To Live Longer, What We Should Do?

2000692

1/3/2021

Abstract

We wonder, for the well-being of a certain human, which factor matters more. Our conclusion based on data from OECD data set of Better Life, 2017, suggests that people in OECD countries feeling 1% safer walking at night, will have a longer life expectancy of 0.23 year. This result will shed some light on the study of sociologists and biologists, who can improve road security status and people's psychological construction, while aiming for a longer life span.

Data

We concentrate on nationals of OECD countries in year 2017. We retrieved the data from OECD data set, which include indicators such as housing expenditure, employment rate, years in education, air pollution and so forth. These official data have quite high quality, so measurement error can be neglected.

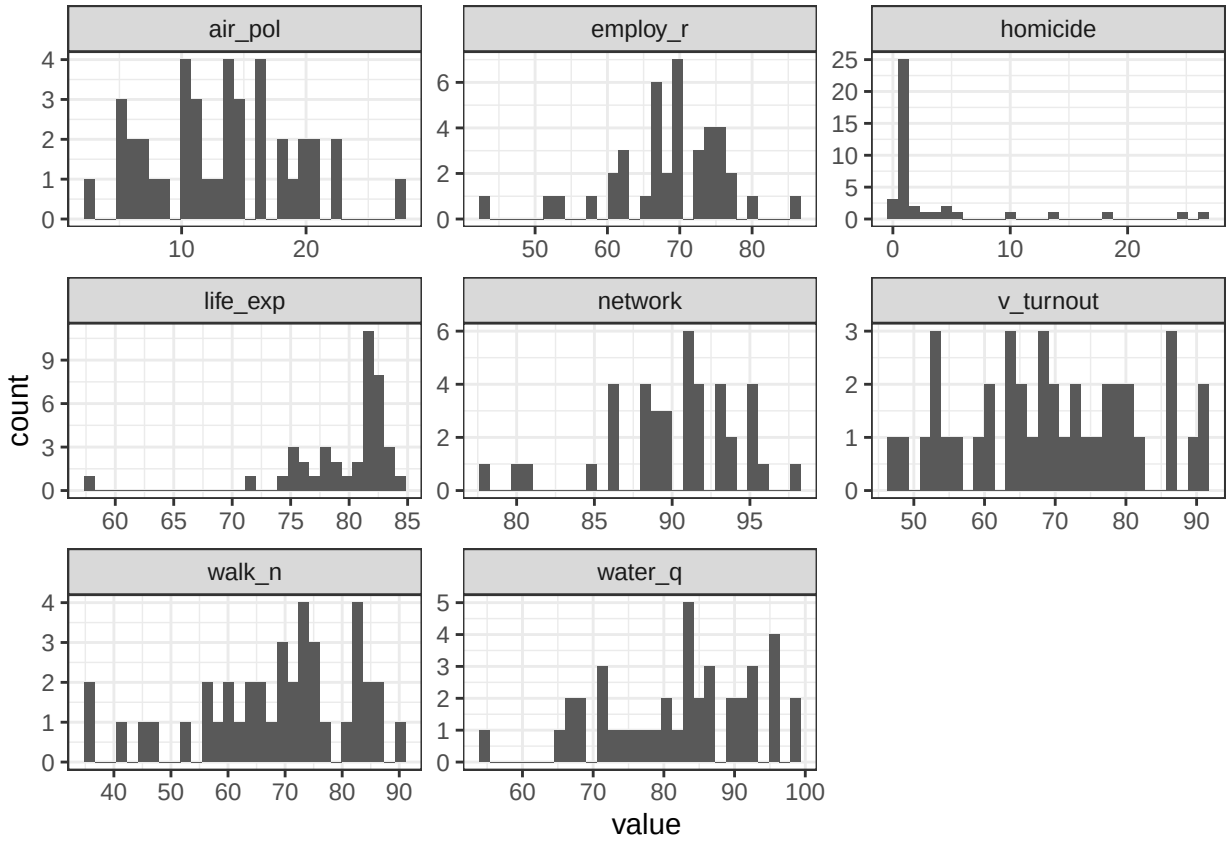
To keep our study precise, we removed variables with less than 40 observations; in the end, we obtained 9 variables: country names, life expectancy, employment rate, air pollution, water quality, supportive network, voter turnout, safety of walking at night and homicide. In the examination of these variables, we found some extreme values in homicide rate, which will affect our regression result, considering the size of our sample(40 observations, only). So, we decide not to drop these 5 values, but rather replace them with mean value of remaining 35 observations.

We intend to find the association of pattern between life expectancy and other factors. Our aim, is to find which factor contribute most to people's life span. Thus, we will analyze the statistics of these variables and build our model.

The following table shows the descriptive statistics of these variables. As we notice, 3/4 of our variables(except air pollution and life expectancy) are in percentage unit, which can not offer us further relative differences even if log transformation is applied. We also check all histograms, most of which do not have right-tailed pattern, except one, the homicide rate.

Table 1: Summary Statistics of Variables

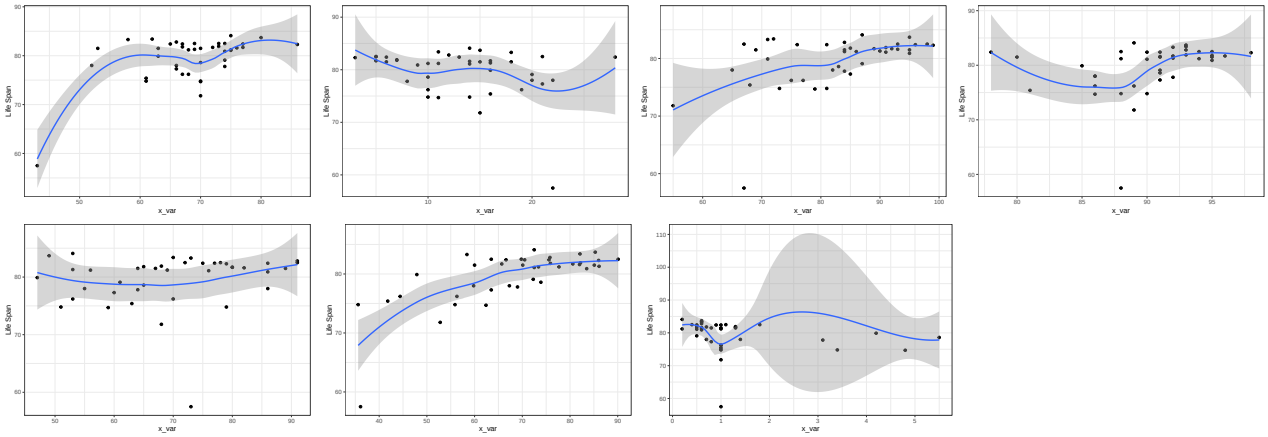
Name	n	Min	1st IQR	Median	3rd IQR	Max	Mean	Std.	Skew
Air pollution	40	3.0	9.75	14.00	16.50	28.0	13.32	5.77	0.26
Employment rate	40	43.0	65.75	69.50	74.00	86.0	68.47	7.97	-0.87
Homicide rate	40	0.2	0.60	0.95	2.12	26.7	3.43	6.33	2.64
Life expectancy	40	57.5	77.95	81.40	82.40	84.1	79.58	4.67	-2.85
Quality of support network	40	78.0	88.00	91.00	93.00	98.0	90.12	4.30	-0.86
Voter turnout	40	47.0	60.75	69.50	79.00	91.0	69.58	12.21	-0.02
Feeling safe walking alone at night	40	35.6	59.95	70.30	78.43	90.1	68.25	14.02	-0.67
Water quality	40	55.0	74.50	84.00	91.00	99.0	82.38	10.55	-0.48



Recall the extreme values of homicide rate: considering the small size of our sample, obviously these extreme values will affect our regression result. However, if we simply remove these values, our sample size will become even smaller, and we will be further from precise result as well. So, we decide to replace these 5 extreme values with the mean value of homicide rate.

Model

We intend to regress all variables on life expectancy. First, let's check lowess smoother graph to have a general idea about the pattern of life expectancy and other variables. Apart from employment rate and homicide rate, there are clear linear trend between life expectancy and other variables.



We try to capture this connection with a linear model:

(life expectancy)E = beta0 + beta1(employment_rate) + beta2(air pollution) + beta3(water quality) + beta4 (supportive of network) + beta5(voter turnout) + beta6(feeling safe walking at night) + beta7(homicide rate)

Note here that we are not taking all 40 observations for regression, but rather randomly filter out 8 observations for subsequent prediction and leave 32 observation for regression.

Our first model include all variables, in which we check the p value of all variables and eliminate one with p value larger pre-set significant level(10%). Our second model contains 6 explanatory variables only, in which we check again and delete unacceptable variable, regarding its p value. One by one, in the end we find that only the p values of "feeling safe walking at night", voter turnout and water quality are significant. Model 4 is our final choice. To be more specific, we can say that, comparing OECD nationals living under:

- same air pollution level, voter turnout rate and same percentage for feeling safe walking at night, people with ten percent higher water quality, on average, live less than 0.6 years.
- or, same same air pollution level, voter turnout rate and same level water quality, people with one higher percentage for feeling safe walking at night, on average, live more than 0.27 years.
- same air pollution level, same water quality and same percentage for feeling safe walking at night, people with ten percent higher voter turnout, on average, live less than 0.3 years.
- same voter turnout rate, same water quality and same percentage for feeling safe walking at night, people with ten percentage higher air pollution level, on average, live less than 0.5 years.

Based on model 4, we can declare with 95% confidence that the 51% of association is composed of linear regression. "Feeling safe walking at night has positive, while others have negative association with OECD national life span. All are significant at 10%. Model 4 is a better fit.

Why not model 5? You may ask. Sure we get a higher adjusted R-square in model 5, at the cost of substituting one more variable, which is suspect of "over-fitting". Besides, we would like to study the possible interaction of water quality and air pollution for rbustness check, so it is better to keep air pollution in this model.

Residual Analysis

For these 3 countries in table 2, the model overestimated life expectancy, as the actual value is smaller than the predicted value; in another word, these countries have short life span than average. The explanation could be extreme weather(temperature) or worsened social safety conditions.

For these 3 countries in table 3, the model underestimated life expectancy, as the actual value is larger than the predicted value; in another word, these countries have longer life span than average. The explanation could be healthy(Mediterranean and Eastern Asian diet) or complete social security administration.

Also, we check the y and y_hat plot to examine the model fit. We can see that most scatters fall aside the line, indicating a good fit of the model.

Prediction and Robustness analysis

We are concerned whether our analysis and model are true for other OECD countries; also, we wonder if the model we chose is truly "robust". Therefore, we check for two alternative specification:

- We test these 8 countries that are set aside for prediction analysis.
- We try to find the interaction between air pollution and water quality, as well as possible piecewise linear spline pattern as an alternative.

As we have only 8 observations for prediction analysis, it is difficult to draw a scatter plot and tell from the graph whether our prediction is precise, and also we are not certain if any of these 8 countries have extreme value, which will drag our analysis even further. Therefore, We choose to get the mean value of each variable and calculate the predicted value of this one "average" country. We check the residual of the regression model on the average value of 8 countries. The residual is also acceptable. Thus, we can say that our regression model is reasonable.

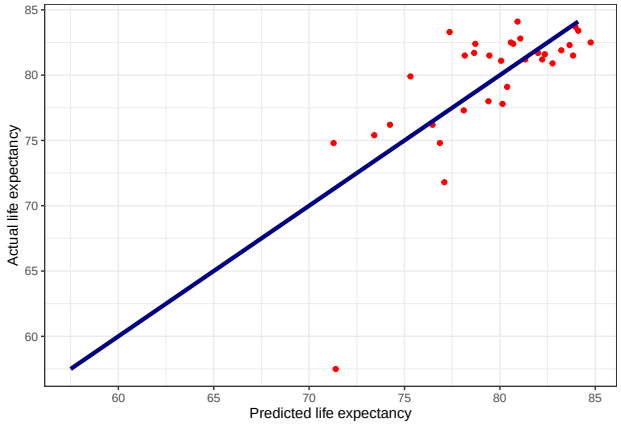
Also, we wonder whether our model performance will remain unchanged, taking interaction of explanatory variables and alternative models. First, we add interaction of air pollution and water quality as a new variable, which provides us with a new model. Also, from the scatter plot of life expectancy and air pollution, we find there are two "turning points" at x = 10 and 20; so we make a piecewise linear regression at these points, to see if this model will better fit the pattern.

As we can see from the figure, indeed using piecewise linear spline and interactions of certain x slightly increases R-square of the model, and they do not decrase p value of each variable to a significant degree. Furthermore, both models are more complicated to interpret. So, we decide to stick to the original model.

Summary

We study the relationship between life expectancy and air pollution, water quality, voter turnouts, the percentage of people feeling safe walking at night, in OECD countries. We build a multiple linear regression model, where we reach a conclusion that "Feeling safe walking at night has positive, while others have negative association with OECD national life span. Changing the model does not seems to have a significant effect on model fit. Also, the conclusion might be referential to sociologists: more effort in road safety and people's psychological health will possibly lead to longer life span.

Appendix


Actual life expectancy
Predicted life expectancy
85
80
75
70
65
60
60
65
70
75
80
85

	Model 1	Model 2	Model 3	Model 4	Model 5
(Intercept)	97.57 ***	97.63 ***	62.30 ***	68.65 ***	66.81 ***
	(19.93)	(20.49)	(11.87)	(8.17)	(5.92)
air_pol	-0.17	-0.17	-0.00	-0.05	
	(0.16)	(0.16)	(0.10)	(0.14)	
employ_r	0.15	0.15	0.14		
	(0.25)	(0.25)	(0.27)		
homicide	0.01				
	(0.90)				
network	-0.43	-0.43			
	(0.33)	(0.33)			
v_turnout	0.01	0.01	-0.01	-0.03	-0.02
	(0.04)	(0.04)	(0.05)	(0.06)	(0.06)
walk_n	0.27 **	0.27 **	0.24 **	0.27 *	0.27 *
	(0.08)	(0.08)	(0.07)	(0.11)	(0.10)
water_q	-0.08	-0.08	-0.10	-0.06	-0.05
	(0.20)	(0.19)	(0.21)	(0.11)	(0.09)
nobs	33	33	33	33	33
r.squared	0.59	0.59	0.53	0.51	0.51
adj.r.squared	0.48	0.50	0.45	0.44	0.46
statistic	3.17	3.80	4.29	3.53	4.37
p.value	0.02	0.01	0.01	0.02	0.01
df.residual	25.00	26.00	27.00	28.00	29.00
nobs.1	33.00	33.00	33.00	33.00	33.00
se_type	HC2.00	HC2.00	HC2.00	HC2.00	HC2.00

*** $p < 0.001$; ** $p < 0.01$; * $p < 0.05$.

Table 2: List of Coutnries with largest negative errors, Top 3

Country	life_exp	exp_pred	reg4_res
Estonia	77.8	80.14045	-2.340450
Russia	71.8	77.09081	-5.290814
South Africa	57.5	71.39627	-13.896267

Table 3: List of Coutnries with largest positive errors, Top 3

Country	life_exp	exp_pred	reg4_res
Chile	79.9	75.31331	4.586690
Italy	83.3	77.36488	5.935117
Korea	82.4	78.71501	3.684988

Table 4: prediction of regression model 4

life_exp	exp_pred	reg4_res
81.5	79.92731	1.572687

	Model 1	Model 2
(Intercept)	88.39 **	64.55 ***
	(26.24)	(9.30)
air_pol	-1.53	
	(1.83)	
v_turnout	-0.00	-0.02
	(0.04)	(0.05)
walk_n	0.28 *	0.27 *
	(0.11)	(0.11)
water_q	-0.31	-0.05
	(0.35)	(0.11)
air_pol:water_q	0.02	
	(0.02)	
lspline(air_pol, c(10, 20))1		0.30
		(0.36)
lspline(air_pol, c(10, 20))2		-0.17
		(0.24)
lspline(air_pol, c(10, 20))3		0.06
		(1.37)
nobs	33	33
r.squared	0.55	0.52
adj.r.squared	0.46	0.41
statistic	5.29	2.70
p.value	0.00	0.04
df.residual	27.00	26.00
nobs.1	33.00	33.00
se_type	HC2.00	HC2.00

*** $p < 0.001$; ** $p < 0.01$; * $p < 0.05$.