Data Analysis Report on Life Expectancy Models

Wenjun He, Krit Kasikpan

STA 302H1/1001H: Methods of Data Analysis I

Professor Antonio Herrera Martin

June 17th, 2024

Check the .rmd attachment for full process of model derivation

Introduction

This research aims to explore the relationships between life expectancy and various influencing factors at the individual, national and global levels. The inspiration comes from the author's interest in health and various practices for keeping healthy. It is known that having a high metabolism would decrease lifespan. Therefore, it seems that having a smaller food intake would allow people to live longer. Recent research such as Li et al. (2020) also has supported such an argument by drawing attention to the bonds between health-related behaviors and certain diseases. The research showed that a healthy lifestyle would decrease the risk of chronic disease, hence improving life expectancy. In contrast to Li et al. (2020), who focus on individual lifestyle options, our study will also consider the impact of national economic conditions. In the real world, the countries with less food intake are poorly developed countries with a low GDP per capita, substandard healthcare, and weaker education systems, which correlate with reduced life expectancy. This argument is supported by Miladinov (2020), who found a positive correlation between GDP per capita and life expectancy.

Unlike many studies that focus on a single factor, this research takes a unique approach by integrating personal health behaviors and national economic indicators. By examining the combined impact of these factors on life expectancy, this research aims to provide a more comprehensive analysis of the potential cause and effect of life expectancy. Specifically, this research will explore the relationship between individual-wide daily average food consumption of calories, protein, and fats, and country-wide socio-economic indicators, including GDP per capita and average years of schooling, and their effects on life expectancy.

This comprehensive research on life expectancy holds significant global implications. Life expectancy is a key focus in Development Economics and a top priority for every country and the United Nations (UN). It is a crucial component of the Human Development Index (UN, 2020), a widely used measure of human development. Moreover, Goal 3 'Good Health and Well-being' of the Sustainable Development Goals (SDG), adopted by the UN, includes sub-targets related to reducing mortalities, improving health coverages, and fighting diseases (UN, 2015). These are all vital factors that influence life expectancy. Therefore, this research not only offers individuals some personal health suggestions, but also provides guidelines for countries and international organizations to achieve high life expectancy.

Method

The primary aim of this research is to develop a robust linear model to predict life expectancy based on various predictors we come up with. We will first collect our data.

Data Collection and Description: We sourced the most recent data available in 2021 using data from multiple reputable sources. We did not collect data from different years as the past data may influence the same data collected in the future, where analyzing this requires the knowledge of times series. The predictors with corresponding sources of data and units are shown as below

- **Individual daily average consumption of Calories:**
    - Source: Our World in Data, affiliated with the University of Oxford. Unit: Calories. Name of the predictor: Calorie
- **Individual daily average consumption of Protein, Fat**

    - Source: From Our World in Data, affiliated with the University of Oxford. Unit: grams. Name of the predictors: Protein, Fat, respectively
- **Individual average years of schooling**:
    - Source: Our World in Data, affiliated with the University of Oxford. Unit: years. Name of the predictor: School
- **GDP per capita:**
    - Source: World Bank. Unit: USD. Name of the predictor: GDP
- Additional data gaps were filled using Statista

Note that data from various sources are combined into one whole data set since each data sample represents a country where each predictor stands for a country's statistics, which is logical that we can combine the data from multiple sources.

Data Preparation: Data cleaning is first performed to remove any row with missing data. Then, it is randomly divided into a training set and a testing set. To ensure each set has at least 10 samples per predictor, we split the data into 60% training and 40% testing sets. The training set is used to build the model, while the testing set is used to evaluate its accuracy. Descriptive statistics (minimum, maximum, median, mean, quartiles, and standard deviation) are calculated for each dataset to ensure they can be compared and evenly distributed, and we can perform the data analysis.

Data Analysis: The initial phase of our exploratory data diagnostics involves a linearity check where we plot the relationship between the response variable, life expectancy, and each predictor. This process helps to assess whether a linear model is appropriate for the data. If these plots show any non-linear patterns, we consider applying power or BoxCox transformations to better meet the linear model requirements.

Following the linearity check, we will check the two conditions on our data. We first use the lm() function to generate the full model with all the obtained transformed variables. Testing the first condition involves plotting the observed life expectancy against the predicted values from the linear model and checking whether there is a functional pattern. If there is, then condition 1 is satisfied. Then, pairwise plots of all predictors are drawn to check condition 2. If each pair of predictors has either a linear or random relationship, then condition 2 is satisfied.

After visually inspecting these relationships, we check the four assumptions for the linear regression. First, we use residual versus fitted value plots to check for linearity, independence and homoscedasticity assumptions. Primarily, we are looking for any observed patterns in the plot. If there are no systematic patterns, the linearity assumption is satisfied. If there is no large clustering, then independence is satisfied. If there is no fanning pattern or varying spread of the variances, then homoscedasticity is satisfied. Finally, we also use Quantile-Quantile (QQ) plots of the residuals used to verify if the normal distribution of residuals is satisfied, which we can then proceed to model construction.

We then proceed to make a model selection. We will use backward stepwise selection, removing one possible predictor one at a time. The predictor(s) to be removed are determined holistically using VIF, the significance of the predictors, and the F-test for initial assessment and model refinement. We will first examine the multicollinearity of the full model. To guarantee VIF < 5, we will choose to remove predictors with VIF > 5. We then generate the reduced model, and the linear assumptions and conditions are checked to ensure a valid comparison between the model

and the previous one. Then, whether such predictors need to be removed is judged holistically by the changes in RSE, adjusted R^2, F-test and the smallest information criteria, including AIC and BIC of the new model. We will continue the backward stepwise selection until no further predictors can be removed without worsening the new model.

After selecting the best model, we will examine leverage points, outliers, and influential points. We will measure influential points using Cook's distance. If any point in the plot falls outside of Cook's distance = 1 in the Residuals versus Leverage plot, it is considered a highly influential observation. For outliers, we will follow the standard practice and check if the standardized residual for the point falls outside the interval [-2, 2]. We will continue with the model validation if there are no compelling reasons to exclude these problematic observations. This involves building the same model on a testing dataset and comparing whether there are significant changes in model assumptions, adjusted R^2, coefficients estimates and predictors' significance. The final model is highly likely to be validated if there is no significant change.

Results

The dataset comprises 173 samples, each representing a different country with their relevant statistics. The training set contains 103 samples, while the testing set has 70.

Table 1: Numerical Summary of Variables in Training Dataset

Variable	Minimum	1st Quarter	Median	Mean	3rd Quarter	Maximum	Standard Deviation
Life Expectancy	52.53	67.58	72.54	71.91	77.39	83.78	7.098849
Daily Calorie Intake	1775	2641	2989	2992	3358	3911	468.7336
Daily Protein Intake	28.59	74.70	92.29	90.52	107.31	145.62	22.72955
Daily Fat Intake	15.51	61.89	94.90	94.50	116.73	184.40	39.06187
GDP per capita	419	5188	13839	21424	31947	104672	20725.41
Average schooling years	1.341	6.483	9.550	9.016	11.697	14.256	3.259871

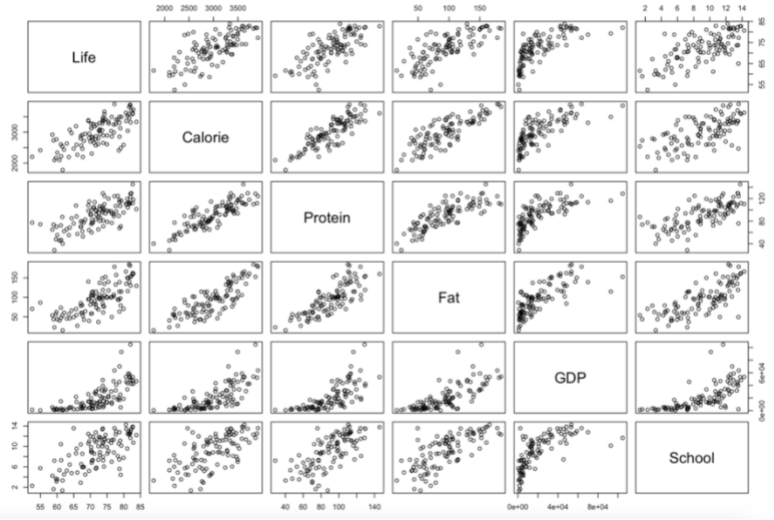
Table 2: Numerical Summary of Variables in Test Dataset

Variable	Minimum	1st Quarter	Median	Mean	3rd Quarter	Maximum	Standard Deviation
Life Expectancy	52.68	63.47	70.65	70.45	76.43	85.47	8.547081
Daily Calorie Intake	1785	2533	2840	2842	3171	3825	470.8284
Daily Protein Intake	39.21	63.53	86.92	86.07	107.66	142.81	25.03017
Daily Fat Intake	22.52	61.29	85.22	91.69	122.21	162.71	34.97328
GDP per capita	838.2	3804.3	10784.1	18604.9	25902.0	118510.0	21186.3
Average schooling years	2.319	6.691	9.217	8.989	11.412	13.904	3.042229

Tables 1 and 2 show similar descriptive statistics for both datasets, indicating that the randomization accurately split the dataset into two datasets with similar values.

The initial full model (it will be called as full_model) generated with lm() function gives an adjusted R^2 of 0.6252, with a showing only 62.52% of the variance can be explained by the model. Considering such a low adjusted R^2, we check the linearity of the predictors and the response variable, which pairwise plot is shown below.

Figure 1: Pairwise graph between the predictors and the response variable



It can clearly be seen that some predictors, such as Fat and GDP, have a nonlinear relationship with all other predictors. Hence, a transformation is needed to resolve this issue. By using common transformations and avoiding complicated transformations, the powerTransform function gives us the result that we need to square the Life, take the log of the GDP (this predictor is called logGDP), and square root the Fat (this predictor is called SqFat).

After transforming the data, the pairwise plot for the new predictors shows that all graphs are either linear or random. Hence, we create another full model using the transformed predictors (it will be called full_model1). The resulting model gives a slightly higher adjusted R^2 of 0.7013. However, RSE increases massively to 547.3 (increases more than 100 times). This is problematic as it may show that the original transformation is preferred or that we need a different transformation. Notice that in Figure 1, Life already has a linear relationship to most attributes, except Fat and GDP. However, these need to be transformed as they have a non-linear relationship to all variables. Hence, it is possible that a transformation is not needed for Life.

We create a new model with only transformed predictors (it will be called full_model2). We can see from Figure 2 that the full_model2 pairwise plot contains only linear or random relationships.

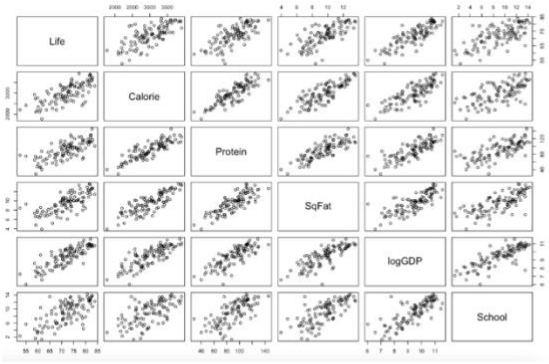


Figure 2: Pairwise Correlation Plot of Second Transformation

Table 3: Model Comparisons for Life Expectancy Predictions

Model	Adjusted R^2	RSE	AIC	BIC
full_model	0.6252	4.346	602.7963	621.2394
full_model1	0.7013	547.3	1598.935	1617.378
full_model2	0.7016	3.878	579.3193	597.7624

Moreover, the comparison of adjusted R2, RSE, AIC, and BIC of the models shows that full_model2 is preferred. Finally, the assumptions and conditions for full_model2 are satisfied. Condition 2 can be checked in Figure 2; all predictors have either linear or random relationships. The assumption checking is shown in Appendix A. Hence, we can conclude that our second transformation with our full_model2 is preferred and can best represent our model. With all assumptions being satisfied, we can proceed with the backward stepwise selection.

Table 4: VIF and P-value for Predictors in full_model2

Predictor	VIF	P-value
Calorie	4.911	0.967
Protein	4.739	0.756
SqFat	4.267	0.109
logGDP	4.755	9.59×10^{-8}
School	3.107	0.814

As shown in Table 4, all predictors in full_model2 have VIF < 5, showing little multicollinearity. However, the P-value of Calorie is the highest being 0.967, showing it is very insignificant. Since it also has the highest VIF, we will see if a reduced model without Calorie is preferred.

A reduced model without calories (it is called reduced_model1) is created. All the adjusted R^2, RSE, AIC and BIC, as well as the F-test by calling anova(reduced_model1, full_model2), shows that reduced_model1 is a more preferred model than full_model2, as shown in Table 5.

Table 5: Statistical Comparison of full_model2 and reduced_model1

Model	Adjusted R^2	RSE	AIC	BIC	F-test (p-value)
full_model2	0.7016	3.878	579.3193	597.7624	0.0018 (0.9666)
reduced_model1	0.7046	3.858	577.3212	593.1296	-

After another check of all the assumptions and conditions for our new reduced_model1, we reject full_model2 and conclude that reduced_model1 is more suitable. We repeat until we can no longer remove a predictor without generally worsening adjusted R^2, RSE, AIC, and BIC. The final model satisfies all the assumptions and contains predictors SqFat and logGDP.

Now, we proceed to look for outliers and influential points. Detailed plots of outliers and influential points for both training and testing data can be seen in Appendix B. Some outliers exist in the training model. They are slightly out of the acceptable range of standard residual. However, considering there is no high influential point (with Cook's distance < 0.5), we conclude that we can keep the outlier as such point does not significantly affect our result. Therefore, we can continue validating our model using the final data.

We first check that all assumptions and conditions are satisfied. Despite some data being slightly out of the acceptable range of standard residual, we conclude that we can keep the outlier as these points do not significantly because there is also no high influential point in the test data (with Cook's distance < 0.5). Finally, we can generate the model for test data using the predictors we obtained in the final model and give the following table for the estimates, standard errors and significance, and the adjusted R^2 and RSE for both training and testing data.

Table 6: Detailed Model Comparison for Training and Test Data

Model	Adjusted R^2	RSE	Predictor Details			
			Statistics	Intercept	SqFat	logGDP
Training	0.7099	3.823	Estimate	27.6961	0.6708	4.0228
			Std. Error	3.1143	0.2940	0.5043
			P-Value	2.63×10^{-14}	0.0246	2.55×10^{-12}
Test	0.6917	4.746	Estimate	17.2808	1.0636	4.6766
			Std. Error	4.8030	0.5448	0.8677
			P-value	0.000609	0.055095	9.86×10^{-7}

Even though not all coefficient estimates in both models are similar up to their standard error, most coefficients still have close estimates within one standard error, and changes in estimates and the P-value for all coefficients aren't significant. Moreover, adjusted R^2 and RSE also give

similar values in both models. Based on these results, we conclude that the final model will likely be validated.

Discussion:

Interpretation

In this model, we observe that one unit increase of log of GDP per capita resulting in 4.6766 increase of average population's life expectancy of the country and one unit increase of square root of daily fat intake resulting in an increase in the same response variable of 1.0636. The intercept of 17.2808 represents that a country with all the people having no daily fat consumption and a GDP per capita of 1 would have an average population life expectancy of 17.2808 years.

While it may seem counterintuitive that dietary fat was included as a significant positive factor in increasing life expectancy, the recent study of the Japanese community can offer some perspective. Tamura et al. (2023) discovered that consuming a low-fat diet increases mortality rate from all causes, including cancer in women. We believe our study helps to better understand the relationship between fat consumption and longevity, leading to dietary consumption that can improve the overall quality of life of the population. It also reveals how economic situations affect people's lifespan.

Limitations

Sample Size: As our study used data representing the general national average, our sample size is limited to the number of countries with existing datasets. Also, this data collection method results in our model, which can only predict the average life expectancy of a country based on its GDP per capita and daily fat consumption rather than at an individual level. This is because obtaining a large amount of data at the individual level is difficult and impractical. Therefore, future studies could employ more specific data at a local or individual level to increase the sample size and could potentially provide improved and more specific results.

Variance Assumption: Although homoscedasticity assumption is satisfied in our model, we cannot rule out variance anomalies due to the nature of aggregated national data. More importantly, as we are taking the average of each country's data, such data does not apply to every person in this country. Such inequality, which influences the variance of our data, is different for each data sample. Moreover, varying population for each country gives unequal weights for these unequal variances. Therefore, techniques such as collecting data on the individual level could be considered in future studies to account for the variance in the data collection that cannot be detected through conventional assumption testing.

Data Extrapolation Impracticality: When using our linear regression model for predicting, the intercept and extrapolation features demonstrate necessary restrictions when such a model is used on value outside the practical data scope. For example, if we consume 100 units of square root grams of fat daily (10,000 grams of fat per day, 20 times the standard recommendation), we will exceed 100 years, which is very unlikely. It is important to note that our model works with a practical data set with a very limited range for the predictors. Moreover, under such a small range of predictors, the graph does behave close to a line, which can be told from the data that it follows the assumptions and conditions of a linear regression.

References:

Li, Y., Schoufour, J., Wang, D. D., Dhana, K., Pan, A., Liu, X., Song, M., Liu, G., Shin, H. J., Sun, Q., Al-Shaar, L., Wang, M., Rimm, E. B., Hertzmark, E., Stampfer, M. J., Willett, W. C., Franco, O. H., & Hu, F. B. (2020). Healthy lifestyle and life expectancy free of cancer, cardiovascular disease, and type 2 diabetes: Prospective cohort study. *BMJ*, l6669. https://doi.org/10.1136/bmj.l6669

Miladinov, G. (2020). Socioeconomic development and life expectancy relationship: Evidence from the EU Accession Candidate countries. *Genus*, *76*(1). https://doi.org/10.1186/s41118-019-0071-0

Tamura, T., Wakai, K., Kato, Y., Tamada, Y., Kubo, Y., Okada, R., Nagayoshi, M., Hishida, A., Imaeda, N., Goto, C., Ikezaki, H., Otonari, J., Hara, M., Tanaka, K., Nakamura, Y., Kusakabe, M., Ibusuki, R., Koriyama, C., Oze, I., … Matsuo, K. (2023). Dietary carbohydrate and fat intakes and risk of mortality in the Japanese population: The Japan multi-institutional collaborative Cohort study. *The Journal of Nutrition*, *153*(8), 2352–2368. https://doi.org/10.1016/j.tjnut.2023.05.027

United Nations. (2015b). *Transforming our world: The 2030 agenda for sustainable development department of economic and social affairs*. United Nations. https://sdgs.un.org/2030agenda

United Nations. (2020, September 15). Human development report 2020. Human Development Reports. https://hdr.undp.org/content/human-development-report-2020

Appendix A: Predicted values versus observed values plot (A1), residuals versus fitted values plot (A2), Normal QQ plot (A3) on full_model2

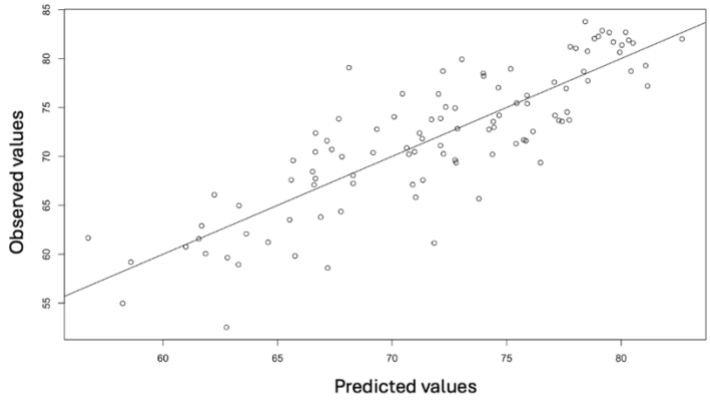


Figure A1: Predicted versus Observed value of full_model2

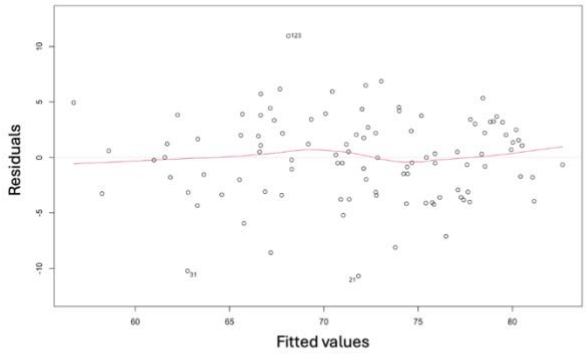


Figure A2: Residual versus Fitted values of full_model2

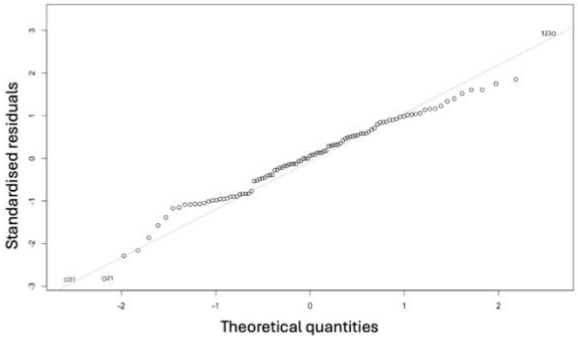


Figure A3: Q-Q Plot of full_model2

Appendix B: Training data standardized residuals versus fitted values plot (B1) and standardized residuals versus leverage plot (B2). Testing data standardized residuals versus fitted values plot (B3) and standardized residuals versus leverage plot (B4).

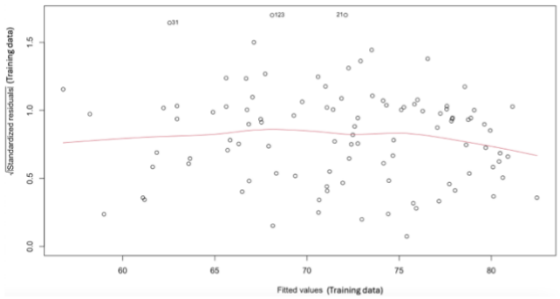


Figure B1: Residual versus Fitted values of training

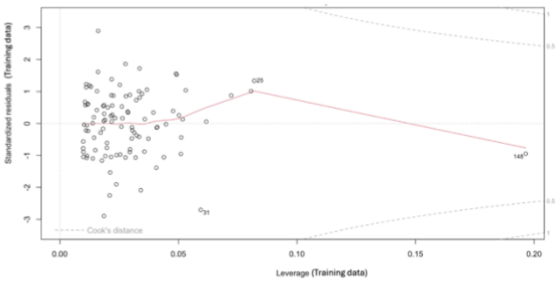


Figure B2: Residual versus Leverage of training

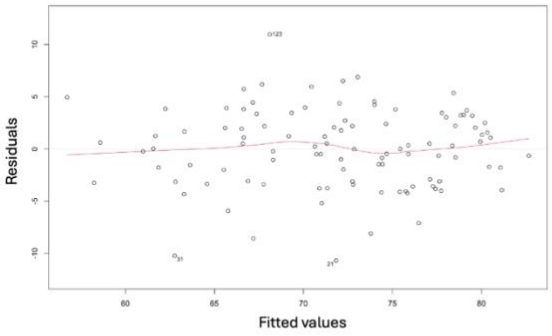


Figure B3: Residual versus Fitted values of testing

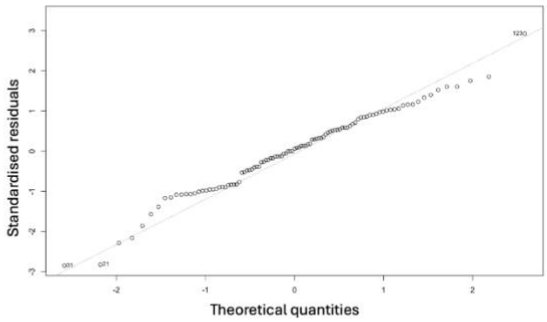


Figure B4: Residual versus Leverage of testing