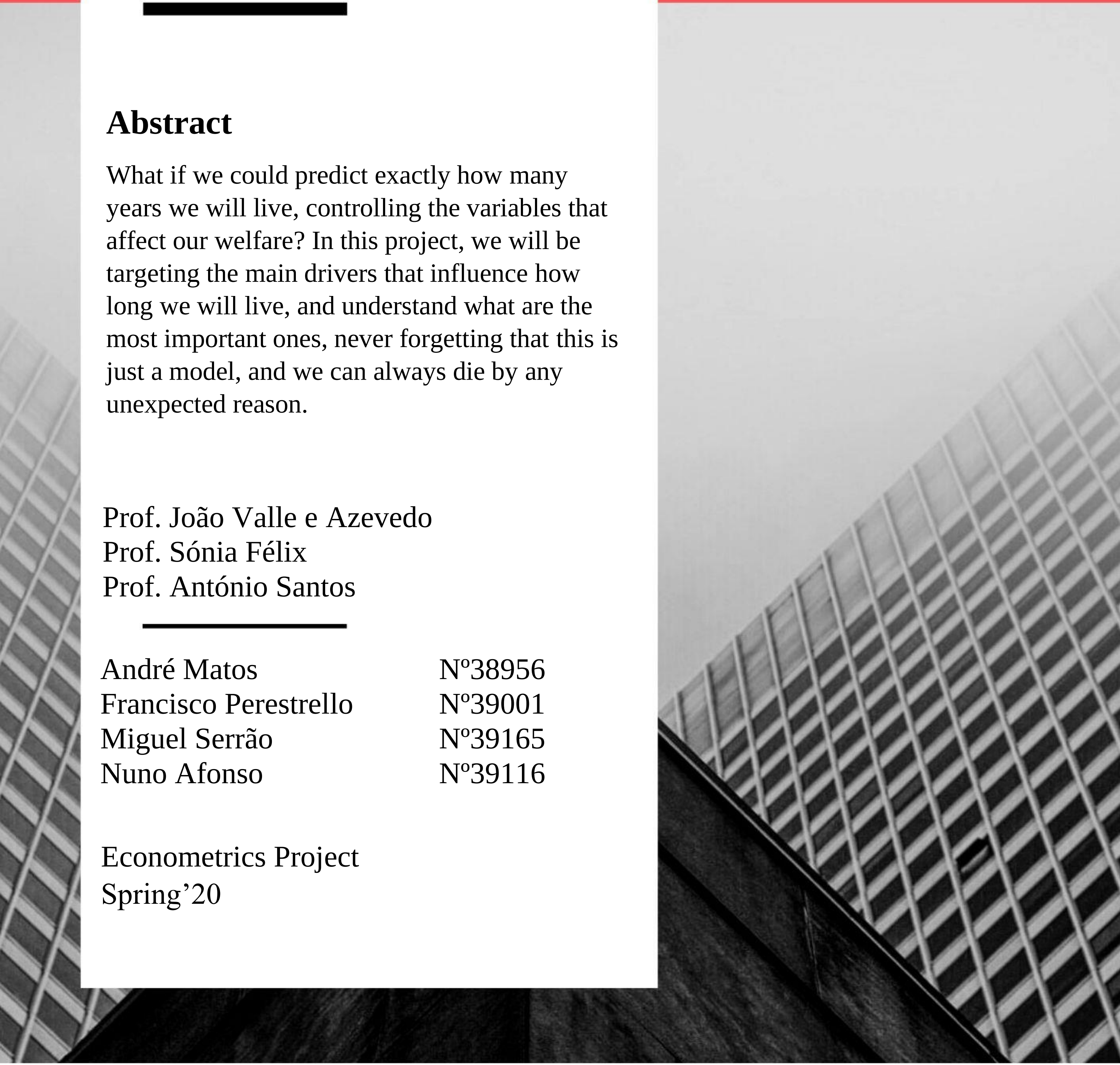
NOVA SCHOOL OF
BUSINESS & ECONOMICS

Does Education Affect Life Expectancy?



Abstract

What if we could predict exactly how many years we will live, controlling the variables that affect our welfare? In this project, we will be targeting the main drivers that influence how long we will live, and understand what are the most important ones, never forgetting that this is just a model, and we can always die by any unexpected reason.

Prof. João Valle e Azevedo
Prof. Sónia Félix
Prof. António Santos

André Matos	Nº38956
Francisco Perestrello	Nº39001
Miguel Serrão	Nº39165
Nuno Afonso	Nº39116

Econometrics Project
Spring'20

Introduction

Does education affect our life expectancy? Is it viable to study for more years to increase the time we live? That's the main question we try to answer with this econometrics project. Each country has its life expectancy, which itself depends on a lot of variables such as air quality, income distribution, or even the percentage of smokers, for example. Of course, multiple variables can help us predict life expectancy, but our work will focus on the effect of education on this variable throughout multiple countries. Based on econometric tools, we want to predict the relevance of education in this model, hence we will conduct statistical tests on this variable as well as many others to infer on the relevance of education, both individually as well as when compared to other variables.

We decided to explore this topic because we believe that education could have a great influence on life expectancy. Also, besides our personal interest, it could be useful for governments to define some important issues. This model could, for example, help governments define how much to invest in education, as life expectancy could be seen as a welfare measure. On another note, it could also help make decisions concerning the retirement age, as the average age of death is a big factor in deciding where to draw the line for it. This is important because it too has big implications on the state budget, as governments must pay pensions to all retired workers. Furthermore, the model we are about to present could also help in the decision-making around human capital investment for firms.

Throughout the last couple of decades, investment in education has been in a constant upswing, as well as life expectancy all over the world. What we want to address with this work is whether the latter is explained by the first or not. Other socioeconomic factors can explain life expectancy but, focusing on education, we will try to define whether this relation exists and how strong it is.

By a conceptual framework, the countries with the lowest life expectancy usually display low household income levels, low health care levels, and low levels of education.

Another important factor to account for is the average education of the household, as parents with high levels of education tend to provide their children with the same opportunities. On the other hand, if we think about our grandparents in Portugal, most of them never had access to a good health care service nor to a lot of years of education, they usually had bad eating habits, went to war and many were smokers and still they lived until they were eighty or ninety years old. This brings back our question, is education really relevant to our life expectancy? Looking at our grandparents' example, probably not. However, if we look at other countries like Japan, where life expectancy has always been one of the highest in the world, we can see that their investment on education is massive, or even Turkey, where we observe the biggest increase in the last thirty years, with this indicator increasing more than twelve full years. This poses our question one final time, for which we will try to answer with the coming model.

Taking more of a theoretical approach, what we exactly pretend to do is to test our model, understand if any of the Gauss-Markov assumptions are violated, and, if that's the case, try to explain the reasoning. Additionally, we intend to test statistical significance to try to reach a conclusion about education, as the available information is a bit contradictory: some studies conclude one thing, and others the opposite. To do this, we will interpret the t-values and p-values as well as discuss whether our model is homoskedastic or not. Our main goal is to understand if studying more will, indeed, increase life expectancy, ceteris paribus.

Literature Review

After some research about the estimators of life expectancy, we found some papers that discussed this matter. For our study, we focused on two papers, the first being the "*Variations in Life Expectancy in Organization for Economic Co-operation and Development countries*", by H. Zare, D. J. Gaskin, and G- Anderson in 2015, and the second "*The Determinants of Life Expectancy: An analysis of the OECD Health Data*", by J. W. Shaw, W. C. Horrace, and R. J. Vogelf back in 2005. Even though education is not the main focus of these papers, they gave us some key ideas to understand some concepts and some details about the formulation of their models, which helped us create our own. Moreover, these papers also contributed to our work as we were able to find informative sources through the references of these papers, which lead us to other econometrist papers regarding this topic.

We encountered a study from Gradstein and Kaganovich from 2004 in which they focused on the main factors that explained the increases in longevity over the twenty years before the study and the relation of public funding of education and economic growth, explaining how economic growth influenced how long their population lived. The increase in longevity introduces political and economic implications, and this paper depicts well some important relations, such as the inverse relationship between the fraction of the elderly population and the spending on public education across the states and communities in the United States of America. This study also shows that government expenditure on education is based on the collective interest in the future productivity of younger adults in the economy. The increase in life expectancy, according to this article, generates two main effects: a negative effect created by the increase of the old population and a positive effect created by the increase in support for education funding among the younger population. In the article, they prove that the latter overcomes the first, resulting in a positive effect on society. All in all, the model they presented allows us to conclude there is a causal effect between increasing the investment in education and the increase in life expectancy.

Another relevant study we found was conducted by Cremer, Lozachmeur, and Pestieau also in 2004, which regards the age of retirement. In the article, they point that given life expectancy is increasing and keeping the fertility rate relatively constant, the population of a given country will grow (it is based on this theory that worldwide population has increased over the last decades) and governments, knowing this, should prepare all public expenses based on this. Another important thing that they expose regards the productivity of public workers. This closely relates to the increased incentives that governments have been creating recently to avoid old workers falling out of the labor market, as most of them are averse to new technologies, for example. Furthermore, the retirement age also has a big influence on whether a worker decides to early retire. Finally, the life expectancy model presented is primarily predicted by some main drivers, such as environmental measures, lifestyle measures, and consumption of health care measures. The thing that all these variables have in common, though, is the impact of the income level. With more income, better health care is obtained, consumption increases, and it becomes a lot easier to practice healthy living habits.

To help us bring all these articles together, it makes sense to talk about one last study, the Peltzman study, in which the author created a model that helped him find that by the regression of GLS, what essentially matters when trying to estimate the longevity of random countries is the income level of households, which is directly correlated with education, as usually more years of education translate into higher income levels, which then will lead to higher life expectancy. On a side note, an American study by Freeh H.E. and Miller R. concluded that smokers live for the same average time as non-smokers, which is not intuitive but is something we will try to understand by also including work on this variable in our model.

Empirical Model and Data

Sample and Data

Our sample consists of 78 countries (both developed and undeveloped) from multiple continents. Some countries were not included in our sample because of the unavailability of data.

All the data collected is from the same year (2016) and collected from the World Bank, with the exception of the countries' classification as developed or developing economies which was gathered from the IMF.

Empirical Model

To access how education affected a country's total life expectancy on birth, we developed the following model based on our literature review:

$$\begin{aligned} LE = \beta_0 &+ \beta_1 AdvancedEduc + \beta_2 \ln\,(GDP_{per\ capita}) + \beta_3 DomesticSavings \\ &+ \beta_4 Unemployment + \beta_5 HealthExpenditure \\ &+ \beta_6 AlchoholConsumption + \beta_7 Smoking + \beta_8 Urban \\ &+ \beta_9 Developed + u \end{aligned}$$

Our **dependent variable** is total life expectancy at birth in 2016. According to the World Bank, it consists of the "average number of years a newborn is expected to live if mortality patterns at the time of its birth remain constant in the future"[1]. As such, our variable measures the "mortality characteristics" of a certain country in 2016.

Initially our model had one additional variable (government expenditure per student, tertiary (% of GDP per capita)). However, we decided not to include it since there was a lack of information. Using this variable would reduce our sample size which could mean some of our assumptions would no longer be valid. It should also be noted that dropping this variable will not violate either the MLR.3 assumption nor the MLR.4 assumption.

The regressors in our model can be divided into 2 categories: macroeconomic-related or social-related.

- **Macroeconomy-related**

$GDP_{per\ capita}$: GDP per capita, PPP (current international $). Measures a country's total output divided by its total population considering the relative cost of local goods, services, and inflation rates of the country. It's measured in current international dollars based on the 2011 ICP round. This was the only independent variable that we decided to log since it made more economic sense to measure the effect of a 1% increase in GDP per capita rather than a 1$ increase. We predict that this variable will have a positive effect on life expectancy.

$DomesticSavings$: Gross Domestic Savings. It consists of savings of the household sector, the private corporate sector, and the public sector and is measured as % of GDP. We expect it to have a positive effect.

$Unemployment$: Unemployment. It refers to the share of the labor force that is employable and seeking a job but is unable to find a job. It is measured as a % of the total labor force and we hypothesize that it will negatively affect our explained variable.

$HealthExpenditure$: Current health expenditure. Measures the level of current health expenditure expressed as a % of GDP. It includes healthcare goods and services consumed during each year but not capital health expenditures. It will likely be positively correlated with life expectancy.

[1] Sources – Link 1

$Developed$: A dummy variable that is equal to 1 if the country is included in the group “advanced economies” in the IMF’s classification in the *World Economic Outlook*, and 0 otherwise. Advanced economies will likely have a bigger life expectancy, thus this regressor is positively correlated with the regressand.

- **Social-related**

$AdvancedEduc$: Labor force with advanced education. Describes the amount of people in the labor force that have a short-cycle tertiary education, a bachelor’s degree or equivalent education level, a master’s degree or equivalent education level, or doctoral degree or equivalent education level according to the International Standard Classification of Education 2011 (ISCED 2011). It is measured as a % of the total working-age population. The relation between this variable and life expectancy is one of the focuses of our study. We hypothesize that this relation is positive.

$AlchoholConsumption$: Total alcohol consumption per capita. This variable is defined as the sum of the recorded and unrecorded amount of alcohol consumed per person (15 years of age or older) over a calendar year, in liters of pure alcohol, adjusted for tourist consumption. Because of the health issues related to alcohol consumption, we believe it will have a negative effect on the dependent variable.

$Smoking$: Total smoking prevalence. It’s the percentage of men and women ages 15 and over who currently smoke any tobacco product on a daily or non-daily basis, excluding smokeless tobacco use. It will also likely have a negative correlation with life expectancy.

$Urban$: Urban population. Refers to people living in urban areas, measured in % of the total population. Since these areas often have better infrastructures and more job opportunities, we believe that this will have a positive effect on life expectancy.

This information is summarized in Table 1 as well as each variable’s descriptive statistics.

Variable	**Definition**	**Max**	**Min**	**Mean**	**SD**	**Expected Sign of Coeficcient**
$GDP_{per\ capita}$	GDP per capita, PPP (current international $)	108165,761	1269,88622	27120,4201	19858,4382	Positive
$DomesticSavings$	Gross Domestic Savings	54,4112276	58,759129	21,0658886	13,7236394	Positive
$Unemployment$	Unemployment	26,5365009	0,68800002	7,88502057	5,62910062	Negative
$HealthExpenditure$	Current health expenditure	17,1972603	2,31177993	7,39554299	2,50181901	Positive
$Urban$	Urban population	100	18,311	67,8775366	20,5637473	Positive

$AdvancedEduc$	Labor force with advanced education	92,6893997	59,8406982	78,5429451	5,96633062	Positive
$AlchoholConsumption$	Total alcohol consumption per capita	15,2	0	8,09638554	4,02563268	Negative
$Smoking$	Total smoking prevalence	43,4	2	22,7831325	9,55391401	Negative
$Developed$	Advanced economy	-	-	-	-	Positive

Violation of Assumptions

In this segment of our project, we focus on the potential violations of the Gauss-Markov assumptions under the Classical Linear Model (assumption MLR.1 to MLR.6). We have our basic empirical model specified and we explained some particular issues and from here we will test and ensure the consistency and unbiasedness of the results.

- **MLR.1 (Linearity in Parameters)**

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + u$$

The linear regression equation must respect the linearity in parameters, the model must have linear coefficients. Implying that no parameter appears as a non-linear form, for example, a parameter multiplied by another parameter.

In our model MLR.1 holds, every parameter is presented in linear form, this can be deducted only by looking at the model.

- **MLR.2 (Random Sampling)**

$$\text{Random sample of size n, } \{(x_{i1}, x_{i2}, \ldots, x_{ik}, y_i)\colon \text{ i} = 1, 2, \ldots, \text{n}\}$$

The data constitutes a random sample of n observations. The model must have randomly selected observations, meaning that the values of independent variables should not be correlated. Due to random sampling, observations and errors are independent for different i.

In our project we had to reduce the countries used to the ones we had available data for each regressor of the model. This Random Sampling assumption thus is partly violated. The model is, however, still valid as Random Sampling holds asymptotically.

- **MLR.3 (No perfect collinearity)**

There are no linear relationships between the regressor variables. In the sample, none of the independent variables is constant. None of the independent variables have a perfect correlation with a linear combination of the others, otherwise, the variable would be redundant. MLR.3 is not violated for non-linear relations.

This assumption holds in our model, just by observing the correlation matrix[2] we can conclude that there is no perfect linear correlation between any two variables, no correlation is equal to 1 or even close to it. Before defining the model used, we excluded one independent variable as it didn't have sufficient observations, nonetheless, MLR.3 still holds[3].

- **MLR.4 (Zero Conditional Mean)**

$$E(u|x_1, x_2, \ldots, x_k) = 0 \Longrightarrow E(u) = 0$$

The error u has an expected value of zero given any values of the independent variables. There is no correlation between the error term (which captures all the external factors and omitted variables) and our independent variables. In a perfect scenario there are no omitted variables excluded that correlate with our independent variables. If that doesn't happen, then we have an omitted variable bias, as we omitted an explanatory part. Meaning we could infer erroneous conclusions (a frequent error regarding regression analysis "correlation does not imply causation").

MLR.4 also implies that $cov(x_j, u) = 0$ (MLR.4'). When studying the distribution of the error term[4], we can see that its mean will be approximately zero.

It is virtually impossible to include all the variables that explain the dependent variable in our model. Thus, the existent omitted variables are included in the error term. The independent variables will be correlated with the error term, violating MLR.4. When constructing our model, we incorporated all the important variables we had available in order to minimize omitted variable bias. However, due to data constraints we were not able to include variables related to physical exercise, eating habits of the population among others which we expect to have an explanatory influence in the dependent variable (life expectancy). We also excluded one variable prior to defining the model used due to the insufficient number of observations, the respective variable was Government Expenditure per student.

As we have a large number of observations[5], the distribution of the estimators will approximate the true parameter values and we can conclude they are consistent, thus MLR.4' holds true.

[2] Appendix – Figure 1

[3] Appendix – Figures 1 and 8

[4] Appendix – Figure 6

[5] Given that our population are the world countries, our sample is considerably large

- **MLR.5 (Homoskedasticity)**

$$Var(u|x_i, \dots, x_k) = \sigma^2$$

The error u has the same variance given any value of the explanatory variables, u does not significantly change its variance when the values of the independent variables change. If the model is correctly defined, the variation of the dependent variable will predominantly be explained by the independent variables.

If MLR.5 is violated, the variance of u and the variance of the dependent variable will change with the different values of the explanatory variables. When omitting significant variables, an additional part of the variance is explained by factors extrinsic to the model. This indicates the variance of u may be heteroskedastic and that the coefficient estimates are more distant from the true population parameters.

A test of Homoskedasticity of error terms determines whether a regression model is capable of predicting the dependent variable consistently across values of independent variables or not.

To verify if Homoskedasticity holds in this model, we performed the Breusch-Pagan Test[6] for Heteroskedasticity, the White-Test[7] for Heteroskedasticity, and the Alternative White-Test[8] for Heteroskedasticity. The hypothesis stated in every test were:

$$H_0: Var(u|x_1, x_2, \dots, x_k) = \sigma^2 \qquad H_1: Not\ H_0$$

After completing the three tests for a 5% significance level, we observed that the null hypothesis is rejected in all three tests. Thus, we conclude the error term is heteroskedastic.

MLR.5 is then violated, so we should use Robust Standard Errors in order to conduct inference.

Under assumptions MLR.1 through MLR.5 the OLS estimators are BLUE (Best Linear Unbiased Estimator).

- **MLR.6 (Normality of the error term)**

The population distribution of the error u is independent of $x_1, x_2, \dots, x_k$. MLR.6 states that u follows a normal distribution with mean 0 and variance σ^2: $u \sim Normal\ (0, \sigma^2)$.

[6] Appendix – Figure 3
[7] Appendix – Figure 4
[8] Appendix – Figure 5

MLR.6 implies both MLR.4 and MLR.5. In fact, the independence assumption is stronger than MLR.4, and normality and independence imply MLR.5. Hence, all the previous results with respect to unbiasedness and variance of the estimators are still valid.

Assumptions MLR.1 through MLR.6 form the Classical Linear Model, which states that the OLS is not only BLUE, but it is the minimum variance unbiased estimator (no other unbiased estimator has a variance smaller than OLS).

Through the *Central Limit Theorem*, a sampling distribution follows approximately a normal distribution as *n* goes to infinity. In our model, the 78 observations are sufficient to apply this theorem and ensure a normal distribution of the error term. Through the Kernel Density graph[9] we can verify that normality of the error term holds in our model.

Results

Our model aims to explain the impact of the aforementioned variables on life expectancy across countries. To evaluate the impact of the regressors, we will interpret the coefficients obtained in the corrected model, as well as test these results and talk about the overall fit of the model.

As previously stated, our initial model had an additional variable which we decided to drop. However, dropping this variable did not come without a cost. As you can see from comparing both regressions[10], our coefficients changed. As we can observe, the coefficients that were most affected by this drop were Unemployment, Current Health Expenditure, and Smoking Prevalence, with two of these variables even changing the signal of their coefficient. This change in the regressors' coefficients comes because of the known problem of the omitted variable bias. Additionally, we could discuss the overall fit of the two models, by comparing the two R^2. Even though they can't give us an absolute decision as to which model is better given that the two models have a different number of observations, they could suggest which fitted the reality better. As we can see, the unrestricted model has a higher overall fit (84.13%)[11] than the restricted model (72.20%)[12]. Moreover, both models' variables are statistically jointly significant, which is assured by the fact that Prob > F = 0.000 (H0: B1=B2=...=Bk=0).

Focusing now on our final model, we can start by looking at the results obtained from performing test statistics to our regressors[13]. We can observe that, for a 5%

[9] Appendix – Figure 6
[10] Appendix – Figures 7 and 9
[11] Appendix – Figure 9
[12] Appendix – Figure 7
[13] Appendix – Figure 7

significance level, we were not able to reject the null hypothesis (H0: Bj=0) for *Labor Force with Advanced Education, Gross Domestic Savings, Unemployment, Alcohol Consumption, Smoking Prevalence, Urban Population,* and our dummy *Developed.* Even though many of our variables turned out to be statistically insignificant for a 5% significance level, that does not mean they shouldn't be included in the model. In fact, the 5% significance level is not a rigorous threshold that states a rule over significance. Economic theory tells us that all of these variables are, indeed, important to our model. As such, they should remain in the model.

As to study for the presence of heteroskedasticity, we conducted three different tests to our model. Firstly, to test for linear forms of heteroskedasticity, we performed the Breusch-Pagan test[14], which gave us an F-statistic(9,68) of 2.53, associated with a p-value of approximately 1.44%, which means that we should reject the null that the model was homoskedastic for a 5% significance level. In order to test for non-linear forms of heteroskedasticity, we carried out a White-test to our model[15], which resulted in an LM-Statistic of 74.31, associated with a p-value of 2.83%, meaning that for the same 5% significance level, we again reject the null (H0: homoskedasticity). The third test we performed was the Alternative White-Test[16], which netted us with a p-value of 0%, reinforcing once again the heteroskedasticity of our model. From the beginning, it was very likely that our model would be heteroskedastic, as we are dealing with real-life examples and it is expected that, for instance, the variance of life expectancy varies for, say, different levels of education. Given that the presence of a heteroskedastic error term makes inference invalid (MLR.5 fails), we used Robust Standard Errors to conduct inference, whose results we discussed above. Since our sample is somewhat big, using Robust Standard Errors does not invalidate our conclusions.

Furthermore, we can assess the impact each regressor has on life expectancy by looking at the table below:

Variable	Coefficient	Interpretation
AdvancedEduc	0.092497	On average, ceteris paribus, when the Labor Force with Advanced Education increases by 1%, Life expectancy increases by 0.092497 years.
ln(GDP per capita)	4.299745	On average, ceteris paribus, if GDP per capita increases by 1%, it is predicted that Life expectancy increases by 4.290745 years.
DomesticSavings	0.0250817	On average, ceteris paribus, when Domestic Savings increase by 1%, we expect that Life expectancy increases by 0.0250817 years.

[14] Appendix – Figure 3
[15] Appendix – Figure 4
[16] Appendix – Figure 5

Unemployment	-0.2166734	On average, ceteris paribus, if Unemployment increases by 1%, Life expectancy is going to drop by 0.2166734 years.
HealthExpenditure	0.7287658	On average, ceteris paribus, when Health Expenditure increases by 1%, Life expectancy increases by 0.7287658 years.
AlcoholConsumption	-0.3240001	On average, ceteris paribus, when Alcohol Consumption per capita increases by 1 liter, Life expectancy is expected to decrease by 0.3240001 years.
Smoking	0.1238926	On average, ceteris paribus, if Smoking Prevalence increases by 1%, the model predicts Life expectancy will increase by 0.1238926 years.
Urban	-0.0276591	On average, ceteris paribus, if the Urban Population of a country increases by 1%, Life expectancy will decrease by 0.0276591 years.
Developed	1.293814	On average, ceteris paribus, if a country is Developed instead of Developing, the Life expectancy will increase by 1.293814 years.
_cons	22.94739	If we assume all the independent variables take the value of zero, the overall Life expectancy for our countries would be of 22.94739 years. However, as none of our variables takes this value, this number has no intrinsic meaning.

As we can see, only two of our regressors ended up having a different sign from what we initially expected[17], those being the percentage of Urban Population and the Smoking Prevalence. We originally predicted that having more percentage of the population living in urban areas would lead to an increase in Life expectancy, given that access to better health care, higher income, and higher education is easier when living in these conditions. However, the effect of this variable ended up being the opposite. This coefficient can maybe be explained by the fact that urban areas tend to have a lot more pollution, for example, than rural areas, increasing the propensity of diseases.

It is also interesting to point out what happened with the coefficient for Smoking Prevalence, as at the beginning we all believed it should have a negative influence on Life expectancy. However, this shows exactly what the study from Freeh H.E. and Miller R. found out - smokers tend to live, on average, around the same time as non-smokers. One last important thing to note is that we can see that, on average, people living in developed countries tend to live longer than those living in developing countries, ceteris paribus.

[17] Empirical Model and Data – Table 1

Conclusion

The present paper tried to explore the effect that a more educated population had on a country's life expectancy at birth using a multilinear regression model. Nine variables were used as regressors to examine not only the significance but also to estimate the impact that education had in determining the life expectancy of these countries.

Life expectancy is a complex statistical measure because it is affected by a multitude of factors. However, it is important to study it and what exactly changes it since it is a key metric in assessing population health. We selected many variables, from where we would like to highlight the variable of AdvancedEducation, where we can see that it does not considerably affect our model, so the purpose of our project is partially reached. Our conclusion is that, in general, education does not affect life expectancy by much, unlike what was expected, as was said in the introduction, and based on our research.

However, we tried to understand not only if education is important for our life expectancy, but also other factors that can affect it. In the literature review, based on those papers, we conclude that the main drivers that influence life expectancy are directly related to environmental measures, lifestyle habits, and consumption (which is related to income, since more income means more consumption). Given that, and after all of the research made, we decided to allocate the measures of GDPpercapita and HealthConsumption as a consumption or income measure, the AlcoholConsumption and Smoking as measures of lifestyle habits and finally the Development as an environmental measure (since the more developed countries have, in general, higher levels of air pollution, aquifers pollution, etc.). Therefore, based on our model, we can corroborate the other research projects, because the factors that most contribute to longer life expectancy are GDP per capita, Development, and Health Expenditure. This conclusion makes sense as it is what is initially expected. Other factors that also positively influence life expectancy are Education and the Savings Level, but the influence of these variables is really small. Hence, their contribution is not very relevant.

On the other hand, we found that Unemployment and Consumption of Alcohol have a negative impact on life expectancy, as was expected, and also exposed in other papers.

The biggest surprise was that we realized that smokers on average, live the same as non-smokers, which is a huge surprise since it does not seem to make sense. The reasoning behind this could rally behind the fact that most smokers live in developed countries, and have access to better health care to treat for any possible diseases.

To conclude, the results of our model were satisfactory, as we had, at the beginning of this project, some difficulty in finding sufficient observations to compute the model, respecting all the assumptions. However, by evaluating the R^2, we can see that we have a positive result since, based on our model, we can guarantee that life expectancy could be explained 72.20% by the inputs of our model, which is a relatively high fit.

References

Shaw, J., Horrace, W., & Vogel, R. (2005). The Determinants of Life Expectancy: *An Analysis of the OECD Health Data*

Bayati M, Akbarian R, Kavosi Z. Determinants of life expectancy in eastern mediterranean region: *A health production function. Int J Health Policy Manag.*

Kabir, M. (2008). Determinants of Life Expectancy in Developing Countries. *The Journal of Developing Areas*

Cremer, H., Lozachmeur, J. M., & Pestieau, P. (2004). Social security, retirement age and optimal income taxation. *Journal of Public Economics*, *88*

Gradstein, M., & Kaganovich, M. (2004). Aging population and education finance. *Journal of Public Economics*, *88*(9–10)

Peltzman, S. (1987). Regulation and health: The case of mandatory prescriptions and an extension. *Managerial and Decision Economics*, *8*(1), 41–46.

Zare, H., Gaskin, D. J., & Anderson, G. (2015). Variations in life expectancy in Organization for Economic Co-operation and Development countries – 1985–2010.

Dhrymes, P. (2017). Introductory econometrics. *Introductory Econometrics*, 1–626.

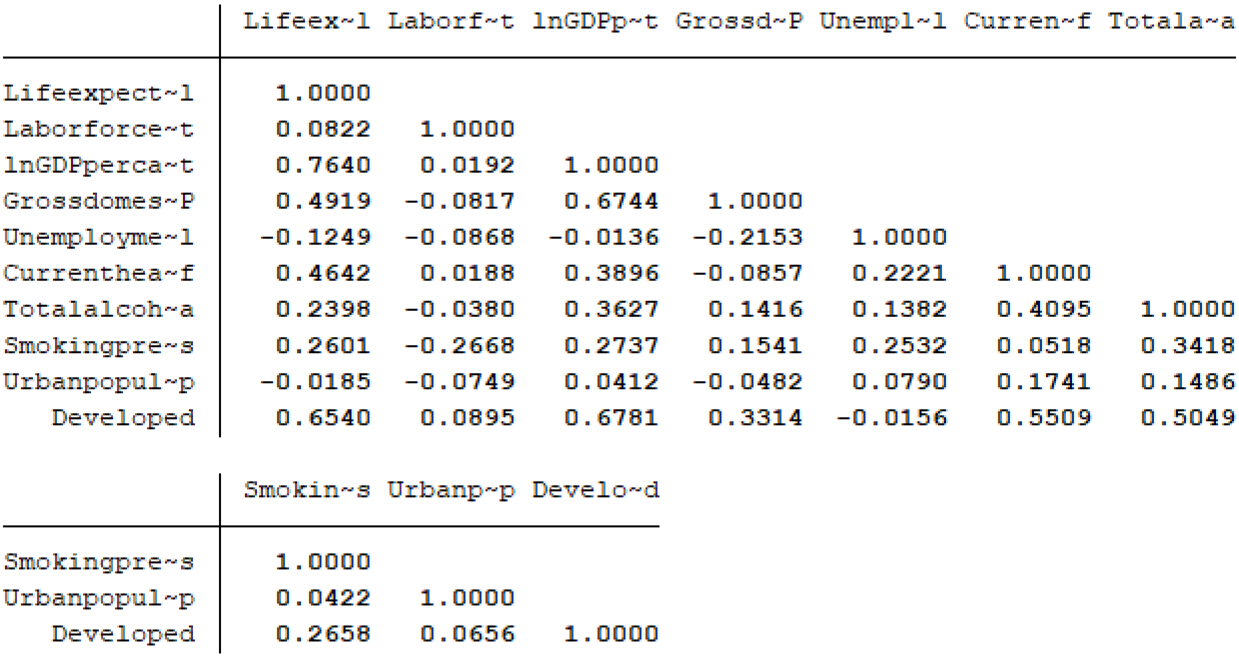
Sources

Link 1 - https://blogs.worldbank.org/opendata/what-does-life-expectancy-birth-really-mean

https://sites.google.com/site/nationalekonomigrunder/regression-analysis/assumptions

World Bank – Consulted in May 2020 (www.worldbank.org)

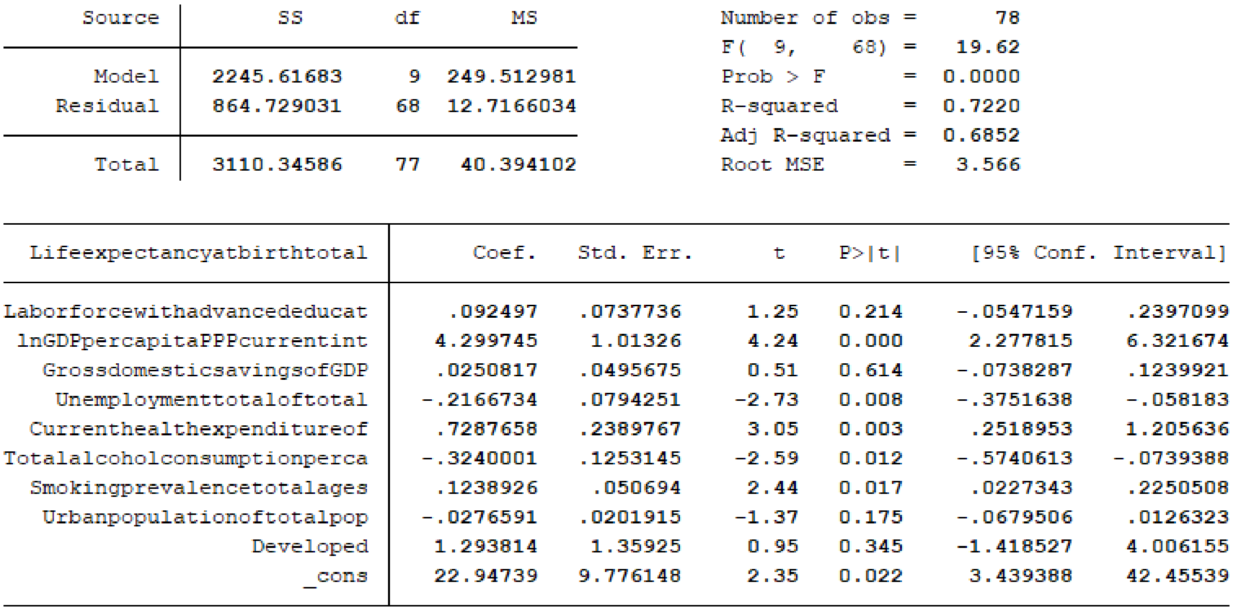
Appendix



```
             | Lifeex~l Laborf~t lnGDPp~t Grossd~P Unempl~l Curren~f Totala~a
-------------+---------------------------------------------------------------
Lifeexpect~l |   1.0000
Laborforce~t |   0.0822   1.0000
lnGDPperca~t |   0.7640   0.0192   1.0000
Grossdomes~P |   0.4919  -0.0817   0.6744   1.0000
Unemployme~l |  -0.1249  -0.0868  -0.0136  -0.2153   1.0000
Currenthea~f |   0.4642   0.0188   0.3896  -0.0857   0.2221   1.0000
Totalalcoh~a |   0.2398  -0.0380   0.3627   0.1416   0.1382   0.4095   1.0000
Smokingpre~s |   0.2601  -0.2668   0.2737   0.1541   0.2532   0.0518   0.3418
Urbanpopul~p |  -0.0185  -0.0749   0.0412  -0.0482   0.0790   0.1741   0.1486
   Developed |   0.6540   0.0895   0.6781   0.3314  -0.0156   0.5509   0.5049

             | Smokin~s Urbanp~p Develo~d
-------------+---------------------------
Smokingpre~s |   1.0000
Urbanpopul~p |   0.0422   1.0000
   Developed |   0.2658   0.0656   1.0000
```

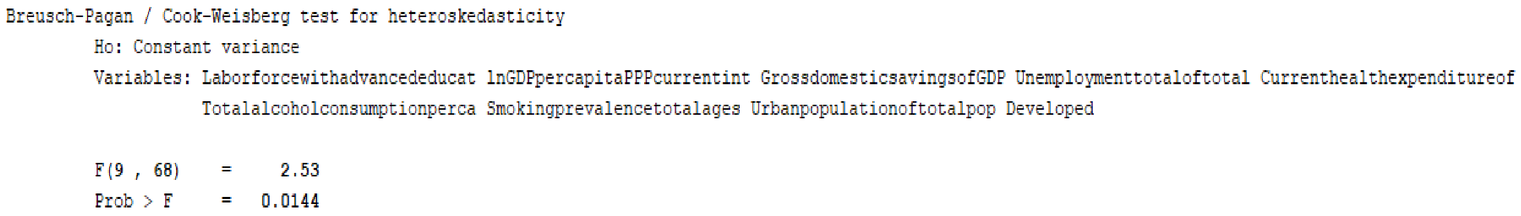
Figure 1- Restricted Model Correlation Matrix



```
      Source |       SS       df       MS              Number of obs =      78
-------------+------------------------------           F(  9,    68) =   19.62
       Model |  2245.61683     9  249.512981           Prob > F      =  0.0000
    Residual |  864.729031    68  12.7166034           R-squared     =  0.7220
-------------+------------------------------           Adj R-squared =  0.6852
       Total |  3110.34586    77   40.394102           Root MSE      =   3.566

------------------------------------------------------------------------------
Lifeexpectancyatbirthtotal |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
Laborforcewithadvancededucat |    .092497   .0737736     1.25   0.214    -.0547159    .2397099
lnGDPpercapitaPPPcurrentint |   4.299745    1.01326     4.24   0.000     2.277815    6.321674
GrossdomesticsavingsofGDP |   .0250817   .0495675     0.51   0.614    -.0738287    .1239921
Unemploymenttotaloftotal |  -.2166734   .0794251    -2.73   0.008    -.3751638    -.058183
Currenthealthexpenditureof |   .7287658   .2389767     3.05   0.003     .2518953    1.205636
Totalalcoholconsumptionperca |  -.3240001   .1253145    -2.59   0.012    -.5740613   -.0739388
Smokingprevalencetotalages |   .1238926    .050694     2.44   0.017     .0227343    .2250508
Urbanpopulationoftotalpop |  -.0276591   .0201915    -1.37   0.175    -.0679506    .0126323
   Developed |   1.293814    1.35925     0.95   0.345    -1.418527    4.006155
       _cons |   22.94739   9.776148     2.35   0.022     3.439388    42.45539
------------------------------------------------------------------------------
```

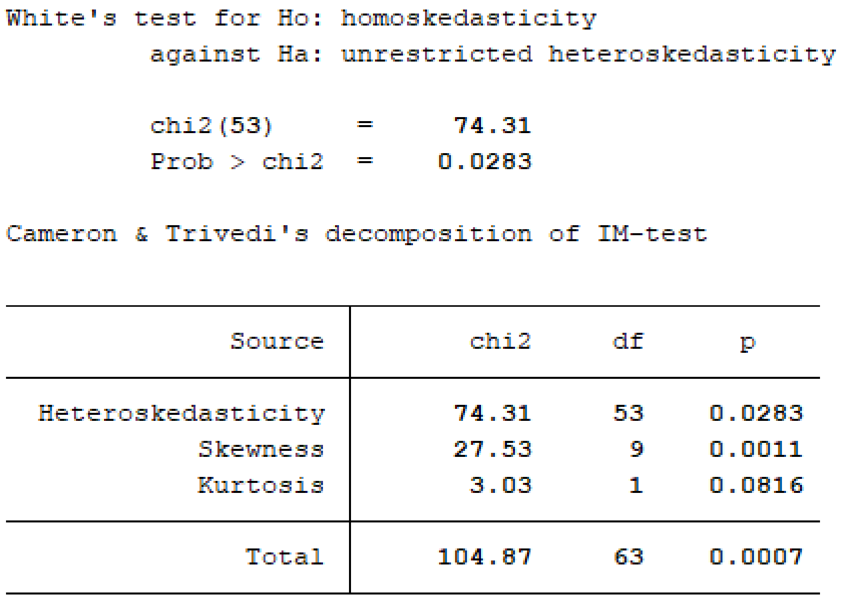
Figure 2 – Restricted Model Regression



```
Breusch-Pagan / Cook-Weisberg test for heteroskedasticity
         Ho: Constant variance
         Variables: Laborforcewithadvancededucat lnGDPpercapitaPPPcurrentint Grossdomesticsavingsofgdp Unemploymenttotaloftotal Currenthealthexpenditureof
                    Totalalcoholconsumptionperca Smokingprevalencetotalages Urbanpopulationoftotalpop Developed

         F(9 , 68)    =     2.53
         Prob > F     =   0.0144
```

Figure 3 – Restricted Model Breusch-Pagan Test for Heteroskedasticity



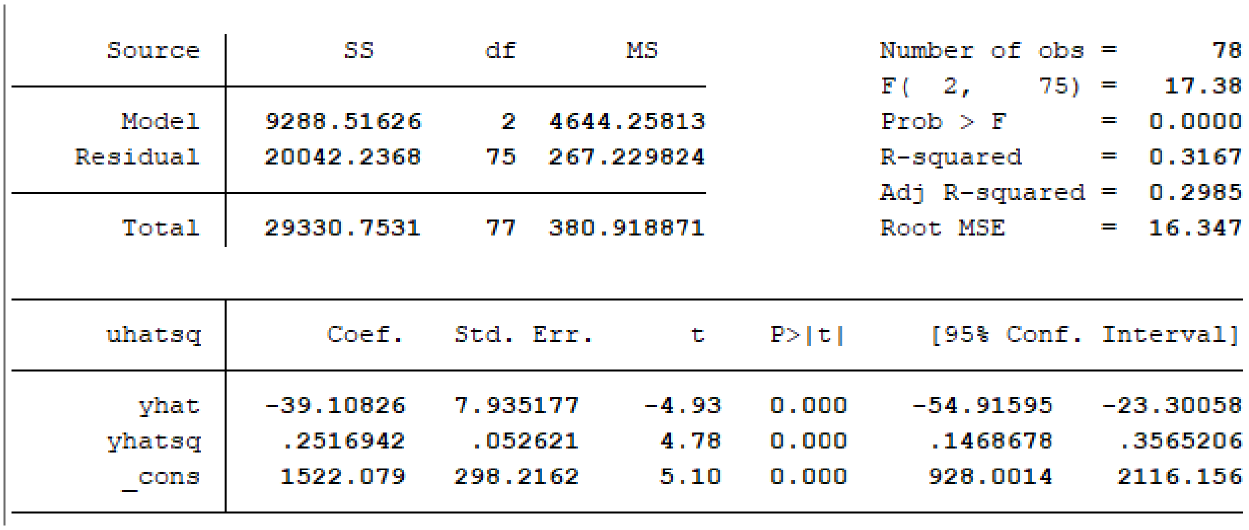
```
White's test for Ho: homoskedasticity
         against Ha: unrestricted heteroskedasticity

         chi2(53)      =     74.31
         Prob > chi2   =    0.0283

Cameron & Trivedi's decomposition of IM-test

---------------------------------------------------
              Source |       chi2     df      p
---------------------+-----------------------------
  Heteroskedasticity |      74.31     53    0.0283
            Skewness |      27.53      9    0.0011
            Kurtosis |       3.03      1    0.0816
---------------------+-----------------------------
               Total |     104.87     63    0.0007
---------------------------------------------------
```

Figure 4 – Restricted Model White-Test for Heteroskedasticity



```
      Source |       SS           df       MS      Number of obs   =        78
-------------+----------------------------------   F(  2,     75)  =     17.38
       Model |  9288.51626         2  4644.25813   Prob > F        =    0.0000
    Residual |  20042.2368        75  267.229824   R-squared       =    0.3167
-------------+----------------------------------   Adj R-squared   =    0.2985
       Total |  29330.7531        77  380.918871   Root MSE        =    16.347

------------------------------------------------------------------------------
      uhatsq |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        yhat |  -39.10826   7.935177    -4.93   0.000    -54.91595   -23.30058
      yhatsq |   .2516942    .052621     4.78   0.000     .1468678    .3565206
       _cons |   1522.079   298.2162     5.10   0.000     928.0014    2116.156
------------------------------------------------------------------------------
```

Figure 5 – Restricted Model Alternative White-Test for Heteroskedasticity

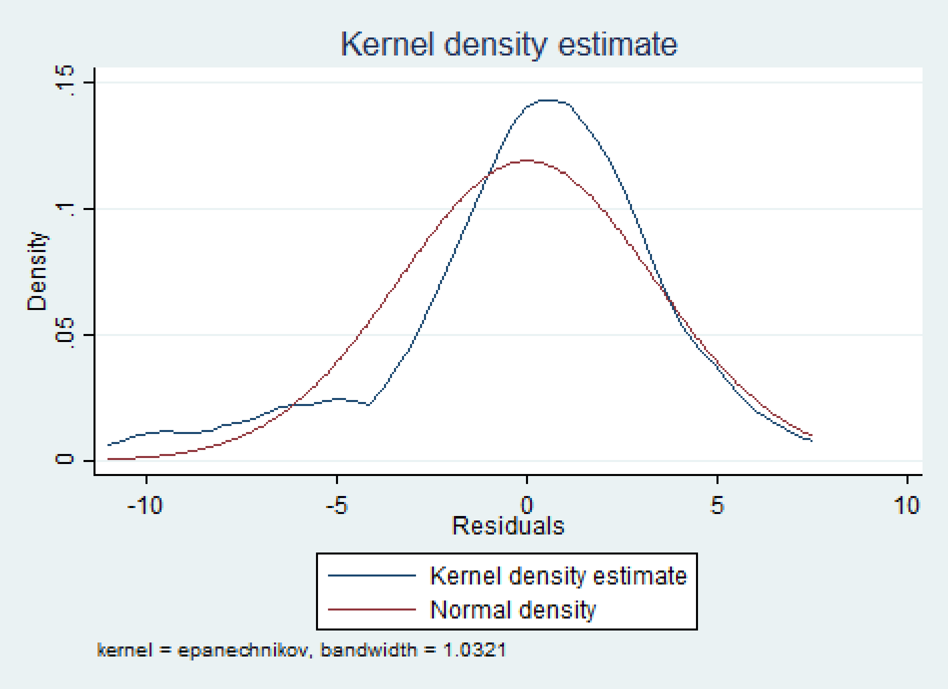
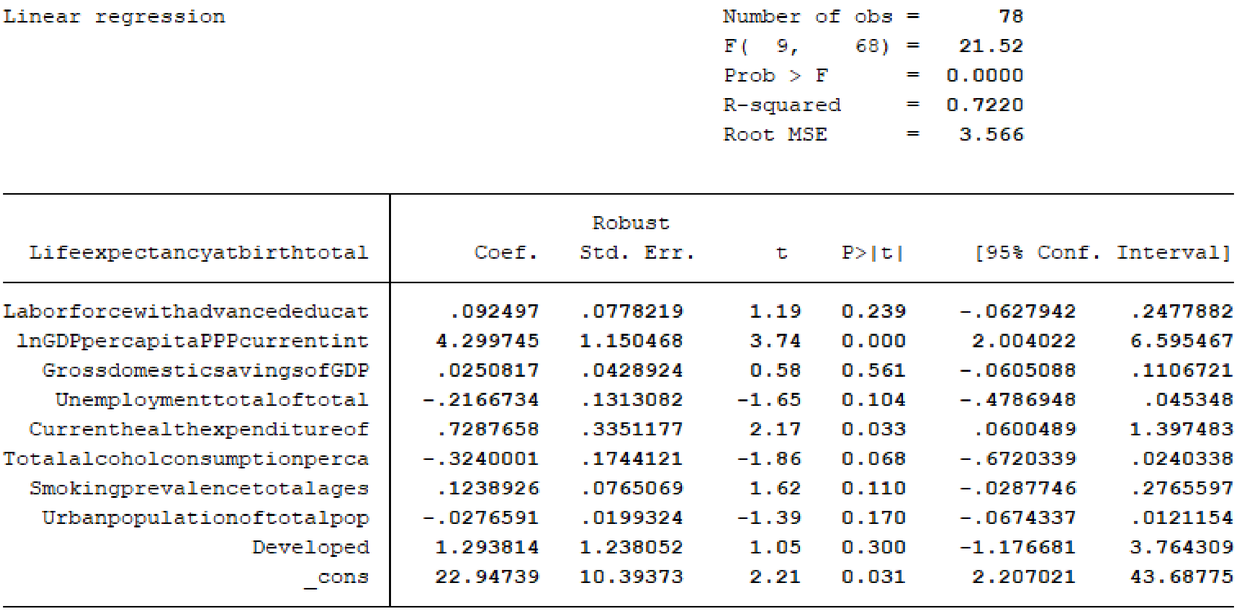


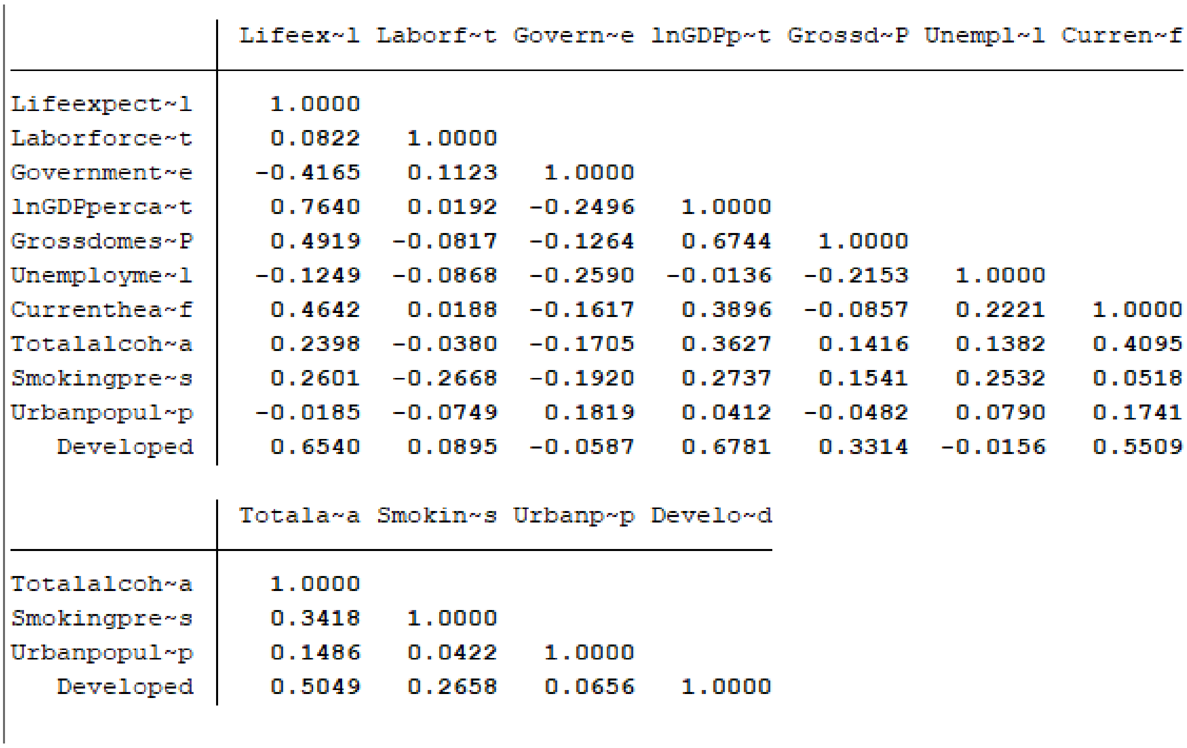
Figure 6 – Restricted Model Kernel Density Graph



```
Linear regression                                      Number of obs =      78
                                                       F(  9,     68) =   21.52
                                                       Prob > F      =  0.0000
                                                       R-squared     =  0.7220
                                                       Root MSE      =   3.566

------------------------------------------------------------------------------------------
                            |               Robust
 Lifeexpectancyatbirthtotal |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
----------------------------+-------------------------------------------------------------
Laborforcewithadvancededucat |    .092497   .0778219     1.19   0.239    -.0627942    .2477882
 lnGDPpercapitaPPPcurrentint |   4.299745   1.150468     3.74   0.000     2.004022    6.595467
   GrossdomesticsavingsofGDP |   .0250817   .0428924     0.58   0.561    -.0605088    .1106721
    Unemploymenttotaloftotal |  -.2166734   .1313082    -1.65   0.104    -.4786948     .045348
  Currenthealthexpenditureof |   .7287658   .3351177     2.17   0.033     .0600489    1.397483
Totalalcoholconsumptionperca |  -.3240001   .1744121    -1.86   0.068    -.6720339    .0240338
  Smokingprevalencetotalages |   .1238926   .0765069     1.62   0.110    -.0287746    .2765597
   Urbanpopulationoftotalpop |  -.0276591   .0199324    -1.39   0.170    -.0674337    .0121154
                   Developed |   1.293814   1.238052     1.05   0.300    -1.176681    3.764309
                       _cons |   22.94739   10.39373     2.21   0.031     2.207021    43.68775
------------------------------------------------------------------------------------------
```

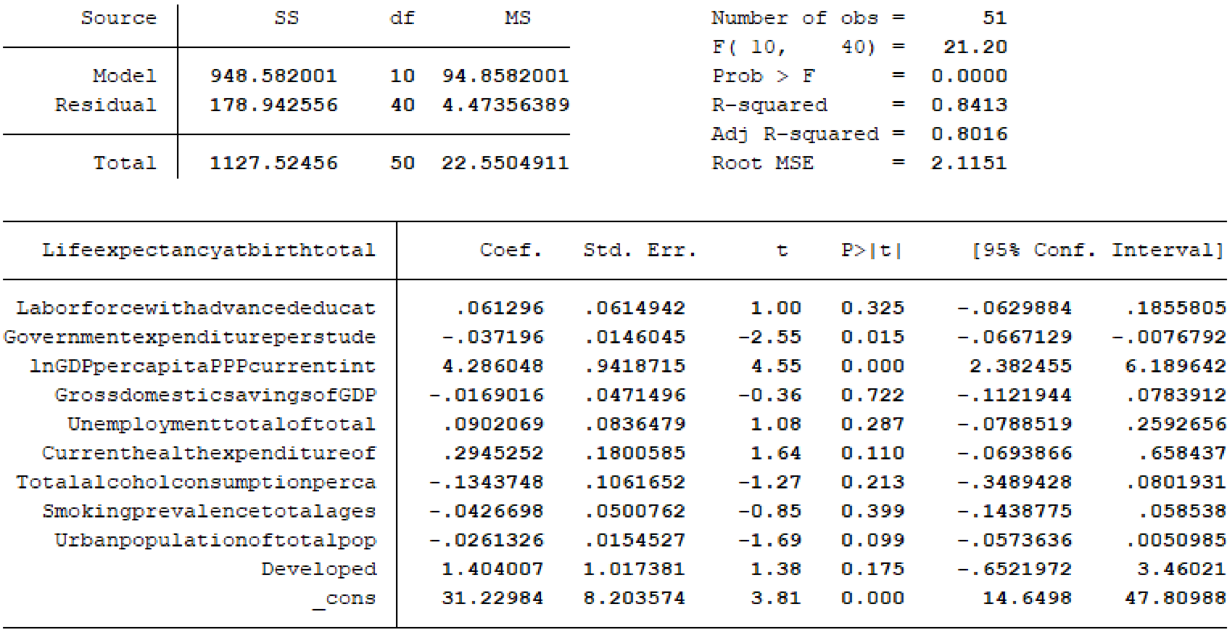
Figure 7 – Restricted Model Regression with Robust Standard Errors



```
             | Lifeex~l Laborf~t Govern~e lnGDPp~t Grossd~P Unempl~l Curren~f
-------------+---------------------------------------------------------------
Lifeexpect~l |   1.0000
Laborforce~t |   0.0822   1.0000
Government~e |  -0.4165   0.1123   1.0000
lnGDPperca~t |   0.7640   0.0192  -0.2496   1.0000
Grossdomes~P |   0.4919  -0.0817  -0.1264   0.6744   1.0000
Unemployme~l |  -0.1249  -0.0868  -0.2590  -0.0136  -0.2153   1.0000
Currenthea~f |   0.4642   0.0188  -0.1617   0.3896  -0.0857   0.2221   1.0000
Totalalcoh~a |   0.2398  -0.0380  -0.1705   0.3627   0.1416   0.1382   0.4095
Smokingpre~s |   0.2601  -0.2668  -0.1920   0.2737   0.1541   0.2532   0.0518
Urbanpopul~p |  -0.0185  -0.0749   0.1819   0.0412  -0.0482   0.0790   0.1741
   Developed |   0.6540   0.0895  -0.0587   0.6781   0.3314  -0.0156   0.5509

             | Totala~a Smokin~s Urbanp~p Develo~d
-------------+------------------------------------
Totalalcoh~a |   1.0000
Smokingpre~s |   0.3418   1.0000
Urbanpopul~p |   0.1486   0.0422   1.0000
   Developed |   0.5049   0.2658   0.0656   1.0000
```

Figure 8 – Unrestricted Model Correlation Matrix



```
      Source |       SS       df       MS              Number of obs =      51
-------------+------------------------------           F( 10,    40) =   21.20
       Model |  948.582001    10  94.8582001           Prob > F      =  0.0000
    Residual |  178.942556    40  4.47356389           R-squared     =  0.8413
-------------+------------------------------           Adj R-squared =  0.8016
       Total |  1127.52456    50  22.5504911           Root MSE      =  2.1151

------------------------------------------------------------------------------
Lifeexpectancyatbirthtotal |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
Laborforcewithadvancededucat |    .061296   .0614942     1.00   0.325    -.0629884    .1855805
Governmentexpenditureperstude |   -.037196   .0146045    -2.55   0.015    -.0667129   -.0076792
 lnGDPpercapitaPPPcurrentint |   4.286048   .9418715     4.55   0.000     2.382455    6.189642
   GrossdomesticsavingsofGDP |  -.0169016   .0471496    -0.36   0.722    -.1121944    .0783912
    Unemploymenttotaloftotal |   .0902069   .0836479     1.08   0.287    -.0788519    .2592656
  Currenthealthexpenditureof |   .2945252   .1800585     1.64   0.110    -.0693866     .658437
 Totalalcoholconsumptionperca |  -.1343748   .1061652    -1.27   0.213    -.3489428    .0801931
  Smokingprevalencetotalages |  -.0426698   .0500762    -0.85   0.399    -.1438775     .058538
   Urbanpopulationoftotalpop |  -.0261326   .0154527    -1.69   0.099    -.0573636    .0050985
                  Developed |   1.404007   1.017381     1.38   0.175    -.6521972     3.46021
                      _cons |   31.22984   8.203574     3.81   0.000      14.6498    47.80988
------------------------------------------------------------------------------
```

Figure 9 – Unrestricted Model Regression