Factors influence health expenditure

Yihang Luo

December 19, 2022

Introduction

Health expenditure has always been a large part of household expenses. With the advance of technology in recent years, families have been required to spend more and more on health. Beside the basic health insurance covered by part of the salary and the expenses given to the government for medical care, additional costs may be incurred to pay clinicians or for unexpected illnesses. As a student from a middle-class background with high personal health conscious, I am more than aware of how the high cost of a serious illness or accident can tear a family with a normal income apart. Also, the article Household catastrophic health expenditure: a multicountry analysis[1] discovered that the proportion of households facing catastrophic payments from out-of-pocket health expenses varied widely between countries, which is why the topic is a great concern to lots of families around the world. Thus, studying the factors that influence spending on health can improve the awareness of people related to health, may to some extent help prevent the excessive health expenditure by families who are not financially well off.

The related journal, the determinants and effects of health expenditure in developed countries[2], re-examined the results of previous work using a larger sample of 560 pooled time-series and cross-section observations. They confirmed the importance of GDP as a determinant of health spending and the importance of some non-income variables such as demographic structure, which let me reassure the inclusion of GDP variables and the try to add some demographic structure factors such as adult mortality, unemployment rate, number of passed abortion laws and education level. Besides, we add life expectancy, alcohol level, BMI as variables related to physical factors. We will use all the variables to perform a linear regression model that predict people's spending on health.

Methods

Our study is used to the best linear regression model for predicting spending on health, including predictors like GDO, life expectancy. adult mortality, etc. All the variables come from Kaggle data for studying life expectancy among 16 years from 2000 to 2015 and unemployment rate in 2007, deriving from WHO data, UNdata and World Bank data[3][4][5][6], including 134 countries as observations. The predictors come from both personal health conditions and demographic structure among varies countries so people and government can focus more on these factors in order to decrease the expenditure on health.

At the beginning, we split the data into two parts: a training dataset containing 70% collected data and a testing dataset containing the rest 30% collected data. We set the seed to 1006871100 to ensure that the training and testing data are always consistent. We then use the training data set to build a full model consisting of all possible predictors: Condition 1 is satisfied if there is a functional pattern in the scatterplot between spending on health and predicted health spending. Condition 2 is satisfied if there is a linear or random pattern in all the scatterplots between numerical predictors. When the models satisfy both conditions, we create residual plots of the fitted response variables, numerical predictors, and normal QQ plots to check the four assumptions.The linearity assumption holds if the residual plots has a random pattern. The independence assumption holds if there is no points significantly aggregated in the residual plot.The homoscedasticity assumption holds if the points in the residual plot are randomly distributed. The normality assumption holds if the normal QQ plot shows that the residuals do not deviate significantly from the standard line. If neither model violates the assumptions, we will proceed with model selection. We will apply the Boxcox transformation to deal with these violations if the models have some violations of the model assumptions. After the transformation, we will compare the two original full models with the transformed model. We will choose the model that violates the smallest amount of assumptions to continue as the full model if only few violations in the converted model are made.

By selecting a model from the previous steps, we could proceed with model selection. Multicollinearity is an important assumption to check the full model, through using Variance Inflation Factor(VIF) for each predictor. When there are predictors with VIF greater than 5, we will keep eliminating the predictor with the largest VIF each time we build a new model, and then check for the other predictors until all predictors have VIF less than 5. Then, we will do manual selection by looking at the summary of the full model to see the significance of the coefficients. Subsequently, we do a partial F-test to check whether the reduced model was better than the full model. We would like to choose one of full model and reduce model as our final model. Then we used the test dataset to fit these preferred models. We compared examine the changes in coefficients, importance of predictors, model assumptions and adjusted R2, AIC, BIC, etc. between testing and training dataset. A model was validated if it looked very similar to its performance in the training dataset. We preferred the model with larger adjusted R^2 and p-values and smaller AIC and BIC values. So we can select the final model based on above. After deciding on the preferred model, we will examine the leverage points, outlier points, and impact points. The impact points will be measured by all three truncation points, i.e., Cook distance, DFFITS, and DFBETAS. if there is no background reason, we will proceed with model validation and remove the problematic observations. Finally, we rechecked the conditions and assumptions of the reduced model through residual plots.

Result

- Training dataset

Variable	Mean	Min	Max	Standard Deviation
Response(percentage health expenditure)	5.5366355	2.1496001	9.0736896	1.6545641
Predictor 1(GDP)	7.9139777	4.9750606	10.9460142	1.532334
Predictor 2(life expectancy)	69.0962799	45.7866667	82.5266667	9.3778959
Predictor 3(adult mortality)	4.9212011	3.8080504	6.1544334	0.5743727
Predictor 4(alcohol)	4.6925881	0.014375	13.4973333	3.8872274
Predictor 5(BMI)	38.5294179	14.2833333	65.9533333	16.0028044
Predictor 6(schooling)	12.0106401	4.9	18.86875	3.1197771
Predictor 7(unemployment rate)	6.4754787	0.58	15.948125	3.769241

- Testing dataset

Variable	Mean	Min	Max	Standard Deviation
Response(percentage health expenditure)	5.7046652	2.6159009	9.2548609	1.9132807
Predictor 1(GDP)	7.9851592	5.4789907	11.0118479	1.6499475
Predictor 2(life expectancy)	70.9948438	48.2466667	82.46	9.5667975
Predictor 3(adult mortality)	4.8164006	3.9902179	5.872681	0.5657321
Predictor 4(alcohol)	4.740876	0.01	12.236	4.0049498
Predictor 5(BMI)	37.3489792	12.5133333	58.16	16.8792303
Predictor 6(schooling)	12.3935521	3.9266667	20.0133333	3.5319073
Predictor 7(unemployment rate)	6.7803594	0.705625	15.701875	3.8597268

Table 1: All numerical variables summary for training and testing dataset

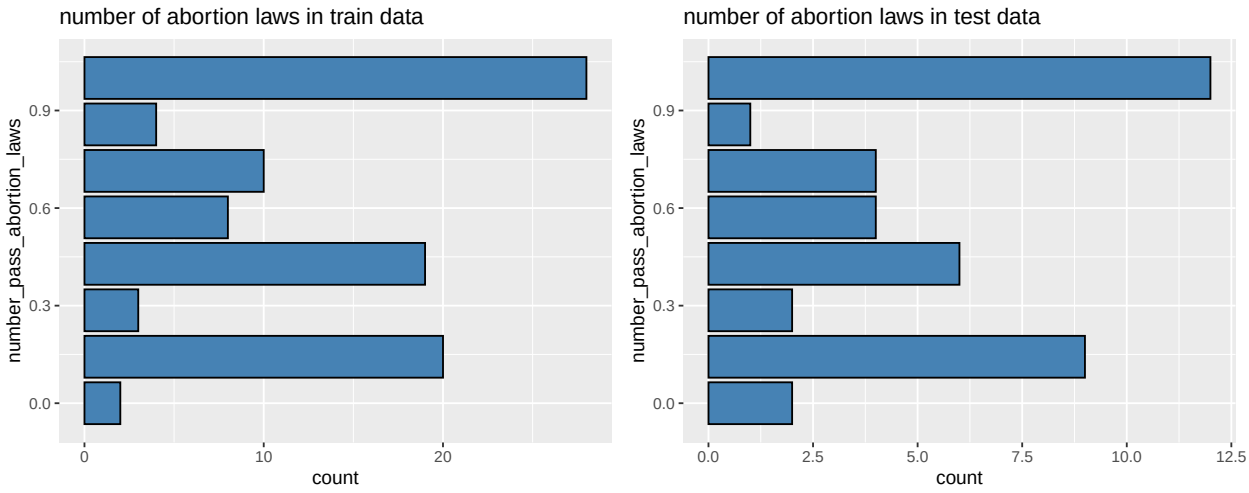

Figure 1: Boxplot of categorical variables in training and testing datadset

There are 1034 data in training dataset and 440 data in testing data. We can see that the distributions of numerical variables in both training and testing datasets are similar from table 1, where the distribution of adult mortality and alcohol are highly skewed to the right, and the distribution of life expectancy and BMI are highly skewed to the left. We can see more clearly through figure 2 in appendix. Figures 1 show that the distributions of categorical variable in both datasets are similar. Also, the introduction of each variables are displayed in table 4 in appendix.

We observe that the data are well clustered around the diagonal in the plot of responses versus fitted values, and all paired predictors appear to be linear in terms of the scatter plot of all paired predictors. This means that the full model in both training and testing dataset satisfy condition 1 and 2. There is no clear linear pattern in the residual plot, while the normal QQ plot has a clear deviation on both sides, denoting that the Normality assumption does not hold and other three assumptions holds. So we transformed life expectancy, alcohol, adult mortality and unemployment rate by Boxplot transformation. However, the transformed model has the VIF exceeds 5, which means it violates the multicollinearity assumption. The two conditions still hold, while the Normality assumption is still violated but the points are less deviated from the line. We then remove the highest VIF variable and find the new full model after transformation. The new transformation model still satisfy 2 conditions and all assumptions except Normality. Subsequently, we manually reduce the model by removing the variables such as BMI, unemployment rate and adult mortality that have large p-value. Also, condition 1 and 2 still holds for the manual reduce model, and the assumption of it satisfies and fits better on Normality compared to the full model as QQ-plot does not deviate too much. Thus, as we can see from table 3, we can compare only the reduce model in train and test data through adjusted r, AIC, BIC, leverage points, outlier and influential points to find a better model. They have neither the outlier and influential points. Additionally, we find that the the manual reduce model is better than the full model after transformation with passed anova test, larger adjusted r and smaller AIC and BIC values. The validation for test data of full model and manual reduce model shows similar identity where manual reduce model in test data has passed anova test, larger adjusted r and smaller AIC, BIC compared to full model after same transformation in test data. Both the full model and manual reduce model satisfy 2 conditions and three assumptions, and neither of them has outlier nor influential points.

Model	Adjusted R square	AIC	BIC	Leverage points	Outliers	Influential points	Annova test (with full model)
Full model in train data	0.9514447	86.70965	109.5993	2 39 113 13 31 49	None	None	
reduce manual model in train data	0.9485602	84.43645	99.69622	2 39 13 31	None	None	p-value = 0.33
Full model in test data	0.9299746	68.139	83.33891	121 35	None	None	
reduce manual model in test data	0.9324359	64.29222	74.4255	27 93 121 9 29 35	None	None	p-value = 0.6261

Table 2: Comparison of full model and fitted reduce model in train and test data

Discussion

As a result, we would choose manual reduce model for our final model. There is no new assumption violations in the final model. From the summary table of the final model in appendix, we also know that all predictors are significant. Thus the final model is supposed to be validated. The linear regression equation of the final model looks like this:

$$\hat{y} = -2.58199 + 0.91102 X_{GDP} + 0.11045 X_{Alcohol_{trans}} + 0.07444 X_{Schooling} - 0.33909 X_{numberofpassedabortionlaws}$$

We would expect 0.91102 increase in health expenditure for one unit increase in GDP, fixing all other variables. We would expect 0.11045 increase in health expenditure for one unit increase in alcohol level, fixing all other variables. We would expect 0.07444 increase in health expenditure for one unit increase in year of schooling, fixing all other variables. We would expect 0.33909 decrease in health expenditure for one more passed abortion law in the area. The findings explain how these four variables affect the spending on health, and have other possible predictors present.

There are limitations exist in the research. We acknowledge that there are extreme values such as leverage points in the data that affect the stability and accuracy of the experimental results, and the Normality assumption is barely satisfied. The conclusions drawn with and without extreme values may be different. In addition, the transformation of variables makes the model difficult to interpret, but without the transformation, the violation of assumptions becomes worse. Finally, it seems contrary to our common sense that more passage of the abortion laws in the summary would lead to a decline in health spending.

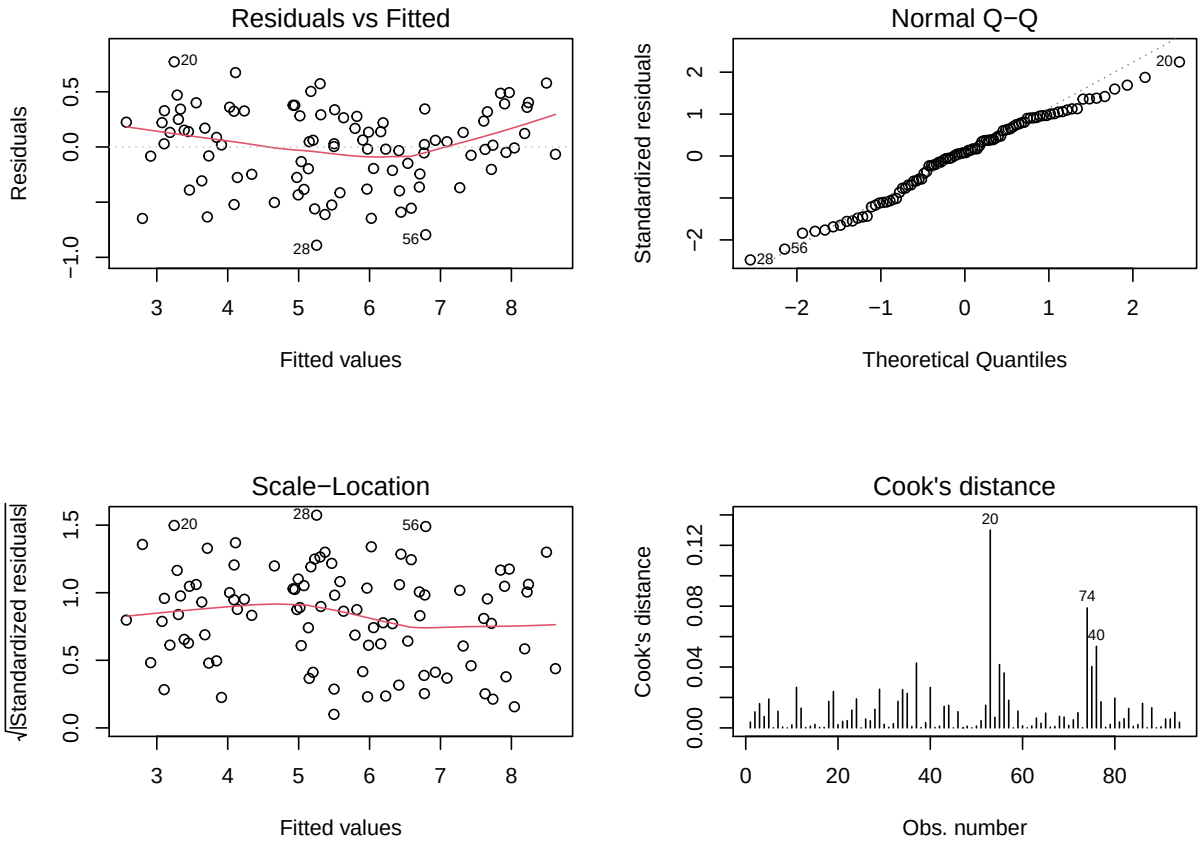

Figure 2: Check four assumptions for final model

All analysis for this report was programmed using `R version 4.0.2`.

Bibliography(APA format)

1. Xu, K., Evans, D. B., Kawabata, K., Zeramdini, R., Klavus, J., & Murray, C. J. L. (2003). Household catastrophic health expenditure: a multicountry analysis. The Lancet, 362(9378), 111–117. doi: 10.1016/S0140-6736(03)13861-5

2. Hitiris, T., & Posnett, J. (1992). The determinants and effects of health expenditure in developed countries. Journal of Health Economics, 11(2), 173–181. doi:10.1016/0167-6296(92)90033-W

3. kaggle datasets download -d kumarajarshi/life-expectancy-who [https://www.kaggle.com/datasets/kumarajarshi/life-expectancy-who]

4. © Copyright World Health Organization (WHO), 2021. All Rights Reserved. [https://www.who.int/data]

5. kaggle datasets download -d pantanjali/unemployment-dataset [https://www.kaggle.com/datasets/pantanjali/unemployment-dataset]

6. Copyright © 2022 - United Nations Statistics Division Version [http://data.un.org/Data.aspx?q=abortion&d=GenderStat&f=inID%3a11]

Appendix

Variable	Variable_Type	Description
GDP	numerical	Gross domestic product
Life.expectancy	numerical	life expectancy
Adult.Mortality	numerical	mortality of adult
Alcohol	numerical	alcohol level
BMI	numerical	Body mass index
Schooling	numerical	year of schooling
number_pass_abortion_laws	categorical	number of passed abortion laws in the area
unemployment_rate	numerical	unemployment rate

Table 3: Introduction to variables

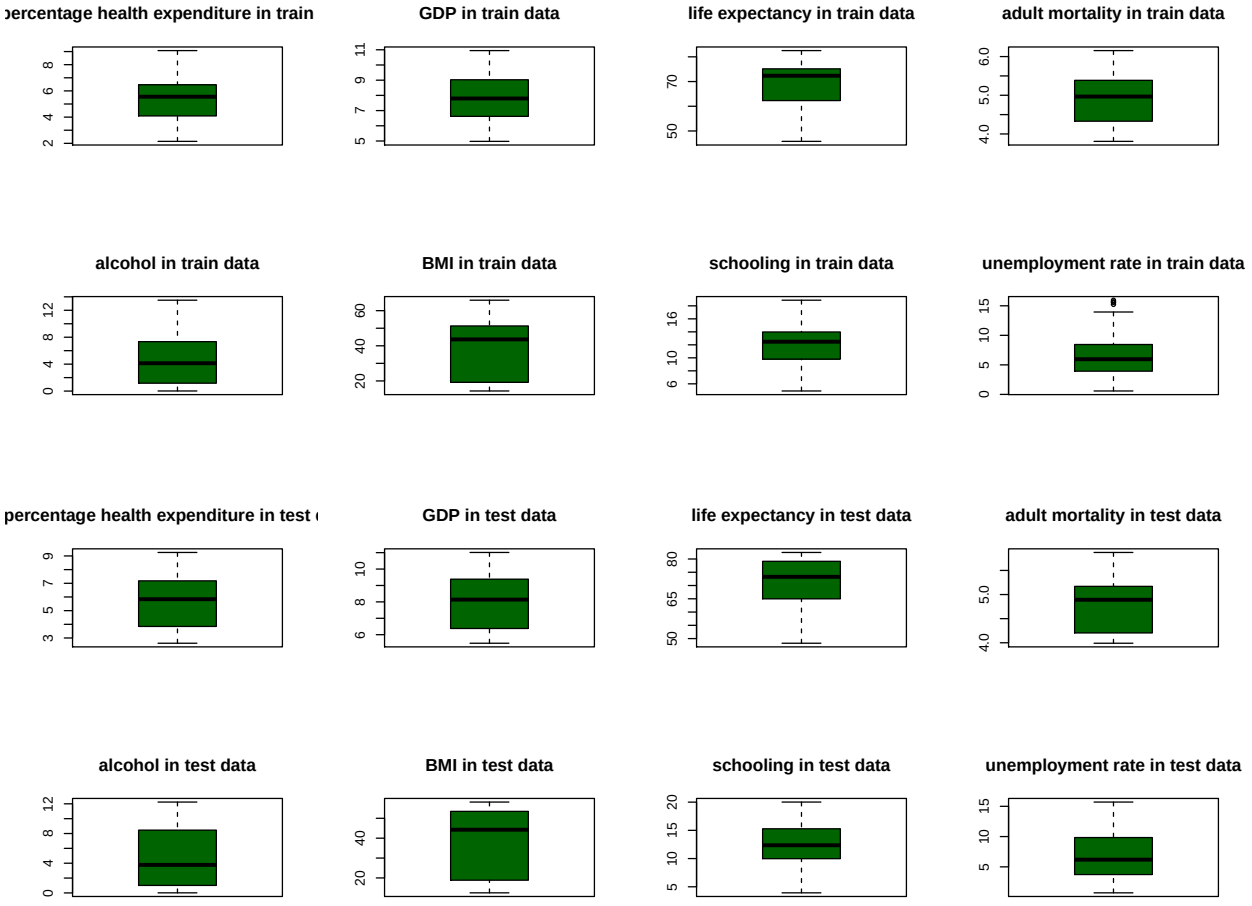

Figure 3: Boxplot of numerical variables in training and testing datadset

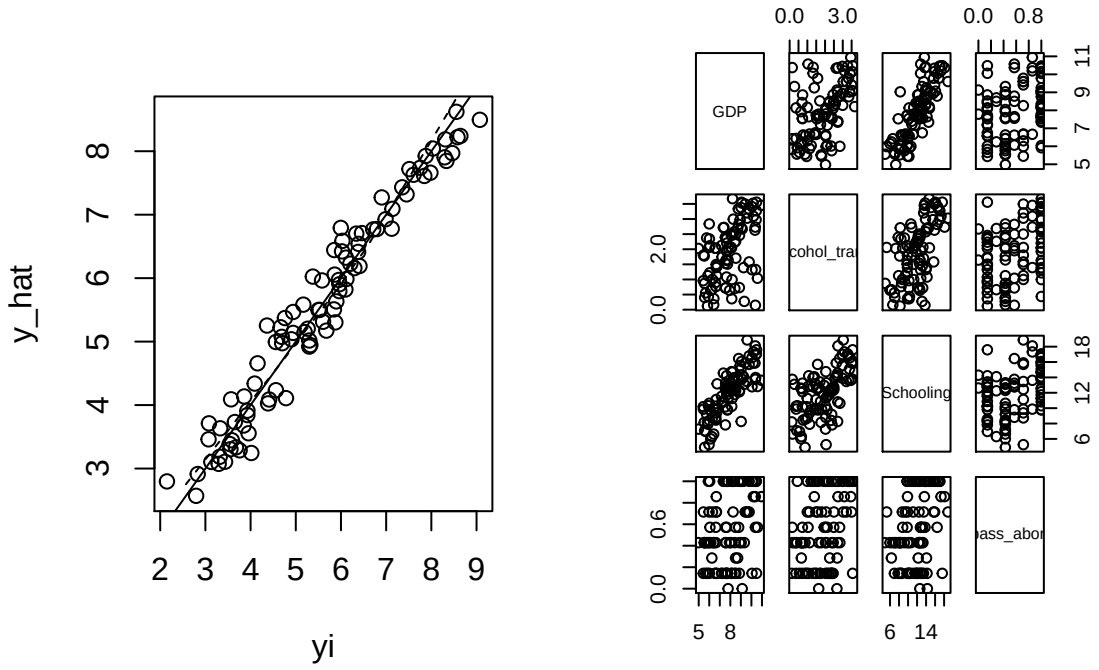

Figure 4: Check Condition 1 and 2 for final model

```
##
## Call:
## lm(formula = percentage.expenditure ~ GDP + Alcohol_trans + Schooling +
##     number_pass_abortion_laws, data = train_trans)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.89014 -0.24717  0.02895  0.28071  0.77154
##
## Coefficients:
##                             Estimate Std. Error t value Pr(>|t|)
## (Intercept)                -2.58199    0.20225 -12.766  < 2e-16 ***
## GDP                         0.91102    0.04338  21.000  < 2e-16 ***
## Alcohol_trans               0.11045    0.04903   2.253  0.02674 *
## Schooling                   0.07444    0.02348   3.170  0.00209 **
## number_pass_abortion_laws -0.33909    0.12466  -2.720  0.00785 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3656 on 89 degrees of freedom
## Multiple R-squared:  0.9533, Adjusted R-squared:  0.9512
## F-statistic:   454 on 4 and 89 DF,  p-value: < 2.2e-16
```

Summary of final model