A Mathematical Essay on Linear Regression

Swathi V
Dept. of Engineering Design
Indian Institute of Technology Madras
Chennai, India
ed18b034@smail.iitm.ac.in

***Abstract*—Incidence and mortality rates of cancer can be influenced by various factors, some of them understood, and others we are still exploring. In this article, we analyse the socio-economic factors that might influence cancer incidence and related mortality in the US. We start by providing a mathematical overview of linear regression, which we then use to understand trends in the cancer rates.**

***Index Terms*—Simple linear regression, multiple linear regression, polynomial linear regression, R^2 score, cancer mortality rate, cancer incidence rate**

I. Introduction

Linear regression models are one of the simplest statistical tools used to predict a target variable using measurable input variables, to quantify the relationship between them. The input variables are considered to be explanatory variables, while the target variable is considered to be a dependent variable.

A linear regression model fits a straight line through the training data points. The most common way to do this is to use the least squares method, or minimize the sum of the squares of the vertical deviations from each data point to the line.

In this article, we perform linear regression on the cancer survey data of the United States in 2015 that measured social and economic factors like income, sex-ratios, ethinicities, etc. This model can be used to understand what socio-economic factors may affect the incidence and mortality rates, and to what extent.

The article starts with a brief overview of linear regression, followed by a description of the problem being solved. The approach used for solving the problem is then highlighted, and conclusions that are drawn from the results are listed out.

II. Linear Regression

Linear regression is a statistical tool used for predicting a quantitative response (value of a dependent variable, Y) from given vector of input variables, $X = (X_1, X_2, \cdots X_p)$. The steps to building a linear regression model are outlined below.

A. Hypothesizing a Model

A **model hypothesis** can be built based on prior knowledge of the data, or by visualizing the data to understand the nature of relationships between the input and response variables.

1) Simple Linear Regression: Simple linear regression predicts the response Y using a single input variable X, assuming that there is approximately a linear relationship between the two. This relationship can be written as

$$Y \approx \beta_0 + \beta_1 X \tag{1}$$

where β_0 and β_1 are constants that represent the intercept and slope of the linear model.

2) Multiple Linear Regression: Multiple linear regression uses a vector of input variables $X = (X_1, X_2, ...X_p)$ to predict the response variable Y. This relationship can be modelled as

$$Y \approx \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots \beta_p X_p \tag{2}$$

where β_j quantifies the relationship between the input variable X_j and response Y.

3) Polynomial Regression: Polynomial regression uses a single or multiple input variables to predict the response , under the assumption that Y can be modelled as a polynomial function of the various input variables $X_1, X_2, \cdots X_p$. For a single input variable X, this relationship is given as

$$Y \approx \beta_0 + \beta_1 X + \beta_2 X^2 \cdots \beta_m X^m \tag{3}$$

where β_j quantifies the relationship between the j^{th} power of X and the response Y.

All linear regression models can thus be represented as

$$Y \approx \beta_0 + \sum_{i=1}^{p} \beta_i X_i \tag{4}$$

where X_j can be either a direct or a mapped (example, polynomial, or log-transformed) input variable, and β_j quantifies the relationship between X_j and the response variable Y.

B. Estimating the Coefficients

After the model has been hypothesized, the optimal coefficients can be predicted to minimize a **cost function**, typically the mean-squared error (MSE). The MSE is given as

$$MSE = \frac{1}{N} \sum_{i=1}^{N} (y_i - \hat{y}_i)^2 \tag{5}$$

where N is the total number of samples used to train the model, y_i is the response variable of each sample, and $\hat{y}_i$ is the response variable predicted by the model using the input variables at each sample.

The values $\beta = (\beta_0, \beta_1, \cdots \beta_p)$ that minimize the cost function (5) are chosen to be the optimal weights. This optimization problem can be solved either analytically, or numerically using techniques such as gradient descent.

1) Regularized Cost Functions: To prevent the model from overfitting to the training data, a regularized cost function can be used, which is given as

$$L = MSE + \frac{\lambda}{m}||\beta||^m \tag{6}$$

where the hyperparameter λ controls the extent of regularization. The most commonly used regularization techniques include Lasso regression ($m = 1$), and Ridge regression ($m = 2$).

C. Assessing the Accuracy of the Model

This article uses the **$\mathbf{R^2}$ statistic** to assess the accuracy of the model. The R^2 statistic is a number between 0 and 1, which indicates the amount of variance in the data that can be explained by the model. The R^2 score is given by

$$R^2 = 1 - \frac{RSS}{TSS} \tag{7}$$

where RSS is the *residual sum of squares*, given by

$$RSS = \sum_{i=1}^{N}(y_i - \hat{y}_i)^2 \tag{8}$$

and TSS denotes the *total sum of squares*, given by

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

where $\bar{y}$ is the mean of all values of the response variable Y.

TSS quantifies the total variability in the data, and RSS quantifies the variability in the predicted values. Thus, a model with R^2 score of 0 does only as good as a model that predicts the mean for every given input, and a model with a higher R^2 score is able to fit the data better, and subsequently make better predictions.

III. The Problem

The objective of this article is to understand if there exists any relationship between cancer incidence and mortality, and socio-economic status. Data about cancer incidence and mortality rates in various counties in the US are obtained from [1]. Socio-economic data of the counties is obtained from [2]-[4]. The merged dataset created using these datasets contains the various fields of data for each county, given in I.

To explore the above-said relationships between the data, the steps followed were:

- Data cleaning
- Exploratory analysis
- Data visualization
- Creating models to describe the data variation
- Comparing various models to understand the best fit
- Conclusions from the best model

TABLE I
Various features of present in the merged dataset

Field	Description
State	
AreaName	
All_Poverty, M_Poverty, F_Poverty	Total number of people below poverty line, classified into number of males and females respectively
FIPS	State + County code
Med_Income	Median income of all ethnicities
Med_Income_White, Med_Income_Black, Med_Income_Nat_Am, Med_Income_Asian, Med_Income_Hispanic	Median income for people of each ethnicity
All_With, M_With, F_With	Total number of people with health insurance, divided into number of males and females
All_Without, M_Without, F_Without	Total number of people without health insurance, divided into number of males and females
Incidence_Rate	Age adjusted lung cancer incidence rate for every 100,000 people
Avg_Ann_Incidence	Number of people diagnosed with lung cancer per year
Recent_Trend	Trend in number of cases of lung cancer (stable, rising, or falling)
Mortality_Rate	Age adjusted death rate (because of lung cancer) per 100,000 people
Avg_Ann_Deaths	Average lung cancer mortalities

A. Data Cleaning

To clean the data, unwanted columns were first dropped. These included *State*, *AreaName*, and duplicate columns of *FIPS*. To quantitatively describe the *Recent_Trend*, rising, stable, and falling were replaced with +1, 0, and -1 respectively. Since poverty and insurance data was given in absolute numbers, new features *Poverty_Rate*, *M_Poverty_Rate*, *F_Poverty_Rate*, *All_Ins_Rate*, *M_Ins_Rate*, *F_Ins_Rate*, and *Total_Population* were created using existing features. Redundant features were then dropped.

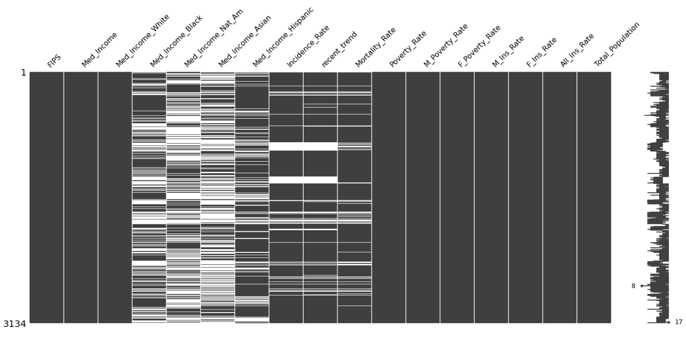


Fig. 1. Matrix of missing data, the white lines indicate that data pertaining to that column in missing

The **missingno** library was used to visualize the missing data points, the results of which are shown in Fig. 1 and Fig.2. Since the columns *Incidence_Rate* and *Mortality_Rate* are the target variables, rows where these variables were missing were discarded. For other fields like Med_Income, missing values were replaced with the median value from the distribution.

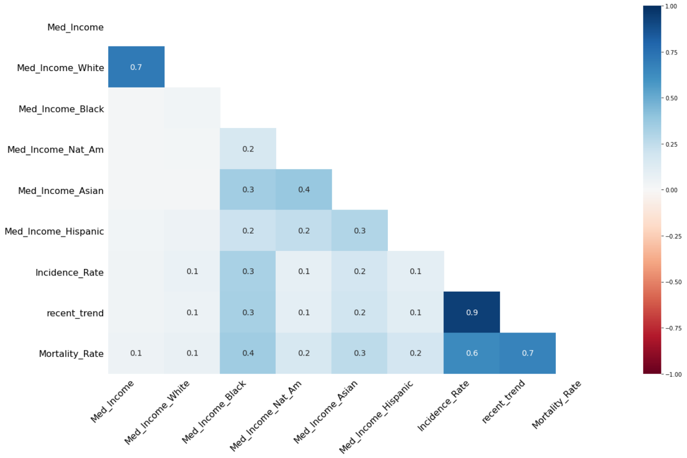


Fig. 2. A heatmap of missing data, a high correlation indicates that data pertaining to same rows is missing in both fields

B. Exploratory Data Analysis

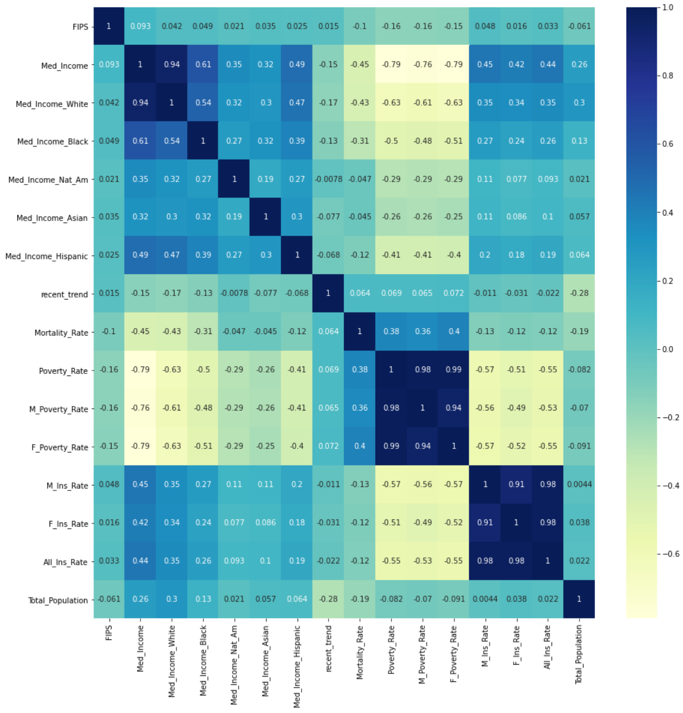


Fig. 3. A heatmap showing the relationship between various features

Relationship between variables was first visualized using a **correlation heatmap**, which is shown in Fig. 3. It can be seen that all the incomes are correlated with each other. Thus, two models will be built, one considering all the input features (except ones which have very high number of missing values), and another using only one income, insurance and poverty parameter. There is also a negative correlation between poverty rates and insurance rates, but since this correlation is not as high as that within poverty rates or insurance rates, both will be considered for the model.

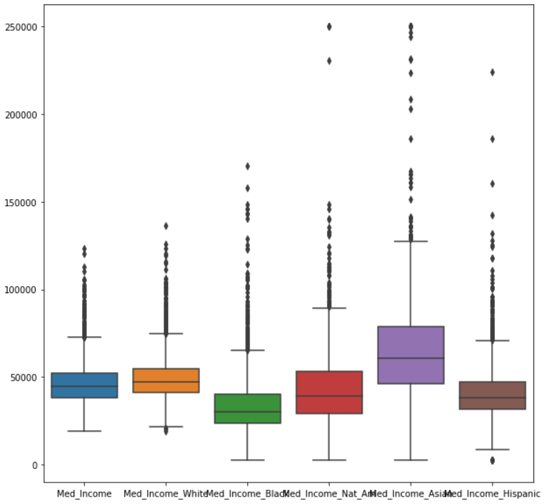


Fig. 4. A box-plot showing the distribution of income with respect to the ethnicity

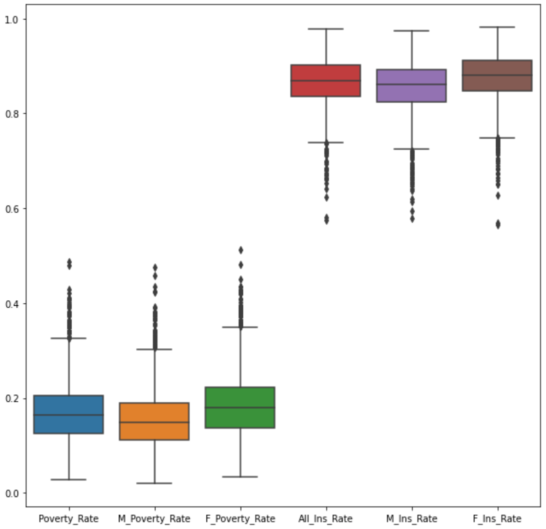


Fig. 5. A box-plot showing the distribution of rates of poverty, and the rates of those with insurance

C. Data Visualization

To visualize the distribution of the data further, box-plots Fig. 4 - Fig. 6 were plotted. The distribution of income, poverty rates, and insurance rates, as well as incidence and mortality rates, all are distributed equally across counties without significant skewness. Another interesting observation from Fig. 5 is that outliers are mostly counties with higher poverty rates, and lesser insurance rates.

To try and understand the relationship between incidence rate, mortality rate, and recent trend, Fig. 7 was plotted. The high correlation between incidence and mortality rates are well evident. The scatter plot is colour coded with respect to the recent trend, which shows that there is no observable relationship between the incidence and mortality rates, and the recent trend.

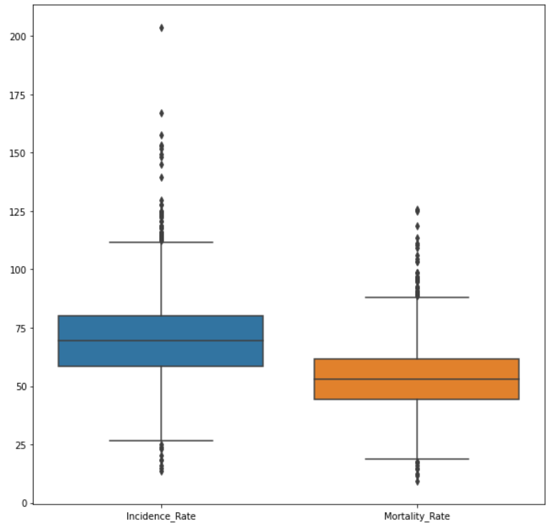


Fig. 6. A box-plot showing the distribution of incidence and mortality rates of cancer

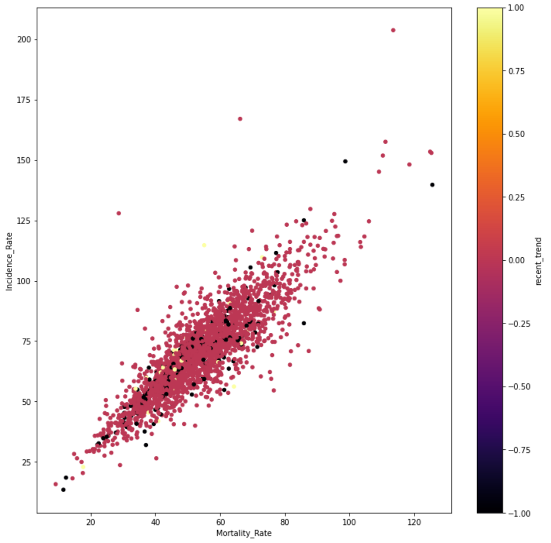


Fig. 7. A scatter plot showing the variation of incidence with mortality, colour coded by the recent trend

Fig. 8 was plotted to observe the variation of the incidence and mortality rates, with respect to the median income of the county. A weak quadratic relationship is observable in this plot. Fig. 9 shows the variation of incidence and mortality rates with respect to the insurance rates, and no observable pattern is seen here. Fig. 10 shows the variation of incidence and mortality rates with the poverty rates, and a linear relationship is observed here.

Similar plots were plotted for each feature including gender-wise insurance and poverty rates, as well as for ethnicity-wise incomes.

In addition to these, in order to understand how various input features vary with respect to each other, a pairwise scatter plot was created, which is shown in Fig. 15. The diagonal of

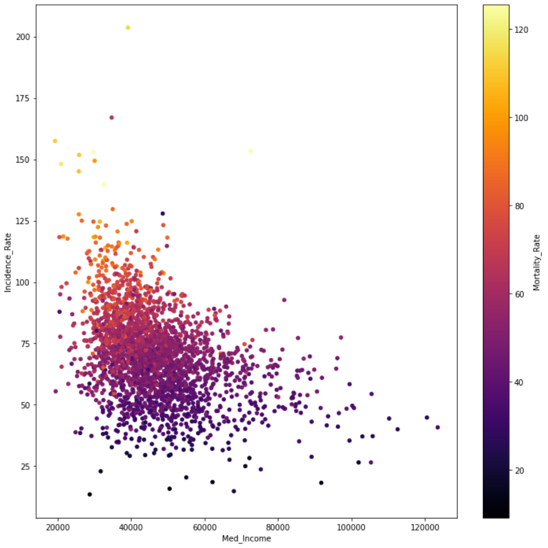


Fig. 8. A scatter plot showing the variation of incidence rates with median income of the county

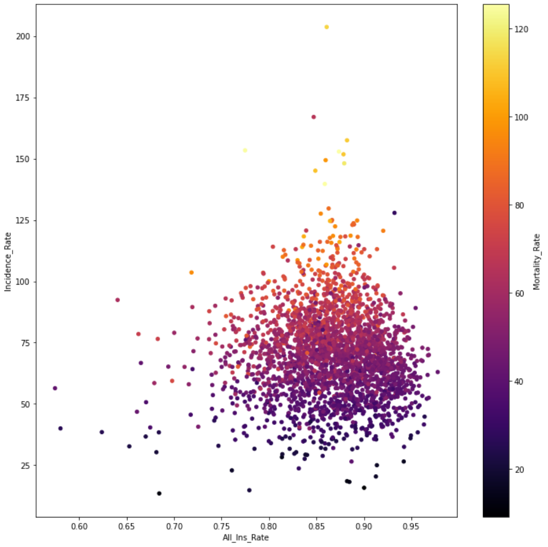


Fig. 9. A scatter plot showing the variation of incidence rates with insurance rates of the counties

the pairwise plot show the histogram of the feature.

D. Creating models to describe the data variation

The data was split into training (80%) and test (20%) datasets. Various models were hypothesized and tested, using the **sklearn** library in python. These models included multiple linear regression, polynomial regression, log transformed multiple linear regression, log transformed polynomial regression, lasso and ridge regression. In order to improve training, all the input features were scaled to values between 0 and 1.

E. Comparing various models to find the best fit

Models which use only three of the given features: Fig.11 shows the R^2 score of various models that try to predict the

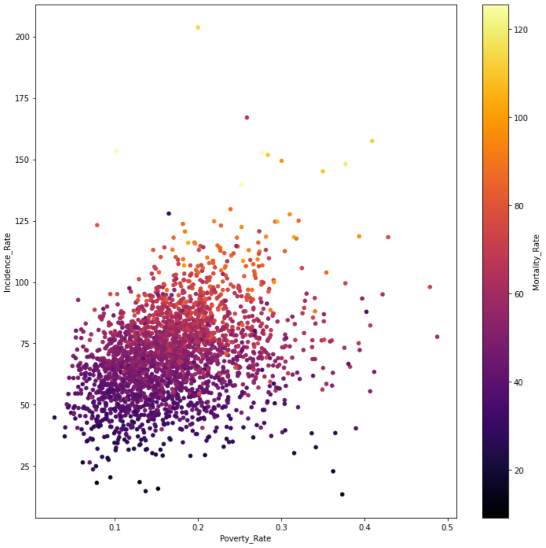
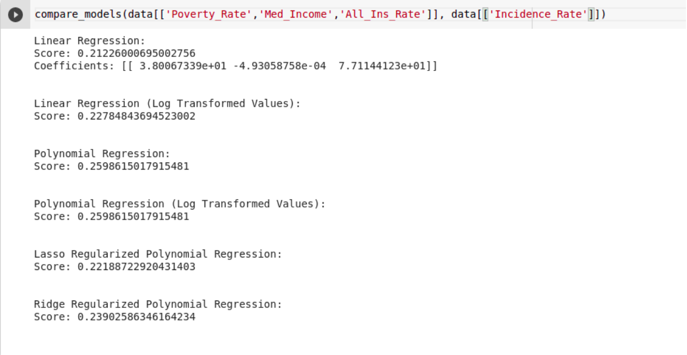


Fig. 10. A scatter plot showing the variation of incidence rates with poverty rates of the county

incidence rate using the three input variables (*Poverty_Rate*, *Med_Income*, and *All_Ins_Rate*). The coefficients printed for the linear model show that there is a weak negative relationship between median income and incidence rate, whereas there is a strong positive relationship between incidence rate and poverty rates, as well as with insurance rates. Similar trends can be observed in Fig. 12 for models predicting the mortality rates.

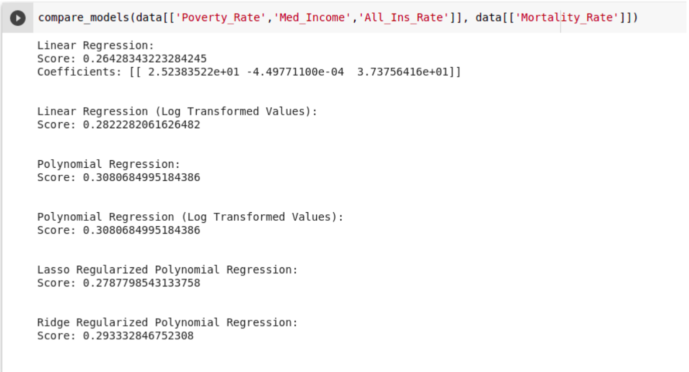
Models which use all the features given: Fig. 13 and Fig. 14 show the R^2 scores obtained for various models that take all the given features as inputs, to predict the incidence and mortality rates respectively. We observe that even though we use features with high multicollinearity, the results obtained are much better than the models which only uses a portion of the input features.



```
compare_models(data[['Poverty_Rate','Med_Income','All_Ins_Rate']], data[['Incidence_Rate']])

    Linear Regression:
    Score: 0.21226000695002756
    Coefficients: [[ 3.80067339e+01 -4.93058758e-04  7.71144123e+01]]


    Linear Regression (Log Transformed Values):
    Score: 0.22784843694523002


    Polynomial Regression:
    Score: 0.2598615017915481


    Polynomial Regression (Log Transformed Values):
    Score: 0.2598615017915481


    Lasso Regularized Polynomial Regression:
    Score: 0.22188722920431403


    Ridge Regularized Polynomial Regression:
    Score: 0.23902586346164234
```

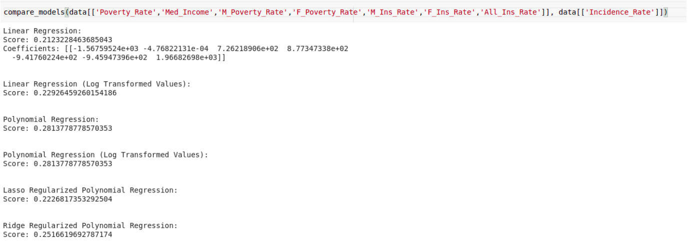
Fig. 11. R^2 scores obtained for various models trying to predict the incidence rate using the poverty rates, median income, and insurance rates.

The model that best fits the data is a **polynomial model that uses log-transformed values**, which is able to explain 28% of the variation in incidence rates and 31% of the variation in mortality rates using the poverty rates, median income, and insurance rates.



```
compare_models(data[['Poverty_Rate','Med_Income','All_Ins_Rate']], data[['Mortality_Rate']])

    Linear Regression:
    Score: 0.26428343223284245
    Coefficients: [[ 2.52383522e+01 -4.49771100e-04  3.73756416e+01]]


    Linear Regression (Log Transformed Values):
    Score: 0.2822282061626482


    Polynomial Regression:
    Score: 0.3080684995184386


    Polynomial Regression (Log Transformed Values):
    Score: 0.3080684995184386


    Lasso Regularized Polynomial Regression:
    Score: 0.2787798543133758


    Ridge Regularized Polynomial Regression:
    Score: 0.293332846752308
```

Fig. 12. R^2 scores obtained for various models trying to predict the mortality rate using the poverty rates, median income, and insurance rates.



```
compare_models(data[['Poverty_Rate','Med_Income','M_Poverty_Rate','F_Poverty_Rate','M_Ins_Rate','F_Ins_Rate','All_Ins_Rate']], data[['Incidence_Rate']])

Linear Regression:
Score: 0.2123228463685043
Coefficients: [[-1.56759524e+03 -4.76822131e-04  7.26218906e+02  8.77347338e+02
  -9.41760224e+02 -9.45947396e+02  1.96682698e+03]]


Linear Regression (Log Transformed Values):
Score: 0.22926459260154186


Polynomial Regression:
Score: 0.2813778778570353


Polynomial Regression (Log Transformed Values):
Score: 0.2813778778570353


Lasso Regularized Polynomial Regression:
Score: 0.2226817353292504


Ridge Regularized Polynomial Regression:
Score: 0.2516619692787174
```

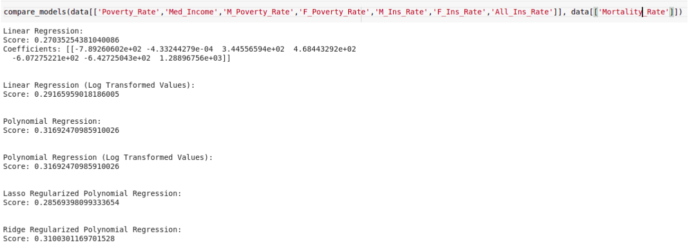
Fig. 13. R^2 scores obtained for various models trying to predict the incidence rate using all the input features

IV. Conclusions

As shown from Fig.8 and Fig.10, weak relationships can be noticed visually between incidence and mortality rates, and socio-economic status (poverty rates and median income). The R^2 scores shown in Fig. 11 and Fig. 12, especially for the polynomial model that uses log-transformed values, shows theat these relationships can be proved statistically. Thus, to some extent, socio-economic factors do influence the incidence and mortality rates of lung cancer.

References

[1] statecancerprofiles.cancer.gov
[2] https://data.world/uscensusbureau/acs-2015-5-e-poverty
[3] https://data.world/uscensusbureau/acs-2015-5-e-income
[4] https://data.world/uscensusbureau/acs-2015-5-e-income
[5] Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani, "Linear Regression" in *An Introduction to Statistical Learning*, New York, Springer, 2013.



```
compare_models(data[['Poverty_Rate','Med_Income','M_Poverty_Rate','F_Poverty_Rate','M_Ins_Rate','F_Ins_Rate','All_Ins_Rate']], data[['Mortality_Rate']])

Linear Regression:
Score: 0.27035254381040086
Coefficients: [[-7.89260602e+02 -4.33244279e-04  3.44556594e+02  4.68443292e+02
  -6.07275221e+02 -6.42725043e+02  1.28896756e+03]]


Linear Regression (Log Transformed Values):
Score: 0.29165959018186005


Polynomial Regression:
Score: 0.31692470985910026


Polynomial Regression (Log Transformed Values):
Score: 0.31692470985910026


Lasso Regularized Polynomial Regression:
Score: 0.28569398099333654


Ridge Regularized Polynomial Regression:
Score: 0.3100301169701528
```

Fig. 14. R^2 scores obtained for various models trying to predict the mortality rate using all the input features

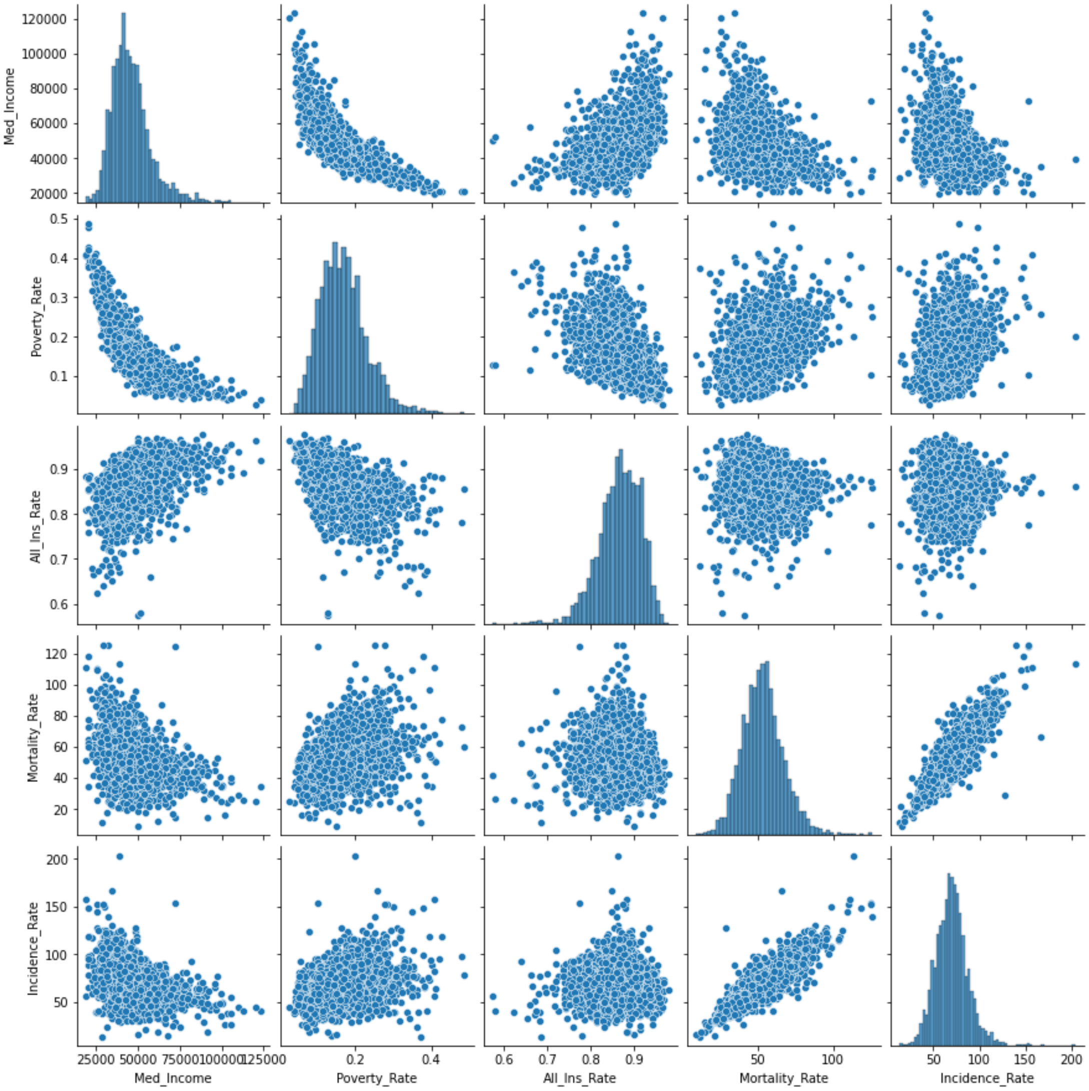


Fig. 15. Scatter plots showing the pairwise relationship between uncorrelated input features