11th MALAYSIA STATISTICS CONFERENCE 2024
Data and Artificial Intelligence: Empowering the Future
Sasana Kijang, Bank Negara
19th September 2024

Malaysian Household Poverty Determinant by Using Model Selection & Model Averaging

Siti Aisyah Mohd Padzil[1*] & Khuneswari Gopal Pillay[1]

[1] Universiti Tun Hussein Onn Malaysia

Abstract: Income in Malaysia is categorized into three classes: B40, M40, and T20. According to the Department of Statistics Malaysia (DOSM) in 2022, the average incomes for these groups are RM3,401 for B40, RM7,971 for M40, and RM19,750 for T20. Within each income group, there are further subcategories, with poverty primarily affecting the B40 group. Poverty status is determined by the Poverty Line Income (PLI), which has shown a significant increase over the years. The PLI rose from RM960 per month in Peninsular Malaysia in 2016 to RM2,208 in 2019, and by 2022, an income below RM2,589 on average was considered poor.

Given the rising trend in the Malaysian PLI, this research aims to identify the determinants of poverty using the 2019 Household Income & Expenditure data provided by DOSM. The study conducts a comparative analysis using three statistical modelling methods: model selection, model averaging, and variable selection model averaging. The findings highlight the key factors contributing to poverty based on the final best model.

Keywords:
Logistic Regression; Model Averaging; Model Selection; Malaysian Poverty

1. Introduction:

Poverty in Malaysia is determined by poverty line income which was revised once in several years. It is revised to ensure that the poverty measurement is in line with Malaysia's economic development. Incident of poverty reported by (DOSM, 2020) figured that poverty incidence had increased from 405.4 thousand in 2019 to 639.8 thousand in the following year. By the increasing poverty rate, Malaysia has come out with a prosperity vision which aims for a poverty eradication in year 2030 through job opportunity and career progression plan (Percetakan Nasional Malaysia Berhad. 2020).

Previous study by (Saidatulakma 2014 and Aisyah et al. 2019) summarized similar poverty factor which are state, household age, household gender, household marital, household education, household activity, household size and net income. Referring to household education and activity, most of B40 head of household tends to have poor education thus results in having lower-level job positions or might even be unemployed (Darshana et al. 2021). Besides that, states with most rural areas such as Sabah and Sarawak are more prone to poverty when compared to states which is more developed in industries and had more job opportunities.

According to Khaled et al. (2020), increasing incident of poverty will impact the individual life expectancy. Factors such as employment and income have strong correlations with

health, thus relates with persons longevity. Higher poverty percentage in a country will also contributes to malnutrition due to lack of necessary daily nutrition intake (Faareha et al. 2020).

To prevents worst poverty impact from becoming reality, Malaysian government had come out with poverty eradication plan since 1996 which is called as Seventh Malaysia Plan (Economic Planning Unit, Prime Minister Department Malaysia, 2017). The plan includes the development of Bumiputra Commercial and Industrial Community which focus on both rural and urban areas. This research intended to help policy makers in Malaysia in organizing strategies to eradicate poverty by summarizing the most influential determinants of poverty in year 2019.

2. Methodology:

Three types of modeling techniques were applied to the household data which are logistic regression using model selection, model averaging, and a newly proposed method named variable selection-model averaging. Two popular information criteria, the Corrected Akaike Information Criterion (AICc) and the Bayesian Information Criterion (BIC), were used to determine the most parsimonious model. Additionally, Brier scores were computed to compare the predictive performance of each modeling method.

Logistic regression is a statistical modelling technique specifically designed for working with binary outcomes data where the respondent variable is taking on value 1 (Yes) and 0 (No).The basic general model of logistic regression was defined by (Kutner *et al.*, 2008). Unlike regression which with a continuous number for dependent variable, logistic regression presents the binary response variable in form of probability which falls between 0 to 1. For example, probability of 0.7 will explain that there are 70% chances of “Poor” and remaining 30% “Not Poor” to occur based on the coefficient and covariates.

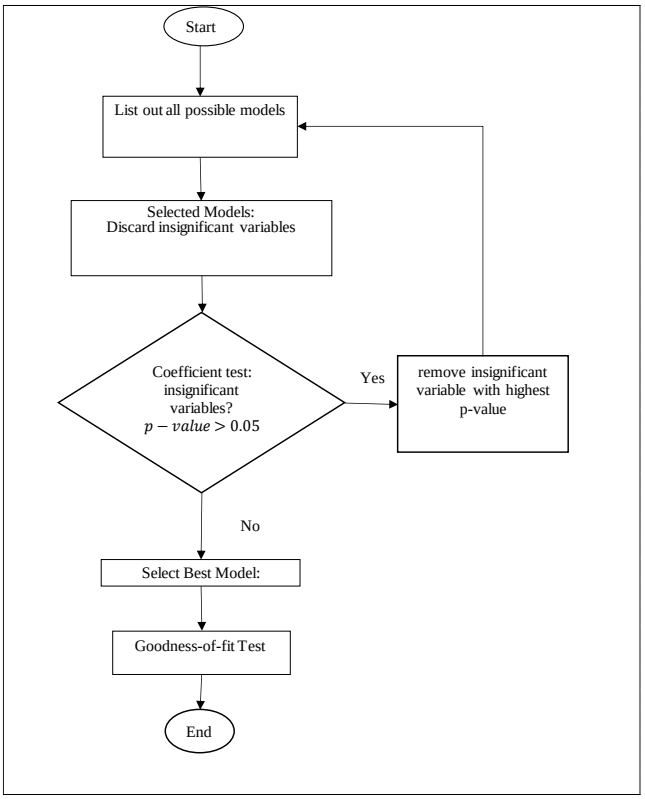

FIGURE 1: Model-building of Logistic Regression using Model Selection

Model selection refers to the procedure of selecting the best models from a set of candidate models. Zainodin & Khuneswari (2010) summarized the procedure in model selection for linear regression into four phases. A similar procedure also can be applied to build logistic regression model and had been clearly explained step by step in (Khuneswari & Aisyah, 2018). Figure 1 summarizes the flows of model selection.

Model averaging is another modelling approach which was created to overcome uncertainty issues in model selection. The idea of model averaging is to assign weights on all possible models, so that no variables are omitted. Independent variables with greater significance will receive more weight while less important variables will have a weaker weight. In other words, model averaging incorporates information and prediction from all possible models to produce final model with better accuracy while accounting for model uncertainty caused by variable removal in model selection. The common model averaging process starts from listing out all possible models until forming model using average coefficient and finally computing the Goodness-of-fit. Figure 2 visualizes the modelling flows in model averaging and the variable selection- model averaging. For the newly improvised method, the arrow with dashes is the proposed additional step by (Aisyah et al.2019). The modelling process will rerun until no more insignificant variables appear in the final model.

Finally, to determine the best predictive performance among models, Brier Score as in (Steyrberg *et al.,* 2001) was used. The formula is similar with MSE but according to (Steyrberg *et al.,* 2001), Brier Score range model from 0 (perfect) to 0.25 (worthless). Hence, Brier Score near to 0 indicated a better model and as it exceeds 0.25, the model is said to be worthless.

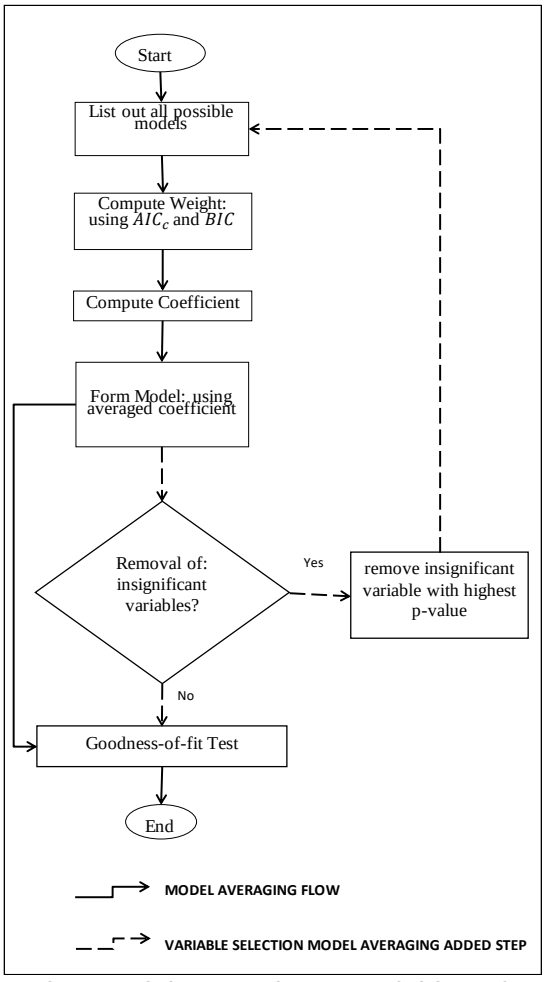

FIGURE 2: Modelling using Model Averaging & Variable Selection Model Averaging

TABLE 1: Data Description

Variables	Description	
Y	**Poverty Status**	
	0: Poor	1: Not Poor
X1	**Household Size**	
X2	**State**	
	1: Johor	9: Perlis
	2: Kedah	10: Selangor
	3: Kelantan	11: Terengganu
	4: Melaka	12: Sabah
	5: Negeri Sembilan	13: Sarawak
	6: Pahang	14: W.P. Kuala Lumpur
	7: Pulau Pinang	15: W.P. Labuan
	8: Perak	16: W.P. Putrajaya
X3	**Strata**	
	1: Urban	2: Rural
X4	**Ethnic**	
	1: Bumiputera	3: Indian
	2: Chinese	4: Others
X5	**Gender**	
	1: Male	2: Female
X6	**Age**	
X7	**Highest Certificate**	4: STPM
	1: Degree/Advance	5: SPM/ SPMV
	2: Diploma	6: PMR/SRP
	3: Diploma / certificate	7: No Certificate
X8	**Activity Status**	
	1: Employer	8: Student
	2: Government employee	9: Government pensioner
	3: Private employee	10: Private pensioner
	4: Own account worker	11: Elderly
	5: Unpaid family worker	12: Persons with
	6: Unemployed	15: Others
	7: Housewife	
X9	**Occupation**	
	1: Manager	7: Craft and related trades work
	2: Professional	8: Plants and machine operators
	3: Associate professionals	9: Elementary occupations
	4: Clerical support workers	10: Not classified
	5: Services and sales workers	
	6: Skilled agricultural	
X10	**Working Industry**	
	1: Agriculture, forestry and fishing	11: Financial
	2: Mining and quarrying	12: Professional
	3: Manufacturing	13: Administrative
	4: Electricity supply	14: Public administration
	5: Water supply	15: Education
	6: Construction	16: Human health
	7: Wholesale and retail trade	17: Arts
	8: Transportation and storage	18: Other service activities
	9: Accommodation and food services	19: Household as employers
		20: Organizations
	10: Information and communication	21: Industries not classified

The data of Household Income & Expenditure 2019 utilized in this study is the courtesy from Department of Statistic Malaysia (DOSM). This data comprises of 16354 samples with ten variables as summarized in Table 1. The poverty status for each sample is determined by comparing head of household income with Malaysian poverty line income

(PLI) provided by DOSM in year 2019 which is RM2208 (DOSM, 2019). Household income which falls under this PLI is considered poor.

3. Result:

General overview regarding poverty in Malaysia as well as areas that was most affected by poverty were visualized using several graphs.

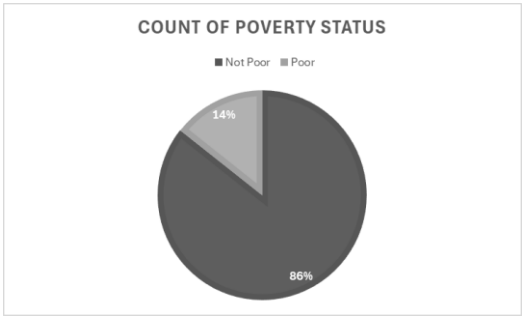

FIGURE 3. Percentage of Poverty

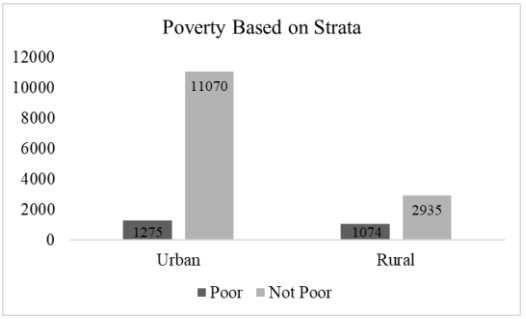

FIGURE 4. Poverty based on Strata

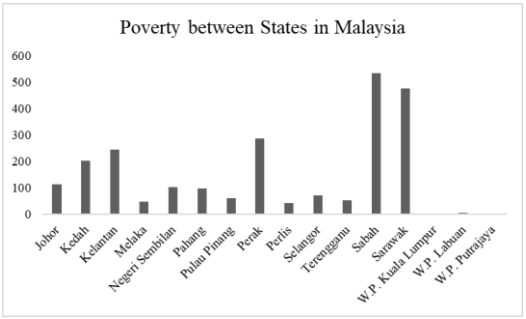

FIGURE 5. Poverty based on States

Out of 16534 sample for head of household data in Malaysia, 14% or 2349 respondents have income below Malaysian's poverty line income. Examining the data in Figure 3, it is evident that within this 14%, the majority come from urban areas. However, when comparing the sample sizes of each stratum, 27% of individuals in rural areas are living in poverty, whereas only 10% of those in urban areas are affected by poverty. Relating the results from Figure 4 and Figure 5, the urban areas or areas with most industries such as Kuala Lumpur, Labuan and Putrajaya have slight to no poverty whereas Sabah & Sarawak which have the most rural areas in Malaysia have a significant number of poverties compared to other states.

Table 2 compares the predictive performance among models by using Brier Score.

TABLE 2: Predictive Performance

Modelling Methods	Selection Criteria	Brier Score
Model Selection	Backward Selection	0.10008
Model Averaging	**AICc**	**0.10001**
Model Averaging	BIC	0.10015
Variable Selection-Model Averaging	**AICc**	**0.10002**
Variable Selection-Model Averaging	BIC	0.10015

The outcomes demonstrate that both best model formed using model averaging and variable selection model averaging with AICc slightly have better compared to model bulit by using model selection. When evaluating predictive performance between information

criteria, both traditional model averaging and the proposed method shows a better performance when using AICc as the selection criteria.

TABLE 3: Final Model

Modelling Methods	Models
Model Selection	$\widehat{\boldsymbol{Y}} = -5.15248 - 0.53426\boldsymbol{X1} + 0.88162\boldsymbol{X3} - 0.164513\boldsymbol{X4} - 0.33862\boldsymbol{X5} - 0.01688\boldsymbol{X6} + \boldsymbol{X7} + 0.11652\boldsymbol{X8} + 0.57489\boldsymbol{X9}$
Model Averaging (AICc)	$\widehat{\boldsymbol{Y}} = -5.1589 - 0.5335\boldsymbol{X1} + 0.004676\boldsymbol{X2} + 0.88195\boldsymbol{X3} - 0.16437\boldsymbol{X4} + 0.34405\boldsymbol{X5} - 0.01671\boldsymbol{X6} + 0.11658\boldsymbol{X7} + 0.575567\boldsymbol{X8} + 0.254525\mathbf{X9} + 0.001605\mathbf{X10}$
Model Averaging (BIC)	$\widehat{\boldsymbol{Y}} = -5.10572 - 0.5308\boldsymbol{X1} + 0.004712\boldsymbol{X2} + 0.8799\boldsymbol{X3} - 0.16862\boldsymbol{X4} + 0.389802\boldsymbol{X5} - 0.01555\boldsymbol{X6} + 0.117049\boldsymbol{X7} + 0.579096\boldsymbol{X8} + 0.255064\mathbf{X9} + 0.001607\mathbf{X10}$
Selection-Model Averaging (AICc)	$\widehat{\boldsymbol{Y}} = -5.15248 - 0.5338\boldsymbol{X1} + 0.88131\boldsymbol{X3} - 0.16508\boldsymbol{X4} + 0.345582\boldsymbol{X5} - 0.0167\boldsymbol{X6} + 0.11653\boldsymbol{X7} + 0.57548\boldsymbol{X8} + 0.255015\mathbf{X9}$
Selection-Model Averaging (BIC)	$\widehat{\boldsymbol{Y}} = -5.10525 - 0.5308\boldsymbol{X1} + 0.8799\boldsymbol{X3} - 0.1686\boldsymbol{X4} + 0.38952\boldsymbol{X5} - 0.01555\boldsymbol{X6} + 0.11705\boldsymbol{X7} + 0.5791\boldsymbol{X8} + 0.25508\mathbf{X9}$

Referring to Table 3, for both information criteria, variable selection model averaging had eliminated three insignificant variables which are $X2$: State and $X10$: Working Industry.

4. Discussion and Conclusion:

By examining the results of predictive performance, it is found that the predictive accuracy value had a slight difference between model averaging and variable selection model averaging. If the decimal values are rounded to one decimal place, the error would be identical. Several past studies have argued that model selection introduces uncertainty in parameter estimates (Schomaker & Heuman, 2014). Burnham and Anderson, 2002 developed model averaging to cater the under-estimation issues, hence averaging is a more preferred modelling technique.

Based on the result, since the accuracy value for model averaging and variable selection-model averaging is similar, the best model can be determined by the researcher's objectives of study. If the aim is to point out the determinants or most contributing covariates, variable selection model averaging using AICc can be applied. Hence, the most preferred model is

$$\widehat{\boldsymbol{Y}} = -5.15248 - 0.5338\boldsymbol{H.Size} + 0.88131\boldsymbol{Strata} - 0.16508\boldsymbol{Ethnic} + 0.345582\boldsymbol{Gender} - 0.0167\boldsymbol{Age} + 0.11653\boldsymbol{Certificate} + 0.57548\boldsymbol{Activity} + 0.255015\mathbf{Occupation}$$

Based on the best model, there are eight determinants of poverty which are household size, strata, ethnic, head of household gender, head of household age, head of household activity status, certificate and head of household type of occupation. The results are almost similar with factors of Malaysian poverty in 2016 by (Aisyah et al., 2019) except for marital status. Apart from the final model, the descriptive statistics had visually proven strata as one of the reasons of poverty among Malaysian.

Overall, this study conducted the application and comparison between three logistic regression modelling methods on household income and expenditure data to highlight the determinants of poverty in Malaysia for the year 2019. In statistics fields, this

research offers an improved modelling method which helps in reducing model uncertainty and summarizes contributing factors at the same time. Also, this paper offers insights to policymakers for designing programs aimed at alleviating poverty in Malaysia based on the highlighted determinants.

References:

1. Burnham and Anderson. 2002. Model Selection and Multimodel Inference:A practical Information-theoretic Approach. 2nd Ed. NewYork: Springer-Verlag.
2. Darshana Darmalinggam , Maniam Kaliannan and Mageswari Dorasamy. 2021. Proactive measures to eradicate Malaysia's poverty in IR4.0 era: a shared prosperity vision. F1000 Research. 28;10:1094. doi: 10.12688/f1000research.73330.2. PMID: 35237432; PMCID: PMC8790706.
3. Department of Statistic Malaysia. 2020. Household Income & Basic Amenities Survey Report 2019. Reference: https://v1.dosm.gov.my/v1/index.php?r=column/cthemeByCat&cat=120&bulid=TUTmRhQ1N5TUxHVWN0T2VjbXJYZz09&menu_id=amVoWU54UTl0a21 NWmdhMjFMMWcyZz09.
4. Economic Planning Unit, Prime Minister Department Malaysia. 2017. Malaysia Succes Story in Poverty Eradication.
5. Faareha Siddiqui, Rehana Salam, Zohra Lassi and Jai Das. 2020. The Intertwined Relationship Between Malnutrition and Poverty. Front Public Health. 28;8:453. DOI: 10.3389/fpubh.2020.00453
6. Kutner, M. H., Nachtsheim, C. J. and Neter, J. (2008). Applied Linear Regression Models. 4th edition. Singapore: McGraw-Hill Inc.
7. Khuneswari Gopal Pillay, Siti Aisyah Mohd Padzil & Noraini Abdullah. 2018. Model Selection and Model Averaging on Mortality of Upper Gastrointestinal Bleed Patients IOSR Journal of Dental and Medical Sciences: 17, 68–78. https://doi.org/10.9790/0853-1711086878
8. Khaled Thafran, Makmur Tumin and Ahmad Farid Osman. 2020. Poverty, Income and Unemployment as Determinant of Life Expectency: Empirical Evidence from Panel Data of Thirteen Malaysian States. Iranian Journal of Pubic Health
9. Mu Y, See I, Edwards JR. 2019. Bayesian model averaging: improved variable selection for matched case-control studies. Epidemiol Biostat Public Health. 16(2):13048. DOI: 10.2427/13048. PMID: 31772926; PMCID: PMC6879006.
10. Micheal Schomaker and Christian Heumann. 2014. Model selection and model averaging after multiple imputation. Computational Statistics and Data Analysis (71)758-770.
11. Percetakan Nasional Malaysia Berhad. 2020. Shared Prosperity Vision 2030. Restructuring the priorities of Malaysia Development.
Saidatulakmal. 2014. Poverty Issues Among Malaysian Elderly. Proceeding of the Social Sciences Research: 123-132
12. Siti Aisyah Mohd Padzil, Khuneswari Gopal Pillay, Mohd Saifullah Rusiman, & Rohayu Mohd Salleh. 2019. New model averaging approach in predicting mortality rate of intensive care unit patients. Journal of Physics: Conference Series, 1366(1). DOI: https://doi.org/10.1088/1742-6596/1366/1/012123
13. Steyerberg, Harrell, Borsboom et al. 2001. Internal Validation of predictive models: Efficiency of some procedures for logistic regression analysis. Journal of Clinical Epidemiology, 54, pp. 774-781.
14. Zainodin H J and Khuneswari Gopal Pillay. 2010. Model-Building Approach in Multiple Binary Logit Model using Coronary Heart Disease. Malaysian Journal of Mathematical Sciences vol 4) pp 107-133