ISSN 2959-6157

Potential food factors affect the fatality rate of COVID-19: an analysis using multiple linear regression

Zihao Yang

Qilu University Of Technology
13307159560@163.com

1. Abstract

The COVID-19 pandemic has caused great harm to people around the world. This article explores the relationship between healthy diets and the fatality rate of COVID-19 from the perspective of healthy diets. This article collects data from Kaggle, pre-processes and standardizes it, conducts principal component analysis(PCA), divides it into five different dietary patterns(The contribution rate of these five dietary patterns is 72.917%, which can be interpreted as the principal component), and then constructs a multiple linear regression model. In the construction of this model, this article uses the idea of stepwise regression to remove variables that are not significant, leaving the parameters of the first and second principal components significantly non-zero. The results show that the estimated value of the parameter for the first dietary pattern in linear regression is 0.013, whereas the estimated value of the parameter for the second dietary pattern in linear regression is -0.005. Vegetable oils and vegetable products are negatively correlated with the fatality rate of COVID-19, while starchy root foods and animal products are positively correlated with it. Finally, the conclusion is drawn: In normal meals, it is more important to pay attention to the intake of vegetable foods, and try to ensure that the intake of meat foods should also be accompanied by a large amount of different types of vegetables. In the process of cooking, try to use vegetable oils such as soybean oil, olive oil, etc., which contain a lot of vitamin E and unsaturated fatty acids. For starchy root foods and certain high-protein, high-fat animal products, try to eat less in life. On the premise of ensuring good eating habits, it will have a good protective effect on COVID-19.

Keywords: healthy diets, regression model, stepwise regression

2. Introduction

Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) was first reported from the Huanan seafood market in Wuhan, Hubei province, China [1]. It has now affected over 164 million people worldwide, with nearly 3.5 million deaths reported globally as of May 18, 2021. SARS-CoV-2 has rapidly spread across the world as a result of the efficient human-to-human transmission [2]. So how to improve the immune system, that is, how to enhance the ability to resist the SARS-CoV-2s, is the key topic of this research.

Interestingly, the risk of developing SARS-CoV-2 varies between countries [3]. Not only the fatality rate, there are noticeable differences in fatality rates among countries having raised the question of whether COVID-19 is influenced by cultural factors such as nutrition and healthy dietary habits [3]. These kinds of habits are called the recommended dietary pattern. The recommended dietary pattern involves eating more plant-based food and less animal-based food [3]. For instance, it recommends eating at least 200 g of vegetables, 200 g of fruit, 15 g of unsalted nuts, 90 g of brown bread, whole-grain bread, or other whole-grain products, and taking few portions of dairy produce daily while limiting the consumption of red meat, particularly processed meat, and minimizing the consumption of sugar-containing beverages [4]. This article mainly considers alcoholic beverages, animal fats, nuts, cereals, fish, meat, starches, dairy products, sugars, fruit categories, vegetable categories, and egg categories to research the fatality rate of SARS-CoV-2.

Early in the SARS-CoV-2 pandemic, the fatality of SARS-CoV-2 was notably selective for the elderly and those with chronic diseases such as hypertension, diabetes, Alzheimer's, and cardiovascular diseases [5]. These diseases are mostly related to healthy diets in their daily life. Therefore, eating excessive animal fats makes it more likely to suffer from hypertension and diabetes, and thus the fatality of COVID-19 is stronger. However, the correlation between healthy diets and fatality and recovery

from COVID-19 is obscure. This correlation may include other potential factors, like whether these people are alcoholics or lack of fruits and vegetables, etc. Dietary patterns' influence on recovery from infectious diseases in general has been investigated in previous studies but is not well documented [6]. Very few studies have been conducted to investigate the association between diet and fatality or recovery rates in some countries. So this article has opened up a completely new perspective: linking the fatality rate of COVID-19 with people's usual daily eating habits and dividing them into several parts using principal component analysis, which is also a major new field in treating and resisting COVID-19, and also, it eliminates the influences on potential factors in some degrees. In the data set of this article, it shows the proportion of different types of food consumed by people in 170 countries, as well as fatality and recovery rates. Therefore, by studying the nutritional composition of those countries that have done a good job in epidemic prevention (countries with a lower fatality rate and a higher cure rate), the dietary structure can be modified and improved.

While countries around the world are actively researching and developing vaccines and effective drugs for COVID-19, the COVID-19 pandemic has resulted in behavioral changes in food purchasing and consumption; people panic, and overstock food supplies as they fear food insecurity [7]. So Some people started to have a healthy and balanced diet to maintain the correct nutrition status and reduce health risks [3].

3. Materials and Methods

3.1 Data collection and cleaning

This article utilizes data sourced from the Kaggle website(COVID-19 Healthy Diet Dataset (kaggle.com)), which includes information on 170 countries regarding the proportions of obesity or undernourishment related to various foods, as well as death and recovery rates. The data used in this article is from the third dataset in its dataset (Food_Supply_Quantity_kg_Data.csv). This data for different food group supply quantities, nutrition values, obesity, and undernourished percentages are obtained from Food and Agriculture Organization of the United Nations FAO website (https://www.fao.org/home/en), and data for population count for each country comes from Population Reference Bureau PRB website(https://www.prb.org/), Data for COVID-19 confirmed, deaths, recovered and active cases are obtained from Johns Hopkins Center for Systems Science and Engineering CSSE website (https://systems.jhu.edu/). The USDA Center for Nutrition Policy and Promotion Diet intake guideline information is found in ChooseMyPlate.gov (http://www.foodpyramid.com/myplate/).

To ensure that the data can be directly used in subsequent models, the data should be cleaned. First, delete the variables(Aquatic Products, Other, Miscellaneous, Offal, Oilcrops, Pulses, Spices, Stimulants, Obesity, Undernourished) that are not needed for the study. Then, in order to eliminate the impact of dimensionality on the experiment, each variable that needed to be studied should be standardized. This standardization process helps enhance the effectiveness of subsequent model building and improves the accuracy of model fitting.

3.2 Dietary patterns

Due to the complexity of handling 15 variables, SPSS statistical software(IBM SPSS Statistics 25.0) was used to conduct principal component analysis (PCA) on these variables. This method successfully reduces the correlations between the 15 variances, and correlations between the variables were computed, and the Kaiser-Meyer-Olkin (KMO) and Bartlett's tests were performed. The KMO test statistic is an indicator used to compare the simple correlation coefficients and partial correlation coefficients between variables. The rationale behind its test is that if there are indeed common factors present in the original data, the partial correlation coefficients between variables should be small. In such cases, the KMO value approaches 1, indicating that the original data is suitable for factor analysis. The chi-square approximation value of Bartlett test of sphericity refers to the chi-square statistic calculated during the testing process, which is used to measure the difference between the sample data and a spherical distribution. The 15 variables were categorized into 5 dietary patterns for study. Through the eigenvalues and eigenvectors of these five principal components, linear combinations of the 15 variables were derived for each dietary pattern. The corresponding coefficients can be calculated by dividing the component matrix by the square root of their respective eigenvalues. Finally, the magnitudes of the parameters in these linear combinations determined that the 5 principal components roughly represented the relevant factors among the 15 variables, achieving a simplification effect.

3.3 Linear regression and hypothesis testing

To explore the effects of these five dietary patterns on COVID-19, the R language(4.3.3) is used to establish a multiple linear regression model. Using these five dietary patterns as independent variables and the fatality rate of COVID-19 as the dependent variable, a multiple linear regression model is constructed. The relationship between each of these five independent variables and the COVID-19 fatality rate is determined by the positive or

negative sign of the corresponding parameter, which is a good way to identify the correlation between the independent variables and dependent variables. In order to assess the overall goodness-of-fit of the fitted model and the significance of the parameters corresponding to the five independent variables, the F-statistic and t-statistic respectively conduct hypothesis testing on the overall fit and the significance of individual parameters. When not all parameters are significant, this paper employs the method of stepwise regression to eliminate the insignificant variables, which can also mitigate the impact of multicollinearity in the model. Generally speaking, bidirectional elimination is considered one of the best approaches for stepwise regression. Its principle is roughly an improvement on the basic approach. After introducing a variable, this model should be examined whether this variable significantly alters. If there is a significant change, the t-test is conducted on all variables. If a previously introduced variable no longer shows significant change due to the addition of a subsequent variable, it is removed. This process ensures that only significant variables are included in the regression equation before introducing a new variable. The process continues until no more significant explanatory variables can be selected for inclusion in the regression equation, and no insignificant explanatory variables remain to be removed. Eventually, an optimal set of variables is obtained.

4. Results

In the dataset, it conducts descriptive statistics on 15 variables and the recovery rate and fatality rate of 162 countries (8 countries were excluded due to missing values in their variables). After standardization, a principal component analysis should be performed on the 15 variables to obtain their component matrix and the total variance explained.

4.1 Explanatory data analysis

It can be concluded from Table 1 that the recovery rate and fatality rate in this data, as well as the fifteen variable units, are all between 0 and 1, so there are no outliers or missing values in the data.

4.2 Correlation matrix and Bartlett test

As shown in Figure 1 there is a correlation between variables. For example, the correlation between the third and ninth variables is almost 1, while the correlation between the third and fifteenth variables is approximately -1, etc. Therefore, there is a strong correlation between these 15 variables, which requires further analysis through principal component analysis to eliminate their multicollinearity. In the Bartlett test, the significance is 0.000, so it can be concluded from these two aspects that the dataset is suitable for principal component analysis.

4.3 Total variance explained and component parameters

According to the PCA, these fifteen variables are divided into five principal components, as their cumulative variance contribution rate reaches 72.917, which is sufficient to represent the entire set of variables using these five principal components.

After concluding the results, the five dietary patterns(Table 2) are: eating more animal products and fewer vegetable foods; regularly using vegetable oil and not eating starchy root foods; regularly drinking alcoholic beverages, fish, and seafood, and eating fewer vegetables; eating less cereal, eating more vegetables, and eat more tree nuts and fewer fruits.

4.4 linear regression and stepwise regression

From the initial linear regression model, only the parameters of the first principal component and intercept term were significant in this regression model, so stepwise regression was used to extract the first, second, and third principal components, while the remaining variables were eliminated and then subjected to multiple linear regression with the dependent variable.

In the stepwise regression(Table 3), the overall regression effect is very good(p-value=0.0000), the first and second principal components are significantly better(p-value=0.0000,0.0173 respectively), and from the fitted image(Figure (2,3)) the first and second principal components have better effects. So the first and second principal components should be mainly studied.

The first principal and second components are dietary patterns named “eating more animal products and fewer vegetable foods” and “regularly using vegetable oil and eating starchy root foods in moderation”. It is found that when this multiple linear regression analysis is performed on it, the coefficient of the first principal component is positive, while the coefficient of the second principal component is negative. Therefore, it can be concluded that the first dietary pattern will increase the fatality of COVID-19, while the second dietary pattern will alleviate its symptoms.

5. Discussion

In the discussions of the experimental results presented in this article, there are significant consistencies with existing research. The plant-based diet (PBD) is strictly defined as consisting of “all minimally processed fruits, vegetables, whole grains, legumes, nuts and seeds, herbs and spices, and excludes all animal products, including red meat, poultry, fish, eggs and dairy products [8].” Here, PBDs will be defined as those that minimize the

consumption of animal products while prioritizing plant-based foods, particularly fruits, vegetables, whole grains, legumes, and nuts [9]. So in a case-controlled study of healthcare workers across six countries found a 73% or 59% reduction in the risk of severe COVID-19 in those consuming either a PBD or PBD/pescatarian diet [10]. And there is another research exploring the correlation between vegetables and COVID-19 recovery time, the vegetables comprised the only food group that showed a significant effect on recovery from COVID-19 in countries [11]. These items are sources of glutathione, an antioxidant tripeptide, in addition to vitamin C, mainly citrus fruits, broccoli, tomatoes, and leafy greens [12, 13]. Moreover, vegetable oils presented significant contributions in most countries, where there was a greater consumption of oils, which are mainly composed of unsaturated fatty acids (rapeseed oil, sunflower oil) [11]. There was also a large consumption of soybean and olive oils in some countries [11]. The aforementioned vegetable oils are important sources of vitamin E (a-tocopherol) [14]; soybean or rapeseed oils are particularly important sources of omega-3 unsaturated fatty acids (a-linolenic acid) [11]. Another factor is the correlation between vitamin D deficiency and the fatality of COVID-19. Vitamin D receptors are present in many immune cells and modulate the response to viral lung diseases. Vitamin D is thus an important factor in protection against infectious respiratory diseases and plays an important role in the prevention of COVID-19 [15]. So there is research searching the relationship between vitamin D and COVID-19 morbidity, severity, and mortality. That research indicates vitamin D plays an important immunomodulatory role in the innate and adaptive immune systems [16]. One of its most important functions is the downregulation of pro-inflammatory cytokines, such as interleukin (IL)-1, IL-6, IL-8, IL-12, and tumor necrosis factor (TNF)-α [17]. Moreover, vitamin D level was inversely correlated with the clinical outcomes of COVID-19, independent of inflammatory markers (e.g., IL-6 and C-reactive protein [CRP]), age, or the presence of major comorbidities such as obesity, diabetes, and hypertension. In another study, from Los Angeles, vitamin D deficiency was identified as a risk factor for positive COVID-19 tests [18]. The authors of a study from Cincinnati found some correlations between vitamin D deficiency and hospitalization, disease severity, and death among patients in primary care and specialist clinics [15]. These conclusions are in line with those of previous studies[19, 20]. Starchy root foods such as sweet potatoes and cassava usually do not contain vitamin D, especially in the nutritional content table of potato starch, where the vitamin D content is nearly 0 micrograms, this directly explains that a large amount of starchy root foods in this article can lead to vitamin D deficiency, which can lead to an increase in the fatality rate of COVID-19, which is also consistent with the conclusion of this article.

A few meta-analyses additionally reported a significant (two- to threefold) increase in fatality in people with diabetes and COVID-19 [21]. Animal products, especially their offal, and fatty tissues have a lot of fat and cholesterol, and eating too much of them can increase the risk of diabetes. In the obesity and Type 2 diabetes mellitus fields, it is proven that mouse models have proven invaluable in the basic science of the diseases by identifying the roles of inflammation, insulin resistance, fat content of the diet, pAMPK, exercise, and potential treatments [22]. In addition, and very importantly, what has been learned from the mouse models has faithfully been carried over into the human patients [23]. These physiological similarities between the two species are due to the genetic homology between the two species [24]. So this mouse model indicates potentially that animal products are bad for the cure for COVID-19.

This model clearly identified the correlation between the fatality rate of COVID-19 and healthy diets, and to eliminate the multicollinearity, PCA is used, classifying the dietary patterns and simplifying this model, which was a good way to analyze these kinds of problems. However, in this model, this analysis did not notice the impacts on other different aspects (continents, ethnicities, ages, etc).

6. Conclusion

In this article, a regression analysis was conducted on the five dietary patterns extracted from the eating habits of different countries and the corresponding COVID-19 mortality rates in these countries. It was found that eating too many animal products and starchy root products can promote the mortality rate of COVID-19 while using vegetable oils and eating more vegetable products can alleviate the symptoms of COVID-19.

7. Reference

1. Dhama, K., et al., *Coronavirus Disease 2019-COVID-19.* Clin Microbiol Rev, 2020. 33(4).

2. Sharun, K., R. Tiwari, and K. Dhama, *COVID-19 and sunlight: Impact on SARS-CoV-2 transmissibility, morbidity, and mortality.* Ann Med Surg (Lond), 2021. 66: p. 102419.

3. Alamri, F.F., et al., *Association of Healthy Diet with Recovery Time from COVID-19: Results from a Nationwide Cross-Sectional Study.* Int J Environ Res Public Health, 2021. 18(16).

4. Kromhout, D., et al., *The 2015 Dutch food-based dietary guidelines.* Eur J Clin Nutr, 2016. 70(8): p. 869-78.

5. Wang, H., et al., *The role of high cholesterol in SARS-CoV-2 infectivity.* J Biol Chem, 2023. 299(6): p. 104763.

6. van der Gaag, E., et al., *Influence of Dietary Advice Including Green Vegetables, Beef, and Whole Dairy Products on Recurrent Upper Respiratory Tract Infections in Children: A Randomized Controlled Trial.* Nutrients, 2020. 12(1).
7. Galanakis, C.M., *The Food Systems in the Era of the Coronavirus (COVID-19) Pandemic Crisis.* Foods, 2020. 9(4).
8. Ostfeld, R.J., *Definition of a plant-based diet and overview of this special issue.* J Geriatr Cardiol, 2017. 14(5): p. 315.
9. Campbell, J.L., *COVID-19: Reducing the risk via diet and lifestyle.* J Integr Med, 2023. 21(1): p. 1-16.
10. Kim, H., et al., *Plant-based diets, pescatarian diets and COVID-19 severity: a population-based case-control study in six countries.* BMJ Nutr Prev Health, 2021. 4(1): p. 257-266.
11. Cobre, A.F., et al., *Influence of foods and nutrients on COVID-19 recovery: A multivariate analysis of data from 170 countries using a generalized linear model.* Clin Nutr, 2022. 41(12): p. 3077-3084.
12. Jayawardena, R., et al., *Enhancing immunity in viral infections, with special emphasis on COVID-19: A review.* Diabetes Metab Syndr, 2020. 14(4): p. 367-382.
13. Polonikov, A., *Endogenous Deficiency of Glutathione as the Most Likely Cause of Serious Manifestations and Death in COVID-19 Patients.* ACS Infect Dis, 2020. 6(7): p. 1558-1562.
14. Maras, J.E., et al., *Intake of alpha-tocopherol is limited among US adults.* J Am Diet Assoc, 2004. 104(4): p. 567-75.
15. Skrajnowska, D., et al., *Covid 19: Diet Composition and Health.* Nutrients, 2021. 13(9).
16. Zhang, J.J., et al., *Risk and Protective Factors for COVID-19 Morbidity, Severity, and Mortality.* Clin Rev Allergy Immunol, 2023. 64(1): p. 90-107.
17. Bouillon, R., et al., *Skeletal and Extraskeletal Actions of Vitamin D: Current Evidence and Outstanding Questions*. Endocr Rev, 2019. 40(4): p. 1109-1151.
18. Chang, T.S., et al., *Prior diagnoses and medications as risk factors for COVID-19 in a Los Angeles Health System.* medRxiv, 2020.
19. Panagiotou, G., et al., *Low serum 25-hydroxyvitamin D (25[OH]D) levels in patients hospitalized with COVID-19 are associated with greater disease severity.* Clin Endocrinol (Oxf), 2020. 93(4): p. 508-511.
20. Hernández, J.L., et al., *Vitamin D Status in Hospitalized Patients with SARS-CoV-2 Infection.* J Clin Endocrinol Metab, 2021. 106(3): p. e1343-e1353.
21. Singh, A.K. and K. Khunti, *COVID-19 and Diabetes.* Annu Rev Med, 2022. 73: p. 129-147.
22. Heydemann, A., *An Overview of Murine High Fat Diet as a Model for Type 2 Diabetes Mellitus.* J Diabetes Res, 2016. 2016: p. 2902351.
23. Islam, M.S. and T. Loots du, *Experimental rodent models of type 2 diabetes: a review.* Methods Find Exp Clin Pharmacol, 2009. 31(4): p. 249-61.
24. Peltonen, L. and V.A. McKusick, *Genomics and medicine. Dissecting human disease in the postgenomic era.* Science, 2001. 291(5507): p. 1224-9.

Table 1: Explanatory data analysis of the variables studied(There is no standardization)

	Number	Minimum	Maximum	Average	Variance
AlcoholicBeverages	162	0	15.37	3.0397	5.708
Animalfats	162	0.0018	1.3559	0.221723	0.078
AnimalProducts	162	1.7391	26.8865	12.171808	35.13
CerealsExcludingBeer	162	3.4014	29.8045	11.819564	34.738
Eggs	162	0.0239	1.696	0.466422	0.113
FishSeafood	162	0.0342	8.7959	1.33766	1.434
FruitsExcludingWine	162	0.6596	19.3028	5.657533	10.372
Meat	162	0.356	8.17	3.316145	3.027
MilkExcludingButter	162	0.0963	20.8378	6.618765	25.603
StarchyRoots	162	0.6796	27.7128	5.404408	32.309
SugarSweeteners	162	0.3666	9.7259	2.797319	2.398
Treenuts	162	0	0.8	0.118	0.022
VegetableOils	162	0.0915	2.2026	0.852935	0.203
Vegetables	162	0.857	19.2995	6.047441	12.845
VegetalProducts	162	23.1132	48.2585	37.824711	35.134

Table 2: Five principal components and the parameters corresponding to each variable

Principal component coefficient	1	2	3	4	5
AlcoholicBeverages	0.21	-0.26	0.32	-0.16	0.24
Animalfats	0.32	-0.23	-0.01	-0.06	0.17
AnimalProducts	0.44	-0.06	-0.07	-0.00	-0.01
CerealsExcludingBeer	-0.26	0.22	-0.20	-0.50	0.07
Eggs	0.27	0.34	0.04	0.05	0.11
FishSeafood	0.00	0.32	0.32	0.18	0.50
FruitsExcludingWine	-0.08	0.10	0.26	0.57	-0.46
Meat	0.30	0.20	0.30	0.06	-0.00
MilkExcludingButter	0.37	-0.23	-0.27	-0.06	-0.14
StarchyRoots	-0.26	-0.39	0.20	0.13	0.23
SugarSweeteners	0.18	0.32	0.30	-0.19	-0.29
Treenuts	0.05	0.22	-0.29	0.33	0.52
VegetableOils	-0.01	0.39	0.03	-0.34	-0.03
Vegetables	0.05	0.21	-0.54	0.29	-0.09
VegetalProducts	-0.44	0.06	0.07	0.00	0.01
AlcoholicBeverages	0.21	-0.26	0.32	-0.16	0.24

Table 3: Stepwise regression

	coefficient	standard deviation	T-value	P-value
Intercept	0,038922	0.003053	12.748	<2e-16
Factor1	0.012983	0.001385	9.374	<2e-16
Factor2	-0.005307	0.002205	-2.407	0.0173
Factor3	-0.003520	0.002405	-1.464	0.1452

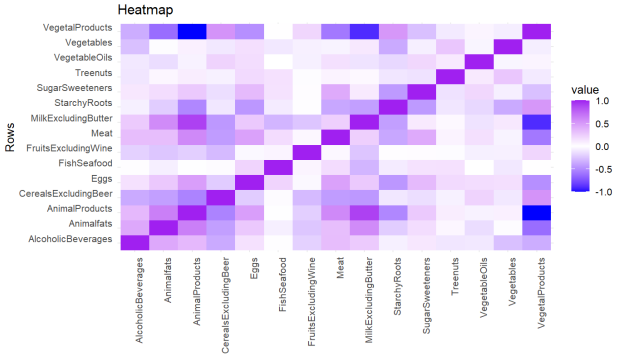

Figure 1:Heatmap for the 15 variances

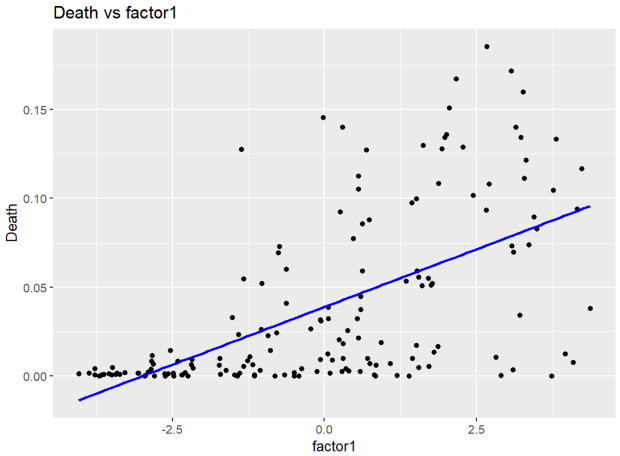

Figure 2: The fitted image of stepwise regression(regard factor1 as independent variance)

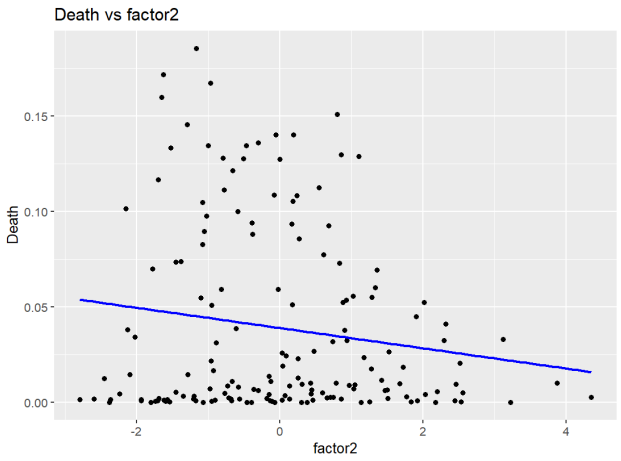

Figure 3: The fitted image of stepwise regression(regard factor2 as independent variance)