MULTIPLE LINEAR REGRESSION AND K-MEAN CLUSTERING OF POST-COVID ON HOUSEHOLD INCOME IN MALAYSIA

Muhammad Akmal Danial bin Mohd Sabri, Wan Nurshazelin binti Wan Shahidan
Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA Perlis Branch, Malaysia

e-mail: akmaldanialsabri@gmail.com, shazelin804@uitm.edu.my

Abstract

This study explores the socio-economic factors that influence household income in Malaysia after the COVID-19 pandemic by using a combination of Multiple Linear Regression and K-Means clustering. Data were taken from the 2022 Household Income and Expenditure Survey provided by the Department of Statistics Malaysia. The regression analysis found that household size, age, gender, and education level significantly affect income. The model achieved an adjusted R-squared value of 0.499 and an F-value of 35.727, showing a moderately strong fit. These significant variables were then used in the clustering process, which grouped households into three distinct segments. The first group includes older, low-income, and less-educated households. The second group represents middle-income households with moderate education. The third group consists of younger, highly educated, and higher-income households. This hybrid approach gave more detailed insights than using regression alone and supports more targeted and effective policy planning. The findings are consistent with earlier research by Yee et al. in 2023, which highlighted the importance of clustering in understanding income differences in Malaysia.

***Keywords**: Multiple Linear Regression, K-Means clustering, household income, Malaysia*

1. Introduction

This study aims to examine the demographic and socio-economic factors that influence household income in Malaysia's post-pandemic landscape. Drawing on 2022 survey data from the Department of Statistics Malaysia, the research applies a hybrid method combining Multiple Linear Regression (MLR) and K-Means clustering (Yee et al., 2023). The MLR model identifies statistically significant variables that explain income variation, while the clustering technique segments households into distinct socio-economic groups for clearer profiling. This approach enhances understanding of income inequality beyond traditional classifications like B40, M40, and T20. By segmenting households based on shared characteristics, the study builds on work by Yee et al. (2023), offering a more data-driven framework for targeted interventions and policy formulation.

2. Methodology

This study utilizes a quantitative approach based on secondary data from the 2022 Household Income and Expenditure Survey (HIES). From the available dataset, nine key variables were selected, including income, household size, education level, gender, age, marital status, employment type, strata (urban/rural) and state. First, a Multiple Linear Regression (MLR) model was developed to determine which variables significantly affect income. Model fit was evaluated using adjusted R-squared and F-statistics (Chen et al., 2023). Next, K-Means clustering was applied using only the significant predictors identified in the regression model. The optimal number of clusters was determined through the Within-Cluster Sum of Squares (WCSS) method (Saji, 2024). The resulting clusters were then analyzed to understand demographic and socio-economic differences, providing insights into hidden household segments. Visualizations such as boxplots and bar charts were used to aid interpretation. All statistical analysis and visualization were performed using SPSS software.

3. Results / Findings

The findings of this study highlight key demographic and socio-economic factors influencing household income in Malaysia. Using Multiple Linear Regression (MLR), variables such as education level, gender, age of the head of household, and household size were found to significantly affect income. Households with higher education, male household heads, older age, and larger size were generally associated with higher income levels. The MLR model produced an adjusted R^2 of 0.499, indicating a moderate fit. To further explore income differences, K-Means clustering was applied using the significant variables identified from the regression model. This process resulted in three distinct household segments. Cluster 1 consisted of older, less-educated, self-employed individuals with the lowest income. Cluster 2 included younger, moderately educated rural families with average income. Cluster 3 represented highly educated, urban professionals with the highest income. These results reveal clear socio-economic distinctions among Malaysian households and demonstrate the value of combining regression and clustering techniques for deeper insights into income inequality.

4. Novelty of Research / Product / Project

This study introduces a hybrid analytical approach by integrating Multiple Linear Regression (MLR) with K-Means clustering to examine household income disparities in Malaysia after the COVID-19 pandemic. While previous research has analyzed income determinants using traditional regression methods, this study goes a step further by segmenting the population into distinct socio-economic clusters before reapplying regression models. This two-step method allows for the identification of group-specific income patterns that would be overlooked in a general model (Yee et al., 2023). By using 2022 data from the Department of Statistics Malaysia (DOSM), the study reflects the most recent post-pandemic economic conditions, offering up-to-date insights into the factors that influence household income. The clustering process helped reveal hidden intra-group differences within standard income classifications like B40, M40, and T20, making the analysis more precise and meaningful. This innovative approach not only improves the understanding of income inequality but also enhances the potential for developing targeted and effective policy interventions tailored to the needs of different household types in Malaysia.

5. Conclusion or Future Work

This study concludes that strata, education level, gender, age, and household size significantly influence household income in Malaysia. The hybrid method combining MLR and K-means clustering provided deeper insights into socio-economic differences. Future research should include more variables and updated datasets to improve model relevance and support better-targeted policymaking.

REFERENCE LIST

Chen, S., Fan, R., & Lin, Y. (2023). Research on Income and Expenditure Based on Multiple Linear Regression: A Case of Filipinos (pp. 117–125). https://doi.org/10.2991/978-94-6463-030-5_13

Saji, B. (2024, December 4). Elbow method for optimal cluster number in K-Means. Analytics Vidhya. https://www.analyticsvidhya.com/blog/2021/01/in-depth-intuition-of-k-means-clustering-algorithm-in-machine-learning/

Yee, G. P., Rusiman, M. S., Ismail, S., Suparman, Hamzah, F. M., & Shafi, M. A. (2023). K-means clustering analysis and multiple linear regression model on household income in Malaysia. IAES International Journal of Artificial Intelligence, 12(2), 731–738. https://doi.org/10.11591/ijai.v12.i2.pp731-738