A Mathematical Essay on Linear Regression

Prajwal Dinesh Sahu
ME18B114
Dept. of Mechanical Engineering
Indian Institute of Technology, Madras
me18b114@smail.iitm.ac.in

***Abstract*—This paper gives a brief introduction of multivariate linear regression and the basic mathematical theory behind it. We also look at some metrics to evaluate the performance of linear regression on the data-set. Further, we apply linear regression on a real world data-set containing the cancer incidence and mortality rates in various regions of USA, along with social and economic data from that region, to investigate whether the lower income groups are at a higher risk of cancer deaths. We employ exploratory data analysis methods and linear regression to find trends in our data. Our findings show a clear relationship between the socioeconomic status and the cancer mortality rates.**

I. Introduction

Linear regression is a supervised machine learning algorithm used to train a model to predict the behaviour of our data based on some variables. In linear regression, we find out the relationship between the inputs and the target variable by trying to find a best fit line.

In simple linear regression, we model our data in a linear equation –

$$\mathrm{y}_p = \alpha + \beta_1 \mathrm{X}_1 + \beta_2 \mathrm{X}_2 ..... + \beta_k \mathrm{X}_n + \epsilon \tag{1}$$

We try to minimize the least square error of the various data points from our fit line.

$$Minimize \sum_{i=1}^{n} (y_p - y_i)^2 \tag{2}$$

Based on the cancer incidence, death rate and socio-economic data for various US counties, we have to determine whether low-income groups are at greater risk for being diagnosed and dying from cancer, using linear regression.

The paper will first give a brief explanation of linear regression. Then, we will try to solve the problem and model our data-set by
1. Cleaning the data.
2. Dealing with Categorical variables.
3. Exploratory and visual analysis.
4. Choosing the correct variables.
5. Training the model.
6. Evaluating the model.
7. Interpreting the results.

II. Linear Regression Theory

A. Simple Linear Regression

Linear regression is the process of predicting a quantitative response Y on the basis of a single predictor variable X. It assumes that there is approximately a linear relationship between X and Y. We can write this relationship as –

$$\mathrm{y} = \beta_0 + \beta_1 \mathrm{X}_1 + \epsilon \tag{3}$$

The core idea is to obtain a line that best fits the data. The best fit line is the one for which total prediction error (all data points) are as small as possible.

B. Estimating the coefficients

In a linear regression problem, we have n paired data points of (x, y) containing collected data. We try to find the best fit of a line through these points in such a way that the distance of the line from the points in minimized. The most common approach involves minimizing the least squares criterion.

For error of point k_{th}

$$\epsilon_k = (y_k - y_i)^2 \tag{4}$$

On minimizing the sum of errors . we get the following result -

$$\beta_1 = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sum (X_i - \bar{X})^2} \tag{5}$$

$$\beta_0 = \bar{Y} - \beta_1 \bar{X} \tag{6}$$

C. Evaluating the model

In model evaluation, we measure the extent to which the model fits the data. The quality of a linear regression fit is typically assessed using two related quantities: the residual standard error (RSE) and the R22 statistic.

*1) **RMSE**:* To calculate the root mean squared error, we first calculate the residual sum of squares that is the sum of squares difference between the ith observed response value and the ith response value that is predicted by our linear model and then we divide it by the number of observations and then take it's square root.
It is one of the most commonly used metric for linear regression.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(y_i - y)^2}{n}} \tag{7}$$

2) $\boldsymbol{R^2}$: The R2 statistic provides an alternative measure of fit. It takes the form of a proportion—the proportion of variance explained—and so it always takes on a value between 0 and 1, and is independent of the scale of Y

$$R^2 = 1 - \frac{RSS}{TSS} \tag{8}$$

where

$$TSS = \sum_{i=1}^{n}(y_i - \bar{y})^2 \tag{9}$$

D. Multivariate Linear Regression

In practice, we often have more than one predictor. We extend the simple linear regression model so that it can directly accommodate multiple predictors. A multivariate regression is an extension of multiple regression with one dependent variable and multiple independent variables. Based on the number of independent variables, we try to predict the output.

$$\mathrm{y} = \beta_0 + \beta_1 \mathrm{X}_1 + \beta_2 \mathrm{X}_2 ..... + \beta_k \mathrm{X}_n + \epsilon \tag{10}$$

To estimate the coefficients of multivariate regression, we use the same least square error method to minimize the following equation -

$$\sum_{i=1}^{n}(\mathrm{y}_i - \beta_0 + \beta_1 \mathrm{X}_1 + \beta_2 \mathrm{X}_2 ..... + \beta_k \mathrm{X}_n)^2 \tag{11}$$

III. THE PROBLEM

A NGO is advocating for better health outcomes for low-income populations in the United States. To help with lobbying and fundraising, we need to examine and analyse whether low-income groups are at greater risk for being diagnosed and dying from cancer.

A. The Variables

The data given to us has these significant variables -
FIPS - Federal Information Processing Standard. Each area has a unique code.
Incidence rates - Cases per 100,000 population per year which are age-adjusted to the 2000 US standard population.
Mortality Rate - The measure of number of deaths.
All Poverty - Population For whom income in the past 12 months is below poverty level.
M Poverty - Male Population For whom income in the past 12 months is below poverty level.
F Poverty - Female Population For whom income in the past 12 months is below poverty level.
Med Income - Median household income in the past 12 months.
Med Income X - Median household income in the past 12 months for the X ethnicity (X - White, Black, Asian, Hispanic etc.).
All With - Population covered by health insurance.
All Without - Population not covered by health insurance.
M With, M Without - Male population covered and not covered by health insurance respectively.
F With, F Without - Female population covered and not covered by health insurance respectively.
Avg Ann Deaths - Average lung cancer mortalities.
Avg Ann Incidence - Average lung cancer incidence rate.

Population Data - We acquired a data-set from data.world for population estimate in counties in 2015, the year our dataset was collected We will need this data to normalize our variables.

Finally, we merge our socio-economic ,population and mortality rate data to get a merged dataframe with all suitable variables.

B. Cleaning the Data

We check for null values in all columns and find that the a significant percentage of Median Income data for various ethnicities is missing. So, we decide to drop these columns and only retain the overall median income data.

We check for non-numeric values in our data and notice that the following variables contain strings.
Incidence Rate - 499 strings including *, -
Mortality Rate and Avg. Ann Deaths – 325 strings including *,-

Since Mortality Rate is the target variable, we will drop all the string columns present in it. For Mortality Rate and Avg. Ann Deaths, we first remove the strings and impute the missing values with median values.

C. *Categorical Variables*

We have the recent trend variable which we have to convert to numerical values using dummy variables.

We have the following strings in recent trend so we can use 0 and 1 in place of them –
Falling – 197
Rising – 39

D. *Normalizing by population*

We create new columns with Population Normalization for Poverty and Health Insurance Data. We use the Population estimates we acquired.

$$Variable.PN = \frac{Variable}{PopulationEstimate} * 10^5 \qquad (12)$$

E. *Visualising and selecting data*

1) ***Income and Poverty*** *:* We have four variables relating to poverty data – All Poverty, M and F Poverty and Median Income.
We draw a pair plot and correlation heat map between these variables. (Fig.1 and 2)

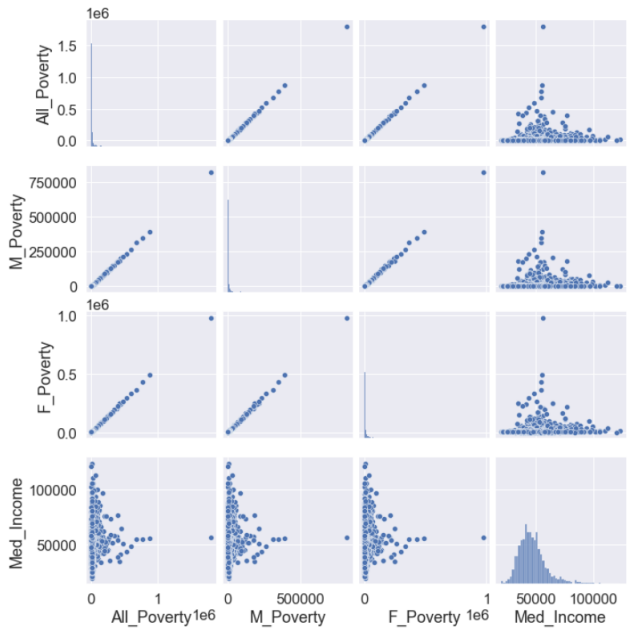

Fig. 1. Poverty Variables Pairplot

We can drop M Poverty and F Poverty as we see very high correlation between the gender related Poverty data. We retain the All Poverty column.

Through the scatterplot between Mortality Rate and All Poverty (Fig.3), we can see a clear positive relationship between them.

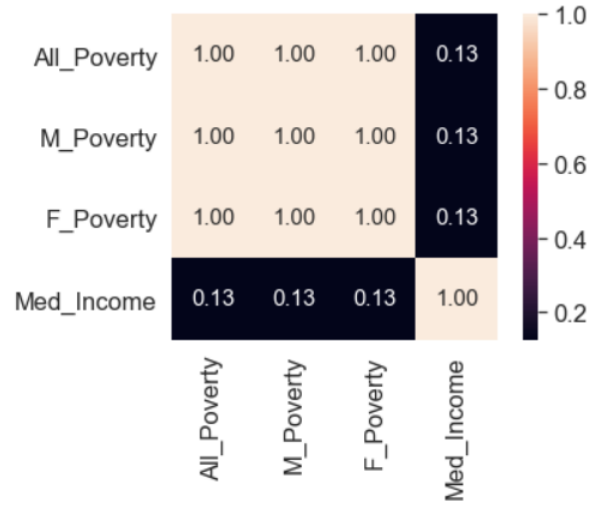

Fig. 2. Poverty Variables Heatmap

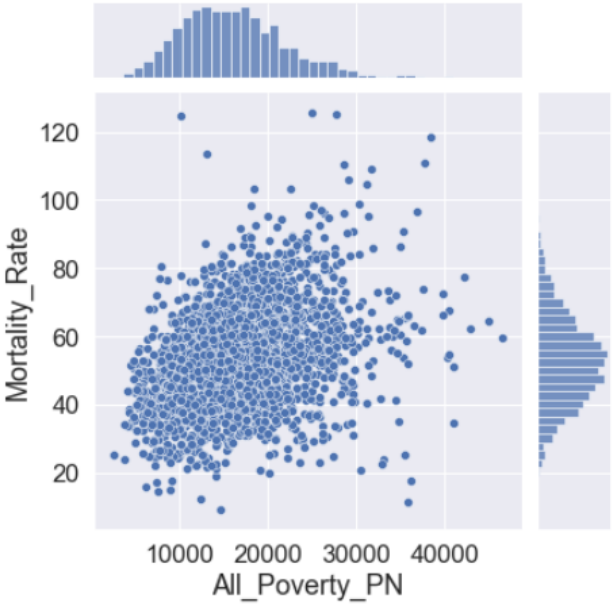

Fig. 3. Mortality Rate vs All Poverty

Through the scatterplot between Mortality Rate and Median Income (Fig. 4), we can see a clear negative relationship between them.

2) ***Health Insurance****:* We have 6 variables relating to poverty data – All With and Without, M and F With and Without.

We draw a pair plot and correlation heatmap between these variables. (Fig. 5 and 6)

Similar to poverty data, we can drop the gender related data

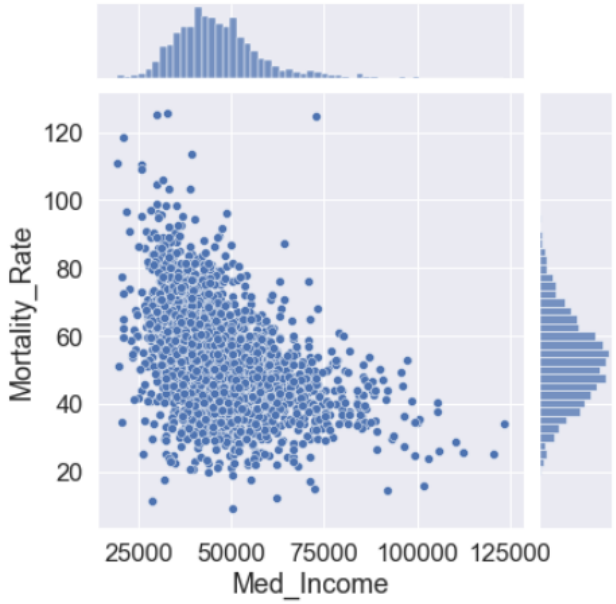

Fig. 4. Mortality vs Med Income

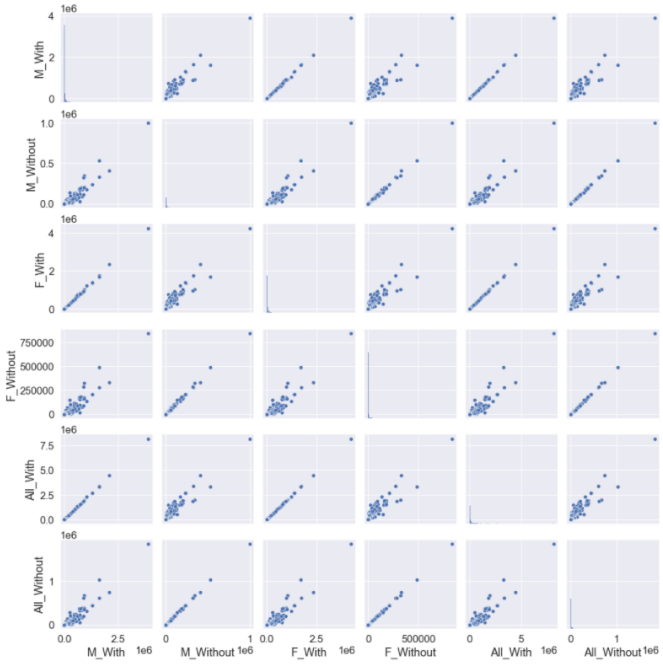

Fig. 5. Health Insurance Pairplot

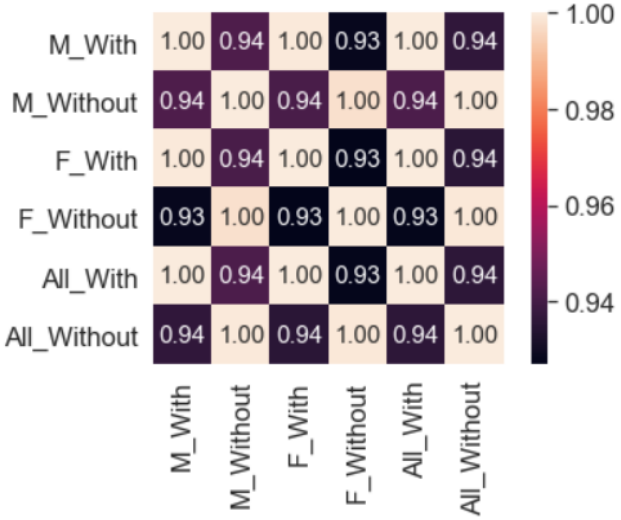

Fig. 6. Health Insurance Variables Heatmap

as we see very high correlation between them. We retain the All With and All Without column.

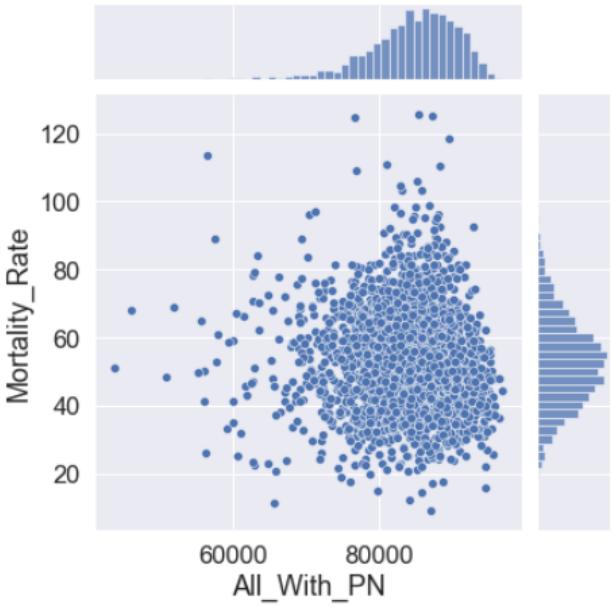

Fig. 7. Mortality vs All With Health Insurance

Through the scatterplot between Mortality Rate and Health Insurance Data (Fig. 7 and 8), we witness some pattern between mortality and health insurance data and therefore it is advisable to include them in our model.

3) ***Death and Incidence Rate****:* These variables are closely interrelated so we should keep only one of these variables. Our hypothesis is also confirmed by the correlation heat map and pair plot. (Fig. 9 and 10). We keep only incidence rate for our model.

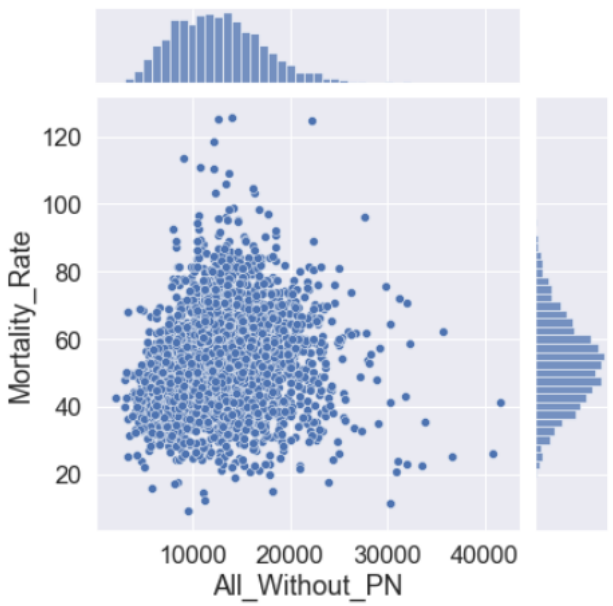

Fig. 8. Mortality vs All Without Health Insurance

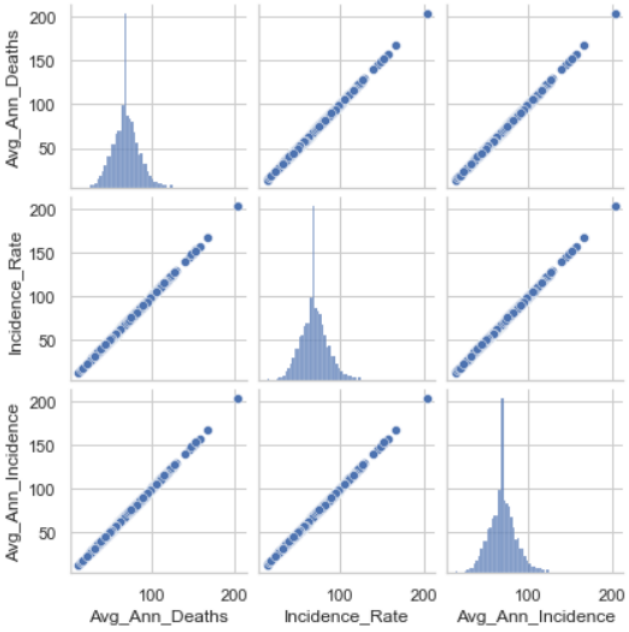

Fig. 9. Incidence Variables Pairplot

This also indicates high correlation between Incidence Rate and Mortality Rate. We can use either of these variables in our regression model.

There is a definite positive correlation with mortality rate, as visible in scatterplot. (Fig. 11)

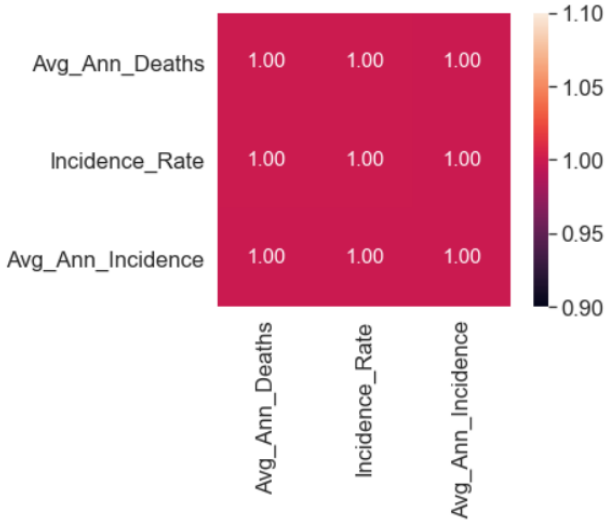

Fig. 10. Incidence Variables Heatmap

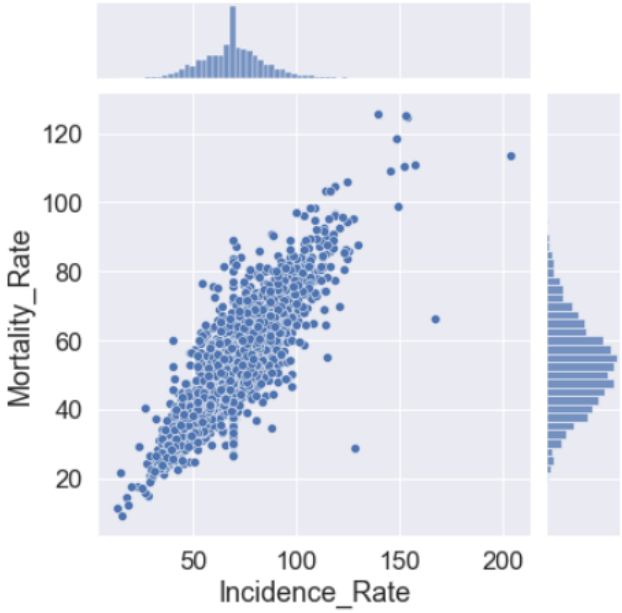

Fig. 11. Mortality Rate vs Incidence Rate

F. Visualizing incidence rate with poverty and income variables

Incidence rate represents the number of cancer cases reported in a particular area. We can also visualize incidence rate's relation to poverty and income data and check for any trends.

Through the below given plot between Incidence Rate and Median Income (Fig.12), we can see a slight negative trend, which confirms our suspicion that lower income groups are more vulnerable to cancer.

Similarly, through the plot between Incidence Rate and Poverty stats (Fig. 13), we can see a positive trend which

shows that in areas with higher number of people below the poverty line, we see higher incidence of cancer cases

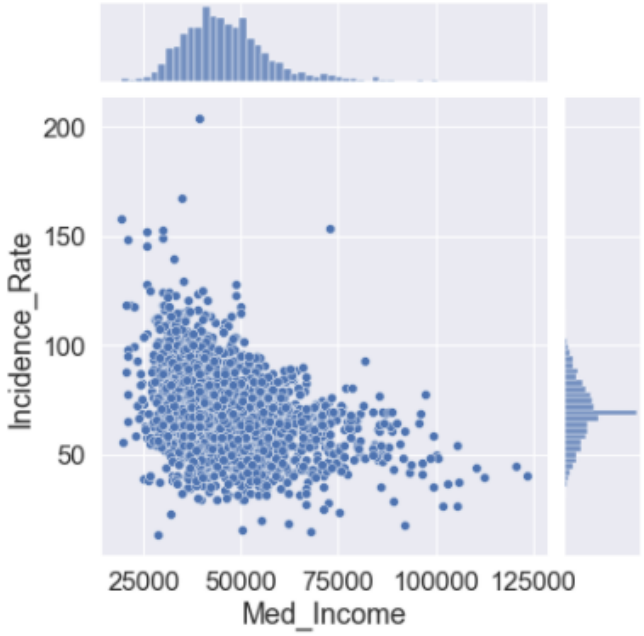

Fig. 12. Incidence Rate vs Median Income

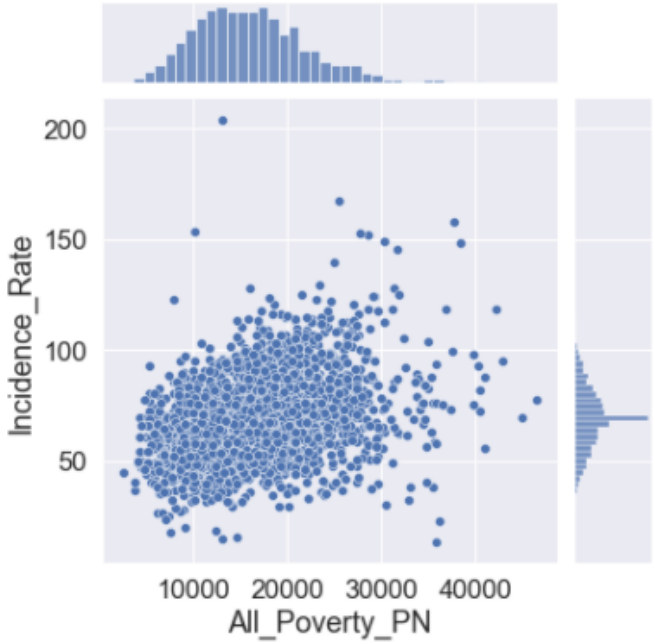

Fig. 13. Incidence Rate vs Poverty

G. *Final Variables for our model*

We have these final variables with us - 'All Poverty PN', 'Med Income', 'All With PN', 'All Without PN', 'Incidence Rate','rising' and 'falling'.

We will draw final correlation heat map to ensure no correlated variables.

We find no correlated variables and therefore, we are ready to go ahead with the model. (Fig. 12)

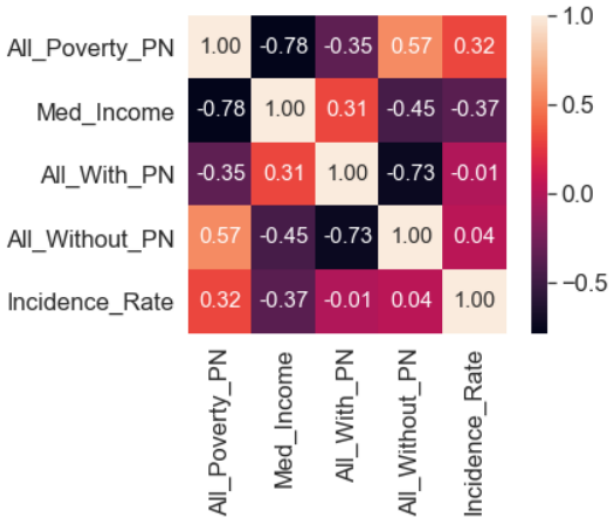

Fig. 14. Final Variables Pairplot

IV. The Model

A. *Fitting our Data*

We have the following X and y for our final model.

X= All Poverty PN , Med Income, All With PN, All Without PN, rising , falling.
y = Mortality Rate.

Using scikit-learn library's LinearRegression() function on our X and y variables, we develop a multivariate linear regression model -

$$\begin{aligned}\text{Mortality.Rate}_i = \beta_0 &+ \beta_1 \text{All.Poverty.PN}_i \\ &+ \beta_2 \text{Med.Income}_i \\ &+ \beta_3 \text{All.With.PN}_i \\ &+ \beta_4 \text{All.Without.PN}_i \\ &+ \beta_5 \text{rising}_i \\ &+ \beta_6 \text{falling}_i + \epsilon\end{aligned}$$

B. *Intercepts and Coefficients*

After fitting our data, we get the following values - Intercept -

$$\beta_0 = 6.144721764929308 \tag{13}$$

The variable-wise coefficients are tabulated below

	Coefficient
All_Poverty_PN	0.000057
Med_Income	-0.000116
All_With_PN	0.000028
All_Without_PN	0.000240
Incidence_Rate	0.656103
rising	-0.986183
falling	0.682052

Fig. 15. The Coefficients of Model

C. *Model Evaluation and Metrics*

The Root Mean Squared Error (RMSE) of the model is 0.736365906887327

The R2 value for the model comes out to be 7.208664437398365

D. *Interpretations of the coefficients in the model*

*1) **Poverty and Income Variables**: -*

Med Income = -0.000116

We have a strong negative relationship. This can be interpreted as on 1000 dollar decrease in median income of the region, mortality rate goes up by 1.16 . Therefore we can report the lower income groups are having a higher mortality rate and in danger.

All Poverty PN = 0.000057

We have a positive coefficient which can be interpreted as if the number of people below the poverty line in an area increases, the risk of cancer death is higher.

*2) **Health Insurance Data**: -*

All Without = 0.00024

We have a positive coefficient which means that if the area has higher number of people without health insurance, the risk of cancer death is higher. Having an insurance can affect the treatment a person gets and therefore we witness a higher coefficient for All Without (0.000240) than All With (0.000028). This adds weight to the fact that lower income groups who can't afford health care are at a greater risk.

*3) **Incidence Rate**: -*

Incidence Rate = 0.656103

As we saw before, the incidence and mortality rate are highly correlated as logically, more instances of cancer cases means more deaths.

V. Conclusions

According to our findings, the incidence rate and mortality rate are linked to the socioeconomic background. As we expected, the incidence rate and mortality rate are highly correlated as more cases mean more death instances. We have both visual and mathematical evidence to support our claim.

Through mathematical evidence based on our linear regression model with mortality rate as our target variable, we can report-

- A negative coefficient for median income, indicating that lower income groups have higher mortality risk.
- A positive coefficient for poverty statistic, indicating that areas with higher number of people below poverty line have higher mortality rate.
- A higher positive coefficient for people without health insurance than those with health insurance, indicating that people who cannot afford insurance are at higher risk.

Through visual evidence ,the data indicates a negative exponential link between death /incidence rate and median income. This can be explained by the fact that initially as the income decreases, the death rate goes up, but after a certain threshold the curve flattens down as lowering the income further won't affect the treatment and lifestyle.

The positive link between death/ incidence rate and people below poverty line also has a similar explanation.

Therefore we can conclude that lower income groups have higher incidence and mortality rates and they are definitely at an higher risk of cancer.

References

[1] James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013) An Introduction to Statistical Learning with applications in R, www.StatLearning.com, Springer-Verlag, New York

[2] Swaminathan, S. (2019, January 18). Linear regression - detailed view. Medium. Retrieved September 16, 2021, from https://towardsdatascience.com/linear-regression-detailed-view-ea73175f6e86.

[3] Bureau, U. S. C. (2020, February 13). Acs poverty data tables. The United States Census Bureau. Retrieved September 16, 2021, from https://www.census.gov/topics/income-poverty/poverty/data/tables/acs.html.

[4] ACS2015 5 E Income - dataset BY USCENSUSBUREAU. data.world. (2018, April 2). Retrieved September 16, 2021, from https://data.world/uscensusbureau/acs-2015-5-e-income.

[5] US population Estimates 2015 - dataset BY NRIPPNER. data.world. (2017, January 27). Retrieved September 17, 2021, from https://data.world/nrippner/us-population-estimates-2015/.