Determination of Recent In-Migrants in Yogyakarta using Bayesian Regression Logistics

.

ABSTRACT

Aims: This research aims to find out the status determinant of recent in-migrants entering the province of Yogyakarta, Indonesia in 2021 with Bayesian logistic regression combined with a nonlinear principal component analysis in the process of forming a latent variable.
Study design: Quantitative Design.
Place and Duration of Study: Sample: Department of Medicine (Medical Unit IV) and Department of Radiology, Services Institute of Medical Sciences (SIMS), Services Hospital Lahore, between June 2009 and July 2010.
Methodology: The data were obtained from the results of national socio-economic surveys (SUSENAS) KOR in March 2021.
Results: The research results indicate that particular variables such as age, resident latest education, the status of main activities, the status of residential ownership, housing quality, and asset ownership have significant influences on recent in-migrants entering the Province of Yogyakarta, Indonesia.
Conclusion: This research concludes that the Bayesian approach in logistic regression with iteration 1,000,000, 4 thinning interval, and 500.000 burn-in indicates that of seven variables, six variables with significant influences on the status of recent migrants entering the Province of Yogyakarta consist of age (X1), latest education (X3), main activities (X4), house ownership (X5), housing quality (X6), and asset ownership (X7). Of these six variables, the younger resident, the resident with high school or equal as their latest education, the resident currently working as their main activity, the resident renting a house, and the resident with high housing quality score and high asset ownership score are more likely to do recent migration to the Province of Yogyakarta.

Keywords: Recent In-migrants, Bayesian Logistic Regression, Yogyakarta

1. INTRODUCTION

Logistic regression is a method in statistical analysis to describe the connection between independent variables having two or more categories of dependent variables with reference to a categorical or interval scale[1]. The logistic regression comprises binary, multinominal, and ordinal logistic regressions. The binary logistic regression is intended to analyze the connection between one response variable and several predictor variables. The response variable consists of dichotomous qualitative data with the value of 1 to indicate the existence of a characteristic and 0 to indicate the absence of a characteristic.

Several cases could be analyzed based on logistic regression, but this regression is not adequate to analyze the latent variable formed by several indicators. If the indicators forming the latent variable are analyzed partially, they will result in more variables being studied, leading to inefficiency of the research, especially in analysis and interpretation. It is not possible to apply arithmetic operations in the indicators forming the latent variable due to

varied measurements. To tackle this issue, logistic regression can be combined with the nonlinear principal component analysis to transform latent variable data.

Solimun-Fernandes[2] stated that, generally, latent variable is defined as a variable that cannot be directly measured but should involve the reflecting or forming indicators. To measure the latent variable data composed by other indicators, it can use a principal component scoring method obtained from the nonlinear principal component analysis. The result of the transformation utilizing the nonlinear principal component analysis (PCANL) could be referred to as data input for the following analyses, such as logistic regression.

In this logistic regression modeling, parameter estimation is regarded as a vital stage. The performance of this estimation is often affected by the sample size and data characteristics. An unbalanced dependent variable is often seen in logistic regression when one of the classes determined is uncommon[3]. This condition could affect the performance of the estimation method used, and to deal with this problem, the Bayesian method can be employed as an estimation.

The research once conducted by King & Zeng[4] indicates that the Bayesian method is unbiased for unbalanced data. They also add that the parameter estimation that refers to Bayesian yields a more relevant result than the conventional method often used in parameter estimation in logistic regression, namely the Maximum Likelihood, to model the case with an unbalanced dependent variable. That is, this research adopted the Bayesian method to estimate the parameter of the logistic regression.

One of the cases that requires nonlinear principal component analysis to transform data and is used for logistic regression input with Bayesian estimation method is to find out the determinant of the recent migrant status in Yogyakarta. Migration serves as the response to the variation in the condition of a neighborhood where the population resides. Lee[5] argues that there are several matters affecting migration, and economic motive is among others. Recent migration, one of the types of migration, according to the Statistics Centre, refers to recent migrants whose province in the past five years was different from the province during the census. In other words, the recent migrants entering the province of Yogyakarta refer to the residents living in another province other than Yogyakarta before the census took place. The data on recent migrants were obtained from National Socio-Economic Survey (SUSENAS) KOR in March 2021.

The study conducted by Syairozi & Wijaya[6] implies that age has a significant influence on the decision made by informal migrant workers in Pasuruan, while education, sex, number of dependents, land ownership, and real wage differences are significant to the decision to migrate. Furthermore, Dustmann & Glitz[7] argue that migration and education are inseparable. Sarmita & Simamora[8] studied the social and economic characteristics of migrants from Java using descriptive statistic methods, while Statistics and Data Center of Education and Culture defines component variables forming the socio-economic status of households by referring to the characteristics of housing quality and asset ownership. Of these two studies, it is obvious that the socio-economic variable is constructed by other indicators such as the characteristics of housing quality and asset ownership.

Departing from previous studies, this research adopted pre-existing variables to identify factors affecting the status of recent migrants entering the province of Yogyakarta 2021 by employing a Bayesian approach of logistic regression.

2. MATERIAL AND METHODS

2.1 Migration

In a broader term, migration is defined as permanent or semi-permanent residential change (Tjiptoherijanto, 2009). Mantra (2012) argues that a person can be said to migrate when the person moves to another residential place permanently or relatively permanently (for a minimum period of time) by reaching a particular minimum distance or moving from one geographical unit involving the residential change from the place of origin to another point of destination. According to the Statistics Center, to define the term migration, it is essential to refer to administrative boundaries that cover provinces, regencies, villages, sub-districts/hamlets, and the scope of minimum time of six months or less than six months, if the person concerned has a plan to reside in the place of destination. According to Mantra (2012), recent migration is among other migration types, where a person can be categorized as a migrant in the residential area during the census is different from five years ago before the census took place.

Lee (1976) argues that there are four factors that need attention in the process of population migration such as the place of origin, factors existing in the place of destination, obstacles between the place of origin and destination, and the factors existing in the place of origin and destination

2.2 Non-linear Principal Component Analysis

A principal component analysis is a multivariate analysis used to transformed the original variable set into new and smaller variables explaining the majority of the variety of the original variable set (Dillon and Goldstein, 1984). Gifi (1981) defines the specific multivariate analysis as linear and non-linear. Mixed scale indicators (metric and non-metric) used the non-linear principal component analysis to transform data. The non-linear transformation refers to the principal component analysis with optimal scaling from qualitative scale to quantitative value (Markos, et. al., 2010 and Meulman, et. al., 2004).

In the non-linear principal component analysis, the category of all variables with the scale other than numeric scale is labelled with categorical quantification with relevant numeric scale; the non-linear principal component analysis aims to optimize or find the quadratic mean of optimal correlation between variables labelled with the quantification of categorical components.

Gifi (1981) opines that the observation analyzed using a non-linear principal component will form matrix H of $n \times m$ size. This matrix $\boldsymbol{H}$ is then broken down into vector h_j which is transformed and normalized in PRINCALS in package Homals of software R. The transformation result of matrix $\boldsymbol{H}$ to $\boldsymbol{G}$ via vector h_j can be written in the block matrix in equation (1).

$$\boldsymbol{G} \triangleq [\boldsymbol{G}_1 \vdots \boldsymbol{G}_2 \vdots \cdots \vdots \boldsymbol{G}_m] \quad (1)$$

From matrix $\boldsymbol{G}_j$, the following process refers to equation (2).

$$\boldsymbol{D}_j \triangleq \left[\boldsymbol{G'}_j \boldsymbol{G}_j\right] \quad (2)$$

where:

$\boldsymbol{D}_j$: diagonal matrix $k_j \times k_j$ with the relative frequency of variable j in the main diagonal

$\boldsymbol{G}_j$: indicator matrix for each indicator

$\triangleq$ means defined.

Matrix of quantification category of variable *j* is formulated with equation (3):

$$\boldsymbol{Y}_j=\boldsymbol{D}_j^{-1}\boldsymbol{G}_j'\boldsymbol{X} \quad (3)$$

with

$$\boldsymbol{X}=m^{-1}\boldsymbol{G}\boldsymbol{Y} \quad (4)$$

$\boldsymbol{Y}_j$: multicategory calculation ($k_j \times p$)
$\boldsymbol{D}_j$: diagonal matrix $k_j \times k_j$ (relative frequency of the variable *j* in the main diagonal)
$\boldsymbol{G}_j$: indicator matrix for each indicator
$\boldsymbol{X}$: matrix score of object component ($n \times m$)
m : number of variables
$\boldsymbol{G}$: block compound matrices of $\boldsymbol{G}_j$
$\boldsymbol{Y}$: a set of multiple and single category quantification

In the non-linear principal component analysis, the optimal linear combination model used to calculate the principal component score refers to equation (5):

$$\begin{aligned} \boldsymbol{z} &= \sum_{j=1}^{m} \boldsymbol{a}_j \boldsymbol{G}_j \boldsymbol{Y}_j \\ &= \sum_{j=1}^{m} \boldsymbol{a}_j \boldsymbol{q}_j \end{aligned} \quad (5)$$

where:

$\boldsymbol{z}$: principal component score
$\boldsymbol{a}_j$: component weight with the order $p \times 1$
$\boldsymbol{G}_j$: indicator matrix j of the size $n \times k_j$
$\boldsymbol{Y}_j$: calculation of multicategory with order $k_j \times p$
$\boldsymbol{q}_j$: transformation data

2.3 Logistic Regression Analysis

Logistic regression is a basic classification method initially intended for response variable or dependent variable with two classes namely binary logistic regression, which further develops with a dependent variable that consists of Multinomial Logistic Regression. Binary logistic regression is a method of analysis used to find out the connection between a dependent variable that is binary or dichotomous and a predictor variable or independent variable that is polychotomous (Hosmer and Lemeshow, 2013).

If there is an observation (***X, Y***) in which the ***X*** represents the independent variable and ***Y*** is the dependent one with the (6) and can be transformed as in equation (7):

$$\pi(\boldsymbol{X}) = \frac{exp(\beta_0+\beta_1 x_1+\cdots+\beta_m x_m)}{1+exp(\beta_0+\beta_1 x_1+\cdots+\beta_m x_m)} \quad (6)$$

$$ln\left(\frac{\pi(\boldsymbol{X})}{1-\pi(\boldsymbol{X})}\right) = (\beta_0 + \beta_1 x_1 + \cdots + \beta_m x_m) = \beta^T \boldsymbol{X} \quad (7)$$

2.3.1 Parameter Estimation in Logistic Regression

One of the methods commonly used to estimate the parameter in logistic regression is the maximum likelihood (MLE) and Bayesian method. The Bayesian method was developed according to the Bayes theorem. In this case study, the application of this method is intended to compound information coming from data with prior probability in terms of model validity level, so that the best model can be selected with the highest posterior probability and an average sum is obtained (Rachev, et. al., 2008).

In Bayes theorem, B_i, with $i = 1, 2, \ldots, n$ as sample space S with $P(B_i) \neq 0$ and representing an independent event; thus, for random event A where $P(A) \neq 0$, probability B_i with condition A is given as follows:

$$P(B_i|A) = \frac{P(A|B_i)P(B_i)}{\sum_{i=1}^{n} P(B_i)P(A|B_i)} \quad (8)$$

Furthermore, if $\sum_{i=1}^{n} P(B_i)P(A|B_i)$ is regarded as constant, equation (8) will turn to equation (9):

$$P(B_i|A) \propto P(A|B_i)P(B_i) \quad (9)$$

According to Ghosh, et. al., (2007), in addition to the model $f(x|\theta)$ or the likelihood, Bayesian requires the distribution for θ, or known as prior. Liu and Powers (2012) suggest that non-informative prior can be referred to without initial knowledge of the parameter distribution to determine the prior distribution, while Genkin, et. al., (2007) argue that the prior distribution for the parameter in the binary logistic regression model of Bayesian follows the normal distribution. According to Walpole, et. al., (2012), estimating parameter θ may refer to the distribution $f(x|\theta)$ and $\pi(\theta)$, with $\pi(\theta)$ as prior distribution for θ. This refers to θ given X (observed data) called posterior distribution given in the following formula:

$$\pi(\theta|x) = \frac{f(x|\theta)\pi(\theta)}{g(x)} \quad (10)$$

so:

$$\pi(\theta|x) \propto f(x|\theta)\pi(\theta) \quad (11)$$

2.3.2 Markov Chain Monte Carlo (MCMC)

Markov Chain Monte Carlo or MCMC is based on the construction of Markov Chain that is convergent with posterior distribution as the target distribution $f(\theta|x)$. The MCMC method is also known as an iterative model, considering that each stage yields a score contingent upon the previous stage. One of the algorithms in MCMC, in which conditional distribution has the known form, uses Gibbs sampling.

The convergence checking in the MCMC method is intended to figure out whether the data is relevant to the prior distribution. This convergence checking method of MCMC can refer to Trace Plot, MC error, and autocorrelation (Ntzoufras, 2011).

- Trace Plot; if the model has converged, the trace plot result does not form a particular pattern.
- MC error; if the model has converged, the score of MC Error is very low (less than 5% of standard deviation).
- Autocorrelation; if the model has converged, the first lag the autocorrelation score is close to one and the following lag shows an autocorrelation score heading for zero.

Moreover, parameter testing is required to investigate whether the predictor variable significantly affects the response variable. In the Bayesian method, the parameter test refers to credible intervals of 2.5% and 97.5% quantile of the distribution. If the credible intervals do not indicate a score of 0, the predictor variable significantly affects the response variable.

2.3.3 Interpretation.

In logistic regression modelling, parameter interpretation is aimed to find out the value estimation of the predictor variable. Interpreting the logistic regression parameter of the categorical variable uses Odds Ratio (Hosmer and Lemeshow, 2013). The odds ratio represents a ratio between the probability of success and the probability of failure, leading to relative probability from the probability of success towards the probability of failure. The odds

ratio is also referred to as the exposure association (risk factor) of an event. The following is equation of the Odds and Odds Ratio (Azen and Walker, 2011):

$$odds = \frac{\pi}{1-\pi} \quad (12)$$

$$Odds\ Ratio = \frac{Odds1}{Odds2} \quad (13)$$

2.4 Material

This research employed secondary data gathered from Migrant Profiles of Socio-Economic Survey Results KOR in March 2021 downloaded from Silastik BPS. The criteria of respondents covered 17 to 64-year-old Indonesian citizens residing in Yogyakarta during the census. The variables and indicators used in this research are presented in Table 1.

Table 1. Variable Outlines

Variable	Indicator	Answer	Data Scale
Recent Migrants (Y)	-	(0) No (1) Yes	Nominal
Age (X1)	-	15-64 years old	Ratio
Sex (X2)	-	(0) Men (1) Women	Nominal
Latest Education (X3)	-	(0) Not going to school (1) Primary- Secondary School (2) High School (3) University Qualifications	Ordinal
Main Activities (X4)	-	(0) Others (1) Studying (2) Working (3) Taking care of the household	Nominal
Home ownership(X5)	-	(0) Not under their ownership (1) Under their ownership	Ordinal
Housing quality (X6)	Widest roof type (I6.2)	(0) Fibers and others (1) Not fibers	Ordinal
	Widest wall type (I6.3)	(0) Others (1) Brick	Ordinal
	Widest floor type (I6.4)	(0) Ground (1) Not ground	Ordinal
	Defecation facility (I6.5)	(0) Public property (1) Own/shared	Ordinal
	Toiler type (I6.6)	(0) Others (1) Gooseneck Toilet	Ordinal
	Lighting source (I6.7)	(0) Others (1) Electricity with meter	Ordinal
	Drinking water facilities (I6.8)	(0) Others (1) Bottled water/refill	Ordinal
	Cooking fuel (I6.9)	(0) Others	Ordinal

Variable	Indicator	Answer	Data Scale
		(1) Electricity/gas	
Asset Ownership (X7)	Motorcycle (I7.1)	(0)No (1)Yes	Ordinal
	Gold (min.10 gr) (I7.2)	(0)No (1)Yes	Ordinal
	Flat screen TV (min. 30 inch) (I7.3)	(0)No (1)Yes	Ordinal
	Air Conditioner (I7.4)	(0)No (1)Yes	Ordinal
	Water heater (I7.5)	(0)No (1)Yes	Ordinal
	Gas cylinder >5.5 kg (I7.6)	(0)No (1)Yes	Ordinal
	Refrigerator (I7.7)	(0)No (1)Yes	Ordinal
	Laptop (I7.8)	(0)No (1)Yes	Ordinal
	Car (I7.9)	(0)No (1)Yes	Ordinal
	Land (I7.10)	(0)No (1)Yes	Ordinal
	Landline (I7.11)	(0)No (1)Yes	Ordinal

3. RESULTS AND DISCUSSION

3.1 Results of Non-linear Principal Component Analysis

One of the outputs obtained from the non-linear principal component analysis with software R and homals package is the component loading score. The component loading can be seen in Table 2.

Table 2. Component Loading

Variable	Indicators	Loading
X6	I6.1	-0.00546
	I6.2	-0.17883
	I6.3	-0.2073
	I6.4	-0.21008
	I6.5	-0.25341
	I6.6	-0.12467
	I6.7	-0.08904
	I6.8	-0.12449

Variable	Indicators	Loading
X7	I7.1	-0.0812
	I7.2	-0.18931
	I7.3	-0.19984
	I7.4	-0.21103
	I7.5	-0.14889
	I7.6	-0.21292
	I7.7	-0.16007
	I7.8	-0.17334
	I7.9	-0.21122
	I7.10	-0.05051
	I7.11	-0.1562

As in Table 2, the model for X_6 and X_7 can be given as follows:

$$X_6 = -0.00546\, I_{1.1} - 0.17883\, I_{1.2} - 0.20730\, I_{1.3} - 0.21008\, I_{1.4} - 0.25341 I_{1.5} - 0.12467 I_{1.6} - 0.08904 I_{1.7} - 0.12449 I_{1.8}$$

$$X_7 = -0.08120\, I_{2.1} - 0.18931\, I_{2.2} - 0.19984\, I_{2.3} - 0.21103\, I_{2.4} - 0.14889\, I_{2.5} - 0.21292\, I_{2.6} - 0.16007\, I_{2.7} - 0.17334\, I_{2.8} - 0.21122\, I_{2.9} - 0.05051\, I_{2.10} - 0.15620\, I_{2.11}$$

The next output obtained from the non-linear principal component analysis is category quantification score. Table 3 shows the quantification of categories obtained from the analysis results of the non-linear principal component analysis. This quantification of categories was used to replace the respondent qualitative data and for the calculation of transformation data and principal component score in line with equation (5).

Table 3. Category quantification score

Variable	Indicators	Category quantifications	
		0	1
X6	I6.1	0.003857	-0.0000019
	I6.2	0.011322	-0.0006985
	I6.3	0.019054	-0.0005577
	I6.4	0.02715	-0.000402
	I6.5	0.028231	-0.0005625
	I6.6	0.008002	-0.0004803
	I6.7	0.000782	-0.002508
	I6.8	0.003509	-0.0010922
X7	I7.1	0.002845	-0.000573
	I7.2	0.001597	-0.0055467
	I7.3	0.001328	-0.0074347
	I7.4	0.00096	-0.0114818
	I7.5	0.000386	-0.0142123
	I7.6	0.001331	-0.0084216
	I7.7	0.003118	-0.0020285
	I7.8	0.001816	-0.0040918
	I7.9	0.001455	-0.0075821
	I7.10	0.00126	-0.0004987
	I7.11	0.000474	-0.0127457

The following step is to find out the score of the quantification of categories and the component loading that required the calculation of the principal component score. To obtain this score, the multiplication of the category quantification score and the component loading of each indicator of the dimension used was required. The principal component score was obtained by adding up the multiplication result of the transformed data of each respondent to the component loading in each indicator.

Table 4. Combined Data

Subject	Y	X_1	X_2	X_3	X_4	X_5	X_6	X_7
1	0	41	0	2	1	0	0.005938	0.013365
2	1	39	1	3	3	0	0.005938	0.013365
3	1	54	0	3	3	0	0.005938	0.078863
…	…	…	…	…	…	…	…	…
8731	0	61	1	0	1	1	-0.034718	-0.01411

Subject	Y	X_1	X_2	X_3	X_4	X_5	X_6	X_7
8732	0	21	0	1	1	1	-0.034718	-0.01411

3.2 Multicollinearity Test

The multicollinearity test was performed by referring Pearson-Spearman correlation score. The correlation score of each variable can be seen in Table 5.

Table 5. Pearson-Spearman Correlation Score

Corr.	Y	X_1	X_2	X_3	X_4	X_5	X_6	X_7
Y	1.00	-0.17	-0.00	0.09	0.08	-0.26	0.03	0.01
X_1	-0.17	1.00	0.02	-0.28	-0.12	0.17	-0.03	-0.00
X_2	-0.00	0.02	1.00	-0.01	0.25	0.01	0.01	0.03
X_3	0.09	-0.28	-0.00	1.00	0.04	-0.19	0.18	0.41
X_4	0.08	-0.12	0.25	0.04	1.00	-0.05	0.05	0.08
X_5	-0.26	0.17	0.01	-0.12	-0.06	1.00	-0.00	0.09
X_6	0.03	-0.03	0.01	0.18	0.05	-0.00	1.00	0.21
X_7	0.01	-0.00	0.03	0.41	0.08	0.09	0.21	1.00

Table 5 shows that there was no correlation between variables higher than 0.6, meaning that the assumption of non-multicollinearity was fulfilled

3.3 Bayesian Logistic Regression

Bayesian method aims to obtain posterior distribution from the multiplication of prior distribution and likelihood. Several previous studies using the Bayesian method often used non-informative prior to determining prior distribution, considering that there has not been any prior knowledge regarding the parameter distribution. With the absence of this parameter, previous studies referred to normal distribution. This type of distribution was picked due to the two parameters, namely mean (μ) showing the true parameter score and standard deviation (σ) showing the uncertainty of the score of a parameter. Therefore, in this research, the prior was determined to have a normal distribution with the mean zero and variance 1. In this research, Gibbs sampling was used as the MCMC algorithm with the iteration of 1,000,000 + 500,000 burn in + 4 thin.

3.3.1 Convergence Test

In the Bayesian method, a convergence test is required to find out if the generated score is in accordance with the posterior distribution. The convergence test in MCMC used a trace plot, MC Error, and autocorrelation plot. The convergence test using MC Error is presented in Table 6.

Table 6. Convergence Test using MC Error

Parameter	SD	1% SD	MC Error	Result
β_0	0.2162	0.0022	0.00165	Convergence
β_1	0.0041	0.0000	0.00003	Convergence
β_2	0.0990	0.001	0.00030	Convergence
β_3	0.0588	0.0006	0.00028	Convergence
β_4	0.0622	0.0006	0.00032	Convergence
β_5	0.0974	0.001	0.00030	Convergence
β_6	0.17070	0.00171	0.00052	Convergence
β_7	0.02917	0.00029	0.00010	Convergence

The trace plot from the analysis result is presented in Figure 1.

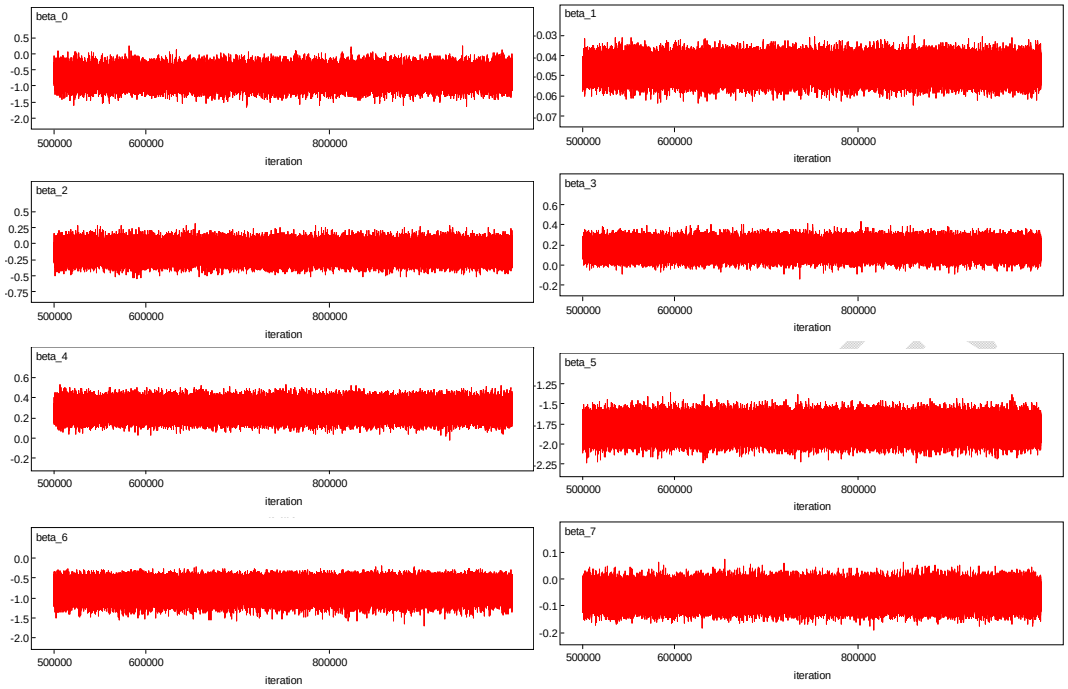

Fig. 1. Trace Plot

Figure 1 shows that the trace plot did not show any strong pattern or periodicity. The autocorrelation plot is presented in Figure 2.

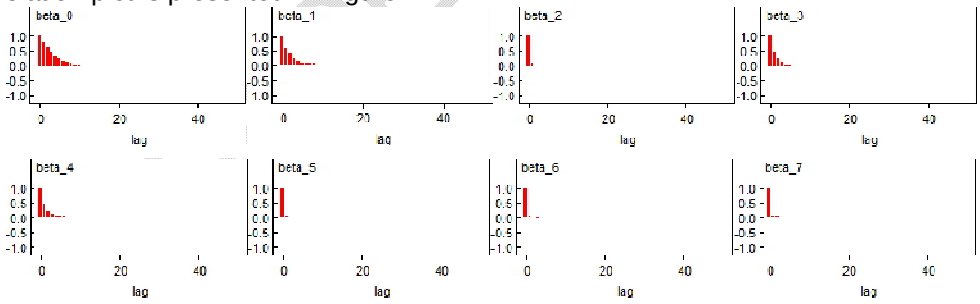

Fig. 2. Autocorrelation Plot

Figure 2 shows that autocorrelation between parameters was low, resulting in generated independent samples. According to the result of the convergence test of MCMC, each parameter was convergent or the generated samples were from the expected posterior distribution

3.3.2 Parameter Significance Test

In a Bayesian method, a parameter significance test is performed by examining the credible interval. A parameter is deemed to be significant if the credible interval does not show zero in the percentile interval of 2.5% to 97.5%. The credible interval for each parameter is presented in Table 7.

Table 7. Credible Interval

Parameter	2.50%	97.50%	Result
β_0	-1.140	-0.293	Significant
β_1	-0.055	-0.039	Significant
β_2	-0.321	0.067	Insignificant
β_3	0.046	0.277	Significant
β_4	0.159	0.403	Significant
β_5	-1.997	-1.615	Significant
β_6	-1.134	-0.469	Significant
β_7	-0.119	-0.005	Significant

Table 7 shows that of the seven variables used, only one variable was proven insignificant based on the credible interval, namely variable $\boldsymbol{X_2}$ or sex, while all the six other variables gave significant results based on credible interval.

3.3.3 Classification Accuracy

The accuracy of a model in classifying data is useful to find out the goodness of a Bayesian logistic regression model. The higher the classification of a model is formed, the better the model is obtained. This accuracy in the binary logistic regression is presented in Table 8.

Table 8. Classification Accuracy

Classification Accuracy		Prediction Class		Precisely Predict
		Not as recent migrants	Recent Migrants	
Actual Class	Not as recent migrants	6554	10	0.9985
	Recent migrants	390	16	0.0394
		%		**0.9426**

This table indicates that the model can precisely predict the research subject not as recent migrants, accounting for 6,554 or 99.85% (sensitivity) and this model can precisely predict the recent migrants as the research subject for as much as 16 or 3.9% (specificity). Overall, this model can give an accurate prediction of 94.26%.

In this research, the ROC curve was used to test the relevance of the model used in addition to the analysis using a classification table. The ROC curve of the analysis result is presented in Figure 3.

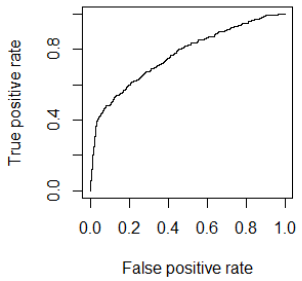

Fig. 3. Kurva ROC

Figure 3 indicates that the model is relevant since the curve generated was close to one. This is in line with the area under the curve or commonly abbreviated as AUC for as much

as 0.768. Of this score, this model is deemed to be appropriate to explain the model with a fair discrimination category.

3.3 Interpretation

The analysis result based on the odds Ratio is presented in Table 9:

Table 9. Odds

Variable	Category		Odds ratio
X1	17 years old (reference)		1 (reference)
	18 years old		0.9552 (0.9499, 0.9606)
X2	0	Men	1 (reference)
	1	Women	0.8808 (0.88, 0.8751)
X3	0	Not going to school	1 (reference)
	1	Primary/Secondary school	1.94 (1.11, 3.41)
	2	High school	4.79 (2.79, 8.23)
	3	University	4.425 (2.42, 7.46)
X4	0	Others	1 (reference)
	1	Studying	1.57 (0.58, 4.27)
	2	Working	8.71 (3.18, 23.8)
	3	Taking care of household	2.35 (0.85, 6.48)
X5	0	Not under their ownership	1 (reference)
	1	Under their ownership	0.14 (0.12, 0.17)
X6	Housing quality score -0.1928 (lowest)		1 (reference)
	Housing quality score -0.1428		0.964 (0.961, 0.967)
X7	Asset ownership score -0.026 (lowest)		1 (reference)
	Asset ownership score 0.024		0.9975 (0.9972, 0.9979)

4. CONCLUSION

This research concludes that the Bayesian approach in logistic regression with iteration 1,000,000, 4 thinning interval, and 500.000 burn-in indicates that of seven variables, six variables with significant influences on the status of recent migrants entering the Province of Yogyakarta consist of age (X_1), latest education (X_3), main activities (X_4), house ownership (X_5), housing quality (X_6), and asset ownership (X_7). Of these six variables, the younger resident, the resident with high school or equal as their latest education, the resident currently working as their main activity, the resident renting a house, and the resident with high housing quality score and high asset ownership score are more likely to do recent migration to the Province of Yogyakarta.

This research is expected to give merit to the local government to anticipate matters regarding migration to the Province of Yogyakarta such as the improvement of school systems and facilities, procurement of student-friendly public transport, and many more. Furthermore, future researchers can examine through the point of view of economy that

influence migrants to migrate to DI Yogyakarta Province, which has the lowest Province Minimum Wage (UMP), such as its comfort of life, low cost of living, etc.

Suggestion for further research is to use rare event logistic regression analysis to reduce the bias level in the prediction of minority data, or latent class regression analysis can also be considered as another option.

ETHICAL APPROVAL

All authors hereby declare that all experiments have been examined and approved by the appropriate ethics committee and have therefore been performed in accordance with the ethical standards laid down in the 1964 Declaration of Helsinki.

REFERENCES

1. Hosmer, D. W., & Lemeshow, S. (2013). *Applied Logistic Regression* (Third Edition). John Wiley & Sons.
2. Solimun and Fernandes, A.A.R. (2017), Investigation the mediating variable: What is necessary? (case study in management research), International Journal of Law and Management, 59(6), 1059-1067
3. Owen, A. B. (2007). Infinitely Imbalanced Logistic Regression. In *Journal of Machine Learning Research* (Vol. 8).
4. King, G., & Zeng, L. (2001). Logistic Regression in Rare Events Data. *Political Analysis*, *9*(2), 137–163.
5. Lee, E. S. (1976). *Teori Migrasi*. Pusat Penelitian Kependudukan UGM.
6. Syairozi, M., & Wijaya, K. (2020). Migrasi Tenaga Kerja Informal: Studi Pada Kecamatan Sukorejo Kabupaten Pasuruan. *Seminar Nasional Sistem Informasi (SENASIF)*, *4*(1), 2383–2394.
7. Dustmann, C., & Glitz, A. (2011). Migration and Education. In *Handbook of the Economics of Education* (pp. 327–439). Elsevier. https://doi.org/10.1016/B978-0-444-53444-6.00004-3
8. Sarmita, I. M., & Simamora, A. H. (2018). Karakteristik Sosial Ekonomi Dan Tipologi Migrasi Migran Asal Jawa Di Kuta Selatan-Bali. *Jurnal Ilmiah Ilmu Sosial*, *4*(2).
9. Tjiptoherijanto, P. (2009). Dimensi Kependudukan Dalam Pembangunan Berkelanjutan. Bappenas.
10. Mantra, I. B. (2012). *Demografi Umum*. Pustaka Pelajar.
11. Dillon, R., & Goldstein, M. (1984). *Multivariate Analysis: Methods and Applications*. John Wiley & Sons.
12. Gifi, A. (1981). *Nonlinear Multivariate Analysis*. Universitas Leiden.
13. Markos, A. I., Vozalis, M. G., & Margaritis, K. G. (2010). An Optimal Scaling Approach to Collaborative Filtering Using Categorical Principal Component Analysis and Neighborhood Formation. In *IFIP Advances in Information and Communication Technology* (pp. 22–29). Springer. https://doi.org/10.1007/978-3-642-16239-8_6
14. Meulman, J., van der Kooij, A., & Heiser, W. (2004). Principal Components Analysis With Nonlinear Optimal Scaling Transformations for Ordinal and Nominal Data. In *Handbook of Quantitative Methodology for the Social Sciences* (pp. 50–71).

SAGE Publications, Inc. https://doi.org/10.4135/9781412986311.n3

15. Rachev, S. T., Hsu, J. S., Bagasheva, B. S., & Fabozzi, F. J. (2008). *Bayesian Methods in Finance*. John Wiley & Sons.

16. Ghosh, J. K., Delampady, M., & Samanta, T. (2007). *An Introduction to Bayesian Analysis: Theory and Methods*. Springer Science and Business Media.

17. Liu, H., & Powers, D. A. (2012). Bayesian Inference for zero inflated poisson regression modes. . *Journal of Statistics : Advance in Theory and Aplications*, *7*(2), 155-188.

18. Genkin, A., Lewis, D. D., & Mandigan, D. (2007). Large Scale Bayesian Logistic Regression for Text Categorization. *Technometrics*, *49*(3), 291–304.

19. Walpole, R. E., Myers, R. H., Myers, S. L., & Ye, K. (2012). *Probability & Statistics for Engineers & Scientist* (9th ed.). Pearson Education, Inc.

20. Ntzoufras, I. (2011). *Bayesian Modeling Using WinBUGS* . John Wiley & Sons.

21. Azen, R., & Walker, C. M. (2011). Categorical data analysis for the behavioral and social sciences. Routledge.

22. Zaiceva, A. (2014). The impact of aging on the scale of migration. *IZA World of Labor*.

23. United Nations General Assembly. (2019). The impact of migration on migrant women and girls: a gender perspective.

24. The European Institute for Gender Equality (EIGE). (2020). Gender mainstreaming Sectoral Brief: Gender and Migration. *The European Institute for Gender Equality*.

25. Dustmann, C., & Glitz, A. (2011). Migration and education. In *Handbook of the Economics of Education* (Vol. 4, pp. 327-439). Elsevier.

26. Buffel, T., Handler, S., & Phillipson, C. (2018). *Age-Friendly Cities and Communities*. Policy Press: Bristol, UK.

27. Hagen-Zanker, J., & Mallett, R. (2016). Journeys to Europe. *The Role of Policy in Migrant Decision-Making ODI Insights*.

28. Helderman, A. C., Ham, M., & Mulder, C. H. (2006). Migration and Home Ownership. *Tijdschrift Voor Economische En Sociale Geografie*, *97*(2), 111–125. https://doi.org/10.1111/j.1467-9663.2006.00506.x