Journal of Statistics and Its Application, Likelihood Vol 1 No 2, December 2025

Best Regression Model Selection for Factors Affecting Life Expectancy in Indonesian Provinces in 2024

Muhammad Zulfadhli [1]

[1] Institut Teknologi Sepuluh Nopember, Indonesia, email: muhammadzulfadhli23@gmail.com

Abstract

Life expectancy reflects the welfare and quality of life of the population. BPS data shows an increase in Indonesia's life expectancy from 73.39 years to 74.15 years (+0.298%). Factors that are thought to influence this include average net salary, expected length of schooling, poverty severity index, and GRDP. The analysis was conducted using multiple linear regression, then the best model was selected using Forward Elimination, Backward Elimination, and Stepwise Selection. This study used secondary data obtained from publications and BPS statistical tables, all of which had a ratio measurement scale. There are differences in life expectancy between provinces, as well as high economic inequality. There are no indications of multicollinearity. The best model includes expected years of schooling and GRDP with residuals that meet the assumptions of normality, independence, and homoscedasticity. The model is not yet adequate because it can only explain less than half of the variability in life expectancy in Indonesia in 2024.

Keywords: *Backward Elimination, Forward Elimination, Multiple Linear Regression, Stepwise Selection, Life Expectancy*

Introduction

Every human being has a lifespan that has been determined by God since conception, more commonly referred to as age. In population studies, however, age is not only understood as an individual biological condition, but also as part of a broader demographic indicator that reflects the health and welfare of a society. One of the most widely used indicators in this context is life expectancy. Life expectancy refers to the estimated average number of years a person is expected to live under prevailing mortality conditions. Therefore, life expectancy does not merely describe survival, but also reflects the quality of public health, access to education, economic opportunity, and the overall standard of living of a population. In this sense, life expectancy can illustrate the level of welfare and quality of life of a population (Țarcă et al., 2024).

In many developing countries, including Indonesia, life expectancy is often used as an important benchmark in evaluating development outcomes. A higher life expectancy generally indicates better living conditions, improved nutrition, broader access to health services, and stronger social support systems. For this reason, life expectancy is also closely related to the Human Development Index and is frequently discussed in studies of public policy, regional inequality, and socio-economic development. Based on statistics obtained from Statistics Indonesia, the life expectancy of Indonesians increased by 0.298 percent from 73.39 years in the previous year to 74.15 years in 2024 (Badan Pusat Statistik, 2024). This increase is encouraging, yet it should not be interpreted as evidence that all provinces in Indonesia experience the same level of progress. In reality, disparities in life expectancy across provinces remain an important issue, indicating that the determinants of life expectancy may vary substantially across regions.

Changes in life expectancy are thought to be influenced by various social and economic factors, including the average monthly net salary of workers, expected length of schooling, poverty severity index, and gross regional domestic product (GRDP). These variables are relevant because they represent different but interconnected dimensions of human development. Income, for example, can affect people's ability to access nutritious food, health care, and decent housing.

https://doi.org/10.30605/likelihood.1123

Education can shape health awareness, decision-making, and long-term productivity. Poverty severity reflects the depth of deprivation experienced by the poor and may capture unequal access to essential services more effectively than simple poverty incidence. Meanwhile, GRDP is often used to describe regional economic capacity and the extent to which economic activity supports social welfare. Previous studies have shown that socio-economic conditions play an important role in shaping life expectancy patterns, both within and across countries (Fadri, 2024; Kabir, 2008).

Because life expectancy is commonly assumed to have a linear relationship with several explanatory variables, the effect of these factors can be analyzed using multiple linear regression. Multiple linear regression is one of the most widely used statistical methods for explaining the relationship between a response variable and several predictor variables simultaneously. The magnitude and direction of the influence of each factor are commonly estimated using Ordinary Least Squares (OLS), a method that provides parameter estimates by minimizing the sum of squared residuals (Wooldridge, 2020). The use of multiple linear regression and OLS is attractive because the method is relatively simple, easy to interpret, and effective for identifying significant predictors when the classical assumptions are reasonably satisfied. In studies of life expectancy, regression-based approaches remain relevant because they allow researchers to quantify the contribution of each socio-economic factor and assess whether the observed relationships are statistically meaningful (Jackowska et al., 2024).

Nevertheless, the use of multiple linear regression does not end with estimating one full model. When several candidate predictors are available, different combinations of variables may produce different regression models, and not all of them are equally efficient or substantively meaningful. A model with too many variables may become unnecessarily complex and difficult to interpret, whereas a model with too few variables may omit important information. For that reason, selecting the best regression model becomes an essential step in the analysis. From the many regression models that can be formed, the best model should ideally balance statistical adequacy, simplicity, and interpretability. In regression analysis, this issue is often addressed through variable selection procedures such as Backward Elimination, Forward Selection, and Stepwise Selection (Hastie et al., 2009; James et al., 2023).

Several previous studies have shown that regression model selection plays an important role in accurately identifying factors affecting life expectancy. Research on life expectancy in various countries, including Indonesia, commonly applies multiple linear regression with the OLS method due to its simplicity, interpretability, and effectiveness in explaining linear relationships between socio-economic variables such as income, education, and poverty (Țarcă et al., 2024). However, differences arise in the techniques used to select the best regression model. The Backward Elimination method starts with a full model and removes statistically insignificant variables one by one until only significant predictors remain. In contrast, Forward Selection begins with no predictors and gradually adds variables that provide the greatest contribution to the model. Stepwise Selection combines both approaches by allowing variables to be added or removed at different stages of the procedure, which makes it more flexible than using only backward or forward strategies.

Although these three methods are widely used, each of them has strengths and limitations. Backward Elimination is often useful when the number of candidate predictors is not too large and when the researcher wants to begin with a theoretically complete model. Forward Selection may be more practical when the researcher prefers a gradual construction of the model and wants to prioritize predictors with the strongest initial contribution. Stepwise Selection is often considered more adaptive because it evaluates the entry and removal of variables throughout the modeling process. However, variable selection procedures also receive criticism in statistical practice. Heinze et al. (2018) emphasized that automated selection methods may lead to instability, particularly when small changes in the data produce different selected models. Similarly, overly data-driven model selection may increase the risk of overfitting and reduce the substantive interpretability of the final model if not supported by theoretical reasoning. Therefore, comparing these three methods in the present study is important not only for obtaining a statistically good model, but also for identifying a parsimonious model that remains meaningful in explaining the determinants of life expectancy across Indonesian provinces in 2024.

This study is important for at least two reasons. First, from a substantive perspective, identifying the main factors that influence life expectancy can provide useful evidence for policy formulation, especially in reducing provincial disparities in welfare and human development. Second, from a methodological perspective, this study offers a direct comparison of three commonly used regression model selection methods in the context of provincial life expectancy data. Such a comparison is useful because it can show whether different procedures lead to the same final model or produce different conclusions regarding the most influential predictors. If the resulting model is consistent across methods, confidence in the selected predictors may become stronger. On the other hand, if the results differ, the comparison itself becomes valuable for understanding the sensitivity of model selection to methodological choice.

This study was conducted with the hope that the results would reveal the factors that influence life expectancy in Indonesian provinces in 2024 and identify a parsimonious model, namely a model that is simple but still meaningful for predicting life expectancy. In addition, this study is expected to provide comprehensive and detailed references regarding the selection of the best regression model, while also improving understanding of the advantages and disadvantages of various regression models that can be formed. Thus, this research contributes not only to the empirical discussion on life expectancy in Indonesia, but also to the practical application of multiple linear regression model selection in regional socio-economic analysis.

Method

The data analyzed in this study is secondary data obtained from the appendix of the 2024 education statistics publication by the Central Statistics Agency (BPS) and statistical tables available on the Indonesian Central Statistics Agency (BPS) website. The data consists of 1 response variable (y) and 4 predictor variables (x). Explanations of each variable are presented in Table 1.

Table 1. Research Variables

Variable	Description	Operational Definition	Measurement Scale	Unit
y	Life Expectancy	The estimated average length of time a person can live during their lifetime	Ratio	year
x_1	Old School Hopes (HLS)	Opportunities for the community to pursue formal education	Ratio	year
x_2	Average Net Salary of Workers	Average income earned from work after income tax is deducted	Ratio	Rupiah/ month
x_3	Poverty Severity Index (P2)	Disparity in expenditure among people living below the standard of living	Ratio	-
x_4	Gross Regional Domestic Product (PDRB)	Indicators of community income and describing the level of welfare enjoyed by residents as an output of economic activity	Ratio	Billion Rupiah

The selection of the best model among the regression models that can be formed by factors suspected of affecting the life expectancy of provinces in Indonesia in 2024 follows these steps, Collect data on life expectancy and expected years of schooling from the 2024 Human Development Index publication, and the average net wage of workers, poverty severity index, and gross regional domestic product obtained from the statistics table on the BPS website. Describe the characteristics of life expectancy, expected years of schooling, average net worker income, poverty severity index, and gross regional domestic product based on the mean, variance, maximum, and minimum values. Select the best model using the Forward Selection, Backward Elimination, and Stepwise Selection methods. Examine and test the classical assumptions of the components in the best model. Interpret the best model obtained based on the parameter estimate values for factors 1 to k. Test the significance of the model parameters partially to ensure that all factors are significant for life expectancy. Measure how well the best model is selected based on the coefficient of determination.

Results and Discussion

Descriptive Analytics

Descriptive analysis is used to determine the characteristics of the data for each variable as follows Table 2.

Table 2. Descriptive Statistics of Research Variables

Variabel	*Mean*	Varians	Min	Max
y	72,875	5,296	67,390	75,99
x_1	13,21	1,044	9,630	15,7
x_2	3318	644945	2365	5807
x_3	0,3329	0,0593	0,06	1,42
x_4	343844	2,804	13799	2151041

Berdasarkan Tabel 2, secara umum terlihat bahwa setiap variabel penelitian memiliki karakteristik sebaran data yang berbeda. Variabel terikat (Y) menunjukkan kecenderungan nilai yang relatif stabil, dengan rentang skor yang tidak terlalu lebar, sehingga mengindikasikan homogenitas hasil pengukuran pada subjek penelitian. Variabel X_1 memiliki nilai rata-rata yang cukup representatif dengan variasi data yang rendah, yang menandakan bahwa responden cenderung memiliki karakteristik yang hampir seragam pada variabel ini. Sementara itu, variabel X_2 dan X_4 memperlihatkan tingkat variasi data yang lebih tinggi dibandingkan variabel lainnya, yang menunjukkan adanya perbedaan karakteristik responden yang cukup mencolok pada kedua variabel tersebut. Adapun variabel X_3 memiliki nilai rata-rata yang relatif kecil dengan rentang nilai yang cukup lebar, mengindikasikan adanya ketidakteraturan distribusi data serta perbedaan intensitas antar responden. Secara keseluruhan, statistik deskriptif ini memberikan gambaran awal bahwa data penelitian bersifat variatif namun masih berada dalam batas yang wajar untuk dianalisis lebih lanjut pada tahap pengujian inferensial.

Pemilihan Model Terbaik Menggunakan Metode Forward Elimination, Backward Elimination, dan Stepwise

Selection of the Best Model using the Forward Elimination Method

The regression model equation formed according to the forward elimination method is shown in Table 3.

Table 3. Regression Model using the Forward Elimination Method

Regression Model
$\hat{y} = 59{,}73 + 0{,}939x_1 + 0{,}000002x_4$

Table 3 shows that from the results of the forward elimination method model formed with an alpha to enter of 0.05, there are only two predictor variables included in the model, namely x_1, which is expected length of schooling, and x_4, which is gross regional domestic product.

Best Model Selection using the Backward Elimination Method

The regression model equation formed according to the backward elimination method is shown in Table 4.

Table 4. Regression Model using the Backward Elimination Method

Regression Model
$\hat{y} = 59{,}73 + 0{,}939x_1 + 0{,}000002x_4$

Table 4 shows that from the results of the backward elimination method model formed with an alpha to remove of 0.05, there are only two predictor variables included in the model, namely x_1, which is expected length of schooling, and x_4, which is gross regional domestic product.

Best Model Selection using the Stepwise Method

The regression model equation formed according to the stepwise method is shown in Table 5.

Table 5. Regression Model using the Stepwise Method

Regression Model
$\hat{y} = 59{,}73 + 0{,}939x_1 + 0{,}000002x_4$

Table 5 shows that from the results of the forward elimination method model formed with alpha to enter and alpha to remove of 0.05, there are only two predictor variables included in the model, namely x_1, which is expected length of schooling, and x_4, which is gross regional domestic product.

Determining the Best Model from the Forward Elimination, Backward Elimination, and Stepwise Methods

After analyzing the factors that affect life expectancy by province in Indonesia in 2024 using forward elimination, backward elimination, and stepwise models, the results are shown in Table 6.

Table 6. Determination of the Best Model

Criteria	*Forward*	*Backward*	*Stepwise*
S	1,785	1,785	1,785
R-sq	43,03%	43,03%	43,03%
Mallow's Cp	1,07	1,07	1,07
AICc	158	158	158

Based on Table 6, it can be seen that the three methods have the same standard deviation value of 1.785, an R-sq or model goodness value of 43.03%, a Mallows' Cp value of 1.07, and a Corrected Akaike Information Criterion (AICc) value of 158. Therefore, the three methods (forward, backward, stepwise) produce the same and optimal final model, with good prediction quality, efficient model structure, and low and equivalent prediction errors.

Conclusion

The selection of the best regression model among all possible models for factors affecting life expectancy in Indonesia by province in 2024, using forward elimination, backward elimination, and stepwise selection, resulted in the same final model composition. In all three methods, the best model consistently retained two predictor variables, namely expected years of schooling and gross regional domestic product (GRDP). This result indicates that these two variables are the most relevant factors among the predictors considered in explaining provincial variation in life expectancy in Indonesia in 2024.

The selected regression model shows that an increase in expected years of schooling and GRDP tends to be associated with an increase in life expectancy. Substantively, this finding confirms that education and regional economic performance play an important role in shaping population welfare and longevity. Higher expected years of schooling may reflect better access to education, stronger health awareness, and improved long-term living conditions, while higher GRDP may indicate greater economic capacity to support health services, infrastructure, and social development. Therefore, improvements in both education and regional economic conditions are likely to contribute positively to better life expectancy outcomes.

However, the findings also suggest that the model explains only part of the variability in life expectancy across provinces. This means that there are still other factors outside the present model that may also influence life expectancy, such as health service access, environmental quality, nutrition, sanitation, and other demographic or social conditions. Therefore, future studies are recommended to include additional relevant variables and consider alternative modeling approaches in order to obtain a more comprehensive explanation of life expectancy differences across Indonesian provinces.

Acknowledgment

The authors would like to express their sincere gratitude to Institut Teknologi Sepuluh Nopember (ITS) for the academic support, research facilities, and conducive research environment provided throughout the completion of this study. The authors also acknowledge the valuable guidance and support from lecturers and staff of Institut Teknologi Sepuluh Nopember, which greatly contributed to the improvement and completion of this research.

References

Al Azies, H., & Dewi, V. M. (2021). Factors affecting life expectancy in East Java: Predictions with a Bayesian model averaging approach. *Jurnal Perencanaan Pembangunan: The Indonesian Journal of Development Planning, 5*(2), 283–295.

Arum, P. R., Gautama, R. P., Fitriani, I., & Naza, F. (2023). Identifying factors that influence life expectancy in Central Java using spatial regression models. *Jurnal Ilmiah Teori dan Aplikasi Statistika, 16*(2), 606–613.

Badan Pusat Statistik. (2024). *Life expectancy at birth and Human Development Index components in Indonesia, 2024 (Press release).*

Bahadure, N. B., Khomane, R., Raut, D., Bhagwatkar, D., Bakshi, H., Bawse, P., & Vishwakarma, M. (2025). Comparative analysis of life expectancy prediction using regression algorithms. *Intelligent Decision Technologies, 19*(1), 429–442.

Bahadure, N. B., Khomane, R., Raut, D., Chendake, Y., Routray, S., & Mishra, D. P. (2024). Regression model selection for life expectancy prediction: A comparative analysis of imputation techniques. In *Proceedings of the 4th International Conference on Advanced Research in Computing (ICARC)* (pp. 49–54). IEEE.

Boyko, N., & Moroz, O. (2021). Comparative analysis of regression regularization methods for life expectancy prediction. In *MoMLeT+ DS* (pp. 310–326).

Choudhury, T., Bharti, S. K., Gourisaria, M. K., Jena, J. J., Behera, D. K., & Bandyopadhyay, A. (2024). Predictive modeling of life expectancy using machine learning algorithms. In *Proceedings of the Global Conference on Communications and Information Technologies (GCCIT)* (pp. 1–6). IEEE.

Fadri, F. (2024). Socio-economic and environmental factors as determinants of life expectancy: Statistical approach with quantile regression. *Lebesgue: Jurnal Ilmiah Pendidikan Matematika, Matematika dan Statistika, 5*(1), 319–328. https://doi.org/10.46306/lb.v5i1.562

Fox, J. (2008). *Applied regression analysis and generalized linear models* (2nd ed.). Sage Publications.

Gazilas, E. T. (2024). Factors influencing life expectancy in low-income countries: A panel data analysis. *Journal of Applied Economic Research, 23*(3), 580–601.

Harrell, F. E. (2006). *Regression modeling strategies: With applications to linear models, logistic regression, and survival analysis*. Springer.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction* (2nd ed.). Springer.

He, X., Hu, J., Liu, C., & Zhang, Y. (2022). Analysis on relevant factors affecting life expectancy. In *Proceedings of the IEEE Asia-Pacific Conference on Image Processing, Electronics and Computers (IPEC)* (pp. 569–572). IEEE.

Heinze, G., Wallisch, C., & Dunkler, D. (2018). Variable selection – A review and recommendations for the practicing statistician. *Biometrical Journal, 60*(3), 431–449. https://doi.org/10.1002/bimj.201700067

Jackowska, B., et al. (2024). Effects of lifestyle-related risk factors on life expectancy: A comprehensive model for use in early prevention of premature mortality from noncommunicable diseases. *PLoS ONE, 19*(3). https://doi.org/10.1371/journal.pone.0298696

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2023). *An introduction to statistical learning: With applications in* R (2nd ed.). Springer.

Kabir, M. (2008). Determinants of life expectancy in developing countries. *The Journal of Developing Areas, 41*(2), 185–204.

Lakshmanarao, A. (2022). Life expectancy prediction through analysis of immunization and HDI factors using machine learning regression algorithms. *International Journal of Online & Biomedical Engineering, 18*(13).

Mun, J., Jalal, M. S., & Al Mamun, K. A. (2023). Analysis on top factors affecting life expectancy in Bangladesh. In *Proceedings of the IEEE Region 10 Humanitarian Technology Conference (R10-HTC)* (pp. 772–777). IEEE. https://doi.org/10.1109/R10-HTC57504.2023.10461848

Nair, R. (2025). Statistical analysis on factors influencing life expectancy. *Authorea Preprints.*

Otok, B. W., Zulfadhli, M., Pangesti, R. D., Kurniawan, M. I., Haryanto, A. E. P., & Kurniawan, I. (2026). Comparison of binary probit regression and Fourier series nonparametric logistic regression in modeling diabetes status at Hajj General Hospital Surabaya. *Barekeng, 20*(1), 255–270.

Pandey, A., & Chhikara, R. (2020). Analysis of life expectancy using various regression techniques. In *Proceedings of the International Conference on Advances in Computing, Communication Control and Networking (ICACCCN)* (pp. 209–213). IEEE. https://doi.org/10.1109/ICACCCN51052.2020.9362914

Țarcă, V., Țarcă, E., & Moscalu, M. (2024). Social and economic determinants of life expectancy at birth in Eastern Europe. *Healthcare, 12*(11), 1148. https://doi.org/10.3390/healthcare12111148

Teh, W. S., Ling, G. L. H. T., Rusli, N., Teh, H. H., Tan, K. P., Faisal, S., & Mohamed, M. S. A. (2025). Factors influencing life expectancy: Global evidence from 125 countries. *International Journal of Built Environment and Sustainability, 12*(2), 17–28. https://doi.org/10.11113/ijbes.v12.n2.1294

Trpkova-Nestorovska, M. (2018). Determinants of life expectancy: Comparative analysis for different groups of countries using panel regression. *Annual of the Faculty of Economics-Skopje, 53*, 581–591.

Wang, Y. (2021). The greatest factors affecting life expectancy: A research based on different continents and countries. In *Proceedings of the International Conference on Machine Learning, Big Data and Business Intelligence (MLBDBI)* (pp. 531–541). IEEE. https://doi.org/10.1109/MLBDBI54094.2021.00107

Wooldridge, J. M. (2020). *Introductory econometrics: A modern approach* (7th ed.). Cengage Learning.

ISSN: 3089-9729

Zulfadhli, M., Budiantara, I. N., Ratnasari, V., Suriaslan, A. S., & Dhewy, R. C. (2025). A statistical inference framework for FSNBLR: Modeling underdeveloped regional status in Eastern Indonesia. *MethodsX*, 103746. https://doi.org/10.1016/j.mex.2023.102983