Zachary Zusin
Linear Regression Models
Professor Steven Campbell
December 12, 2023

Introduction

The primary aim of this report is to assess whether we can construct a statistical model which accurately depicts the relationship between life expectancy and a variety of other variables (all of which pertain to a nation's social, economic, political, or geographic status). Such a model would allow us to evaluate whether there is in fact a statistically significant relationship between life expectancy and democracy index in particular, our first research question, and give us the ability to predict life expectancy based on values for our other variables more generally, our second research question. Understanding this connection will provide key insights into the social, political, geographic, and economic determinants of health outcomes at a national level.

Data Description

The dataset, which mostly comes from the CIA's "The World Factbook," comprises information on 166 countries, each of which is characterized by a diverse set of variables. These variables range from demographic indicators like population and birth rates to economic metrics such as GDP and health expenditure.

Table 1. Variables included in the initial dataset

Variable Name	*Variable Description*
name	country name
population	population
birth_rate	number of births per 1,000 people
death_rate	number of death per 1,000 people
infant_mortality_rate	infant deaths per 1,000 live births
life_exp_at_birth	overall life expectancy in years
life_exp_at_birth_m	male life expectancy in years
life_exp_at_birth_f	female life expectancy in years

gdpPPP	gross domestic product purchasing power parity in USD
gdpPPP_percap	gross domestic product purchasing power parity per capita in USD
labor_force	domestic labor force
land_area	land area in square kilometers
coastline	coastline in kilometers
land_use_agricultural	percent of total land used for agriculture
urbanization	percent of total population living in urban area
refined_petrol_consumption	petrol consumption in barrels per day
co2_emmissions_energy_consumption	CO_2 emissions from energy consumption in metric tons
airports	number of airports
region	geographic region
roadways	total length of roadways in kilometers
democracy_index	average of the 5 measures that follow
electoral_process_and_pluralism	measure of democracy
function_of_government	measure of democracy
political_participation	measure of democracy
political_culture	measure of democracy
civil_liberties	measure of democracy
regime_type	classification of government type
continent	continent
health_spend_pct_gdp	health expenditure as a percentage of GDP

However, to enhance the model's predictive power, many of the original variables were removed and others were augmented. Firstly, preliminary modifications were made to enhance the relevance and informativeness of the regression analysis. The variables `life_exp_at_birth_m`, `life_exp_at_birth_f`, and `infant_mortality_rate` were excluded as they are obviously highly correlated with life expectancy and would not provide any meaningful insights if included in the model. Also, the `name` variable was removed for the names of countries deemed irrelevant to the analysis. To avoid redundancy, the `gdpPPP` variable was omitted due to its similarity with and being less informative than `gdpPPP_percap`. Additionally, variables with non-numerical entries, such as `region`, `regime_type`, and `continent`, were eliminated as it would be difficult to perform a regression on them. Lastly, the removal of `electoral_process_and_pluralism`, `function_of_government`, `political_participation`, `political_culture`, and `civil_liberties` was justified by their redundancy since `democracy_index` already represented their average.

Secondly, many of the variables had a lot of values clumped together, so it made sense to apply a logarithmic transformation to them, spreading out the values more evenly. Specifically, these variables were `population`, `gdpPPP_percap`, `labor_force`, `land_area`, `coastline`, `land_use_agricultural`, `refined_petrol_consumption`, and `co2_emisssions_energy_consumption`. However, for `coastline`, since many of the values were zero (landlocked nations), taking the logarithm results in a value of *-inf* which cannot be regressed on, so ultimately `coastline` was removed given these difficulties and the unlikeliness that this variable is greatly correlated with life expectancy.

Also, since `airports` and `roadways` are essentially both measures of transportation infrastructure within a nation, the joint effect of them seemed to be more reasonable to investigate the effect of each one separately, so an interaction variable of their product `airports_roadways_interaction` was used instead of either.

Lastly, the datapoint for North Korea was excluded due to a missing value in the `health_spend_pct_gdp` variable. Rather than remove an entire variable, along with the 165 values associated with it, less information would be removed from the model if only a single country was removed instead. These modifications aim to ensure that only meaningful and non-redundant variables contribute to the subsequent regression analysis.

Methodology

After narrowing down the covariates to those which could reasonably be included in our model, the best subsets procedure was employed with three different selection criteria independently. Using the Mallows' C_p criterion, the chosen predictors were:`[population_log, birth_rate, death_rate, gdpPPP_percap_log, labor_force_log, land_area_log, land_use_agricultural, urbanization, co2_emisssions_energy_consumption_log, airports_roadways_interaction_log, democracy_index, health_spend_pct_gdp]`. When Akaike's Information Criterion was used instead, the same set of variables were chosen, but when the Bayesian Information Criterion was used, the following smaller set of covariates were selected: [birth_rate, death_rate, gdpPPP_percap_log, urbanization, democracy_index, health_spend_pct_gdp]. Thus, we now have two models which we can further examine, the better of which will be chosen as our recommended model. We then fitted two linear regression models, each expressing life_exp_at_birth as a function of our two sets of independent variables respectively.

Results

Upon examining both our models, we can see that they exhibit similar performance based on various metrics, including their adjusted R^2 values and p-values. Given their nearly identical predictive capabilities, our preference leans towards the model with fewer dependent variables as it attains equivalent results while maintaining a higher degree of simplicity. We present the exact values of our model below.

Table 2. Calculated values of the linear regression model

Variable	*Coefficient Estimate*	*P value*
Intercept	69.26774	$< 2 \times 10^{-16}$
birth_rate	-0.46057	$< 2 \times 10^{-16}$
death_rate	-0.82047	$< 2 \times 10^{-16}$
gdpPPP_percap_log	1.24391	1.23×10^{-3}
urbanization	0.03351	2.7054×10^{-2}
democracy_index	0.46746	2.53×10^{-4}
health_spend_pct_gdp	0.34504	3.58×10^{-4}

Upon examining the model, it is evident that the selected variables, including birth rate, death rate, log-transformed GDP per capita, urbanization, democracy index, and health expenditure as a percentage of GDP, collectively contribute significantly to explaining variations in life expectancy at birth. The model demonstrates a strong statistical significance, as indicated by the low p-values for each predictor. The adjusted R^2 value of 0.891 implies that approximately 89.1% of the variability in life expectancy at birth is accounted for by the chosen variables. These findings suggest that this simple model effectively captures the essential relationships between the selected predictors and life expectancy at birth, providing a concise yet powerful tool for understanding and predicting life expectancy trends across the observed countries. For a detailed exploration of technical aspects and additional information, please refer to the accompanying appendix.

Discussion and Conclusion

Our findings affirm the existence of a statistically significant relationship between the democracy index and life expectancy. The coefficient estimate for the `democracy_index` variable is 0.46746, and its associated p-value is 2.53×10^{-4}, which means that, holding all other factors constant, a one-unit increase in the democracy index is associated with an average increase of 0.46746 years in life expectancy at birth. The statistical significance of the coefficient implies that this relationship is unlikely to have occurred by chance, so we can confidently conclude that countries with higher democracy index scores tend to exhibit higher life expectancies. Thus, the findings suggest that democratic governance plays an important role in positively influencing the life expectancy of a nation's population.

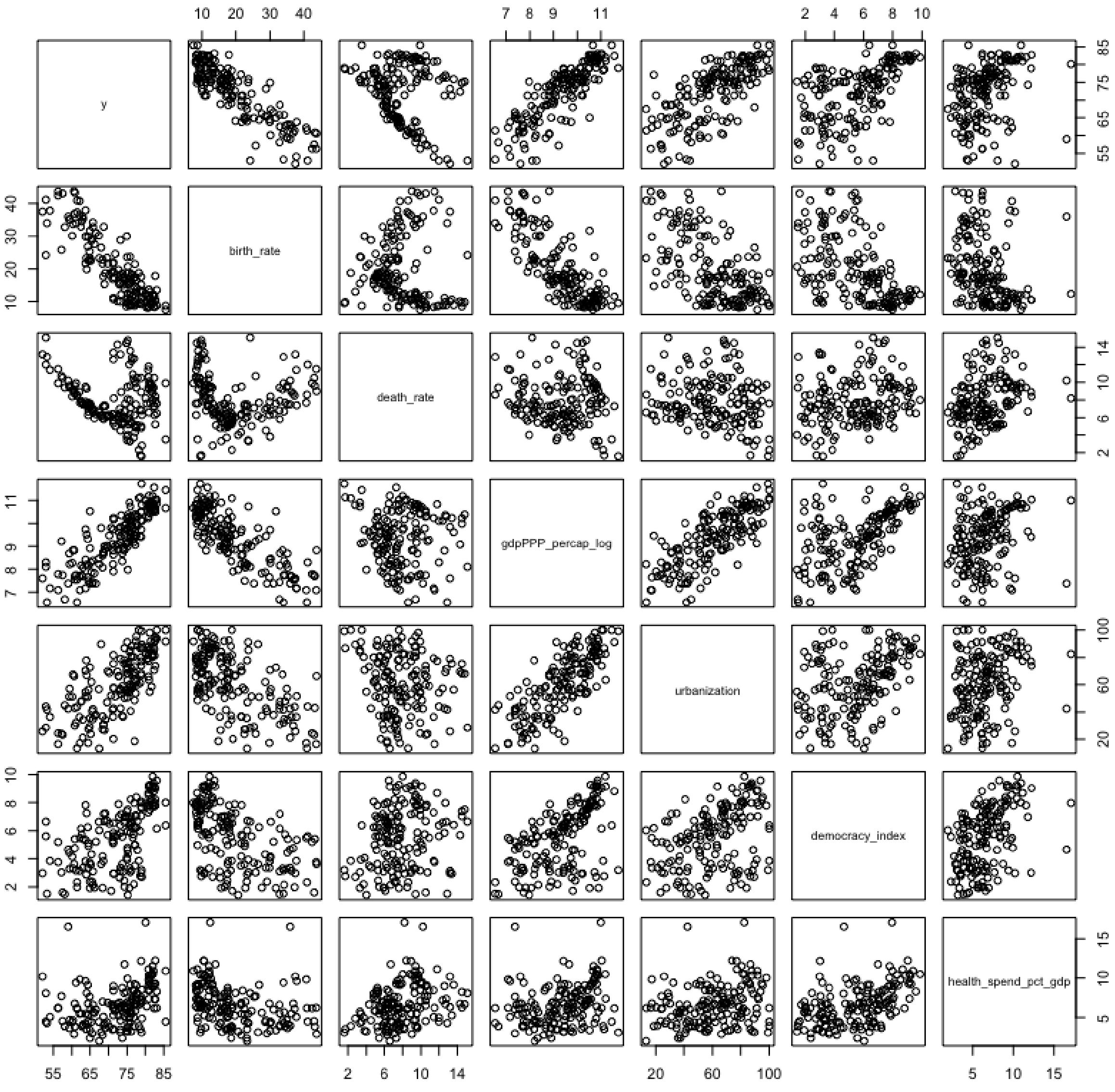
The same model which allowed us to conclude there is a statistically significant relationship between democracy index and life expectancy can be used to predict life expectancy. Our rather simple model makes use of just six key predictors but still very accurately estimates an individual's life expectancy given their nation's birth rate, death rate, GDP per capita, degree of urbanization, degree of democratization, and health expenditure as a percentage of GDP.

While the constructed regression model provides valuable insights into the relationship between life expectancy and various socio-economic factors, it is crucial to acknowledge certain limitations in our analysis. Firstly, the chosen predictors represent only a subset of potential determinants, and there may be variables that significantly impact life expectancy that we chose to exclude from the model. Additionally, our model assumes a linear relationship between the predictors and life expectancy, potentially oversimplifying more complex interactions that may be better modeled with a different form of regression. Despite these limitations, the models contribute valuable insights, and future research can build upon these foundations to enhance the understanding of the many determinants of life expectancy.

Appendices

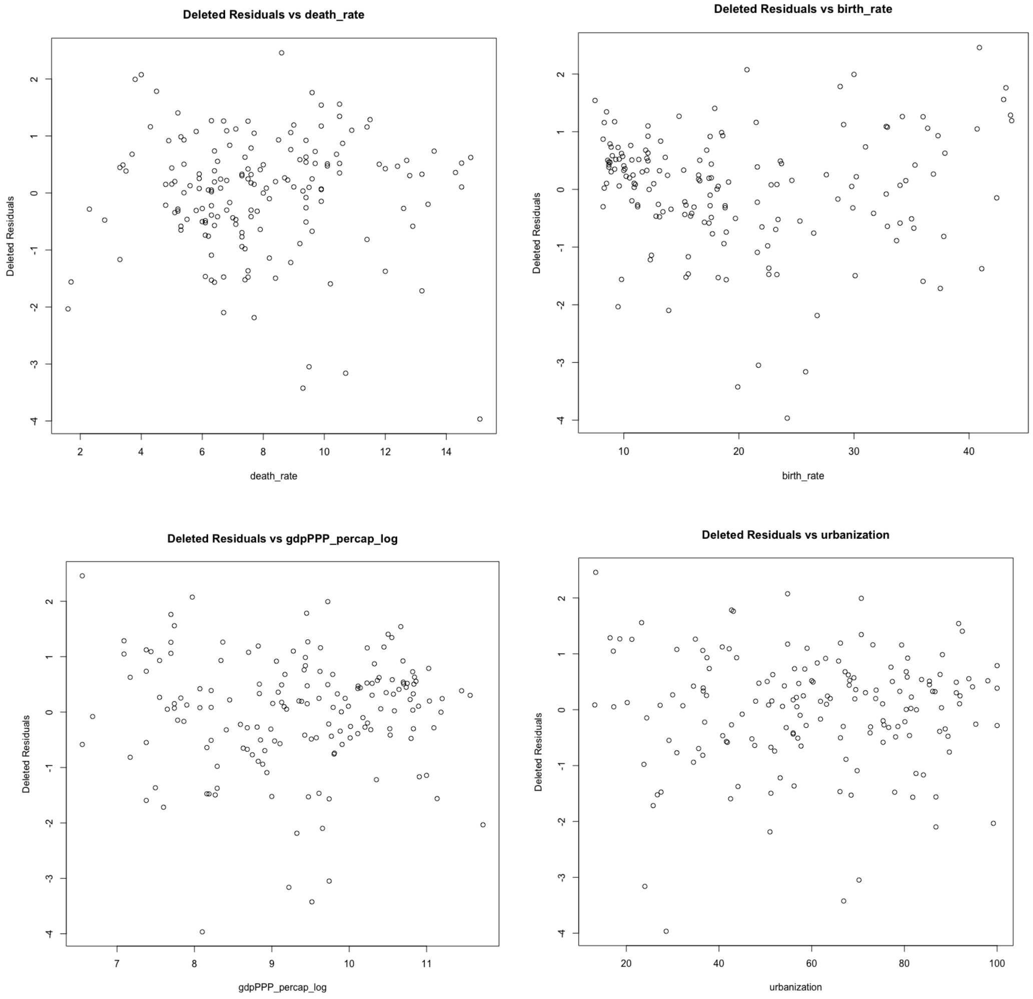
Appendix A: Model Selection

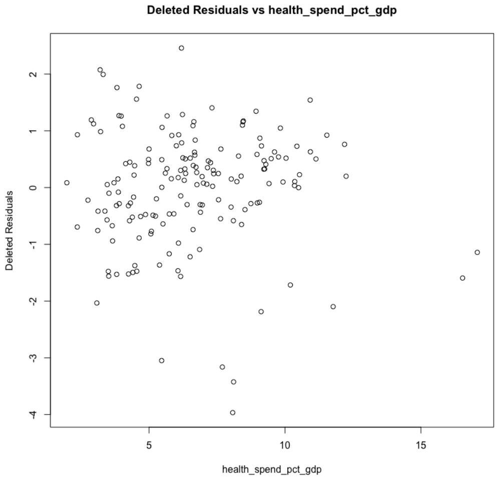
Figure 1: Scatter plots of all pairs of variables in the model

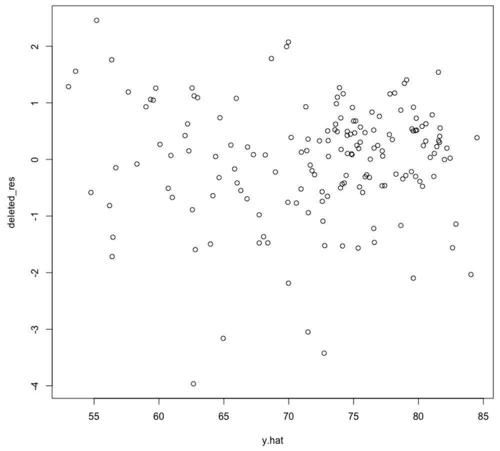


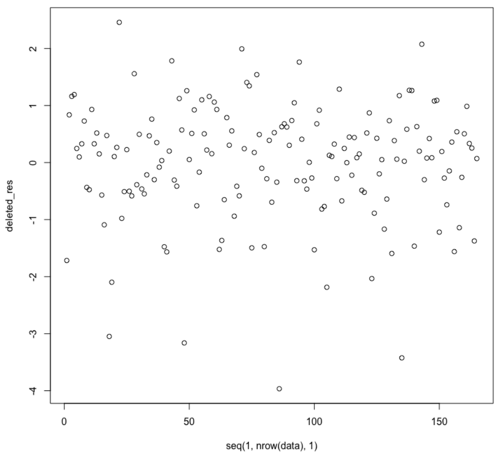
When it comes to our model selection, several diagnostic plots were generated to examine how our covariates were related to our response variable and to each other so as to ensure our regression would make sense. We see that the graphs in the first row, those plotting our response variable against each of the dependent variables, exhibit mostly linear relationships which is desirable as it means our choice of a linear regression model will be appropriate. Now, looking at the plots to the right of the diagonal and below the first row, we ideally would have no clear relationships of any kind so as to avoid multicollinearity. This is true for most of the plots, however some still exhibit a linear relationship. We have done the best we can by transforming many of the variables with a logarithmic transform and using the best subsets procedure, but sometimes multicollinearity is not completely avoidable if we wish to keep our selected set of covariates.

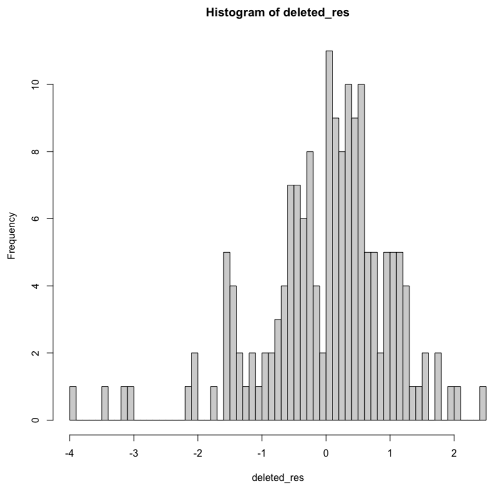
Figure 2: An assortment of scatter plots comparing deleted residuals and our covariates

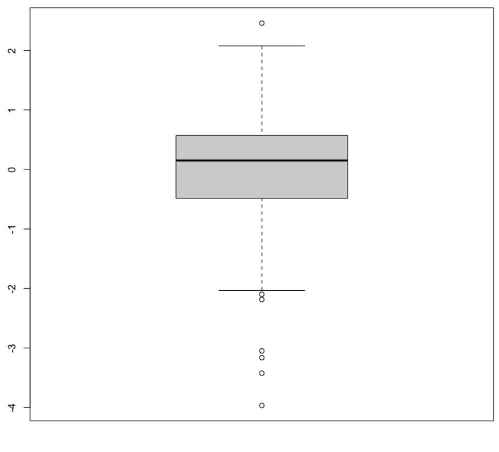


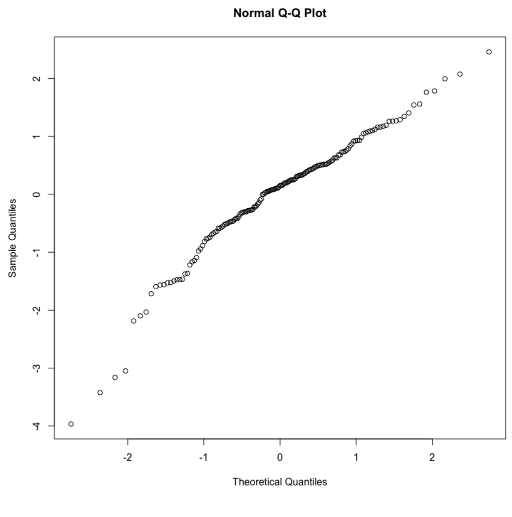












The diagnostic plots above of deleted residuals vs each of the covariates (and time) exhibit a uniform band of points, the histogram of the deleted residuals is normally distributed, and the QQ plot is almost linear. All of these are ideal as indicate the residuals are normally distributed, linear, have constant variance, and are independent of each other.

Table 3. Full summary of statistics generated by the model

Variable	*Standard Error*	*Coefficient Estimate*	*T value*	*P value*
Intercept	3.67633	69.26774	18.842	< 2 x 10^{-16}
birth_rate	0.03544	-0.46057	-12.996	< 2 x 10^{-16}
death_rate	0.08421	-0.82047	-9.743	< 2 x 10^{-16}
gdpPPP_percap_log	0.37795	1.24391	3.291	1.23 x 10^{-3}
urbanization	0.01502	0.03351	0.027054	2.7054 x 10^{-2}
democracy_index	0.12486	0.46746	3.744	2.53 x 10^{-4}
health_spend_pct_gdp	0.09458	0.34504	3.648	3.58 x 10^{-4}

Residual Standard Error	2.624 on 158 degrees of freedom
Multiple R-squared	0.895
Adjusted R-squared	0.891
F-statistic	224.5 on 6 and 158 DF
P-value	< 2.2 x 10^{-16}

The above table indicates that estimates for the intercept and coefficients of `birth_rate`, `death_rate`, `democracy_index`, `health_spend_pct_gdp` are significant with alpha = 0.001, the estimate for the coefficient of `gdpPPP_percap_log` is significant with alpha = 0.01, and the estimate for the coefficient of `urbanization` is significant with alpha = 0.05. So, with any reasonable choice of alpha value, all of our coefficient estimates are statistically significant.

Appendix B: Model Validation

After performing data splitting, training the model on 80% of our data and testing it on the other 20%, the model was able to achieve a mean squared prediction error of 7.0131. This means on average, our predictions of life expectancy were off by approximately 2.64 years, which is a rather small amount in the context of the average 80 year old lifespan. Additionally, after conducting k-fold cross validation with k = 5, an average mean squared prediction error of 7.1762 was achieved, further reinforcing the fact that our model will on average predict a life expectancy between 2 and 3 years off of the actual value.