Science Program Comprehensive Assessment (ESP Project)

Massimo Caruso & Nicholas Gambino
2072289 & 2070757

Probability and Statistics

201-HTH-05 Section 00001

Presented to Ivan Ivanov

Thursday, December 15, 2022

Table of Contents

1.0 Abstract

It has become increasingly common for humans to live past 70 years of age in contemporary times, irrespective of gender, which was not as common only a few decades ago. Through the thorough analysis of data provided by The Global Health Observatory (GHO) data repository, a branch under the World Health Organization (WHO), it is possible to determine which factors significantly impact human life expectancy. This paper considers data from 193 countries collected between 2000 and 2015 and considers 20 variables that are suspected to influence the life expectancy of a human being. Using multiple linear regression, the experiment aims to identify the characteristics that affect human life expectancy and shorten people's lifespans. Including numerous nations will make it easier to identify the elements contributing to the region's lower life expectancy. This will enable nations to determine what policies need to be implemented to raise their citizens' average life expectancy. The dataset's multiple missing values must be dealt with to keep our conclusions as accurate as possible. Most of the missing data pertains to population, Hepatitis B, and GDP. The missing data were from less-known countries like Vanuatu, Tonga, Togo, Cabo Verde, etc. Finding consistent data for these countries was difficult; hence, we decided to exclude these countries from the final model dataset. The raw dataset consists of 22 columns and 2938 rows, meaning 20 predicting variables. All predicting variables were then divided into several broad categories: immunization-related factors, mortality factors, economic factors, and social factors, all of which will be covered throughout the analysis. Detecting and removing null values allowed us to clean our data and get rid of outliers.

2.0 Introduction:

Life expectancy can be affected by many different factors in a person's life. In this paper, data from 193 countries are observed, recording 22 external variables that impact the life expectancy of humans. Factors such as alcohol consumption are expected to have an impact on a human being's life expectancy. By using a multiple linear regression model, we will be able to identify which variables have a significant correlation with and have an effect on life expectancy.

Linear regression is used to evaluate the relationship between two given variables by constructing a straight-line model based on a set of data. However, since we have more than one variable, we must use a multiple linear regression model, which is used to estimate the relationship between a quantitative dependent variable and many independent variables. Prior to doing the multiple linear regression model, a collection of data must be chosen, and a proper study must be conducted on the topic under evaluation in order to understand why such a variation could be happening. Using several variables will allow us to make better predictions and possibly allow for more accurate results. Some variables may have to be removed, and only those that help obtain the most accurate results for our model will be kept. To build a suitable model, we will use backward elimination in order to ensure that our variables correspond to our model. Such variables that will have to be removed are those that were multicollinear. Multicollinearity arises when the independent variables chosen have a relationship with one another. Variables may also have to be removed in order to avoid fake modeling, otherly known as overfitting, which occurs when many variables are put into play. When doing calculations, these prior procedures will allow for a more precise adjusted R^2 value.

2.1 Variable interpretation

A measurement variable is an unknown attribute that can take one or more values and measure a specific entity. In statistics, measurement variables, unlike in mathematics, can take both quantitative and qualitative values. Statistical variables can be measured using measurement tools, algorithms, or even human judgment. The scale of measurements refers to how we measure variables, and it influences the type of analytical techniques that can be used on the data and the conclusions that can be drawn from it. Nominal, ordinal, interval and ratio variables are the four types of measurement variables. It is important to note that, for the model to have functionality, every qualitative variable will be transformed into a quantitative variable using a programming software called Python. These variables will then be used to configure our model.

Nominal Variable

A nominal variable is a categorical variable that is used to name, label, or categorize specific attributes being measured but does not possess any numerical properties.

Ordinal Variable

An ordinal variable is a measurement variable that accepts values in a specific order or rank.

Interval Variable

The interval variable is a measuring variable that is used to specify values measured along a scale, with each point positioned at an equal distance from the others.

Ratio Variable

The ratio variable is one of two types of continuous variables, the other being the interval variable. It's the most common variable type being used in this model.

2.2 Defined Variables:

- Country (Nominal) - the country in which the indicators are from (i.e., Canada or Italy)
- Year (Ordinal) - the calendar year the indicators are from (ranging from 2000 to 2015)
- Status (Nominal) - whether a country is considered to be 'Developing' or 'Developed' by WHO standards.
- Life expectancy (Ratio) - the life expectancy of people in years for a particular country and year
- Adult mortality (Ratio) - the adult mortality rate per 1000 population (i.e., number of people dying between 15 and 60 years per 1000 population); if the rate is 263, then that means 263 people will die out of 1000 between the ages of 15 and 60; another way to think of this is that the chance an individual will die between 15 and 60 is 26.3%
- Infant deaths (Ratio) - number of infant deaths per 1000 population; similar to above, but for infants
- Alcohol (Ratio) - a country's alcohol consumption rate measured as liters of pure alcohol consumption per capita
- Percentage expenditure (Ratio) - expenditure on health as a percentage of Gross Domestic Product (GDP)
- Hepatitis B (Ratio) - number of 1-year-olds with Hepatitis B immunization overall 1-year-olds in the population
- Measles (Ratio) - number of reported Measles cases per 1000 population
- BMI (Interval/Ordinal) - average Body Mass Index (BMI) of a country's total population
- Under five deaths (Ratio) - the number of people under the age of five deaths per 1000 population
- Polio (Ratio) - the number of 1-year-olds with Polio immunization over the number of all 1-year-olds in the population
- Total expenditure (Ratio) - the government expenditure on health as a percentage of total government expenditure
- Diphtheria (Ratio) - Diphtheria tetanus toxoid and pertussis (DTP3) immunization rate of 1-year-olds
- Hiv/aids (Ratio) - deaths per 1000 live births caused by HIV/AIDS for people under 5; the number of people under five who die due to HIV/AIDS per 1000 births
- GDP (Ratio) - Gross Domestic Product per capita
- Population (Ratio) - the population of a country
- Thinness 1-19 years (Ratio) - the rate of thinness among people aged 1-19
- Thinness 5-9 years (Ratio) - the rate of thinness among people aged 5-9
- Income composition of resources (Ratio) - Human Development Index in terms of income composition of resources (index ranging from 0 to 1)
- Schooling (Ratio) - the average number of years of schooling of a population

3.0 Multiple Linear Regression equations

The following equation represents the multiple regression model's generic form:

(1) $$Y=\beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k + \varepsilon$$

Y will be the dependent variable, and all of the distinct x's will be the independent variables. Since there are k regressors in this multiple linear regression, we assume the error term ε is centered at zero and relatively small. Unknown parameters $\beta_0, \beta_1, \beta_2 ..., \beta_k$ are linear functions of the equation (1) given above.

The intercept of the hyperplane is represented by β_0 on a surface produced by a multiple linear regression model or equation, and the partial regression coefficients are the regressors' coefficients. If x_2,..., and x_k are all maintained constant, β_1 quantifies the anticipated change in Y per unit change in x_1.

3.1 Hypothesis Testing:

The purpose of this test is to evaluate the strength of the correlation between each coefficient and the dependent variable. This will allow us to filter out which predictors are useful and which aren't to our model. The closer this t-value test is to zero, the less significant this variable is as a predictor of life expectancy. Since there is no correlation between them when the value is 0, this particular variable may be eliminated.

The two following hypotheses below are made. If we are unable to reject the null hypothesis, it will follow that we have evidence that the predictor is not significantly correlated with life expectancy. Therefore, it may be eliminated from the model. If the null hypothesis is rejected, then we would accept the alternative hypothesis that claims the predictor is significant.

$$H_0 : \beta_k = 0 \qquad\qquad H_1 : \beta_k \neq 0$$

Where:

H_0 =Null Hypothesis

H_1=Alternate Hypothesis

In order for the null hypothesis to be rejected, a two-tailed test must be conducted, which will ensure that our result falls within the rejection region. The following equation below is used to calculate :

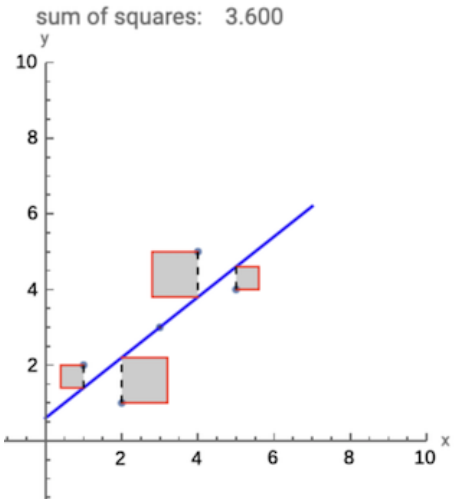
(2) $$-\ t_{\frac{\alpha}{2}, n-2} < T < \ t_{\frac{\alpha}{2}, n-2}$$

A confidence interval must be established in order to obtain the t critical values from both tails. Once these values are calculated, we can then make conclusions about whether we can reject the null hypothesis.

3.2 Least Squares Estimation of Parameters :

The following figure shows a visual of the explanation below to better visualize:

Figure 1. *Least Squares Criteria for the Least Squares Regression Line (*Maughan, Mariel, and Bruce Torrence., 2011)



This figure shows a random set of data that uses the least square method based on a single variable. The graph above demonstrates a linear best-fit line with the sum of squared errors. In all, minimizing the area found of each square from the line by finding the best fit line will allow us to minimize the sum of squared errors. Although, in this case, a linear example is shown, the least squares optimization works out very well and may be used for a variety of regression models, including linear, quadratic, exponential, logarithmic, sinusoidal, etc.

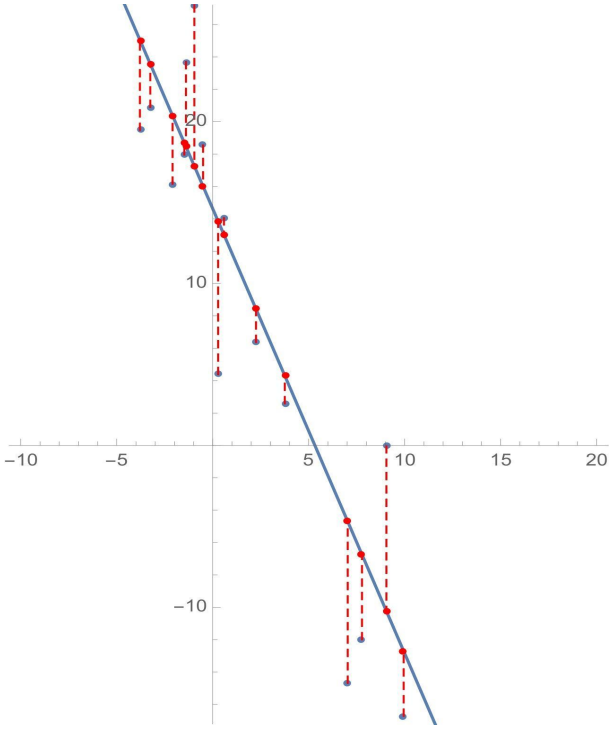


Figure 2. *Least-squares fitting using vertical offsets*

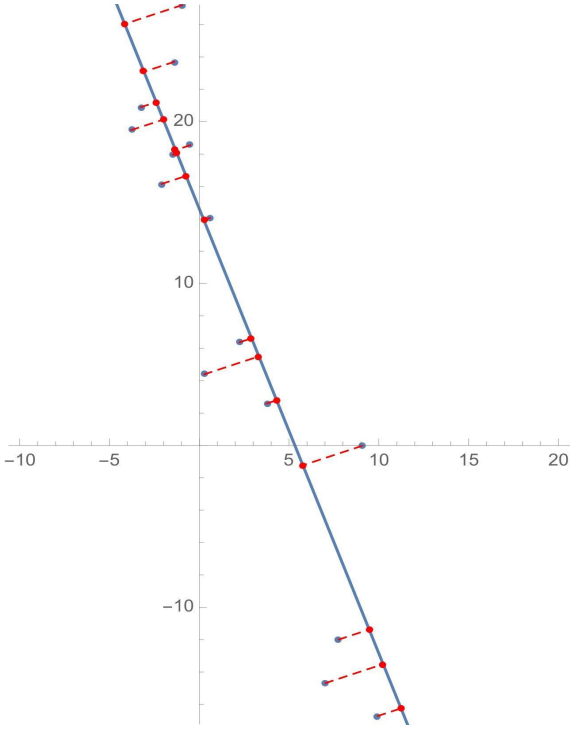


Figure 3. *Least-squares fitting using perpendicular offsets*

The figures above indicate two forms of error margins, better known as "offsets."Instead of being measured perpendicularly from the best fit line, the error margins are measured vertically. Vertical offsets are far more useful in practice than perpendicular offsets, even though there isn't much of a difference in how well they fit compared to each other. This is because they make it easier to account for the uncertainties of data points along both the x and y axes and because it's easier to analyze the parameters involved in fitting vertical offsets than perpendicular offsets.

3.3 Least Square Function for Multiple Linear Regression:

The regression coefficients in the multiple regression equation are estimated using the least squares method. Let x_{ij} display the i^{th} observation of variable x_j if n>k observations are available.

The observed data points are shown below.

$$(x_{i1}, x_{i2}, \ldots, x_{ik}, y_i), \qquad i = 1,2\ldots, n \text{ and } n > k$$

The data is often shown similarly, as in the table below from a multiple linear regression set of data:

y	x_1	x_2	...	x_k
y_1	x_{11}	x_{12}	...	x_{1k}
y_2	x_{21}	x_{22}	...	x_{2k}
⋮	⋮	⋮		⋮
y_n	x_{n1}	x_{n2}	...	x_{nk}

Figure 4.

The data fits the model as follows:

(3)
$$\begin{aligned} y_i &= \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \epsilon_i \\ &= \beta_0 + \sum_{j=1}^{k} \beta_j x_{ij} + \epsilon_i \quad i = 1, 2, \ldots, n \end{aligned}$$

The least squares function shown in the following equation captures the discrepancy between the predictors and the data.

(4)
$$L = \sum_{i=1}^{n} \epsilon_i^2 = \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{k} \beta_j x_{ij} \right)^2$$

The least squares estimates of these parameters must satisfy the following two equations (5&6) to minimize L with respect to $\beta_{0,}\ \beta_{1,}\ldots.\beta k$

(5)
$$\left.\frac{\partial L}{\partial \beta_j}\right|_{\hat{\beta}_0,\hat{\beta}_1,\ldots,\hat{\beta}_k} = -2\sum_{i=1}^{n}\left(y_i - \hat{\beta}_0 - \sum_{j=1}^{k}\hat{\beta}_j x_{ij}\right) = 0$$

(6)
$$\left.\frac{\partial L}{\partial \beta_j}\right|_{\hat{\beta}_0,\hat{\beta}_1,\ldots,\hat{\beta}_k} = -2\sum_{i=1}^{n}\left(y_i - \hat{\beta}_0 - \sum_{j=1}^{k}\hat{\beta}_j x_{ij}\right) x_{ij} = 0 \qquad j = 1, 2, \ldots, k$$

The scalar least squares equations can then be found after simplifying equations 5 and 6:

(7)
$$\begin{array}{lllll}
n\hat{\beta}_0 & +\hat{\beta}_1\sum_{i=1}^{n}x_{i1} & +\hat{\beta}_2\sum_{i=1}^{n}x_{i2} & +\cdots \quad +\hat{\beta}_k\sum_{i=1}^{n}x_{ik} & = \sum_{i=1}^{n}y_i \\
\hat{\beta}_0\sum_{i=1}^{n}x_{i1} & +\hat{\beta}_1\sum_{i=1}^{n}x_{i1}^2 & +\hat{\beta}_2\sum_{i=1}^{n}x_{i1}x_{i2} & +\cdots \quad +\hat{\beta}_k\sum_{i=1}^{n}x_{i1}x_{ik} & = \sum_{i=1}^{n}x_{i1}y_i \\
\vdots & \vdots & \vdots & \vdots & \vdots \\
\hat{\beta}_0\sum_{i=1}^{n}x_{ik} & +\hat{\beta}_1\sum_{i=1}^{n}x_{ik}x_{i1} & +\hat{\beta}_2\sum_{i=1}^{n}x_{ik}x_{i2} & +\cdots \quad +\hat{\beta}_k\sum_{i=1}^{n}x_{ik}^2 & = \sum_{i=1}^{n}x_{ik}y_i
\end{array}$$

The equations above are the least squares estimators of the coefficients of regression and can be easily obtained by using techniques involving linear algebra.

3.4 Matrix notation

Solving for the regression coefficients can be computed using matrix notation. We can rewrite equation (1) given the observed data points expressed through the following form. $(x_{11}, x_{12}, \ldots, x_{1k}, y_1)$, as shown below:

(8)

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon}$$

$$\boldsymbol{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \quad \boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

$$\boldsymbol{X} = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1k} \\ \vdots & \vdots & \ldots & \vdots \\ 1 & x_{n_1} & \cdots & x_{nk} \end{bmatrix} \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix}$$

This formula can then be simplified by using a matrix notation into expressions for **ˆβ**, the vector of the least square estimates concerning $\beta_0, \beta_1, \ldots, \beta_k$ as shown:

(9)

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X})^{-\mathbf{1}}\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{y}$$

The prediction equation can then be expressed as:

(10)

$$\widehat{\boldsymbol{y}} = \boldsymbol{X}\widehat{\boldsymbol{\beta}}$$

The residuals, portrayed as **ε**, are known as the difference between an observed data value and a predicted data value and can be expressed as a vector under the form:

(11)

$$\boldsymbol{e} = \boldsymbol{y} - \widehat{\boldsymbol{y}}$$

The underlying assumptions made towards residuals denoted as **ε** state that its values are independent, normally distributed, and have a mean centered around zero with a common variance (σ^2). If these conditions are met, then **ˆβ** is an unbiased estimator for **β.**

3.5 Estimating σ^2

In a regression model, whether linear or multiple linear, σ^2 represents the variance of the random error. The formula in order to calculate σ^2 is demonstrated below:

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n} e_i^2}{n-p} = \frac{SSE}{n-p} = \frac{SSE}{n-k-1} \quad (12)$$

SSE is the residual sum (Sum of Squares Due to Error). Due to many parameters, since it is a multiple linear regression model, the number of parameters is denoted as p, unlike a linear regression model with only two parameters.

3.6 R^2 and adjusted R^2

Unlike in a simple regression model, it is critical to calculate the adjusted R^2 when there are multiple variables in a regression model. In a Linear regression model, the R^2 value is used to determine the variation between the y values and the mean, which can easily be explained by the independent variable. Given numerous predictor values in multiple linear regression, the SSE is reduced, and the R^2 value is artificially raised. Although the R^2 value can rise as more predictors are put into play, at certain times, the rise in the R^2 value can result in fake modeling or overfitting, reducing the effectiveness of our model. The R^2 adjusted is often used to see whether the rise in R^2 is legitimate. It allows for the right predictors to be used and for those that are not significant to be removed. The R^2 adjusted value is found using the equation below:

(13) $$\text{Adjusted } R^2 = 1 - \frac{(1-R^2)\,(N-1)}{N-p-1}$$

Where:

R^2 = the sample R^2

N = the total sample size

P = number of independent variables

Therefore, the higher the adjusted R^2 value, the more accurately the predictors are helping predict the dependent variable. R^2 adjusted will only increase if a new variable is added, and it reduces the mean square error.

4.0 Data and Analysis

Using Google Colab and the Jupyter Notebook that was given in class, a model will be built from a dataset containing data from 193 countries that were collected from 2000-2015, including 22 variables that are suspected of leading to a decreased life expectancy. This data will help us build a model that will be used to determine which factors have a correlation with life expectancy and which don't. Accumulating as much data as possible allows us to choose which predictors benefit our model the best. Having data from many countries will also ensure that many of them have to have a correlation with one of the predictors in order for it to have a valid correlation with life expectancy.

An example of the data from one country (Afghanistan), including 7 of the 22 variables, is shown below :

Figure 5. Sample data from 2000-2015 in Afghanistan

	Country	Year	Status	Life expectancy	Adult Mortality	infant deaths	Alcohol	percentage expenditure
2	Afghanistan	2015	Developing	65	263	62	0.01	71.27962362
3	Afghanistan	2014	Developing	59.9	271	64	0.01	73.52358168
4	Afghanistan	2013	Developing	59.9	268	66	0.01	73.21924272
5	Afghanistan	2012	Developing	59.5	272	69	0.01	78.1842153
6	Afghanistan	2011	Developing	59.2	275	71	0.01	7.097108703
7	Afghanistan	2010	Developing	58.8	279	74	0.01	79.67936736
8	Afghanistan	2009	Developing	58.6	281	77	0.01	56.76221682
9	Afghanistan	2008	Developing	58.1	287	80	0.03	25.87392536
10	Afghanistan	2007	Developing	57.5	295	82	0.02	10.91015598
11	Afghanistan	2006	Developing	57.3	295	84	0.03	17.17151751
12	Afghanistan	2005	Developing	57.3	291	85	0.02	1.388647732
13	Afghanistan	2004	Developing	57	293	87	0.02	15.29606643
14	Afghanistan	2003	Developing	56.7	295	87	0.01	11.08905273
15	Afghanistan	2002	Developing	56.2	3	88	0.01	16.88735091
16	Afghanistan	2001	Developing	55.3	316	88	0.01	10.5747282
17	Afghanistan	2000	Developing	54.8	321	88	0.01	10.42496

4.1 Backwards Elimination:

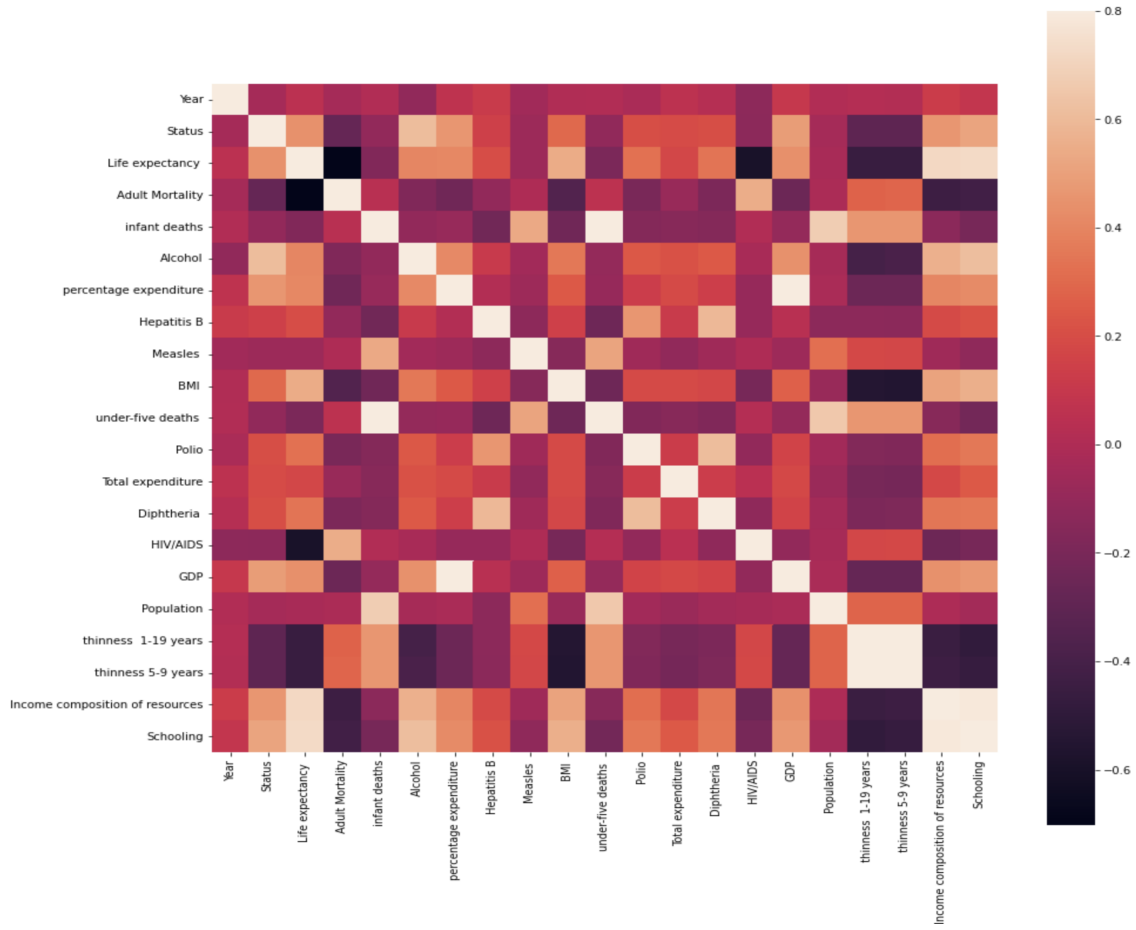
Although there were many ways to clean our data, the most efficient for our model was backwards elimination. This process involved manually reviewing data and removing data that led to a decrease in the R^2 and R^2 adjusted value in the hypothesis testing. By examining the data, those with the highest p-values were removed. This process was continued until all the p-values under 0.05 were removed, and we were satisfied with our R^2 and adjusted R^2 values. It allowed us to remove predictors that didn't have a strong correlation or data that was multicollinear, thus reducing the strength of our model.

4.2 Linear Independence and Multicollinearity:

When two or more independent variables in a regression model have a strong correlation with one another, multicollinearity arises. This implies that, in a regression model, one independent variable may be predicted from another independent variable. It was clear that multicollinearity in the model affected coefficients and the p-value, which made our model weaker and less effective in predicting life expectancy. Thus, linear dependency is important, and in our model, we aimed to make sure that the correlations were linear-dependent and not multicollinear. Linear independence ensures that the matrix has an inverse and that its determinant is not equal to zero. This allows for an individual variable to have its own correlation rather than being correlated to another predictor variable.

4.3 Heat Map

Figure 6. Correlations between variables



Heatmaps are used to show correlations between two variables. Looking at each cell's color, can help distinguish to what extent two variables are correlated. A scale is shown on the right of the heat map above in order to indicate the strength of the relationship by color; while light cells have the strongest correlation, the darkest ones have the weakest. Observing heat maps can help indicate patterns and help visualize correlations. For example, 'GDP' and 'percentage expenditure' are highly correlated; meanwhile, 'Adult Mortality' and 'Life expectancy' have a low correlation.

4.4 Scatter plot

Figure 7. Scatter plot of life expectancy in correlation to each predictor

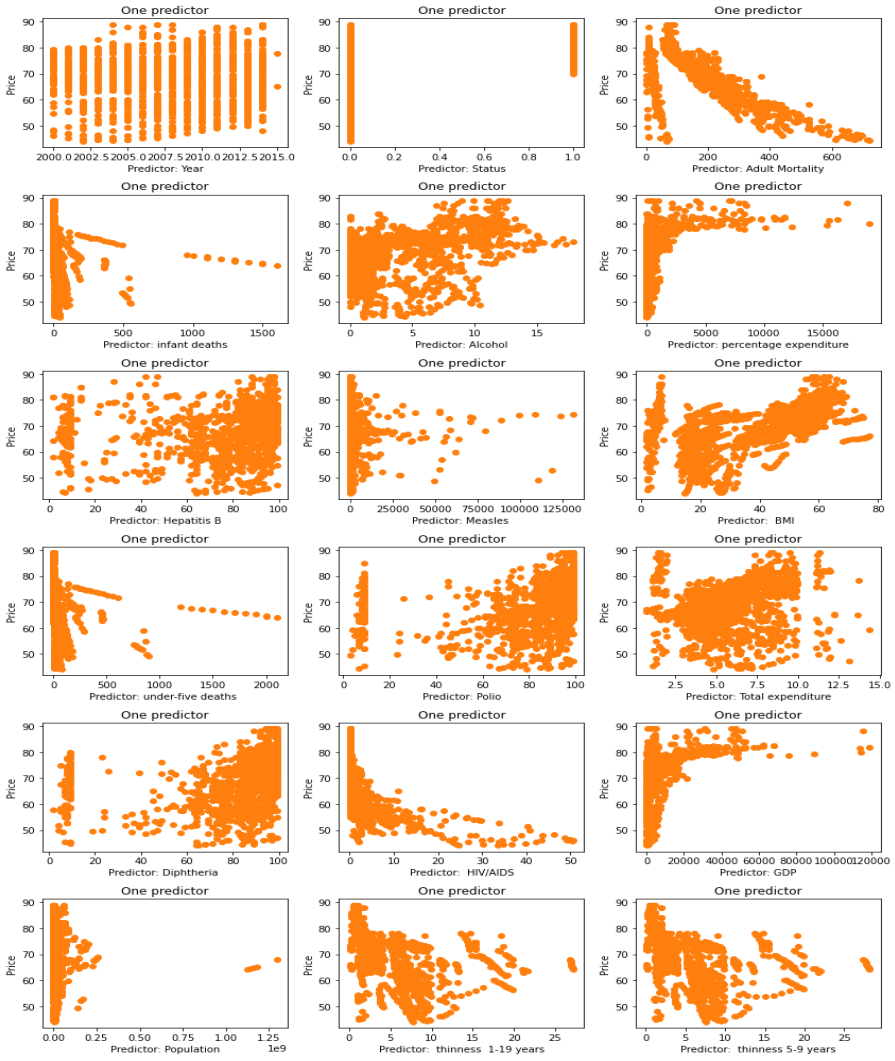


Figure 5 shows the relationship between each predictor and life expectancy. This allows for a quick view in order to visualize the relationship between each of the two variables. The exponential relationships indicate the correlation of the variables, which allows for outliers to be easily seen. With these graphs, we can also observe the linear relationships between each variable. A small number of graphs have straight lines due to the discrete values of that specific predictor, for example, which categorizes a country as either developed or developing with

nothing in the middle. As this is a nominal variable, it is transformed into a quantitative variable in order to fit into our model. Such that there is nothing between developing and developed, they are given a numerical value of 0 and 1, respectively.

4.5 Building the Model

Figure 8. The initial model comprised all 20 predictors

OLS Regression Results

Dep. Variable:	Life expectancy	**R-squared:**	0.839
Model:	OLS	**Adj. R-squared:**	0.837
Method:	Least Squares	**F-statistic:**	422.9
Date:	Tue, 06 Dec 2022	**Prob (F-statistic):**	0.00
Time:	02:57:36	**Log-Likelihood:**	-4421.2
No. Observations:	1649	**AIC:**	8884.
Df Residuals:	1628	**BIC:**	8998.
Df Model:	20		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Year	-0.1272	0.023	-5.510	0.000	-0.172	-0.082
Status	0.8865	0.335	2.644	0.008	0.229	1.544
Adult Mortality	-0.0162	0.001	-17.171	0.000	-0.018	-0.014
infant deaths	0.0887	0.011	8.376	0.000	0.068	0.110
Alcohol	-0.1313	0.034	-3.901	0.000	-0.197	-0.065
percentage expenditure	0.0003	0.000	1.691	0.091	-4.83e-05	0.001
Hepatitis B	-0.0033	0.004	-0.732	0.464	-0.012	0.005
Measles	-1.033e-05	1.07e-05	-0.966	0.334	-3.13e-05	1.07e-05
BMI	0.0318	0.006	5.345	0.000	0.020	0.044
under-five deaths	-0.0666	0.008	-8.682	0.000	-0.082	-0.052
Polio	0.0058	0.005	1.132	0.258	-0.004	0.016
Total expenditure	0.0922	0.040	2.281	0.023	0.013	0.171
Diphtheria	0.0140	0.006	2.387	0.017	0.002	0.026
HIV/AIDS	-0.4481	0.018	-25.174	0.000	-0.483	-0.413
GDP	2.451e-05	2.83e-05	0.867	0.386	-3.09e-05	7.99e-05
Population	-6.085e-10	1.73e-09	-0.351	0.726	-4.01e-09	2.79e-09
thinness 1-19 years	-0.0058	0.053	-0.111	0.912	-0.109	0.097
thinness 5-9 years	-0.0501	0.052	-0.966	0.334	-0.152	0.052
Income composition of resources	10.4497	0.833	12.549	0.000	8.816	12.083
Schooling	0.8949	0.059	15.142	0.000	0.779	1.011
const	308.1207	46.223	6.666	0.000	217.457	398.784

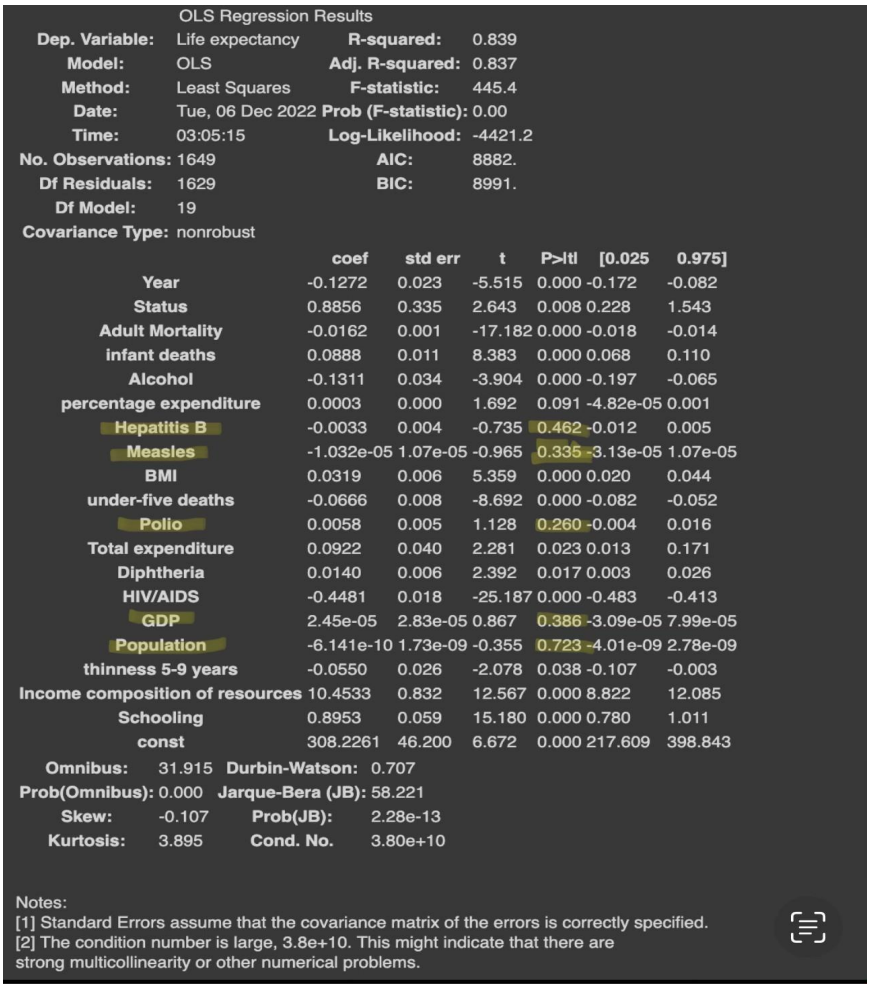
Omnibus:	31.845	**Durbin-Watson:**	0.707
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	58.052
Skew:	-0.107	**Prob(JB):**	2.48e-13
Kurtosis:	3.894	**Cond. No.**	3.80e+10

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 3.8e+10. This might indicate that there are strong multicollinearity or other numerical problems.

As the title suggests, this is the first defined life expectancy model, which has 20 predictors and an adjusted R^2 value of 0.837. At first glance, this seems extremely favorable; however, the extremely high conditional number of 3.80×10^{10} suggests that there is strong

multicollinearity. This is evidently bad for our model and must be adjusted. In order to begin adjusting this value, we must drop the least predictive variable - 'thinness 1-19 years.

Figure 9. Second model with 19 predictors



OLS Regression Results

Dep. Variable:	Life expectancy	**R-squared:**	0.839
Model:	OLS	**Adj. R-squared:**	0.837
Method:	Least Squares	**F-statistic:**	445.4
Date:	Tue, 06 Dec 2022	**Prob (F-statistic):**	0.00
Time:	03:05:15	**Log-Likelihood:**	-4421.2
No. Observations:	1649	**AIC:**	8882.
Df Residuals:	1629	**BIC:**	8991.
Df Model:	19		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Year	-0.1272	0.023	-5.515	0.000	-0.172	-0.082
Status	0.8856	0.335	2.643	0.008	0.228	1.543
Adult Mortality	-0.0162	0.001	-17.182	0.000	-0.018	-0.014
infant deaths	0.0888	0.011	8.383	0.000	0.068	0.110
Alcohol	-0.1311	0.034	-3.904	0.000	-0.197	-0.065
percentage expenditure	0.0003	0.000	1.692	0.091	-4.82e-05	0.001
Hepatitis B	-0.0033	0.004	-0.735	0.462	-0.012	0.005
Measles	-1.032e-05	1.07e-05	-0.965	0.335	-3.13e-05	1.07e-05
BMI	0.0319	0.006	5.359	0.000	0.020	0.044
under-five deaths	-0.0666	0.008	-8.692	0.000	-0.082	-0.052
Polio	0.0058	0.005	1.128	0.260	-0.004	0.016
Total expenditure	0.0922	0.040	2.281	0.023	0.013	0.171
Diphtheria	0.0140	0.006	2.392	0.017	0.003	0.026
HIV/AIDS	-0.4481	0.018	-25.187	0.000	-0.483	-0.413
GDP	2.45e-05	2.83e-05	0.867	0.386	-3.09e-05	7.99e-05
Population	-6.141e-10	1.73e-09	-0.355	0.723	-4.01e-09	2.78e-09
thinness 5-9 years	-0.0550	0.026	-2.078	0.038	-0.107	-0.003
Income composition of resources	10.4533	0.832	12.567	0.000	8.822	12.085
Schooling	0.8953	0.059	15.180	0.000	0.780	1.011
const	308.2261	46.200	6.672	0.000	217.609	398.843

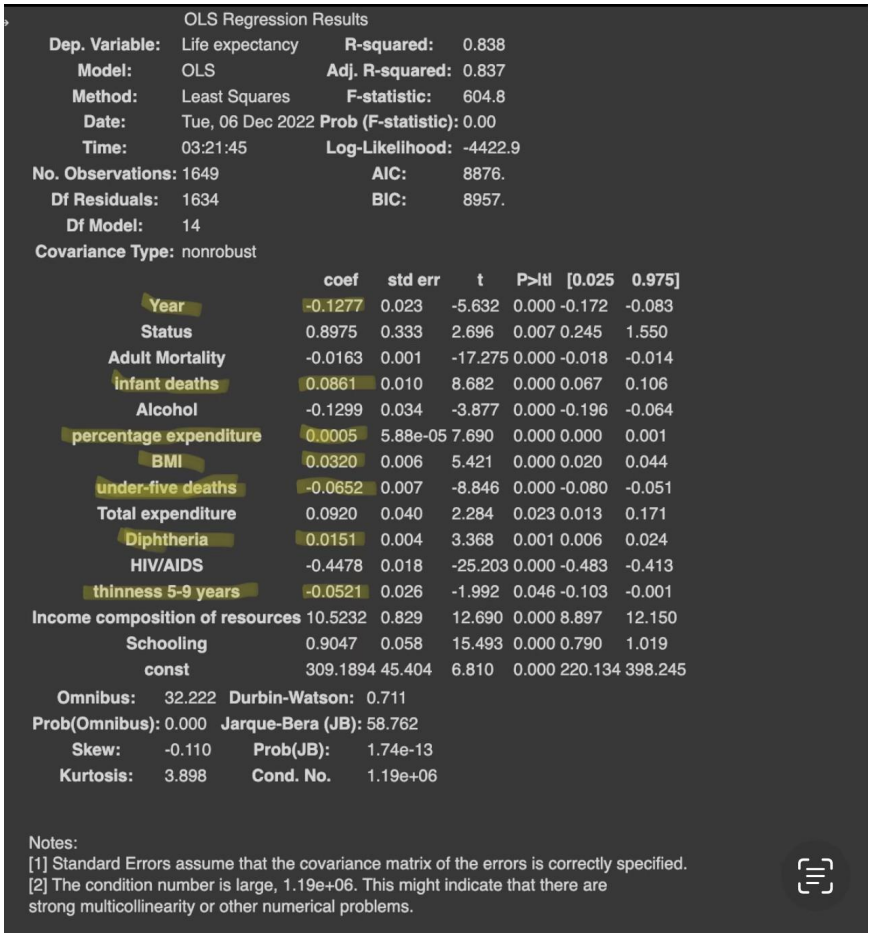
Omnibus:	31.915	**Durbin-Watson:**	0.707
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	58.221
Skew:	-0.107	**Prob(JB):**	2.28e-13
Kurtosis:	3.895	**Cond. No.**	3.80e+10

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 3.8e+10. This might indicate that there are strong multicollinearity or other numerical problems.

After dropping the least predictive variable, it is noted that there is no significant change in the model. Moreover, there is no change in the R^2 adjusted value nor the conditional number. After thorough analysis, we noted that the extremely high p-values should be dropped in order to

attain a more stable model. This stability is achieved through dropping the 5 highlighted predictors : Hepatitis B, Measles, Polio, GDP, and Population.

Figure 10. Third model, with 14 predictors



OLS Regression Results

Dep. Variable:	Life expectancy	R-squared:	0.838
Model:	OLS	Adj. R-squared:	0.837
Method:	Least Squares	F-statistic:	604.8
Date:	Tue, 06 Dec 2022	Prob (F-statistic):	0.00
Time:	03:21:45	Log-Likelihood:	-4422.9
No. Observations:	1649	AIC:	8876.
Df Residuals:	1634	BIC:	8957.
Df Model:	14		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Year	-0.1277	0.023	-5.632	0.000	-0.172	-0.083
Status	0.8975	0.333	2.696	0.007	0.245	1.550
Adult Mortality	-0.0163	0.001	-17.275	0.000	-0.018	-0.014
infant deaths	0.0861	0.010	8.682	0.000	0.067	0.106
Alcohol	-0.1299	0.034	-3.877	0.000	-0.196	-0.064
percentage expenditure	0.0005	5.88e-05	7.690	0.000	0.000	0.001
BMI	0.0320	0.006	5.421	0.000	0.020	0.044
under-five deaths	-0.0652	0.007	-8.846	0.000	-0.080	-0.051
Total expenditure	0.0920	0.040	2.284	0.023	0.013	0.171
Diphtheria	0.0151	0.004	3.368	0.001	0.006	0.024
HIV/AIDS	-0.4478	0.018	-25.203	0.000	-0.483	-0.413
thinness 5-9 years	-0.0521	0.026	-1.992	0.046	-0.103	-0.001
Income composition of resources	10.5232	0.829	12.690	0.000	8.897	12.150
Schooling	0.9047	0.058	15.493	0.000	0.790	1.019
const	309.1894	45.404	6.810	0.000	220.134	398.245

Omnibus:	32.222	Durbin-Watson:	0.711
Prob(Omnibus):	0.000	Jarque-Bera (JB):	58.762
Skew:	-0.110	Prob(JB):	1.74e-13
Kurtosis:	3.898	Cond. No.	1.19e+06

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 1.19e+06. This might indicate that there are strong multicollinearity or other numerical problems.

After analyzing our more stable model, it is noted that the R^2 adjusted value remained constant throughout the dropping of the predictors. However, the conditional number dropped by a significant exponent of four. This is looking a lot better at this point, as the model has less multicollinearity issues while keeping the same R^2 adjusted value. Also, the p-values have become much more reasonable now. The next step in stabilizing this model will come from

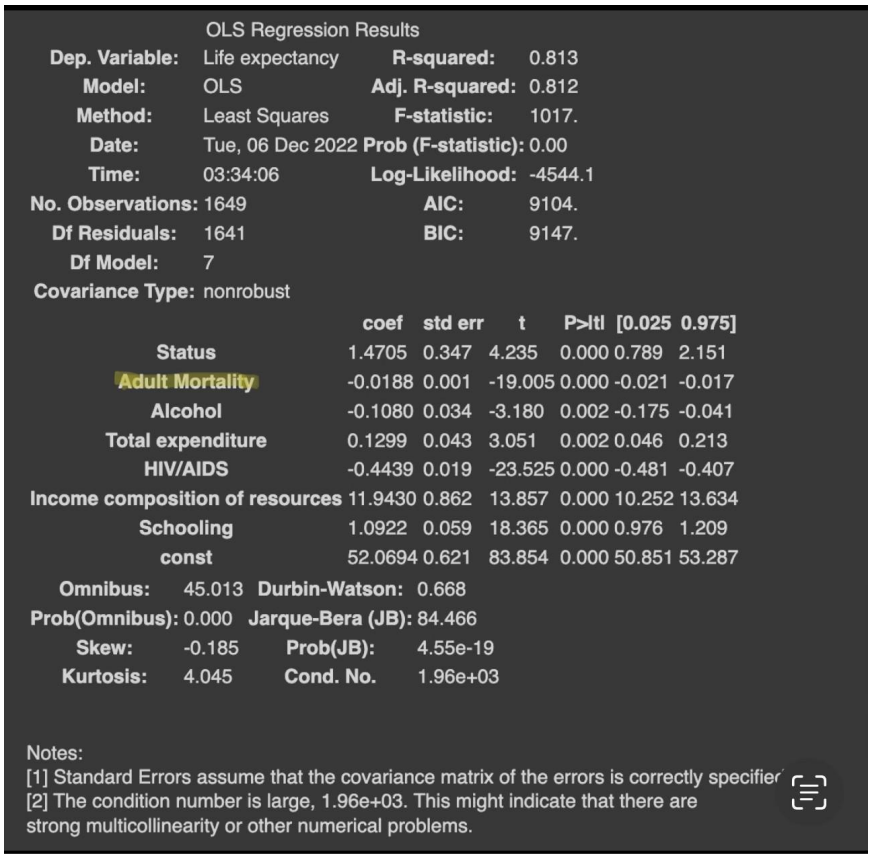
eliminating predictors with unreasonable coefficients. These predictors include year, infant deaths, percentage expenditure, body mass index (BMI), under-five deaths, diphtheria, and thinness between 5 and 9 years old.

The following analysis of the dropped predictors' coefficients will explain the reasoning behind each elimination. A positive coefficient signifies that the predictor, being the independent variable, has a positive correlation with life expectancy, the dependent variable. The opposite is also true regarding negative coefficients. A negative coefficient demonstrates that the predictor has a negative correlation with life expectancy.

Firstly, it is shown that the year is negatively correlated with life expectancy, which is immediately suspected to be wrong since it's been proven that humans live longer as the years pass. For this reason, the predictor will be dropped. Infant deaths are shown to have a positive correlation with life expectancy, which is another odd occurrence since a large rate of infant deaths would likely cause life expectancy to be shorter. However, there are much fewer infant deaths in modern times as compared to a century ago and do not prove to be a good fit for this model. This predictor poses a significant multicollinearity issue with other predictors and must therefore be dropped. Percentage expenditure is another predictor which causes a lot of instability in the model due to the multicollinearity it imposes on it. BMI has a positive correlation with life expectancy, which shouldn't be the case. BMI does not distinguish between excess fat, muscle, or bone mass, nor does it provide any indication of the distribution of fat among individuals. It is inaccurately representative of muscle mass and serves to the relationship between height and weight. With that said, BMI should be negatively correlated. Under-five deaths are also shown to have a positive correlation with life expectancy. It should be negative

since infant mortality is subject to decreased life expectancy. Diphtheria is shown to be slightly multicollinear, and this slight correction will help create a better model. Lastly, thinness between 5 and 9 year olds is subject to be removed since most kids between 5 and 9 are thin, thus making it an insignificant predictor for our model.

Figure 11. Fourth model with 7 predictors



OLS Regression Results

Dep. Variable:	Life expectancy	**R-squared:**	0.813
Model:	OLS	**Adj. R-squared:**	0.812
Method:	Least Squares	**F-statistic:**	1017.
Date:	Tue, 06 Dec 2022	**Prob (F-statistic):**	0.00
Time:	03:34:06	**Log-Likelihood:**	-4544.1
No. Observations:	1649	**AIC:**	9104.
Df Residuals:	1641	**BIC:**	9147.
Df Model:	7		
Covariance Type:	nonrobust		

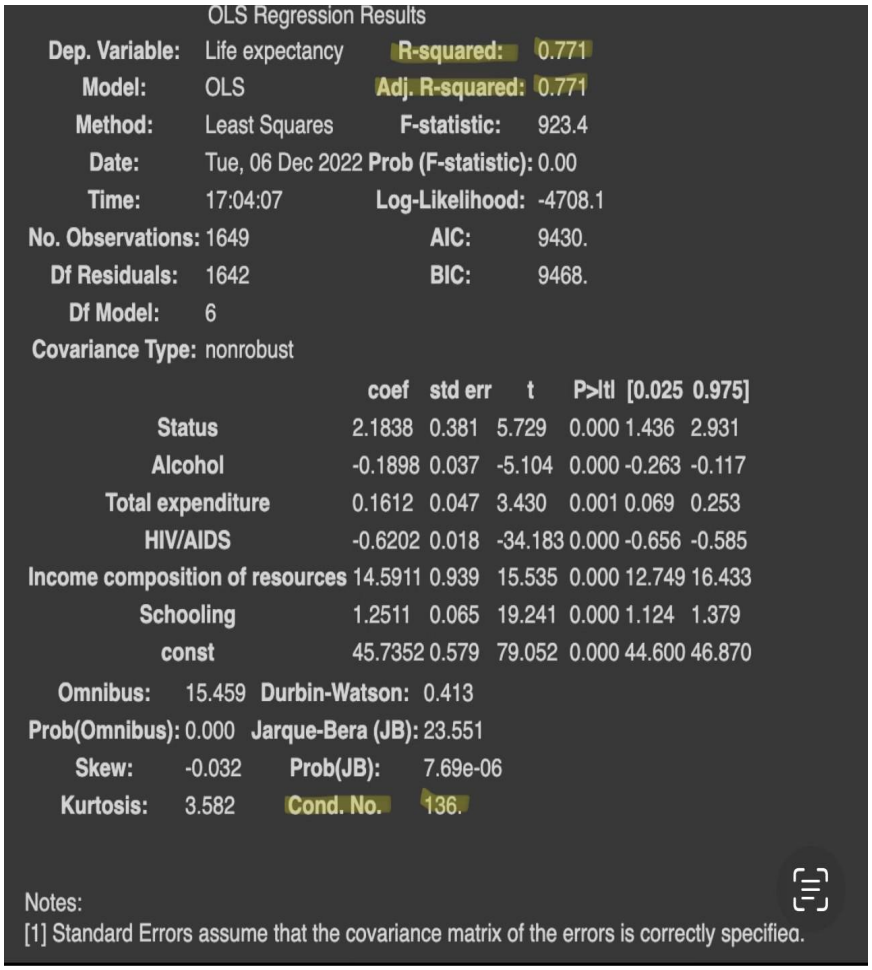
	coef	std err	t	P>|t|	[0.025	0.975]
Status	1.4705	0.347	4.235	0.000	0.789	2.151
Adult Mortality	-0.0188	0.001	-19.005	0.000	-0.021	-0.017
Alcohol	-0.1080	0.034	-3.180	0.002	-0.175	-0.041
Total expenditure	0.1299	0.043	3.051	0.002	0.046	0.213
HIV/AIDS	-0.4439	0.019	-23.525	0.000	-0.481	-0.407
Income composition of resources	11.9430	0.862	13.857	0.000	10.252	13.634
Schooling	1.0922	0.059	18.365	0.000	0.976	1.209
const	52.0694	0.621	83.854	0.000	50.851	53.287

Omnibus:	45.013	**Durbin-Watson:**	0.668
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	84.466
Skew:	-0.185	**Prob(JB):**	4.55e-19
Kurtosis:	4.045	**Cond. No.**	1.96e+03

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specifie
[2] The condition number is large, 1.96e+03. This might indicate that there are strong multicollinearity or other numerical problems.

After negating predictors with illogical coefficients, the model is observed to look notably better but still indicates strong multicollinearity (with a conditional number of 1.96×10^3). The one predictor that's proving to be a big problem for the linear dependency of this model is adult mortality which will therefore be dropped.

Figure 12. Final model with 6 predictors



OLS Regression Results

Dep. Variable:	Life expectancy	R-squared:	0.771
Model:	OLS	Adj. R-squared:	0.771
Method:	Least Squares	F-statistic:	923.4
Date:	Tue, 06 Dec 2022	Prob (F-statistic):	0.00
Time:	17:04:07	Log-Likelihood:	-4708.1
No. Observations:	1649	AIC:	9430.
Df Residuals:	1642	BIC:	9468.
Df Model:	6		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Status	2.1838	0.381	5.729	0.000	1.436	2.931
Alcohol	-0.1898	0.037	-5.104	0.000	-0.263	-0.117
Total expenditure	0.1612	0.047	3.430	0.001	0.069	0.253
HIV/AIDS	-0.6202	0.018	-34.183	0.000	-0.656	-0.585
Income composition of resources	14.5911	0.939	15.535	0.000	12.749	16.433
Schooling	1.2511	0.065	19.241	0.000	1.124	1.379
const	45.7352	0.579	79.052	0.000	44.600	46.870

Omnibus:	15.459	Durbin-Watson:	0.413
Prob(Omnibus):	0.000	Jarque-Bera (JB):	23.551
Skew:	-0.032	Prob(JB):	7.69e-06
Kurtosis:	3.582	Cond. No.	136.

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

In the final model, the conditional number is seen to have improved drastically. While sacrificing a slight measure of fit, or more specifically, the R^2 value, the model now denotes more accurate linear dependency. With both an R^2 and adjusted R^2 value of 0.771, and a conditional number of 1.36×10^2, we were prompted to stop adjusting the model. Further modifications would only hinder the accuracy of the model and not measure what we were hoping for, so we will stop here.

5.0 Conclusion

In summary, our fifth and final model highlights the best values attainable while maintaining the integrity of our predictors, along with the data they contain. The R^2 and adjusted R^2 value of 0.771 coupled with a reasonable condition number satisfies the model. Accurate and rational coefficients are also present in making this model superior to its predecessors. By analyzing the initial model, it was made abundantly clear to us that there was a big issue at hand due to it being multicollinear. This allowed us to identify insignificant and weak predictors, which were then removed to create a more linear dependent model.

The final equation for this model is as shown below:

Life expectancy (years) = (45.735238) + (2.183772 × Status) - (0.189813 × Alcohol consumption) + (0.161154 × Total expenditure) - (0.620225 × HIV/AIDS) + (14.591108 × Income composition of resources) + (1.251054 × Schooling)

The coefficients in this model are all reasonable given the role they play in a person's life. Alcohol consumption and HIV/AIDS are known to be some deterrents for a long life. Status, total expenditure, income composition of resources, and schooling are all sustaining factors in the longevity of a human being's life.

Given data from the WHO dataset, we can estimate the average life expectancy of a region using our model. We decided to use data from Canada in 2013 in order to demonstrate the accuracy of our model.

Given that Canada is a developed country, its status (ranging numerically from 0 to 1) is given a quantitative value of 1. The alcohol consumption recorded per capita has a numerical value of 8.2. The general government expenditure on health, as a percentage of total government expenditure, is 11.67. Deaths per 1000 live births related to HIV/AIDS is 0.1. The Human Development Index, in terms of income composition of resources (with an index ranging from 0 to 1) is 0.907. Lastly, the number of years of schooling is an impressive 15.9 years. Given these values, we can compute the expected life expectancy of a region with these data points.

Note: Bold values portray data points.

Life expectancy (years) = (45.735238) + (2.183772 × **1**) - (0.189813 × **8.2**) + (0.161154 × **11.67**) - (0.620225 × **0.1**) + (14.591108 × **0.907**) + (1.251054 × **15.9**) = 81.3 years

Our model has computed that the average life expectancy in Canada, given the data from 2013, is 81.3 years. The dataset provided by WHO states that the life expectancy is 81.8 years in Canada in 2013. This means that our model was within 0.5 years of the World Health Organization's life expectancy for Canada.

This model proves to be effective in estimating life expectancy given the 6 predictors it has been configured to compute with. This multiple linear regression model is an effective tool for making predictions similar to this data (such as for other countries). It is important to retain that quality beats quantity and that more variables are not necessarily better for a linear regression model. This has been proven through our analysis of the data and through the removal of insignificant or highly multicollinear predictors. We started out with 20 variables and ended with 6; with the removal of 14 predictors came a fairly accurate model.

In data science, there are always ways to improve a model. For example, we can use polynomial terms to model the nonlinear relationship between an independent variable and a target variable in the future. Moreover, interaction terms can also be implemented to simulate how two or more independent variables influence the target variable.

6.0 Bibliography

(1) No author. “How to Derive the Least Square Estimator for Multiple Linear Regression?” *Cross Validated*, February 1 1960.

https://stats.stackexchange.com/questions/46151/how-to-derive-the-least-square-estimator-for-multiple-linear-regression

(2) Maughan, Mariel, and Bruce Torrence. “Wolfram Demonstrations Project.” *Least Squares Criteria for the Least Squares Regression Line*, March 12 2011.

https://demonstrations.wolfram.com/LeastSquaresCriteriaForTheLeastSquaresRegressionLine/

(3) Ivanov, I.T. (2021). §27: Linear Regression. *Probability and Statistics*, 201-HTH-05, Lectures.

(4) Frost, Jim. “How to Interpret Adjusted R-Squared and Predicted R-Squared in Regression Analysis.” *Statistics By Jim*, February 27 2022,

https://statisticsbyjim.com/regression/interpret-adjusted-r-squared-predicted-r-squared-regression

(5) Bhandari, Aniruddha. “Multicollinearity: Detecting Multicollinearity with VIF.” *Analytics Vidhya*, April 16 2020,

https://www.analyticsvidhya.com/blog/2020/03/what-is-multicollinearity/#:~:text=Multicollinearity%20occurs%20when%20two%20or,variable%20in%20a%20regression%20model

(6) Team, The Investopedia. “R-Squared vs. Adjusted R-Squared: What's the Difference?” *Investopedia*, Investopedia, 14 June 2022,

https://www.investopedia.com/ask/answers/012615/whats-difference-between-rsquared-and-adjusted-rsquared.asp.

(7) No author. “Linear Regression (Part-3)- the Underlying Assumptions !” *Medium*, Towards Data Science, 24 Feb. 2022,

https://towardsdatascience.com/linear-regression-part-3-the-underlying-assumptions-82a66d5d5dd5

(8) Reed, Todd. “Perpendicular Distance Least-Squares Fitting for Arbitrary Lines.” *Perpendicular Distance Least-Squares Fitting for Arbitrary Lines-Todd Reed*,

https://www.toddreed.name/articles/line-fitting/

(9) Kumar, Rajarshi. “Life Expectancy (WHO).” *Kaggle*, 10 Feb. 2018,

https://www.kaggle.com/datasets/kumarajarshi/life-expectancy-who/versions/1?resource=download

(10) Gadige, Harshini. “Life Expectancy Cleaning,EDA,Feature Engineering.” *Kaggle*, Kaggle, 18 May 2019,

https://www.kaggle.com/code/harshini564/life-expectancy-cleaning-eda-feature-engineering