Assignment 1: A Mathematical essay on Linear Regression

Vishal Rishi MK
Department of Chemical Engineering
IIT Madras
ch18b013@smail.iitm.ac.in

***Abstract*—Linear regression is a statistical procedure for calculating the value of a dependent variable from an independent variable. Linear regression measures the association between two variables. It is a modeling technique where a dependent variable is predicted based on one or more independent variables. Linear regression analysis is the most widely used of all statistical techniques. Linear regression can not only be viewed as an estimation technique. It also helps us to understand the strength of relationships between variables and their statistical significance (causal analysis). In this article, we explore the relationships between socioeconomic status and cancer mortality (and incidence) rates in several states in the USA using regression analysis.**

Keywords—linear regression, significance, estimation

I. Introduction

The concept of linear regression was first proposed by Sir Francis Galton in 1894. Linear regression is a statistical test applied to a data set to define and quantify the relation between the considered variables. Univariate statistical tests such as Chi-square, Fisher's exact test, *t*-test, and analysis of variance (ANOVA) do not allow taking into account the effect of other covariates/confounders during analyses. However, linear regression allows the researcher to control the effect of confounders in the understanding of the relation between two variables. Linear regression assumes a linear relationship (in parameters) between the independent variables and the dependent variable. It also assumes a random, unpredictable, zero mean error term that is being added to the dependent variable. Using the least squares technique, we estimate the coefficients of the independent variables (including the intercept). During the process of estimation, we make several assumptions on the data and the residuals. Residuals are the difference between the observed values and the predicted values of the dependent variable. If some of the assumptions are broken, our estimates of the coefficients might not be reliable. Hence, it is important to assess the significance of our model and estimates through various testing methods before we use the model to derive insights and conclusions about the data. Otherwise, we would be arriving at spurious correlations between variables.

In this article, we explore relationships between cancer incidence (and mortality) and socioeconomic status in several states in the USA. We have cancer incidence and mortality data for 51 states in the USA along with socioeconomic data for several counties in those states. The features that describe socioeconomic status of a county are 1) Median income 2) Number of people below the poverty line (male and female) 3) Median income of native Americans, Black, White, Asian and Hispanic. Cancer mortality (or incidence) is best described by the Incidence and Mortality rates in a county. Incidence rate of a disease is the number of new cases per 100,000 people at risk. Mortality rate of a disease is the number of new deaths per 100,000 people at risk. Along with this, we also have the average annual incidence and average annual deaths due to cancer in each county.

Section 2 of this article gives a mathematical background to linear regression and the statistical assumptions behind the estimation of the parameters. Section 3 of the article describes the data processing pipeline and several observations, insights, visualisations. We also provide quantitative results after applying regression analysis. Finally, section 4 draws conclusions from our explorations and analysis.

II. Linear Regression

In this section, we develop the linear regression methodology for building a model of the relation between two or more variables of interest on the basis of available data. An interesting feature of this methodology is that it may be explained and developed simply as a least squares approximation procedure, without any probabilistic assumptions. Yet, the linear regression formulas may also be interpreted in the context of various probabilistic frameworks, which provide perspective and a mechanism for quantitative analysis. Suppose that our data consist of triples of the form (***xi*** , ***yi***, ***zi***) and that we wish to estimate the parameters ***θj*** of a model of the form

$$y \approx \theta_0 + \theta_1 x + \theta_2 z$$

We then seek to minimize the sum of the squared residuals

$$\sum_{i=1}^{n} (y_i - \theta_0 - \theta_1 x_i - \theta_2 z_i)^2 \qquad \text{---------- (1)}$$

over all ***θj, j = 0, 1, 2.***

A. Approximate Bayesian LMS estimation (linear model)

Let the pairs ***(Xi , Yi, Zi)*** be random and i.i.d (identical and independently distributed random variables). Let us also make the additional assumption that the pairs satisfy a linear model of the form

$$Y_i = \theta_0 + \theta_1 X_i + \theta_2 Z_i + W_i$$

where the ***Wi*** are Li.d. , zero mean noise terms, independent of ***Xi, Zi***. From the least mean squares property of conditional expectations, we know that **E[*Yi*|*Xi,Zi*]** minimizes the mean squared estimation error. Under our assumptions,

$$E[Y_i|X_i, Z_i] = \theta_0 + \theta_1 X_i + \theta_2 Z_i$$

Thus, the true parameters minimize

$$E[(Y_i - \theta_0 - \theta_1 X_i - \theta_2 Z_i)^2]$$

By the weak law of large numbers, this expression is the limit as **n**→∞ of

$$\frac{1}{n}\sum_{i=1}^{n}(Y_i - \theta_0 - \theta_1 X_i - \theta_2 Z_i)^2 \quad \text{---------- (2)}$$

This indicates that we will obtain a good approximation of the minimizers of $E[(Y_i - \theta_0 - \theta_1 X_i - \theta_2 Z_i)^2]$ (the true parameters), by minimizing the above expression (2) (with ***Xi, Yi*** and ***Zi*** replaced by their observed values ***xi***, ***yi***, and ***zi*** respectively) . But minimizing this expression is the same as minimizing the sum of the squared residuals given in (1).

III. APPLICATION TO THE PROBLEM

The data that we have contains data records for 3134 counties in the USA. The counties belong to one of the 51 states. We have 23 features (categorical and numerical) that describe the socioeconomic status and cancer incidence (and mortality) of the counties.

A. Data Preparation

For linear regression, one of the main assumptions is that the variables are independent and identically distributed. Hence, we sort all the 51 states in descending order based on the number of counties in a particular state. The top eight states having the most number of counties (and hence, the most number of data points) is shown in Table 1 . Figure 1 shows the boxplot distribution of Med_Income (refer Table 1) for the top eight states. From Figure 1, we can see that the states GA and KY have a similar distribution of Med_Income. Figure 2 shows the histogram of Med_Income for the states GA and KY. From this, we can say qualitatively that Med_Income is identically distributed in the states GA and KY. We also need to assure the independence of the observations. Though independence is a strong statement and is hard to prove, we can fairly assume that there is no autocorrelation between the observations. Figure 3 shows the autocorrelation plot and the normal probability plot of Med_Income for the states GA and KY. They show that the assumption of independence is justifiable. We group the data records for the counties in the states GA and KY in a separate dataset and analyse this further.

B. Feature selection and engineering

Along with the median income of the county, we also have the median income of several ethnicities (eg. Hispanic, White, Black, Asian etc). But without the actual composition of the ethnic groups in the entire population, they would not create meaningful correlations with the incidence or mortality rates (which does not take ethnicity into account). The same argument goes with gender. Hence, we can discard variables that indicate ethnicity or gender. Figure 4 shows a scatter plot between All_Poverty and Mortality (and Incidence rates). There are no obvious patterns or trends visible. On the other hand, there is an increasing trend when we plot Mortality (and Incidence rates) against All_Poverty_rate. All_Poverty_rate is the ratio of the number of people below the poverty line in a county. This shows that rates are better features than absolute counts. Hence, we also introduce another variable All_Without_rate, which is the ratio of the number of people without health insurance. Our dataset contains 279 data points and we do encounter some missing values. The missing entity is replaced by the mean of the corresponding variable. Missing values are encountered only in Incidence_Rate and Mortality_Rate. The main features that we worked with are as follows 1) Poverty_rate, 2) All_Without_rate, 3) Med_Income, 4) Incidence_Rate, and 5) Mortality_Rate.

C. Exploratory Data Analysis

Figure 5 shows scatter plots between several combinations of the filtered variables. An important point to note is that the features are highly correlated. Poverty_rate and Med_Income have a very correlation of -0.87. Including both these variables in our regression can lead to unstable coefficients and unreliable results due to multicollinearity. In Figure 5, the scatter plot between Med_Income and Poverty_rate shows qualitatively that Med_Income is a good indicator of Mortality and Incidence rates. Counties with a higher value of Med_Income have higher mortality rates and those with a lower value of Med_Income have lower mortality rates. An interesting thing to note is that the correlation between All_Without_rate and Mortality_Rate is -0.25. This might seem counter intuitive. This might be due to the effect of unaccounted confounding variables. In Figure 5, in the scatter plot between Med_Income and All_Without_rate, for a fixed value of Med_Income, we still observe that the value of Mortality_Rate decreases as All_Without_rate increases. This interesting phenomenon might be due to specific local effects present only in these selected counties. Also in the scatter plot between Incidence_Rate and All_Without_rate, we cannot see any relationship between All_Without_rate and Mortality_Rate, for a fixed value of Incidence_Rate. This indicates that Incidence_Rate could well be one of those confounding variables. We should make sure to include them in our model. Otherwise, we would get biased coefficients leading to spurious correlations. Figure 6 shows the distribution of Incidence_Rate. We can see a considerable number of outliers in the left tail of the distribution. If these outliers are not removed, they might lead to residuals that are not normally distributed. Hence, these outliers are removed, reducing our data points to 255.

D. Regression analysis - Mortality rate

We tried several regression models with different feature combinations. Table 2 shows the results for all the regression models. The model that gave reliable results, in terms of accuracy and statistical significance is shown at the bottom of the table. This model used Med_Income, All_Without_rate and Incidence_Rate as the features and Mortality_Rate as the target variable. It achieved an R - squared value of 0.798. This means that model was able to explain 79.8% of the variance in Mortality_Rate. The statistical significance of the measure is given by the F - score, which is, in our case, equal to 329.9 (p - value ≈ 0). A higher F - score leads to rejecting the null hypothesis that the coefficients of the model are insignificant. The statistical significance of each of the coefficients are given in Table 2. We can see that the coefficient of All_Without_rate is insignificant (p - value = 0.198). This can be qualitatively

verified by the scatter plot between Incidence_Rate and All_Without_rate in Figure 5. We cannot see any relationship between All_Without_rate and Mortality_Rate, for a fixed value of Incidence_Rate. This can be quantitatively verified by looking at the model that uses only Med_Income and Incidence_Rate to model Mortality_Rate (refer Table 2). Though this model does not use the variable All_Without_rate, it's R - squared value does not drop significantly from the best model. This shows that the variable All_Without_rate does not add significant explanatory power to the model. The coefficient of the variable Med_Income in our best model is -1.61. The p - value suggests that it is statistically significant (p - value = 0.007). This suggests that the median income of a county is inversely related to the cancer mortality rate (as seen in Figure 5). The variable Poverty_rate is not included in our final model because of multicollinearity. Table 2 shows such a model with all the variables included. Since Poverty_rate is highly correlated with Med_Income, we can see that the p - values of these variables have been distorted. This is because of the high standard error of the coefficients. To avoid such unreliable estimates, we did not include Poverty_rate in the final model. The reason for including Incidence_Rate in the final model can be easily understood by looking at the model that includes Med_Income and All_Without_rate as the features (refer Table 2). The coefficient of All_Without_rate in this model is -7.5610 (p - value = 0). This is clearly a biased estimate and indicates a spurious relationship. This is due to omitted variable bias. The omitted variable, in this case, is the Incidence_Rate. Though we need to model the relationship between socioeconomic status and cancer mortality (or incidence), we need to include Incidence_Rate in our model so that we do not model incorrect relationships between variables. Once we include it in the final model, we can see that there is no clear relationship between All_Without_rate and Mortality_Rate.

Figure 7 shows the autocorrelation plots of residuals, plots to check no endogeneity, and plots to check the normality of the residuals. The normal probability plot shows that the errors are normally distributed. From the autocorrelation plot of the residuals, we can see that the residuals are linearly independent of each other. Also, the scatter plot between the residuals and each of the explanatory variables show a random scatter. It indicates that the explanatory variables are not able to further explain the residuals (no endogeneity). This also assures us that we are not leaving any confounding variables in our model.

E. Effect of outliers

The presence of outliers can severely affect the results in our model. The results in Table 2 are conducted with the outliers in Incidence_Rate removed from the data. Table 3 shows the results obtained when these outliers are not removed. Once again, we can see the biased estimate for the coefficient of All_Without_rate, even when the confounding variable Incidence_Rate is included. The coefficient of All_Without_rate is also statistically significant (p - value = 0). These are wrong results. This clearly depicts the importance of removing outliers from our data.

IV. Conclusion

This article illustrates the use of linear regression on a practical dataset - cancer mortality. The mathematical background for linear regression is explained. We further employ linear regression to explore relationships between cancer mortality and socioeconomic data. The results show that in the states GA and KY, there is strong evidence that the median income has a strong influence on cancer mortality. Also, the ratio of the number of people without health insurance does not affect the mortality rate in the states GA and KY. The effect of outliers on the results is explained by including the outliers in our analysis and computing the estimates. There is room for improvement. We can extend our analysis to other states as well. We might be able to find different patterns in different states. Similar to modelling mortality rate, we can also try to model incidence rate. But mortality rate cannot be used as an explanatory variable because mortality does not cause incidence. The racial and gender composition of the population can be included to study the socioeconomic effects on cancer mortality within an ethnic or gender group. We can also include several other factors like medical infrastructure, public awareness, food insecurity, lifestyle habits like smoking, obesity etc. These factors might be able to explain cancer incidence which in turn helps us to understand cancer mortality.

States	No. of Counties
TX	254
GA	159
VA	132
KY	120
MO	115
KS	105
IL	102
NC	100

Table 1: The top eight states having the most number of counties (and hence, the most number of data points)

References

[1] D. Bertsekas, and J. Tsitsiklis, Introduction to Probability, 2nd ed., Massachusetts Institute of Technology, 2008, pp.475-483

[2] D. Montgomery, C. Jennings, and M. Kulahci, Introduction to Time Series Analysis and Forecasting, Wiley, 2008, pp.73-138

[3] J. O'Connor, T. Sedghi, M. Dhodapkar, M. Kane, and C. Gross, "Factors associated with cancer disparities among low-, medium-, and high-income US counties ", 2018

[4] J. Kim, "Multicollinearity and misleading statistical results", 2019

[5] G. Wilkes et al., "Cancer and poverty: breaking the cycle", 1994

	coef	p val	std err	F val	R squared
cons	65.1	0	0.47	246.5	0.798
x0	-0.1	0.96	0.957	246.5	0.798
x1	-1.6	0.008	0.966	246.5	0.798
x2	0.81	0.198	0.63	246.5	0.798
x3	14.6	0	0.60	246.5	0.798

	coef	p val	std err	F val	R squared
cons	65.1	0	0.87	59.32	0.32
x1	-9.1	0.0	0.936	59.32	0.32
x2	-7.5	0.0	0.936	59.32	0.32

	coef	p val	std err	F val	R squared
cons	65.1	0	0.47	492.8	0.796
x1	-2.1	0.00	0.51	492.8	0.796
x3	14.2	0.0	0.51	492.8	0.796

	coef	p val	std err	F val	R squared
cons	65.1	0	0.47	329.9	0.798
x1	-1.6	0.007	0.59	329.9	0.798
x2	0.79	0.198	0.61	329.9	0.798
x3	14.6	0	0.60	329.9	0.798

Table 2: x0 : Poverty_rate, x1 : Med_Income, x2 : All_Without_rate, x3 : Incidence_Rate. The final model is the last table

	coef	p val	std err	F val	R squared
cons	64.9	0	0.72	83.9	0.478
x1	-6.6	0	0.83	83.9	0.478
x2	-4.9	0	0.84	83.9	0.478
x3	7.07	0	0.79	83.9	0.478

Table 3: x1 : Med_Income, x2 : All_Without_rate, x3 : Incidence_Rate. Model built with outliers

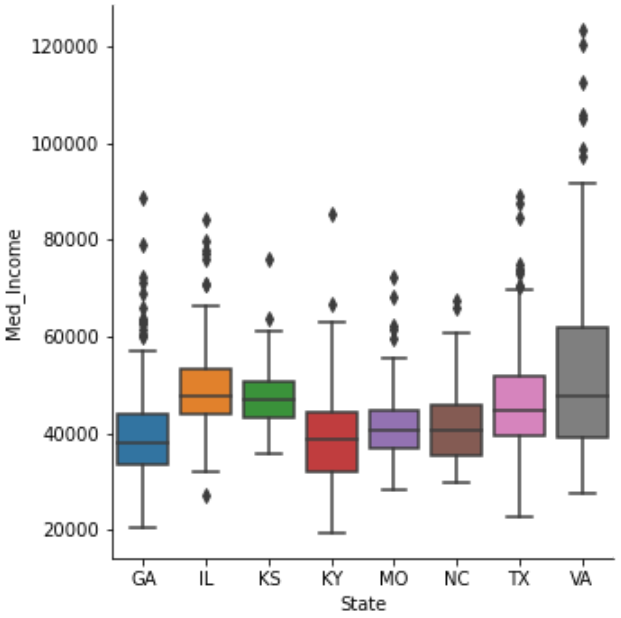
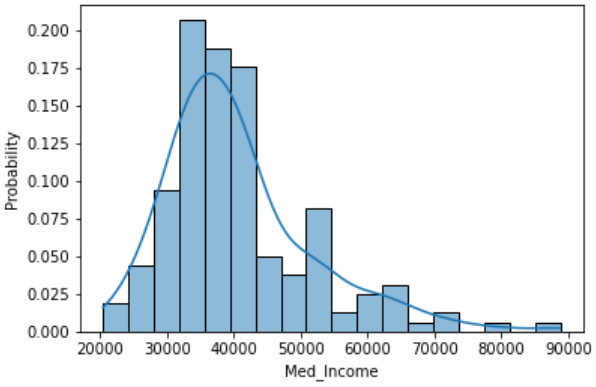


Figure 1: Boxplot distribution of Med_Income for the top eight states

a)



b)

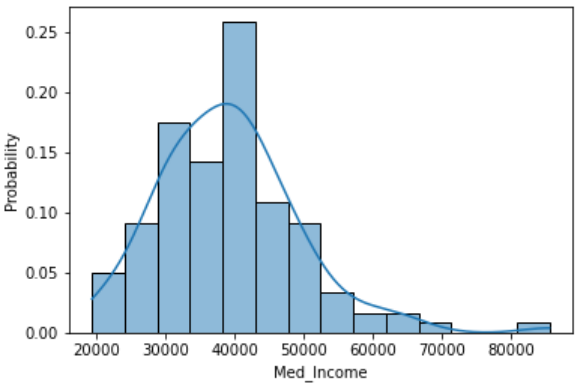
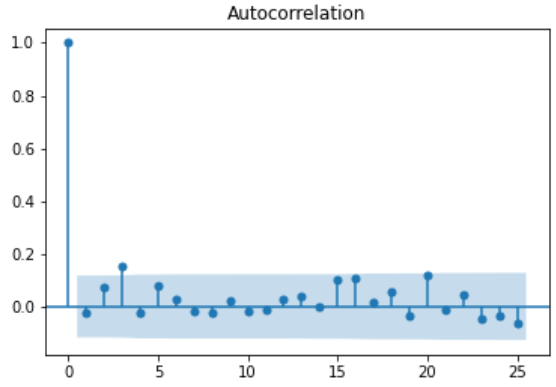


Figure 2: a) Histogram of Med_Income for the state GA b) Histogram of Med_Income for the state KY



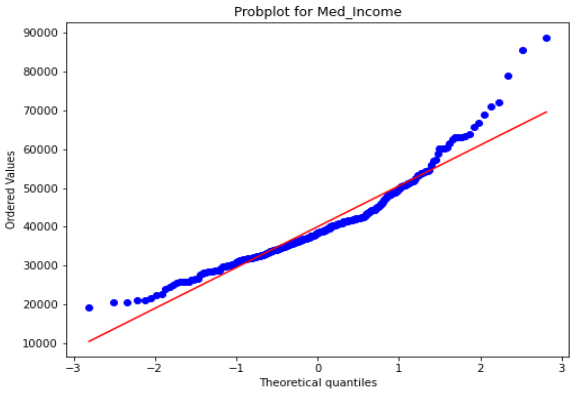
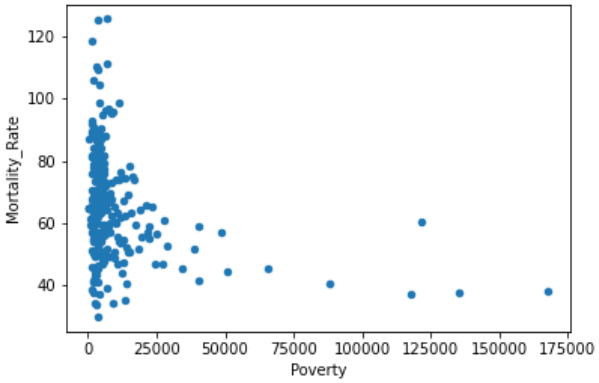


Figure 3: Autocorrelation plot and the normal probability plot of Med_Income for the states GA and KY



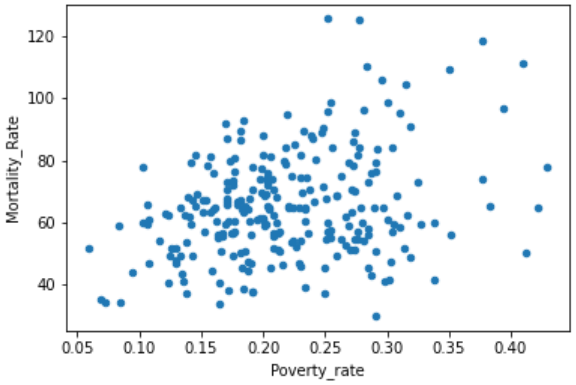
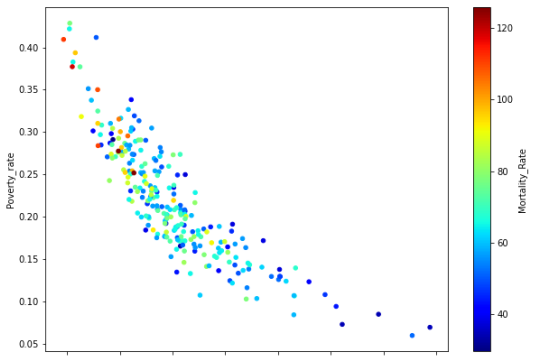
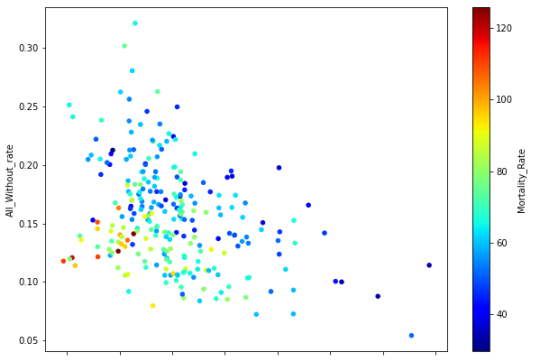


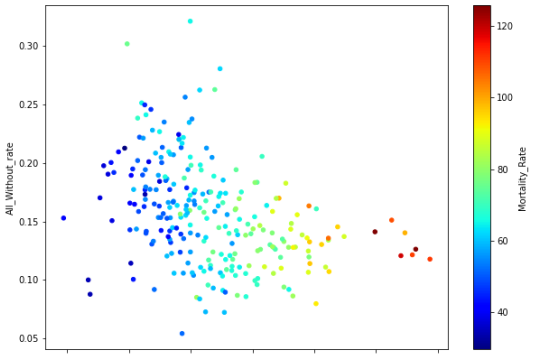
Figure 4: Effect of All_Poverty and Poverty_rate on Mortality rate



a) Poverty_rate vs Med_Income



b) All_Without_rate vs Med_Income



c) All_Without_rate vs Incidence_Rate

Figure 5: a), b), c) Scatter plots between several combinations of the filtered variables

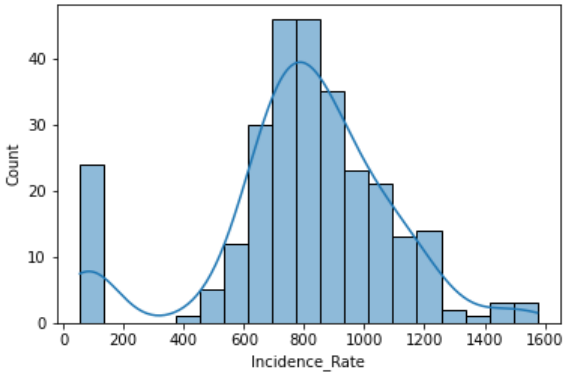
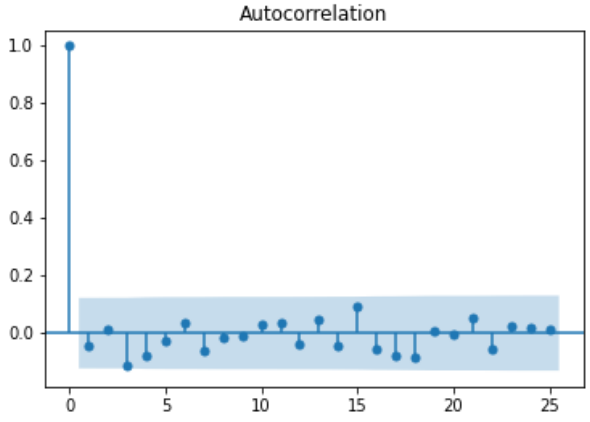
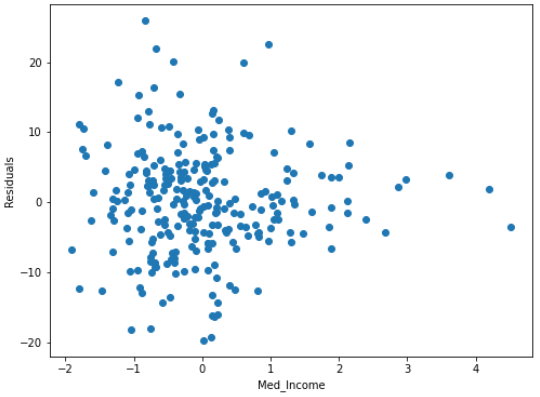


Figure 6: Distribution of Incidence_Rate. .

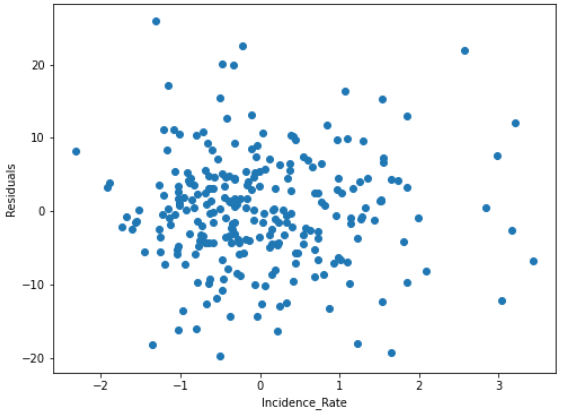
a)



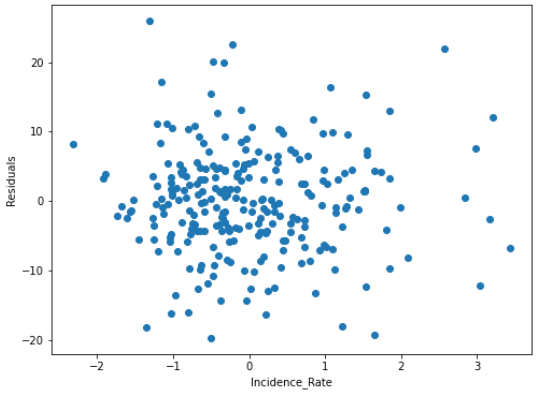
b)



c)



d)



e)

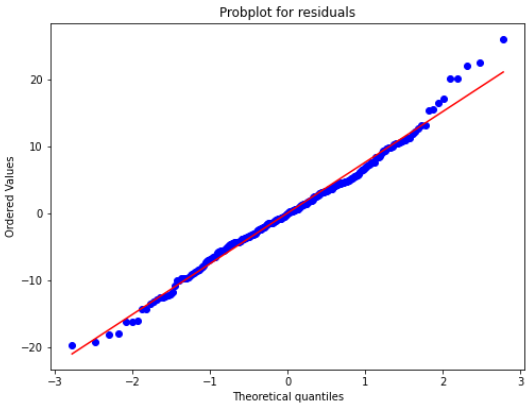


Figure 7: a) Autocorrelation between residuals,
b), c), d) scatter plots between explanatory variables and residuals,
e) Normal probability plot of residuals