Communications on Applied Nonlinear Analysis
ISSN: 1074-133X
Vol 32 No. 8s (2025)

Mathematical Modeling and analysis of the Relationship Between Educational Achievements and Income

[1]Archana Ratnaparkhi, [2]Abhishek Mallav, [3]Abhijit Chitre, [4]Sandeep Kore

[1]Electronics and Telecommunication Dept Vishwakarma Institute of Information Technology Pune
archana.ratnaparakhi@viit.ac.in
[2]Electronics and Telecommunication Dept Vishwakarma Institute of Information Technology Pune
abhishek.22310511@viit.ac.in
[3]Electronics and Telecommunication Dept, Vishwakarma Institute of Information Technology
Pune,
abhijit.chitre@viit.ac.in
[4]Mechanical Engg Dept, Vishwakarma Institute of Information Technology
Pune,
sandeep.kore@viit.ac.in

Article History:

Received: *10-11-2024*

Revised:*17-12-2024*

Accepted:*04-01-2025*

Abstract: The correlation between income and education has been studied in many fields and literature. But, in socio-economic context, other factors like demographic aspects, work experience, industry, and job satisfaction are necessary. In this study, we created a synthetic dataset that contains educational levels, work experience, industries, sex, marital status, and work satisfaction. We performed supplementary data analysis to show how empirical data of different educational levels obtained increased income did increased income. The findings obtained suggest that education and work experience greatly affect income, which shows that higher earnings are linked to the level of schooling completed. In particular, a very weak negative relationship was established between influence and income, but this effect was noted among men. The change in the influence that was observed differed marginally by marital status. From this evidence, one can infer that any affirmative policy should focus on the subsidizing of barriers to accessibility of higher education because it is said to resolve income inequality gaps in the country.

Keywords: Gender Pay Gap, Two-Sample Z-Test, Linear Regression Model for Salary Prediction, Feature Reduction Techniques, Last Regression, and Employee Salaries.

I. INTRODUCTION

The research's connection to the sociological work per- formed on human capital explains why social and economic sciences have regarded education as a determinant of income so much.

Mincer (1974) established a significant connection when he noted that there was a positive linkage between the amount of schooling undertaken and the earning received. This link is driven by the rational expectation of employers of increasing productivity that comes with higher levels of education, which creates better employment opportunities. This understanding has affected the policy on how education is financed. The theory contends that the absence of education severely con- strains one's ability to earn, indicating the investment value of education. Mincer and others argue that, as a rule, more education leads to earning an

income, meaning that education provides a return on investment.

On the other hand, multiple interrelated factors have bearing on the education – income relationship.

According to Thropp and Wideronn (2020), "other main factors like work experience, industry and demographic factors makes the correlation with income complex because education greatly affects these factors." On the other hand, the findings of Coady and Dizioli suggest that the inequality deviates because some forms of education attribute lower barriers of access, and some forms increase it if the education of lower quality is provided.

It suggests that there are relations between education and earnings; however, in the wider context of sociology, a per- son's demographic, work history, industry, and even job morale count too.

It is suggested that researchers take on different novel approaches to study the factors mentioned earlier. Therefore, the objective is to create synthetic data that are as close as possible to the real data recorded during the distribution and relationships. This way of data synthesis will anonymize all the respondents while allowing controlled research to be done on the education, work experience, demographic data and income correlation. In this document, the above discussed method is applied to create a synthetic data set which includes educational qualification, years of work experience, industry, sex, marital status, and job satisfaction.

This research focuses on the degree to which novel artificial data is able to capture the existing patterns and relationships in education and income. Estimation of income is done through the application of the models on the generated data.

The factors regarding a more complete understanding of income are examined more thoroughly using linear regression and then clues to income are sought through PCA with the use of fewer variables.

These techniques has the ability to extract relevant informa- tion from extensive datasets by reducing them to their principal components to indicate the key sources of variance within the dataset.All the same, because this work aims at making linear regression analysis, it is important to evaluate the degree in which each element seeks to explain income. This fusion of techniques creates an understanding of synthetic data that is informed by the methods used in real world data empirical studies. In essence, it allows assessment of synthetic works with the existing body of literature. To put it very briefly, this study wishes to find out and examine how education and more variables affect income in terms of synthetic data. It would try to sustain the claimed methodology and increase the refinement in substantiating the importance of education, work experience, and other demographic variables on income. At this point, there already exists considerable intellectual debate about what is education and the scope of its economic implications. Studies are pieces of evidence that are developed through the synthesis of various data and attempts in drawing the real world population trends. Therefore, these results have been used to shape decisions directed toward optimizing issues such as schooling and the gap between earnings.

II. LITERATURE SURVEY

Researchers worldwide have consistently focused on the relationship between educational levels and wages earned. Incomes earned as a result of educational qualifications have always garnered great focus from researchers around the globe. This review of citations focuses on regression analysis, principal component analysis (PCA), and SES indices as the most important studies that have been done on this topic.

Jacob Mincer in his publication, Schooling, Experience, and Earnings, is considered the first author to research the concord between earning and education. In the 1950s, Mincer paid attention to the available data concerning the relationship between earning and education and created a model of human capital. This model, based on the relationship between educa- tion attended and productivity measured, is very significant. From this basic alone, we can deduce that there is high economic productivity the higher levels of education one attains. This analysis lays the groundwork for more in-depth investigations into the relationship between education and the economy [1].

Through the use of regression analysis on aggregated data collected from several Swedish municipalities, Thropp and Wider¨onn in their thesis, The Effect of Education on Income, presented employment patterns and the importance of educa- tion, work experience, and demographic factors in wage levels. Their results strongly illustrate that education is a major factor when predicting income, and

earnings tend to be more complicated because other factors like work experience and demographics also play a role. As Coady and Dizioli claim in their paper "Income Inequality and Education: A Review of the Evidence", there exists a cross-section of habitual studies on the consequences of schooling towards the economy gap within different nations. Their explanation is that education serves two purposes: it is capable of reducing income dif- ferences by providing chances for economic growth, but it also increases the difference in the case where education that is offered is not of good quality. This accounts for a fasci- nating feature of socio-economic inequalities across different countries. The social-economic status of a country can also be incorporated by analyzing household wealth accumulated through income and property. Keller's work focuses on the construction of socio-economic status indices based on the PCA approach and explains how they deal with this issue. Through PCA, Keller shows how to easily make SES indices that take into account education and income assimilation. This technique is particularly useful for the understanding of many other conditions that go together with earning a living aside from schooling. In the East African Evidence on Education, Productivity, and Inequality, Knight and Sabot look at the relationships between education, productivity, and income inequality in East Africa.

Their empirical analysis suggests that an expansion in education relates to an increase in productivity, which can help reduce the gaps in income distribution. This work provides useful evidence on how educational policies can influence economic processes in developing countries [5]. Spence's theory in "Job Market Signaling" argues that education in- creases a worker's possible productivity indirectly because it acts to a certain degree as a filter to the employers. This view adds another dimension in understanding the relation between schooling and income because it suggests that the qualifications obtained in school can significantly determine wage rates and employment positions offered [6]. The conse- quences of Yang and Qiu's study on the parental investment on education and its implications for income inequality in China are accentuated. Their findings indicate that parental aspirations and expectations significantly affect the educational success of the children and subsequently the child's earning capacity. This research illustrates that there is a family context which significantly determines the educational achievement which is beyond the linear relationship between income and education attainment [7]. Together, all these studies show that education impacts a person's income the most, although other factors such as a person's socio-economic status, work experience, demographic profile, and even the methods used in market signaling influence the effect of education on income. Understanding these interactions is necessary to devise educa- tional policies that aim to improve economic productivity for different groups.To achieve our objectives, we set out to create an actual pretend dataset, which was constructed with the goal of analyzing the relationship between income, education, and demographics.

III. METHODOLOGY

This set contains multiple attributes that showcase the person's level of education, work experience, age, industry, gender, marital status, and their level of job satisfaction. Each of these attributes sought to capture different elements that are most common related to the distribution of income. For example, "High School," "Associate," "Bachelor," "Master," and "PhD" would be used as categories to define "Education" along with base incomes and sets of actual data that depict the relationship between education and income. Experience, age, and the other demographic elements were generated within realistic ranges that reflected typical levels of sophistication of a given population. We also synthesized social class while taking other considerations into account. This ensures that this pretend dataset incorporates socioeconomic aspects that have been shown to affect income and conducts controlled analysis of the relationships between the variables within the dataset.

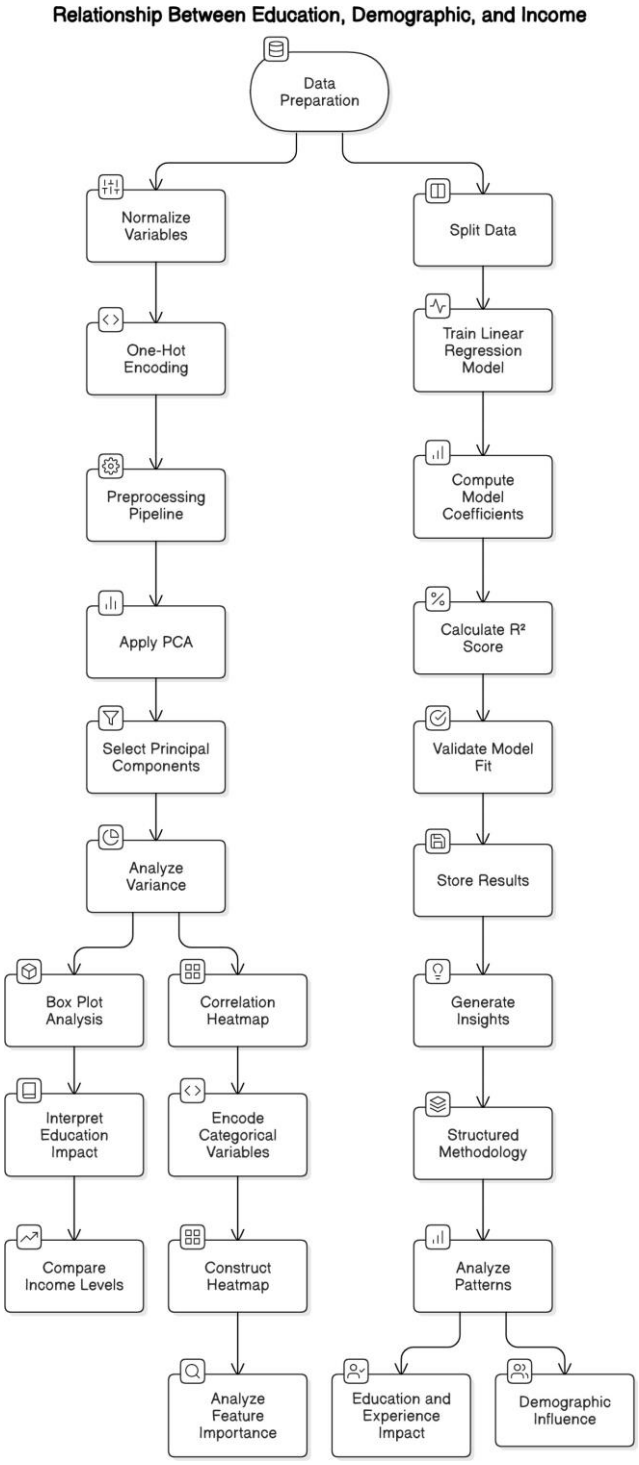

Fig. 1. Methodology

The dataset was designed to allow for statistical computa- tions to be performed.

Both categorical and numerical variables were used to define industry and education level, while the age and experience continous variables were normalized using Standard Scaler. One hot encoding was used to define the categorical variables of education and industry.

The extraction and compilation of information into the end product is a complex process, particularly regarding the use of PCA models or linear regression analysis on categorical data. All the objectives of the project were completed by em- ploying Column Transformer and Pipeline classes provided by scikitlearn. This approach makes it possible to simplify the processes of changing and analyzing data to

ensure that results are consistent and reproducible. These decisions around preprocessing are consistent with best practices of undertaking complex analyses of socio-economic data and were made doing research on the effects of education on income. The data set was subjected to a reduction of dimensionality while the few initial components that have a significant contribution of variation were selected through PCA.

Excluding the income parameter, the remaining feature was processed and submitted to the PCA stage with the aim of preserving four principal components, and extracting the fore- most primitive patterns in the data. Meeting this target is done by selecting components that strike a fine balance between dimension reduction and retention of additional information. As a result, we are able to understand the variance that each component elucidates.

Lastly, we may be able to spot the key characteristics that explain the difference in income. After isolating all aspects of ethnicity and sex, it is possible that education or industry are likely to exhibit greater impacts than the previous ones. This is very similar to the techniques that were used for constructing socio-economic indices like those built by Keller with the use of principal component analysis (PCA), which are meant to aggregate complex phenomena but still contain an adequate explanation. At this level, we added more data for analysis to show the hoped-for relationship between education and income, especially how income can be appreciably increased through proper empirical research and education at different levels. This set of analyses was done by box graphs.

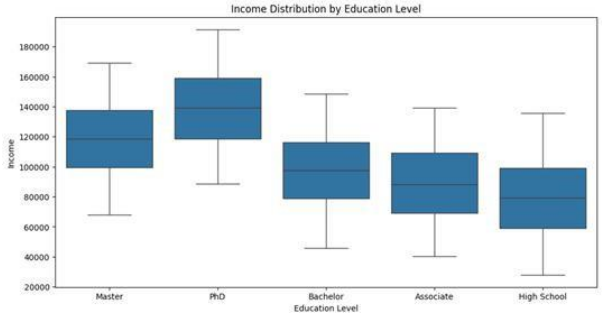

Fig. 2. Box Plot Analysis of Income by Education Level

Moreover, the heatmap developed to represent the correlation was used to demonstrate the effectiveness or ineffectiveness of income correlations with all variables in the data base. To grasp the basic nature of the relationship between income and other features, there was need to one-hot encode all categorical variables and make them part of the correlation matrix before building the heat map.This new approach to this type of analysis helped in understanding the relative importance of all aspects that identified factors associated with education, experience and other demographic variables.The qualitative effect of the different factors on income was captured using linear regression analysis.

Once the categorical variables were encoded, the data was divided into training and testing subsets. We employed a linear regression model to estimate income based on available features such as education, experience, industry, and other demographics. The model coefficients offered better insight on the critical determinants of income, especially education and experience. The R^2 score of the model was computed to check the coverage of the selected underlying income distribution. This is a figure of merit for the performance of the model against the test set. This will indicate how effective the selected features are in representing the actual income distribution. The strength of our method of constructing synthetic data is in the inference from our conclusions from the synthetic data to the established relationships of real studies. To have all the collected information to carry out the more sophisticated analysis, we also split part of the CSV files using PCA- transformation and regression analysis results since we knew we are going to need that information.

Therefore, we are working on a methodological approach to create, manipulate, and analyze synthetic income datasets that puts the patterns of education, experience, and demographics into sharper focus.

IV. RESULT

Box plot analysis brought out several important findings in the income distribution at different levels of education Fig2

- Education Level PhD: The median income is about $140,000.
- Education Level Master's: The median income is about $120,000.

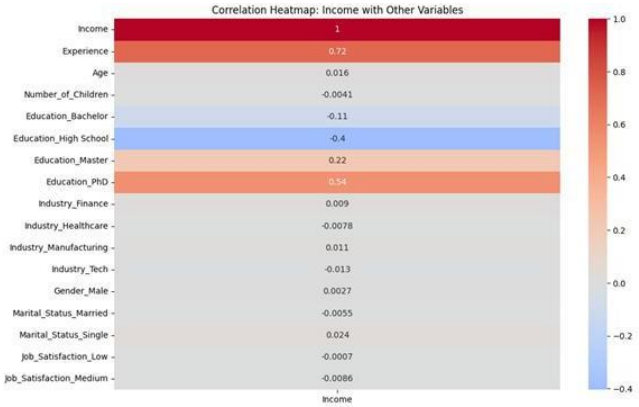

Fig. 3. Box Plot Analysis of Income by Education Level

- Education Level Bachelor's: The median income is about $100,000.
- Associate's Education Level: The median income is about $80,000.
- High School Education Level: The median income is approximately $60,000.

The interquartile range (IQR)represents the middle 50% of the information; that is, it varies between education levels-the distribution of incomes is different at each level of education. The whiskers of the box plots range until the minimum and maximum value within 1.5 times the IQR of the quartiles. From the above graphs you would find that income level distribution prevails in each educational qualification. Which is enhanced by the fact that the education level correlates with earning ability positively, wherein the earning ability tends to improve monetarily for a particular higher education qualification levels, whereby one can observe instantly that earnings depend upon educational qualification. The corre- lation heatmap Fig.3 continued to present further insights as follows with respect to how income varies with other demographics and professions:

1) Experience: Very strong positive relation with the in- come variable r = 0.72, thereby indicating that as experience rises, so does the amount of money earned.
2) Education Level, PhD: Shows a moderate positive rela- tion to the amount of money earned, i.e., r = 0.54, in that persons with doctoral levels of education tend to make more.
3) Master's Education Level: The association was weak positive (r = 0.22), meaning that it exerts minimal influence on income.
4) High School Education Level: There was moderate negative association at r = -0.4, which signifies that the lower the level of education, the lower the level of income.
5) Other Demographic Variables (Age, Number of Chil- dren, Industry, Gender): They had very weak positive and negative associations with minimal influence on income.

6) Marital Status and Job Satisfaction: Both of these factors correlated weakly with income, thus having a minimal influence on income inequality.

The statistical results confirm that education and work experience have strong relationships with income, meaning that academic achievements and work experience are related to higher earning potential. In contrast, the demographic factors tested did not have any notable relationships with income, thereby implying that they have very minimal influence on earning abilities.

A. Key Takeaways

1) More Higher Median Incomes at Each Level of Educa- tion: Advanced education is very highly correlated with higher median incomes, indicating that an investment in education can generate very high returns.
2) Experience as a Strong Predictor: This implies that experience remained the best predictor of income, and this fact is consistent with most agreement that those holding longer experiences tend to have higher incomes because of more skill and expertise accumulated by those experienced workers.
3) Education Level and Income Relationship: There is a well-defined relationship between the income and level of education. The highest-paid are those with PhDs, followed by Master's, then Bachelor's, and the lowest- paid are those with high school diplomas.
4) Industry Influence on Income: Income differences ap- pear to be influenced less by industry variables, and the industries of finance, health care, and manufacturing vary very little. For these industries, education and experience are more critical income determinants.
5) Job Satisfaction and Income: The relationship between job satisfaction and income is rather weak. Although higher income is a slight booster for job satisfaction, it is secondary.
6) Impact of Gender and Marital Status: The obtained results showed that gender and marital status barely influence the income while there was a very weakly negative correlation of the influence with the income, where this influence was recorded among males; the influence shows minor variations in marital statuses.

B. Model Accuracy

The analytical model did well, and the R^2 was 0.98; this implies that the model explains the variance of income in the synthetic data set to a great extent. On the other hand, the score is high, which increases the likelihood of overfitting, especially since the data used is synthetic.

V. FUTURE WORK

Because this is artificial data, the analysis should be repli- cated using real-life data to confirm these findings. We can better understand the dynamics of income by exploring non- linear models and interaction effects. Further, multicollinearity checks may be done perhaps through VIF analysis.

CONCLUSION

Education and experience play a vital role in the determi- nation of income level; better education is highly associated with income, and the degrees obtained at higher education levels, such as PhD and master's degrees, influence income greatly. Experience remains a determining factor; acquired professional knowledge and skill are of value. Industry type and job satisfaction did influence the outcome, but only very weakly, suggesting that income differences are largely an artifact of education and experience. Also, the fact that gender and marital status don't seem to have much of an effect may mean that income is becoming more evenly distributed in this fake data, but real-world data would be needed to confirm this. Implications for Stakeholders The implications of the results are that any policy initiative should focus more on making higher education generally accessible to bridge income disparity. The employer can also use the insights to structure compensation models, seriously considering educational quali- fications and years of experience as major factors towards fair and competitive wage policies.

ACKNOWLEDGMENT

We would like to express our gratitude to all individuals and organizations that contributed to this research. Additionally, we acknowledge the data sources and tools used, which played a critical role in facilitating our analysis. This study would not have been possible without the collaboration and insights shared by everyone involved.

REFERENCES

[1] J. Mincer, Schooling, Experience, and Earnings. National Bureau of Economic Research, 1974.
[2] C. Thropp and M. Wideronn,¨ "The Effect of Education on Income," DiVA portal, 2020.
[3] D. Coady and A. Dizioli, "Income Inequality and Education: A Re- view of the Evidence," International Monetary Fund Working Paper WP/18/65, 2018.
[4] A. Keller, "Constructing socio-economic status indices: How to use principal components analysis," Health Policy, vol. 21, no. 6, pp. 459–469, 2008.
[5] J. Knight and R. Sabot, Education, Productivity and Inequality: The East African Evidence. Oxford University Press, 1983.
[6] A. M. Spence, "Job Market Signaling," Quarterly Journal of Economics, vol. 87, no. 3, pp. 355–374, 1973.
[7] H. Yang and L. D. Qiu, "Parental Investment in Education: Evi- dence from China," Journal of Economic Studies, vol. 43, no. 5, pp. 1049–1060, 2016.
[8] Blau, F. D., & Kahn, L. M. (2016). The Gender Wage Gap. Journal of Economic Perspectives, 30(2), 3–28.