International Journal of Innovative Technology and Exploring Engineering (IJITEE)
ISSN: 2278-3075 (Online), Volume-9 Issue-12, October 2020

B40 Group Income Household Trend in Malaysia

Humaida Banu Samsudin, Norsyasya Aina Mohd Mokhtar

Abstract*: Income inequality is crucial issue in the Malaysian economy. This issue has a great impact especially on the B40 group income household because of the rising cost of living today. Therefore, modelling of income data is done to look at income pattern of B40 group in Malaysia. Household income data for Malaysia in year 2007, 2009, 2012, 2014 and 2016 have been used in this study. The income distribution used in this study is a two-parameter distribution of Weibull, Log Normal, Fisk and Gamma. This study uses only two parametric distributions to suit the income data because the simplest model is better than the complex model. The best distribution selection is performed with the fitting of statistical distribution through maximum likelihood estimation (MLE) method. Goodness of fit test has been done to model B40 household income data. The best model for each year used to predict the average income in the future by using regression method. Weibull distribution is the best model for B40 household income data. The study also shows that the average income of the B40 group in the future will increase. Therefore, this study was conducted to assist B40 group to be more sensitive to the Malaysian economy and plan their income wisely.

***Keywords*: B40 group income, goodness of fit test, income distribution, income inequalities**

I. INTRODUCTION

Income inequality is an issue that constantly debated in the economy. Based on the Department of Statistics Malaysia's information, income levels for each group have increased from year to year. This increment was due to economic growth in Malaysia. According to [1], economic growth has affected the inequality of the population in Malaysia. They also stated that the higher the Gross Domestic Product (GDP), the higher the income inequality. [2] also proves that there is an increase in income inequality in Malaysia.

The increase in inequality makes it difficult for the B40s to sustain their lives in the high cost of living. The B40 or Bottom 40 is known as the lowest income group in Malaysia. This group can be defined as a household with average monthly income below RM3860. Therefore, this study can help these lower income groups to understand the Malaysian economy by analyzing the distribution of household income.

The earliest model in measuring income distribution was introduced by Vilfredo Pareto through one-parameter distribution of Pareto distribution [3]. However, fitting through Pareto distribution is generally suitable for upper tail income but this distribution is not suitable for fitting the overall range of income data [4]. Due to the constraints of one-parameter distribution, [5] proposed two-parameter distribution which is Log Normal. Next, there are other two-parametric distributions recommended after the Log Normal distributions which are Fisk distribution [6], Gamma distribution [7] and Weibull distribution [8].

Three-parameter distribution such as Singh-Maddala [9], Dagum [10] and generalized Gamma [11] were introduced and classified as special cases for the distribution of Pareto, Fisk, Normal Log, Gamma and Weibull. [12] introduced the four-parameter model of the first generalized beta distribution (GB1) and the second type of generalized beta (GB2). All of these distributions were used to analyze the distribution of household income data set. [12] used the above distributions for the best distribution fitting to the United States family income data set in 1970, 1975 and 1980. He concluded that the GB2 distribution provided the best fit and the Singh-Maddala distribution gave a better fitting compared to GB1 distribution and all the two-parameter distributions and the three-parameter distributions used in his study. Additionally, [13] have compared the size of income distribution for 23 countries. Their study shown that the GB2 distribution is the best model for a four-parameter distribution, the Dagum distribution is the best model for the three-parameter distribution and the Weibull distribution is the best model of the two-parameter distribution. Furthermore, [14] conducted a parametric comparative study to explore the impact of taxes and transfer payments on income data for 13 countries with different years. They found that the Weibull distribution is the best two-parameter distribution of income data. For the three-parameter distribution, the Dagum distribution is the best distribution of income data. The GB2 distribution is best fitted to income data compared to GB1. [15] also conducted studies on parametric modeling for household income in Punjab, Pakistan. The data used are individual income per capita data in 2003-2004 and 2007-2008. The goodness of fit test is used to determine the most appropriate distribution model according to the data. This study found that distribution of GB2 was the most suitable distribution of income data in 2003-2004 and 2007-2008.

Revised Manuscript Received on October 30, 2020.
* Correspondence Author
Humaida Banu Samsudin*, Actuarial Science Programme, Faculty of Science and Technology, Universiti Kebangsaan Malaysia, 43600 Bangi, Selangor, Malaysia. Email: humaida@ukm.edu.my
Norsyasya Aina Mohd Mokhtar, Actuarial Science Programme, Faculty of Science and Technology, Universiti Kebangsaan Malaysia, 43600 Bangi, Selangor, Malaysia. Email: syasyamokhtar97@gmail.com

Therefore, this study aims to identify the suitable distribution of B40 household income data in Malaysia for year 2007, 2009, 2012, 2014 and 2016.

II. METHODOLOGY

A. Income Distribution

The income distribution used in this study is a two-parameter distribution of Weibull, Log Normal, Fisk and Gamma. This study uses only two parametric distributions to suit the income data because the simplest model is better than the complex model [16].

B. Quantile-quantile plot (Q-Q plot)

The Q-Q plot is a graphical tool that often used by researchers to assess whether a set of data comes from the theoretical distribution or not. The advantages of using the graphic method through Q-Q plot is the size of the sample data need not be the same [17]. Therefore, this method is applicable for this study because the size of household income data for year 2007, 2009, 2012, 2014 and 2016 is not same. However, this graphic method does not show strong evidence that such data is having the same distribution as the theoretical distribution. Therefore, statistical test criteria are performed to reinforce the distribution assumption through the Q-Q plot.

C. Maximum likelihood estimation (MLE)

Maximum likelihood estimation is used to estimate the parameter of predicted distribution from Q-Q plot. Maximum likelihood estimation variables, in a given function are defined as value that maximise the likelihood [18]. Therefore, MLE is done to assess whether the parameter's value corresponds to the distribution function or not.

D. Log likelihood

The log likelihood will be maximized to determine the optimum value of the estimated coefficients [19]. Therefore, this method can be used to compare the sensitivity and different fitted coefficients [19]. The large log likelihood will result in a better distribution fitting because this method maximizes the likelihood function.

E. Akaike's Information Criteria (AIC)

AIC is one of the most commonly used measures for comparing certain models. AIC is designed to select models that produce probability distribution with the smallest difference from the actual distribution. Therefore, the low AICs provide a better model [20].

F. Bayesian Information Criteria (BIC)

BIC is one of the methods to compare models. BIC is an approximation of the posterior probability function for a model that is considered true, under certain Bayesian supplies, so that lower BICs are more likely to be the real model [21].

G. Kolmogorov-Smirnov (K-S)

The Kolmogorov-Smirnov test is used in this study to evaluate the probability distribution of household income data whether the sample data comes from a population with a certain distribution. This test compares the cumulative distribution of observation data and fitted data [22].

H. Anderson-Darling (A-D)

Anderson-Darling test is used to assess whether the sample data is derived from a population with a specific distribution. This test is modified from the K-S test by focusing more on the tail part of the K-S test [23].

I. Cramer von Misses (CvM)

CvM is one of the alternatives to the K-S test. This CvM criterion is used to assess the goodness of fit by comparing the cumulative distribution function, $F(x)$ with empirical distribution function, $F_n(x)$.

J. The weight of ranks

The goodness of fit test will be compared for each statistical distribution by using the weight of ranks. This method works by comparing each distribution using rank [24]. The distribution of the highest rank weights is the most fitted statistical distribution to the data. This rank begins with the highest criterion value. The highest rank is for the lowest criterion value. The lowest criterion values always show the most fit distribution to sample data [24]. The formula for the weight of ranks is defined as follows:

$$w_i = \frac{1}{k}\sum_{i=1}^{k} r_i \quad , \quad i \le w_i \le k$$

Where r_i is rank, k is the number of criteria used, dan w_i is the relative weight of ranks [24].

In this study, the AIC, BIC, K-S, A-D and CvM test values will be compared between the distributions. The lowest value indicates that the data suits to the theoretical distribution. Log likelihood values are chosen based on the highest values that also show good distribution fitting. The income distribution can be applied to look at the pattern of household income in Malaysia.

K. Income trend

The changes in parameter of the best distribution of household income will lead to the change in the average household income. As a result, the expected future average income in year 2017, 2018, 2019 and 2020 are predictable. This forecast is done using the simple linear regression method between the income distribution parameters with time. The formula for regression equation is:

$$y = \beta_0 + \beta_1 t$$

where, y is the parameter to be predicted, t is year, β_0 is the y-intercept dan β_1 is a slope [25].

Hypothesis testing was also conducted to evaluate either the regression equation is significant or not. This regression linear equation is said to be significant if the *p*-value is less than the significance level (α). The significance level used in this study is 0.05. The significant regression results prove that there is enough evidence to reject the null hypothesis [25]. The forecast parameter value of the best income distribution will be used to predict the average income of B40 households in year 2017, 2018, 2019 and 2020 by using the mean of the best income distribution function.

III. RESULT AND DISCUSSION

A. The best distribution selection

The best distribution selection has been performed with the fitting of statistical distribution through MLE method.

The MLE will be used in the goodness of fit test. Each distribution has been fitted to the income data set and goodness of fit test has been calculated. Based on Table-I, in year 2007, 2009, 2012, 2014 and 2016, the Weibull distribution yields the highest maximum likelihood value and the lowest AIC, BIC, K-S, A-D and CvM values. This led to the highest weights of ranks as compared to other distributions. Therefore, the Weibull distribution is the income distribution for the B40 group for each year involved. This result is also supported by Q-Q plot that shows Weibull distribution is the most suitable for B40 income data.

K-S, A-D, and CvM are goodness of fit tests that use hypothesis testing to determine whether the data is distributed as the desired distribution. The null hypothesis for this test is a set of data that is distributed by a certain distribution. However, K-S, A-D, and CvM values for each distribution rejected the null hypothesis at every level of confidence. This indirectly indicates that the data does not follow the desired distribution. According to [12], the rejection of the null hypothesis is due to large sample size of data. Thus, the comparison of the goodness of fit test values between the distributions is done to obtain the best distribution for income data.

Table-I: Goodness of fit test for each distribution in 2007, 2009, 2012, 2014 and 2016.

Year	Distribution	Log Likelihood	r1	AIC	r2	BIC	r3	K-S	r4	A-D	r5	CvM	r6	W
2007	Lognormal	-36538.89	1	73081.77	1	73094.75	1	0.12366	1	136.35028	1	22.62852	1	1
	Gamma	-36156.43	3	72316.86	3	72329.84	3	0.09426	3	86.37886	3	13.88064	2	2.8
	Weibull	**-35733.09**	**4**	**71470.18**	**4**	**71483.16**	**4**	**0.04340**	**4**	**30.84942**	**4**	**3.80859**	**4**	**4**
	Fisk	-36341.8	2	72687.59	2	72700.57	2	0.11230	2	89.22892	2	9.26361	3	2.2
2009	Lognormal	-39457.8	1	78919.6	1	78932.7	1	0.09722	2	106.71710	1	16.21080	1	1.2
	Gamma	-39104.25	3	78212.49	3	78225.59	3	0.07310	3	60.59981	3	8.59763	2	2.8
	Weibull	**-38750.64**	**5**	**77505.28**	**4**	**77518.38**	**4**	**0.03903**	**4**	**18.82528**	**4**	**1.96940**	**4**	**4**
	Fisk	-39295.58	2	78595.16	2	78608.26	2	0.09936	1	65.80360	2	5.70692	3	2
2012	Lognormal	-41469.6	1	82943.21	1	82956.36	1	0.08715	2	74.55198	1	11.02057	1	1.2
	Gamma	-41194.78	3	82393.57	3	82406.71	3	0.06630	3	46.18007	3	6.62170	3	3
	Weibull	**-40950.91**	**4**	**81905.82**	**4**	**81918.97**	**4**	**0.04636**	**4**	**27.10076**	**4**	**3.83977**	**4**	**4**
	Fisk	-41435.36	2	82874.72	2	82887.87	2	0.09432	1	61.09109	2	6.82318	2	1.8
2014	Lognormal	-80307.87	1	160619.7	1	160634.1	1	0.11061	2	226.03017	1	34.94941	1	1.2
	Gamma	-79740.6	3	159485.2	3	159499.6	3	0.08976	3	162.34122	3	24.37797	2	2.8
	Weibull	**-79094.5**	**4**	**158193.0**	**4**	**158207.4**	**4**	**0.08794**	**4**	**105.95770**	**4**	**15.17967**	**4**	**4**
	Fisk	-80240.74	2	160485.5	2	160499.9	2	0.12002	1	182.83527	2	21.35421	3	2
2016	Lognormal	-77824.55	1	155653.1	1	155667.4	1	0.10238	2	205.10611	1	32.61700	1	1.2
	Gamma	-77326.25	3	154656.5	3	154670.8	3	0.09562	3	146.46230	3	22.86730	2	2.8
	Weibull	**-76739.57**	**4**	**153483.1**	**4**	**153497.4**	**4**	**0.07570**	**4**	**88.12408**	**4**	**13.1935 7**	**4**	**4**
	Fisk	-77806.63	2	155617.3	2	155631.6	2	0.11258	1	170.04765	2	20.48668	3	2

B. Income Trend

The parameters for the Weibull distribution are the shape parameters (β) and the scale parameters (α). Changes in shape and scale parameters will affect the average change in income. This can be seen in Figure 1.

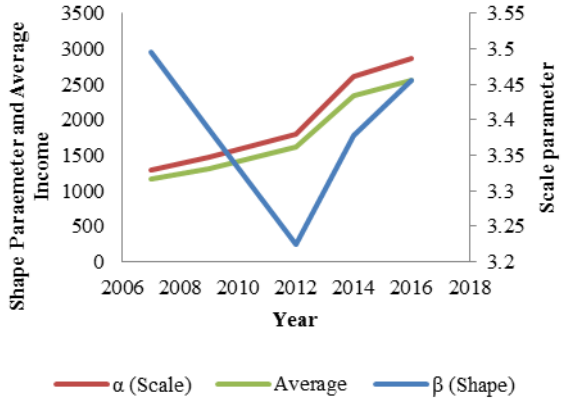

Fig. 1. The graph of scale parameter changes, shape parameters and average household income from 2007 to 2016.

Predictions for shape and scale parameters can be made through regression. The regression results are shown in Table-II. The results of Table-II show that the relationship between scale parameters with time and shapes parameter with time is significant at the significance level. Subsequently, the forecasts for the average income in year 2017, 2018, 2019 and 2020 can be made through Weibull's average income formula:

$$average = \alpha\Gamma\left(1+\frac{1}{\beta}\right)$$

Where $\Gamma\left(1+\frac{1}{\beta}\right)$ is gamma function for $1+\frac{1}{\beta}$.

Table-II: Table of linear equations and *p* values for parameter β and α.

Parameter	Linear Equations	*p*-value
β	$\frac{\beta}{t^2} = -0.553 + 3.95\frac{1}{t}$	0.001
α	$\alpha = 978 + 184t$	0.007

The forecast results are shown in Table-III. The table shows the average income of the B40 group increasing in the future.

Table-III: Table of forecast β and forecast α for year 2017, 2018, 2019 and 2020

Time	Year	Forecast β	Forecast α	Forecast Average (RM)
11	2017	-27.086	3001.741	3069.90
12	2018	-32.234	3185.717	3245.89
13	2019	-42.110	3369.694	3417.81
14	2020	-53.091	3553.670	3593.58

IV. CONCLUSION

In this study, the distribution of household income in Malaysia for year 2007, 2009, 2012, 2014 and 2016 was built. The comparison of the statistical criteria values has been done. The results show that the Weibull distribution is the best distribution that corresponds to the B40 household income data in Malaysia from year 2007 to 2016. The Weibull distribution has the highest log likelihood value and the lowest AIC, BIC, KS, AD and CvM values which in turn produces the highest weight of ranks. These results are used in income trend that shows an increasing pattern for the average B40 income in the future.

REFERENCES

1. Nor Fatimah Che Sulaiman, Ishak Yussof, Mohd Azlan Shah Zaidi & Noorasiah Sulaiman. 2017. Long run relationship between income inequality and economic growth : Evidence from Malaysia 7(6): 73–88.
2. Fatimah Abdul Razak & Faridatulazna Ahmad Shahabuddin. 2018. Malaysian household income distribution: A Fractal Point of View. Sains Malaysiana 47(9): 2187–2194
3. Pareto, V. 1897. Cours d'Économie Politique. The Annals of the Academy of Political and Social Science 9(3): 128-131. Lausanne: F. Rouge.
4. Boccanfuso, D., Richard, P. & Savard, L. 2013. Parametric and nonparametric income distribution estimators in CGE micro-simulation modeling. *Economic Modelling* 35: 892–899.
5. Gibrat, R. 1931. Les Inégalités économiques. Paris : Recueil Sirey.
6. Fisk, P.R. 1961. The graduation of income distributions. Econonometrica: Journal of the Econometric Society 29(2): 171–185.
7. Salem, A.B.Z. & Mount, T.D. 1974. A Convenient descriptive model of income distribution: The Gamma Density. *Econometrica: Journal of the Econometrica Society* 42(6): 1115–1127.
8. Bartels, C.P. & Van Metelen, H. 1975. Alternative probability density functions of income. *Vrije University Amsterdam: Research Memorandum No. 29*, hlm. 30.
9. Singh, S.K. & Maddala, G.S. 1976. A function for size distribution of incomes. *Econometrica: Journal of the Econometrica Society* 44(5): 963–970.
10. Dagum, C. 1977. A new model of personal income distribution: Specification and estimation. *Economie Appliquée* 30: 413–437.
11. Taillie, C. 1981. Lorenz ordering within the generalized Gamma family of income distributions. *Statistical Distribution in Scientific Work: Volume 6- Application in Physical, Social, and Llife Sciences*, hlm. 181–192.
12. McDonald, J.B. 1984. Some generalized functions for the size distribution of income. *Econometrica: Journal of the Econometrica Society* 52(3): 647–665.
13. Bandourian, R., McDonald, J.B. & Turley, R.S. 2002. A Comparison of Parametric Models of Income Distribution Across Countries and Over Time.
14. Dastrup, S.R., Hartshorn, R. & McDonald, J.B. 2007. The impact of taxes and transfer payments on the distribution of income: A parametric comparison. *The Journal of Economic Inequality* 5: 353–369.
15. Shakeel, M., Hussain, I., Arif, M.M., Ameen, M. & Haq, M.A. ul. 2015. Parametric modeling of household income distribution in the Punjab, Pakistan. *Sci.Int.(Lahore)* 27(5): 4161–4170.
16. Klugman, S.A., Panjer, H.H. & Willmort, G.E. 2012. Model selection. Dlm. J.Balding, D., Cressie, N. A., Fitzmaurice, G. M., Goldstein, H., Johnstone, I. M., Molenberghs, G., Scott, D. W., Smith, A. F. M., Tsay, R. S., & Weisberg, S. (pnyt.). *Loss Models: From Data to Decisions*, hlm. 323–350. Edisi ke-4. John Wiley & Sons, Inc: United States of America.
17. NIST/SEMATECH e-Handbook of Statistical Methods. 2012. https://www.itl.nist.gov/div898/handbook/eda/section3/eda33o.htm [1 Mei 2019].
18. Myung, I.J. 2003. Tutorial on maximum likelihood estimation. *Journal of Mathematical Psychology* 47: 90–100.
19. Anon. 2019. Regression models: What is log-likelihood. https://support.minitab.com/en-us/minitab/18/help-and-how-to/modeling-statistics/regression/supporting-topics/regression-models/what-is-log-likelihood/ [1 Mei 2019].
20. Akaike, H. 1974. A new look at the statistical model identification.
21. *IEEE Transactions on Automatic Control* 19(6): 716–723.
22. Stone, M. 1979. Comments on model selection criteria of Akaike and Schwarz. *Journal of the Royal Statistical Society. Series B: Methodological* 41(2): 276–278.
23. Frank J. Massey, J. 1951. The Kolmogorov-Smirnov test for goodness of fit. *Journal of the American Statistical Association* 46(253): 68–78.
24. Nelson, L.S. 1998. The Anderson-Darling test for normality. *Journal of Quality Technology* 30(3): 298–299.
25. Al-Dhurafi, N.A., Ahmad Mahir Razali, Nurulkamal Masseran & Zamira Hasanah Zamzuri. 2016. The probability distribution model of air pollution index and its dominants in Kuala Lumpur. *THE 2016 UKM FST POSTGRADUATE COLLOQUIUM: Proceedings of the Universiti Kebangsaan Malaysia, Faculty of Science and Technology 2016 Postgraduate Colloquium*. American Institute of Physics: Selangor.
26. Bowerman, B.L., O'Connell, R.T. & Koehler, A.B. 2005. Simple linear regression. Dlm. Bowerman, B. L., O'Connell, R. T., & Koehler, A. B. (pnyt.). *Forecasting, Time Series, and Regression*, hlm. 79–125. Edisi ke-4. Curt Hinrichs: United States of America.

AUTHORS PROFILE

Humaida Banu Samsudin (Dr)*, Bachelor of Science (Hons) Actuarial Science, National University of Malaysia (UKM). MSC in Risk Management and Insurance, City University, London, UK. PhD in Risk Mangement, University of Salford, Greater Manchester, UK. Work place: Senior Lecturer in Actuarial Science Program, Faculty of Science and Technology, National University of Malaysia (UKM), Bangi, Selangor, Malaysia. Area of Expertise: Risk management, Insurance, Mortality and Survival studies/analysis. Membership: PERSAMA (Mathematical Society of Malaysia – Vice Treasury), MARIM (Malaysian Association of Risk and Insurance Management – member). Email: humaida@ukm.edu.my.

Norsyasya Aina Mohd Mokhtar obtained her Bachelor of Science (Hons) Actuarial Science from the National University of Malaysia (UKM) in year 2019. Besides studying, she also involved herself in many of the activities, not only limited to those related to Actuarial Science Club of UKM, but also in education volunteering. As a fresh graduate, she is currently working as a risk management executive in a local statutory team of a foreign insurance company. Her current job is mainly focusing on the operational risk portfolio – risk control self-assessment (RCSA), preparing and validating the data for Bank Negara Malaysia submission.