Exploring UN Data: Correlation Between Education and Population

Abstract

This paper explores the question of: to what extent do education rates of certain countries and regions correlate with life expectancy and other important health outcomes? We found that there was a statistically significant relationship between regions of the world and life expectancy levels using a Chi Squared Test. We then turned to investigate the relationships between education enrollment and life expectancy in countries. We also created a linear model to demonstrate the relationship, which indicated there is a positive linear association between the log of proportion enrolled and life expectancy. These findings seem to support the notion that UN efforts to expand education globally and to specific populations, such as women, have been met with success. This makes clear that education is strongly related to life expectancy and perhaps can be used to understand and address other important measures of public health.

I. Background and Introduction

Since its founding in 1945, the United Nations has been regarded as the world's foremost intergovernmental organization for international peacemaking, development, and record keeping. As such, the UN's data collection provides vast knowledge about the state of our world and its most pressing issues. Our team is concerned specifically with the question of: to what extent do education rates of certain countries and regions correlate with life expectancy? Through this investigation, we explore relationships that touch on these varying topics of gender, population and education levels in order to better understand if education factors are associated with a country or region's population statistics. The findings of this project could be used as evidence for advocating for education initiatives especially in countries with low life expectancies. By analyzing the UN data, we draw attention to the inequalities in education throughout the different regions of the world and show the importance of education of both males and females through the correlations to life expectancy.

II. Data and Exploratory Analysis

Four datasets were joined to address our research question. Two datasets, containing education and population data for countries and regions, were obtained through UNdata [1–3]. A dataset containing mappings between country code and region was obtained through Kaggle [4]. A dataset containing country and territory populations from 2015 was obtained through the UN Department of Economic and Social Affairs World Population Prospects site [5]. The education and population datasets were accessed from UNdata, a web-based data service that brings international statistical databases through a single-entry point [6]. The data itself was compiled by the United Nations (UN) statistical system and other international agencies.

The education dataset contains variables describing overall enrollment in three education levels (primary, secondary, tertiary) as well as enrollment ratios between females and males for 213 countries and territories. We focused on tertiary enrollment, which is generally defined as college or university. The population dataset contains variables describing life expectancy and other factors affecting population growth for 234 countries and territories. These datasets were joined to allow for the analysis of relationships between education and population between the 204 countries and territories present in both datasets. These datasets also contain data by year, primarily for 2005, 2010, and 2015. The dataset containing mappings to regions was incorporated to allow for grouping of countries and territories by region. The total populations dataset from 2015 was incorporated to allow for further analysis of population trends in 2015.

We wanted to understand the distribution of life expectancy across regions in our dataset. The visualizations in Figure 1 allowed us to identify some preliminary trends regarding the relationship between education and life expectancy.

A more detailed breakdown of tertiary enrollment ratio and life expectancy is found in Appendix A. For every year for which we have data, higher enrollment ratios are associated with higher life expectancy.

III. Results

We exclusively focused on data from 2015, the most recent year for which we have data, and used a significance level of $\alpha = 0.05$ for each of our statistical tests.

Based on the visualization in Figure 1 and a Chi Squared test (See Appendix B), we saw that there

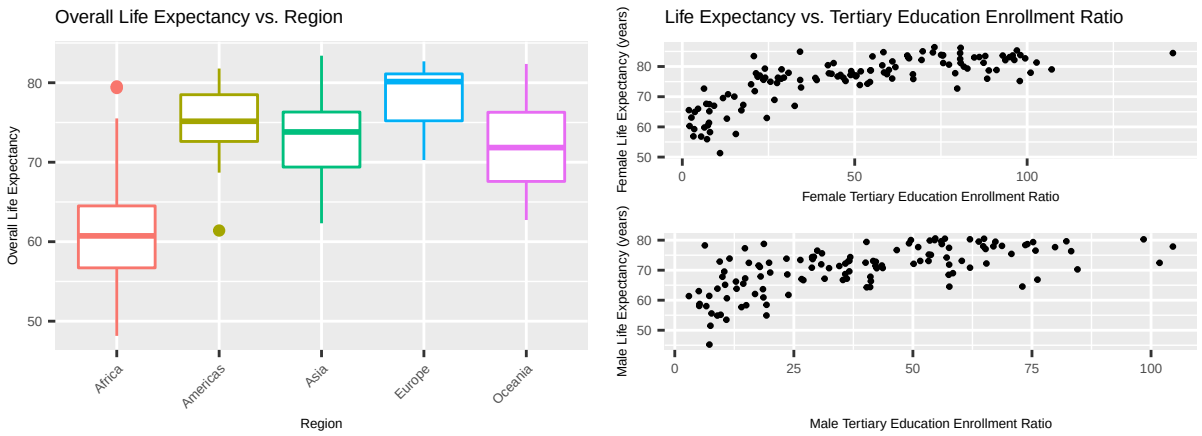


Figure 1: *These visualizations display data from 2015. There are some clear differences that emerge when comparing life expectancy in different regions. Life expectancy also seems to be positively correlated with tertiary enrollment ratio for both males and females.*

is a relationship between life expectancy and region of the world. We wanted to further investigate if this relationship could be affected by education rates of enrollment. Based on our earlier exploration of tertiary enrollment ratio and life expectancy for males and females (See Appendix A), we expected that there would be a relationship between life expectancy and tertiary education enrollment. In order to study this relationship, we fit a linear model. We can understand the enrollment level in tertiary education by taking the proportion of the population enrolled in tertiary education for a given country (See Figure 2).

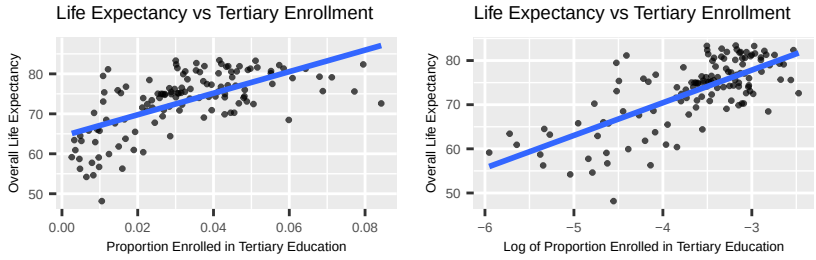


Figure 2: *Original and Transformed Tertiary Education vs. Life Expectancy Visualization.*

The raw data has a logarithmic shape, so the proportion enrolled values were transformed using the natural log, which makes the trend appear more linear. The linear model developed is using this transformed proportion enrolled data. The calculated model indicates that there is a significant positive linear relationship between the two variables (See Appendix C for the mathematical model). According to the model, if the percent enrolled in tertiary education multiplies by a factor of 2, the life expectancy of a country is expected to increase by $7.398 * \ln(2) = 2.23$ years.

After establishing a positive linear relationship between life expectancy and the log proportion enrolled in tertiary education, we then investigated disparities in education by checking for statistical significance regarding the difference in average tertiary enrollment ratios between males and females.

We then used two-sample inference to test the null hypothesis that the average tertiary education enrollment ratio of females is less than or equal to the average tertiary education enrollment ratio of males. This t-test assessed whether the difference in these means is actually significant.

We first calculated the observed difference in mean tertiary education enrollment ratio between males and females as 8.426 years. We then conducted the one-sided t-test to observe if this difference in means is significant. From the test we receive a test statistic of 2.839 (See Appendix D). We obtained the p-value of 0.002 which is smaller than $\alpha = 0.05$. Therefore, we rejected the null hypothesis, meaning there is sufficient evidence to conclude that the average tertiary education enrollment ratio of females is greater than the average tertiary education enrollment ratio of males.

IV. Discussion and Conclusion

We began our investigation with the question, "to what extent do education rates of certain countries and regions correlate with life expectancy?" To break down this question, we first found that there was a statistically significant relationship between regions of the world and life expectancy levels using a Chi Squared Test. We then turned to investigate the relationships between education enrollment and life expectancy in countries. We created a linear model which indicated there is a positive linear association between the log of proportion enrolled and life expectancy. Our model is consistent for the range of log proportions in our data, so life expectancy values should not be extrapolated using much larger proportion values. This model may also benefit from more explanatory variables that are not education-related. We then found sufficient evidence to conclude that females, on average, have a higher life expectancy than males and, on average, are more educated than males at the tertiary level.

For further analysis, we may want to look into data that is from a more recent year than 2015, especially as the COVID-19 pandemic has significantly changed the way that students learn, and different countries have different approaches to handling coronavirus and education. We could have also explored different variables other than those involving tertiary education, such as other levels of education. Tertiary enrollment may also vary in definition across regions or countries, so it could limit the validity of comparing this by region and country. We also mostly focused on life expectancy, but there are other important trends in population, such as infant mortality rate and fertility rate (See Appendix E) in our original dataset, that may be interesting to explore. We could further explore the correlation between life expectancy and enrollment ratio to confirm that the relationship between the overall life expectancy and enrollment proportion extends when analyzing by gender.

In conclusion, our exploration shows strong evidence that there is a positive correlation between education and population trends. It also provides important insights into how and where there exist disparities in achievement, which in turn can help inform policy decisions and resource allocation. These findings seem to support the notion that UN efforts to expand education globally and to specific populations, such as women, have been met with success. We did not explore other underlying factors, such as countries' wealth, as a driving force of the relationship. We can expect that wealthier countries have higher proportions of people enrolled in tertiary education and higher life expectancies, thus maintaining that education is an important factor among others. Although our analysis does have limitations, our findings can support advocacy for better public health through the advancement of education around the world.

References

1. *UNdata*. (n.d.-). United Nations Statistics Division. Retrieved from http://data.un.org

2. *Population growth and indicators of fertility and mortality*. (2019). United Nations Statistics Division. Retrieved from http://data.un.org/_Docs/SYB/PDFs/SYB62_246_201907_Population%20growth%20and%20indicators%20of%20fertility%20and%20mortality.pdf

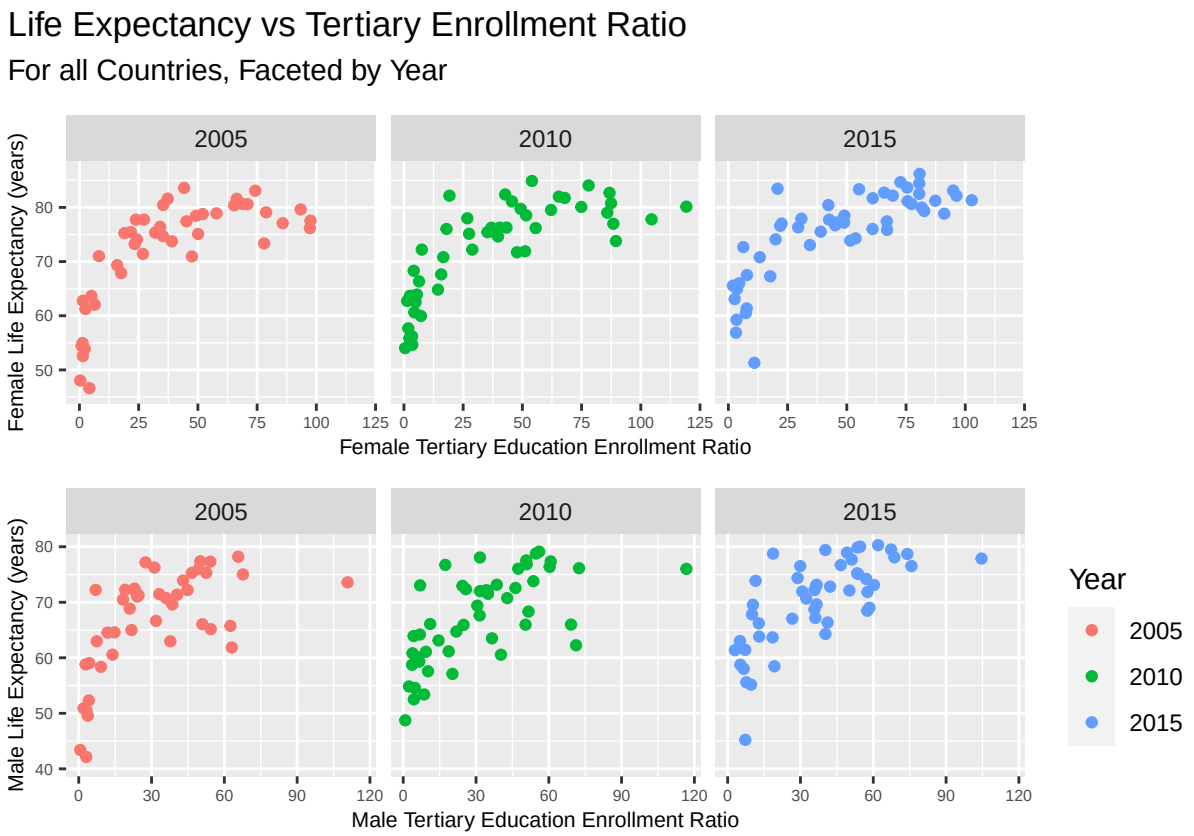
3. *Enrollment in primary, secondary, and tertiary education levels*. (2019). United Nations Statistics Division. Retrieved from http://data.un.org/_Docs/SYB/PDFs/SYB62_309_201906_Education.pdf

4. *Country mapping - iso, continent, region*. (2019). Kaggle. Retrieved from https://www.kaggle.com/andradaolteanu/country-mapping-iso-continent-region

5. *Total population*. (2019). United Nations Department of Economic; Social Affairs. Retrieved from https://population.un.org/wpp/Download/Standard/CSV/

6. *UNdata | about us*. (n.d.-). United Nations Statistics Division. Retrieved from http://data.un.org/Host.aspx?Content=About

Appendix A



These visualizations shows that as female tertiary education rates increase, so does female life expectancy, and that as male tertiary education rates increase, so does male life expectancy. And despite several outliers, the general trend shows that overall life expectancy is increasing through the years for both males and females.

Appendix B

statistic	chisq_df	p_value
77.71568	4	0

According to the UN's world population prospects in 2015, the average life expectancy throughout every included country in the world was 71 years of age [1]. We use a chi-squared test to determine whether or not there is a relationship between life expectancy and region of the world. The null hypothesis is that there is no relationship between life expectancy and region of the world, and the alternative is that there is a relationship. We will categorize the variable of life expectancy as either greater than or equal to the global average ("High") of 71 or less than that global average ("Low").

Our chi-squared test statistic is 77.7 with 4 degrees of freedom with an extremely small p-value approaching zero. This tells us that we can reject the null hypothesis, meaning there is sufficient evidence to suggest that there is a relationship between region of the world and life expectancy of the population on average. This makes sense in accordance with our original exploration of the data that showed that the average life expectancies of each region of the world varied (See Figure 1).

Appendix C

The resulting linear model with response variable life expectancy and explanatory variable log of tertiary enrollment proportion is as follows:

$$\widehat{\text{lifeexpectancy}} = 100.032 + 7.398 \times (\ln(\text{tertiaryenroll}))$$

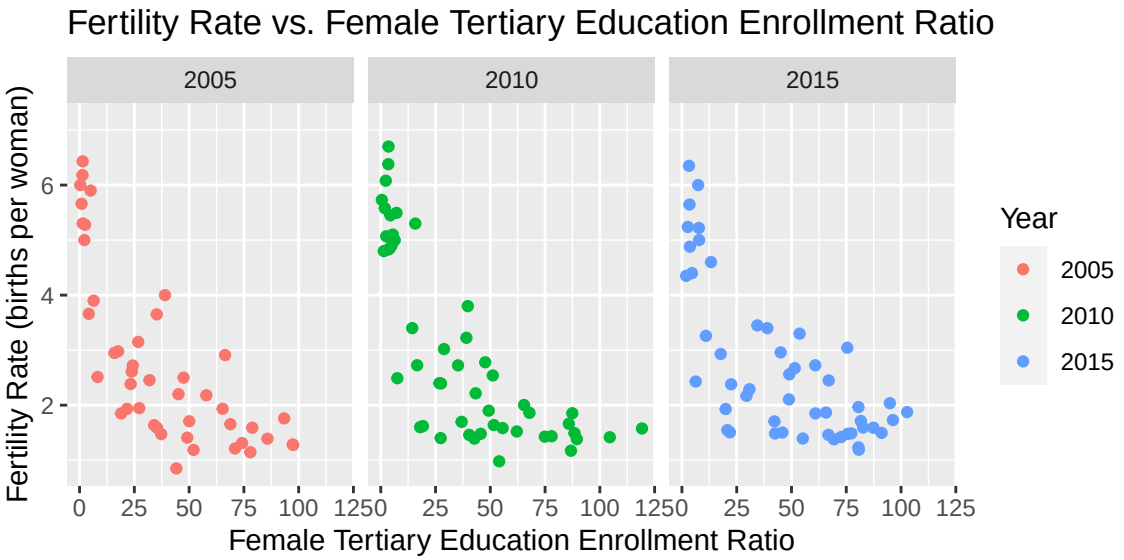
For the model, $R^2 = 0.514$, indicating that 51.4% of the variance in overall life expectancy is explained by the log of the percent tertiary enrollment. For every unit increase in the log proportion of enrollment in tertiary education, the life expectancy of a country is predicted to increase, on average, by 7.398 years. In other words, if the percent enrolled in tertiary education multiplies by a factor of 2, the life expectancy of a country is expected to increase by $7.398 * \ln(2) = 2.23$ years. The y-intercept of this model defines the life expectancy when the log proportion of enrollment is 0, which is also when the proportion of enrollment is $(0 = \ln(1))$ 100%. This means that if a country has 100% of its population enrolled in tertiary education, the life expectancy is expected to be, on average, 100.032 years.

Appendix D

statistic	t_df	p_value	alternative	lower_ci	upper_ci
2.839	225.87	0.002	greater	4.306	Inf

We conducted the one-sided t-test to observe if the observed difference in means between males and females is significant. From the test we receive a test statistic of 2.839, which is a slight difference from our observed difference in means due to the nature of the t-test equation which accounts for the difference divided by standard error of the difference. We obtained the p-value of 0.002, a very small number approaching zero that is smaller than our alpha, $\alpha = 0.05$. Therefore, we reject the null hypothesis $H_0 : \mu_{female} = \mu_{male}$, meaning there is sufficient evidence to conclude that the average tertiary education enrollment ratio of females is greater than the average tertiary education enrollment ratio of males.

Appendix E



This visualization demonstrates that higher education levels for women seem to have a relationship with female fertility levels as we can see that higher education levels for women are associated with lower female fertility rates.

For the purposes of this paper, we did not further investigate the trends in the visualization.