A mathematical essay on linear regression

Jeshlin Donna J
Department of Metallurgical and Materials Engineering
Indian Institute of Technology Madras
Chennai, India
mm20b029@smail.iitm.ac.in

***Abstract*—Linear regression is the mathematical technique of fitting data to a linear model. In this paper we explain the theory behind linear regression and implement the technique on a real-world dataset to capture the relationship between cancer incidence and mortality with socioeconomic status in the United States.**

***Index Terms*—Linear regression, fitting data, cancer incidence, mortality, socioeconomic status**

I. Introduction

Cancer has a major impact on society in the United States and across the world. And is the second leading cause of death in the United States. Monitoring and reducing health disparities according to socioeconomic status and race/ethnicity have long been an important health policy goal in the United States. The major behavioural determinants of cancer, such as smoking, diet, alcohol use, obesity, physical inactivity, reproductive behaviour, occupational and environmental exposures, and cancer screening, are themselves substantially influenced by individual-level and area-level socioeconomic factors. Analysing socioeconomic and racial/ethnic patterns in cancer mortality and incidence allows us to quantify cancer-related health disparities between the least- and most-advantaged social groups and to identify areas or population groups that are at greatest risk of cancer diagnosis and mortality and who may therefore benefit from targeted social and medical interventions. [1,2,6]

Linear regression analysis is used to predict the value of one or more dependant variables based on the values of other related explanatory variables(or independent variables) by modeling the relationship between them. The variables we want to predict is called the dependent variables. The variables used for the prediction are called the independent variables. This form of analysis estimates the coefficients of the linear equation, involving one or more independent variables that best predict the value of the dependent variable. Geometrically, linear regression fits a straight line or surface that minimizes the discrepancies between predicted and actual output values.

In this paper, we aim to examine the socioeconomic and racial/ethnic disparities in US cancer mortality and assess racial/ethnic and area-based socioeconomic patterns in cancer incidence and survival using the NPCR-NSS and SEER datasets. Such a comparison of cancer trends across population groups or areas may also provide important insights into the impact of cancer control interventions, such as smoking cessation, cancer screening, physical activity campaigns, and cancer treatment.

Socioeconomic and racial/ethnic disparities in US mortality, incidence, and survival rates from all-cancers combined were analyzed using the linear regression technique. And the correlation between them was discovered. The paper also provides quantitative and visual evidence that can be used advocate for better health outcomes for the more vulnerable population in the United States.

II. LINEAR REGRESSION

A. Regression

Regression is a statistical process for estimating the relationships among variables. It includes many techniques for modeling and analyzing several variables, where the focus is on the relationship between the dependent variable and one or more independent variables (or 'predictors'). More specifically, regression helps one understand how the value of the dependent variables change when the independent variables are varied.

Most commonly, regression analysis estimates the conditional expectation of the dependent variable given the independent variables – that is, the average value of the dependent variable when the independent variables are fixed. Less commonly, the focus is on a quantile, or other location parameters of the conditional distribution of the dependent variable given the independent variables. In all cases, the estimation target is a function of the independent variables called the regression function.

B. Linear Regression

Linear regression is an approach for modeling the relationship between a scalar dependent variable y and one or more explanatory variables (or independent variables) denoted X into a linear form. The case of one explanatory variable is called simple linear regression. For more than one explanatory variable, the process is called multiple linear regression.

The idea behind simple linear regression is to "fit" the observations of two variables into a linear relationship

between them. Graphically, the task is to draw the line that is "best-fitting" or "closest" to the points (x_i,y_i),(xi,yi), where x_i, xi and y_i, yi are observations of the two variables which are assumed to depend linearly on each other.

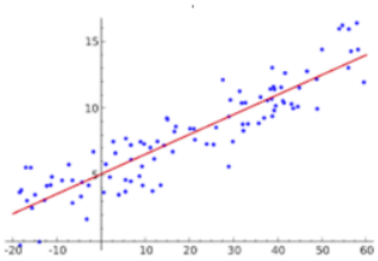

Fig. 1. The best fitting linear relationship between the variables x and y.

The general formula for linear regression is:

$$\mathrm{Y} = \mathrm{Z}\beta + \epsilon \tag{1}$$

where β represents the weights/coefficients of the model and ϵ represents the bias.

Suppose we have a sample of size n.

$$Y^{j^r}_{n\times m} = \begin{bmatrix} Y_{11} & Y_{12} & \cdots & Y_{1p} \\ \vdots & \vdots & \vdots\!\vdots & \\ Y_{21} & Y_{22} & \cdots & Y_{2p} \\ Y_{n1} & Y_{n2} & \cdots & Y_{np} \end{bmatrix} = [\mathbf{Y}_{(1)}\mathbf{Y}_{(2)}\cdots\mathbf{Y}_{(p)}], \tag{2}$$

where Y(i) is the vector of n measurements of the ith variable. Also,

$$\beta_{(r+1)\times m} = \begin{bmatrix} \beta_{01} & \beta_{02} & \cdots & \beta_{0m} \\ \beta_{11} & \beta_{12} & \cdots & \beta_{1m} \\ \vdots & \vdots & \vdots & \vdots \\ \beta_{r1} & \beta_{r2} & \cdots & \beta_{rn} \end{bmatrix} = [\beta_{(1)}\beta_{(2)}\cdots\beta_{(m)}], \tag{3}$$

where (i) are the (r+1) regression coefficients in the model for the ith variable. Finally, the p n-dimensional vectors of errors C=(i),i=1,...,p are also arranged in an n×p matrix

$$\epsilon = \begin{bmatrix} \epsilon_{11} & \epsilon_{12} & \cdots & \epsilon_{1p} \\ \epsilon_{21} & \epsilon_{22} & \cdots & e_{2p} \\ \vdots & \vdots & \cdots & \vdots \\ \epsilon_{n1} & \epsilon_{n2} & \cdots & \epsilon_{n\iota)} \end{bmatrix} = [\epsilon_{(1)}\epsilon_{(2)}\cdots\epsilon_{(p)}] = \begin{bmatrix} \epsilon'_1 \\ \epsilon'_2 \\ \vdots \\ \epsilon'_n \end{bmatrix}, \tag{4}$$

C. Learning a linear regression model

Learning a linear regression model involves estimating the values of the coefficients used in the linear representation using the training data.

1) Simple Linear Regression Model: If the data matrix X contains only two variables, a constant and a scalar regressor x_i, then this is called the "simple regression model". The parameters are denoted as α, β:

$$y_i = \alpha + \beta x_i \tag{5}$$

The least squares estimates in this case are given by:

$$\hat{\beta} = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{n\sum x_i^2 - (\sum x_i)^2} \tag{6}$$

$$\hat{\alpha} = \overline{y} - \hat{\beta}\,\overline{x} \tag{7}$$

2) Ordinary Least Squares: The Ordinary Least Squares procedure seeks to minimize the sum of the squared residuals. This means that given a regression line through the data, we calculate the distance from each data point to the regression line, square it, and sum all the squared errors together to get the quantity that ordinary least squares seeks to minimize.

The goal is to find the coefficients β which fit the equations "best", in the sense of solving the quadratic minimization problem.

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} S(\boldsymbol{\beta}), \tag{8}$$

where the objective function S is given by

$$S(\boldsymbol{\beta}) = \sum_{i=1}^{n}\left|y_i - \sum_{j=1}^{p} X_{ij}\beta_j\right|^2 = \left\|\mathbf{y} - \mathrm{X}\boldsymbol{\beta}\right\|^2 \tag{9}$$

This approach treats the data as a matrix and uses linear algebra operations to estimate the optimal values for the coefficients.

3) Gradient Descent: Gradient descent is a first-order iterative optimization algorithm for finding a local minimum of a differentiable function. The idea is to take repeated steps in the opposite direction of the gradient (or approximate gradient) of the function at the current point, because this is the direction of steepest descent.

The gradient descent technique starts with random values for each coefficient. The sum of the squared errors are calculated for each pair of input and output values. A learning rate(γ) is used as a scale factor and the coefficients are updated in the direction towards minimizing the error. The process is repeated until a minimum sum squared error is achieved or no further improvement is possible.

The learning rate (γ) parameter determines the size of the improvement step to take on each iteration of the procedure.

Gradient descent is based on the observation that if the multi-variable function $F(x)$ is defined and differentiable in a neighborhood of a point a , then $F(x)$ decreases fastest if one goes from a in the direction of the negative gradient of F at a, -∇F(a).

It follows that, if

$$a_(n+1) = a_n - \gamma\nabla F(a_n) \tag{10}$$

for a $\gamma \in_+$ small enough, then F(a_n)$\geq$ F(a$_(n+1)$). In other words, the term $\gamma \nabla F(a)$ is subtracted from a because we want to move against the gradient, toward the local minimum. With this observation in mind, one starts with a guess x_0 for a local minimum of F, and considers the sequence $x_0, x_1, x_2, \ldots$ such that

$$x_(n+1) = x_n - \gamma_n \nabla F(x_n), \ n \geq 0 \tag{11}$$

We get a monotonic sequence

$$F(x_0) \geq F(x_1) \geq F(x_2) \geq \cdots, \tag{12}$$

so, hopefully, the sequence (x_n) converges to the desired local minimum. With certain assumptions on the function F and particular choices of γ ,

$$\gamma_n = \frac{\left| (\mathbf{x}_n - \mathbf{x}_{n-1})^T \left[\nabla F(x_n) - \nabla F(x_(n-1)) \right] \right|}{\left\| \nabla F(x_n) - \nabla F(x_(n-1)) \right\|^2} \tag{13}$$

convergence to a local minimum can be guaranteed. When the function F is convex, all local minima are also global minima, so in this case gradient descent can converge to the global solution.

4) Regularization: The extensions of the training of a linear model is called regularization. This aims to minimize the sum of the squared error of the model on the training data (using ordinary least squares) but also to reduce the complexity of the model (like the number or absolute size of the sum of all coefficients in the model) at the same time.

Two important examples of regularization methods for linear regression are:
1. Lasso Regression(or L1 regularization): where ordinary least squares is modified to also minimize the absolute sum of the coefficients.
2. Ridge Regression(or L2 regularization): where ordinary least squares is modified to also minimize the squared absolute sum of the coefficients.

These methods are espicially effective to use when there is collinearity in the input values and ordinary least squares would overfit the training data.

III. USING LINEAR REGRESSION TO SOLVE THE PROBLEM

A. The problem

Analyzing socioeconomic and racial/ethnic patterns in cancer mortality and incidence allows us to quantify cancer-related health disparities between the least- and most-advantaged social groups and to identify areas or population groups that are at greatest risk of cancer diagnosis and mortality and who may therefore benefit from targeted social and medical interventions. Comparison of cancer trends across population groups or areas may provide important insights into the impact of cancer control interventions, such as smoking cessation, cancer screening, physical activity campaigns, and cancer treatment.

We examined disparities in cancer mortality, incidence, and survival using three national data sources: the national mortality database, the 1979–2011 National Longitudinal Mortality Study (NLMS), and the SEER cancer registry database. The national mortality database has been the primary source of mortality analyses and surveillance by age, sex, race/ethnicity, cause of death, and place of residence for over a century. Using prospectively linked census and mortality records, we analyze individual-level racial/ethnic and socioeconomic inequalities in mortality from cancer.

B. Applying linear regression to solve the problem

1) Preparing the data for analysis: We merge the three data-frames provided to us to form the final data-frame. The rows at the end of this merged data-frame contain significant number of missing values in various columns/features. This has been handled by deleting these rows completely. The missing values in the columns/variables which had low variance of values within them were imputed using their mean value.

It was also noted that there is notably a large difference between 75th %tile and max values of all the variables. This suggests that there are extreme values(outliers) in the data.

The other missing values present in the data-frame were imputed using the correlation between the various features/variables in the data-frame. It was also found that multiple columns of the data-frame had high correlation between each other, these columns were hence removed from the data before training with the help of SimpleImputer from the datawig library to avoid redundancy.

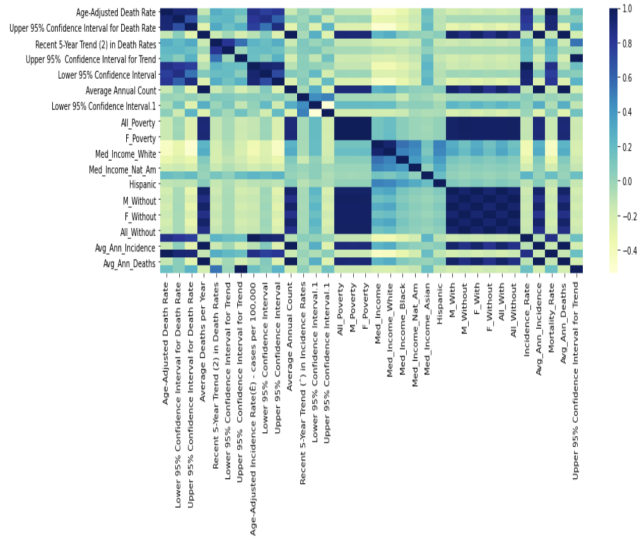

Fig. 2. Correlation heat-map between the various features of the data

2) Building the Linear Regression Model: A linear regression model had been built using the inbuilt $nn.Linear$ method from the Pytorch library. The features- Male Poverty, Female Poverty, Median Income, Median Income White, Median Income Black,Median Income Native Americans, Median Income Asian, Hispanic and Met Objective of 45.5 were taken as the independent input variables for the linear regression model. These features were used to predict the dependant variables Average Annual Incidence and Average Annual Deaths.

C. Insights and Observations

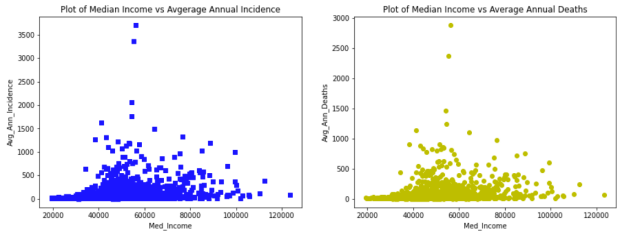

Fig. 3. Variation of cancer mortality and incidence with income status

From (Fig.2.) The plot between Mortality rate and Income shows a general trend that mortality rate is lesser for those people having high income. This can be due to the fact that people having a higher income can afford sophisticated cancer treatments like chemotherapy.

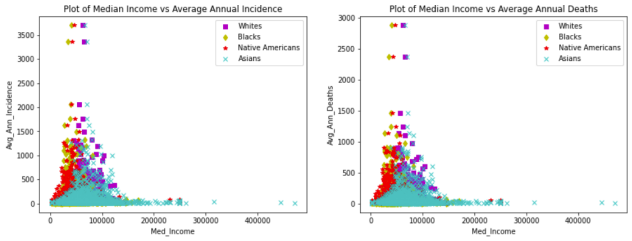

Fig. 4. Variation of cancer mortality and incidence with ethnic group

From (Fig.3.) we infer that the average incidence and mortality also depends on the kind of ethnic group. Taking any particular value of income, we can infer that the average incidence and mortality increases with the type of ethnic group in the order: Native American, Blacks, Asian, Whites.

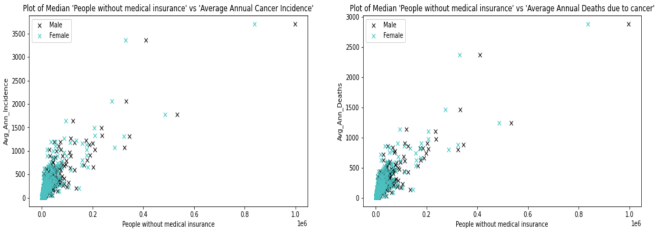

Fig. 5. Variation of cancer mortality and incidence with gender identity

From (Fig.4.) we infer that the the average incidence and mortality also depends on the gender. Taking any particular value of "number of people of the considered gender without medical insurance", we can infer that the average incidence and mortality of male is greater than that of female.

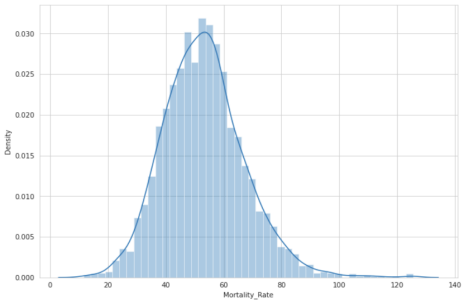

Fig. 6. Distribution of Mortality Rate

From plotting the distributions of the different variables, it has been inferred that the variables "Upper 95% Confidence interval", "Mortality Rates" (Fig.6.) and "Recent 5-Year Trend in Incidence Rates" columns appear to be normally distributed. And all other variables are either positively or negatively skewed.

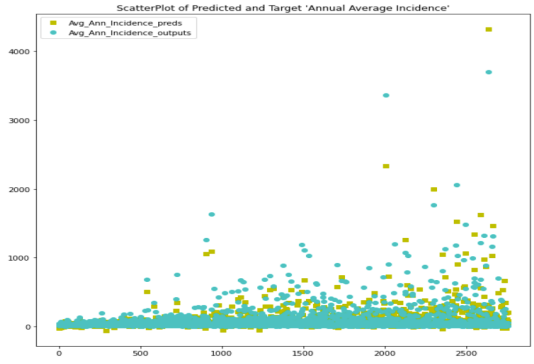

Fig. 7. Predicted and target values of 'Annual Average Incidence'

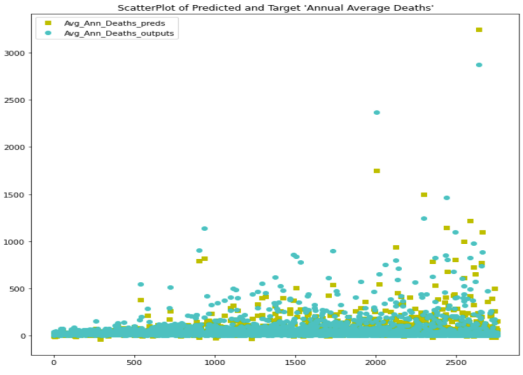

Fig. 8. Predicted and target values of 'Annual Average Deaths'

From (Fig.7.) and (Fig.8.), we can infer that our linear model surprisingly fits the data quite well. The mean square error obtained in the linear regression model was 63.0418. This implicitly means that there is a linear correlation of cancer incidence and mortality with income levels, gender and ethnicity.

IV. CONCLUSIONS

We analyzed socioeconomic and racial/ethnic disparities in US mortality, incidence, and survival rates from all-cancers combined. The average incidence and mortality of males were found to be greater than that of female individuals. Individuals in more deprived areas or lower education and income groups had higher mortality and incidence rates than their more affluent counterparts. Mortality was higher among Blacks and lower among Asians and Whites. Cancer patient survival was significantly lower in more deprived neighborhoods and among most ethnic-minority groups. Cancer mortality and incidence disparities may reflect inequalities in smoking, obesity, physical inactivity, diet, alcohol use, screening, and treatment.

Cancer is a leading cause of death in the US and the most prominent cause of death in terms of years of potential life lost. Evidence presented in this paper indicates how cancer disparities contribute greatly to the overall health inequalities in the US. With large socioeconomic and racial/ethnic inequalities in smoking, obesity, and physical inactivity among young people continuing to persist, inequalities in US cancer mortality and incidence are not expected to diminish in the foreseeable future. Healthcare inequalities have also risen in both absolute and relative terms and socioeconomic and racial/ethnic disparities in stage at diagnosis and survival from major cancers are present. These trends would also imply continued social inequalities in cancer mortality and incidence in the future. Health policies therefore should enhance access to cancer screening programs among the disadvantaged populations and underserved areas. Lastly, social policy measures aimed at improving the broader social determinants, such as material living conditions and the social and physical environments, are needed to tackle health inequalities in cancer outcomes

Further improvements in the model are highly possible. Using the population data in different areas of the United States is likely improve the model. While imputing for missing values, SimpleImputer was directly used, instead of this, different methods of imputation can also be tried out based on further research on the data. In the present model, outliers were not accounted for, further work can done on addressing the outliers present in the data. Other methods of fitting a linear regression model like grid search, gradient descent can be done as well. More emphasis should also be given to hyperparameter tuning to improve the accuracy of the linear model to better fit the data.

REFERENCES

[1] US Department of Health and Human Services, Healthy People 2020, http://www.healthypeople.gov/2020/default.aspx.

[2] National Center for Health Statistics, Health, United States, 2011 with Special Feature on Socioeconomic Status and Health, US Department of Health and Human Services, Hyattsville, Md, USA, 2012.

[3] G. K. Singh, B. A. Miller, B. F. Hankey, and B. K. Edwards, Area Socioeconomic Variations in U.S. Cancer Incidence, Mortality, Stage, Treatment, and Survival, 1975–1999, NCI Cancer Surveillance Monograph Series No. 4, NIH Publication No. 03-5417, National Cancer Institute, Bethesda, Md, USA, 2003, http://seer.cancer.gov/publications/ses/index.html.

[4] G. K. Singh and A. Jemal, "Socioeconomic inequalities in cancer incidence and mortality," in American Cancer Society's Clinical Oncology Textbook, T. Gansler, Ed., Wiley, New York, NY, USA, 2017.

[5] F. Faggiano, T. Partanen, M. Kogevinas, and P. Boffetta, "Socioeconomic differences in cancer incidence and mortality," IARC Scientific Publications, no. 138, pp. 65–176, 1997.

[6] D. L. Blackwell, J. W. Lucas, and T. C. Clarke, "Summary health statistics for U.S. adults: national health interview survey, 2012," Vital and Health Statistics, vol. 10, no. 260, pp. 1–161, 2014.