SOÇIO-ECONOMIC AND ENVIRONMENTAL FACTORS AS DETERMINANTS OF LIFE EXPECTANCY: STATISTICAL APPROACH WITH QUANTILE REGRESSION

Firda Fadri¹
¹Program Studi Matematika, FMIPA, Universitas Jember
Email: firdafadri@unej.ac.id

ABSTRACT
The use of quantile regression in analyzing determinants of life expectancy provides a deep understanding of the influence of predictor variables across various quantiles of the data distribution. The analysis results reveal significant differences in predictor variables affecting life expectancy at each quantile, with socio-economic, environmental, and health program factors having varying impacts. Model evaluation is conducted considering several goodness-of-fit measures such as MAPE, AIC, and coefficient of determination, selecting the best quantile based on a combination of these values. Based on the analysis, the optimal quantile is chosen based on the lowest MAPE and AIC, and the highest coefficient of determination, considering the efficiency of significant predictor variables. The analysis indicates that the best model is found at the 60th quantile, with two significant predictor variables, namely mean years of schooling and labor force participation. This underscores the significant role of socio-economic factors in determining life expectancy at specific quantiles.

Keywords: Life Expectancy, Quantile Regression.

Received: February 19, 2024 / Accepted: March 18, 2024 / Published Online: April 30, 2024
INTRODUCTION

Life Expectancy (LE) is a crucial indicator used to measure the effectiveness of government efforts in enhancing the health and living standards of its society. LE reflects the average number of years that a person is expected to live. An increase in LE not only signifies progress in healthcare but also serves as a significant measure in evaluating the effectiveness of welfare policies and social advancement (Sulaeman, E.S., 2020). In the context of human development, LE is a vital component of the Human Development Index (HDI), utilized to assess the level of development in a region or country (Indah Sari & Warsitasari, 2023). High levels of LE indicate better access to quality healthcare services, adequate nutrition, proper sanitation, and a healthy environment conducive to human life. Therefore, efforts to enhance LE become a priority for governments in achieving sustainable development goals and improving the quality of life for their society (Montez et al., 2020).

According to The World Factbook CIA, Indonesia had a life expectancy of 73.08 years in 2022, with life expectancy for males around 70.86 years and approximately 75.4 years for females (CIA, 2023). Many factors influence the height or low of life expectancy in a region. These factors play a crucial role in determining the quality and length of an individual's life. For example, economic development reflected in per capita income (GDP per capita), income inequality index (Gini ratio), unemployment rate, and the percentage of the population living in poverty have significant impacts on life expectancy (Bidarti, A., 2020)(Tafran et al., 2020). Studies conducted by various researchers have highlighted the relationship between these economic factors and LE (Arum et al., 2023)(Nurfitri & Yanti, 2023)(Oluwaseyi et al., 2023). High income levels and labor force participation have a positive impact on LE because they allow better access to healthcare services and healthier lifestyles (Nica et al., 2023). Additionally, a high GDP per capita can serve as an indicator of economic well-being in a region (Şenturk & Ali, 2021). Higher income levels enable communities to access better healthcare services and live in healthier environments. This aligns with research indicating that a high GDP per capita contributes to increased LE. Therefore, a comprehensive understanding of the socio-economic factors that determine LE is very important for designing effective development policies and programs to improve society's welfare and LE (Chen et al., 2021).

Education also plays a significant influence in determining LE. Education not only provides knowledge about health but also opens up wider access to information and health services (Singh & Lee, 2021). The combination of income and education facilitates access to health services, which can extend life and increase the number of elderly people (Budiono & Rivai, 2021). In addition to education and income factors, the availability of physical health resources also supports health services for the society (Felangi & Yasa, 2021). The number of primary health facilities, such as community health centers and clinics, the percentage of coverage under health insurance (PBI and non-PBI), are important factors in ensuring wider access to quality health services (Cervantes et al., 2020) (Van den Heuvel WJ, Olaroiu M., 2017). Preparation of adequate physical health resources can support providing health services to the community. Health programs such as reducing infant mortality rates and increasing the percentage of infants receiving breastfeeding are also important strategies for increasing LE, especially among toddlers and infants (Aburto et al., 2020)(Miladinov, G., 2020).
Another significant factor that affects LE is the environmental factor. Access to clean water and adequate sanitation is a key factor that affects public health (Rahman et al., 2022; Setyadi et al., 2023). Additionally, demographic factors such as population density also impact LE (Tanadjaja et al., 2017). High population density is often associated with public health problems, such as the spread of infectious diseases, limited access to health resources, poor sanitation, and social instability, which have negative impacts on health and LE (Cohen et al., 2023).

Multiple linear regression analysis is an effective quantitative statistical method for analyzing the determinants of socio-economic and environmental factors in LE (Rubi et al., 2021; Agarwal et al., 2019). This approach allows researchers to identify how variables representing these factors influence LE and control for the influence of other variables. Thus, multiple linear regression analysis enables a more accurate assessment of the relative contributions of each factor to LE variation. The complexity of LE determinants makes the use of multiple linear regression analysis insufficient. Therefore, several previous researches have used the quantile regression method to obtain a more holistic understanding (Bilal et al., 2021).

Quantile regression allows researchers to examine how certain factors affect various levels of LE distribution. In other words, quantile regression enables the identification of the impact of independent variables on LE not only at the mean values but also at various percentiles of the LE distribution (Bai, et al., 2018) (Karimi, et al., 2023). This is important because certain factors may have different impacts depending on where a person is within the LE distribution. By using quantile regression analysis, researchers can gain a deeper understanding of how specific factors influence LE at various distribution levels, thereby providing a stronger basis for decision-making in planning and implementing more inclusive and effective healthcare policies.

METHOD
Data sources and research variables
The data for this research were derived from secondary sources and collected from the Central Bureau of Statistics and the National Social Security Council of Indonesia. Among them are the profile and statistics of the 2022 East Java Provincial Health Service Report and the 2022 East Java Province Annual Report. This research focuses on 38 regencies/cities in East Java Province as observation units. The response variable used in this research is life expectancy (LE) which is measured in years. The growth of economic and financial development can be measured by examining the Gross Regional Domestic Product (GRDP) per capita on the basis of current prices, the Gini ratio, and real expenditure per capita (in thousands of IDR). Meanwhile, social factors can be observed from the Percentage of population below the poverty line, mean years of schooling, unemployment rate, and labor force participation rate. Environmental factors include percentage of households with access to adequate sanitation, percentage of households with access to safe drinking water, and population density (per km²) which are considered in this research to determine their influence on LE. The success of health programs measured by infant mortality rate, percentage of infants (aged 0-23 months) receiving breastfeeding (in percent), the number of primary health facilities (FKTP), and percentage of health insurance coverage (PBI and non-PBI). The variables in this research are logarithmically transformed to deal with nonlinearity, heteroskedasticity, outliers, and skewness in the data.
**Table 1. Predictor variables**

<table>
<thead>
<tr>
<th>Variables name</th>
<th>Symbol</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gross Regional Domestic Product (GRDP) per capita at current prices</td>
<td>LnGRDP</td>
</tr>
<tr>
<td>Gini ratio</td>
<td>LnG</td>
</tr>
<tr>
<td>Real per capita expenditure (in thousand IDR)</td>
<td>LnRCE</td>
</tr>
<tr>
<td>Percentage of population below the poverty line</td>
<td>LnP</td>
</tr>
<tr>
<td>Mean years of schooling</td>
<td>LnMYS</td>
</tr>
<tr>
<td>Unemployment rate</td>
<td>LnUE</td>
</tr>
<tr>
<td>Labor force participation rate</td>
<td>LnLFP</td>
</tr>
<tr>
<td>Percentage of Households with Access to Adequate Sanitation</td>
<td>LnHAS</td>
</tr>
<tr>
<td>Percentage of Households with Access to Safe Drinking Water</td>
<td>LnHSD</td>
</tr>
<tr>
<td>Population density (per km2)</td>
<td>LnPD</td>
</tr>
<tr>
<td>Infant mortality rate</td>
<td>LnIMR</td>
</tr>
<tr>
<td>Percentage of infants (aged 0-23 months) receiving breastfeeding</td>
<td>LnIRB</td>
</tr>
<tr>
<td>Number of primary healthcare facilities (FKTP)</td>
<td>LnPHF</td>
</tr>
<tr>
<td>Percentage of health insurance coverage (PBI and non-PBI)</td>
<td>LnHI</td>
</tr>
</tbody>
</table>

**Research method**

The analytical steps to achieve the purpose of research are as follows:

1. Conduct descriptive analysis using thematic map of LE rates in East Java Province.
2. Transform each predictor variable into ln form, to ensure that they all have the same data type and format for consistency in the analysis.
3. Determine the quantile values to be estimated in the model. In this research, the quantile values used are the 0.05th, 0.1st, 0.2nd, 0.3rd, 0.4th, 0.5th, 0.6th, 0.7th, 0.8th, 0.9th, and 0.95th quantiles. Based on these quantile values, models will be obtained to estimate the likelihood of each quantile and extreme values in explaining LE.
4. Estimate the LE equation for each quantile.
5. Conduct model testing for each quantile value to determine which factors influence LE.
6. Calculate Mean Absolute Percentage Error (MAPE) and AIC values for each quantile to assess the goodness of fit from model.
7. Determine the best model based on the goodness of fit and interpret the model.

**RESULT AND DISCUSSION**

**Life expectancy in East Java**

Life expectancy rates in East Java Province is 71.76 years, ranking tenth in Indonesia in 2022, slightly different from Indonesia's overall life expectancy of 71.88 years. This indicates that the average LE of East Java society is nearly the same as most other provinces. A visualization of the LE distribution in East Java based on regencies/cities divided into three groups: low, medium, and high, it is shown in Figure 1. Regencies/cities with darker colors indicate higher LE rates, while lighter colors indicate lower LE rates. The range of LE rates is 7.25, with the highest LE rates in Tulung Agung Regency at 74.54 years and the lowest LE rates in Bondowoso Regency at 67.29 years.
In Figure 1, regencies/cities such as Gresik, Sidoarjo, and Surabaya City are included in the high LE group, which is reasonable as these areas are industrial and economic centers in East Java Province. Meanwhile, regencies such as Ponorogo, Trenggalek, Tulung Agung, and Blitar also have high LE even though they are geographically distant from the economic center of East Java province. Therefore, there are other factors besides economic factors, which require deeper analysis to understand the determinants of these factors.

The relationship between life expectancy, socio-economic and environmental factors.

Quantile regression is used to model LE for various quantile values that are influenced by predictor variables and to identify factors that have a significant effect. The estimated values of the quantile regression parameters are presented in Table 2. The results of quantile regression estimates show significant differences in predictor variables that influence LE. In Q5, almost all predictor variables are significant except the LnGRDP variable. Additionally, Q10, Q20, Q30 experienced a decrease in the number of predictor variables even though they still numbered 6 to 11 variables from each socio-economic, environmental, and health program factor. Q40, Q50, Q60 only have two or three significant predictor variables that impact LE. There are no predictor variables that significantly effect on LE in Q70. Entering Q80, Q90, Q95, there are changes with the increasing number of predictor variables from socio-economic, environmental, or health program factors.

An interesting thing to discuss is the change of significant predictor variables at each quantile. The economic factors such as GRDP and the number of population below the poverty line have a significant influence on both low and high quantiles. Population density as an environmental factor and all health program factors are also significant in the low and high quantiles. Meanwhile, Gini ratio, real expenditure per capita, unemployment rate, labor force participation, households having access to adequate sanitation and safe drinking water are significant only in the low quantile. The variable that is always significant at low to high quantiles is the mean years of schooling, which is not significant at Q70 because the model formed is not significant.
Table 2. Estimation of quantile regression parameters

<table>
<thead>
<tr>
<th>Quantile</th>
<th>MAPE</th>
<th>AIC</th>
<th>( R^2 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Q5</td>
<td>0.1941</td>
<td>-205.23</td>
<td>63.77</td>
</tr>
<tr>
<td>Q6</td>
<td>0.1993</td>
<td>-197.05</td>
<td>70.74</td>
</tr>
<tr>
<td>Q7</td>
<td>0.2093</td>
<td>-212.06</td>
<td>64.01</td>
</tr>
<tr>
<td>Q8</td>
<td>0.1953</td>
<td>-211.87</td>
<td>60.60</td>
</tr>
<tr>
<td>Q9</td>
<td>0.1953</td>
<td>-210.52</td>
<td>63.80</td>
</tr>
</tbody>
</table>

To evaluate the model, several measures can be used such as MAPE, AIC, and \( R^2 \). Each measure provides different insights into the model's performance in different parts of the distribution. Based on table 3, it can be seen that there are variations in the values of MAPE, AIC, and coefficient of determination between various quantiles (Q5, Q10, ..., Q95). The optimal quantile is the one with the lowest MAPE value, because this indicates that the model has a smaller prediction error at that quantile. The MAPE values range from about 0.19 to 0.38, indicating variations in the prediction error rate among different quantiles. The quantile with the lowest MAPE value is Q50.

AIC is a measure of a model's relative goodness, with lower AIC values indicating a better model. The optimal quantile is the one with the lowest AIC value, indicating that the model has a better fitting quality to the data at that quantile. AIC values ranged from -203.50 to -190.81, indicating that some models were better than others in fitting the data. The quantile with the lowest AIC value is Q70, but because there is no significant predictor variable at that quantile, the model at Q70 cannot be the best model. Therefore, the quantile with the next lowest AIC is chosen, namely Q80.
The next measure of model goodness is the coefficient of determination. The optimal quantile is the one with the highest coefficient of determination, indicating that the model is able to explain the variability of the data well at that quantile. The coefficient of determination values range from about 11.34% to 70.74%, indicating variation in how well the model explains the data across different quantiles. Q60 has the highest determination coefficient.

By combining all measures of the goodness of the model, the quantile with the lowest MAPE and AIC and the highest coefficient of determination is selected, taking into account the efficiency of the predictor variables that are significant at that quantile. Considering this, Q60 was chosen as the best model because it has a small MAPE value, the smallest AIC and the largest coefficient of determination with two significant predictor variables, namely mean years of school and labor force participation.

CONCLUSION

By using quantile regression to analyze the determinants of socio-economic, environmental, and health program factors on life expectancy, we can draw several conclusions, for the first, quantile regression allows to understand the influence of these factors on life expectancy at various quantiles of the data distribution. It provides deeper insights than ordinary linear regression, as it can reveal heterogeneity in the effect of predictor variables on different quantiles. Second, based on the goodness-of-fit measure, socio-economic factors appear to have the most significant impact on life expectancy at the 60th quantile of the data distribution. This indicates that socio-economic factors play a strong role in determining life expectancy levels in the middle part of the distribution. And the third, while environmental or health program factors also influence life expectancy at other quantiles, the goodness-of-fit measure for these factors is not yet optimal. This suggests that the model may not be strong enough to capture the relationship between environmental or health program factors and life expectancy at the extreme parts of the distribution.
SOCIO-ECONOMIC AND ENVIRONMENTAL FACTORS AS DETERMINANTS OF LIFE EXPECTANCY: STATISTICAL APPROACH WITH QUANTILE REGRESSION
Firda Fadri
Vol. 5, No. 1, April 2024 hal. 319-328
DOI Artikel: 10.46306/fb.v5i1.562

REFERENCES


SOCIO-ECONOMIC AND ENVIRONMENTAL FACTORS AS DETERMINANTS OF LIFE EXPECTANCY: STATISTICAL APPROACH WITH QUANTILE REGRESSION

Firda Fadri
Vol. 5, No. 1, April 2024 hal. 319-328
DOI Artikel: 10.46306/lb.v5i1.562

Pembangunan Universitas Udayana, 10(4), 1447–1477.
