STA302 Fall 2023 Methods of Data Analysis 1
Final Project Proposal (Part 1)

Names of Group Members	Contribution to Proposal
Nida Li	Peer review articles, Part A3-4, C3, B2-3, data cleaning
Dawson Li	Group leader, R code, part A1-2, B1, C1-2, C4
Betty Liu	Data selector, R code, B2, C1, C5, data cleaning
Haozhe Wang	part C3-C4

A. Research Question and Supporting Literature

1. *What is the research question you will be studying in this project? Be sure to explicitly refer to the variables under study and avoid using vague language to describe your study question.*

The research question we examine is how we can predict life expectancy in different regions or countries based on previous years and a variety of variables. The response variable is obvious lifespan times. The predictor variables included: the year between 2000 to 2015, different countries and regions, probability of death among adults aged 15 to 60, baby newborn death, alcohol consumption over the years, education qualities, economic strength and prosperity. Furthermore, predictor variables related to health: immunization of Hepatitis B, the infection of HIV/AIDS and the collective weight status Body Mass Index(BMI). In addition, the categorical variables are the status of countries which is whether the country is developed or developing.

2. *Provide an explanation for why a linear regression model would allow you to answer your research question. What aspect of your fitted model would give you the answer.*

Once a linear regression model is fit to the data, it can be used for prediction and estimation. We can use a model to predict the response variable's value based on the predictor variable's value. Furthermore, in multiple linear regression, we can assess the relative importance of different predictor variables according to the changes in response variables. By examining the size and significance of each coefficient and the p-values we can determine which variables have the greatest impact on the results.

3. *Provide proper citations for 3 peer-reviewed academic research articles related to your specific research question or your topic of interest. For each, describe how the results of the article relate to your research question. Further, rank each article on a scale of 1 to 3 (1=not useful, 2=slightly useful, 3=very useful) based on how useful the article is in providing insight into the population relationship you wish to estimate. Justify this ranking.*

Citation	Description, ranking and justification
Meara, E. R., Richards, S., & Cutler, D. M. (2008). The gap gets bigger: Changes in mortality and life expectancy, by education, 1981 - 2000. Health	Usefulness: 1 The research findings of the study indicate that during the period from 1981 to 2000, the life expectancy of individuals with higher education consistently exhibited greater growth compared to those without higher education, with educational disparity being a significant contributor to the observed disparity in life expectancy. Furthermore, the factor of smoking is closely associated with the level of education, as smoking-related diseases constitute one of the primary risk factors for mortality. Prior to the implementation of tobacco control policies, public health issues stemming from smoking-related diseases led to a shorter life

Affairs, 27(2), 350–360. https://doi.org/10.1377/hlthaff.27.2.350	expectancy. These findings align with two determinants in our project, namely "Schooling" and "Public Health." The utility of this article is somewhat limited for several reasons. Firstly, the study's scope is restricted as it primarily focuses on non-Hispanic black and white populations from the 1980s and 90s, whereas our research pertains to global life expectancy. Secondly, the variables of educational attainment and public health address only two of the determinants we are investigating in our study.
Lin, R.-T., Chen, Y.-M., Chien, L.-C., & Chan, C.-C. (2012). Political and social determinants of life expectancy in less developed countries: a longitudinal study. BMC Public Health, 12(1), 85–85. https://doi.org/10.1186/1471-2458-12-85	Usefulness: 2 The research findings of the article conclude that life expectancy in less developed countries is influenced by various determinants, including GDP, literacy rate (educational resources), nutritional status, and political system (degree of democracy), among others. Among these determinants, GDP and educational attainment emerge as the primary factors affecting life expectancy. The utility of this article ranks second because it aligns more closely with our research. In comparison to the first article, the study population in this article is more relevant as it encompasses approximately half of the countries worldwide. Furthermore, the countries selected in this study are all developing nations, which is related to the determinant of "status," one of our key factors of investigation.
Kabir, M. (2008). Determinants of life expectancy in developing countries. The Journal of Developing Areas, 41(2), 185–204. https://doi.org/10.1353/jda.2008.0013	Usefulness: 3 The research findings of the article indicate that increases in GDP, educational resources, and healthcare quality are crucial for social development, but do not necessarily lead to an improvement in life expectancy. Significant gains in life expectancy are associated with improvements in public health, such as access to safe drinking water, reduction in malnutrition, and an increase in the number of healthcare professionals, among other factors. The utility of this article ranks first. The determinants mentioned in the article align with those in our project. Furthermore, this article examines life expectancy across a vast majority of regions worldwide, both in terms of scope and relevance to our project. The article provides a clear delineation of the contribution of each determinant to life expectancy, including factors that have a negative impact on life expectancy.

4. *Provide the database/library where you located the above academic papers. List the search terms used to find these papers, in addition to the number of results for each search term.*

Database/library searched	Search terms used	Number of results for each
Library Search U of T	Determinants of Life Expectancy Life Expectancy Determinants of Life Expectancy in Developing Countries	5,015 152,768 678
Google Scholar	Life Expectancy	562,000
Ulrich's Web	Life Expectancy	305

B. Data Description, Justifications and Summary

1. *Provide the website from which your chosen data was obtained/downloaded.*

Website:**	https://www.kaggle.com/datasets/kumarajarshi/life-expectancy-who/data

*** If your data was obtained from a data repository (e.g. Kaggle, UCI Repository, etc.), please state how your research question differs from the original purpose of these data.*

- The dataset comes from Kaggle, which shows that the data we use has very high accuracy. The research question initially aims to answer the following questions: do the various predictors initially chosen actually affect life expectancy? And how various factors influence life expectancy. However, Kaggle's method of analyzing data is very complex. The predictor variables we selected were not all identical. This is because we only get part of the data set, not all of it. On the other hand, we will interpret the data based on linear regressions. In addition, we want to do more research on how can we predict future life expectancy based on these data.

2. *List the variables you have selected to be part of your preliminary model (minimum of 5 with at least one a categorical variable). Please give an understandable name to each variable rather than writing the name that appears in R.*

   For <u>each</u> variable, justify why you have chosen to use this variable over others in the dataset, and what the role of each variable will be (e.g., predictor of interest, predictor informed by literature, confounder, etc.).

Variable Name	Justification for Use	Role in Model
Lifespan time --- Numerical variable	We are interested in predicting future life expectancy based on our data set. Also, we want to examine how life expectancy is affected by other factors.	Predictor of interest
Year between 2000 to 2015--- Numerical variable	We are interested in how life expectancy is changing from the past.	Predictor of interest
Developing & Developed status of a country --- Categorical variable	We are interested in the question, do people in developed countries generally live longer on average?	Predictor of interest
Adults' deaths before 60th birthday --- Numerical variable	We are interested in studying the impact of age structure on life expectancy which is an important factor.	Predictor of interest
Newborn death --- Numerical variable	Similar to Adult Mortality, life expectancy is definitely related to age structure.	Predictor of interest
The consumption of alcohol --- Numerical variable	Drinking alcohol is a common behaviour among people nowadays, and many minors are also drinking. We are curious whether alcohol intake will have an impact on people's life expectancy.	Predictor of interest
Immunization Hepatitis B --- Numerical Variable	Informed by the literature that Immunization of hepatitis B is a very important aspect of public health. It is essential for human health, and to prevent diseases.	Predictor informed by literature
BMI --- Numerical Variable	Average body mass index which is Body fat measurement based on height and weight. We are interested in examining people's health for life expectancy.	Predictor of interest
HIV/AIDS death rate --- Numerical Variable	HIV/AIDS, a virus that attacks the immune system has impacted people's lives heavily. Informed by the literature that this is an important health factor that contributes to human life expectancy.	Predictor informed by literature

Measurement of wealth --- Numerical Variable	A measure of the added value a country creates by producing goods and services. The degree of wealth freedom in a country greatly affects the quality of life.	Predictor informed by literature
The population density. --- Numerical Variable	Under the same circumstances, a country's resources are limited, and population density largely determines the resources available to each person.	Confounder
Education levels --- Numerical Variable	It is known from the literature that people's level of education has a relationship with life expectancy.	Predictor informed by literature

3. *Produce a table of numerical summaries of the variables listed above. Summaries should be appropriate to the type of variable, and interesting/important characteristics about variables should be mentioned in an informative caption. Include your summary table below.*

Summary Table of Numerical Variables for Life Expectancy

	Life Expectancy	Year	Status Developing|ed	Adult Mortality	Infant Deaths	Alcohol	Hepatitis B	BMI	HIV/AIDS	GDP	Population	Schooling
Variable Type	Continue response variable	Continue numerical variable	Discreate Categorical variable	Continue numerical variable	Continue numerical variable	Continue numerical variable	Continue numerical variable	Continue numerical variable	Continue numerical variable	Continue numerical variable	Continue numerical variable	Continue numerical variable
Min	44.0	2000	0|0	1	0	0.01	2.0	2.0	0.1	2	3.40e+01	4.2
1st Qu	63.3	2005	1|0	77	1	0.79	75.00	19.7	0.1	463	1.88e+05	10.4
Median	71.7	2008	1|0	148	3	3.82	91.00	43.9	0.1	1631	1.38e+06	12.3
Mean	69.3	2008	0.852|0.148	169	24	4.55	79.6	38.3	2.0	5601	1.08e+07	12.1
3rd Qu	75.0	2011	1|0	228	22	7.38	96.0	55.8	0.7	4773	7.42e+06	14.0
Max	89.0	2015	1|1	723	556	17.87	99.0	77.1	50.6	119173	2.55e+08	20.7
Total Data Used	1637	1637	1395|242	1637	1637	1637	1637	1637	1637	1637	1637	1637

From the above table, we observe a great difference between the minimum and maximum values of life expectancy. Although the averages for adult and infant mortality are quite low, the maximum values for each of those categories are quite large. Furthermore, there is a huge gap between the highest and lowest levels of HIV/AIDS, which means that medical protection in many countries is not very good. In addition, the gap between rich and poor is very large, and most of the world's countries are in the developing stage. The GDP of the poorest countries is only a few thousandths of the average GDP, as well as the maximum education level is almost 5 times the minimum.

C. Preliminary Model Results

1. *Fit your preliminary multiple linear model and present the estimated relationship. Present this information carefully so that it is easily readable and understandable.*

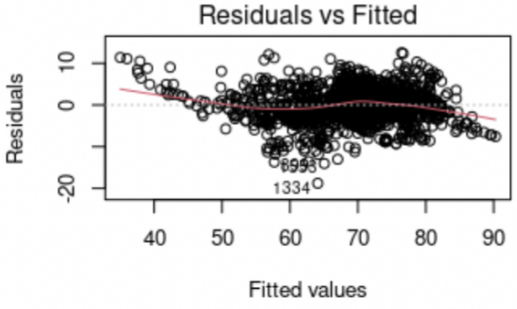
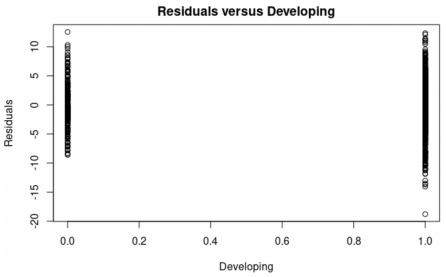


Table for Multiple Linear Model Estimated Coefficients

Coefficients	Intercept	Year	Developing	Developed	Adult Mortality	Infant deaths	Alcohol	Hepatitis B	BMI	HIV/AIDS	GDP	Population	Schooling
Estimate	2.89e+02	-1.16e-01	-8.64e-01	NA	-1.97e-02	-4.40e-03	-8.77e-02	9.18e-03	4.42e-02	-4.56e-01	8.22e-05	4.49e-09	1.36e+00
Std.Error	4,93e+01	2.46e-02	3.64e-01	NA	1.01e-03	1.71e-03	3.51e-02	3.96e-03	5.98e-03	1.92e-02	1.01e-05	3.63e-09	5.51e-02
T value	5.86	-4.73	-4.73	NA	-19.47	-2.57	-2.50	2.32	7.39	-23.71	8.12	1.24	24.72
Pr (> | t |)	5.5e-09	2.5e-06	0.018	NA	< 2e-16	0.010	0.013	0.020	2.3e-13	< 2e-16	9.1e-16	0.217	< 2e-16

The Residual vs Fitted plot gives the residuals on the fitted value where when the fitted value reaches around 60, the variance of the residual reaches the maximum. The Normal Q-Q plot shows the high normality of the dataset. The Scale-Location plot shows how equally the residuals spread along the ranges of predictors, with the range being mostly between 0.0 to 2.0. The Cook's distance plot gives possible outliers, where the unusual peaks are located.
After summarizing the preliminary multiple linear model. The Intercept indicated that when all other predictor variables are set to a constant 0, the response variable which is life expectancy would result in a positive change in 289. The p-value of the intercept is less than 0.001. The estimated relationship is that the predictor variables Year, Adult Mortality, BMI, HIV/AIDS, GDP and Schooling are highly significant in explaining the variation of the response variable because they all have extremely low p-values smaller than 0.001. Other predictor variables: infant deaths, alcohol and Hepatitis B are also statistically significant but have less impact because their p-values indicate they are in the range between 0.01 to 0.05. Lastly, the population has the least influence on the response variable because its p-value is 0.217.

2. *Justify your choice of how you included the categorical variable in your preliminary model. How does this choice contribute to answering your research question?*



We included categorical variables by dividing life expectancy into two groups. One is a developed country and the other is a developing country, this is because citizens have better healthcare, earn more income, have higher education levels, etc. According to our model output, the development is a dummy variable, and its coefficient is negative. This means that holding all other predictor variables constant, being a developing country will result in a negative change of 0.864 in the response variable. Therefore, generally speaking, life expectancy is higher in developed countries because living conditions are better in these countries. Furthermore, the developed shows all the values with NA, this is because it is collinear with Developing. By picking our categorical variable, we can use this option to better understand how life expectancy will change in the future and will increase if the country continues to prosper and grow.

3. *Do your estimated coefficients align/agree with the results of your three peer-reviewed articles? Explain in what way they differ/agree and provide a reason why this might be the case.*

The first article demonstrates that higher education levels or we can say longer school years lead to a relatively long life expectancy. By observing our data's coefficients, we can conclude that the predictor variable schooling and our response variable life expectancy are in a positive relationship. As well as the coefficients of BMI and Hepatitis B also indicated the positive relationship between public health and life expectancy. On the other hand, public health is closely related to education level. People may overdose on cigarettes due to the low level of education with no cognition about how cigarettes may influence your health. The p-value we estimated for schooling is around 2e-16 which is a really low value, hence the effect of education level is closely related to life expectancy. Thus, the points made in the article are consistent with the data we derived from our code predictions.

Focus on the second article, indicates that life expectancy is related to GDP and the status of a country. It matches two of our variables, according to the article, it mainly focused on developing countries. It states that GDP plays an important role when we predict life expectancy. For those underdeveloped countries' life expectancy is relatively low compared to developed countries. By the estimated coefficients of GDP, we conclude that life expectancy increases with the increase in GDP. Further, If we take a look at the p-value for GDP which is 7.6e-16, we can tell that it is extremely small which is much smaller than 0.001, which proves that GDP is a significant factor when we consider life expectancy. This perfectly matches the article.

Last but not least, the third article we found elaborated that the increase in GDP, education level or other factors has a great influence on social development, but not 100% on expanding life expectancy, the variables in this article perfectly match the majority of our predictor variables. We may conclude that it somehow increases life expectancy but not directly. The variable mentioned in this article as well as in our research contains GDP, schooling, Health care and developing countries. However, it does not state that GDP and schooling have direct influences on life expectancy which is different from our statistical results. We believe there may be several reasons such as differences in sample size, differences in time, or it can be heavily influenced by random variability.

4. *Perform a complete assessment of the assumptions of your preliminary model. Do you observe violations of assumptions or conditions? Describe how you came to this conclusion, making explicit reference to any plots or other information that is relevant.*

If we take a look at the scatter plot we can conclude that the relationship between our response variable and some of the predictor variables is linear. On the other hand, the QQ plot also shows a linear relationship, the slope of the trend is positive. Since the line is around 45 degrees, this indicates the data is approximately normally distributed. Also, the points in the QQ plot are below the straight line at the lower end and above the straight line at the higher end, which indicates positive skewness in the data. Hence, we conclude that we have followed the normality assumption. In addition, the residual is a relatively good model in this situation, after observation of the Residual vs Fitted plot, about half of the residuals are positive and half negative, and they are symmetrically distributed about the zero line. So the constant variance assumption is obeyed. Since the residual plot indicated that our data points have distributed well, it leads to a well-fitting model and shows the linear relationship in the coefficients, as well as low p-values. Therefore, our preliminary model indicated that we have followed the linearity assumptions. Finally, each data point in the population is uncorrelated or connected to any other data point, indicating that we are following the uncorrelated error assumption.

5. *Include all relevant plots created for assessing model assumptions below, with appropriate axis labels and captions.*