

# Methodological Insights: Factor Analysis and Variable Selection for Reliable Modeling

Jongchan LEE

## Abstract

This paper presents a structured approach to data analysis with a focus on Factor Analysis (FA) and variable selection techniques. These methods address common issues in highdimensional datasets, such as multicollinearity, and demonstrate how they can be effectively used to reduce the complexity of a model while retaining its interpretability and reliability. Although life expectancy data is used as an illustrative example, the methods discussed are broadly applicable across various fields. A two-step process is outlined: first, variables are reduced using the Akaike Information Criterion (AIC), and second, further dimensionality reduction is performed through FA. The results reveal how FA can provide insights into the relationships between variables by simplifying them into lower-dimensional factors. This paper highlights the importance of understanding both the data and statistical methods to avoid misleading conclusions and to derive accurate and meaningful results from analysis.

*Keywords:* Factor Analysis, Variable Selection, Multicollinearity, Akaike Information Criterion, Life Expectancy Data

## 1 Introduction

In general, the primary goal of data analysis is to estimate population parameters using sample data and to draw accurate and meaningful conclusions. To achieve these goals, a well-designed analysis plan is essential, which includes steps such as data sampling, data cleaning, descriptive statistics, and inferential statistics. While each step in the data analysis process is important, particular care must be taken during the descriptive and inferential analysis stages, because most statistical methods in inferential statistics rely on probabilistic assumptions that must be considered. For example, in a two-sample t-test, the results are valid under the assumption of normally distributed populations. More precisely, if this assumption is violated, alternative

methods should be employed to ensure the validity of the analysis results, or the analysis should at least be supplemented to account for the violation of the assumption.

A good understanding of the assumptions that underlie certain statistical methods is critical not only for making the results more reliable in a theoretical sense, but also for determining which statistical methods should be used, avoided, or applied with caution. This is why most typical textbooks include chapters related to probability theory as well.

However, in real-world data analysis, final conclusions are often incorrectly derived by focusing solely on specific indices, such as p-values or  $R^2$ , provided by statistical software, without deep consideration of underlying assumptions or mathematical principles. This kind of analysis can lead to undesirable results, which should be avoided.

In real data analysis, another set of issues often arises. Suppose the goal is to model the relationship between a continuous response variable  $y$  and exploratory variables  $x_1, \dots, x_p$ , where the dimensions of the data are  $n \times 1$  for the response variable and  $n \times p$  for the exploratory variables, respectively. However, problems often occurs when  $p$ , the number of predictors, is large relative to  $n$ , the number of observations.

To control this, variable selection methods are often considered to identify a subset of predictors that contribute most significantly to the model. One common approach is stepwise regression, where variables are added or removed from the model based on statistical criteria such as the Akaike Information Criterion (AIC) or p-values (Akaike, 1974). Another approach is regularization techniques like Lasso (Least Absolute Shrinkage and Selection Operator), which imposes a penalty on the size of the coefficients, effectively shrinking some of them to zero and thus performing variable selection (Hastie et al., 2009). At times, it is necessary to address multicollinearity problems, where two or more predictor variables are highly correlated, which can result in inflated standard errors and, consequently, unreliable coefficient estimates (Hoerl and Kennard, 1970). Another potential issue is overfitting, where the model becomes too complex, capturing noise rather than the true underlying relationships in the data, leading to poor predictive performance on new data (Agresti, 2013). Depending on the research objective, more advanced techniques such as Principal Component Analysis (PCA) and Factor Analysis (FA) should be considered for variable dimension reduction while still capturing the overall information in the data (Johnson et al., 2014).

As discussed above, even in seemingly simple problems like regression analysis, there are numerous combinations of data analysis methods that can be applied to achieve the same objective, due to the wide range of available statistical techniques. This often leads to the misconception that data analysis can vary widely depending on the analyst and that there is a subjective aspect to it. However, a sound understanding of each statistical method from a theoretical perspective is essential for making informed decisions about which methods to use and for effectively planning the entire data analysis process. Even in simple scenarios, thorough consideration of the underlying theory is necessary to ensure a reliable and robust analysis.

In this paper, the data analysis process using real data is discussed, where the number of predictors ( $p = 160$ ) greatly exceeds the number of observations ( $n = 47$ ). While some results from the analysis are reported, the primary focus is on the methodology—what was used, why it was used, and how it can be applied. This approach aims to provide clear guidance for those who may

be unsure of how to proceed with their data analysis.

This paper is organized as follows: The data structure is outlined, and the analysis objectives are defined in Section 2. The discussion then transitions to Factor Analysis (FA) in Section 3, which plays a crucial role in illustrating results and aiding in their correct interpretation in Section 6. In Section 4, the data is explored through correlation as the first step in data exploration, providing insights for the subsequent steps. Section 5 addresses the variable selection process, focusing on the initial screening of predictors. This is followed by a discussion on dimension reduction through FA (Section 6). Finally, linear regression is applied to identify relationships within the data (Section 7) based on the results from this analysis.

## 2 Data and Variables Used in the Analysis

To simplify the discussion, let us assume that the objective of this data analysis is to identify the factors that influence life expectancy using publicly available data.

To illustrate the real-world data analysis process, the average life expectancy (LE) of males from 47 prefectures in Japan is used as the response variable (mean = 80.652, ranging from 78.67 in Aomori to 81.78 in Shiga). A total of 162 explanatory variables for these prefectures were collected, covering a wide range of areas, including not only health and medical variables but also social and economic variables. All data used in this study is publicly available and can be downloaded from e-Stat, the portal for Japanese Government Statistics<sup>1</sup>.

In this paper, matrix notation will be used where appropriate to simplify the explanation. The explanatory dataset is represented as  $\mathbf{X}_{47 \times 162}$ , where the rows correspond to the 47 prefectures and the columns represent the 162 explanatory variables. The response variable,  $\mathbf{y}_{47 \times 1}$ , denotes the male average life expectancy (LE) for each prefecture. Where context allows, subscripts indicating matrix dimensions will be omitted for clarity.

## 3 Quick Overview of Factor analysis from the Matrix Decomposition

In Section 6, Factor Analysis (FA) is used both for variable dimension reduction and as a tool for visualization. Unlike elementary visualization methods such as bar plots or scatter plots, Factor Analysis requires a deeper understanding of its underlying mathematical principles to ensure accurate interpretation. For this reason, this section is placed immediately after the data section to provide the necessary mathematical foundation before proceeding with the analysis.

Let  $\mathbf{X}_{n \times p}$  be data matrix, where each column represents a variable, and each row represents an observation. The standardized data matrix  $\mathbf{Z}$  is obtained by subtracting the mean vector  $\boldsymbol{\mu}$  from each observation and scaling by the standard deviation (or covariance matrix  $\boldsymbol{\Sigma}$ ).

The standardized data matrix  $\mathbf{Z}$  is given by:

$$\mathbf{Z} = (\mathbf{X} - \mathbf{1}_n \boldsymbol{\mu}^T) \boldsymbol{\Sigma}^{-1/2}$$

---

<sup>1</sup> <https://www.e-stat.go.jp/en>

where the vector  $\mathbf{1}_n$  is an  $n \times 1$  vector of ones,  $\boldsymbol{\mu}$  is the  $p \times 1$  vector of means for each variable,  $\Sigma$  represents the  $p \times p$  covariance matrix of the variables, and  $\Sigma^{-1/2}$  denotes the inverse of the square root of the covariance matrix. After standardization, the resulting matrix  $\mathbf{Z}$  has variables with mean 0 and standard deviation 1.  $\mathbf{Z}$  can be represented as a set of  $p$  variable vectors  $\mathbf{z}_j, j = 1, \dots, p$ ,

$$\mathbf{Z} = \begin{bmatrix} \mathbf{z}_1 & \mathbf{z}_2 & \cdots & \mathbf{z}_j & \cdots & \mathbf{z}_p \end{bmatrix}$$

The objective in factor analysis is to minimize the sum of squared projections of the standardized column vectors  $\mathbf{z}_j$  onto the factor vector  $\sqrt{n}\mathbf{u}$ , subject to the constraint that the length of each  $\mathbf{z}_j$  is  $\sqrt{n}$  (Figure 1). This can be formulated as:

$$\arg \max_{\mathbf{u}} \sum_{j=1}^p \left\| \frac{\mathbf{z}_j' \mathbf{u}}{\sqrt{n}} \right\|^2 = \arg \max_{\mathbf{u}} \frac{\mathbf{u}' \mathbf{Z} \mathbf{Z}' \mathbf{u}}{n}$$

subject to  $\|\mathbf{z}_j\| = \sqrt{n}$ ,  $\|\mathbf{u}_j\| = 1$  for all  $j = 1, 2, \dots, p$ .

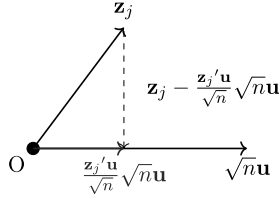


Figure 1: Projection of vector  $\mathbf{z}_j$  onto vector  $\mathbf{u}$  with perpendicular component  $\mathbf{z}_j - \frac{\mathbf{z}_j' \mathbf{u}}{n} \mathbf{u}$

To solve this optimization problem, a Lagrange multiplier  $\lambda$  is introduced to enforce the constraint  $\mathbf{u}' \mathbf{u} = 1$ . The corresponding Lagrange function is:

$$\mathcal{L}(\mathbf{u}, \lambda) = \frac{\mathbf{u}' \mathbf{Z} \mathbf{Z}' \mathbf{u}}{n} - \lambda(\mathbf{u}' \mathbf{u} - 1)$$

Taking the derivative of  $\mathcal{L}$  with respect to  $\mathbf{u}$  and setting it to zero gives:

$$\frac{\partial \mathcal{L}}{\partial \mathbf{u}} = \frac{2}{n} \mathbf{Z} \mathbf{Z}' \mathbf{u} - 2\lambda \mathbf{u} = 0$$

This process simplifies to the following eigenvalue equation:

$$\frac{\mathbf{Z} \mathbf{Z}'}{n} \mathbf{u} = \lambda \mathbf{u}$$

where  $\lambda$  represents the eigenvalue corresponding to the eigenvector  $\mathbf{u}$ . In matrix notation, the eigenvalue decomposition of the covariance matrix  $\frac{\mathbf{ZZ}'}{n}$  can be written as:

$$\frac{\mathbf{ZZ}'}{n} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}'$$

where:

- $\mathbf{U}$  is an  $n \times p$  orthogonal matrix containing the eigenvectors  $\mathbf{u}_1, \mathbf{u}_2, \dots, \mathbf{u}_p$ ,
- $\mathbf{\Lambda}$  is an  $p \times p$  diagonal matrix having the eigenvalues  $\lambda_1, \lambda_2, \dots, \lambda_p$  in diagonal elements.

The eigenvectors  $\mathbf{u}_i$  are an orthonormal basis ( $\mathbf{u}'\mathbf{u} = 1$ ). By solving the eigenvalue equation, the directions (factors) that maximize the variance of the original data are identified, which is the primary objective of Factor Analysis. The only difference in Factor Analysis is that the factor vectors should be scaled by  $\sqrt{n}$ , as they are assumed to have a length of  $\sqrt{n}$ . Therefore, the factor vectors are represented as  $\sqrt{n} \times \mathbf{u}_i$  instead of  $\mathbf{u}_i$ , to maintain the appropriate scaling in the factor space.

Additionally, the same result can be derived through the Singular Value Decomposition (SVD) of  $\mathbf{Z} / \sqrt{n}$ . SVD decomposes the matrix  $\mathbf{Z} / \sqrt{n}$  into three components.

$$\frac{\mathbf{Z}}{\sqrt{n}} = \mathbf{U}\mathbf{D}\mathbf{V}' \quad (1)$$

where:

- $\mathbf{U}$  is an  $n \times p$  orthogonal matrix containing the left singular vectors,
- $\mathbf{D}$  is an  $p \times p$  diagonal matrix containing the singular values,
- $\mathbf{V}$  is a  $p \times p$  orthogonal matrix containing the right singular vectors (which are also the principal components).

The singular values in  $\mathbf{D}$  are the square roots of the eigenvalues of the covariance matrix  $\mathbf{ZZ}' / n$ .

Although Factor Analysis (FA) is rarely explained through Singular Value Decomposition (SVD) in typical statistical texts, this approach provides a more intuitive understanding of the principles behind biplots, as it directly reveals the geometric relationships between variables and observations in a lower-dimensional space (Huh, 2012). To understand the connection between Factor Analysis and Singular Value Decomposition (SVD), consider the following forms instead of Eq.(1):

$$\mathbf{Z} = \sqrt{n}\mathbf{U}\mathbf{D}\mathbf{V}' = \mathbf{F}\mathbf{L}'$$

In the context of Factor Analysis, define:

$$\mathbf{F} = \sqrt{n}\mathbf{U} \quad \text{and} \quad \mathbf{L} = \mathbf{V}\mathbf{D}$$

Here,  $\mathbf{F}$  corresponds to the factor scores matrix, and  $\mathbf{L}$  corresponds to the factor loadings matrix. The factor scores  $\mathbf{F}$  represent the coordinates of the observations in the reduced factor space, while the factor loadings  $\mathbf{L}$  represent how the original variables relate to the factors. This result allows for the creation of a plot known as the biplot (Gabriel, 1971).

### **Biplot Representation:**

Using the results of this decomposition, we can construct a *biplot*, which simultaneously visualizes both the observations (rows of  $\mathbf{Z}$ ) and the variables (columns of  $\mathbf{Z}$ ) in the same factor space. In the biplot:

- The **observations** are plotted using the rows of the factor scores matrix  $\mathbf{F}$ . Each observation is represented as a point in the factor space.
- The **variables** are plotted using the rows of the factor loadings matrix  $\mathbf{L}$ . Each variable is represented as a vector in the same factor space.

The biplot provides a graphical representation of how the observations (e.g., prefectures) are positioned in relation to each other and how the variables (e.g., health, social, economic indicators) are correlated. The biplot is illustrated in Section 7.2.

## **4 Correlation analysis as the first step in the Data Analysis Process**

The data analysis process is now revisited, incorporating real-world data analysis results. As mentioned in Section 2, the number of variables ( $p$ ) is significantly larger than the number of prefectures ( $n = 47$ ). Given the size of the data, handling the correlations between all pairs of variables can be overwhelming. Nonetheless, exploring the correlations between the response variable and the explanatory variables provides valuable insights and serves as a useful *first step* in the analysis.

In this case, the correlation between male life expectancy (the response variable) and each of the 162 explanatory variables ( $x_1, x_2, \dots, x_{162}$ ) is explored. This allows for the identification of potential relationships between life expectancy and various factors, guiding the subsequent steps in the analysis process.

### **4.1 Correlation Between Life Expectancy and Various Factors**

Among all the explanatory variables, those that show high correlations with the response variable, life expectancy, are listed in Table 1. These variables are not only correlated with health and

medical factors, such as malignant neoplasms and cerebrovascular disease mortality rates, but also with non-health variables, including social and economic factors. For instance, life expectancy is positively correlated with variables like sports participation rate ( $r = 0.67$ ), computer ownership ( $r = 0.62$ ), and educational attainment ( $r = 0.62$ ). This suggests that lifestyle factors and socio-economic conditions can significantly influence life expectancy, beyond just medical conditions.

Table 1: Correlation Between Life Expectancy and Explanatory Variables (Males)

Variable Name	Correlation with Life Expectancy (2015)
<b>Health Variables</b>	
Under-75 Age-Adjusted Mortality Rate_Malignant Neoplasms_2019	-0.84
Under-75 Age-Adjusted Mortality Rate_Malignant Neoplasms_2018	-0.83
Malignant Neoplasms (Colon)_Age-Adjusted Mortality Rate 2015	-0.69
Age-Adjusted Mortality Rate_Cerebrovascular Disease_2015	-0.65
Cerebrovascular Disease_Age-Adjusted Mortality Rate 2015	-0.65
Pneumonia_Age-Adjusted Mortality Rate 2015	-0.54
Malignant Neoplasms (Stomach)_Age-Adjusted Mortality Rate 2015	-0.47
Malignant Neoplasms (Trachea, Bronchus, and Lung)_Age-Adjusted Mortality Rate 2015	-0.44
Outpatient Rate_Cerebrovascular Disease_2017	-0.59
Hospitalization Rate_Malignant Neoplasms_2017	-0.41
<b>Economic Variables</b>	
Household_Current Savings Balance	0.55
Total Cash Earnings_2016	0.41
Administrative Foundation_Fiscal Power Index	0.40
Labor_Primary Industry Employment Rate	-0.58
Total Working Hours_2016	-0.45
Labor_Unemployment Rate	-0.44
<b>Social Variables</b>	
Culture and Sports_Sports Participation Rate	0.67
Household_Number of Computers Owned (per thousand households)	0.62
Education_Percentage of University/Graduate School Graduates	0.62
Culture and Sports_Travel and Leisure Participation Rate	0.62
Household_Number of Tablet Devices Owned (per thousand households)	0.58
Household_Number of Smartphones Owned (per thousand households)	0.54
Residence_Sewerage Coverage Rate	0.47
Population and Households_Elderly Population Percentage 2020	-0.41
Population and Households_Crude Death Rate 2020	-0.55

One might hesitantly conclude that, for example, the sports participation rate ( $r = 0.67$ ) and the unemployment rate ( $r = -0.44$ ) significantly affect life expectancy. However, interpretation must be done carefully to avoid misleading conclusions. One important consideration is that conclusions drawn from pairwise correlations, and selectively picking certain variables, can be misleading due to the presence of spurious relationships.

Rather, the figures shown in table 1 should be regarded as providing some overall descriptive insights (not definitive evidence) and as an *intermediate step* toward more advanced analysis. The key insight from this descriptive analysis is the hypothesis that life expectancy is not only influenced by health-related factors but potentially by other non-health factors as well.

## 5 One Further Step Towards Robust Modelling: Variable Selection

As discussed in Section 1 and Section 2, the dataset includes 162 explanatory variables for only 47 observations. It is essential to reduce the number of variables before proceeding with modeling to ensure that the model remains interpretable and reliable. Given the high dimensionality of the data, applying variable selection is an inevitable step before any causal effect model analysis involving the response and explanatory variables. This process is necessary to prevent overfitting and to improve the robustness of the model.

In this paper, a 2-step variable selection process is proposed before applying the model. The first step involves applying the Akaike Information Criterion (AIC) for an initial reduction in the

number of variables. This process is carried out in accordance with the algorithm shown in Figure 2, which is suggested in this study. In the second step, the result from the AIC selection is further refined by applying Factor Analysis (FA), with the goal of reducing the dimensions to two. This 2-step approach ensures that the most relevant explanatory variables are retained while simplifying the model for better interpretability and accuracy. The second step will be covered in more detail in the next section.

In step 1, the aim is to simplify the model problem from the situation shown in Figure 3 to that in Figure 4, following the algorithm outlined in Figure 2. response variable effects



Figure 3: Before model selection



Figure 4: After model selection

Once the problem is simplified by the algorithm in Figure 2, as shown in Figure 4, where  $n > p$ , a linear model, such as linear regression model, can be applied. A linear regression model combines  $p$  explanatory variables  $x_1, \dots, x_p$  with weights  $\beta_1, \dots, \beta_p$  and a constant term  $\beta_0$ , as expressed in the following form:



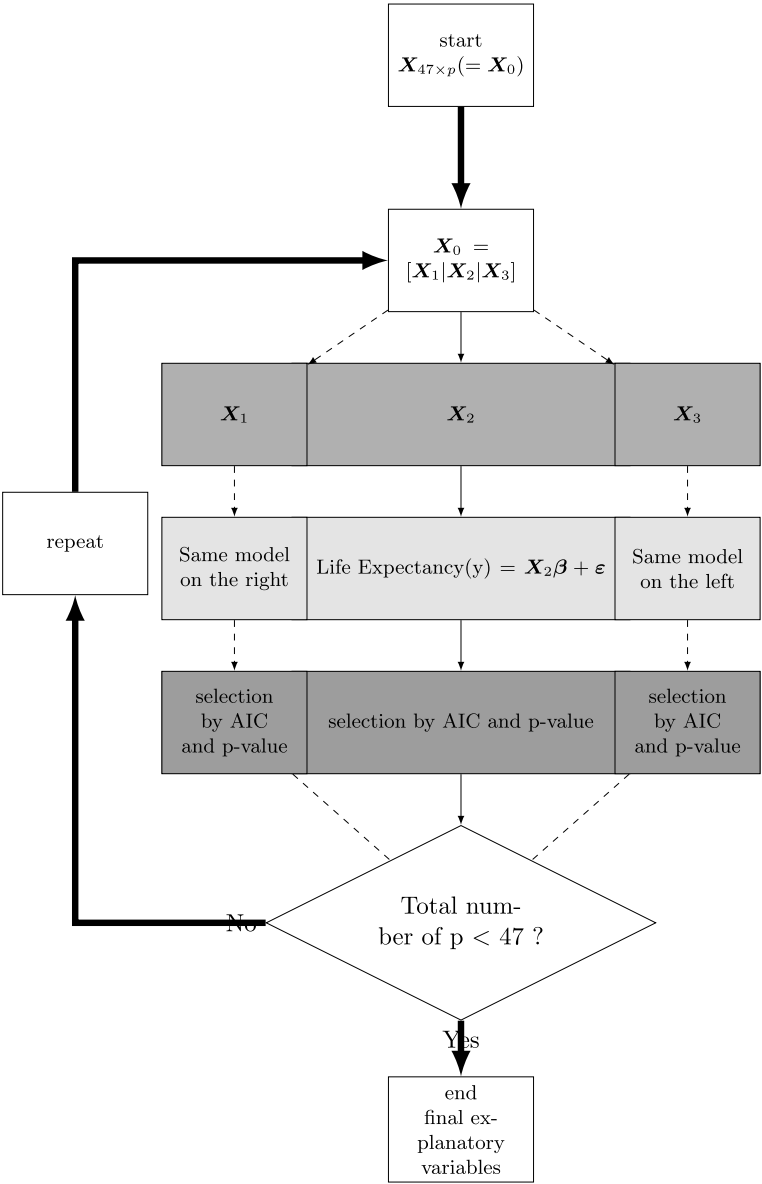


Figure 2: Variable Selection Algorithm

$$\mu = \beta_0 + \beta_1x_1 + \cdots + \beta_px_p$$

This model explains the probabilistic relationship between the expected value of  $y$ , denoted as  $\mu$ , and the dependent variable  $y$ . While simple, it is robust and widely used due to its ability to estimate the effects of changes in explanatory variables (e.g., how life expectancy might change with a 1% increase in volunteer participation). Provided that  $p$  is sufficiently small and the explanatory variables are nearly linearly independent, linear regression can be confidently applied without encountering issues such as multicollinearity or overfitting.

However, if  $p$  is large or if multicollinearity exists, additional steps may be necessary to ensure reliable regression results. In such cases, techniques like Factor Analysis (FA) can be applied to transform the explanatory variables into a set of linearly independent components, thus addressing potential multicollinearity and improving the model's stability.

In the case of the illustrated results in Table 2, which presents the variables selected by the process outlined in Figure 2 using life expectancy data, this corresponds to the latter case. Although the number of variables has been drastically reduced from 162 to 10, this still represents approximately one-fourth of  $n$  (the total number of observations). Furthermore, several pairs of explanatory variables exhibit non-trivial correlations, suggesting potential issues with multicollinearity.

Table 2: Selected Variables After Variable Selection (Male)

Variable Name
<b>Health Variables</b>
Hospitalization Rate for Heart Disease (2017)
Age-Adjusted Mortality Rate for Colorectal Cancer (2015)
Number of Public Health Nurses (per 100,000 population)
<b>Social Variables</b>
Self-Development/Training - Computer and Information Processing
Self-Development/Training - Arts and Culture
Self-Development/Training - Foreign Languages (Other than English)
Proportion of Elderly Single-Person Households
<b>Economic Variables</b>
Current Household Savings
Barrier-Free Rate (2018)
<b>Environmental Variables</b>
Average Annual Temperature

For example, the correlation between *Self-Development/Training-Computer and Information Processing* and *Self-Development/Training-Arts and Culture* is 0.82, suggesting a multicollinearity problem in the modeling analysis. To mitigate this issue and reduce the complexity of the problem, making it more suitable for modeling, additional data reduction is performed via Factor Analysis (FA).

## 6 Variable Reduction Using Factor Analysis

In this section, the focus shifts to further reducing the dimensionality of the explanatory variables using Factor Analysis (FA). The goal is to reduce the number of explanatory variables to two or three factors. This reduction simplifies the model while retaining most of the relevant information.

Additionally, by reducing the dimensions to two or three, it becomes easier to visualize the relationships between the variables and gain insights into the underlying structure of the data.

Before illustrating the data results, a brief overview of FA will be provided. Although the mathematical concepts have already been discussed in Section 3, FA will now be approached in a more data-friendly manner to facilitate its practical application in this analysis.

Factor Analysis (FA) is a multivariate statistical method used to reduce the dimensionality of large datasets by identifying underlying latent factors. It assumes that a small number of unobserved latent factors influence the observed explanatory variables linearly, as shown in the equation:

$$X_i = l_{i1}F_1 + l_{i2}F_2 + \cdots + l_{im}F_m + \epsilon_i,$$

where  $F_1, F_2, \dots, F_m$  are the latent factors,  $l_{ij}$  are the factor loadings, and  $\epsilon_i$  represents the unique factor for each observed variable. This concept is illustrated in Figure 5.

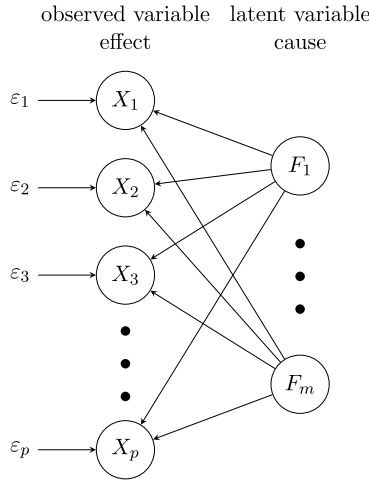


Figure 5: Conceptual Diagram of Factor Analysis

For instance, consider the observed variables  $X_1, X_2, \dots, X_5$  representing subject scores in areas such as mathematics, language, and science. The latent factors  $F_1$  and  $F_2$  could represent underlying learning abilities, such as analytical skills and verbal skills, which influence these scores. The relationship between the observed scores and the latent factors can be written as:

$$X_1 = l_{11}F_1 + l_{12}F_2 + \epsilon_1 \quad (2)$$

$$X_2 = l_{21}F_1 + l_{22}F_2 + \epsilon_2$$

$$\vdots$$

$$X_5 = l_{51}F_1 + l_{52}F_2 + \epsilon_5$$

Here,  $l_{ij}$  represents the factor loadings, which indicate how strongly each subject score is influenced by each learning ability. The factors  $F_1$  and  $F_2$  are the unobserved latent variables (learning abilities), and  $\varepsilon_i$  represents the unique factors or noise specific to each subject score.

The factor scores for each observation, i.e. the *estimated* values of  $F_1$  and  $F_2$ , quantify their levels of the underlying latent factors, while the factor loadings  $l_{ij}$  indicate how much each factor contributes to the observed variables for a given subject.

## 6.1 Factors, Factor Loadings, and Their Interpretation

The factor analysis results, yielding two factors with a cumulative variance proportion of 0.58, are presented in Table 3. The first column in Table 3 lists the 10 variables from the life expectancy data selected through the variable selection process described in Section 5, while the remaining columns display the factor loadings for the two factors derived from the Factor Analysis.

Table 3: Factor Loadings for Male Variables

	Variable Name	$F_1$	$F_2$
1	Hospitalization Rate for Heart Disease 2017	0.02	<b>-0.61</b>
2	Annual Average Temperature	0.50	0.16
3	Public Health Nurses per 100k Population	-0.36	<b>-0.76</b>
4	Household Savings	-0.43	0.65
5	Ratio of Elderly Single-Person Households	0.20	-0.51
6	Colorectal Cancer Mortality Rate 2015	<b>0.65</b>	-0.11
7	Self-Development: Computer and Information Processing	0.08	<b>0.90</b>
8	Barrier-Free Rate 2018	<b>-0.93</b>	0.07
9	Self-Development: Arts and Culture	-0.11	<b>0.87</b>
10	Self-Development: Foreign Languages Other Than English	0.17	<b>0.82</b>

Factor 1 has high loadings on variables such as *Colorectal Cancer Mortality Rate (2015)* (0.65) and *Barrier-Free Rate (2018)* (-0.93). These variables are closely related to health outcomes and accessibility infrastructure, suggesting that Factor 1 represents a *Health and Accessibility* factor. The negative loading for the barrier-free rate indicates that improved accessibility may be associated with better health outcomes.

Factor 2 shows strong loadings for *Self-Development: Computer and Information Processing* (0.90), *Self-Development: Arts and Culture* (0.87), and *Self-Development: Foreign Languages Other Than English* (0.82). These variables are linked to personal growth and learning activities, indicating that Factor 2 represents *Self-Improvement and Educational Motivation*, focusing on educational attainment and personal development rather than direct health measures.

These results align with the insights derived from the correlation analysis in Section 4, where life expectancy was shown to be influenced by both health-related variables and factors associated with personal development and social engagement. While the correlation analysis focused on pairwise relationships between variables, Factor Analysis confirms and extends these insights by highlighting broader underlying patterns and connections among the variables.

## 7 Regression Model Estimation and Visualization

The dimension reduction results from Factor Analysis in Section 6 are not only useful for simplifying the model but also for visualizing the observations (e.g., prefectures) in a more interpretable factor space. In this section, the two factors are used as explanatory variables for both modeling and visualization, providing a deeper understanding of the factors influencing life expectancy.

### 7.1 Factor-Based Regression Model

After applying Factor Analysis, the model for  $LE$  can now be constructed using the explanatory variables  $F_1$ ,  $F_2$ , which satisfy the assumption of linear independence. This approach addresses potential issues with multicollinearity and improves the model's reliability as follows:

$$LE = \beta_0 + \beta_1 F_1 + \beta_2 F_2 + \epsilon$$

where  $\beta_0$  is the intercept, and  $\beta_1$  and  $\beta_2$  are the coefficients corresponding to Factor 1 and Factor 2, respectively.

The estimation results are summarized in Table 4. The model shows a statistically significant relationship between both factors and life expectancy, with  $F_1$  having a negative effect on life expectancy and  $F_2$  having a positive effect. It suggests that health and accessibility factors (represented by  $F_1$ ) are negatively related to life expectancy. In contrast, self-improvement and educational motivation factors (represented by  $F_2$ ) have a positive impact on life expectancy, indicating that personal development and social engagement contribute to longer lifespans.

Table 4: Regression results of Life Expectancy on Factors  $F_1$  and  $F_2$  (Males)

	Coefficient	Estimate	Std. Error	t-value	p-value
Males	Intercept	80.65	0.06	1393.83	0.00 <sup>†</sup>
	Factor $F_1$	-0.25	0.06	-4.29	0.00 <sup>†</sup>
	Factor $F_2$	0.34	0.06	5.89	0.00 <sup>†</sup>

The model explains approximately 52.7% of the variation in life expectancy ( $R^2 = 0.547$ , adjusted  $R^2 = 0.527$ ). The  $F$ -statistic ( $= 26.6$ , with a p-value of  $< 0.001$ ) indicates a statistically significant model, supporting the conclusion that the factors  $F_1$  and  $F_2$  are significantly related to life expectancy.

### 7.2 Visualization of Observations in Factor Space

As discussed in Section 3, one of the key advantages of Factor Analysis is its ability to reduce complex, high-dimensional data into a smaller set of interpretable factors. This dimensionality reduction allows for effective visualization of the data, making it easier to identify patterns and trends among the observations.

Figure 6 presents a scatter plot of the observations (prefectures) in the two-dimensional factor

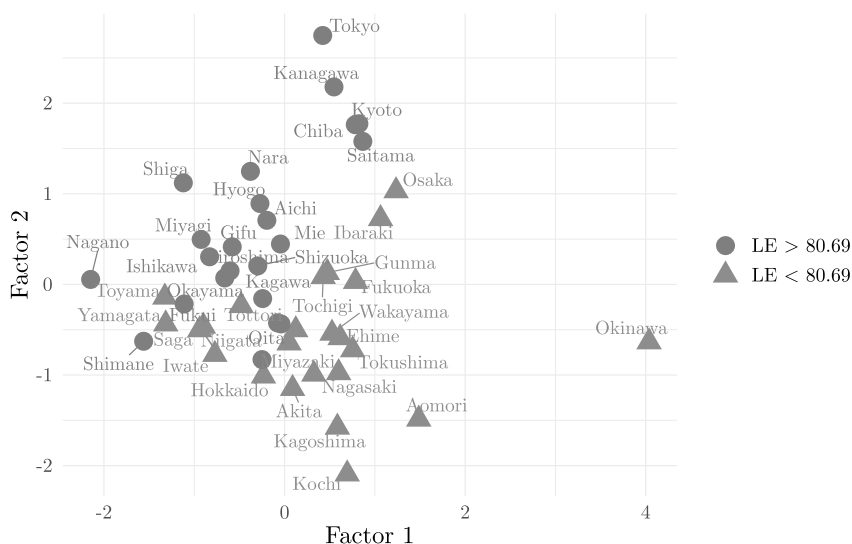


Figure 6: Scatter plot of prefectures in factor space

space defined by  $F_1$  and  $F_2$ . Prefectures are classified based on the median life expectancy ( $= 80.69$ ) as a threshold, with different symbols used to distinguish those above and below the median. This grouping of observations allows for an easy identification of trends in life expectancy.

Factor  $F_1$ , reflecting *health and accessibility*, shows a negative relationship with life expectancy. Prefectures with higher  $F_1$  values tend to have lower life expectancies, likely due to higher mortality rates from certain health conditions. In contrast, factor  $F_2$ , associated with *personal development and social engagement*, exhibits a positive relationship, where higher  $F_2$  values correspond to longer life expectancies, reflecting the benefits of education, self-development, and social activities.

In a broader interpretation,  $F_1$  can be seen as a *direct health factor*, influencing life expectancy through tangible health outcomes, while  $F_2$  acts as an *indirect health factor*, promoting Factor 1 longevity through social and educational channels that indirectly contribute to better health and well-being.

### 7.3 Exploring Differences: Comparing Shiga and Wakayama in Factor Space

By plotting the specific prefectures in factor space, a clearer understanding of the differences in factors is gained. A comparison between Shiga and Wakayama prefectures serves as a suitable case example for this type of plot, due to their notable differences in life expectancy, despite both being located in the Kansai region of Japan. According to the 2015 National Census, Shiga Prefecture

has been reported the highest male life expectancy among all prefectures in Japan, at 81.78 years, while Wakayama Prefecture shows a significantly lower average male life expectancy of 79.94 years, ranking 4th from the bottom. Table 7.3 illustrates this comparison between Shiga and Wakayama.

In the factor space, Shiga, with its higher life expectancy, is located in the upper-right region (the desirable region of good health condition and high social activity rate) of the plot, indicating high values on both  $F_1$  and  $F_2$ . In contrast, Wakayama is positioned lower, particularly along the  $F_2$  axis, reflecting lower levels of social engagement and personal development.

The plot highlights how effective visualization in Factor Analysis can bring to light key differences between specific observations by simply masking unrelated observations. At times, this small technique can greatly enhance the interpretability of the plot, allowing for more focused insights and clearer comparisons between selected data points.

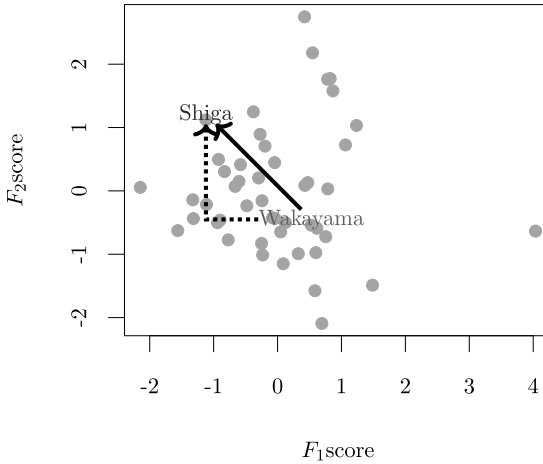


Figure 7: Scatter plot of prefectures in factor space (unrelated prefectures are masked)

## 8 Conclusion

In this paper, life expectancy data was used to illustrate the data analysis process, with particular emphasis on the careful and methodical steps required at each stage. From variable selection to Factor Analysis (FA), a structured approach was demonstrated to reveal statistically significant insights while ensuring the robustness of the model. A key takeaway is the importance of understanding the background knowledge of the data and statistical principle when conducting data analysis.

Although there are no absolute rules or manuals to follow in the steps of data analysis, even when the analysis is based on the same hypothesis, it is essential to keep in mind that approaches

which may lead to misleading outcomes should be avoided. Such outcomes are often the result of a lack of understanding or incomplete knowledge of the data and the statistical methods being applied. Ensuring a thorough grasp of both the data and methods is crucial for obtaining accurate and reliable conclusions.

## References

- Alan Agresti. *Categorical Data Analysis*. Wiley, Hoboken, NJ, USA, 3rd edition, 2013.
- Hirotsugu Akaike. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6):716–723, 1974. doi: 10.1109/TAC.1974.1100705.
- Karl R. Gabriel. The biplot graphic display of matrices with application to principal component analysis. *Biometrika*, 58(3):453–467, 1971. doi: 10.2307/2334381.
- Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, New York, NY, USA, 2nd edition, 2009.
- Arthur E. Hoerl and Robert W. Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67, 1970. doi: 10.1080/00401706.1970.10488634.
- M. H. Huh. *Exploratory Multivariate Data Analysis*. Freedom Academy, Korea, 2012.
- Richard Arnold Johnson, Dean W Wichern, et al. *Applied multivariate statistical analysis*. Pearson London, UK, 2014.