A Mathematical Essay on Linear Regression

Vignesh Kumar S
Department of Chemical Engineering
Indian Institute of Technology, Madras
Email address: ch18b118@smail.iitm.ac.in

***Abstract*—The nexus between incomes and health outcomes is a necessary measure that needs to be established to identify population groups to help with better prognosis. In this work, we examine whether low-income groups are at a greater risk of being diagnosed and dying from cancer. Due to complex nature of the problem and abstractness involved, it is difficult to obtain a good solution using traditional and manual approaches. Adopting a Machine Learning Technique can help in getting insights, analyse different features and large amount of data. The primary goal of this paper is to understand Linear Regression and the Mathematics behind it, use it for modelling and its applicability in real-world scenarios.**

I. Introduction

In this paper, we will study Linear Regression, a technique used for modelling a response variable with one or more explanatory variables in a linear fashion. Then it is used to analyse whether the incidence and cancer mortality occurrences are affected by socio-economic status in the United States. This analysis would aid in obtaining insights among a large population of varying segments and could be used to make decisions involving fundraising and legislation towards this particular cause.

Regression is a technique that attempts to establish a relationship between a target variable and one or more explanatory variables. Linear Regression is a regression method that models the relationship linearly. The best fit line obtained by minimizing the sum squared distance between the model predictions and the outputs is essentially the idea behind this technique. This could also be used to establish if there is any correlation between variables. However, this does not imply the output is the cause of input features but is just a technique to see if there is some relationship between them. Linear Regression is also widely used in time series analysis such as stock market predictions, weather conditions, sales in a particular store.

The dataset used for this problem comprises cancer incidences and mortality across various areas in the United States and the median income split among ethnicities in each area. The data also contains the number of people with health insurance and the trend of cancer incidence in the past 5 years. We use Linear Regression to demonstrate whether cancer incidence and mortality are strongly correlated with socio-economic status. The insights obtained from this could be used by firms and the government to make decisions and solve issues.

In this paper, we will study the concepts behind Linear Regression and its various methods. We then use this technique to establish the presence of a relationship between cancer incidence and socio-economic factors such as income, poverty, and the percentage of people with access to insurance.

II. Linear Regression

Regression is a statistical technique where both the dependent and independent variable takes continuous values and a model is fitted on the explanatory variables. If the model that is learnt is linear, this is called a linear regression. Methods of estimation of the parameters could be done in several ways including Least Squares and statistical techniques such as Maximum Likelihood Estimation, Bayesian Estimation. A few extensions of it also include Ridge Regression, Lasso, Elastic Net and Online based learning such as Recursive Least Squares method. The choice of the model is based on the A priori knowledge of the process, observation and visualization of the data and the fit of the model.

Least Squares Estimators

Let the number of datapoints be N and k be a integer such that $1 \leq k \leq N$ and be used to represent a particular instant in the datapoint. Ordinary least squares problem deals with finding the best prediction of $\vec{y}$ using p explanatory variables (or) regressors $\psi_i[k]$, i = 1,2,...p such that ŷ[k] are collectively at a minimum distance. When the predictor are linear in parameters, we have linear least squares problem.

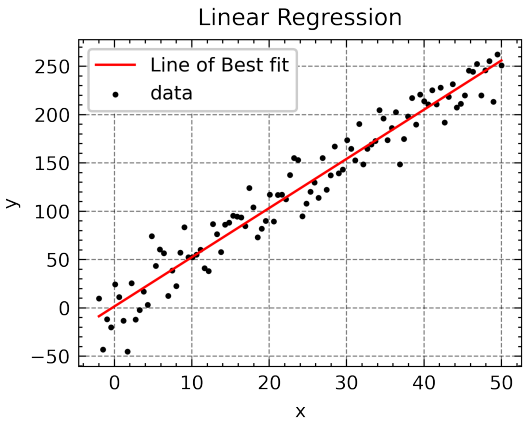


Fig. 1: Example of Linear Regression

The General form of a linear regression model is

$$\hat{y}[k] = \sum_{i=1}^{p} \psi_i[k]\theta_i \tag{1}$$

$$\varepsilon = \vec{y} - \hat{y} \tag{2}$$

where ε is the residual vector.

A. Ordinary Least Squares Method

Let's represent the predictions

$$\hat{y}[k] = \sum_{i=1}^{p} \psi_i[k]\theta_i \text{ in a matrix notation}$$

$$\psi[k] = \begin{bmatrix}\psi_1[k] & \psi_2[k] & .... & \psi_p[k]\end{bmatrix}^T$$

$$\text{Let } \Phi = \begin{bmatrix}\vec{\psi}[1]^T \\ \vec{\psi}[2]^T \\ ... \\ \vec{\psi}[N]^T\end{bmatrix}$$

$$\underset{n\times 1}{\mathrm{Y}} = \underset{n\times p}{\Phi} \times \underset{p\times 1}{\theta} + \underset{n\times 1}{\varepsilon}$$

Cost function (or equivalent to negative reward function) is a measure of penalty between the predicted outputs and true outputs. While several choices could be worked on, Ordinary least squares tries to minimize the sum squared error.

Obj Function:

$$\min_{\theta} \quad J_n(\theta) = \sum_{i=1}^{N} (y[k] - \hat{y}[k])^2$$

In matrix form, this could alternatively be expressed as

$$\min_{\theta} \quad J_n(\theta) = \|\mathbf{y} - \hat{\mathbf{y}}\|_2^2$$

where $\hat{y} = \Phi\theta$

On taking the gradient and solving for the optimum minimum point yields the solution,

$$\hat{\theta}^* = (\Phi^T\Phi)^{-1}\Phi^T y$$

Note that a constant term could be easily included in the model by incorporating

$$y[k] = \Psi^T[k]\vec{\theta} + \beta \quad as$$

$$y[k] = \begin{bmatrix}\Psi[k] & 1\end{bmatrix}\begin{bmatrix}\vec{\theta} \\ \beta\end{bmatrix}$$

B. Gradient Descent

Optimization of minimizing the cost function can be done in an alternate way, Gradient Descent. Gradient is the direction along which a function increases the maximum. Going along the opposite direction of Gradient gives the maximum decrease. On each iteration, the cost is reduced by moving along this direction till a certain step size, which is influenced by the hyper parameter learning rate (α). This is continued till a minima is reached or a stopping criterion is met.

1) Parameter Initialization: Parameters can be assigned random values to begin with. In linear regression, often, the parameters are initialized to zero.

2) Stopping Criteria: The convergence to local minima is checked by evaluating if a stopping criterion is met. Some of the choices for stopping critera include:

- Change in cost function is less than a threshold
- Change in gradient is less than a threshold
- Change in parameters is less than a threshold
- Number of iterations has reached its threshold

ALGORITHM: Initialize θ
Repeat till convergence
{
for $j \in \{1, 2, .., p\}$
$\theta_j := \theta_j - \alpha \frac{\partial J(\theta_1, \theta_2, ..., \theta_p)}{\partial \theta_j}$
}
where α is the learning rate.

3) Effect of Learning Rate: Small Values of α may take too many iterations to reach minima. In the algorithm above, small values of α barely updates the model parameters while a large α produce significant changes. This means that very small values will take too many iterations and if α is large enough, it may oscillate and even diverge from the optimum point.

This itself could be set up as an additional optimization problem to obtain the best α at each iteration to converge much faster. However, in practice, α is held constant and is set at appropriate levels which ensures convergence at a reasonable rate.

Though gradient descent for a random function with multiple local minimas may end up in different convergence, but for a Least squares objective function J, we always reach the global minima provided the learning rate is set at appropriate levels.

Gradient descent could be of multiple types. Ordinary, Stochastic, Mini-Batch, Batch e.t.c. These are not discussed as it is not under the scope of this paper and could be taken up as an extensive study under optimization course.

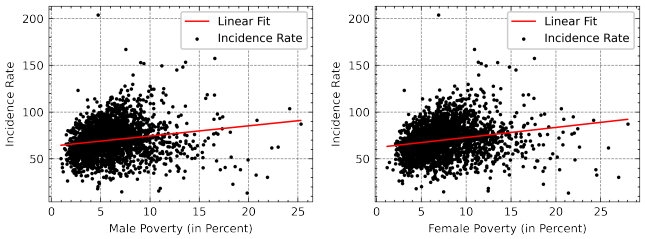
III. The Problem

In this section, We will do a case study on whether socio-economic factors such as income, poverty, subscription to health insurance is related to incidence and mortality of cancer. Linear Regression technique is used to understand our plots better and strengthen our claim.

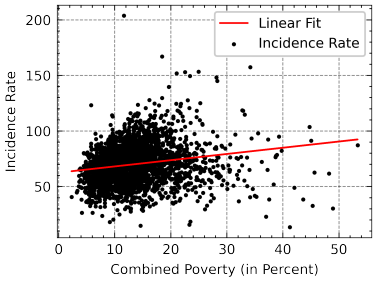
A. Poverty

We attempt to gain insights by visualizing each socio-economic factors versus incidence and mortality rates separately. With the help of linear regression, we try to see if there is a trend associated with the factors.

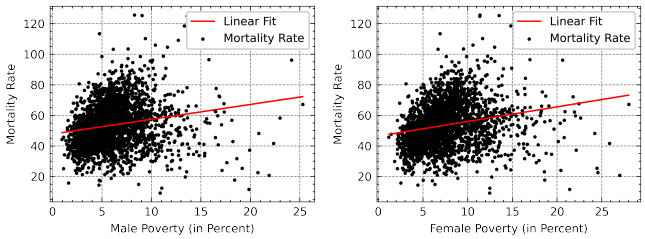
First, let us visualize whether if there is any correlation between Poverty versus Incidence and Mortality Rates. Access



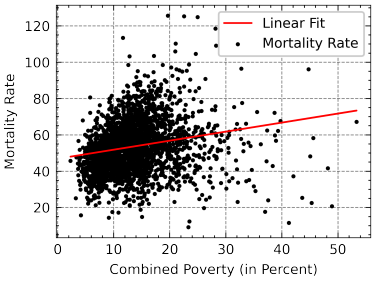
(a) (b)



(c)



(a) (b)



(c)

to good healthcare and proper nutrition is difficult for people living under poverty. In a country like US, where health care is not free unlike UK, by intuition, this seems to have a direct impact on the disease prognosis. The absolute count of people living in poverty may not be a very useful measure. Plots are constructed between Poverty Percentage versus Target Variables where poverty percentage is defined as count of people living under poverty in an area by estimated population (in percent)[1]of that area.

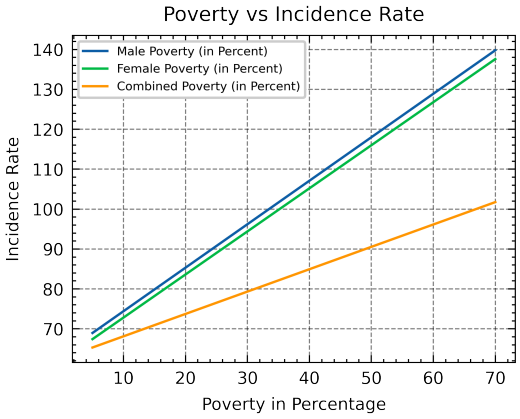


Fig. 2: Influence of Poverty versus Cancer Incidence

Fig.2 strengthens the evidence that as percentage of people living in poverty increases, the incidence rate goes up. Male Poverty and Female Poverty show similar trends implying that gender does not impact much on the incidence rate.

[1]Estimated population = $10^6 * \frac{AnnualizedIncidenceRate}{IncidenceRate}$

Annualized Incidence is the absolute count of people diagnosed
Incidence Rate is the count of people diagnosed per million population in an area.

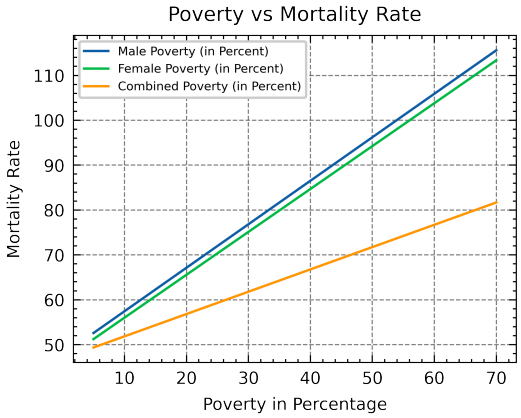


Fig. 3: Influence of Poverty versus Cancer Mortality

Similar trends are oberserved across Poverty versus Mortality Rate. This suggests that people living under poverty are more vulnerable to cancer and related diseases compared to well off people.

B. Income

In this section, we will analyse how incomes impact cancer incidences and Mortality Rates. We will visualize total incomes and incomes based on ethnicities and try to identify if there is any correlation between cancer rates.

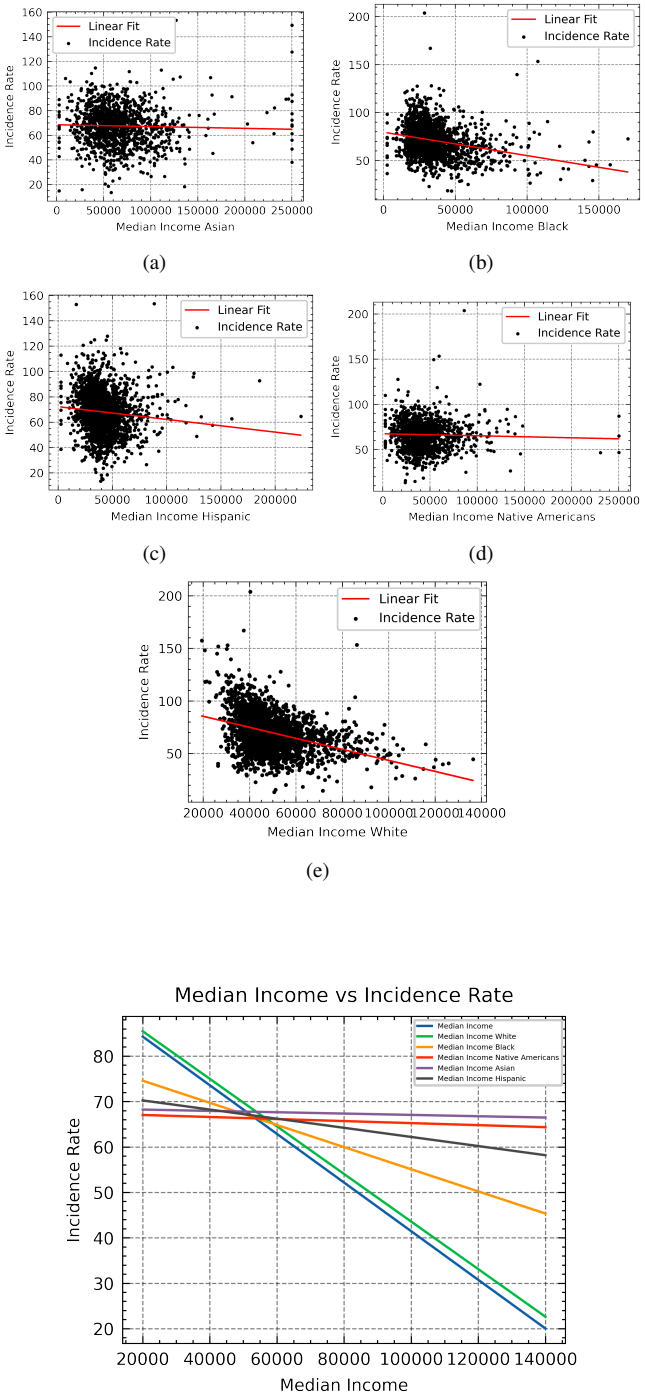


Fig. 4: Influence of Income versus Cancer Incidence

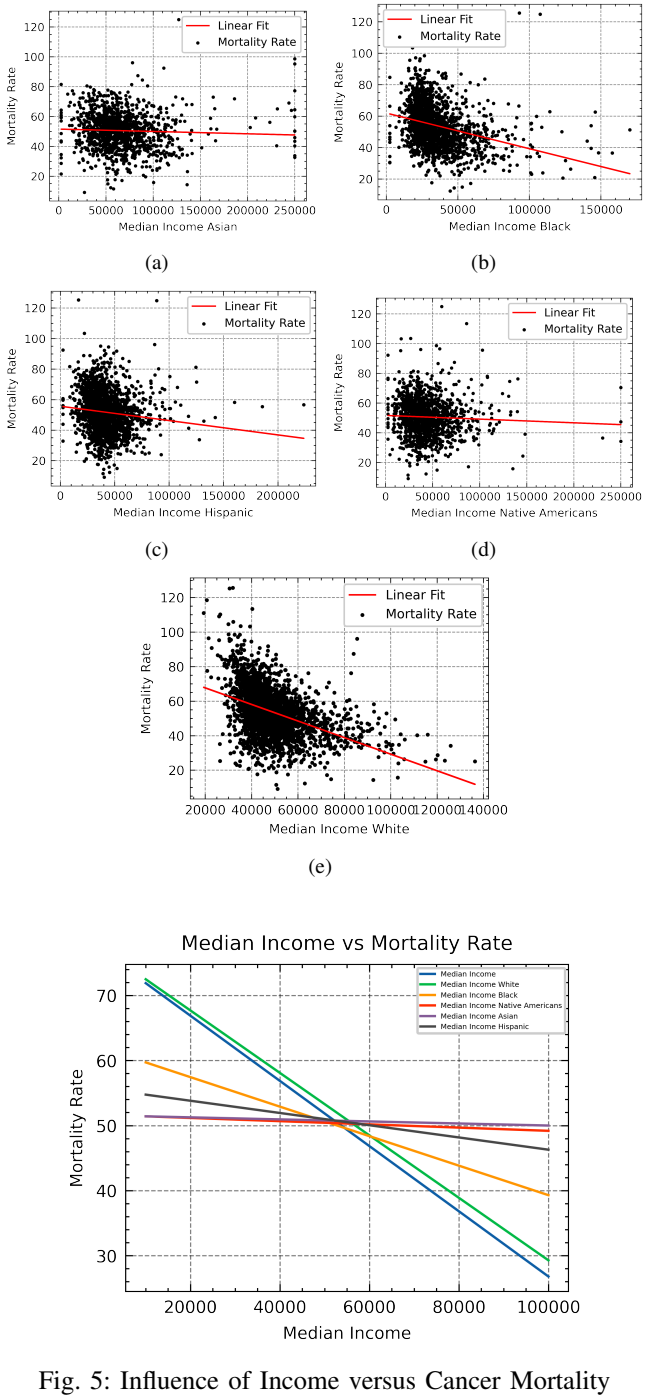


Fig. 5: Influence of Income versus Cancer Mortality

From Fig.4, Income does play a role. People with lower incomes have difficulty in managing health care expenses and obtaining nutritious food. This could also be attributed due to income being closely related with poverty. Hence such trend could have been observed.

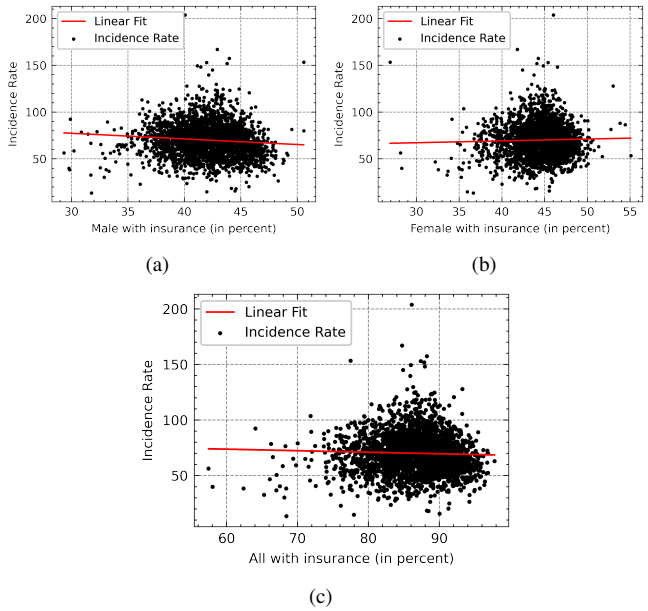
Fig 5. shows mortality rate of lower income people are high and decreases steeply as income increases. This substantiates the fact that people with lower income are more susceptible to cancer.

C. Health Insurance

In this section, we will analyse whether population subscribed to health insurance have an advantage and are less prone to cancer incidence and mortality. We will look across the effects across gender and population as a whole and attempt to interpret the results.

The absolute count of people with Health Insurance may not a useful measure. So plots are constructed for Percentage of people with health insurance versus the Target Variables, Incidences and Mortality Rates. Percentages are number of

people with Health Insurance per estimated population of an area[2]



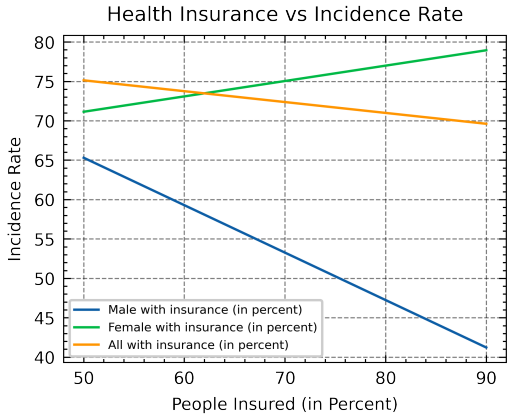
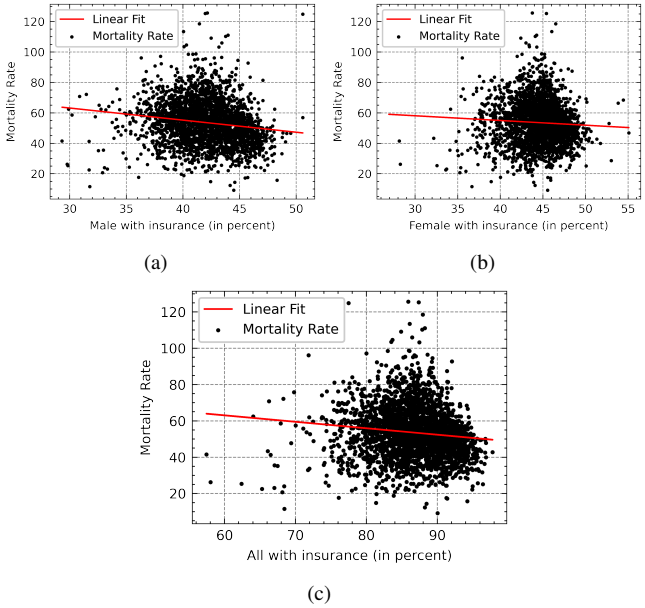


Fig. 6: Influence of Health Insurance versus Cancer Incidence

From Fig.6, The incidences corresponding to areas with All percentage of people with high insurances are almost the same or a little lower than those with low insurance percentages. This makes sense because Insurances are usually availed after a medical treatment and doesn't influence much on the probability of getting cancers. The slight decrease in nature could be speculated due to relatively more number of people subscribing to health insurance who are above poverty line. But, this cannot be justified as people below poverty line also tend to resort to health insurance in case of emergencies. In fact, Incidences of Females with insurance slightly increase (or remain almost constant) with increase in insurance percentage. Hence, it remains inconclusive whether Health Insurance are really effective in reducing the incidence rates.



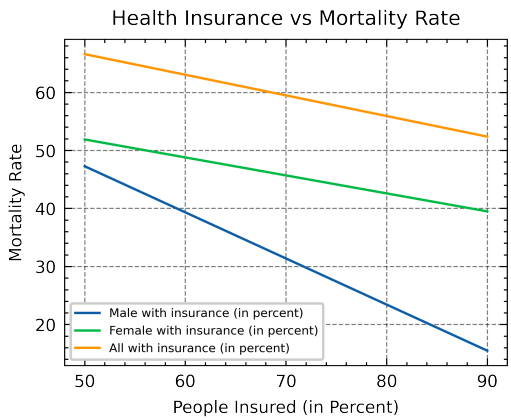


Fig. 7: Influence of Health Insurance versus Cancer Mortality

From Fig.7, Mortality Rate shows down trending values as percentage of people insured increases. This makes sense that people would access health care without the intimidation of financial burden and the chances of survival increases. The findings varies from our previous conclusion on incidence because the probability of getting diagnosed with cancer may not be influenced by Health Insurance but the survival would be improved due to medical facilities and less financial burden.

D. Statistical Evidence

In statistics, the p-value is the probability of obtaining results at least as extreme as the observed results of a statistical hypothesis test, assuming that the null hypothesis is correct. A smaller p-value means that there is stronger evidence in favor of the alternative hypothesis. In a linear regression, the

[2]Estimated Population ≈ People with Insurance + People without Insurance.
The Previous estimate of population[1] could also be used here. In fact, both the estimates are very similar and small differences might be due to census errors

alternate hypothesis being existence of relationship between response and dependent variables, lower the p values, the more significant is that feature in the model.

F-statistic is an indicator of whether there is a relationship between dependent variable and the response variables. Because individual t-tests assume that each variable comes from an independent distribution which is not always the case. This leads to multi-collinearity issues and also accumulation of Type-1 error to falsely conclude the results. Hence, by judging by the F-stat scores, we could conclude if there is a significant relationship between dependent variables and the response variable. The larger the F statistic is away from 1 the better it is.

R-Squared is a statistical measure of fit that indicates how much variation of a dependent variable is explained by the independent variable(s) in a regression model. The closer the R-squared is to 1, the more likely the relationship is statistically linear.

TABLE I: All Poverty Percentage

Metric	Incidence Rate	Mortality Rate
F-stat	99.89	124.6
P-val const	0.001	0.000
P-val feature	0.001	0.000
R^2	0.036	0.045

TABLE II: Median Income

Metric	Incidence Rate	Mortality Rate
F-stat	444.8	652.2
P-val const	0.000	0.000
P-val feature	0.000	0.000
R^2	0.144	0.198

All models except Insurance Percentage vs Incidence Rate have high F-statistic value implying the presence of some relationship between dependent and predictor variables. Though the R-squared value is less which suggests that the actual relationship could be non-linear, one can nevertheless see that pvalues of features are less than 0.05 strengthening the conclusions made in the above section.

TABLE III: Insurance Percentage

Metric	Incidence Rate	Mortality Rate
F-stat	3.903	41.26
P-val const	0.001	0.000
P-val feature	0.058	0.000
R^2	0.001	0.015

For the model between Insurance Percentage vs Incidence Rate, the pvalue lies slightly above 0.05 indicating that the linear regression model may not be statistically significant. In fact, both the F-statistic and the R-squared metrics are among the lowest indicating that the linear model would be inconclusive. This suggests our claim that Health Insurance doesn't have statistical influence on Incidence Rate as Insurances are usually availed after a medical treatment and doesn't influence much on the probability of getting cancers.

IV. Conclusions

Socio-Economic factors such as incomes, poverty and access to health insurance does influence incidence and Mortality Rates. High incomes and Low poverty helps in preventing the disease and also gives more chances of survival due to better access to health care systems. Though Health Insurance is not very effective in preventing cancer incidence, it reduces the Mortality rate and increases the chance of survival.

Targeted programs towards population with low income and regions of high poverty must be planned out to help prevent cancer mortalities. Incentives on Health Insurance plans and decreased hospital charges towards these section of people could be planned to reduce the incidence and the effects of cancer.

Possible avenues of research could include analysing percentage of people having access to good health care system, the reason for not pursuing health insurance which could be, but not limited to, lack of wide range of plans catering to their capability.

References

[1] Aurelian Geron, Hands-On Machine Learning with Scikit-Learn,Keras, and TensorFlow: Concepts, Tools and Techniques to Build Intelligent Systems, pp. 38–111

[2] Arun K. Tangirala, Principles of System Identification, Theory and Practice