A Mathematical Essay on Linear Regression

Shashwat Jangitwar*
Msc DataScience And Artificial Intelligence
Indian Institute Of Technology , Madras
Chennai, Indian
ge23c019@smail.iitm.ac.in

***Abstract*—This document gives rationale on Lower income groups in USA that are at a greater risk from being diagnosed or dying from cancer. Socioeconomic and racial/ethnic patterns of cancer patients discovered during the findings will help non profit organizations gather funds and check for disparity.This paper uses simple linear regressions techniques to find out relations and quantify the objectives.**

I. Introduction

Linear Regression Regression analysis is concerned with describing and evaluating the relationship between a given variable(explained or dependent variable) and one or more variables (explanatory or independent variables).Objectives of using linear regressions include, effects of changing explanatory variables on explained variables, forecast values based on explanatory variables and examine whether any of explanatory variables have significant effect on explained variables. With the help of linear regression we will try to examine whether lower income groups(sex-wise, ethnic) are at risk from diagnosed and dying from cancer. This analysis will help to understand socioeconomic dependence on mortality rates.Data gathered is cleaned first then imputation was done for missing values. Data exploration and visualization was done to find out correlation between variables and better modelling of regression .Model is trained , tested and analysed for obtaining relationship between socioeconomic factors on incident and mortality rates.

This short paper will use linear regression models and data cancer incidence rate and mortality are correlated with socioeconomic status

II. Linear Regression

A. What is Linear Regression

Linear regression provides statistical relationship(which do not give unique value of y for given value of x , but can be described exactly in probabilistic terms). It provides simple relations and easy interpret-able results. This model comes with few underlying assumptions.

1. Linearity -Explained and Explanatory variables should follows linear relations.
2. Homoscedasticity -Variance for error term remains constant.
3. Independence - Error terms to be independent with each other and also with corresponding explanatory variable.
4. Normality -Error terms normally distributed for all values.

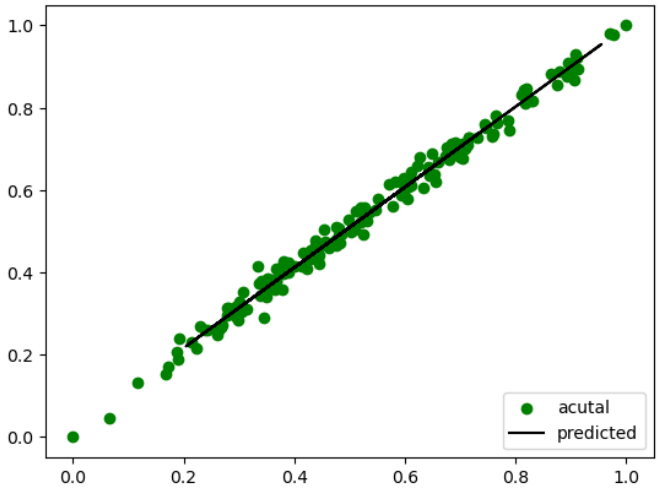

Fig. 1. Simple Linear Regression

5. Multicollinearity - Explanatory variables in case of multiple linear regression should not be highly correlated.

B. Types:

Following are different types of model based on explanatory variables.

1. Linear Regression- single explanatory variable.
2. Multiple Regression - multiple explanatory variables.
3. Polynomial Regression - explanatory variables expressed as higher dimensions.

C. Model

$$y = \alpha + x_1\beta_1 + x_2\beta_2 + x_3\beta_3 .....x_n\beta_n + \varepsilon$$

y= explained variable
$x1, x2, x3...xn$ = explanatory variables
α = intercept coefficient
$\beta_1, \beta_2, ...\beta_n$ = coefficients corresponding to explanatory variables
ε = error term

D. Parameter Estimates and Model Evaluation

Estimation of Parameters is done with the help of method of Least Squares, where squares of difference between actual value and predicted value are computed and minimized with respect to regression coefficients.

$$Q = \sum_{i=1}^{n}(y_i - \hat{\alpha} - \hat{\beta}_1 x_1 - \hat{\beta}_2 x_2 .....\hat{\beta}_n x_n)^2$$

Evaluation of model is done by r^2 also called coefficient of determination. This is computed by ratio of explained sum of Squares with total sum of squares .r^2 values always lies between 0 and 1. r^2 value close to 0 indicates variable x explains very little variation in y , and close to 1 explains most of the variation in y.

E. Hypothesis Testing

$y = \alpha + x_1\beta_1 + x_2\beta_2 + x_3\beta_3 ..... x_n\beta_n + \varepsilon$
Null hypothesis $H_0 : \beta_1 = \beta_2 = \beta_3 .... \beta_{=}0$
There is no linear regression of $x_1, x_2, x_3 ..... x_n$ on y
Alternative hypothesis H_1 : Some or all β values $\neq 0$
There exist some linear relationship

III. Problem Statement

We have data gathered from American Community Survey for the year 2011 to 2015, which contains estimates of population containing data related to heath , income , poverty from US census. Few features present in data set are adjusted as per US standards. Main objective of this paper is to quantify weather lower income groups are at greater risk from identifying and dying from cancer compared to other groups. To achieve this we will evaluate incidence rate and mortality rates dependence on various features such as income , number of people below poverty line , health insured etc. This task will be achieved by using multiple linear regression.

A. Data set

Dataset was merged which contained county wise information on 'State', 'AreaName', 'All Poverty', 'M Poverty', 'FPoverty', 'FIPS', 'Med Income', 'Med Income White', 'MedIncome Black', 'Med Income Nat Am', 'Med Income Asian', 'Hispanic', 'M With', 'M Without', 'F With', 'F Without','All With', 'All Without', 'fips x', 'Incidence Rate', 'Avg Ann Incidence', 'recent trend', 'fips y', 'Mortality Rate', 'Avg Ann Deaths'
This had total of 3134 rows and 47 columns , it contains state wise, gender wise data for people below poverty line and also median income based on ethnicity.

B. Methodology

- "Med Income Black", "Med Income Nat Am", "Med Income Asian", "Hispanic", this all columns contained most of the null values, dropping this values would lead to loss of information, hence this was substituted with median values of the column.
- Incidence rate had multiple missing values , since this was taken as explained variables , no imputation were applied and missing values rows were dropped.
- Plotting the correlation matrix for incomes across groups and average annual deaths , incidence rate to identify correlations amongst columns to avoid multicolinearity

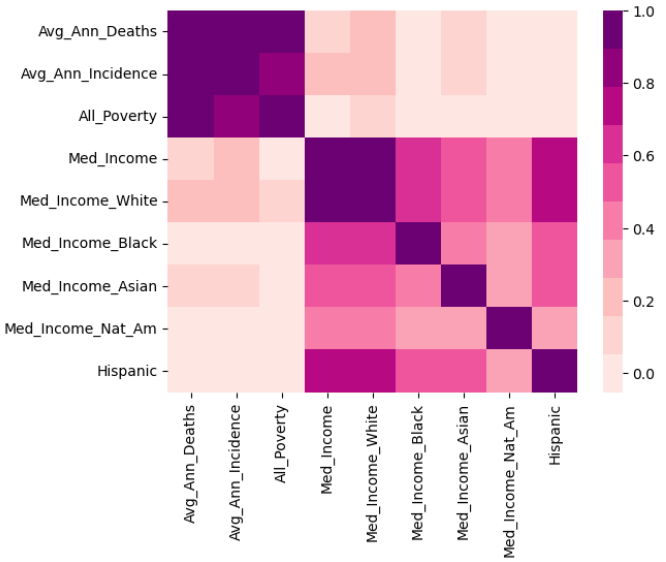

Fig. 2. Correlation between Columns

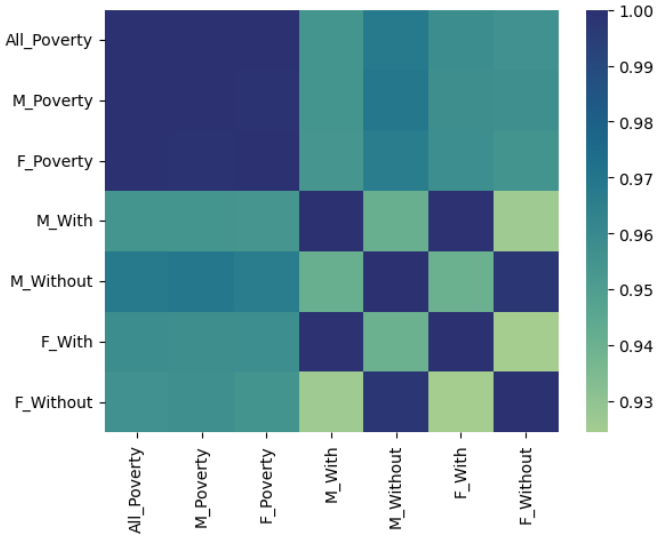

Fig. 3. Gender-wise Correlations

C. Data Analysis and visualization

- Gender-wise: From Fig.3. Men and Women are equally correlated with poverty line , irrespective of insurance coverage. No distinctions based on gender are observed, so not considering gender wise data for model
- Explained variable correlation-from Fig.4 we can see that incidence rate is highly correlated with mortality rate , we can infer that as incidence rate increases , mortality rate also increases.
- Income vs Mortality rate- As per common knowledge as income increases, standard of living improves which further improves awareness in individuals or groups , hence mortality rate should drop. As seen in Fig.4 we can say above statement is valid.

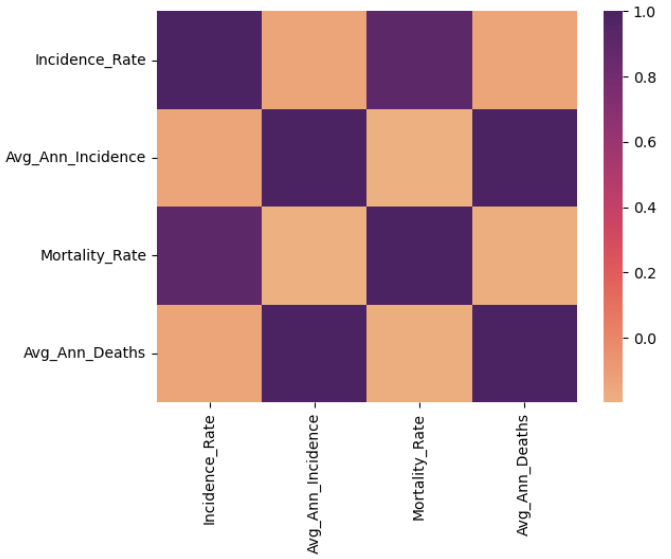

Fig. 4. Incidence Rate V/s Mortality rate

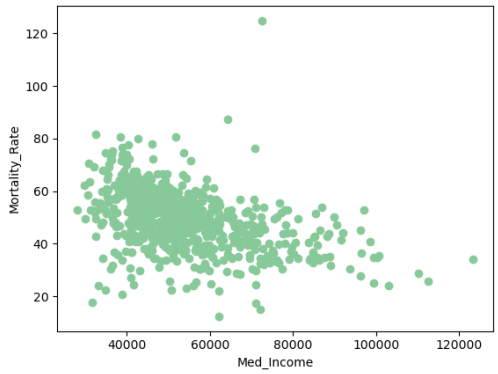

Fig. 5. Mortality Rate v/s Income

D. Linear Regression Model

Explained variable: "Avg Ann Deaths"
Explanatory Variables:"All Poverty","Med Income White", "Med Income Black", "Med Income Asian", "Med Income Nat Am","Hispanic". Performing Multiple Linear Regression with above values, we obtain the following results.

r^2 value = 0.841 , which means the values explanatory variables indicates large variations in explained variables.
Hypothesis Testing and Results
We can see that regression coefficients are non zero, meaning there exist some linear relationship between explanatory and explained variables.Also the p values corresponding to regression coefficients are very low, hence we reject null hypothesis.

IV. CONCLUSION

We can see clear from correlation graphs and above analysis that low income groups have high probability of being diagnosed and death from cancer. Among the low income groups Blacks, Hispanics, and Native Americans have less then median incomes, which means they have high chances of being diagnosed with cancer.No evidence was observed related to gender-wise mortality rate. With this findings non profit organization can use this for policy designs related to specific community and gather funding. This model can be improved when more diverse data is available which includes education , medical treatments, and health related information.

OLS Regression Results

Dep. Variable:	Avg_Ann_Deaths	R-squared:	0.841
Model:	OLS	Adj. R-squared:	0.840
Method:	Least Squares	F-statistic:	617.6
Date:	Fri, 01 Sep 2023	Prob (F-statistic):	2.89e-275
Time:	15:48:01	Log-Likelihood:	-4160.9
No. Observations:	705	AIC:	8336.
Df Residuals:	698	BIC:	8368.
Df Model:	6		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-38.5434	13.738	-2.806	0.005	-65.516	-11.571
All_Poverty	0.0018	3.19e-05	57.931	0.000	0.002	0.002
Med_Income_White	0.0027	0.000	7.643	0.000	0.002	0.003
Med_Income_Black	-0.0008	0.000	-3.430	0.001	-0.001	-0.000
Med_Income_Asian	0.0003	0.000	1.976	0.049	1.89e-06	0.001
Med_Income_Nat_Am	-0.0004	0.000	-2.725	0.007	-0.001	-0.000
Hispanic	-0.0006	0.000	-1.704	0.089	-0.001	9.59e-05

Omnibus:	307.077	Durbin-Watson:	1.708
Prob(Omnibus):	0.000	Jarque-Bera (JB):	4520.365
Skew:	1.549	Prob(JB):	0.00

Fig. 6. Enter Caption

REFERENCES

[1] Bishop, Christopher M., Pattern Recognition and Machine Learning, 2006.
[2] B. N. Ames, L. S. Gold, and W. C. Willett, "The causes and prevention of cancer," PNAS, 1995 Jun 6; 92(12): 5258–5265
[3] Madala G S "Introduction to Econometrics"