Correlation Analysis Using Linear Regression

Assignment 1 : EE4708 Data Analytics Laboratory

Madhur Jindal
Inter Disciplinary Dual Degree Program in Data Science
Indian Institute of Technology (IIT) Madras
Chennai, India
me18b059@smail.iitm.ac.in

***Abstract*—This document is a mathematical essay on Linear Regression wherein it is applied on a dataset involving Income, Poverty & Health insurance data which are the measures of socio-economic status along with Mortality and Incidence data for various states in the United States. The aim for the assignment is to determine whether or not cancer incidence and mortality are correlated with socioeconomic status, which is to be presented both quantitatively and visually.**

I. Introduction

Linear regression is used for finding linear relationship between target and one or more predictors. Linear regression is a task in which the model being learnt is assumed to be a linear function of parameters. There are three types of linear regression - Simple and Multiple, which involve one and more independent variables respectively. The third one is Polynomial regression which is a generalisation case of multiple linear regression in which the input features include higher powers. One is a predictor and other is the response. It looks for statistical relationship but not deterministic relationship. Relationship between two variables is said to be deterministic if one variable can be accurately expressed by the other. For example, using temperature in degrees it is possible to accurately predict Fahrenheit. Statistical relationship is not accurate in determining relationship between two variables. For example, relationship between height and weight.

II. Linear Regression

A. Different Approaches

1) **Generative model approach**
(I) Model p(t, x) = p(x—t)p(t)
(I) Use Bayes' theorem p(t—x) = $\frac{p(x|t)p(t)}{p(x)}$
(D) Take conditional mean/median/mode/any other optimal decision outcome as y(x)

2) **Discriminative model approach**
(I) Model p(t, x) directly
(D) Take conditional mean/median/mode/any other optimal decision outcome as y(x)

3) **Direct regression approach**
(D) Learn a regression function y(x) directly from training data

Identify applicable funding agency here. If none, delete this.

B. Model structure

The most general form of Linear Regression model is:

$$f(x, \theta) = \hat{y} = \theta_0 + \sum_{j=1}^{m-1} \theta_j \phi_j(x)$$

where $\hat{y}$ is the predicted output for a given input, x = $(x_1, ..., x_d)^T$, is a d-dimensional input feature vector, $\phi_j(.)$ represents a basis function, $\theta = (\theta_1, ..., \theta_m)^T$ are the model parameters and m is the order of the linear model.

C. Cost Function

Cost function is a measure of how wrong the model is in terms of its ability to estimate the relation between inputs and outputs. There exist different types of cost functions and the most popular among them is the Mean Squared Error (MSE) between the predicted and observed outputs.

The formula mean squared error for a dataset with N samples is given by:

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

where $\hat{y_i}$ is the predicted output and y_i is the observed or given output.
The objective is to find θ that minimizes the cost function J

D. Solution Approaches

Following are the different approaches for finding the parameters of the desired regression model:

1) **Analytical approach:** An approach to find the parameters directly using an equation called the normal equation.
2) **Maximum Likelihood approach:** An probabilistic approach to estimating the model parameters.
3) **Gradient Descent approach:** An optimization approach to minimizing the cost function and finding the parameters.
4) **Grid Search:** A brute force search approach to finding the optimal parameters.

E. Residual Analysis

Randomness and unpredictability are the two main components of a regression model.

Prediction = Deterministic + Statistic

Deterministic part is covered by the predictor variable in the model. Stochastic part reveals the fact that the expected and observed value is unpredictable. There will always be some information that are missed to cover. This information can be obtained from the residual information.

Let's explain the concept of residue through an example. Consider, we have a dataset which predicts sales of juice when given a temperature of place. Value predicted from regression equation will always have some difference with the actual value. Sales will not match exactly with the true output value. This difference is called as residue.

Residual plot helps in analyzing the model using the values of residues. It is plotted between predicted values and residue. Their values are standardized. The distance of the point from 0 specifies how bad the prediction was for that value. If the value is positive, then the prediction is low. If the value is negative, then the prediction is high. 0 value indicates prefect prediction. Detecting residual pattern can improve the model.

Non-random pattern of the residual plot indicates that the model is,

- Missing a variable which has significant contribution to the model target
- Missing to capture non-linearity (using polynomial term)
- No interaction between terms in model

Characteristics of a residue

- Residuals do not exhibit any pattern
- Adjacent residuals should not be same as they indicate that there is some information missed by system.

F. Metrics for model evaluation

1) R-Squared value: This value ranges from 0 to 1. Value '1' indicates predictor perfectly accounts for all the variation in Y. Value '0' indicates that predictor 'x' accounts for no variation in 'y'.

1. Regression sum of squares (SSR)

This gives information about how far estimated regression line is from the average of the actual output.

$$\text{Error} = \sum_{i=1}^{n}(\hat{y} - \bar{y})^2$$

2. Sum of Squared error (SSE)

How much the target value varies around the regression line (predicted value).

$$\text{Error} = \sum_{i=1}^{n}(y - \hat{y})^2$$

3. Total sum of squares (SSTO)

This tells how much the data point move around the mean.

$$\text{Error} = \sum_{i=1}^{n}(y - \bar{y})^2$$

$$R^2 = 1 - \frac{SSE}{SSTO}$$

2) Null-Hypothesis and P-value: The p-value for each term tests the null hypothesis that the coefficient is equal to zero (no effect). A low p-value (¡ 0.05) indicates that you can reject the null hypothesis. In other words, a predictor that has a low p-value is likely to be a meaningful addition to your model because changes in the predictor's value are related to changes in the response variable.

Conversely, a larger (insignificant) p-value suggests that changes in the predictor are not associated with changes in the response.

III. Problem - Understanding and Modelling

We are presented with a problem involving testing of the hypothesis that cancer incidence and mortality are correlated with socioeconomic status.

We are given Poverty, Income and Insurance data as metrics for socioeconomic status and Incidence rate, Average incidence, mortality rate & average deaths as the dependent variables on which the hypothesis is to be tested, for different areas in the United States identified by the FIPS number.

We have information about the state, the name of the area, poverty - total people, males, females below the poverty line; Income - median income of all people, white, black, native americans, hispanic, asians; Insurance - All people with and without insurance, including males and females with and without insurance. We have a total of 3134 data points.

A. Data pre-processing

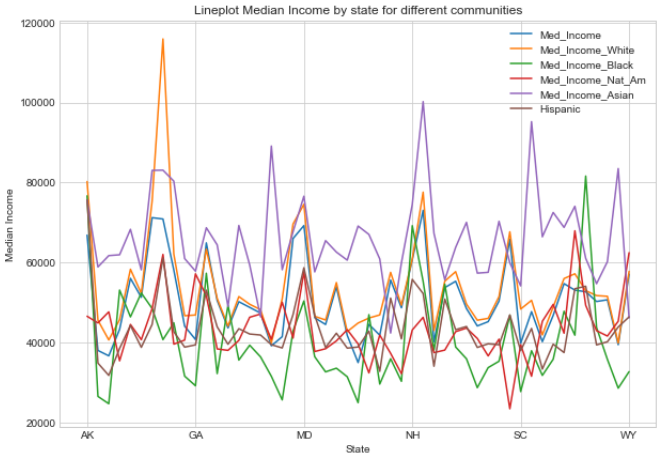

Fig. 1. Line-plot median income by state

We start with checking the effect of social status on the median income as this is the only social data, thus we try to create a line-plot for the median incomes (Figure 1) for the different communities in different states, and thus see a trend wherein the Asians usually are seen to have a higher income, while the income for the blacks is on the lower side. We can see that the different social group have different mean incomes for different states, hence if we can prove that the median income is a valid factor determining the Avg Ann Incidence or Deaths, then we can also assume that social status would also be a valid factor.

We then check on the number of null values in each of the columns, and we find out that the columns with Median Income values for the social groups like Native American, Asian, Black, Hispanic have too many missing values, ranging

from 20 % to 55 % thus we proceed to delete all these columns. We then check for columns with values that are not numeric, and thus find out that the columns Incidence Rate, Avg Ann Incidence, recent trend, mortality rate and Avg Ann deaths has values that are not purely numeric.

A very important observation is that all the independent columns are not normalized by the population and we also do not have population data, thus it is better to train our model on Avg Ann incidence and Avg Ann deaths, rather than using Incidence rate and Mortality rate data as these are just the normalized versions of the average values. Thus we drop these two dependent columns.

On evaluating the data in Avg Ann Deaths column, we see 325 data-points with an asterisk. It represents the data that has been suppressed due to confidentiality when fewer than 16 cases were reported. So we need to deal with this missing data, let's evaluate all the other columns and then proceed to treat the missing data. On evaluating Avg Ann incidence column we see three types of data other than numeric data including '3 or fewer', '_', '__'. which is not much compared to the total data. We replace the '3 or fewer' rows with 3 and the other with null values.

We also use feature extraction to create two new columns corresponding to rising and falling of the recent trend and drop the recent trend column. Now we have some of the data that is null and we need to treat it before training our model which can be done in some ways- One can be to remove the rows (data points) corresponding to the NULL values, other can be imputing the data with the median of the data grouping using the state data. One other method can be using a model that can handle missing data, and this is the method we are going to use in this paper.

B. Visualization

We start by plotting scatter plots for Avg Ann Incidence vs Avg Ann Deaths (Figure 2) and see that there is very high correlation between these, hence testing on one of these would lead to similar results on the second, this assumption can be made. We then follow by creating a pair-plot between all the poverty columns and the median income column. We also see a high correlation (as backed up by the correlation heat-map) between the poverty columns i.e. All poverty, M poverty and F Poverty (Figure 3), thus this could lead to the problem of high multicollinearity and thus we need to take care of this, thus we remove the M poverty and F poverty columns. We then repeat the same process with the insurance columns and then using similar results proceed to dropping M with, M without, F with, F without columns. We then proceed with creating pair-plots (Figure 4) for the remaining All Poverty, Med Income, All with, All without columns and see that the columns All poverty and Insurance columns have high correlation (Figures 5, 6, 7, 8), but we will deal with it later. We then look at the scatter plots of the independent columns with the dependent columns.

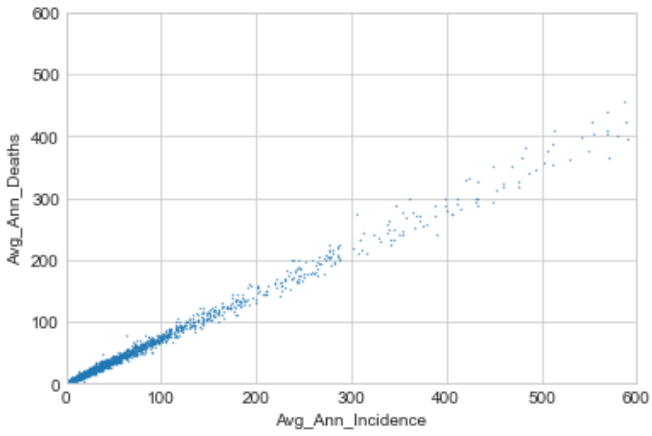

Fig. 2. Scatter-Plot

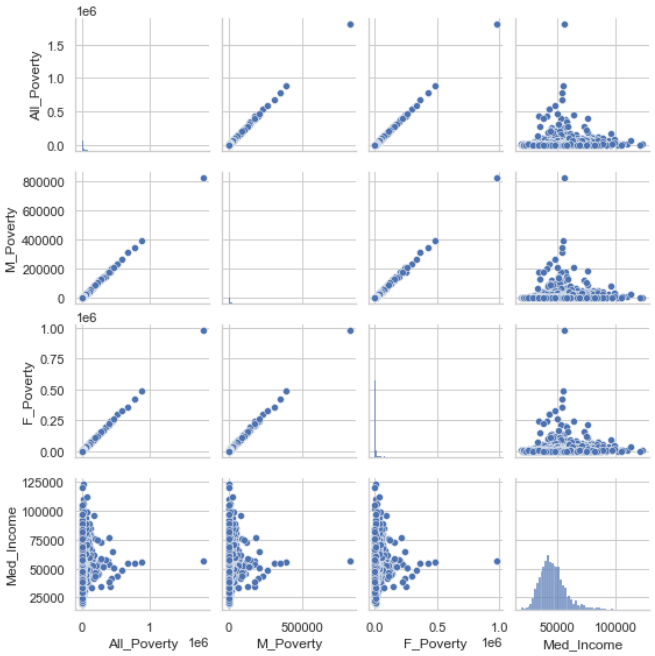

Fig. 3. Pair-plot

C. Statistical Linear Regression Modelling

We use the Statsmodels library in python to model the Linear regression as it gives us the added benefits of getting the significance values of the regression coefficients. We start with modelling the two dependent variables Avg Ann Deaths and Avg Ann Incidence separately into two models, with the independent variables being All poverty, Med income, All With, All without, Rising and Falling. We get a Adjusted R squared value of 0.922 for both the models and all the coefficient being significant except Rising with a P-value greater than 0.05 (for 95 % confidence). Please note that statsmodels library takes care of the Null values in the data-set on its own.

Now we proceed to check if our model satisfies the assump-

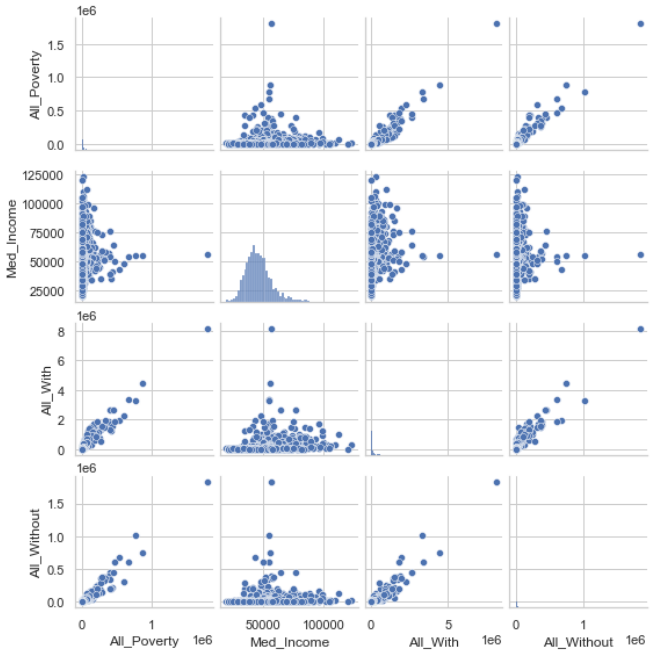

Fig. 4. Pair-plot

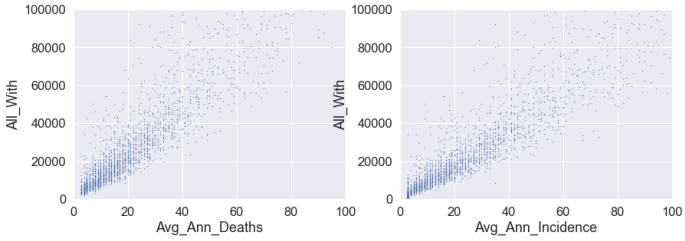

Fig. 5. Scatter-plot

tions for linear regression. We start with computing the VIF (Variance Inflation Factor) for all the independent variables to test for multicollinearity in the features, with the first model giving out All Poverty as the feature with the highest VIF of around 25.7 which might lead to coefficients that are not statistically significant. We then drop the All Poverty column and then iteratively follow the same process till all the VIF values are under the threshold of 5.

Now we proceed to check for the normality of the errors (residuals), on plotting a histogram of our data along with the normal and t distributions (Figure 9), plus creating QQ plots corresponding to the normal and t distribution, we can see that the QQ plot (Figure 10) for the normal distribution is not close to the straight line at the ends and it deviates to the top at the right and to the bottom at the left, we can say that the tails for the residuals are heavier than a normal distribution, hence on following that, we try QQ plot corresponding to the t distribution, which gives a much better QQ plot and hence we can confirm that the residuals come from the t distribution. Fatter tails suggest we have more number of outliers.

Finally, we check for the heteroscedasticity, and start with

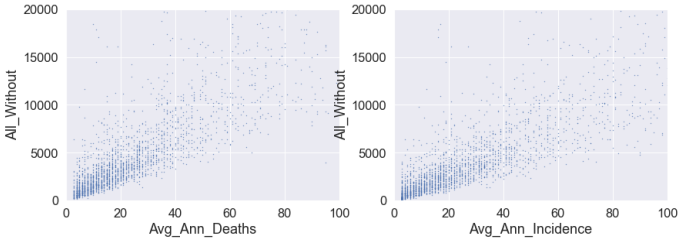

Fig. 6. Scatter-plot

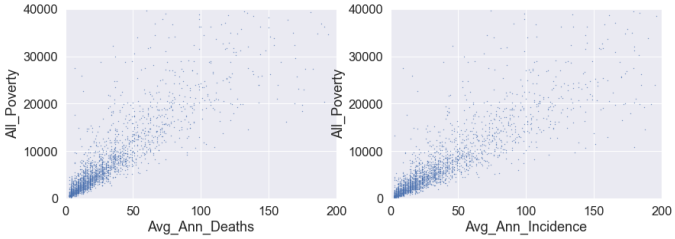

Fig. 7. Scatter-plot

visualizing a regression plot (Figure 11) and get consistent results to high Pearson R coefficient and thus high correlation. We then move on to creating a Lowess curve (Figure 12) and a scatter plot for our residuals (Figure 13) to check for the trend, wherein we see that the Lowess curve is slightly below the y=0 line for lower values which mean our model is overpredicting these values and then goes to the upper side for most of the residuals are about the y=0 line which means our model is underpredicting these values. Finally we proceed to visualizing the residuals and then observe to see any trend of changing variance with the values in the model. We can see a cone shaped residual plot which is common in cases of heteroscadasticity wherein when the fitted value increases, the variance also increases. Hence our model suffers from high heteroscadasticity.

IV. Conclusion

Having modelled a statistical model using different features as representatives for Socio-economic status and the Average incidence and deaths as the independent features, we can see that we get significant correlation between the features including Poverty, Median Income and Insurance which act as economic features, and also using the Median Income data for the different communities and observing different mean values for different states, we can safely say that given median income is a feature that is important in the determination of mortality or death, the difference in median income of the different social groups leads us to believe that the social factor is also significant. Thus we accept our hypothesis that socio-economic status is correlated to cancer incidence and mortality.

Checking for the assumptions of Linear Regression we can see that our model suffers from multicollinearity (was fixed), normality (residuals were more aligned to t distribution) and heteroscedasticity (a cone shaped residual plot), thus the future

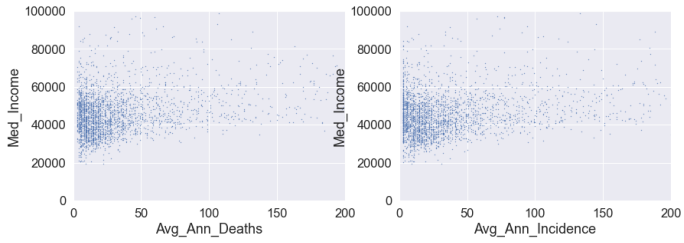

Fig. 8. Scatter-plot

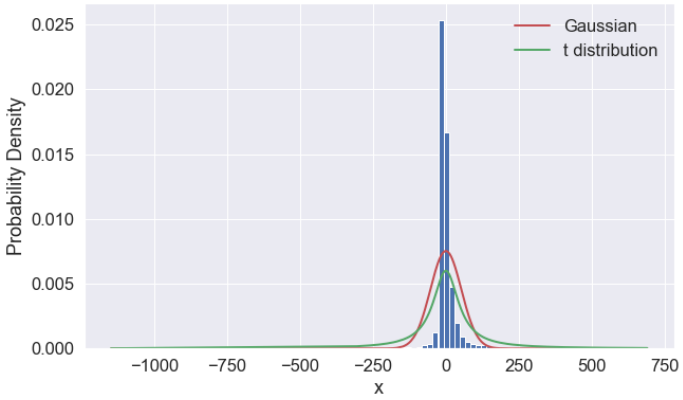

Fig. 9. Probability Density with common distributions

avenues of improvement can be checking for and treating outliers, and also investigating on the cause of this variable variance, some of the possible fixes can be using weighted regression model or using transformations on the dependent variable (one such transformation can be normalizing all the independent variables with population data and then using the Mortality rate and Death rate for modelling purposes which removes the effect of population from the data.)

REFERENCES

[1] Linear regression, Data Analytics Laboratory EE4708, July - November 2021

[2] Linear regression by Prof. Manikandan, Pattern Recognition and Machine Learning, July - November 2021

[3] https://towardsdatascience.com/7-ways-to-handle-missing-values-in-machine-learning-1a6326adf79e

[4] https://blog.minitab.com/en/adventures-in-statistics-2/how-to-interpret-regression-analysis-results-p-values-and-coefficients

[5] https://en.wikipedia.org/wiki/Student%27s_t-distribution

[6] https://www.geeksforgeeks.org/detecting-multicollinearity-with-vif-python/

[7] https://stats.stackexchange.com/questions/76441/interpretation-from-loess-graph

[8] https://stats.stackexchange.com/questions/481413/how-to-interpret-this-shape-of-qq-plot-of-standardized-residuals

[9] https://towardsdatascience.com/q-q-plots-explained-5aa8495426c0

[10] https://medium.com/evidentebm/linear-regression-using-statsmodels-d0db5fef16bb

[11] https://data.world/nrippner/cancer-linear-regression-model-tutorial/workspace/file?filename=OLS_regression_walkthrough.ipynb

[12] https://statisticsbyjim.com/regression/heteroscedasticity-regression/

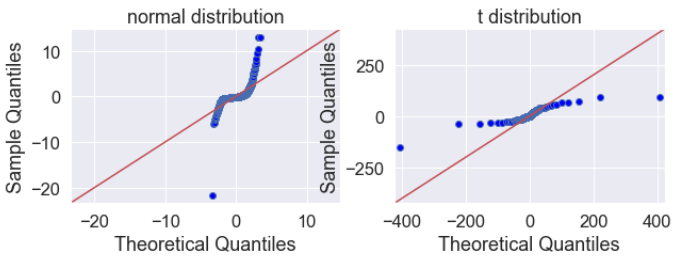

Fig. 10. QQ Plots

Fig. 11. Regression Plot

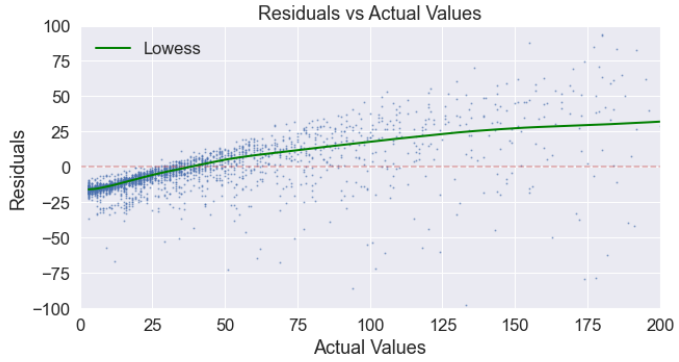

Fig. 12. Residual vs Actual Values

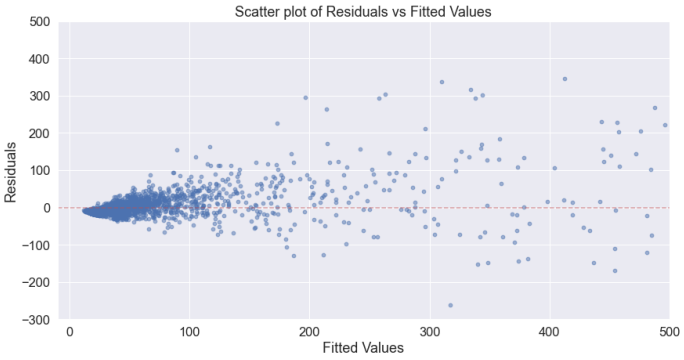

Fig. 13. Heteroscadasticity