent_trend column

all_data<-all_data[,-21]

Converting incidence rate, average annual incidence, average annual death to numeric data

all_data$Avg_Ann_Deaths<-as.numeric(all_data$Avg_Ann_Deaths)

all_data$Mortality_Rate<-as.numeric(all_data$Mortality_Rate)

all_data$Avg_Ann_Incidence<-str_replace_all(all_data$Avg_Ann_Incidence,“3 or fewer”,“3”)

all_data$Avg_Ann_Incidence<-as.numeric(all_data$Avg_Ann_Incidence)

Lots of NAs in Median income by ethnicities, so removed them only kepping Median Income total

all_data<-all_data[,c(-8,-9,-10,-11,-12)]

Impute values using MICE algorithm

all_data_no_catgr_var<-all_data[,c(-1,-2)]

imputed_Data <- mice(all_data_no_catgr_var, m=5, maxit = 50, method = 'pmm', seed = 500)

completeData <- complete(imputed_Data,2)

Visualizatons

Fig 1- ggpairs(completeData[,c(1,2,3)])

Fig 2- completeData%>%ggplot()+geom_histogram(aes(Mortality_Rate,y=..density..),bins=50,fill="grey", color="black")+geom_density(aes(Mortality_Rate),color="red",size=1.5)+ theme_fivethirtyeight()+theme(plot.title = element_text(hjust = 0.5,size=20), axis.title=element_text(size=20), axis.text = element_text(size=14), axis.text.x = element_text(hjust=0.8))+ labs(x="Mortality Rate",y="Denisty",title = "Mortality Rate Values Distribution")+ scale_x_continuous(breaks = seq(0,125,25))+scale_y_continuous(breaks = seq(0,0.04,0.005))

Fig 3- gplot(aes(sample=Mortality_Rate),data=completeData)+stat_qq(distribution=qnorm,size=1.6)+ stat_qq_line(color="blue",size=1.1)+ theme_fivethirtyeight()+theme(plot.title = element_text(hjust = 0.5,size=20), axis.title = element_text(size=20), axis.text = element_text(size=14), axis.text.x = element_text(hjust=0.8))+ labs(x="Theoretical",y="Mortality Rate",title = "q-q plot")

Fig 4- recent_trend%>%ggplot()+geom_bar(aes(trend),fill="darkblue",width=0.5)+coord_flip()+ theme_fivethirtyeight()+theme(plot.title = element_text(hjust = 0.5,size=20), axis.title = element_text(size=20),axis.text = element_text(size=14), axis.text.x = element_text(hjust=0.8))+ labs(x="Trend",y="Count",title = "Recent trend of Cancer diagnosis")

Fig 5-

p1<-all_data%>%group_by(State)%>% summarise(mean_mortality_rate=mean(Mortality_Rate,na.rm=TRUE))%>% mutate(State=reorder(State,mean_mortality_rate))%>%top_n(5)%>% ggplot(aes(State,mean_mortality_rate))+ geom_col(fill="black")+ coord_flip()+theme_fivethirtyeight()+theme(plot.title = element_text(hjust = 0.5,size=20), axis.title = element_text(size=20), axis.text = element_text(size=14), axis.text.x = element_text(hjust=0.8))+ labs(x="Mean Mortality Rate",y="State",title = "5 states with highest mean mortality Rate")

p2<-all_data%>%group_by(State)%>% summarise(mean_incidence_rate=mean(Incidence_Rate,na.rm=TRUE))%>% mutate(State=reorder(State,mean_incidence_rate))%>%top_n(5)%>%ggplot(aes(State,mean_incidence_rate))+ geom_col(fill="red")+ coord_flip()+theme_fivethirtyeight()+theme(plot.title = element_text(hjust = 0.5,size=20), axis.title = element_text(size=20), axis.text = element_text(size=14), axis.text.x = element_text(hjust=0.8))+ labs(x="Mean Incidence Rate",y="State",title = "5 states with highest mean incidence Rate")

grid.arrange(p1,p2,ncol=1)

Fig 6

p11<-completeData%>%filter(All_Poverty)%>%ggplot(aes(y=Mortality_Rate))+

geom_point(aes(x=All_Poverty),color="red")+ theme_fivethirtyeight()+theme(plot.title = element_text(hjust = 0.5,size=20), axis.title = element_text(size=20), axis.text = element_text(size=14), axis.text.x = element_text(hjust=0.8))+ labs(x="Total Poverty",y="Mortality Rate")

p12<-completeData%>%ggplot(aes(y=Mortality_Rate))+geom_point(aes(x=All_With),color="red")+ scale_x_continuous(labels = comma)+theme_fivethirtyeight()+theme(plot.title = element_text(hjust = 0.5,size=20), axis.title = element_text(size=20), axis.text = element_text(size=14), axis.text.x = element_text(hjust=0.8))+ labs(x="People with insurance",y="Mortality Rate")

grid.arrange(p11,p12,ncol=1)

Fig 7

completeData%>%ggplot(aes(y=Mortality_Rate))+geom_point(aes(x=Med_Income),color="red")+ scale_x_continuous(labels = comma)+theme_fivethirtyeight()+theme(plot.title = element_text(hjust = 0.5,size=20), axis.title = element_text(size=20), axis.text = element_text(size=14), axis.text.x = element_text(hjust=0.8))+ labs(x="Income",y="Mortality Rate")

Fig 8 and 9 are screenshot of the summary of the fitted linear regression models

Fig 10

completeData%>%ggplot(aes(y=predicted,x=Mortality_Rate))+geom_point(color="darkblue",size=2,alpha=0.6)+ geom_smooth(method = lm,fill=NA,color="red")+ theme_fivethirtyeight()+theme(plot.title = element_text(hjust = 0.5,size=20), axis.title = element_text(size=20), axis.text = element_text(size=14), axis.text.x = element_text(hjust=0.8))+ labs(x="Actual Mortality Rate",y="Predicted Mortality Rate",title = "Mortality Rate")+ scale_x_continuous(breaks = seq(0,120,20))+scale_y_continuous(breaks = seq(0,160,20))

Fig 11

completeData%>%ggplot(aes(predicted,residual))+geom_point(alpha=0.6,color="darkblue",size=2)+ geom_hline(yintercept = 0,linetype="dashed",size=1.4,color="red")+scale_x_continuous(breaks = seq(20,120,20))+ scale_y_continuous(breaks = c(-60,-40,-20,0,20,40,60),limits =c(-60,60))+theme_fivethirtyeight()+ theme(plot.title = element_text(hjust = 0.5,size=20), axis.title = element_text(size=20), axis.text = element_text(size=14), axis.text.x = element_text(hjust=0.8))+ labs(x="Fitted Values",y="Residual")

For residual data, (data_needed_model_2 defined below)

data_needed_model_2$residual <- -residuals(lin_reg_model_2)$ $data_needed_model_2$predicted<-predict(lin_reg_model_2)

The model

1st model (Fig 8)

data_needed<-completeData[,c(1,4,9,10,11,12,13,14,15,16)]

lin_reg_model<-lm(Mortality_Rate~.,data=data_needed)

summary(lin_reg_model)

2nd model (Fig 9) without insurance data

data_needed_model_2<-data_needed[,c(-3,-4)]

lin_reg_model_2<-lm(Mortality_Rate~.,data=data_needed_model_2)

summary(lin_reg_model_2)

To calulcate mean squared error

mse(data_needed_model_2$Mortality_Rate,data_needed_model_2$predicted)