The Effects of Household Income and Food Expenditure on the Malaysian Household Food Security Index Using the Regularized 2SLS Regression

Nurul Najiha Abdul Rahman Adrin
Department of Mathematics and Statistics, Faculty of Applied Science and Technology, University Tun Hussein Onn Malaysia, Campus Pagoh, Muar, Johor, Malaysia
najihaadrin@gmail.com

Khuneswari Gopal Pillay
Department of Mathematics and Statistics, Faculty of Applied Science and Technology, University Tun Hussein Onn Malaysia, Campus Pagoh, Muar, Johor, Malaysia
khuneswari@uthm.edu.my (corresponding author)

Received: 9 February 2026 | Revised: 14 May 2026 | Accepted: 20 May 2026

 DOI: https://doi.org/10.48084/etasr.18054

ABSTRACT

Household food security remains a critical concern in many Asian countries, including Malaysia, where disparities in income distribution and rising living costs continue to affect food access and stability. While existing studies primarily emphasize food security at the national level, empirical evidence at the household level remains limited. This study examines the effects of household income and food expenditure on the Household Food Security Index (HFSI) in Malaysia using a regularized Two-Stage Least Squares (2SLS) regression framework. By integrating Least Absolute Shrinkage and Selection Operator (LASSO) and Elastic Net techniques into the 2SLS estimation, the proposed approach addresses endogeneity and multicollinearity while enabling robust variable selection. Cross-sectional household data obtained from the Department of Statistics Malaysia (DOSM) for 2022 are utilized. The results indicate that the 2SLS-LASSO specification outperforms the Elastic Net model in terms of Akaike Information Criterion (AIC), Corrected Akaike Information Criterion (AICc), and Bayesian Information Criterion (BIC), providing a more parsimonious and interpretable model. Empirical findings reveal that higher household income, educational attainment, and employment stability are positively associated with food security, whereas income inequality (measured by the Gini coefficient) exerts a significant negative effect. Larger household size is associated with reduced per capita food security, reflecting resource constraints within households. This study contributes to the literature in two key ways: (i) by introducing a regularized 2SLS framework for analyzing household food security, and (ii) by proposing a structured and replicable formulation of the HFSI. The findings provide policy-relevant insights for addressing income inequality and improving targeted interventions for vulnerable households in Malaysia.

Keywords-household food security; household income; household food expenditure; Two-Stage Least Squares (2SLS); regularization regression; LASSO; Elastic Net

I. INTRODUCTION

The concept of food security was introduced in the 1970s to ensure a sufficient food supply for all people, and the literature records over 200 definitions of food security in published papers [1]. The Food and Agriculture Organization (FAO) defines food security as a situation in which all people, at all times, have physical, social, and economic access to sufficient, safe, and nutritious food that meets their dietary needs and preferences for a healthy and active life [2]. As food security spans multiple dimensions, it can be defined by four key components: access, availability, stability, and utilization. The second United Nations Sustainable Development Goal aims to end hunger and malnutrition by 2030; however, recent data suggest that this goal remains unmet, as approximately 2.3 billion people worldwide were moderately or severely food insecure in 2024 [3]. This raises concerns for most regions, including Asian countries. Malaysia, too, aims to address food security at the grassroots level to build a more resilient household economy.

Measuring food security is essential for understanding the dynamics of hunger and malnutrition across populations. Hence, various indicators are introduced that incorporate different dimensions of food security. For example, the Global Food Security Index (GFSI) is a widely used indicator worldwide. In 2013, FAO developed the Food Insecurity Experience Scale (FIES) through the Voice of the Hungry (VOH) project to provide up-to-date information about measuring food security. Unfortunately, there is no single indicator that captures all dimensions simultaneously [4]. Moreover, most indicators focus only on global and national scales rather than on microeconomic indicators, such as household or individual-level indicators [5].

To effectively represent the household level, this study highlights sociodemographic and socioeconomic variables, which are commonly used as key indicators in econometric research. Socioeconomic variables typically include household total income, total expenditure, education level, and occupation type. In contrast, sociodemographic variables capture population diversity through attributes such as age, gender, ethnicity, marital status, household size, and geographic location. These characteristics provide a broad understanding of societal patterns and their implications for household outcomes [6].

Regression is a standard econometric method that effectively captures the relationships between independent and dependent variables. Over time, numerous regression techniques have been established to address various data types, assumptions, and objectives. Classical regression, often estimated by the Ordinary Least Squares (OLS) method, is the basis of regression estimation. However, OLS has limitations, particularly when some explanatory variables are endogenous or correlated with the error term, leading to biased estimates. To address this issue, the Two-Stage Least Squares (2SLS) method uses instrumental variables to provide consistent estimates [7]. However, standard 2SLS can still face challenges when dealing with many predictors, multicollinearity, or model complexity. Therefore, regularized 2SLS approaches, such as 2SLS-Least Absolute Shrinkage and Selection Operator (LASSO) and 2SLS-Elastic Net, are introduced. These methods combine the benefits of instrumental variables estimation with regularization techniques, improving model interpretability, reducing overfitting, and identifying the most relevant predictors, ultimately leading to a more robust and efficient model for analyzing predictor outcomes [8].

There is limited literature that focuses on developing regularized 2SLS regression. This method is useful when instrumental variable methods need to handle high-dimensional data or dynamic endogeneity. Authors in [9] employed Instrumental LASSO (IV-LASSO) to address the simultaneity between female unemployment and fertility decisions. By carefully selecting instruments, such as the male unemployment variable, the model isolates female unemployment from other confounding factors. This reduces bias that can arise from including too many controls or lagged variables. As a result, higher female unemployment leads to lower fertility, underscoring that economic realities often outweigh political rhetoric in shaping family planning. The study demonstrates that regularized 2SLS strengthens standard 2SLS, making estimates more reliable in complex settings.

In a different context, post-LASSO Instrumental Variables (IV) was implemented to handle a large number of possible instruments. Authors in [10] employed a two-stage process that first identifies the most critical control variables and then selects the strongest instruments to address high-dimensional data across 23 European countries. This method enables precise estimation of residential natural gas demand by utilizing many potential instruments, including weather shocks across countries. However, machine-learning-based selection can sometimes retain invalid instruments, which highlights the importance of traditional validation methods.

Existing studies highlight the effectiveness of regularized models in addressing endogeneity and multicollinearity. However, their application in food security research remains limited, indicating a literature gap. By applying a regularized 2SLS framework to household food security data, this study addresses endogeneity between income, food expenditure, and food security outcomes while selecting the most relevant predictors. This study contributes to the literature in three key ways. First, it introduces a regularized 2SLS framework for household food security analysis, enabling the simultaneous treatment of endogeneity and multicollinearity within a unified framework. Second, it proposes a structured, replicable formulation of the Household Food Security Index (HFSI) for consistent measurement at the household level. Third, it provides recent empirical evidence from Malaysia, offering policy-relevant insights into the role of income inequality and socioeconomic characteristics in shaping household food security.

II. METHODOLOGY

A. Data Description

Key variables include household sociodemographic factors, household socioeconomic factors, and the food security indicator. All these variables were obtained through the Department of Statistics Malaysia (DOSM) and cover the year 2022. The dataset was collected through the Household Income and Basic Amenities Survey (HIS & BA) and the Household Expenditure Survey (HES), which are conducted every five years to provide data on household income, expenditure, and basic amenities in Malaysia. Ultimately, the food security indicator was measured using the HFSI, which accurately captures the availability, access, and stability dimensions of household food security using the selected variables. A description of each variable is presented in Table I.

B. Household Food Security Index (HFSI) Formation

Many indicators have been introduced to measure food security, each highlighting different dimensions as its core focus. The HFSI is chosen in this study because it reflects a high level of household involvement in measuring food security. The HFSI applies a composite index methodology at the household level, producing a structurally unique measure for each household, in contrast to the Proteus composite index developed by authors in [11] to assess national food security. Variables are selected to capture all relevant dimensions of

food security, as shown in Table II. Steps 4 and 5 introduce new procedures for estimating the HFSI values. The weights are assigned based on each indicator's relative importance in capturing the dimensions of food security (availability, access, and stability).

TABLE I. VARIABLE DESCRIPTION

Variable	Symbol	Description
Total income	Y_1	Monthly household income
Total food expenditure	Y_2	Monthly household food expenditure
HFSI	Y_3	HFSI for each household
Gender	Z_1	Gender (1 = male, 0 = female)
Age	Z_2	Age of household head
Employment status	Z_3	(1 = employed, 0 = unemployed)
Gini coefficient	Z_4	Gini coefficient measuring income inequality
Total expenditure	X_1	Monthly household total expenditure
Household size	X_2	Number of people in a household
Ethnicity	X_3	Ethnicity of household head
Marital status	X_4	Marital status of household head (1 = married, 0 = not married)
State	X_5	State of the household
Strata	X_6	Strata (1 = urban, 0 = rural)
Citizenship	X_7	Nationality of household head
Education	X_8	Education status of household head
Occupation	X_9	Occupation of household head
Poverty rate	X_{10}	Proportion of households with monthly income below the Poverty Line Income (PLI)

TABLE II. HFSI FORMULATION

Algorithm 1: HFSI computation flow	
Input:	Overall finalized dataset
Output:	Computed values of the HFSI variable
Step 1:	Determine the variables to include in the HFSI.
Step 2:	Apply minimum-maximum normalization to each variable to avoid disproportionate scaling across variables. Positive normalization is calculated as x_{norm}, whereas reverse normalization is applied for a negatively correlated variable as x_{rev}. $X_{norm} = \frac{x - min(x)}{max(x) - min(x)}$, where $0 < X_{norm} < 1$ $X_{rev} = \frac{max(x) - x}{max(x) - min(x)}$, where $0 < X_{rev} < 1$
Step 3:	Apply weights to each normalized variable.
Step 4:	(Proposed) The HFSI for each household is computed as a weighted composite index: $\text{HFSI} = \frac{1}{n}\sum wX_{ij}$ where n = number of variables; w = weight; X_{ij} =normalized value for each household.
Step 5:	(Proposed) Interpret HFSI as follows: **Interpretation** / **HFSI value**: Severe: $0 < \text{HFSI} \leq 0.25$; Moderate: $0.25 < \text{HFSI} \leq 0.5$; Mild: $0.5 < \text{HFSI} \leq 0.75$; Secure: $0.75 < \text{HFSI} < 1$
End	HFSI values

C. Data Preprocessing

The analysis begins with a data preprocessing stage [12]. Outliers are observations that lie substantially farther from the majority of the data and can distort parameter estimation. In large datasets, the performance of regularization methods can be adversely affected by the presence of outliers, making their identification and treatment an important preprocessing step [13]. Accordingly, this study employs adjusted boxplots, modified Z-scores based on the Median Absolute Deviation (MAD), percentiles, and the log-Interquartile Range (log-IQR) to identify outliers. Adjusted boxplots account for skewness through robust measures of central tendency and dispersion. MAD is particularly effective for skewed data distributions, such as those commonly observed in household income and expenditure data [14]. In addition, log-IQR stabilizes variance and reduces the influence of extreme values, whereas percentile-based thresholds focus on observations located in the tails of the distribution [15, 16].

D. Pre-Model Estimation

1) Correlation

Correlation determines the relationship between two variables. Pearson's correlation coefficient is used to explore the correlation between continuous variables, and Spearman's rank correlation coefficient is used for categorical variables. Spearman's rank correlation is applied because it can handle various types of data, including ordinal, interval, and ratio data, making it suitable for Sustainable Development Goals (SDG) indicator analysis [17].

2) Endogeneity

Endogeneity occurs when one or more explanatory variables are correlated with the error term in a regression model, $Cov\ (X_1, \varepsilon) \neq 0$, resulting in biased and inconsistent estimates [18]. To address endogeneity, the Durbin-Wu-Hausman test is conducted. A significant result ($p < 0.05$) indicates endogeneity, justifying the use of instrumental variables in the proposed model for consistent estimation. Following that, variables Y_1 and Y_2 are regressed on Z_i in the first stage.

3) Multicollinearity

Multicollinearity occurs in regression when predictors are highly correlated with each other. The Generalized Variance Inflation Factor (GVIF) proposed by authors in [19] is chosen because the dataset contains categorical variables. If the GVIF exceeds 5, it indicates very high collinearity among the variables. To consider the degrees of freedom associated with categorical variables, the GVIF is often adjusted by taking the $1/(2 \times df)$, allowing for a more meaningful comparison across variables.

4) Heteroscedasticity

Heteroscedasticity occurs when the variance of the error terms is not constant across observations. Authors in [20] describe it as variability among subpopulations, measured by variance or other dispersion metrics. The Breusch-Pagan test, $BP = nR^2$ is utilized, where BP follows a chi-square distribution with k degrees of freedom. A significant p-value indicates the presence of heteroscedasticity.

E. Two-Stage Least Squares (2SLS) Regression

The presence of endogenous variables necessitates the use of instrumental variables in a 2SLS regression. The relevance

and validity of instrumental variables are tested with the Kleibergen-Paap (KP) rank F-test and the Hansen J test, respectively. The 2SLS model is divided into two stages. In the first stage, two endogenous variables are regressed on the set of instruments, whereas in the second stage the fitted values are used to estimate the HFSI outcome. Models A and B represent the first-stage equations of the 2SLS regression, whereas Model C represents the second-stage equation. Variables that exhibit multicollinearity are excluded from the model specification.

$$\text{Model A: } Y_1 = \alpha_i + \beta_2 Z_2 + \beta_3 Z_3 + \gamma_i X_i + \varepsilon_i \quad (1)$$

$$\text{Model B: } Y_2 = \theta_i + \omega_1 Z_1 + \omega_2 Z_2 + \varphi_i X_i + \eta_i \quad (2)$$

$$\text{Model C: } Y_3 = \alpha_i + \beta_1 \hat{Y}_1 + \omega_1 \hat{Y}_2 + \gamma_i X_i + \delta_i \quad (3)$$

F. Regularization Regression

Regularization techniques have been applied across diverse domains, including finance, healthcare, and forecasting, to identify significant factors that influence model predictors.

1) Least Absolute Shrinkage and Selection Operator (LASSO)

LASSO regression employs an L1 regularization penalty to shrink coefficient estimates and perform variable selection simultaneously [21]. Specifically, LASSO minimizes the following penalized least squares objective function:

$$min_\beta \left(\sum_{i=1}^{n} (y_i - X_i^T \beta)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right) \quad (4)$$

where β_j is the regression coefficient of the j-th predictor, and λ is the tuning parameter that controls the strength of the penalty.

2) Elastic Net

Elastic Net regression extends LASSO by combining L1 and L2 regularization penalties, balancing variable selection and coefficient stabilization [22]. The objective function of Elastic Net is represented as:

$$min_\beta \left(\sum_{i=1}^{n} (y_i - X_i^T \beta)^2 + \lambda \left[\sum_{j=1}^{p} |\beta_j| + \sum_{j=1}^{p} \beta_j^2 \right] \right) \quad (5)$$

G. Cross-Validation

This study employs cross-validation in two stages within the proposed framework. First, the finalized dataset is split into training and testing datasets at an 80:20 ratio. The second stage applies k-fold cross-validation on the training dataset during regularization. A k value of 10 is used to determine the optimal λ value. The selected λ is then used to estimate the final models. The research framework is visualized in Figure 1.

III. RESULTS AND DISCUSSION

A. Data Preprocessing and Pre-Model Estimation

Outliers in household income often reflect genuine economic conditions rather than data entry errors. Therefore, most extreme values are retained, as they represent actual high-income households [23]. However, household incomes exceeding RM100,000 are considered excessively large and are removed, as the substantial gap between these values and the rest of the sample could disproportionately influence the analysis. The same procedure is applied to total household expenditure using the corresponding threshold. As a result, the models produce more stable and consistent estimates that better reflect the underlying relationships among variables. Figure 2 illustrates the relationships between variables. Consequently, model construction for the first-stage estimation is based on the relationships among total income, food expenditure, and other explanatory variables.

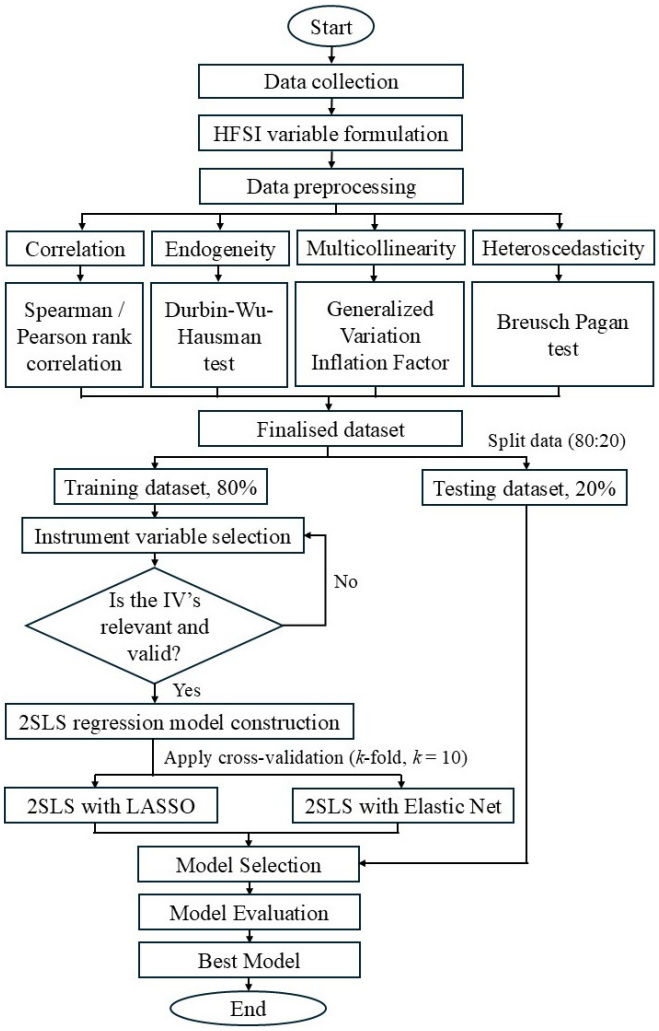

Fig. 1. Research framework.

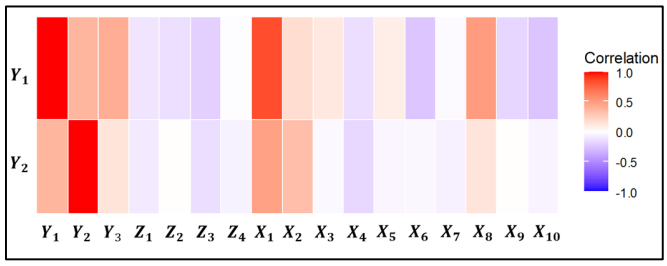

Fig. 2. Correlation matrix heatmap.

Table III shows evidence of endogeneity in both equations of the first-stage model, as the Durbin-Wu-Hausman test p-values are less than 0.05.

TABLE III. DURBIN-WU-HAUSMAN TEST RESULTS

Model	df1	df2	Statistic	p-value
Model A	1	13,679	19,259.814	< 0.05
Model B	1	13,707	6,509.96	< 0.05

Variables exhibiting multicollinearity are removed in Model C, as illustrated in Figure 3. This ensures consistency with regularization-based regression, which automatically addresses multicollinearity within the model [24].

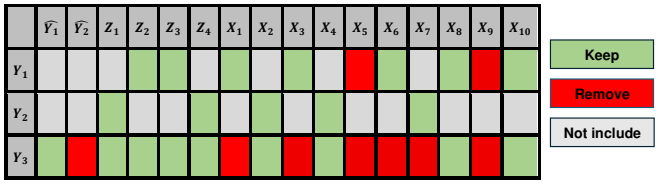

Fig. 3. Variable removal due to multicollinearity.

Table IV indicates heteroscedasticity in the model, as the Breusch-Pagan p-value is below the 0.05 significance level.

TABLE IV. BREUSCH-PAGAN TEST RESULTS

Model	Statistic	df	p-value
Model A	377.95	32	< 0.05
Model B	143.89	4	< 0.05

B. Model Interpretation

1) Two-Stage Least Squares (2SLS) Regression

Age and employment are selected as instrumental variables for Model A, whereas gender and the Gini coefficient are chosen for Model B. Both tests indicate the instrumental variables are valid and relevant. As tabulated in Tables V and VI, the KP rank test statistics exceed 10, and the p-values of the Hansen J test are not statistically significant.

TABLE V. DIAGNOSTIC TESTS FOR RELEVANCE

Test	Model A		Model B	
	Statistic	p-value	Statistic	p-value
KP rank F-test	47.961	$< 2\times10^{-16}$	30.27	$< 7.67\times10^{-14}$

TABLE VI. DIAGNOSTIC TESTS FOR OVERIDENTIFICATION

Model	Model C	
	Statistic	p-value
Hansen J test	7.442	0.242

2) Regularization Regression

Table VII presents the lambda (λ) values obtained through k-fold cross-validation for both stages of model building.

TABLE VII. LAMBDA (λ) VALUES

Model	Model A	Model B	Model C
2SLS-LASSO	7.1360	0.5969	0.0001
2SLS-Elastic Net	20.7062	1.0877	0.0002

The Cross-Validation Mean-Squared Error (CV-MSE) plots illustrate the performance of the regularization models across a range of λ values at each stage. The curves in Figures 4–6 decrease as the penalty is reduced, after which improvements become marginal.

This indicates that cross-validated performance improves steadily and then stabilizes, suggesting a well-defined optimal tuning region. The minimum MSE is reached without a sharp turning point, implying that model performance remains stable across a range of penalty values. Moreover, the curves are relatively smooth, indicating that model performance is less sensitive to tuning choices.

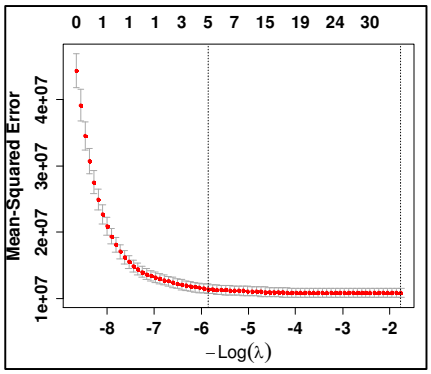

Fig. 4. CV-MSE plot for Model A LASSO.

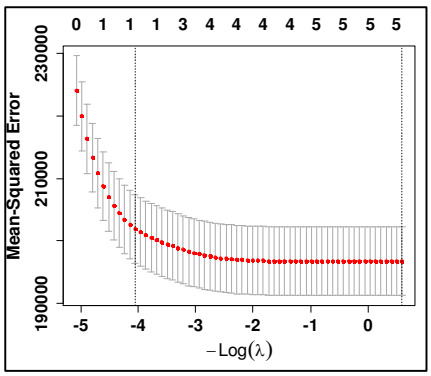

Fig. 5. CV-MSE plot for Model B LASSO.

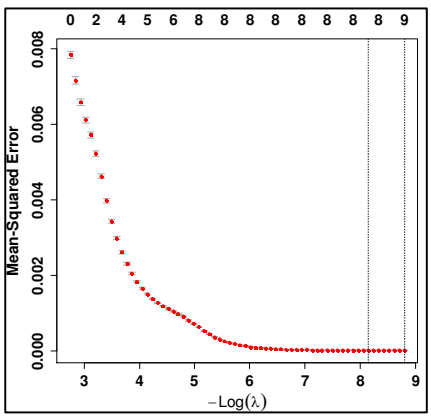

Fig. 6. CV-MSE plot for Model C LASSO.

The number of variables selected in Model A LASSO ranges from 5 to 30, whereas in Model B LASSO it ranges from 1 to 5. Meanwhile, Model C LASSO retains only 8 to 9 variables. These differences arise due to variation in the regularization strength across models.

Figures 7–9 show a similar pattern, except that Model A Elastic Net spans a range from 5 to 31. This is because Elastic Net combines L1 and L2 regularization, enabling it to capture a broader set of predictors in the first stage, reflecting its ability to balance sparsity and stability in the presence of multicollinearity.

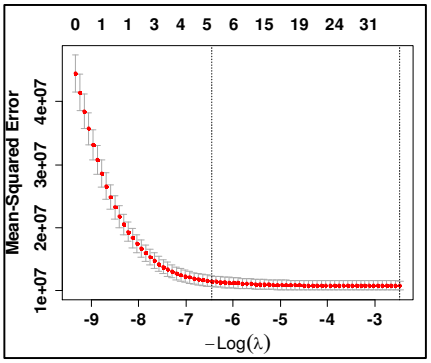

Fig. 7. CV-MSE plot for Model A Elastic Net.

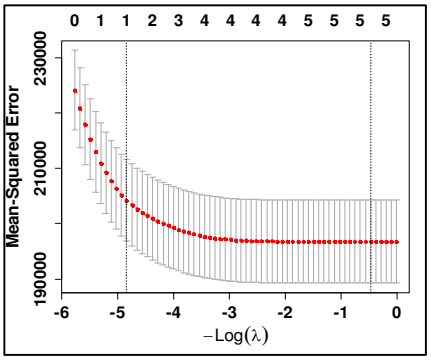

Fig. 8. CV-MSE plot for Model B Elastic Net.

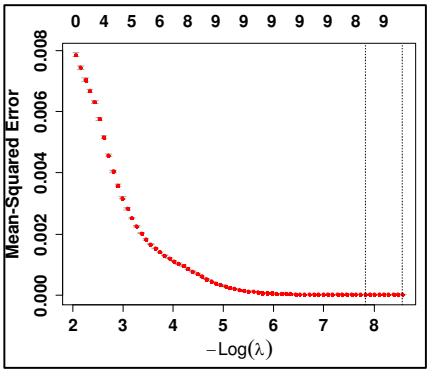

Fig. 9. CV-MSE plot for Model C Elastic Net.

Figures 10 and 11 illustrate the coefficient estimates from regularized Model A. Key predictors, such as state and education, consistently show substantial effects, underscoring their relevance to household income. Households residing in Labuan and Kuala Lumpur, as well as those headed by individuals with higher educational attainment, tend to report higher income levels.

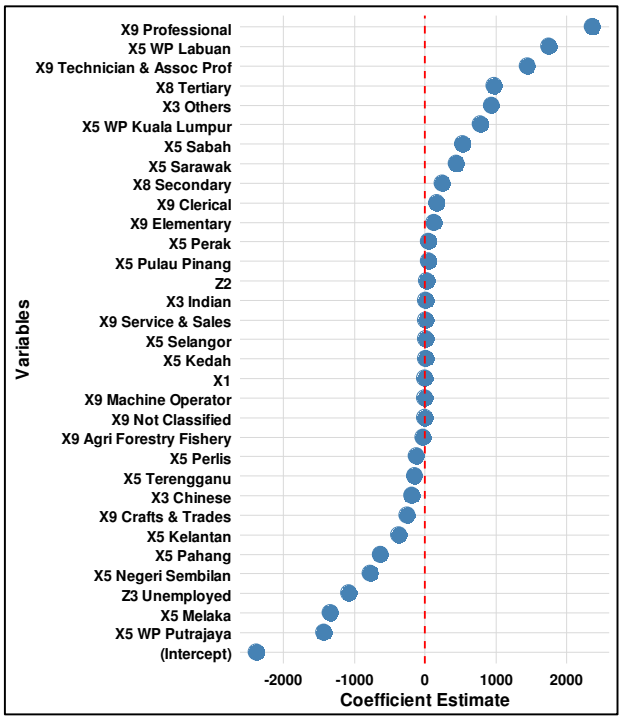

Fig. 10. Coefficient plot for Model A LASSO.

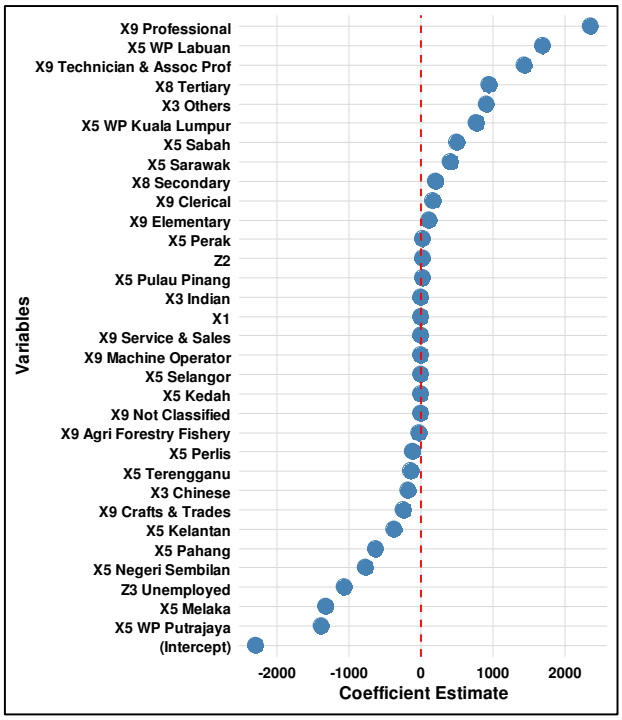

Fig. 11. Coefficient plot for Model A Elastic Net.

In Figures 12 and 13, household food expenditure increases significantly with age, which may be attributed to the accumulation of resources and responsibilities over time. However, most retained variables show negative associations, with the Gini coefficient exhibiting a substantial adverse effect that increases as λ increases.

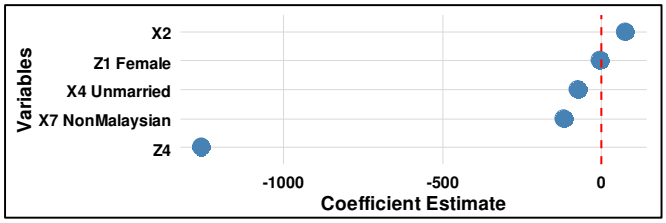

Fig. 12. Coefficient plot for Model B LASSO.

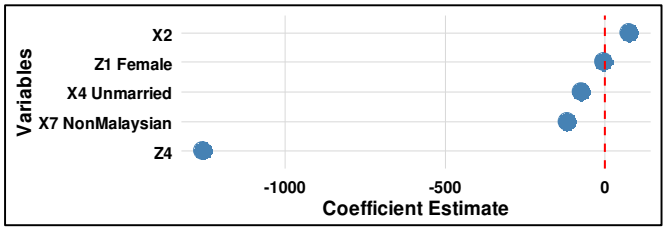

Fig. 13. Coefficient plot for Model B Elastic Net.

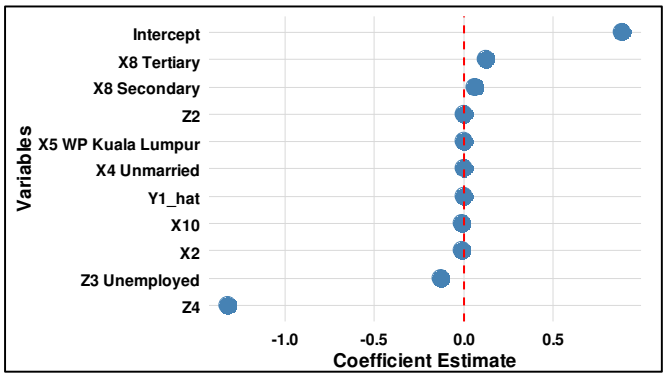

Fig. 14. Coefficient plot for Model C LASSO.

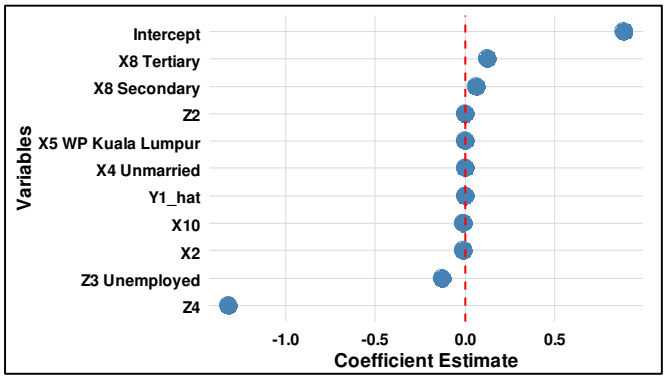

Fig. 15. Coefficient plot for Model C Elastic Net.

Figures 14 and 15 present the coefficients retained in Model C. The Gini coefficient, household unemployment, and tertiary education are consistently selected as key predictors of the HFSI. The positive coefficient for education suggests that households led by more educated individuals tend to experience greater food security, likely due to higher earnings and better resource management. In addition, other predictors have modest impacts, but the Gini coefficient remains dominant, highlighting its central role in explaining household food security.

C. Model Selection and Accuracy Checking

Table VIII compares the performance and accuracy of three models: Model C, Model C LASSO, and Model C Elastic Net. Model C LASSO outperforms the other models, achieving lower information criteria values (Akaike Information Criterion (AIC) = –39,489.38, Corrected Akaike Information Criterion (AICc) = –39,489.30, and Bayesian Information Criterion (BIC) = –39,421.84) than Model C Elastic Net (AIC = –39,474.22, AICc = –39,474.14, and BIC = –39,406.68), consistent with the findings in [17].

TABLE VIII. MODEL SELECTION AND ACCURACY

Metric	Model C	Model C LASSO	Model C Elastic Net
AIC	–39,411.94	–39,489.38	–39,474.22
AICc	–39,411.87	–39,489.30	–39,474.14
BIC	–39,344.40	–39,421.84	–39,406.68
MSPE	1.0129×10^{-5}	9.9026×10^{-6}	9.9465×10^{-6}
RMSE	3.1826×10^{-3}	3.1468×10^{-3}	3.1538×10^{-3}
MAE	1.7952×10^{-3}	1.7556×10^{-3}	1.7669×10^{-3}
R^2	0.9987	0.9988	0.9987
Adjusted R^2	0.9987	0.9987	0.9987

MAE: Mean Absolute Error.

The advantage of LASSO lies in its ability to perform strong variable selection by shrinking less important predictors to zero and retaining only the most influential variables. In contrast, Elastic Net combines LASSO and Ridge penalties, leading to a larger set of retained variables and slightly higher information criteria values. Although predictive errors are very similar across the two models, LASSO offers a better balance between simplicity and predictive performance, making it particularly useful for interpretation and identification of key variables. Moreover, Model C LASSO records the lowest Mean Squared Prediction Error (MSPE) values compared to Model C Elastic Net and Model C.

The Root Mean Squared Error (RMSE) values are very similar, with Model C at 3.1826×10^{-3}, while LASSO and Elastic Net report 3.1468×10^{-3} and 3.1538×10^{-3}, respectively. All three models exhibit exceptionally high coefficient of determination (R^2) and adjusted R^2 values, indicating a strong explanatory capacity for variations in the HFSI. These high R^2 values can be partly attributed to the construction of the HFSI as a composite index derived from variables closely related to income, expenditure, and socioeconomic characteristics. Consequently, a degree of inherent correlation between the predictors and the dependent variable is expected. To mitigate concerns of overfitting, cross-validation and out-of-sample testing were employed, confirming that the models maintain predictive stability. Although differences in predictive error are minimal, the

Model C LASSO specification offers a more parsimonious representation by retaining only the most influential predictors. This balance between model simplicity and predictive performance makes it a preferable choice, particularly in contexts where interpretability and variable selection are essential.

D. Best Model Interpretation

The Model C LASSO specification emerges as the best-performing model and is represented as:

$$\widehat{Y_3} = 0.8829 + 0.00000081555\ \widehat{Y_1} - 0.1244\ Z_{32} - 0.0063727\ X_{10} + 0.001473\ Z_2 + 0.00024291\ X_{514} + 0.0614\ X_{82} + 0.124\ X_{83} - 0.006942\ X_2 + 0.000017604\ X_{42} - 1.3133\ Z_4 \quad (6)$$

Income inequality, measured by the Gini coefficient (Z_4), has the most substantial adverse effect (–1.3133), suggesting that a one-unit increase in income inequality reduces the HFSI ($\widehat{Y_3}$) by 1.3133 units, holding all other factors constant. This result indicates that households in regions with higher income disparity experience lower food security. Additionally, employment status and educational attainment emerge as important factors influencing food security.

For illustration, consider a tertiary-educated, 45-year-old, married household head employed in W.P. Kuala Lumpur, with a household size of 4, and assuming a poverty rate of 12.7% (0.127) and a Gini coefficient of 0.38. Using (6), the estimated HFSI ($\widehat{Y_3}$) is 0.551, indicating a mild level of food insecurity. This model-based HFSI calculation indicates that, for this household, income inequality and household size are the most significant limiting factors, whereas education and age contribute positively to food security.

E. Goodness-of-Fit Test

Figure 16 illustrates the residual-versus-fitted plot. The increasing spread of residuals at higher fitted values indicates heteroscedasticity. Hence, heteroscedasticity-robust standard errors were applied, improving the reliability of inferences even though heteroscedasticity persists.

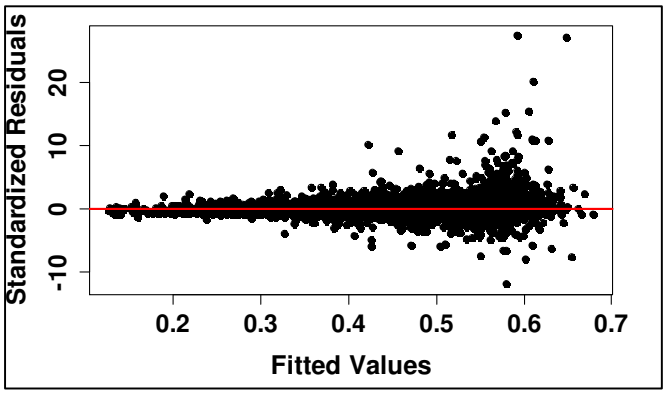

Fig. 16. Residual-versus-fitted plot.

Meanwhile, the Quantile-Quantile (QQ) plot shows deviations from the reference line, particularly in the tails, suggesting non-normality and potential outliers, as illustrated in Figure 17. These findings indicate that classical inference methods may be unsuitable. Therefore, robust standard errors and a diagnostic test were employed to ensure valid inference.

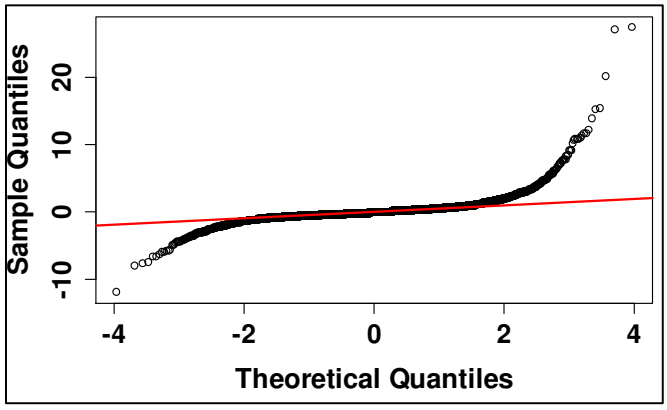

Fig. 17. QQ plot.

IV. CONCLUSION

This study provides empirical evidence on the role of household characteristics in shaping household income, food expenditure, and food security outcomes in Malaysia. Income inequality negatively affects household food security. In other words, the wider the gap between high-income and low-income households, the greater the likelihood that households will experience food insecurity. Conversely, higher education levels and stable employment contribute indirectly to improved food security by ensuring income stability. Meanwhile, larger household sizes are associated with lower per capita food security, likely due to increased dependency burdens and resource constraints within households. Additional determinants, including age, poverty incidence, and marital status, further influence household food security, highlighting the multifaceted nature of food insecurity at the household level.

A key contribution of this study is the integration of Two-Stage Least Squares (2SLS) with Least Absolute Shrinkage and Selection Operator (LASSO) and Elastic Net regularization techniques. This methodological framework effectively addresses endogeneity and multicollinearity while enabling systematic variable selection, thereby improving model parsimony and estimation accuracy. Furthermore, the development of the newly proposed Household Food Security Index (HFSI) offers a transparent and replicable measure of food security, providing policymakers with an analytically robust tool for evidence-based decision-making.

The findings underscore the importance of expanding both data coverage and methodological approaches in household food security research. Future studies may incorporate additional socioeconomic dimensions, such as dietary quality, food price dynamics, and local food availability, to better capture the underlying mechanisms of food insecurity. Extending the analysis beyond two survey periods would enable a more comprehensive examination of long-term trends and structural changes in household food security. From a

methodological standpoint, further exploration of alternative regularization methods or Bayesian estimation frameworks may enhance robustness and interpretability. Improvements in data preparation, including optimized preprocessing sequences, refined treatment of extreme observations, and rigorous validation of instrumental variables, would further strengthen model stability and the reliability of policy-relevant insights.

DECLARATION OF COMPETING INTERESTS

The authors declare that they have no known competing financial interests or personal relationships that could have influenced the work reported.

ACKNOWLEDGEMENT

Communication of this research is made possible through monetary assistance by Universiti Tun Hussein Onn Malaysia and the UTHM Publisher's Office via Publication Fund E15216.

DATA AVAILABILITY

The dataset on income inequality (measured by the Gini coefficient) and the poverty rate is publicly available in [25]. The datasets for the remaining variables are confidential and were made available to the authors through a Memorandum of Understanding (MoU) between the Department of Statistics Malaysia (DOSM) and Universiti Tun Hussein Onn Malaysia (UTHM). Access to this micro dataset is subject to DOSM's data-sharing policies and restrictions.

REFERENCES

[1] G. Makombe, "The food security concept: Definition, conceptual frameworks, measurement, and operationalization," *Africa Development*, vol. 48, no. 4, pp. 53–80, May 2024, https://doi.org/10.57054/ad.v48i4.5574.

[2] FAO, IFAD, UNICEF, WFP, and WHO, *The State of Food Security and Nutrition in the World 2025. Addressing high food price inflation for food security and nutrition*. Rome, Italy: FAO; IFAD; UNICEF; WFP; WHO, 2025, https://doi.org/10.4060/cd6008en.

[3] Martin. "Goal 2: Zero Hunger." United Nations Sustainable Development. https://www.un.org/sustainabledevelopment/hunger/.

[4] I. Manikas, B. M. Ali, and B. Sundarakani, "A systematic literature review of indicators measuring food security," *Agriculture & Food Security*, vol. 12, no. 1, May 2023, Art. no. 10, https://doi.org/10.1186/s40066-023-00415-7.

[5] K. Sajwan and S. P. Singh, "Multidimensionality of Household Food Security: A Systematic Review," *Current Nutrition Reports*, vol. 14, no. 1, Dec. 2025, Art. no. 127, https://doi.org/10.1007/s13668-025-00721-5.

[6] Y. H. Abdi *et al.*, "Regional disparities and sociodemographic determinants of food insecurity in Somalia: a secondary cross-sectional analysis of a National survey," *Journal of Health, Population and Nutrition*, vol. 44, no. 1, Oct. 2025, Art. no. 353, https://doi.org/10.1186/s41043-025-01078-9.

[7] A. Sheikhi, F. Bahador, and M. Arashi, "On a generalization of the test of endogeneity in a two stage least squares estimation," *Journal of Applied Statistics*, vol. 49, no. 3, pp. 709–721, Feb. 2022, https://doi.org/10.1080/02664763.2020.1837084.

[8] C. Chen, M. Ren, M. Zhang, and D. Zhang, "A Two-Stage Penalized Least Squares Method for Constructing Large Systems of Structural Equations," *Journal of Machine Learning Research*, vol. 19, Aug. 2018, Art. no. 2.

[9] Z. Sanioğlu-Tanış, D. Dündar-Öztaşçı, and İ. Özmen, "Fertility and women unemployment: new evidence from Türkiye," *Economic Change and Restructuring*, vol. 58, no. 6, Oct. 2025, Art. no. 89, https://doi.org/10.1007/s10644-025-09934-8.

[10] M. Olmez Turan, B. Gilbert, and T. Flamand, "How good are weather shocks for identifying energy elasticities? A LASSO-IV approach to European natural gas demand," *Journal of Commodity Markets*, vol. 39, Sept. 2025, Art. no. 100498, https://doi.org/10.1016/j.jcomm.2025.100498.

[11] O. M. Caccavale and V. Giuffrida, "The Proteus composite index: Towards a better metric for global food security," *World Development*, vol. 126, Feb. 2020, Art. no. 104709, https://doi.org/10.1016/j.worlddev.2019.104709.

[12] G. Y. V. Tang, K. G. Pillay, and A. Mustapha, "The Impact of Data Preprocessing Order on LASSO and Elastic Net Capabilities," *Engineering, Technology & Applied Science Research*, vol. 15, no. 1, pp. 20264–20270, Feb. 2025, https://doi.org/10.48084/etasr.9611.

[13] H. M. Nayem, S. Aziz, and B. M. G. Kibria, "Comparison among Ordinary Least Squares, Ridge, Lasso, and Elastic Net Estimators in the Presence of Outliers: Simulation and Application," *International Journal of Statistical Sciences*, vol. 24, no. 20, pp. 25–48, Dec. 2024, https://doi.org/10.3329/ijss.v24i20.78212.

[14] V. Tummalapalli, "Outlier Detection & Treatment for Machine Learning Models," *International Journal of Innovative Research and Creative Technology*, vol. 11, no. 3, pp. 1–8, June 2025, https://doi.org/10.5281/zenodo.16500050.

[15] F. Belotti, G. Mancini, and G. Vecchi, "Outlier Detection for Welfare Analysis," World Bank, Washington, DC, USA, Policy Research Working Paper 10231, 2022. https://doi.org/10.1596/1813-9450-10231.

[16] H. P. Vinutha, B. Poornima, and B. M. Sagar, "Detection of Outliers Using Interquartile Range Technique from Intrusion Dataset," in *Proceedings of the 6th International Conference on Frontiers of Intelligent Computing: Theory and Applications*, Bhubaneswar, India, 2017, pp. 511–518, https://doi.org/10.1007/978-981-10-7563-6_53.

[17] T. Bennich, Å. Persson, R. Beaussart, C. Allen, and S. Malekpour, "Recurring patterns of SDG interlinkages and how they can advance the 2030 Agenda," *One Earth*, vol. 6, no. 11, pp. 1465–1476, Nov. 2023, https://doi.org/10.1016/j.oneear.2023.10.008.

[18] S. Ullah, G. Zaefarian, and F. Ullah, "How to use instrumental variables in addressing endogeneity? A step-by-step procedure for non-specialists," *Industrial Marketing Management*, vol. 96, pp. A1–A6, July 2021, https://doi.org/10.1016/j.indmarman.2020.03.006.

[19] J. Fox and G. Monette, "Generalized Collinearity Diagnostics," *Journal of the American Statistical Association*, vol. 87, no. 417, pp. 178–183, Mar. 1992, https://doi.org/10.1080/01621459.1992.10475190.

[20] B. Abdul-Hameed and O. G. Matanmi, "A Modified Breusch–Pagan Test for Detecting Heteroscedasticity in the Presence of Outliers," *Pure and Applied Mathematics Journal*, vol. 10, no. 6, pp. 139–149, Dec. 2021, https://doi.org/10.11648/j.pamj.20211006.13.

[21] R. Tibshirani, "Regression Shrinkage and Selection Via the Lasso," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 58, no. 1, pp. 267–288, Jan. 1996, https://doi.org/10.1111/j.2517-6161.1996.tb02080.x.

[22] H. Zou and T. Hastie, "Regularization and Variable Selection Via the Elastic Net," *Journal of the Royal Statistical Society Series B: Statistical Methodology*, vol. 67, no. 2, pp. 301–320, Apr. 2005, https://doi.org/10.1111/j.1467-9868.2005.00503.x.

[23] D. R. Bhandari, K. Shah, and A. Bhandari, "The Power of Outliers in Research: What actually Works, and Does it Matter?," *Pravaha*, vol. 30, no. 1, pp. 84–91, Dec. 2024, https://doi.org/10.3126/pravaha.v30i1.76894.

[24] K. G. Pillay, F. M. San, R. M. Salleh, and A. Khamis, "LASSO regression to determine risk factors for road accident casualties in Malaysia in the presence of multicollinearity," *AIP Conference Proceedings*, vol. 2465, no. 1, June 2022, Art. no. 030001, https://doi.org/10.1063/5.0078299.

[25] Department of Statistics Malaysia. "Data Catalogue." OpenDOSM. https://open.dosm.gov.my.