Exploring Data Insights through Linear Regression Analysis: A Comprehensive Study

Boneshwar V K
dept. of Biotechnology - IDDD Cyber Physical Sytems
Indian Instituteof Technology, Madras
Chennai, India
bs20b012@smail.iitm.ac.in

***Abstract*—Linear regression is a statistical approach for estimating the value of a dependent variable from an independent variable. A measure of the connection between two variables is linear regression. A dependent variable is predicted using this modeling technique based on one or more independent factors. Of all statistical methods, linear regression analysis is the one that is most frequently utilized.**

***Index Terms*—Linear regression, Exploratory Data Analysis, Cancer death in the United States.**

I. Introduction

The impact of socio-economic factors on cancer incidence rates has emerged as a critical area of investigation within the realm of public health research. In the United States, where cancer remains a prominent health concern, understanding the intricate relationship between socio-economic variables and the occurrence of cancer incidents is pivotal for devising targeted prevention and intervention strategies. This paper delves into the multifaceted connections between socio-economic status and cancer occurrence, shedding light on the mechanisms through which disparities in income, education, healthcare access, and lifestyle choices contribute to varying cancer incidence rates across different segments of the population.

As researchers strive to uncover the intricate interplay of socio-economic factors such as source of income and cancer incidence, it becomes evident that this issue encompasses a spectrum of influences spanning from individual health behaviors to broader societal structures. This paper aims to present a comprehensive overview of the existing literature on this subject, amalgamating insights from epidemiological studies, health disparities research, and socio-economic analyses. By elucidating the intricate pathways through which socio-economic factors impact cancer occurrence, this paper not only contributes to the theoretical understanding of the issue but also provides a foundation for designing evidence-based interventions and policies aimed at mitigating the disparities in cancer incidence related to socio-economic status. Through this exploration, we endeavor to pave the way for a more equitable and effective approach to addressing cancer as a public health challenge in the United States.

II. Impact of Income

A. Income Disparities and Cancer Incidence rate

In the pursuit of comprehending the intricate relationship between income disparities and cancer incidence across diverse racial groups, this study embarks on a rigorous analytical exploration underpinned by statistical modeling. Specifically, we propose the development of a linear regression model where cancer incidence serves as the dependent variable, and income emerges as the independent variable. Leveraging a meticulously curated dataset encompassing finely stratified income brackets and corresponding cancer incidence rates, this research leverages a nuanced statistical framework.

The proposed linear regression model adopts a multi-dimensional perspective, incorporating categorical dummy variables to encapsulate the spectrum of racial identities: Asian, Black, White, Native American, and Hispanic populations. By employing this sophisticated model, we aim to discern potential correlations between income disparities and cancer incidence within each distinct racial cohort. Critically, to ensure the model's robustness and the precision of its findings, we will diligently account for confounding variables such as age, gender, and geographic location.

Following the derivation of regression coefficients and precise calculation of correlation terms for each racial category, the research proceeds to undertake an intricate comparative analysis. This analysis seeks to meticulously gauge the significance and magnitude of the observed correlations. To imbue clarity and immediacy to the findings, the study employs data visualization techniques, manifesting in the form of scatter plots. These plots, uniquely tailored for Asian, Black, White, Native American, and Hispanic populations, artfully depict the dynamic interplay between income and cancer incidence within each racial cohort. Augmenting the visual comprehension, regression lines are superimposed on these scatter plots, succinctly encapsulating the direction and strength of the income-cancer incidence relationship within each racial context. The ensuing correlation analyses and meticulously designed plots synergistically contribute profound insights into potential trends, disparities, and interrelationships.

The following plots aids in visualizing the impact of income in cancer incidence for individual race.

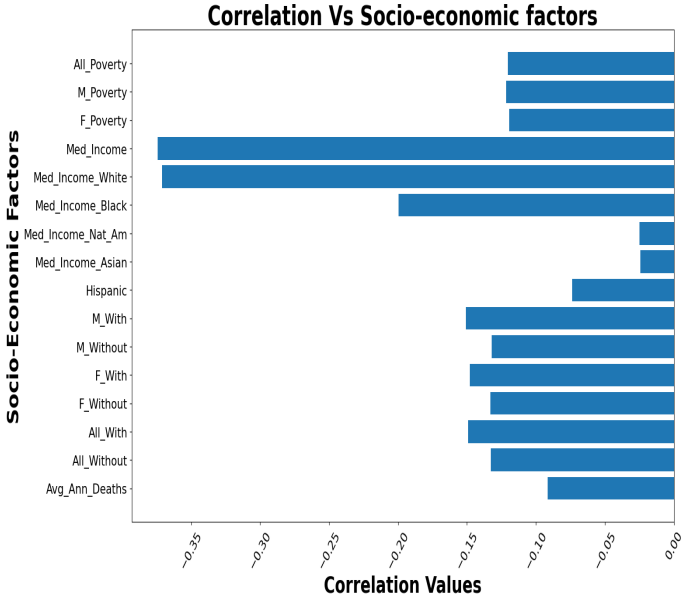

Fig. 1. Visualizing the impact of several socio-economic factors on cancer incidence

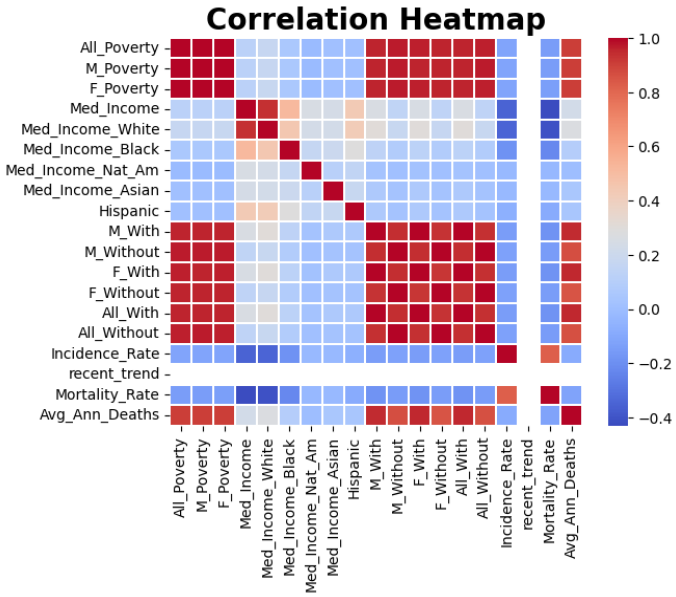

Fig. 2. The correlation heatmap with a higher threshold to visualize indispensable data

It is clearly evident that the median income plays a vital role in cancer occurrence among the citizens of United States. An insightful revelation gleaned from the dataset is the substantial representation of the White community, which not only comprises a significant proportion of the overall population but also exhibits a median income closely mirroring the overall median income. This finding underscores the demographic prominence of the White population within the studied region, as well as their relatively stable economic standing. The parallel nature of the White community's median income to the overall median income is indicative of a socioeconomic equilibrium that has been achieved despite their larger population size. This equilibrium could be attributed to factors such as access to education, employment opportunities, and economic resources, which have likely contributed to sustaining income levels in line with the broader population. This valuable insight highlights the intricate interplay between population size and income dynamics and underscores the need for tailored policies and initiatives aimed at harnessing the economic potential of diverse racial communities while ensuring equitable opportunities and outcomes for all.

B. Influence of Income on Distinct Ethnic Communities

In the presented grouped box plot, a compelling narrative emerges regarding the socioeconomic landscape across different racial groups. Notably, the African American (Black) community exhibits the lowest median income, followed closely by the Hispanic population, and then by Native Americans. This observation underscores persistent income disparities faced by these racial groups, reflecting broader economic challenges within these communities. The median income values for both Black and Hispanic populations signify a pressing need for targeted interventions and policies aimed at addressing income inequality and improving economic prospects for historically marginalized groups.

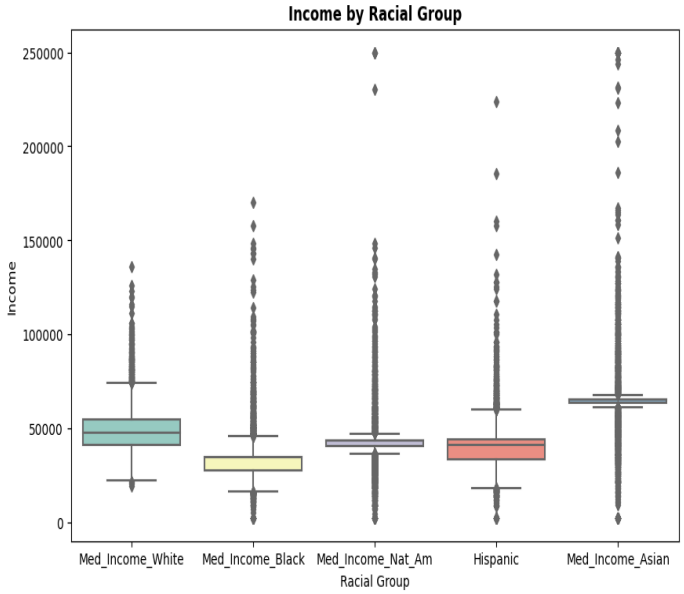

Fig. 3. Insights from visualizing the median income of different communities of United States.

Conversely, the grouped box plot reveals that the White population exhibits a relatively higher median income, indicative of a comparatively improved economic standing. Furthermore, the Asian racial group stands out with the highest median income among the analyzed demographics. This finding underscores the importance of recognizing the distinct socioeconomic experiences of various racial communities, highlighting the need for tailored policy initiatives that address income

disparities, promote equity, and foster economic empowerment for historically disadvantaged groups. The observed disparities in median income among these racial categories warrant further investigation and strategic policy interventions aimed at creating a more equitable economic landscape for all members of society.

C. Impact of insurance

A compelling insight emerges from the data visualization, where the insurance status of individuals has been thoughtfully categorized into six distinct groups based on gender (male, female, and both genders combined) and their insurance coverage (either possessing insurance or lacking it). The scatter plot reveals a nuanced portrayal of healthcare access and insurance prevalence among different gender cohorts. It underscores the multifaceted nature of insurance distribution, indicating variations not only across gender lines but also within each gender category. This granular approach to examining insurance status can be instrumental in identifying specific demographic segments that may be disproportionately affected by insurance disparities, thereby informing targeted interventions and policy measures aimed at achieving comprehensive healthcare coverage and equity for all individuals, regardless of gender. Such insights derived from the scatter plot serve as invaluable inputs for evidence-based decision-making and the formulation of inclusive healthcare policies that cater to the diverse needs of the population.

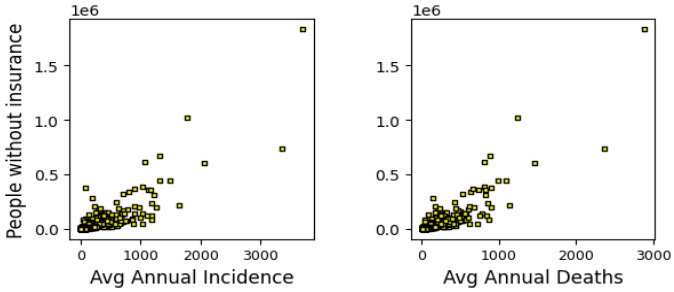

Fig. 4. Insights from visualizing the role of insurance.

The above plots aids in visualizing the influence of insurance on the general population. People with no insurance and fall under the poverty category tend to have cancer caused death rates higher than those who have insurance.

III. Linear Regression

Linear regression is a statistical modeling technique widely employed in data analysis and prediction tasks.It seeks to establish a linear relationship between one or more independent variables and a continuous dependent variable. In this method, the objective is to identify the linear equation that best describes the association between the variables, often represented as y = mx + b, where 'y' is the dependent variable, 'x' is the independent variable, 'm' is the slope, and 'b' is the intercept. Linear regression is instrumental in quantifying the impact of independent variables on the dependent variable, facilitating predictive modeling, trend analysis, and hypothesis testing. It plays a pivotal role in various applications, including engineering, data science, and healthcare, offering valuable insights into the underlying relationships and enabling data-driven decision-making processes.

TABLE I
Feature Values

Features	Values
$M\_Poverty$	-1.8397
$F\_Poverty$	1.6424
Med_Income_White	-0.4207
Med_Income_Black	-0.0637
$Med\_Income\_Nat\_Am$	0.0592
Med_Income_Asian	0.0455
$Hispanic$	0.0979
M_With	-0.2977
$M\_Without$	0.5546
F_With	0.5309
$F\_Without$	-0.6351

In this study, we have designed a linear regression model with the primary objective of predicting the average annual incidence of a particular health condition, leveraging a dataset that incorporates several crucial predictors, including 'All_Poverty,' 'Med_Income,' 'All_With,' and 'recent_trend.' The utilization of a linear regression framework offers a robust means of uncovering relationships and trends within the data, as well as quantifying the impact of each predictor variable on the target variable, i.e., the average annual incidence. 'All_Poverty' and 'Med_Income' represent essential socioeconomic indicators that may exert a substantial influence on health outcomes, while 'All_With' provides insight into the prevalence of insurance coverage—a factor closely tied to healthcare access. Additionally, 'recent_trend' serves as a dynamic variable capturing temporal trends, further enhancing the predictive power of our model. By applying linear regression, we aim to develop a predictive tool capable of informing public health initiatives, healthcare resource allocation, and policy decisions, ultimately contributing to improved healthcare planning and delivery within the studied population.

The following table throws some light on the model and important feature values for each data columns.

This code creates a table with IEEE paper format specifications, including proper caption, label, and column formatting. Remember to include this code in your LaTeX document within the appropriate section to display the table correctly.

IV. Conclusion

We investigated the influence of one's economic status on the occurrence of cancer and found that it plays a substantial role. This pattern extends to other diseases as well, underscoring the significance of accessible healthcare services for all. While the debate around socioeconomic impacts and creating an affordable healthcare system for everyone is pertinent, our focus in this paper is on linear regression and its interpretability.

As previously mentioned, model interpretability tends to decrease as complexity and flexibility increase. Linear regres-

sion stands out as a model that is comparatively easy to interpret when contrasted with tree-based methods or support vector machines, which are more intricate. By examining the coefficients' p-values, we can extract valuable insights from the data. The income plays a vital role in cancer caused deaths. The statistics obtained from the plots state that the race Blacks suffer the most in United States, while on the other side of the coin, Asians drive the median income higher and Whites' median income stays almost the same as overall median income. Hispanic community falls behind the Native Americans, yet they do not suffer as much as the Blacks. Our model here predicts with an R square score of 89.96 % with the training and testing data split typically being $80-20$.
We often prioritize sophisticated models like neural networks due to their enhanced flexibility, but simpler models such as linear regression are sometimes more relevant because they allow for a clearer understanding of the data.

REFERENCES

[1] M. Huang, "Theory and Implementation of linear regression," 2020 International Conference on Computer Vision, Image and Deep Learning (CVIDL), Chongqing, China, 2020, pp. 210-217, doi: 10.1109/CVIDL51233.2020.00-99.

[2] J. Abello and J. Korn, "MGV: A System for Visualizing Massive Multidigraphs", IEEE Trans. on Visualization and Computer Graphics, vol. 8, no. 1, pp. 21-38, 2002.