Jason Sjafrudin
28 November 2018

Introduction

Health has always been an important topic all around the world, especially in today's society. Throughout the history, human beings have been trying all the means to extend life expectancy. In order to do so, one must begin by understanding the factors that may affect their life expectancy and to what degree these factors may affect it. With industrialization, people are enjoying better material lives but at the same time have to live in polluted environment. With the recent wildfire that became the deadliest in California history and some econometrics research papers, I came across one topic in particular that intrigued me. In this paper, I will focus on analyzing **how PM2.5 air pollution is affecting the life expectancy of the population on planet earth, and to what degree it is affecting it**.

Although there are many factors that have the potential to affect the life expectancy of the world population, in order to carry out a statistical analysis, I can only (realistically) cover some of these factors. Therefore, except for PM2.5, I will also include other major factors that may contribute to affecting the life expectancy as well to prevent omitted variable bias; These include Gross National Income per capita, Poverty, Gini Index, smoking consumption, and expenditure on health care.

In summary, the goal of this paper is to mainly analyze how PM2.5, a measurement of air pollution, and other major factors (such as the ones mentioned previously) are affecting the life expectancy of the world population, and the degree that it may be affecting it. The results will most definitely be of value since I hope that this research paper can raise awareness of the air pollution and the importancency for all individuals to collaborate to reduce the overall PM2.5 so that we can all enjoy longer life expectancy.

Model Specification

Air pollution is the main independent variable that I would like to examine. However, life expectancy is complicated and would be affected by many other factors. Therefore, I decided to do multiple linear regression. After doing some researches, I figured out the major factors that would affect life expectancy, including socioeconomic status, the quality of healthcare, lifestyle such as tobacco/smoke consumption.

After some researches, in the end, for socioeconomic indicators, I included 1) Gross National Income per capita in dollars, 2) Poverty, measured by headcount of people earning under 1.90 dollars per day, 3) Gini Index, which measures income inequality. For lifestyle, I included 4) smoking prevalence. I think including smoking would be interesting since there are lots of controversies going on about whether smoking would affect life expectancy, there is also an increasing number of advertisement that informs people to stop smoking "it kills". For the quality of healthcare, I decided to include the 5) dollars amount spent on healthcare per capita.

Overall, with PM2.5, I included 6 independent variable. I have chosen to include these variables because with sufficient amount of research, I have come to find out that these are the major factors that has impactfully affect life expectancy as well. Besides the very fact that these impacts life expectancy, they also have no perfect collinearity with other explanatory variables. In other words they are not perfectly correlated and as a result, I did not violate any multiple regression model assumptions. Also, since each year is different, I wanted to include fixed time effect also to control for time-fixed effects; which are every other factor that change overtime but constant across states since I

am including data from 2011-2015. Therefore, I compiled panel data manually for data from 2011 to 2015 of 264 countries and regions listed on dataset from World Bank.

Our hypothesis on the coefficients; (expected signs for the parameters):

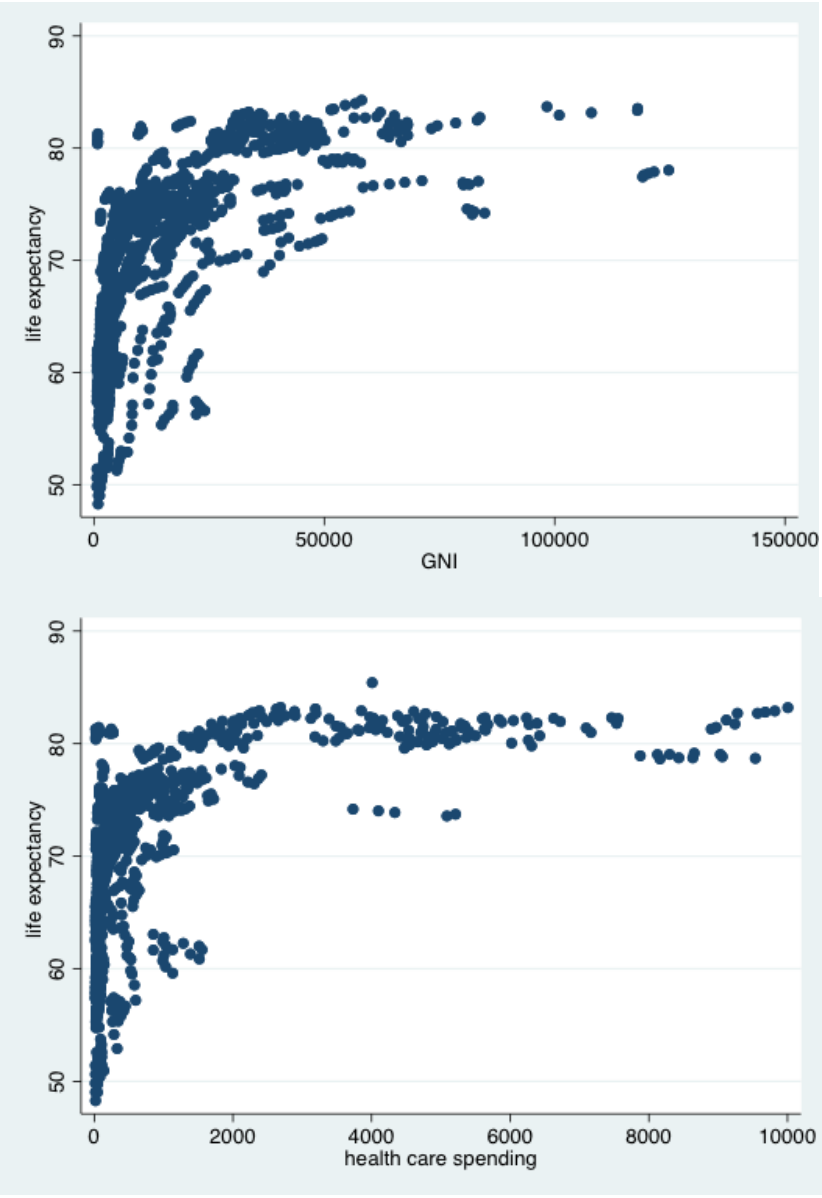
1. PM2.5
    - Negative Relationship with Life expectancy (+ sign for the coefficient)
2. Poverty
    - Negative Relationship with Life expectancy (+ sign for the coefficient)
3. log(Healthcare expenditure)
    - Positive Relationship with Life expectancy (+ sign for the coefficient)
4. log(GNI per capita)
    - Positive Relationship with Life expectancy (+ sign for the coefficient)
5. Smoking
    - Negative Relationship with Life expectancy (+ sign for the coefficient)
6. Gini index
    - Negative Relationship with Life expectancy (+ sign for the coefficient)

Data

Our projects consist of four models. In the first model, I ran the regression of life expectancy on the PM2.5, poverty headcount ratio, GNI per capita, and health care spending per capita. To avoid omitted variable bias, I included two more independent variable, namely smoking and Gini Index in the second model. In the third model, I included time fixed effects to control for omitted variables that are hard to measure but constant for all the countries and vary through years. In the fourth model, I included robust standard error to control for heteroskedasticity.
Here is the summary of sources and abbreviation of data used in the regression. All the data are collected from the World Bank. (Note: The data collected is from 2011 to 2015 from 264 countries.) Based on the assumption that the mortality patterns at birth is constant in the future, the life expectancy is calculated by taking the average number of years that a newborn is expected to live. This method provides a snapshot of the mortality pattern at a specific time. The air quality is measure by the index of PM2.5. PM2.5 is suspended particles with diameter less than 2.5 microns, which will penetrate into the respiratory system and harm the health condition. The PM2.5 index over the 2011-2015 measures the mean annual exposure (micrograms per cubic meter) by weighting average annual level of exposure to the PM2.5 by population in both urban and rural areas. Though it may not be a perfect representative of air quality, it makes a solid basis.

I believe that regional economic conditions may influence an individual's expectancy, omitting these factors may cause omitted variable bias. In this paper, I measured poverty by the headcount ratio of population living on less than 1.90 dollars per day. Preston, S.H (1975) established the Preston curve which indicates that individuals born in richer countries, on average, can expect to live longer than those born in poor countries. Therefore, I include Gross National Income per capita measured in dollars. Literature of Natasha Deshpande(2014) indicates that there exists a strong correlation between the family economic status, health care, and life expectancy. Thus, I include healthcare expenditure which, in this paper, is measured by per capita in dollars. Smoking prevalence is measure by the percentage of men and women age 15 and over who currently smoke any tobacco products. According to Leigh and Jencks's (2011) studies on the relationship of inequality-health, Gini Index represents the standard measure of income inequality. In this paper, the Gini Index coefficient is measured on a 0 to 100 scale.

Furthermore, after running scatterplot of all our independent variables and life expectancy, I found the scatterplots of GNI per capita and Healthcare expenditure are similar to log function. Therefore, I use the logarithm of GNI per capita and healthcare spending per capita in our models.



A. Variable Descriptions (Panal Data)

Variable Name	Description	Unit of Measurement	Type	Frequency	Year	Source
Life expectancy	Life expectancy at birth	Years	Dependent	Annual	2011-2015	World Bank
PM2.5	average annual exposure to PM2.5	Micrograms per cubic meter	Independent	Annual	2011-2015	World Bank
Poverty headcount ratio	the percentage of population living on less than 1.90	Percentage	Independent	Annual	2011-2015	World Bank

	dollars per day at 2011 international price					
Log(GNI)	Logarithm of Gross National Income per capita	Log_dollar	Independent	Annual	2011-2015	World Bank
Log(Healthcare)	Logarithm of healthcare spending per capita	Log_dollar	Independent	Annual	2011-2015	World Bank
Gini Index	The standard measure of income inequality	0-100 scale	Independent	Annual	2011-2015	World Bank
Smoking	Percentage of men and women age 15 and over who smoke	Age-standardized Rates	Independent	Annual	2011-2015	World Bank

B. Gauss-Markov Assumptions

I believe that the Gauss Markov assumptions hold in our model. First, all the parameters in the regression are in the linear relationship. Then, our data are not randomly sampled, since the world bank doesn't have enough sample data for poverty headcount ratio. The third assumption requires no multicollinearity among the regressor. This can be proven by the below table, in which there doesn't exist the same coefficient value. Next, I cannot testify that the expected value of error term is zero.

	lifeex~y	pm25ai~u	poverty	smoking	giniin~x	lnpove~y	lnheal~e
lifeexpect~y	1.0000						
pm25airpol~u	-0.5578	1.0000					
poverty	-0.7123	0.4834	1.0000				
smoking	0.2467	-0.2735	-0.3761	1.0000			
giniindex	-0.1690	0.0550	0.2420	-0.5640	1.0000		
lnpoverty	-0.6574	0.4460	0.7705	-0.4488	0.4746	1.0000	
lnhealthcare	0.6598	-0.5450	-0.5671	0.2965	-0.1341	-0.5518	1.0000

Results:

	Model 1	**Model 2**	**Model 3**	**Model 4**
Dependent variable:	**LifeExp**	**LifeExp**	**LifeExp**	**LifeExp**
PM2.5	-0.0343367	-0.0317352	-0.0325224	-0.0325224

	(0.0125594)	(0.0157415)	(0.0159967)	(0.012171)
Poverty	-0.1068783 (0.0202481)	-0.1339627 (0.0259286)	-0.1324847 (0.0262808)	-0.1324847 (0.0250869)
log(Healthcare expenditure)	1.40596 (0.2251941)	1.097105 (0.2316498)	1.101494 (0.2335312)	1.101494 (0.2884167)
log(GNI per capita)	2.069377 (0.3985106)	2.833123 (0.4246337)	2.810896 (0.4303315)	2.810896 (0.5345581)
Gini index	---	-0.030875 (0.0289266)	-0.0316733 (0.02914)	-0.0316733 (0.0322192)
Smoking	---	-0.102669 (0.0242975)	-0.1015985 (0.0245622)	-0.1015985 (0.0297362)
Intercept	46.99418 (3.263603)	45.38702 (3.805525)	45.39237 (3.852735)	45.39237 (4.083353)
Year fixed effect or no	**No**	**No**	**Yes**	**Yes**
R^2	0.7078	0.7077	0.7081	0.7081
RSS	5369.39527	3557.37831	3552.06095	3552.0609
N	412	314	314	314
F statistic	F(4, 407) = 246.43	F(6, 307) = 123.89	F(10, 303) = 73.52	F(10, 303) = 129.66
Standard Errors				**Robust**

$$\text{Life expectancy}_{it} = \alpha + \beta_1 \text{PM2.5}_{it} + \beta_2 \text{Poverty}_{it} + \beta_3 \log(\text{Healthcare expenditure})_{it} + \beta_4 \log(\text{GNI per capita})_{it} + \beta_5 \text{Smoking}_{it} + \beta_6 \text{Gini index}_{it} + u_{it}$$

i = 264 countries and regions listed on dataset from World Bank

t = 2011, 2012, 2013, 2014, 2015

Model 1:

$$\text{Life expectancy}_{it} = \alpha + \beta_1 \text{PM2.5}_{it} + \beta_2 \text{Poverty}_{it} + \beta_3 \log(\text{Healthcare expenditure})_{it} + \beta_4 \log(\text{GNI per capita})_{it} + u_{it}$$

I performed a multiple linear regression of life expectancy on the four major independent variables. The coefficients of poverty and PM2.5 are negative. With higher poverty and higher PM2.5, life expectancy is expected to decrease. On the other hand, healthcare spending and Gross National Income per capita have positive coefficients. Therefore, with higher healthcare spending and higher income, life expectancy is expected to increase.

All 4 of the variable coefficients(PM2.5, Poverty, log(Healthcare expenditure), log(GNI per capita)) have the sign as I hypothesized and are significant at 1% significance level.

```
. xi: regress lifeexpectancy pm25airpollutionmeanannualexposu poverty log_gni log_he
> althcarespending
```

Source	SS	df	MS	Number of obs	=	412
				F(4, 407)	=	246.43
Model	13004.2379	4	3251.05947	Prob > F	=	0.0000
Residual	5369.39527	407	13.1926174	R-squared	=	0.7078
				Adj R-squared	=	0.7049
Total	18373.6331	411	44.7047035	Root MSE	=	3.6322

lifeexpectancy	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
pm25airpollutio~u	-.0343367	.0125594	-2.73	0.007	-.0590261	-.0096474
poverty	-.1068783	.0202481	-5.28	0.000	-.1466823	-.0670743
log_gni	2.069377	.3985106	5.19	0.000	1.285981	2.852773
log_healthcares~g	1.40596	.2251941	6.24	0.000	.9632712	1.848649
_cons	46.99418	3.263603	14.40	0.000	40.57856	53.4098

Model 2:

$$\text{Life expectancy}_{it} = \alpha + \beta 1\text{PM2.5}_{it} + \beta 2\text{Poverty}_{it} + \beta 3\log(\text{Healthcare expenditure})_{it} + \beta 4\log(\text{GNI per capita})_{it} + \beta 5\text{Smoking}_{it} + \beta 6\text{Gini index}_{it} + u_{it}$$

After model 1, since I want to avoid omitted variable bias, so I included two more variables, namely smoking and Gini Index. After including these two variables, the signs of the previous four independent variables do not change and are still significant. The coefficients for Smoking and Gini index are both negative. Therefore, with tobacco usage and more income inequality, life expectancy are expected to decrease. However, the coefficients for Smoking is significant at 1% significance level and for Gini index is not significant at either 5% or 10% significance level.

5 of the variable coefficients (PM2.5, Poverty, log(Healthcare expenditure), log(GNI per capita), Smoking) have the sign as I hypothesized and are significant at 5% significance level. One variable(Gini index) have the hypothesized sign but is not significant.

```
. reg lifeexpectancy pm25airpollutionmeanannualexposu logexpenditure poverty log_gni smoking
>  giniindex
```

Source	SS	df	MS	Number of obs	=	314
				F(6, 307)	=	123.89
Model	8613.15063	6	1435.52511	Prob > F	=	0.0000
Residual	3557.37831	307	11.5875515	R-squared	=	0.7077
				Adj R-squared	=	0.7020
Total	12170.5289	313	38.883479	Root MSE	=	3.404

lifeexpectancy	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
pm25airpollutionmeanann~u	-.0317352	.0157415	-2.02	0.045	-.0627101	-.0007602
logexpenditure	1.097105	.2316498	4.74	0.000	.6412828	1.552927
poverty	-.1339627	.0259286	-5.17	0.000	-.184983	-.0829425
log_gni	2.833123	.4246337	6.67	0.000	1.997563	3.668684
smoking	-.102669	.0242975	-4.23	0.000	-.1504797	-.0548582
giniindex	-.030875	.0289266	-1.07	0.287	-.0877944	.0260444
_cons	45.38702	3.805525	11.93	0.000	37.89881	52.87523

Model 3:

$$\text{Life expectancy}_{it} = \alpha + \beta_1 \text{PM2.5}_{it} + \beta_2 \text{Poverty}_{it} + \beta_3 \log(\text{Healthcare expenditure})_{it} + \beta_4 \log(\text{GNI per capita})_{it} + \beta_5 \text{Smoking}_{it} + \beta_6 \text{Gini index}_{it} + u_{it}$$

After running model 2, I included time fixed effect to control for omitted variables that are constant for all the countries but change through time. These omitted variables are hard to measure but can be hold constant.

5 of the variable coefficients (PM2.5, Poverty, log(Healthcare expenditure), log(GNI per capita), Smoking) have the sign as I hypothesized and are significant at 5% significance level. The coefficient of Gini index has the sign as I hypothesized, but is not significant at either 5% or 10% significance level.

```
. xi: regress lifeexpectancy pm25airpollutionmeanannualexposu poverty log_gni log_he
> althcarespending smoking giniindex i.year
i.year            _Iyear_2011-2015    (naturally coded; _Iyear_2011 omitted)
```

Source	SS	df	MS	Number of obs	=	314
				F(10, 303)	=	73.52
Model	8618.46799	10	861.846799	Prob > F	=	0.0000
Residual	3552.06095	303	11.7229734	R-squared	=	0.7081
				Adj R-squared	=	0.6985
Total	12170.5289	313	38.883479	Root MSE	=	3.4239

lifeexpectancy	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
pm25airpollutio~u	-.0325224	.0159967	-2.03	0.043	-.0640011	-.0010437
poverty	-.1324847	.0262808	-5.04	0.000	-.1842006	-.0807688
log_gni	2.810896	.4303315	6.53	0.000	1.964079	3.657713
log_healthcares~g	1.101494	.2335312	4.72	0.000	.6419456	1.561042
smoking	-.1015985	.0245622	-4.14	0.000	-.1499325	-.0532645
giniindex	-.0316733	.02914	-1.09	0.278	-.0890157	.0256692
_Iyear_2012	.1780024	.6037692	0.29	0.768	-1.010109	1.366114
_Iyear_2013	.3523122	.6079862	0.58	0.563	-.8440978	1.548722
_Iyear_2014	.1314752	.6085444	0.22	0.829	-1.066033	1.328984
_Iyear_2015	.3325836	.623054	0.53	0.594	-.8934771	1.558644
_cons	45.39237	3.852735	11.78	0.000	37.81087	52.97388

Model 4:

$$\text{Life expectancy}_{it} = \alpha + \beta_1 \text{PM2.5}_{it} + \beta_2 \text{Poverty}_{it} + \beta_3 \log(\text{Healthcare expenditure})_{it} + \beta_4 \log(\text{GNI per capita})_{it} + \beta_5 \text{Smoking}_{it} + \beta_6 \text{Gini index}_{it} + u_{it}$$

This is our final and most complete model. I included robust standard error to take into account of heteroskedasticity. The coefficients of the variables are the same as model 3. With change of standard error, the coefficients of PM2.5 becomes significant at 1% significance level, whereas previously it is only significant at 5% significance level.

As previous model, 5 of the variable coefficients, namely PM2.5, Poverty, log(Healthcare expenditure), log(GNI per capita) and Smoking, have the sign as I hypothesized and are significant at 1% significance level. Gini index has the sign as I hypothesized, but is not significant at either 5% or 10% significance level.

```
. x1: reg lifeexpectancy pm25airpollutionmeanannualexposu poverty logexpenditure log_gni smo
> king giniindex i.year, robust
.year            _Iyear_2011-2015    (naturally coded; _Iyear_2011 omitted)

inear regression                                Number of obs     =        314
                                                F(10, 303)        =     129.66
                                                Prob > F          =     0.0000
                                                R-squared         =     0.7081
                                                Root MSE          =     3.4239
```

lifeexpectancy	Coef.	Robust Std. Err.	t	P>|t|	[95% Conf.	Interval]
m25airpollutionmeanann~u	-.0325224	.012171	-2.67	0.008	-.0564728	-.008572
poverty	-.1324847	.0250869	-5.28	0.000	-.1818514	-.083118
logexpenditure	1.101494	.2884167	3.82	0.000	.5339405	1.669047
log_gni	2.810896	.5345581	5.26	0.000	1.75898	3.862812
smoking	-.1015985	.0297362	-3.42	0.001	-.1601141	-.043083
giniindex	-.0316733	.0322192	-0.98	0.326	-.095075	.0317285
_Iyear_2012	.1780024	.6214032	0.29	0.775	-1.04481	1.400815
_Iyear_2013	.3523122	.6236948	0.56	0.573	-.8750095	1.579634
_Iyear_2014	.1314752	.615517	0.21	0.831	-1.079754	1.342704
_Iyear_2015	.3325836	.5971852	0.56	0.578	-.8425719	1.507739
_cons	45.39237	4.083353	11.12	0.000	37.35705	53.4277

Summary and Conclusion:

Overall, our results agree with our hypothesis. Our results show that PM2.5 negatively affect life expectancy, as I hypothesized. With increasing PM2.5, life expectancy is expected to decrease; 1 unit increase in PM2.5 is expected to decrease life expectancy by 0.0325224 year. The coefficients of poverty and log(GNI) per capita show that socioeconomic conditions do affect life expectancy. Poverty would negatively affect life expectancy, while log(GNI) per capita would positively affect life expectancy. Furthermore, healthcare expenditure is also one of the two independent variable that would also positively affect life expectancy. However, surprisingly, I found that Gini index, even though with a negative coefficient, does not play a really important role in life expectancy since the p-value is large (0.326). Despite all the controversies around whether smoking actually affect people's health, I have come to discover that smoking does negatively affect life expectancy.

References

1. Garcia, Juan, et al. “Modern Day Evaluation of the Preston Curve: The Relationship Between Life Expectancy and Income.” *Georgia Tech Library*, 2016.

2. Deshpande, Natasha, et al. “The Effect of National Healthcare Expenditure on Life Expectancy.” *Georgia Tech Library*, 2014.

3. Preston, S. H. (1975). The changing relation between mortality and level of economic development. Population studies, 29(2):231–248.

Andrew Leigh, Christopher Jencks, and Timothy M. Smeeding. (2011). “The Oxford Handbook of Economic Inequality: Health and Economic Inequality.”