Life Expectancy Regression Analysis

Wenshan Zhang*

*Corresponding author. Email: wzhan779@uwo.ca

School of Statistical and Actuarial Sciences, Western University, London, Ontario N6A 3K7, Canada

Abstract: This paper explores the relationship between life expectancy and a range of factors, including vaccination rates, health status, and socio-economic conditions, through the application of a multiple linear regression model. The study encompasses initial data visualization and model assumption checking, addressing heteroscedasticity through techniques such as removing influential points, Box-Cox transformation of the response variable, Goldfeld-Quandt test, and Weighted Least Squares regression. Two optimal models are derived: one with the best explanatory capacity and another with the best predictive accuracy. In the discussion section, we suggest refining the model using decision trees, stratified by the development status of countries (developed versus developing). This involves fitting distinct regression models for each group to identify more accurate predictive models. Based on these findings, we provide recommendations for improving population life expectancy.

Keywords: Multiple linear regression, Life expectancy, Weighted Least Squares, Gold-Quandt test, Decision tree .

1. Introduction

1.1 Background and Data Resource

With the progress of humanitarian efforts and modern society, governments worldwide are increasingly focusing on the health and well-being of their citizens. “Life expectancy” has become a hot topic. As mentioned in the New England Journal of Medicine, effective analysis of life expectancy can provide crucial information about the population’s health status, useful for planning and decision-making [1]. It also helps identify high-risk populations, enabling targeted preventive healthcare interventions. Based on this inspiration, we have decided to study the factors influencing life expectancy by establishing a multiple regression model.

The data set was collected from the World Health Organization(WTO) and United Nations websites with the assistance of Deeksha Russell and Duan Wan and subsequently uploaded to Kaggle by KumarRajarshi (https://www.kaggle.com/datasets/kumarajarshi/life-expectancy-who/data). It encompasses the average life expectancy from 2000 to 2015 for 193 countries globally, along with associated influencing factors.

1.2 Literature review

Sweden has assessed the impact of smoking and alcohol abuse at different levels of education on changes in life expectancy and mortality and concluded that alcohol can affect a person’s life expectancy, particularly for men [2]. Therefore, this study will further analyze the associa-

MSEA 2024, May 24-26, Jinan, People's Republic of China

DOI 10.4108/eai.24-5-2024.2350121

tion between alcohol and life expectancy, and the extent of the association, to provide better evidence to support alcohol-related policies and actions.

Vaccination of one-year-old children is important for the health and safety of children as they grow into adults. Life expectancy and one-year immunization rates for hepatitis B, diphtheria, and polio were examined to test such assumptions. WHO and UNICEF launched the Expanded Programme on Immunization (EPI) in 1976 to control six childhood diseases: tuberculosis, diphtheria, whooping cough, tetanus, polio, and measles. Children's vaccination and disease prevention and control are health matters of worldwide concern. Studies show that parents' education level and sports level are significantly correlated with children's immunization, and improving parents' education level can improve children's immunization coverage [3]. Measles is one of the most common respiratory infections in children under the age of 5 years. From 2000 to 2016, the percentage of children who received at least one dose of measles vaccine before their first birthday increased from 72% to 85%, and as a result, the global number of measles deaths declined by 84%. From an estimated 550,100 in 2000 to 89,780 in 2016 [4]. With a confirmed association between immunization coverage among 1-year-olds and under-five deaths and populations' life expectancy, we intend to conduct more studies and initiatives to analyze the factors driving vaccination rates to improve them.

1.3 Our objectives

We aim to build and refine a regression model with "life expectancy" as the dependent variable. This model is designed to elucidate the associations between life expectancy and various types of potential influencing factors and to identify the most significant predictors. Finally, we will provide recommendations for countries aiming to improve the life expectancy of their populations.

2. Exploratory data analysis

2.1 Data Structure

This data set consists of 2938 observations and each observation has 21 variables. We first remove all observations with missing values, then randomly select 1000 observations for our project. In our study, "life.expectancy" is the response variable. The table 1 below briefly summarizes the remaining variables in our project.

Table 1. Variable Information

Name	Description	Type
Country	Country name	Character
Year	Year	Character
Life expectancy	Life Expectancy in age	Numerical
Adult Mortality	Adult Mortality Rates of both sexes (probability of dying between 15 and 60 years per 1000 population)	Numerical
Infant deaths	Number of Infant Deaths per 1000 population	Numerical
Status	Developed or Developing status	Categorical
Alcohol	Pure alcohol consumption per capita	Numerical
percentage expenditure	Expenditure on health as a percentage of GDP per capita	Numerical

Name	Description	Type
Hepatitis.B	Hepatitis B (HepB) immunization coverage among 1-year-olds	Numerical
Measles	Measles - number of reported cases per 1000 population	Numerical
BMI	Average Body Mass Index of entire population	Numerical
under-five deaths	Number of under-five deaths per 1000 population	Numerical
Total expenditure	General government expenditure on health as a percentage of total government expenditure	Numerical
Polio	Polio (Pol3) immunization coverage among 1-year-olds	Numerical
Diphtheria	Diphtheria tetanus toxoid and pertussis (DTP3) immunization coverage among 1-year-olds	Numerical
HIV/AIDS	Deaths per 1000 live births HIV/AIDS (0-4 years)	Numerical
Population	Population of the country	Numerical
GDP	Gross Domestic Product per capita	Numerical
thinness 1-19 years	Prevalence of thinness among children and adolescents for Age 10 to 19	Numerical
thinness 5-9 years	Prevalence of thinness among children for Age 5 to 9	Numerical
Income composi-tion of resources	Human Development Index in terms of income composition of resources (index ranging from 0 to 1)	Numerical
Schooling	Number of years of Schooling	Numerical

2.2 Visualization

2.2.1 Relationship between response and single predictor

We aim to discover the relationship between life expectancy and each single predictor variable. Before starting, we need to specify that considering different countries and years is irrelevant. Therefore, we have removed the columns “country” and “year”. All predictors could be divided into 3 categories: immunization-related, health, and economic&social factors. The figure 1 below shows only a few crucial relationships.

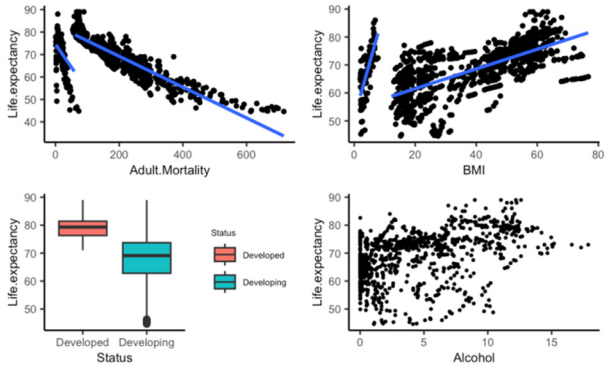

Fig. 1. Relationship between life expectancy and single predictors

Some rough conclusions can be drawn:

1. All immunization-related factors (***Hepatitis.B, Polio, Diphtheria***) have approximately positive relationship with life expectancy. In most of the observations, vaccination rates are high.

2. ***Infant deaths/under-five deaths*** are inversely proportional to life expectancy, which is obvious. However, Adult Mortality seems to exhibit a segmented relationship with life expectancy, there's a break point at Adult Mortality = 60.

3. ***Measles, HIV*** and ***thinness coverage*** have approximately negative linear relationship with life expectancy.

4. Life expectancy has a weak positive relationship with ***Alcohol assumption***. This finding goes against our common sense that alcohol is associated with many chronic diseases. This could be due to imprecise data and the fact that drinking may have different effects on different populations.

5. ***BMI*** seems to exhibit a segmented relationship with life expectancy. Through repeated testing of breakpoints, we found that when $BMI \leq 10$, its impact on life expectancy is relatively strong. When $BMI > 10$, there is also a positive relationship with life expectancy, but it seems not as significant as when BMI is less than or equal to 10. We will discuss the reasons for this phenomenon in the discussion section.

6. ***Status/GDP*** of national development and investment in health (***percentage/total expenditure on health***) are positive related to life expectancy. Life expectancy seems to have no relationship with ***population***. This might be because countries with larger populations often tend to have expansive land areas and abundant resources.

7. ***Income Composition of resources*** and ***Schooling*** have significant positive linear relationship with life expectancy. Higher ICOR indicates optimal utilization of available resources, while higher schooling indicates a higher average education level in a country. These two indicators drive the development of healthcare facilities and underscore the level of importance people place on health, which is good for life expectancy.

2.2.2 Correlation between pairs of predictors

Before constructing a model, we need to discover the correlation between predictors via a correlation coefficient heatmap. Figure 2 only focuses on some crucial relationships between pairs of variables. Positive correlations are represented in blue, while negative correlations are represented in pink.

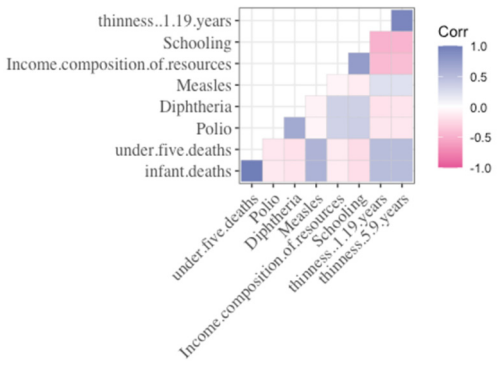

Fig. 2. Correlation heatmap

We could observe that there is a strong correlation between:

1. ***Infant death*** and ***under five deaths:*** Infants and children under five years of age share similar living environments and medical conditions. We hope to remove one of those two to avoid multi-collinearity in the variable selection step.

2. ***Income composition of resource*** and ***schooling***: Resource-rich countries are likely to invest in their education systems to produce higher-level talent, thereby raising overall levels of human capital and income.

3. ***Polio*** and ***Diphtheria***: They have similar modes of transmission under some social and environmental conditions, such as via droplets.

4. ***Measles*** and ***under five death:*** Measles is usually more common in children, whose immune systems are not fully developed, so their immune systems may be more vulnerable when faced with the measles virus.

In summary, the coverage of diseases is closely intertwined with a country's healthcare infra-structure and socioeconomic conditions. If we fit them all into a model, there must exist multi-collinearity. For our further steps, it's important to select representative predictors, so we need to use stepwise selection to avoid over-fitting problems.

3. Modeling

3.1 Full Model

First, we fit a model with all possible predictors. We notice that although we get a high R^2 of 0.84, only 7 predictors(Adult mortality, infant deaths, BMI, under five deaths, HIV/AIDS, Income composition of resources, and Schooling) out of 19 are significantly related with life expectancy, shown as figure 3 below.

	Estimate	Standard Error	t value	Pr(>|t|)	
(Intercept)	53.591	0.915	58.563	0.0000	***
Status1	1.384	0.432	3.200	0.0014	**
Adult.Mortality	-0.017	0.001	-13.488	0.0000	***
infant.deaths	0.107	0.013	8.020	0.0000	***
Alcohol	-0.103	0.042	-2.461	0.0140	*
percentage.expenditure	0.000	0.000	1.162	0.2454	
Hepatitis.B	-0.002	0.006	-0.277	0.7818	
Measles	-0.000	0.000	-1.209	0.2268	
BMI	0.038	0.008	4.839	0.0000	***
under.five.deaths	-0.080	0.010	-8.283	0.0000	***
Polio	0.016	0.007	2.408	0.0162	*
Total.expenditure	0.113	0.052	2.169	0.0303	*
Diphtheria	0.006	0.008	0.748	0.4548	
HIV.AIDS	-0.427	0.024	-18.126	0.0000	***
GDP	0.000	0.000	0.815	0.4156	
Population	-0.000	0.000	-0.765	0.4442	
thinness..1.19.years	-0.064	0.070	-0.912	0.3618	
thinness.5.9.years	0.029	0.069	0.417	0.6771	
Income.composition.of.resources	8.269	0.957	8.643	0.0000	***
Schooling	0.866	0.072	12.061	0.0000	***

*Signif. codes: 0 <= '***' < 0.001 < '**' < 0.01 < '*' < 0.05*

Residual standard error: 3.528 on 980 degrees of freedom

Multiple R-squared: 0.84, Adjusted R-squared: 0.8369

F-statistic: 270.7 on 980 and 19 DF, p-value: 0.0000

Fig. 3. Summary output for full model

3.2 Assumption Checking

Assumptions we need to check:

Linearity(Residual plot, shown in figure 4): By observing the residual plot, we could find that the mean of residuals is roughly zero at any fitted value except a few points at both tails, and residuals do not manifest any discernible systematic patterns, such as discernible trends or curvature. Considering the ample size of our dataset, comprised of 1000 observations, these few points are not sufficient to conclude a violation of the linearity assumption. Therefore, we can conclude that the linearity assumption holds.

Normality(QQ Plot and S-W test, shown in figure 4&5): We could see from Q-Q normal plot, that points at the left tail diverge from the straight line. We also verify that from s-w test, the p-value is smaller than 0.05. Normality assumption is violated!

Equal variance(Residual plot and B-P test, shown in figure 4&5): We could see from the residual plot, the spread of residual varies as the fitted value increase. We also verify that from b-p test, the p-value is much smaller than 0.05. The equal variance assumption is violated.

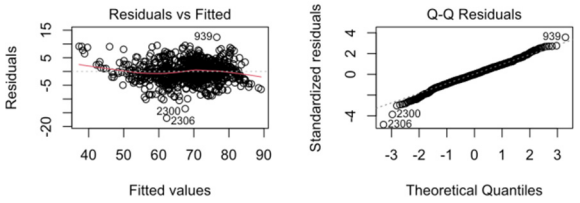

Fig. 4. Residual plot and QQ plot for full model

```
    studentized Breusch-Pagan test

data:  lm.1
BP = 99.068, df = 19, p-value = 7.896e-13


    Shapiro-Wilk normality test

data:  res
W = 0.99197, p-value = 2.943e-05
```

Fig. 5. B-P and S-W test results for full model

3.3 Influential Points

To identify the reasons for violating model assumptions, we need to examine whether there are influential points in the dataset. For this purpose, we consider points with $Cook'sdistance > 4/n$ as influential points, where n represent the total number of observations. We ended up with 67 influential points out of the total 1000. After removing those influential points, we apply the testing/graphical approach again. The results show that equal variance and normality assumptions are still violated. The fortunate aspect is that the p-value in the S-W test has significantly increased, indicating that the removal of influential points is

still somewhat helpful, although it has not achieved the expected effect. Also, the R^2 and $adjustedR^2$ increase about 4% after removing the influential points, so we will try another metric based on the current model.

3.4 Transformation

Since removing influential points did not correct the assumptions, we applied the box-cox method to transform our response variable Life expectancy. Here we choose 1.4 for lambda and construct a new model with transformed y. We apply the testing/graphical approach again. The results in figure 6&7 show that the equal variance assumption is still violated. The good news is that the normality assumption has finally been corrected.

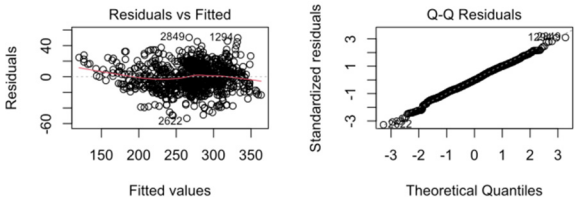

Fig. 6. Residual plot and QQ plot after transformation

```
    studentized Breusch-Pagan test

data:  lm.3
BP = 69.224, df = 19, p-value = 1.237e-07


    Shapiro-Wilk normality test

data:  res
W = 0.99686, p-value = 0.06256
```

Fig. 7. B-P and S-W test result after transformation

3.5 Other Strategies to Fix Equal Variance Assumption

3.5.1 Gold–Quandt Test

Instead of the B-P test, we applied the Gold-Quandt test, and we can observed from figure 8 that the EV assumption was valid, which contradicts the result of the b-p test. The "Journal of Mathematics and Statistics" [5] highlights a distinction between the Breusch-Pagan test and the Goldfeld Quandt Test. While the Breusch-Pagan test assumes equal variances for all data points, the Goldfeld Quandt Test compares variances between two subgroups: one with high values and another with low values. If the variances differ significantly, the test rejects the null hypothesis of constant error variances. Notably, the Breusch-Pagan test is sensitive to deviations from normality, and deviations from a perfect normal distribution can impact the assessment of equal variance. This explains the observed discrepancy in results when using the B-P test in previous work, as the p-value from the Shapiro-Wilk test is not significantly large, just slightly exceeding 0.05.

```
    Goldfeld-Quandt test

data:  lm.3
GQ = 1.0461, df1 = 447, df2 = 446, p-value = 0.3171
alternative hypothesis: variance increases from segment 1 to 2
```

Fig. 8. Gold- Quandt test result

3.5.2 Weighted Least Squares Regression

Since we never dealt with heteroscedasticity before, we conducted an extensive literature review, and ultimately found the solution in "Transformation and Weighting in Regression"[Carroll and Ruppert (1988), Ryan (1997)]. The recommended approach is to employ Weighted Least Squares Regression. In this estimation technique, each observation is weighted proportionally to the reciprocal of its error variance, effectively addressing the challenge posed by non-constant variance.

$W_i \propto \frac{1}{\hat{\sigma}^2}$, where $W_i$is the weight of ith observation, and $\hat{\sigma}^2$ is the sample standard deviation of the response variable at the ith combination of predictor variable values

The fundamental idea behind selecting these weights is to reduce the influence of observations exhibiting larger residuals at certain predictive levels by lowering their weights, thereby minimizing their impact on model fitting.

After applying the WLS regression, figure 9 shows that the equal variance assumption is finally satisfied, and the normality assumption holds as well. Furthermore, the R-square has increased dramatically to an astonishing 0.9995, indicating a potential issue of over-fitting in our model. (We will discuss more about the limitations about WLS regression later)

```
    studentized Breusch-Pagan test

data:  lm.wls
BP = 0.4605, df = 19, p-value = 1


    Shapiro-Wilk normality test

data:  res
W = 0.99683, p-value = 0.05998
```

Fig. 9. B-P and S-W test results for the WLS model

3.6 Variable Selection

Due to the large number of predictors, we aim to avoid an overly complex model. Therefore, we opt to select predictors that are most likely to impact life expectancy by Stepwise selection, which can avoid the problem of collinearity, once the predictor added in the early round be redundant, it allows to be dropped at any trial.

Since we use two distinct methods to address the correction of the equal variance assumption in the step5, two models were derived—(Model A with transformed response term and Model B with both transformed response term and Weighted Least Squares). We will proceed to

conduct a variable selection process based on each of these models and compare the outcomes in the subsequent steps. By observing the summary output, we could find that:

For model A(with transformed response term), only 11 predictors are selected, with R-squared of 0.8665: *Life.expectancy~ Income.composition.of.resources + Adult.Mortality + Schooling + HIV.AIDS + percentage.expenditure + BMI + Polio + Total.expenditure + under.five.deaths + infant.deaths + Status*

For model B(both transformed response term and Weighted Least Squares), 13 predictors are selected: With R-squared of 0.9972: *Life.expectancy ~ Adult.Mortality + Income.composition.of.resources + Schooling + Measles + Total.expenditure + HIV.AIDS + GDP + BMI + Alcohol + Status + Diphtheria + percentage.expenditure + Polio*

3.7 Model Comparison

For this step, we will discuss which model is better in two aspects:

For the explanation purpose: we will use metrics of AIC and BIC to find the most efficient model. From result shown in table 2, Model A preforms better.

Table 2. Model performance in explaining data

Model	AIC	BIC
Model A	7874.046	7936.946
Model B	8235.850	8308.426

For the prediction purpose: we will use the K-fold validation method to find the most accurate model for prediction. From result shown in table 3, Model B returns the lower MSE, Model B is more accurate for prediction new data.

Table 3. Model performance in predicting new data

Model	Mean_RMSE
Model A	186.6515
Model B	186.5834

4. Conclusion

The predictors that are significant related to life expectancy is the predictors in model A. Since we used box-cox transformation here, we need to be careful with the betas in the model. The final model can be written as:$\hat{Y} = (1.4\beta X + 1)^{0.7143}$ where β and X are matrix, which can also be represented as:

$$\begin{aligned} \textit{Life expectancy} = {} & 54.4364 + 22.4048\ \textit{Income Composition of resource} \\ & + 0.8947\ \textit{Adult Mortality} + 3.6847\ \textit{Schooling} - 1.726\ \textit{HIV} \\ & + 1.0681\ \textit{Polio} + 1.9300\ \textit{Total expenditure} + 0.4808\ \textit{under five death} \\ & + 0.594\textit{infant death} + 4.8056 I_{(status=1)} \end{aligned}$$

The most accuracy prediction model is Model B. To enhance future life expectancy, prioritize improving overall education levels in countries. Elevating education fosters health awareness and indirectly encourages vaccine uptake, reducing the likelihood of illness. Addressing the

critical issue of preventing HIV transmission, given its significant impact on life expectancy, can also lower adult mortality rates. Additionally, countries should focus on sustainable development policies, including economic diversification and technological innovation, to boost income composition

5. Discussion

In the process of constructing the model, we encountered several noteworthy issues:

5.1 Weighted Least Squares (WLS)

Assumption:1) Linearity: WLS assumes a linear relationship between the dependent and independent variables, which is a fundamental assumption in regression analysis. WLS is primarily employed in linear regression models. 2) Independence: WLS assumes that observations are independent of each other. This implies that the residual of one observation should not provide information about the residual of another observation. Disadvantages: However, the theoretical foundation of this method relies on the assumption that the weights are precisely known. In real-world applications, this is almost never the case, necessitating the use of estimated weights instead [6]. Therefore, haphazardly using $w = 1/\sigma^2$ in our model is irrational. We need to repeatedly compare and utilize knowledge that may be acquired in the future, combined with data distribution, to choose the most suitable weights.

5.2 Improvement of model: Decision Tree

In Section 2.2.A, we previously mentioned that predictors such as ***BMI*** seem to exhibit a segmented relationship with life expectancy (BMI shows a noticeable breakpoint around 10, while adult mortality is around 60). We hypothesize that this might be due to variations in the level of economic development among countries, leading to an implicit division of the data into two groups. For example, in developing countries, the average BMI may be lower than in developed countries, and the average number of adult deaths may be higher.

To test this hypothesis, we attempted to plot the relationships between predictors and life expectancy, using different colors to distinguish countries at different stages of development (blue representing developed countries and pink representing developing countries). However, we can observed from figure 10 that ***BMI*** and ***adult mortality*** still exhibited breakpoints.

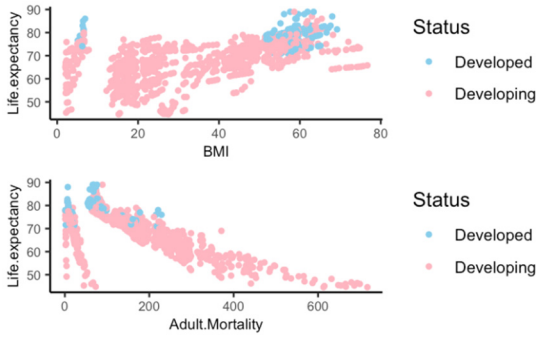

Fig. 10. Breakpoints still exist after grouping by Status

Meanwhile, some breakpoints for predictors were partially explained, such as ***Income Composition of Resources***, ***Diphtheria***, ***Polio***, and ***Infant deaths***. we can clearly see in Figure 11 that the relationship between Income Composition of Resources and life expectancy in developed countries demonstrated a distinct linear positive correlation. However, breakpoints still persisted in developing countries.

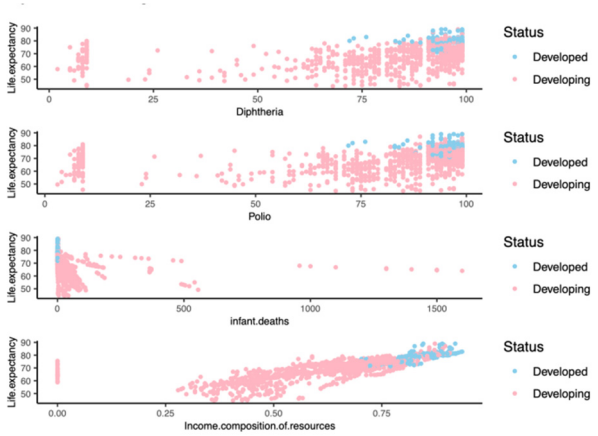

Fig. 11. Categories by Status only has effects on developed country

This finding corroborates our hypothesis, suggesting that we may need to explore additional classification nodes, such as attempting a decision tree, to further understand and delineate the complex relationships within the data. A previous study titled "Life Expectancy Prediction through Analysis of Immunization and HDI Factors using Machine Learning Regression Algorithms" has provided us with additional insights. The research utilized the same dataset as ours but employed four different regression models simultaneously - linear regression, Support Vector Regression, Decision Tree, and Random Forest. The findings indicated that, in comparison to the multiple linear regression model, the decision tree exhibited a higher R-square and a lower Mean Squared Error, suggesting its superior performance in the prediction of life expectancy[7]. In conclusion, by incorporating future learning, we can enhance the model's fitness and conduct a more in-depth analysis of the relationships between variables.

Acknowledgements. This research was supported by the National Natural Science Foundation of China under Grant 70572071.

References

[1] Katz, S. et al. (1983) 'Active life expectancy', New England Journal of Medicine, 309(20), pp. 1218–1224. DOI:10.1056/nejm198311173092005.

[2] Östergren, O., Martikainen, P. and Lundberg, O. (2017) 'The contribution of alcohol consumption and smoking to educational inequalities in life expectancy among Swedish men and women during 1991–2008', International Journal of Public Health, 63(1), pp. 41–48. DOI:10.1007/s00038-017-1029-7.

[3] Siddiqi, N., Khan, A., Nisar, N., & Siddiqi, A. E. (2007). Assessment of EPI (expanded program of immunization) vaccine coverage in a peri-urban area. JPMA. The Journal of the Pakistan Medical Association, 57(8), 391–395.

[4] Revilla, F. (2018) Paho/WHO: Basic measles facts, Pan American Health Organization / World Health Or- ganization, https://www3.paho.org/hq/index.php?option=com_content&view=article&id=14173%3Abasic-measles-facts&Itemid=72231&lang=en#gsc.tab=0 (Accessed: 07 December 2023)

[5] Rana, Md.S., Midi, H. and Imon, A.H.M.R. (2008) 'A robust modification of the goldfeld-quandt test for the detection of heteroscedasticity in the presence of outliers', Journal of Mathematics and Statistics, 4(4), pp. 277–283. DOI:10.3844/jmssp.2008.277.283.

[6] Carroll, R.J., & Ruppert, D. (1988). Transformation and Weighting in Regression (1st ed.). Chapman and Hall/CRC. DOI: https://doi.org/10.1201/9780203735268

[7] Lakshmanarao, A. et al. (2022) 'Life expectancy prediction through analysis of immunization and HDI factors using machine learning regression algorithms', International Journal of Online and Biomedical Engineering (iJOE), 18(13), pp. 73–83. DOI:10.3991/ijoe.v18i13.33315.