MA317 - Modelling Experimental Data

December 14th, 2022

Designing a predictive model for life expectancy in 2020

By: 2200634 and 2200909

Contents

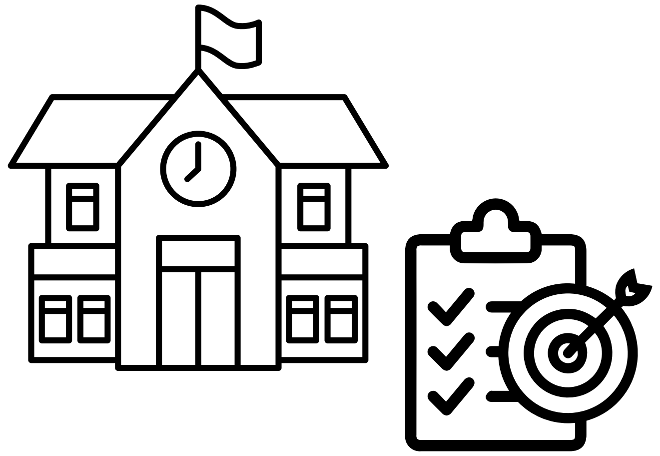
Introduction

What?

- Analyze the most significant **variables** that help us accomplish this and suggest the best **linear model** that predicts **life expectancy** in 2020.

Why?

- Beneficial for institutions, such as governments, to make more suitable and informed decisions.

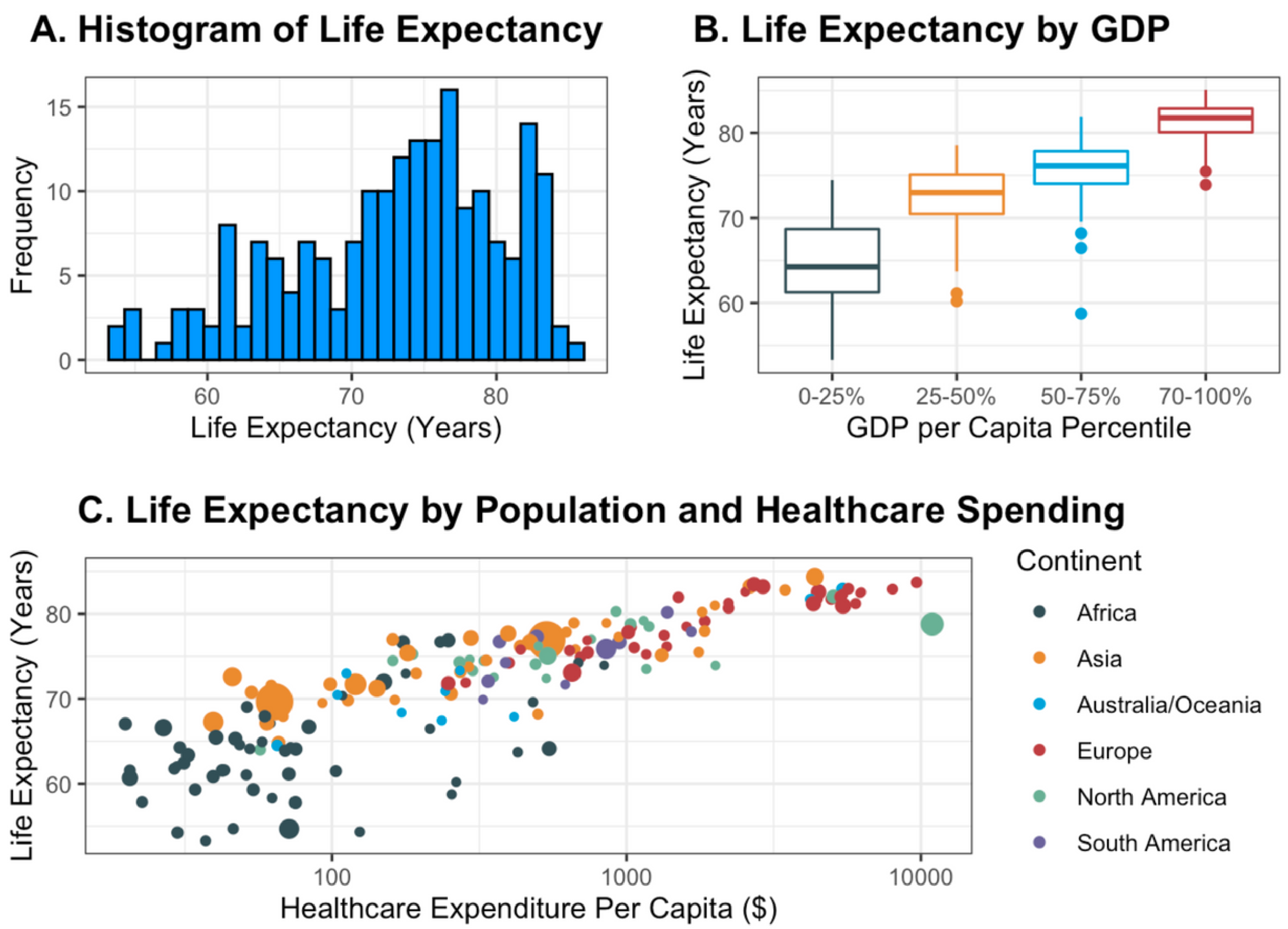


How?

- Firstly, by describing and analyzing the data to solve any problems we might face - such as missing values. Secondly, by investigating the collinearity between the predictor variables to see which method we can use. Finally, by employing an appropriate experimental design to study differences in average life expectancies across the continents.

Descriptive Statistics

- 27 features (including economic, demographic, educational, and health-related areas).
- 217 countries.
- Variable of interest: life expectancy (years).
- 19 (8.8%) of the countries were missing the life expectancy variable.
- Maximum life expectancy: 85.08 years (Hong Kong) and minimum is 53.28 years (Central African Republic).
- Countries that spend a higher percentage of their GDPs on healthcare have higher life expectancies.
- The relationship between GDP per capita and life expectancy is log-linear. Thus, the GDP and the health expenditure variables were log-transformed.



Variables	Missing
Children newly infected with HIV	58.5%
Educational attainment, primary	83.4%
Educational attainment, bachelor's	82.5%
Literacy rate	88.5%
Poverty ratio	89.9%
Renewable energy consumption	100%

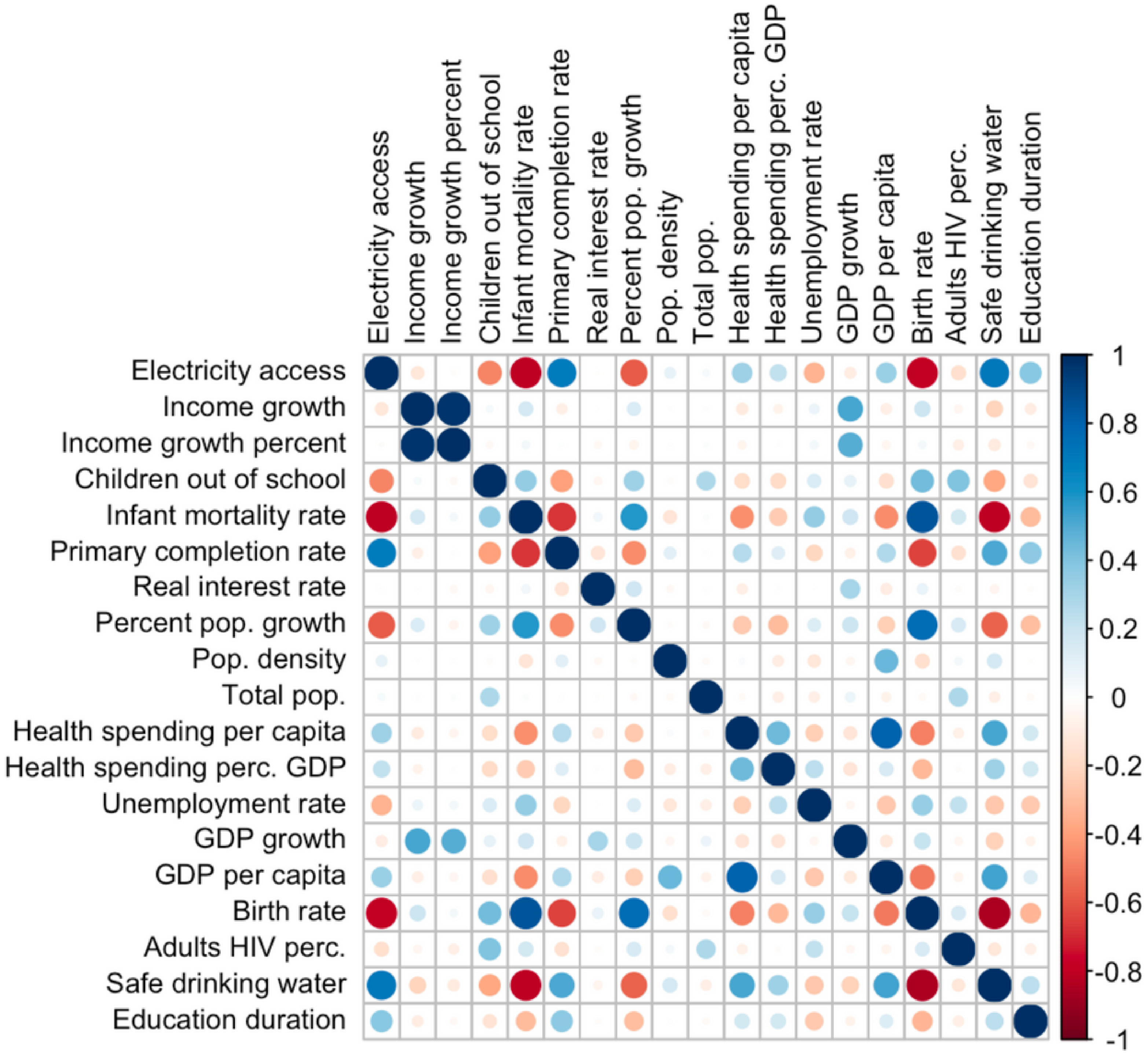
Table 2. Percent missing for five variables with greatest missingness

- Variables contained missing data at high rates.
- Six had missing entries for over 50% of countries.
- These six features were removed before imputation and ensuing analysis because their extreme high rates of missingness

Handling missing values

- The rest of the missing data was addressed using multiple imputation, "mice()" package in R.
- It is important to remember that this method of multiple imputation relies heavily on the assumption that the data are missing at random (MAR).
- Key step: adjust the predictor matrix to remove life expectancy as a predictor for any of the other features (prevents baking a direct predictor-response relationship into the data + keeps the imputation applicable for future predictions)

Investigating Collinearity



- If collinearity left unaddressed, it could lead to models with inflated variances and hinder statistical inference.
- The pairwise correlations between features were analyzed in addition to the variance inflation factors.
- This analysis revealed a number of variables that have high VIFs and correlations with |value| > 0.8.
- Values removed:
  - National net income growth - high correlation with national net income growth per capita (r = 0.979)
  - Birth rate because of its high correlation with population growth rate (r = 0.76), infant mortality rate (r = 0.86) and safe drinking water rate (r = - 0.85)
- Removing these helped to address the issue of multicollinearity which makes the coefficients and their standard errors more appropriate.

A linear model to predict life expectancy

TECHINQUE

Forward stepwise regression technique was implemented on the 10 imputed datasets using the "stats()" R package.

ANALYSIS

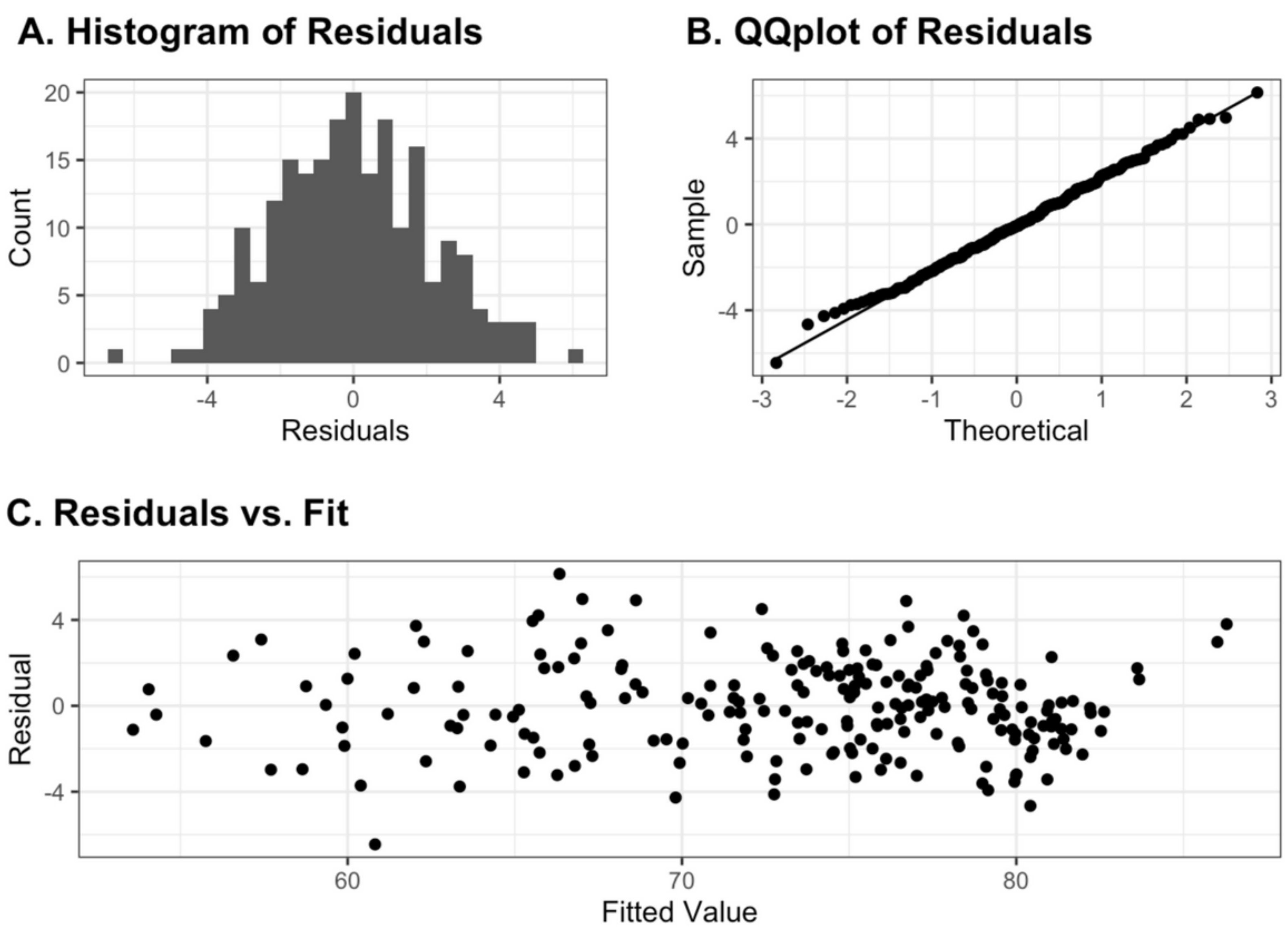
We analyzed the proportion of times each feature appeared in the subsets and that is how the "best" model was selected.

THE VARIABLES

Five characteristics: health expenditure, infant mortality rate, % drinking water, population density, and log(GDP) appeared in 100% of the subsets and, furthermore, four of the five characteristics were associated in a statistically significant way with the life expectancy at the 5% level.

RESIDUALS

The shown figure illustrates that the residuals are approximately normally distributed and shows that there is no clear pattern. Together, these results reassure that the linearity, homoscedasticity, and multivariate normality assumptions for linear regression are met.

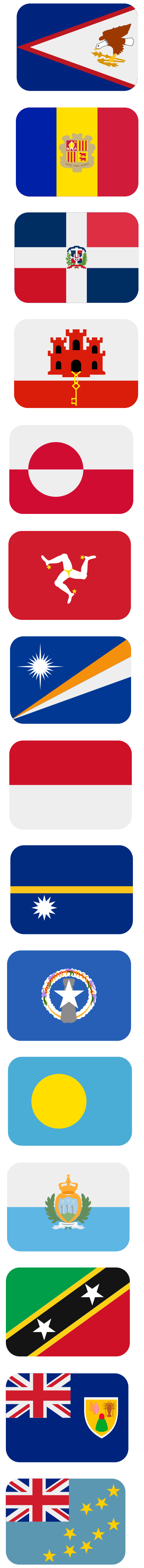


Predicting life expectancy in other countries

To demonstrate how this linear prediction model can be implemented, predictions for life expectancy were generated for the 19 countries that were missing the variable.

Thus, predictions were made using the best model detailed above on the imputed datasets because some of the countries were initially missing predictor values.

This table shows the mean predicted life expectancy across the 10 imputations for these 19 countries.



Country	Predicted life expectancy (years)
American Samoa	77.11
Andorra	80.19
British Virgin Islands	74.92
Cayman Islands	81.04
Curacao	77.11
Dominica	70.31
Gibraltar	78.36
Greenland	79.11
Isle of Man	80.55
Marshall Islands	72.36
Monaco	87.94
Nauru	73.08
Northern Mariana Islands	77.53
Palau	76.96
San Marino	80.88
Sint Maarten (Dutch part)	75.77
St. Kitts and Nevis	76.55
Turks and Caicos Islands	78.35
Tuvalu	75.57

Table S2. Predictions for countries missing life expectancy variable based on pooled coefficients from best linear model

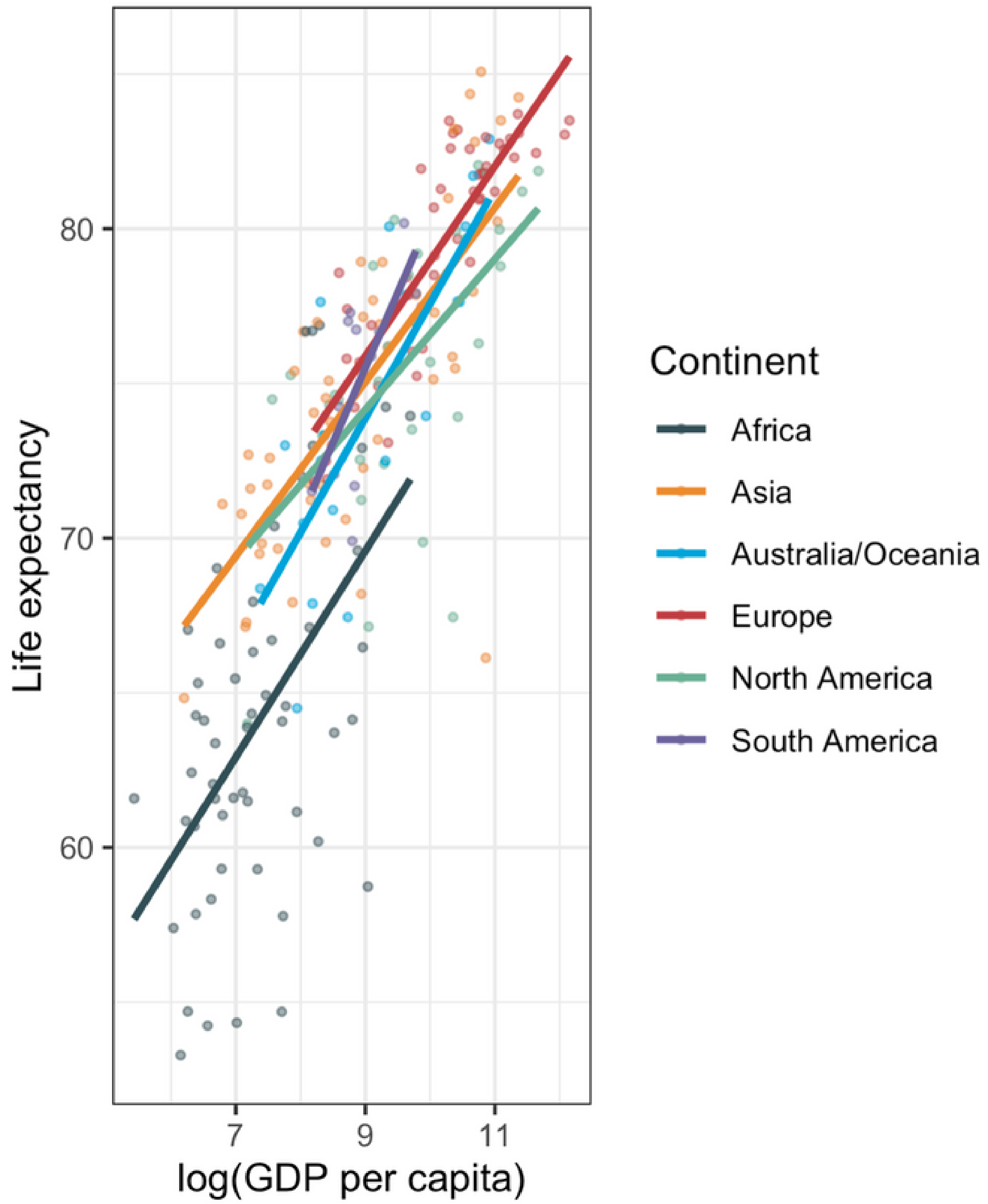
Experimental design to study life expectancy across continents

- To study the differences in life expectancy across continents (a factor variable), an ANCOVA experimental design was employed.
- ANCOVA is designed specifically for testing whether there is a difference in means for a continuous response variable (in this case, life expectancy) across a categorical predictoor (continent), and it has the additional benefit of controlling for covariates that are independent of the predictor.
- In the experimental design, an ANCOVA was performed with GDP per capita as a covariate.
- There was a clear association between GDP per capita and life expectancy. Thus, they were used as covariates in the experimental design to ask the question whether there is a difference in mean life expectancy across continent even after adjusting for GDP.

Source of Variation	Sum of Squares	df	Mean Squares	F	p-value
Log(GDP per capita)	7682	1	7682	581.82	< 0.0001
Continent	1190	5	238	18.03	< 0.0001
Residuals	2273	2273	13		
Total	11145	2279			

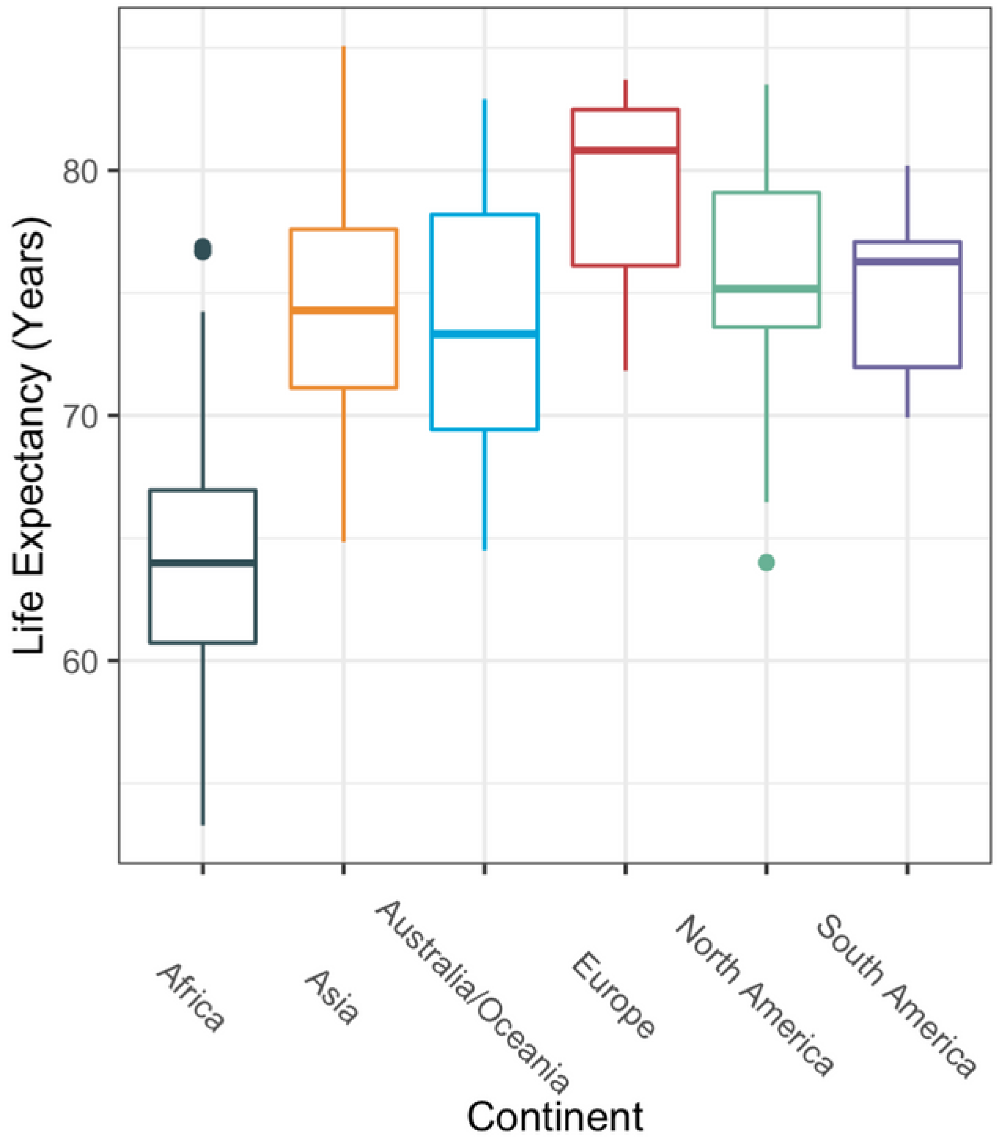
Table 4. ANCOVA Results

Results: life expectancy across continents



After log-transforming GDP per capita, it appears to have a linear relationship with life expectancy that is homogenous across continents.

There appear to be some differences in life expectancy by continent, and the variance looks fairly similar across continents.



Conclusion

- This report has described the methods and results used to create a linear predictive model to predict life expectancy among countries and compare life expectancy across countries in different continents.
- A linear model using the listed variables was found to explain 89.0% of the variance in life expectancy in the dataset.

INFANT MORTALITY RATE

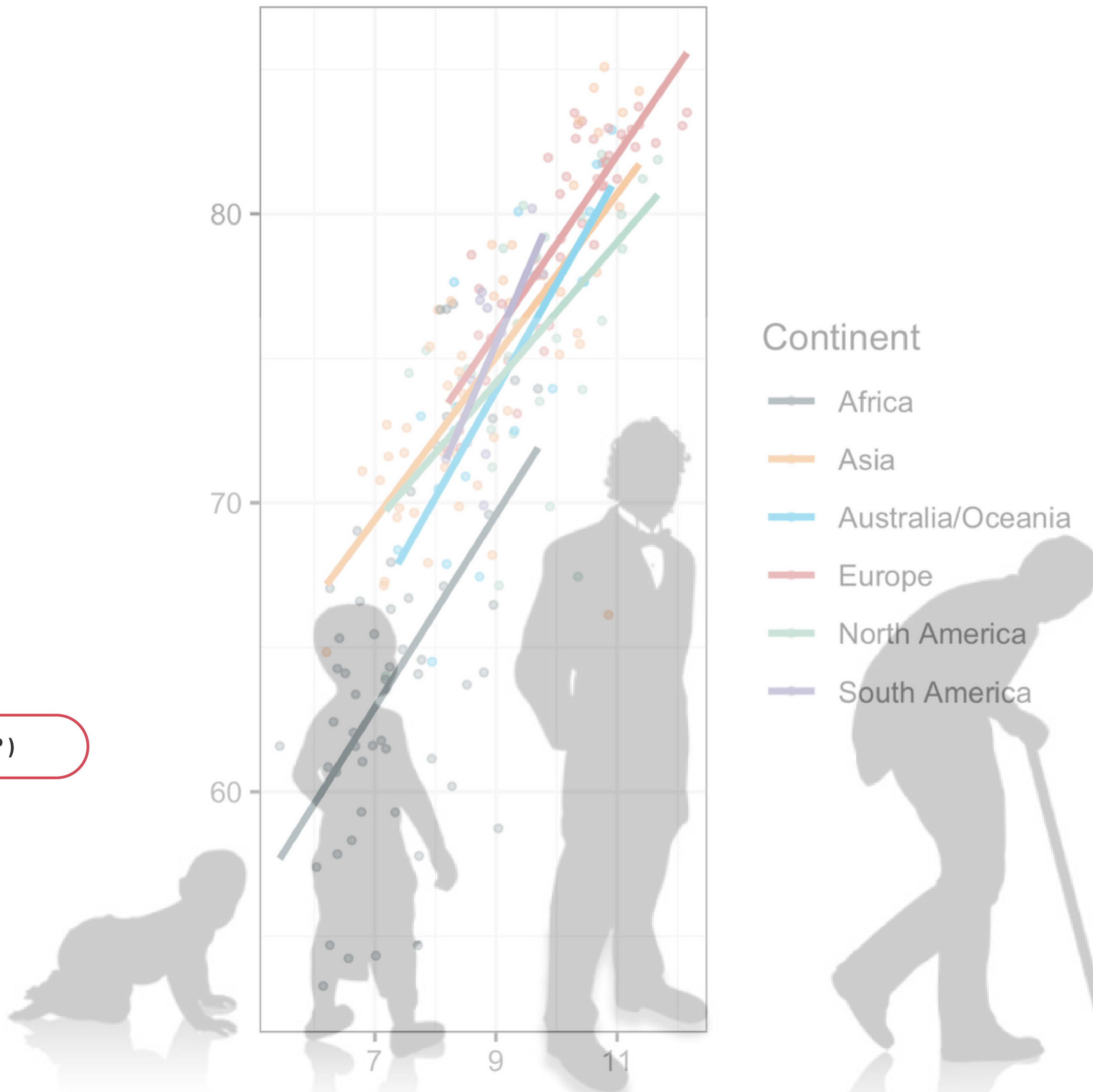
SAFE DRINKING WATER %

POPULATION DENSITY

LOG(GDP PER CAPITA)

HEALTH EXPENDITURE (AS % OF GDP)

- It was then shown that even after controlling for log(GDP per capita), there is evidence that mean life expectancy differs across continents (below 5% significance level).
- It is important to note that although the models described in this report were shown to be useful for *prediction* the observational nature of data collection means that the findings cannot be interpreted as *causal*.



Thank you!

References

- Janssen, K. J., Donders, A. R. T., Harrell Jr, F. E., Vergouwe, Y., Chen, Q., Grobbee, D. E., Moons, K. G. (2010). Missing covariate data in medical research: to impute is better than to ignore. Journal of clinical epidemiology, 63(7), 721-727.
- Azur, M. J., Stuart, E. A., Frangakis, C., Leaf, P. J. (2011). Multiple imputation by chained equations: what is it and how does it work?. International journal of methods in psychiatric research, 20(1), 40-49.
- Van Buuren, S., Groothuis-Oudshoorn, K., Robitzsch, A., Vink, G., Doove, L., Jolani, S. (2015). Package 'mice'. Computer software.
- Van Buuren, S. (2018). Flexible imputation of missing data. CRC press.
- R-bloggers (2021). How to perform ANCOVA in R. Available at: https://www.r-bloggers.com/2021/07/how-to-perform-ancova-in-r/

Images:

- Global Education Project Earth. Available at: Life Expectancy, Food and Hunger, Access to Safe Water, AIDS, Population, and Human Conditions, drugs, suicide - The Global Education Project
- Josoft (2017). Steemit. HUMAN LIFE CYCLE. Available at: https://steemit.com/education/@josoft/human-life-cycle