What factors cause the difference in life expectancy among states

Tomoki Okuno[1]
[1]UCLA Fielding School of Public Health, Dept. of Biostatistics

I. INTRODUCTION

Life expectancy, or the average remaining number of years at birth, is the key indicator for assessing population health. With the advancement of medical science and public health activities, the global life expectancy has increased from 50 to 73 years between 1950 and 2019.[1] Accordingly, the disparity among countries is gradually narrowing.

According to *Our World in Data*[1], only three countries have declined in life expectancy from 2013 to 2019: Venezuela, Grenada, and, surprisingly, the United States. In the United States, life expectancy began a 4-year decline in 2014 for the first time since 1943.

In addition, life expectancy varies from state to state in the US. It is critical, therefore, for policymakers to comprehend the factors that contribute to this relative decline in life expectancy and the significant differences across the US. Individual health issues such as high smoking rates, low physical activity, and lifestyle diseases, and inequalities such as income can affect life expectancy, and they can have a complex relationship with one another.[2, 3]

The aim of my study is to investigate which differences across states might affect life expectancy. Rather than compare one state to another, as past studies have done[4], I instead summarized states into a few clusters to achieve a macro perspective.

II. METHODS & MATERIALS

A. Data Set

I created the dataset for this study by merging more than 20 state-by-state datasets publicly available from the Kaiser Family Foundation. These datasets were collected or estimated in various ways, such as Census and the Current Population Survey, from 2014 to 2021. All variables were placed into three categories: Demographics, Health Investment, and Health Status of individuals. The specific variables for each category are as follows:

(1) Demographics
Total Population, Population by Age (%) (-18, 19-25, 26-34, 55-64, 65+) and by Race (%) (White, Black, Hispanic, Asian, Others), Percent of Citizens, Household Income, and Unemployment Rate.
(2) Health Investment
Health Spending per Capita, Health Insurance Coverage Rate (Employer, Medicaid, Medicare, Military, Uninsured), Distribution of Healthcare Expenditure (Hospital, Physician, Prescription, Others).
(3) Health Status of individuals
Life Expectancy, Deaths Rate by Race, Suicide Rate, Alcohol Deaths Rate, Drug Overuse Deaths Rate, Heart Disease Deaths Rate, Cancer Rate, Kidney Disease Rate.
*The total death rate was not included because of its obvious impact on life expectancy.

B. Study Population

For this study, the initial observation number is 51 (states). The mean, standard deviation, median, minimum, and maximum life expectancy for the entire state are 78.21, 1.701, 78.60, 74.40, and 81.00 years, respectively. However, as described below, I employed 37 variables to divide 51 states into some groups in similar states instead of conducting direct tests.

C. Statistical Methods

I conducted Principal Component Analysis (PCA) for the standardized variables except for “Life Expectancy” to obtain lower-dimensional variables while preserving as much of their validation as possible. The 51 states were then classified with the k-means clustering, known as a centroid-based clustering algorithm, based on their principal components.[5]

The k-means algorithm is randomized in its starting points, meaning that the outputs are different every time. It is, therefore, necessary to determine the optimal (or highly stable) number of clusters. In this study, I focused on the Calinski-Harabasz (CH) index

$$CH = \frac{Between\ Cluster\ Sum\ of\ Squares\ (BSS)}{Within\ Cluster\ Sums\ of\ Squares\ (WSS)},$$

and searched for the k-value with a large CS index.

After summarizing the study population, I performed a Pairwise t-test and obtained the Bonferroni adjusted p-value to observe whether there is a significant difference in the average life expectancy between any two groups. The Bonferroni method is a typical technique to avoid the increased errors caused by multiple comparisons. In this process, Bartlett's test was performed before the Pairwise t-test to check if equal variance could be assumed, although the multiple testing problem occurs.

Finally, if significant differences between any two groups were found, I described the characteristics statistically between them based on the principal components.

The statistical tool used was the latest version of Rstudio (2021.09.0 Build 351).

III. RESULTS

The results of PCA are shown in Table 1 with the eigenvalues, % of the variance (contribution ratio), and cumulative % of the variance. In this study, I adopted up to the first two principal components (PCs) because they accounted for around 70% of the total variance (Note: it is generally recommended to have at least 70-80%).

Table 1: Results of PCA
(Correlation between each variable and PCs)

Variable	PC1	PC2
White Deaths Rate	0.898	0.037
Fair or Poor Health Status Rate	0.881	0.333
Exercise Rate	-0.88	-0.224
Diabetes Rate	0.875	0.299
Household Income (Median)	-0.864	0.041
Smoking Rate	0.843	-0.133
Heart Disease Rate	0.813	0.249
Kidney Disease Rate	0.757	0.264
% of Asian Population	-0.564	0.132
Black Deaths Rate	0.517	0.529
% of 26-34 aged Population	-0.516	0.097
% of Citizen	0.512	-0.63
Hispanic Deaths Rate	-0.464	0.192
% of Employer Insurance	-0.46	-0.237
% of Medicare	0.457	-0.599
Cancer Rate	0.364	-0.395
AmIndAlaNat Deaths Rate	-0.353	-0.461
Medicaid Expenditures per Total	0.353	0.006
% of 55-64 aged Population	0.324	-0.462
% of Hispanic Population	-0.31	0.652
% of White Population	0.286	-0.77
%of Black Population	0.279	0.471
% of Uninsured	0.258	0.533
% of 65+ aged Population	0.251	-0.612
Unemployment Rate	-0.239	0.47
% of Medicaid	0.232	0.271
Health Spending per Capita	-0.225	-0.493
% of 35-54 aged Population	-0.214	0.376
Drug Overuse Death Rate	0.198	-0.123
Alcohol Death Rate	-0.166	-0.236
% of 19-25 aged Population	-0.164	0.553
Suicide Rate	0.141	-0.278
Total Population	-0.118	0.552
Eigenvalue	28.421	23.109
% of var.	38.406	31.228
Cum. % of var.	38.406	69.634

(1) Interpretation of PC1

Looking at the first principal component (PC1), which represents 40% of the total variable, one can see that the most positive contributing variables were Disease Rate for diabetes, heart disease, and kidney, Smoking Rate, and Fair or Poor Health Status Rate while the strong negative variables were Household Income and Exercise Rate. This indicates that this component is considered an “economic health poverty level.”

(2) Interpretation of PC2

On the other hand, the second principal component (PC2) had a strong relationship with population-related variables such as Hispanic, Black, and 19-25 aged Population positively, and White and 65+ aged Population negatively. Moreover, PC2 had a strong positive correlation with Uninsured and Unemployment Rate, and a negative correlation with Health Spending. For this reason, one can say PC2 means the "racial disparity level."

The results of cluster analysis are drawn in Figure 1. The line graph above shows WSS (the denominator of the CH formula) and CH as y-axis for the number of k (x-axis) from 1 to 10 by the k-means algorithm based on PC1 and PC2. Although CH continued to increase with k, the decrease in WSS (the denominator of CH) slowed down after k=4, which had a significant impact on CH's increase. Therefore, I set k=3, which had the largest CH between k=1 to k=4.

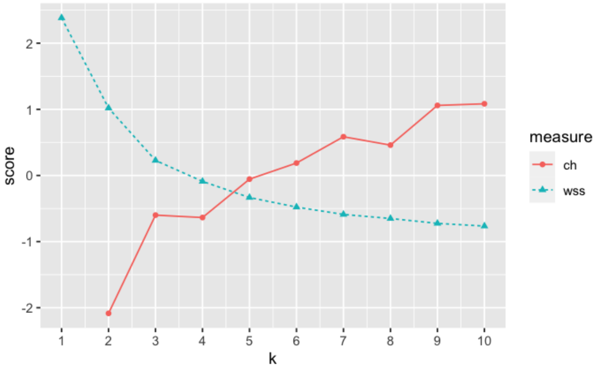

Figure 1: WSS and CH vs. the number of clusters

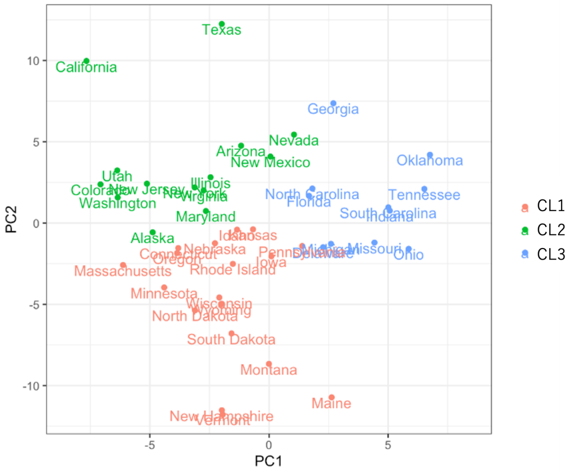

Figure 2: Results of k-means clustering

The scatter plot of Figure 2 and Table 2 indicate the three clusters of the 51 states (the result is not always exactly the same).

Table 2: States in each cluster

CL1	CL2	CL3
Connecticut	Alaska	Alabama
Idaho	Arizona	Arkansas
Iowa	California	Delaware
Kansas	Colorado	Florida
Maine	Dist. of Columbia	Georgia
Massachusetts	Hawaii	Indiana
Minnesota	Illinois	Kentucky
Montana	Maryland	Louisiana
Nebraska	Nevada	Michigan
New Hampshire	New Jersey	Mississippi
North Dakota	New Mexico	Missouri
Oregon	New York	North Carolina
Pennsylvania	Texas	Ohio
Rhode Island	Utah	Oklahoma
South Dakota	Virginia	South Carolina
Vermont	Washington	Tennessee
Wisconsin		West Virginia
Wyoming		

The results of testing these clusters to see if there is a difference in the life expectancy between any two clusters are shown in Table 3. Note that the prior Bartlett's test result was not significant at 5% level (p=0.0934), so the variances were assumed to be equal. It can be seen that there is sufficient evidence to show that the mean life expectancy differed between CL1 and CL3, and CL2 and CL3, unlike CL1 and CL2. In fact, the average life expectancy of CL3 was about three years lower than that of CL1 and CL2.

Table 3: Results of Pairwise t-test for life expectancy

p-values	CL1	CL2
CL2	1	-
CL3	9.20E-10	3.60E-09

Life expectancy		
CL1	CL2	CL3
79.178 yr	79.119 yr	76.329 yr

Finally, we discuss the differences between CL1 and CL3 and between CL2 and CL3, statistically different in life expectancy. Looking at Figure 2 again, CL3, such as Oklahoma and Tennessee, consists mainly of states where PC1 is positive. In other words, the “economic health poverty level” is high. In contrast, most of the states in CL1 (Montana and Maine) and CL2 (California and Utah) have both negative PC1. On the other hand, PC2, “racial disparity level,” does not appear to affect life expectancy, as there was insufficient evidence that life expectancy differed between CL1 and CL2, which have the opposite sign of PC2.

IV. CONCLUSIONS

I conclude that life expectancy is lower in the cluster of states with a high "economic health poverty level," which implies an increased number of people in the cluster indicating poor health, low household income, and lack of physical activity than in the other clusters.

V. DISCUSSION

Although conclusions can be made from the analysis, additional steps can be taken to further confirm these results. With respect to PCA, more meaningful results might be possible by checking normality beforehand and performing power transformations if necessary. In addition, regarding standardization (centering and scaling) of the variables, all the means of the variables need to be zero, while the operation to set the standard deviation to one is not necessarily needed. For example, if one wants to increase the weight of a variable on the principal components in proportion to the magnitude of the variance, scaling may be inappropriate, which is a point of consideration in further research.

In addition, since this study was analyzed from a macro perspective rather than focusing on one state at a time, it is necessary to go into more detail and depth based on the suggestions and research objectives obtained here. Furthermore, since the conclusions of this study are only correlational and do not indicate any causal relationship, causal inference methods provide another promising avenue for future work.

REFERENCES

1. Max Roser, Esteban Ortiz-Ospina and Hannah Ritchie, “Life Expectancy,” *Our World in Data*, https://ourworldindata.org/life-expectancy
2. Diane Whitmore Schanzenbach, Ryan Nunn, and Lauren Bauer. “The Changing Landscape of American Life Expectancy,” *The Hamilton Project*, JUNE 2016
3. Montez JK Beckfield J Cooney JK et al. “US state policies, politics, and life expectancy.” *Milbank Q*. 2020; 98: 668-699
4. Elena Falcettoni & Vegard Nygaard, 2020. "A Comparison of Living Standards Across the States of America," *Board of Governors of the Federal Reserve System (U.S.)*. FEDS Notes 2020-05-28-1
5. A. Ben-Hur and I. Guyon. Detecting stable clusters using principal component analysis. *In Functional Genomics: Methods and Protocols. M.J. Brownstein and A. Kohodursky (eds.) Humana Press*, 2003 pp. 159-182

Appendix

I. Source Data File

State Health Facts, Kaiser Family Foundation, https://www.kff.org/statedata/

(1) Demographics:
state-category/demographics-and-the-economy/

(2) Health Investment:
state-category/health-costs-budgets/
state-category/health-coverage-uninsured/

(3) Health Status:
state-category/health-status/

The following two items were uploaded via CCLE.

II. Final Clean Data set

Proj2_Merged_clean_dataset_TomokiOkuno.csv
Proj2_Summary_Output_TomokiOkuno.csv

III. R Code

Proj2_R_Code_TomokiOkuno.Rmd
Proj2_R_Code_TomokiOkuno.html