**Table of contents**

[Introduction 3](#_bookmark0)

[Literature Review 5](#_bookmark1)

[Section 1. A first look at data 13](#_bookmark2)

[Section 2. Empirical analysis 16](#_bookmark3)

* 1. [Graphical analysis 16](#_bookmark4)
  2. [Descriptive statistics 17](#_bookmark5)
  3. [Econometric modeling 19](#_bookmark6)

[Section 3. Conclusions 28](#_bookmark10)

[Appendix A 32](#_bookmark11)

[Appendix B 35](#_bookmark12)

[Appendix C 36](#_bookmark13)

[Appendix D 38](#_bookmark14)

[Appendix E 39](#_bookmark15)

[Appendix F 40](#_bookmark16)

Appendix G 41

[References 42](#_bookmark17)

**Inception**

To start with, I’d like to say that my paper may not fit all the requirements as it uses not only econometric analysis, but machine learning techniques as well. I did receive some aid from Data scientists during my internship due to insufficient knowledge in ML stuff. Also, my calculations were done in Pyhton with next packages: Pandas and Numpy for data analysis, Sklearn for machine learning, Pyplot for visualization, Scipy for some mathematical calculations, Statsmodels and Linearmodels for econometric analysis. Of course, there were some Stata inclusions where I either couldn’t do or didn’t find the method for some obstacle.

Next, in our highly paced world actual and cutting edge can morph into obsolete not even within a year, but within months or weeks, that’s why taking some thorny subject in my case wasn’t the best move, from my angle of view, adding the fact that I tried to combine econometrics with programming and machine learning. The aim is to observe population from macro level and fathom which factors may have the biggest impact on it and which are less substantial. My topic is connected with social issues, i.e. socio-economic development of the countries under various factors like economic, healthcare, social facets. It’ll be an eternal theme as all of governments are concerned with population and economic development, but the crux of their commotion differs depending on the country and situation. So, my task here is to create hypothesis and infer regarded conclusions. The main object of my studies is population of countries and the subject of my study is factors that can influence somehow grow or decline of population.

Speaking about data, I used multiple sources as it’s nearly impossible to get all the data from one resource due to disparity in ways of calculation of various factors and as some data can be inaccessible because of country peculiarity. Main resources were OECD, The world bank, sites of government. Initially data was collected in Excel and transformed into csv to lessen the overall size.

**Literature review**

In modern world with it’s high paced development there are various issues that are emerging and can lead to deplorable upshots. One of them is population. It’s a multi-pronged topic that can be divided into million subtopics with particular gist in it’s core. Some researches focus on overpopulation and it’s effect on our planet with famine, shortage of space and so on, others tend to incline towards lack of people in some part of our world, like Europe or some countries in Asia, Japan, and how to resolve such an issue with proper migration flows and expats socialization. But the main principle of population is birth and death, all in all. I mean, the ratio of people born, and people deceased. And if we dive deeper into this, we’ll find out that there is cornucopia of prerequisites that can impact whether the population is increasing or nosediving, for example, healthcare system or affluence of people and many others.

The very thorough article that unveils the dependance between GDP, emission, energy use and its influence on population growth is:Carbon dioxide emissions, GDP, energy use, and population growth: a multivariate and causality analysis for Ghana, 1971–2013 [Asumadu-Sarkodie S., Owusu P. A. Carbon dioxide emissions, GDP, energy use, and population growth: a multivariate and causality analysis for Ghana, 1971–2013 //Environmental Science and Pollution Research. – 2016. – Т. 23. – №. 13. – С. 13508-13520.]

It is one of the most in-depth papers I’ve seen, and it comprises both econometric approach and my chosen problem. With a span of approximately 40 years with starting date in 1971 and finishing in 2016. Pretty long haul, but the results are really astonishing.

The scientists use time series data and such techniques as unit root, then unit roots, Johansen’s multivariate co-integration and the last but not the least is variance decomposition analysis using Cholesky’s technique. This research also implements another article that were written before or during current analysis as they wanted figure out whether their result. They decide to apply log on to provide more stable data variance. In result, VECM and ARDL that all the variables are co-integrated. What’s important, long-run eclecticists shows that roughly 1% increase is shown to increase CO2 emissions by roughly 2%. The ARDL bound test also presents the dependence between all the variables.

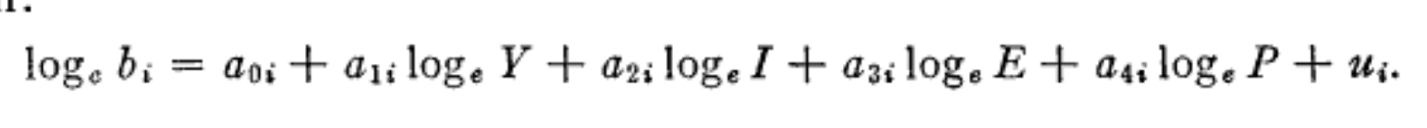
By and large, this research is potent and definitely can be claimed to be sane, but it doesn’t show off something groundbreaking or doesn’t peek into more subtle intricacies. There are a couple more papers I’ve analyzed to define my eventual topic.

Other articles that attracted me was An Econometric Analysis of Population Growth [Adelman I. An econometric analysis of population growth //The American Economic Review. – 1963. – Т. 53. – №. 3. – С. 314-339.] It was initially published in 1963 and can seem to be pretty old, but I took on purpose as wanted to look at how the analysis was done during those old ages and how it differentiates from current time, plus it was crucial for me to find out their result so as to consider how the sentiment of people and scientists specially have altered from those ages.

Their aim was to figure out how fertility and mortality patterns are affected by economic development as research considered that it can be somehow intertwined. Then they were deliberating which data set to choose and how to construct data, they stipulated that time series is pretty unreliable as age specific data for decease and birth cannot be easily obtained even for US, let alone some more obscure countries. And the choose was made in cross sectional direction.

They opted for some variables which were income, countries, years of study, years of schooling, population density, level of urbanization, fertility level, depths among infants.

Their next step was to calculate basis for least square regression:

****

# In their regression they opted for special weights to be implemented as there was statistical difference between variables gained from different resources. So, every weight was a composition of the following:

1. The reciprocals of the Whipple index
2. Measure of the so-called birth registration procedure
3. Plus, measure of the completeness of data was introduced

So, the data was corrected by those weights to produce better results.

# Their dependent variable was number of live births per 1000 females. Also, they pondered over the idea to halve the regression between developed countries and undeveloped.

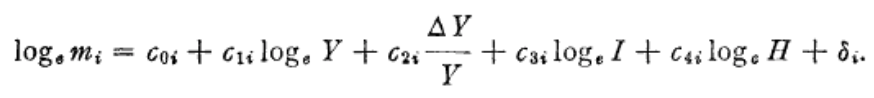
The results were the following: R^2 is significant between all variables and standard error is small, more precise info can be observed on the table:

Table

Description automatically generated

One of the most groundbreaking ideas was that desired family size and timing of childbearing times tends to react to economic behavior. The next groundbreaking thing they uncovered was reverse correlation between urbanization and number of children whilst apparent negative correlation between education and number of posterities was checked one more time.

The next step was to dive deeper into mortality factors. As it is considered to be mortality can be halved: due to standards of living and due to the state of healthcare system. In a nutshell, let’s look at variables that were used: income per capita, density of population, years of schooling. The regression was the following:



The same weights were implemented here as well as the issue with data appears to resemble the one above.

Upshots mainly are plain, like positive correlation between urbanization and decreasing amount of deaths or negative correlation between death and economic conditions, but scrutinizing the work precisely you’ll find one interesting, I’d rather say, cutting edge discovery: nosediving importance of social, medical and economic factors on inter-country death rate differentials at higher ages.

The pinnacle of their investigation was that influence of socioeconomic variables on the demographic ones of a society are smaller than the impact of population growth on the economic development.

The next article that was pretty fresh and also covered in some instance my theme is **Does Education Matter for Economic Growth [**Delgado M. S., Henderson D. J., Parmeter C. F. Does education matter for economic growth? //Oxford Bulletin of Economics and Statistics. – 2014. – Т. 76. – №. 3. – С. 334-359].

The sheer interest of researcher here was to fathom whether there is a correlation between schooling and economic growths. He admits that theme is pretty corny and may seem to be battered at a first glance, but he wishes to check it. The main difference of it’s research is to model which allows nonlinearities in the growth process for, practically, all of the variables in the model, hence results won’t be hugely dependent on the correct specification of the functional form of the model. Next, the two non-parametric procedures to peek out the importance of schooling in the non-parametric growth model.

Their regression model is non-parametric ones:

urn:x-wiley:03059049:media:obes12025:obes12025-math-0002

Where dependent variable is country economic growth.

They opted for dual non-parametric approach to check for robustness of schooling in their non-parametric growth model.   
Data which was used for research they took from open repositories. But one issue they met was various educational measures in various countries if dataset was small, that’s why they chose bigger ones.

In regard to estimation with parametric models, there were: OLS, Fixed Effects, First differences, between, first differences.

Some of the most interesting result are that many years of schooling isn’t a robust determinant of economic growth. Then researcher moves to explaining why aforewritten factor can have zero effect from mistakes in research to general conclusions.

And the last paper that also attracts me was: A spatial econometric analysis of the regional growth and volatility in Europe. [Falk M., Sinabell F. A spatial econometric analysis of the regional growth and volatility in Europe //Empirica. – 2009. – Т. 36. – №. 2. – С. 193-207.]

In the paper above the author focuses mainly on how fluctuations can impact growth in Europe. It isn’t closely related to my topic, but I opted for it as I had a desire to see what other research papers were published in regard to population.

Mainly, there are harsh debates about population swings and volatility as countries like Greece, Netherlands are high paced ones with immense volatility, but such as Italy with the same feature are just schlepping. Plus, there are multiple ways to analyze data, either via Panel datasets or time series ones.

Regional growth can be explained as the following from the paper:

Δ¯𝑌𝑖=𝛽0+𝛽1log(𝑌1995,𝑖)+𝛽2𝜎𝑖+𝛽3𝑠ℎ𝑎𝑔𝑟𝑖+𝛽4𝑠ℎ𝑖𝑛𝑑𝑖+𝛽5log(𝑑𝑒𝑛𝑠𝑖𝑡𝑦𝑖)+∑𝑗=1𝐽𝜃𝑗𝑐𝑜𝑖𝑗+𝑒𝑖,Δ¯Yi=β0+β1log(Y1995,i)+β2σi+β3shagri+β4shindi+β5log(densityi)+∑j=1Jθjcoij+ei

The next formula calculates whether there is a difference of volatility between countries.

Δ¯𝑌𝑖=𝛽0+𝛽1log(𝑌𝑖,95)+𝛽2𝜎𝑖+𝛽3𝑠ℎ𝑎𝑔𝑟𝑖+𝛽4𝑠ℎ𝑖𝑛𝑑𝑖+𝛽5log(𝑑𝑒𝑛𝑠𝑖𝑡𝑦𝑖)+∑𝑗=1𝐽𝜑𝑗𝑐𝑜𝑖𝑗𝜎𝑖+𝑒𝑖,Δ¯Yi=β0+β1log(Yi,95)+β2σi+β3shagri+β4shindi+β5log(densityi)+∑j=1Jφjcoijσi+ei,

There is one important caveat: OLS estimation of the parameters can be biased or even inefficient in the presence of the so-called spatial dependence.

Below will be growth equation that takes into consideration spatial error dependance:

Δ¯𝑌𝑖=𝛽0+𝛽1log(𝑌𝑖,95)+𝛽2𝜎𝑖+𝛽3𝑠ℎ𝑎𝑔𝑟𝑖+𝛽4𝑠ℎ𝑖𝑛𝑑𝑖+𝛽5log(𝑑𝑒𝑛𝑠𝑖𝑡𝑦𝑖)+∑𝑗=1𝐽𝜃𝑗𝑐𝑜𝑖𝑗+𝑢𝑖,Δ¯Yi=β0+β1log(Yi,95)+β2σi+β3shagri+β4shindi+β5log(densityi)+∑j=1Jθjcoij+ui,

Δ¯𝑌𝑖=𝛽0+𝛽1log(𝑌𝑖,95)+𝛽2𝜎𝑖+𝛽3𝑠ℎ𝑎𝑔𝑟𝑖+𝛽4𝑠ℎ𝑖𝑛𝑑𝑖+𝛽5log(𝑑𝑒𝑛𝑠𝑖𝑡𝑦𝑖)+∑𝑗=1𝐽𝜃𝑗𝑐𝑜𝑖𝑗+𝑢𝑖,Δ¯Yi=β0+β1log(Yi,95)+β2σi+β3shagri+β4shindi+β5log(densityi)+∑j=1Jθjcoij+ui,

The dataset was taken from Eurostat database REGIO, which have multiple variables, starting from GDP per capita, total population, purchasing power standards. In the flow of research, he found out that there is relationship between standard deviation of growth and per capita GDP growth. Plus, to detect the appropriate form of spatial autocorrelation they implemented Lagrange Multiplier (LM) tests and their robust versions.

So, there were 1084 European regions with a span from 1995 to 2004 he investigated positive correlation between growth and volatility. But he was very precautious in his conclusions as, firstly, large span of data should be tested, secondly, test of country data also must be reverified, thirdly, more macro analysis is to be done as situation can turn upside down due to another figures.

In the end, I would like to mention that most papers analyzed by me focuses somehow elaborates different topics and of course are all swell, bit they didn’t take the dataset which I’d like to implement: first of all, it’ll be much larger, secondly, I’d implement some other techniques, not general econometric tools, then I’d like to take some data which on the foresight may definitely should have positive correlation, but on the flipside, as were in some articles, may morph into something intriguing.

**Section 1. A first look at data**

**Data and methodology**

Now, we can switch to another point of the research: data. It was, on the one hand, not so arduous as data about population and everything with it is accessible, but to accumulate data from various countries and with a particular span is, on the other hand, a drudgery.

Also, I needed a data that would represent desired factors which I’d like to describe in my analysis. I’ve opted for Panel data as it makes easier for analysis made either in Python or Stata. Also, on the forefront it’s much easier to concoct such a dataset. I’ve used OECD, sites of government, the world bank with official statistics and it took a lot of time to compilate the data as there were various issues connected with either missing points or some odd figures that were pretty leery.

In case of instruments, I was distraught as I prefer Python with stellar Pandas, Numpy, Sklearn and others much more to Stata, but some methods are not easier or even impossible to do in Python, in this moments Stata came into play. My task was to analyze, first of, data in Pandas and then implement some basic ML techniques to create regression, teach the model, also using Matplotlib plot some data. And Stata was implemented in the moments when some methods were pretty arduous. Also, it’s much faster to do any kind of analysis in Python due to intrinsic performance of packages.

Span is average: 15 years but taking into consideration the fact that my data has lots of countries, which totaled in roughly 3000 points. Then, I’d like to have a sneak peek on my data: population, gdp, adult mortality, infant deaths, life expectancy, schooling, percentage expenditure, alcohol. Next one is a description of my variables.

There is one salient difference between analysis in Stata or Python. The former has variables which you input when creating models. But the latter doesn’t have such a thing, at least not the same as in Stata.

Name of the variable Description

|  |  |
| --- | --- |
| 1. Population | Total population |
| 1. Adult mortality | Adult Mortality Rates of both sexes (probability of dying between 15 and 60 years per 1000 population) |
| 1. Life expectancy | Life Expectancy in age |
| 1. Infant deaths | Number of Infant Deaths per 1000 population |
| 1. Alcohol 2. Percentage expenses | Alcohol, recorded per capita (15+) consumption (in litres of pure alcohol)  Expenditure on health as a percentage of Gross Domestic Product per capita(%) |
| 1. GDP | Gross domestic product |
| 1. Schooling | Years in school |
|  |  |

There was a meagre amount of data which was missing. Let’s calculate the number of values that were missing:

|  |  |
| --- | --- |
| Country | 0 |
| Year | 0 |
| Status | 0 |
| Life expectancy | 0 |
| Adult Mortality | 0 |
| infant deaths | 0 |
| Alcohol | 0 |
| percentage expenditure | 0 |
| Hepatitis B | 553 |
| Measles | 0 |
| BMI | 34 |
| under-five deaths | 0 |
| Polio | 19 |
| Total expenditure | 226 |
| Diphtheria | 19 |
| HIV/AIDS | 0 |
| GDP | 0 |
| Population | 0 |
| thinness 1-19 years | 34 |
| thinness 5-9 years | 34 |
| Income composition of resources | 167 |
| Schooling | 0 |

From data above we can observe that throughout the variables I’ve chosen the one that does have missing ones is Total expenditure, others have 0.

So as to solve the issue I implemented mean() on the column. There are other machine learning algorithms, but I prefer not to use them due to too much pressure on time. On the whole, data was pretty decent and significant amount of Nan variables wasn’t found.

**Graphical analysis**

It’s insanely important to analyze data via plots as it can help you to recheck yourself as it’s obvious that you can miss something while creating those models and without scatter plots it’s sometimes difficult to catch outliers. Population was chosen as dependent one and all the aforementioned are undependable variables.

I decided not to deal with multiple regressors, but to pick and choose the ones that may seem tricky. You can observe scatter plots in Appendix A chunk.

Inception of our plot sequence will be with Population vs percentage expenditure. It’s pretty homogenous as there are not so many outliers and they’re concentrated very dense, but

Look at Population vs GDP, this kind of plot is actually bereft of outliers, only some accidental ones, on the whole data is very dense. But if we switch to another plot, Population vs Infant deaths. And now data is sparser.

The last plot is Population versus Alcohol, here there are some dots that seem to be weird, but on the whole here data is also dense.

The whole gestalt of data was formed throughout great span and the data gained was also taken from zillions of resources which accidentally can foster some issues with my data. I can suppose that my models will have either autocorrelation or heteroscedasticity, but it’s just for the future.

**Descriptive statistics and correlational analysis of variables**

I’ve attached descriptive below and it allow us to cast some light on the very variables I’ve chosen. The standard deviation is much bigger than mean in most cases, but it’s not bad as it just tells us about the data being very sparse. For example, standard deviation of Life expectancy is roughly 70 and it’s mean is 9. If you take a gander at min and max values, there will be a substantial difference between these figures, so it verifies our theory.

Also, I forgot to mention that I implemented log on Population as it is in millions and infant deaths is in deaths per 1000.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Population | Adult mortality | Infant deaths | GDP | Schooling | Percentage expenditure | Life expectancy | Alcohol |
| count | 2938.0 | 2938.0 | 2938.0 | 2938.0 | 2938.0 | 2938.0 | 2938.0 | 2938.0 |
| mean | 1.275338 | 164.796448 | 30.303948 | 7483.158469 | 11.992793 | 738.251295 | 69.224932 | 4.602861 |
| std | 5.381546 | 124.080302 | 117.926501 | 13136.800417 | 3.264381 | 1987.914858 | 9.507640 | 3.916288 |
| min | 3.40 | 1.0 | 0 | 1.68135 | 0 | 0 | 36.30 | 0.01 |
| 25% | 4.189172 | 74.0 | 0 | 580.486996 | 10.3 | 4.685343 | 63.2 | 1.0925 |
| 50% | 3.675929 | 144.0 | 3.0 | 3116.56 | 12.1 | 64.912906 | 72.0 | 4.160 |
| 75% | 1.275338 | 227.0 | 22.0 | 7483.158469 | 14.1 | 441.5341 | 75.6 | 7.390 |
| max | 1.29385 | 723.00 | 1800.0 | 119172.741 | 20.70 | 19479.91161 | 89.0 | 17.870 |

Next move is to scrutinize correlation matrix which is in **Appendix B**. I’ve chosen for Pearson’s correlation coefficient and implemented some packages (actually, Seaborn and Pyplot) in Python to implement it. You can observe it below.

From here on let me give a crash course how to look at data: Population is our dependent variable and there are multiple independent ones. Population is the highest on the left and in the bottom the first one is also population. Also, the more negative correlation the darker the color and vice versa.

For example, the strongest positive correlation between Schooling and Life expectancy, maybe it can seem strange, but let it be. And the strongest negative correlation is between Adult mortality ad Life expectancy which is downright obvious.

Alcohol, adult mortality has negative correlation with population what is obvious, but from data you can see that there is strong positive correlation between infant deaths and population. It may seem bonkers at a first sight, but if you surf the net and read about this phenomenon, you’ll stumble upon the fact: the lower mortality fosters lower fertility and vice versa, it’s strange and unobvious from the beginning, but if you analyze it thoroughly, then you’ll be convinced that it’s right. And it’s our first investigation.

**Econometric modeling with hypothesis and main models**

Now the main meat of the work comes into play: we’ll define hypothesis.

1. *All the variables connected with deaths have negative correlation with dependable variable.*
2. *Economics factors like GDP have positive correlation with population.*
3. *How do healthcare factors impact population?*

There are particular tests that should be applied on the data so as check it’s verifiable or not. In python packages they’re named OLS, PanelOLS, RandomOLS => in Stata they’re Pooled OLS, Fixed Effect and Random Effect models.

But to create Fixed effect and Random effect models I need to convert data from general one to multi-index data frame, I coded it the necessary way. You can check it in an appropriate named column.

I wouldn’t go into many depths how these models work, but superficially note the main brass tacks: Pooled OLS has intercept and coefficient which are constant over years and countries. Fixed Effect has one peculiarity in regard to intercept: it varies across when particular year is taken, and slope doesn’t change at all. The last one, Random effect, slope is constant between all countries across years, but intercept has so-called ‘between entity error’ under it’s belt for every country.

Look at **Appendix C** for OLS and Random OLS models

I attempted to calculate Hausman effect in Python but faced some issues and opted for Stata.

Based on Stata analysis I should take Fixed Effect model: **Appendix D**

So, probability is significant, hence we accept Null Hypothesis and RE is worse than FE, which will proceed with. The table is on the next page.

PanelOLS Estimation Summary

================================================================================

Dep. Variable: Population R-squared: 0.3110

Estimator: PanelOLS R-squared (Between): -0.0773

No. Observations: 2938 R-squared (Within): 0.3110

Date: Tue, Jan 19 2021 R-squared (Overall): 0.3097

Time: 12:12:28 Log-likelihood -5.592e+04

Cov. Estimator: Clustered

F-statistic: 187.97

Entities: 16 P-value 0.0000

Avg Obs: 183.62 Distribution: F(7,2915)

Min Obs: 183.00

Max Obs: 193.00 F-statistic (robust): 12.151

P-value 0.0000

Time periods: 2938 Distribution: F(7,2915)

Avg Obs: 1.0000

Min Obs: 1.0000

Max Obs: 1.0000

Parameter Estimates

==========================================================================================

Parameter Std. Err. T-stat P-value Lower CI Upper CI

------------------------------------------------------------------------------------------

const -3.373e+07 1.579e+07 -2.1367 0.0327 -6.468e+07 -2.777e+06

Adult Mortality 4490.7 8111.9 0.5536 0.5799 -1.141e+04 2.04e+04

infant deaths 2.596e+05 7.344e+04 3.5342 0.0004 1.156e+05 4.036e+05

GDP 93.698 41.974 2.2323 0.0257 11.397 176.00

Schooling 4.414e+05 2.232e+05 1.9780 0.0480 3842.0 8.79e+05

percentage expenditure -1066.2 274.64 -3.8821 0.0001 -1604.7 -527.67

Life expectancy 4.737e+05 1.696e+05 2.7924 0.0053 1.411e+05 8.063e+05

Alcohol -2.605e+04 1.812e+05 -0.1438 0.8857 -3.813e+05 3.292e+05

==========================================================================================

F-test for Poolability: 0.9778

P-value: 0.4759

Distribution: F(15,2915)

Included effects: Entity

Now let’s do an analysis of the above table.

1. P-value signifies whether our variable is significant or not according to 3 levels of significance: 1, 5, 10 % level. Let’s just bundle all the data that is significant and not. Significant: infant deaths, GDP, Schooling, percentage expenditure, life expectancy. Insignificant: alcohol, Adult mortality.
2. F statistic should be compared with P value to define whether we reject Null hypothesis or not. In our case Prob < F statistic, hence we reject Null hypothesis, and our data is good.
3. Standard error tells us how well our sample data represents the overall pool of data. In my model, it’s on the acceptable level.
4. Poolability OLS tells us whether our FE model good or not. Also, a **poolability** test is an F test of the null hypothesis that all fixed effects are jointly 0; it is obtained by comparing fixed-effects estimates to those from pooled regression. In our model p-value < F-test for Poolability => we reject null hypothesis.

First of all, let’s perform multicollinearity check, the so-called rule of thumb is that if VIF < 5 then it’s OK, from 5 to 10 may be problematic and above 10 is critical.

|  |  |
| --- | --- |
| Variables | VIF |
| Population | 1.524480 |
| Adult Mortality | 2.370267 |
| Infant deaths | 1.569348 |
| GDP | 6.595013 |
| Schooling | 33.744588 |
| Percentage expenditure | 5.497802 |
| Life Expectancy | 31.732579 |
| Alcohol | 3.259980 |

Here we have Life expectancy, Schooling that are pretty immense.

That’s why let’s remove these two variables. And due to the reason that those variables are not control ones (in the papers I was observing they were not the pillar ones).

Heteroscedasticity vs homoscedasticity

There are multiple ways also to remove heteroscedasticity, ones of which are Huber-white, Breusch-Pagan and BoxCox. BoxCox was implemented at first by me but a little bit in a machine learning fashion:

1) I transformed my data from DataFrame into Numpy array, i.e. I transformed data with columns into stark array with only figures.

2) Then I picked data that is more than 0 as ordinary boxcox demands data to be strictly positive if we don’t use modified one which is BoxCox1 that adds 1 to the figures that are 0.

3)halved data into train and test one.

4)implemented BoxCox which uses for scaling and normalizing data. And plotted test and train data to compare how BoxCox dealt with those chunks of data. Practically, we have a pretty good distribution that resembles The Bell Curve, i.e. Normal distribution. Look at **Appendix E** for graphs connected with BoxCox.

But, unfortunately, I couldn’t transform back from numeric array to data frame with array => I would put it, but don’t use in the following.

Now let’s do graphical check for heteroscedasticity, you can observe it below.

Steps:

1. Calculate weights/parameters via fit() function of the Fixed Effect model in the moment, when I was calculating my regression model:

fe\_res = fe\_mod.fit(cov\_type='clustered', cluster\_entity=True)

1. Then I’ll make predictions on that calculated stuff above and attach them to a variable:

fe\_res\_resids = fe\_res.predict().fitted\_values

1. Next logical step is to extract residuals and attach them also to a variable

fe\_res\_PAnelOLS = fe\_res.resids

1. Culmination is plotting of the aforewritten calculations which is done by many actions, but the gist is in the following line:

ax.scatter(fe\_res\_resids, fe\_res\_PAnelOLS, color = 'red')

Here I put residuals and predicted weights/parameters.

In my plot I check predicted values on my x axis vs residuals on y axis. Look at **Appendix F** for this plot.

There are few values that are spread out => very low degree of heteroskedasticity. I’ve tried to implement Breusch-Pagan-Test and White-Test but couldn’t get right outcome and receive errors. So, let’s proceed with current model.

Next test is Durbin Watson one which allows to check for autocorrelation. I used residuals and my dataset to check for it. According to my data the figure is 1.3.

If we peruse Durbin Watson test theory: from 0 to 2 is positive autocorrelation, straight 2 is zero autocorrelation, from 2 to 4 is negative autocorrelation.

* Searching for viable ways to tweak such an issue I stumbled **upon cochrane-orcutt** and **hildreth-lu methods**, but they’re pretty contrive, and I decided to get by with the current model + in my model, if you look at **Cov. Estimator: Clustered**

Line I PanelOLS model, you’ll understand that my data is already clustered.

But after analyzing data we can see that Population has immensely great value and compared to everything else I should do something to tweak it as if we see at Parameter column in PanelOLS model, you can see that, for example, 1% improvement in % expenditure will kill 4 billion people. It’s bonkers, let’s implement logarithm on Population and recalculate our model.

log\_pop = np.log(df['Population'])

Enables me to create variable that will be with logarithmic tweak.

PanelOLS Estimation Summary

================================================================================

Dep. Variable: Population R-squared: 0.0726

Estimator: PanelOLS R-squared (Between): 0.1180

No. Observations: 2938 R-squared (Within): 0.0726

Date: Wed, Jan 20 2021 R-squared (Overall): 0.0728

Time: 01:13:36 Log-likelihood -6814.3

Cov. Estimator: Clustered

F-statistic: 32.594

Entities: 16 P-value 0.0000

Avg Obs: 183.62 Distribution: F(7,2915)

Min Obs: 183.00

Max Obs: 193.00 F-statistic (robust): 41.930

P-value 0.0000

Time periods: 2938 Distribution: F(7,2915)

Avg Obs: 1.0000

Min Obs: 1.0000

Max Obs: 1.0000

Parameter Estimates

==========================================================================================

Parameter Std. Err. T-stat P-value Lower CI Upper CI

------------------------------------------------------------------------------------------

const 14.274 0.5370 26.579 0.0000 13.221 15.327

Adult Mortality 0.0012 0.0003 3.6080 0.0003 0.0006 0.0019

infant deaths 0.0042 0.0005 9.0624 0.0000 0.0033 0.0051

GDP 8.167e-05 1.316e-05 6.2069 0.0000 5.587e-05 0.0001

Schooling -0.0124 0.0141 -0.8769 0.3806 -0.0401 0.0153

percentage expenditure -0.0005 8.205e-05 -5.7278 0.0000 -0.0006 -0.0003

Life expectancy -0.0055 0.0079 -0.6961 0.4864 -0.0209 0.0100

Alcohol 0.0218 0.0123 1.7686 0.0771 -0.0024 0.0459

==========================================================================================

F-test for Poolability: 0.7688

P-value: 0.7138

Distribution: F(15,2915)

Included effects: Entity

Nice! Here we have much better p-value figures and Parameter can be interpreted which as well I’ll do near the end. All the other stuff stays the same and I won’t repeat myself here.

Below I’ll analyze VIF (aka multicollinearity check)

|  |  |
| --- | --- |
| Variables | VIF |
| Population | 29.3 |
| Adult Mortality | 3.05 |
| Infant deaths | 1.57 |
| GDP | 6.72 |
| Schooling | 33.87 |
| Percentage expenditure | 5.64 |
| Life Expectancy | 56.86 |
| Alcohol | 3.26 |

At start, we’ll jettison population’s VIF here as it’s normal that such immense figure appeared. We implemented logarithm + FE model gives us room for doing so. On the whole data stays the same as before => I remove Life expectancy and Schooling as before.

Next move is check for heteroscedasticity. The steps are all the same as with the first model. In my plot I check predicted values on my x axis vs residuals on y axis. You can observe the new graph in **Appendix G.** We have a little bit worse situation with it here, but not critically.

And the last but not the least is Durbin Watson. Now we have 0.88 => as before it’s positive autocorrelation which can be a little bit stronger than in the previous case as I put log on population and our data has become closer, i.e. Population’s millions were adjusted and now our dataset has become more representable. As before, **cochrane-orcutt** and **hildreth-lu methods** are impossible for me to implement as I don’t have neither understanding of their under-hood operations nor could understand their formula and clustering is already implemented in my model which I also mentioned earlier.

**Conclusion**

In closing I can make various stipulations, but let’s start with my model: we’ll proceed with the latter one as it’s overall score is much better and if, at least, we apply our logic, population cannot be compared with our dataset without right adjustment what was demonstrated in parameter in regard to p-value, like: 1% of Adult mortality could foster population growth by 4 billion, it’s nonsense.

Second model is pretty sturdy one and was able to show us results that I’ll cast light on further.

Firstly, schooling and Life expectancy doesn’t have a positive impact on my model due to multicollinearity issue and I’ve excluded them. If we turn some logic then we’ll understand that schooling may have some implicit impact on the model, but on the whole it doesn’t directly affect it. Life expectancy should influence by sane checker, but maybe it has such an issue due to disparity in data as it was collected from various resources and the way of calculating it differs => the very life expectancy also from country to country may be not statistically equal. Other values don’t have above problem.

Alcohol and adult mortality also shouldn’t be used in our first model as they are not significant at all, but in second one they’re stellar in regard to probability. We ended up with 5 regressors that can be used: GDP, infant deaths, percentage expenditure, adult mortality, alcohol. Now let’s explain them.

**GDP**: the better GDP the better can be, not in 100% cases of course, the main factor that proves country harps on improvement it’s economic situation, hence people may have better lifestyle. Looking at it’s interpretation: 1% change in it will spawn spike in 8 million people which can be a lot of, but making permission for dataset with various countries it’s acceptable.

**Percentage expenditure**. It is calculated in the following fashion: Expenditure on health as a percentage of Gross Domestic Product per capita(%). It hinges on the GDP: the better GDP the better this figure. And this figure can tell us about the healthcare system as well, I mean, even if GDP is great, but people die a lot, there may be question to how well healthcare state workers allocate money as the calculation itself of the factor doesn’t take into account how really money are used. And it can lead us to some even nefarious theories about bribery and everything of that ilk. Thus, in our model it’s a little bit negative, but practically 1% change has zero to no effect on population => it can partially prove my assumptions above about additional things connected with this variable, hence healthcare system, and bearing in mind that in my dataset Developing countries occupy 2500 datasets from 2900 it may seem true. I mean, ideally it should have positive impact on population. I think here questions will be to state workers and how they use coffers.

**infant deaths.** It is the most controversial one at a first glance because if we just ponder over it in layman fashion: the fewer deaths the better population => it will lead to wrong assumption and results. If you read articles on population and everything about how fertility is regulated or can be influenced, you’ll be astonished that the lower children deaths the lower fertility level. I won’t dig too much into this topic, but overall, it is the debacle that arises from the regression analysis. Here 1% alteration of this variable lead to 4200 people increase.

**Adult mortality:** adult mortality may coincide partially with the above variable in explanations. Here 1% change leads to 1200 growth in population: partially described by fertility theory and partially can spur government to give people additional benefits for giving birth to children which is also obvious => more people -> more taxation.

**Alcohol:** is the only variable that fails to be reasonably described: as if 1% change led to 21800 increase in population. I’m not a doctor or anything like that, but such a result fosters room for though.

Now, let me repeat my hypothesis to make reasonable inferences:

1. *All the variables connected with deaths have negative correlation with dependable variable.*

This assumption was debunked as infant deaths lead to growth of population and adult deaths as well.

1. *Economics factors like GDP have positive correlation with population.*

It was proved and explained clearly above.

1. *How do healthcare factors impact population?*

It is a more complicated question. Percentage expenditure ones more time is calculated in the following fashion: Expenditure on health as a percentage of Gross Domestic Product per capita(%). And if we look at correlation matrix, you’ll see that these variables have strong positive correlation: 0.89. It’s obvious as this variable is a product of GDP. Using sanity checker percentage expenditure should have positive impact on Population as it is a healthcare system issue. But it has a correlation of -0,025 which is strange at first. But looking at our data and remunerating about countries that are in the dataset, you will see there are myriad countries with inscription ‘Developing’ whilst ‘Developed’ is much smaller. Actually, 2426 vs 512. It leads us to one salient conclusion: this factor may impact the population but, in our dataset, the prevalent number of countries are Developing ones and it shatters the crux of the variable. Plus, the ways of calculating this variable may inherently be different or some adulteration also can be implemented to ameliorate the position of the country.

In result I could only add that there are no two similar analysis as datasets, methods and even other odds and ends of another kind have some kind of sporadic behavior and that’s why when doing some serious and really life-depending research you’d better recheck everything. This paper was a nice sneak peek into world of econometrics and machine learning and gave opportunity to ruminate over population issues and their impact on other aspects of life such as economy, society and other.

**Appendix A**

**Chart, scatter chart

Description automatically generated**

**Chart, scatter chart

Description automatically generated**

**Chart, scatter chart

Description automatically generated**

**Chart, scatter chart

Description automatically generated**

**Appendix B**

Chart

Description automatically generated

**Appendix C**

**OLS model**

OLS Regression Results

==============================================================================

Dep. Variable: Population R-squared: 0.310

Model: OLS Adj. R-squared: 0.308

Method: Least Squares F-statistic: 187.9

Date: Tue, 19 Jan 2021 Prob (F-statistic): 1.72e-230

Time: 23:02:55 Log-Likelihood: -55923.

No. Observations: 2938 AIC: 1.119e+05

Df Residuals: 2930 BIC: 1.119e+05

Df Model: 7

Covariance Type: nonrobust

==========================================================================================

coef std err t P>|t| [0.025 0.975]

------------------------------------------------------------------------------------------

const -3.459e+07 1.04e+07 -3.322 0.001 -5.5e+07 -1.42e+07

Adult Mortality 4867.3097 9410.379 0.517 0.605 -1.36e+04 2.33e+04

infant deaths 2.596e+05 7186.007 36.121 0.000 2.45e+05 2.74e+05

GDP 84.4335 140.783 0.600 0.549 -191.611 360.478

Schooling 5.442e+05 3.91e+05 1.393 0.164 -2.22e+05 1.31e+06

percentage expenditure -904.7247 913.173 -0.991 0.322 -2695.250 885.801

Life expectancy 4.724e+05 1.61e+05 2.939 0.003 1.57e+05 7.88e+05

Alcohol -1.129e+05 2.49e+05 -0.454 0.650 -6.01e+05 3.75e+05

==============================================================================

Omnibus: 4744.070 Durbin-Watson: 1.335

Prob(Omnibus): 0.000 Jarque-Bera (JB): 6401614.707

Skew: 10.065 Prob(JB): 0.00

Kurtosis: 230.790 Cond. No. 1.92e+05

==============================================================================

Notes:

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

[2] The condition number is large, 1.92e+05. This might indicate that there are

strong multicollinearity or other numerical problems.

**Random Effect model**

RandomEffects Estimation Summary

================================================================================

Dep. Variable: Population R-squared: 0.3440

Estimator: RandomEffects R-squared (Between): 0.9395

No. Observations: 2938 R-squared (Within): 0.3085

Date: Tue, Jan 19 2021 R-squared (Overall): 0.3440

Time: 23:02:56 Log-likelihood -5.593e+04

Cov. Estimator: Clustered

F-statistic: 219.61

Entities: 16 P-value 0.0000

Avg Obs: 183.62 Distribution: F(7,2931)

Min Obs: 183.00

Max Obs: 193.00 F-statistic (robust): 111.56

P-value 0.0000

Time periods: 2938 Distribution: F(7,2931)

Avg Obs: 1.0000

Min Obs: 1.0000

Max Obs: 1.0000

Parameter Estimates

==========================================================================================

Parameter Std. Err. T-stat P-value Lower CI Upper CI

------------------------------------------------------------------------------------------

Adult Mortality -1.878e+04 7844.8 -2.3943 0.0167 -3.416e+04 -3401.0

infant deaths 2.554e+05 7.178e+04 3.5577 0.0004 1.146e+05 3.961e+05

GDP 124.10 45.930 2.7018 0.0069 34.036 214.15

Schooling 7.525e+05 2.812e+05 2.6760 0.0075 2.011e+05 1.304e+06

percentage expenditure -873.71 279.35 -3.1276 0.0018 -1421.5 -325.97

Life expectancy -1.368e+04 5.346e+04 -0.2558 0.7981 -1.185e+05 9.116e+04

Alcohol -7277.8 1.352e+05 -0.0538 0.9571 -2.725e+05 2.579e+05

==========================================================================================

**Appendix D**

**hausman FE RE, sigmamore**

|  |  |  |  |
| --- | --- | --- | --- |
| **Coefficients** | | | |
|  | **(b) (B)**  **FE RE** | **(b-B)**  **Difference** | **sqrt(diag(V\_b - V\_B))**  **S.E.** |
| **AdultMorta~y** | **-8401.763 4867.31** | **-13269.07** | **4975.944** |
| **infantdeaths** | **6175.81 259562.5** | **-253386.7** | **34752** |
| **GDP** | **-8.922123 84.43348** | **-93.35561** | **114.8237** |
| **Schooling** | **498463.4 544248.6** | **-45785.2** | **666228.3** |
| **percentage~e** | **247.8991 -904.7247** | **1152.624** | **730.1932** |
| **Lifeexpect~y** | **51381.82 472407.6** | **-421025.7** | **317280.8** |
| **Alcohol** | **-215409 -112928.4** | **-102480.6** | **486453.8** |

**b = consistent under Ho and Ha; obtained from xtreg**

**B = inconsistent under Ha, efficient under Ho; obtained from xtreg**

**Test: Ho: difference in coefficients not systematic**

**chi2(7) = (b-B)'[(V\_b-V\_B)^(-1)](b-B) = 66.82**

**Prob>chi2 = 0.0000**

**Appendix E**

**Chart, histogram

Description automatically generated**

**Appendix F**

**Chart, scatter chart

Description automatically generated**

**Appendix G**

**Chart, scatter chart

Description automatically generated**

**References**

1. **Asumadu-Sarkodie S., Owusu P. A. Carbon dioxide emissions, GDP, energy use, and population growth: a multivariate and causality analysis for Ghana, 1971–2013 //Environmental Science and Pollution Research. – 2016. – Т. 23. – №. 13. – С. 13508-13520**
2. **Adelman I. An econometric analysis of population growth //The American Economic Review. – 1963. – Т. 53. – №. 3. – С. 314-339**
3. **Delgado M. S., Henderson D. J., Parmeter C. F. Does education matter for economic growth? //Oxford Bulletin of Economics and Statistics. – 2014. – Т. 76. – №. 3. – С. 334-359**
4. **Falk M., Sinabell F. A spatial econometric analysis of the regional growth and volatility in Europe //Empirica. – 2009. – Т. 36. – №. 2. – С. 193-207**
5. **Lane J. E., McKay D. H., Newton K. Political Data Handbook. OECD Countries. – Oxford University Press, 1992.**
6. **Faubert V. School evaluation: Current practices in OECD countries and a literature review. – 2009.**
7. **Linderfalk U., Hilling M. The Use of OECD Commentaries as Interpretative Aids-The Static/Ambulatory–Approaches Debate Considered from the Perspective of International Law //Nordic Tax Journal. – 2015. – Т. 2015. – №. 1. – С. 34-59.**
8. **Hämäläinen P., Gissler M. OECD: n terveydenhuollon laatu-indikaattorit Suomessa 2011-2012.**
9. **Jiménez-Rodríguez\* R., Sánchez M. Oil price shocks and real GDP growth: empirical evidence for some OECD countries //Applied economics. – 2005. – Т. 37. – №. 2. – С. 201-228.**
10. **Swift R. The relationship between health and GDP in OECD countries in the very long run //Health economics. – 2011. – Т. 20. – №. 3. – С. 306-322.**
11. **Auerbach A. J. et al. The dynamics of an aging population: The case of four OECD countries. – National Bureau of Economic Research, 1989. – №. w2797.**
12. **World Health Organization et al. Infant mortality. – 2020.**
13. **Thakrar A. P. et al. Child mortality in the US and 19 OECD comparator nations: a 50-year time-trend analysis //Health Affairs. – 2018. – Т. 37. – №. 1. – С. 140-149.**
14. **Reinhardt U. E., Hussey P. S., Anderson G. F. Cross-national comparisons of health systems using OECD data, 1999 //Health affairs. – 2002. – Т. 21. – №. 3. – С. 169-181.**
15. **Hansen M. B. The making of modern Japan. – Harvard University Press, 2002.**
16. **Feurer M. et al. Auto-sklearn: efficient and robust automated machine learning //Automated Machine Learning. – Springer, Cham, 2019. – С. 113-134.**
17. **McKinney W. et al. pandas: a foundational Python library for data analysis and statistics //Python for High Performance and Scientific Computing. – 2011. – Т. 14. – №. 9.**
18. **Seabold S., Perktold J. Statsmodels: Econometric and statistical modeling with python //Proceedings of the 9th Python in Science Conference. – 2010. – Т. 57. – С. 61.**
19. **Lemenkova P. Testing Linear Regressions by StatsModel Library of Python for Oceanological Data Interpretation. – 2019.**

**Electronic resource**

1. **Info and data about developed countries:** [**https://www.oecd.org**](https://www.oecd.org)
2. **Modern data about Japan:** [**https://www.japan.go.jp**](https://www.japan.go.jp)
3. **Info about developing countries:** [**https://www.worldbank.orga**](https://www.worldbank.orga)
4. **Docs of linearmodels:** [**https://bashtage.github.io/linearmodels/**](https://bashtage.github.io/linearmodels/)
5. **Python docs:** [**https://www.python.org/doc/**](https://www.python.org/doc/)
6. **Place where I read articles:** [**https://medium.com**](https://medium.com)
7. **Most Q&A site about programming: https://stackoverflow.com**