The Impact of Early-Life Social-Economic Status, Lifestyle, and Personality on the Later-Life Health Condition

Duan Ruxiao 3035534232
Fu Yuming 3035534335
Hua Zhisheng 3035533915
Yang Yuezhi 3035603033

1. Introduction

1.1 Background

Contemporarily, health has gradually become the focus of attention for the public, and it has become more and more critical to explore the factors that potentially influence health conditions. One popular direction is to investigate factors that people had when they were at a young age, which will potentially affect their health in old age.

Many investigations have been done in this direction. Some researchers in NCBI found that early-life socioeconomic status has a positive correlation with the later-life state of physical activity. That is, the better the early-life socioeconomic status, the higher the frequency of physical activity in later years [1]. A group of medical personnel affiliated with Northwestern University researched the relationship between adolescents' lifestyles and their lung function in old age. They found that people's early-life smoking frequency, drinking frequency, and body mass index seriously affected the probability of having cardiovascular disease in their 40s [2]. People's education level at young ages, family background, and employment industry have also been confirmed to have more than 30% impact on health at older ages [3].

Motivated by many similar researches, this study further explores how people's early socioeconomic status, lifestyle, and personality will affect their health in later life, which is characterized by the universality of screening for influencing factors and the long span of longitudinal research. Unlike [1], which uses the frequency of physical activity in old age as the response, this study treats health index as the response due to its comprehensive and quantifiability. This study improves the research span of only 20 years in [2] and turns it into a survey that tracks respondents for up to 54 years. In addition, compared to [3], this study has more emotional considerations and adds corresponding variables such as the Big Five personality traits.

1.2 Objectives

1.2.1 Research Target

This study aims to use different statistical learning models to identify several early-life predictors with the most significant impact on the health of the elderly, to choose the best model for predicting the health index.

1.2.2 Potential Usage

By discovering potential factors at young ages that may improve or impair health conditions in old age, this study could alert young people to develop proper habits and remind people to pay attention to physical and mental health at any stage of life. In addition to predicting one young person's elderly health, this study could also provide parents with advice on the discipline of their children and help young people's life planning. At the same time, this study can offer reference opinions to government policies. If the government wants to improve the national health index of the elderly and realize a virtuous circle, it should take measures such as increasing national education investment, restricting tobacco production and sales, and rationally setting up psychiatrists in various institutions according to the result of this study. Besides, by inferencing models of this study, the relationship of later-life health condition to some early-life factors could be discovered, which could be further investigated by later study to find out the reasons behind it.

2. Data Source & Data Statistics

2.1 Predictors Selection Procedure

We select a reasonable amount of predictors in respondent's early-life (the 1950s - 1970s), which could typically represent their social-economic status and lifestyle then choose a general health indicator to summarize respondents' health conditions in their later life. The selection of predictors is suffered from the dilemma that, on the one hand, an excessive number of predictors would result in a sharp decline in the number of observations available because more respondents would not answer all the predictors we select as the number of predictor increase. As a result, they would be excluded from our final data set, resulting in potential risks of high variance for some statistical learning tools we apply due to small observations. On the other hand, we are aware that a relatively small amounts of predictors may result in an inconclusive summary of the respondent's social-economic status and lifestyle in early life, which would result in significant biases in prediction and meaningless inference of the model. After several trials and errors, we finally decide to set the number of predictors to 25 with 3005 observations to balance this dilemma.

2.2 Data Statistics

2.2.1 Classification of Predictors

The 25 predictors selected provide different kinds of information and we arrange them into different categories, which includes:

(1) Parents' information: parents' level of education, parental income (1957), father's occupation (1957);
(2) General information: the measure of IQ score (1957), the total number of children;
(3) Personality: total scores for the Five-Factor Model of Personality Structure (1992-1993);
(4) Educational background: total years of schooling;
(5) Lifestyle: the average number of cigarettes smoked per day when smoking regularly, the average number of alcoholic drinks per day, frequency of going together with friends (2003-2005), frequency of participating in particular leisure time activities (2003-2005), frequency of participating in vigorous physical exercise (1992-1993), frequency of religious attendance (1992-1993);
(6) physical condition: body mass index (1992-1993); total number of times giving blood for use by others over the lifetime;
(7) psychological well-being: the summary score for psychological distress/depression - modified CES-D (1992-1993), frequency of feeling negative emotions (2011);
(8) job and income: categories of current or last job (1975), occupational income score for current or last job (1975), hours of work per week (1975).

2.2.2 Labels for different values of categorical variables

The meaning of each class for all 5 different categorical variables among 25 predictors are listed below.
(1) Father's occupation in 1957:
0 = "Farmers, farm managers";
1 = "Laborers including farm laborers";
2 = "Private household workers and service workers";
3 = "Operatives and kindred workers";
4 = "Craftsmen, foremen, and kindred workers";
5 = "Clerical, sales and kindred workers";
6 = "Proprietors";

7 = "Managers and officials--salaried";
8 = "Professional, technical and kindred workers";
9 = "Student, housewife, unemployed, or others";
(2) Categories of current or last job in 1975:
1 = "Private company, business or individual for wages, salary or commission";
2 = "Government employee (federal, state or local government)";
3 = "Own business or professional practice; incorporated";
4 = "Own business or professional practice; not incorporated";
5 = "Working without pay in a family business or farm";
(3) Frequency of feeling negative emotions in 2011:
1 = "Rarely";
2 = "Occasionally";
3 = "Often";
4 = "Almost always";
(4) Frequency of participating in vigorous physical exercise during 1992-1993:
1 = "Three or more times per week";
2 = "Once or twice per week ";
3 = "About one to three times per month";
4 = "Less than once per month ";
(5) The total number of times giving blood for use by others over the lifetime:
1-4 = "Corresponding number of times of blood donation";
5 = "Five or more times".

2.2.3 Response

We choose respondents' Health Utilities Index Mark 3 summary score collected in 2011 as the response to represent respondents' general health status in their old age. It provides a compact but comprehensive frameworks within which to describe health status [4]. The health index ranges from -1 to 1 with 1 to describe the best health condition and -2 to describe the worst health condition.

The distribution of the health index of all the observations is shown on the figure below (Fig. 1).

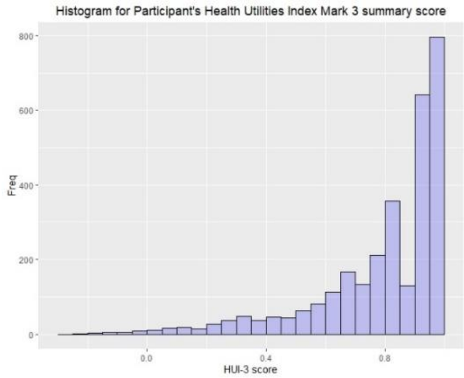

Figure 1. Histogram for Participants' Health Utilities Index Mark 3 Summary Score

3. Data Processing

Before feeding data into models, some preliminary data processing jobs are done to transform the raw data into a desirable data form. Firstly, we summarize some predictors into one predictor to lower the dimension of the dataset and prevent from creating an overwhelming number of dummy variables for qualitative predictors in the model. For example, we merge the predictors about respondents' alcohol drinking status to the predictors about respondents' frequency of drinking alcohol by marking those who do not drink at all to the frequency of 0 times per week. Thus, only the frequency predictors would appear in the dataset after processing. Similar procedures are also done in other places, such as for the smoking frequency. Secondly, we delete the observations with missing values and values of no interest, like those marking respondents, refused to answer, or the questions not applicable to the respondent. Besides, observations with that out of scope of this research like respondent who

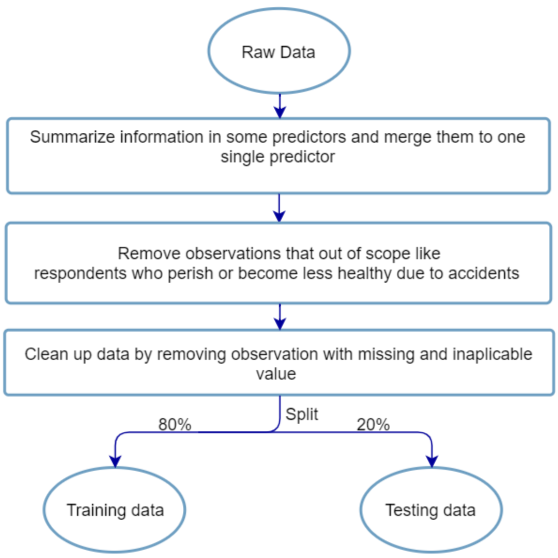

Figure 2. Data Processing Procedure

perish or whose health index decreases due to accidents are also excluded from the data. Lastly, data are split into a training dataset and a testing dataset randomly in a ratio of 8:2, where training dataset are fed to train different models while the testing dataset is used to test the model and generate testing MSE. Noticeably, Python instead of R is used to conduct all the data processing jobs due to its flexibility. A flow chart summary of procedure described above are shown above (Fig. 2).

4. Different Models

4.1 Lasso

4.1.1 Lasso Analytic Steps

The Lasso model finds coefficients that minimize the below equation. It could also perform predictor selection by making the unrelated predictor with coefficients equal to zero. The λ, tuning parameter controls the level of shrinkage and determines the flexibility of the model.

$$\sum_{i=1}^{n}\left(y_i-\beta_0-\sum_{j=1}^{p}\beta_j x_{ij}\right)^2+\lambda\sum_{j=1}^{p}|\beta_j| = \mathrm{RSS}+\lambda\sum_{j=1}^{p}|\beta_j|.$$

After separating the data into a training set and a test set, we would like to select a suitable tuning parameter λ to build the Lasso model. One of the most efficient ways to do so is to apply cross-validation by using the built-in function cv.glmnet() in R.

The best λ (which has got the lowest MSE in cross-validation) is approximately 0.0025 (Fig. 3).

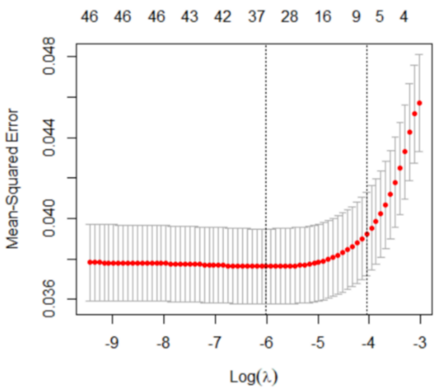

Figure 3. plot of the cross-validation error as a function of λ

Then, we can complete our function by applying the best λ in it. After all the coefficients are computed, we can go on to see the prediction of the model on the testing data. Finally, we get a testing MSE of about 0.0495.

4.1.2 Lasso Results

The most important predictors are vigorous physical exercise frequency (4), respondent's job (3, 4, 5), father's job (2) and depression frequency.

Therefore, according to the Lasso model results (Fig. 4), it can be concluded that people with lower depression frequency tend to have a healthier physical state. Moreover, for respondents whose job is "own business or professional practice (incorporated)" or "working without pay in a family business or farm", and whose father's job is "private household workers and service workers", their HUI scores are generally higher than others, while those whose job is "own business or professional practice (not incorporated)" and who participate in vigorous physical exercise or sports less than once per month tend to have lower HUI scores.

```
R_job1          -8.565527e-04
R_job2           .
R_job3           1.518855e-02
R_job4          -1.893026e-02
R_job5           1.971284e-02
work_hour        8.976737e-06
f_edu            .
m_edu            3.440658e-03
edu_year         1.650482e-03
par_income       3.740070e-05
R_income         2.436241e-05
f_job1           .
f_job2           1.613591e-02
f_job3          -8.603469e-04
f_job4           .
f_job5           .
f_job6          -2.869962e-04
f_job7           .
f_job8           1.847397e-03
f_job9           9.916472e-03
IQ               3.522707e-04
kids_no          .
Smoke           -5.510682e-03
Alco             6.063649e-03
Depress_index   -1.239703e-03
together_freq    .
BMI             -8.825946e-03
depress_freq1   -1.865533e-02
depress_freq2   -1.010430e-01
depress_freq3   -2.928470e-01
depress_freq4   -1.849092e-01
vigorous_ex1     .
vigorous_ex2     .
vigorous_ex3     1.014298e-02
vigorous_ex4    -2.330854e-02
play_freq        .
religion        -1.999128e-05
blood1           6.793347e-03
blood2           .
blood3           .
blood4          -1.017397e-02
blood5           .
extraversion     1.872302e-03
openness         1.475945e-03
neuroticism     -4.471503e-03
conscientious    2.403867e-03
agreeable        .
```

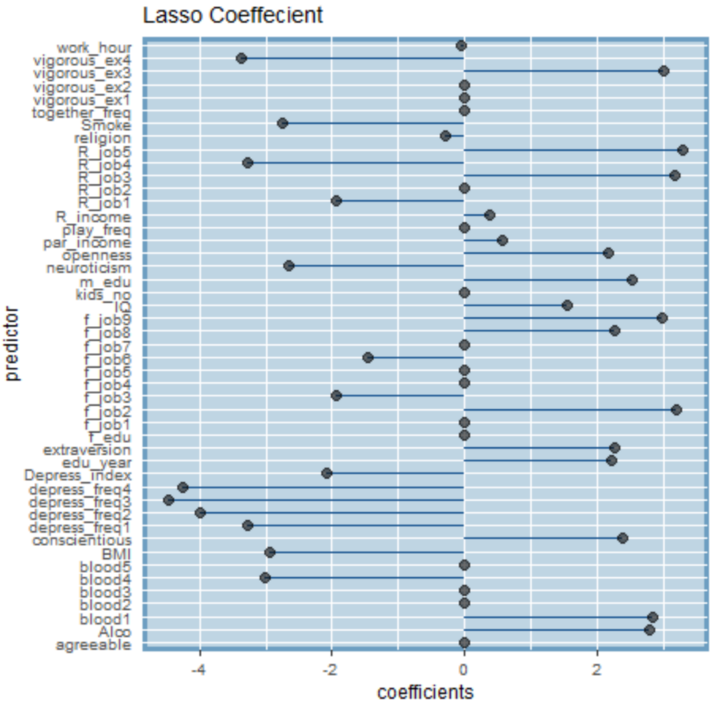

Figure 4. Resulting Coefficients estimates from the lasso model

4.2 Smoothing Splines

4.2.1 Smoothing Splines Analytic Steps

Smoothing Splines is a non-linear statistical learning tool whose goal is to find the estimating function g that minimizes equation, where the first term is known for loss

$$\sum_{i=1}^{n}(y_i - g(x_i))^2 + \lambda \int g''(t)^2 dt$$

function that encourages g to fit the data well, and the second term is a penalty term which penalizes the variability of g. Its flexibility is determined by λ, the tuning parameter in the second term, which represents the relative importance of the second term to the first term. As λ approaches 0, the smoothing spline converges to the interpolating spline while as λ approaches infinity, the second term becomes paramount, and the estimate converges to a linear estimate. The result function g produced would be a natural spline where λ controls the level of shrinkage.

Leave-one-out cross-validation is performed for every quantitative predictor in the training set to determine the suitable λ value and its corresponding degree of freedom. Then, all the predictors are fitted in the generalized additive models (GAM) so that the partial contribution of each predictor to health index is calculated separately, and then all of their contributions are added together. For qualitative variables, dummy variables are created, while for quantitative variables, their smoothing splines with suitable λ value are used in the model.

4.2.2 Smoothing Splines Result

The hypothesis F-test results of the resultant GAM model are shown below (Fig. 5).

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
s(work_hour, df = 2.078519)	1.0	0.079	0.0790	2.1928	0.1387984	
s(f_edu, df = 5.437747)	1.0	0.703	0.7025	19.5060	1.050e-05	***
s(m_edu, df = 7.778929)	1.0	0.290	0.2900	8.0515	0.0045868	**
s(edu_year, df = 4.20491)	1.0	0.251	0.2507	6.9601	0.0083914	**
s(par_income, df = 2.000552)	1.0	0.070	0.0700	1.9438	0.1633914	
s(R_income, df = 3.303216)	1.0	0.095	0.0953	2.6468	0.1038958	
s(IQ, df = 2.000552)	1.0	0.079	0.0788	2.1880	0.1392258	
s(kids_no, df = 15.893626)	1.0	0.003	0.0026	0.0732	0.7867114	
s(Smoke, df = 3.663747)	1.0	0.281	0.2808	7.7957	0.0052807	**
s(Alco, df = 2.000552)	1.0	0.466	0.4657	12.9307	0.0003301	***
s(Depress_index, df = 3.753701)	1.0	4.815	4.8152	133.7004	< 2.2e-16	***
s(together_freq, df = 10.018191)	1.0	0.087	0.0873	2.4229	0.1197128	
s(BMI, df = 8.017229)	1.0	4.111	4.1106	114.1370	< 2.2e-16	***
s(play_freq, df = 9.300598)	1.0	0.005	0.0045	0.1259	0.7227965	
s(religion, df = 2.000552)	1.0	0.046	0.0458	1.2722	0.2594672	
s(extraversion, df = 2.836712)	1.0	0.096	0.0959	2.6620	0.1029108	
s(openness, df = 2.484638)	1.0	0.029	0.0288	0.8003	0.3711110	
s(neuroticism, df = 6.362093)	1.0	0.792	0.7921	21.9931	2.898e-06	***
s(conscientious, df = 4.293779)	1.0	0.093	0.0930	2.5836	0.1081163	
s(agreeable, df = 3.066024)	1.0	0.034	0.0342	0.9496	0.3299370	
blood	5.0	0.134	0.0269	0.7465	0.5886404	
depress_freq	5.0	7.860	1.5719	43.6466	< 2.2e-16	***
vigorous_ex	3.0	0.370	0.1234	3.4268	0.0164846	*
f_job	9.0	0.169	0.0188	0.5228	0.8589883	
R_job	5.0	0.238	0.0477	1.3245	0.2507140	
Residuals	2281.4	82.164	0.0360			

Figure 5. F-test Results from the resultant GAM model

The result shows that respondents' BMI index and predictor associated with depression (i.e. the depressed frequency and depression index) are the most significant predictors to the health index response. The partial effect of BMI on health index against the BMI value is plotted below (Fig. 6).

It could be concluded that BMI is positively correlated with health index when BMI is smaller than 22 while it is negatively correlated with health index when BMI is bigger than 22. The abnormal upwards in the curve when BMI is around 40 could be explained by the fact that only less than 15 observations have a BMI over 40. Fitting smoothing splines to such a small amount of data could result in a large variance. The significant correlations between health index and depression level and depression frequency (Fig. 6) also show that depression in early and middle life could influence respondents' health conditions in later life. In particular, the more depressed with a larger depressed index, the worse the health condition. Again, due to the small number of observations available with depression index larger than 75, the prediction yields large variance resulting in abnormal curve shape after depression index bigger than 75.

Besides, respondents' early life family education background, summarized by the predictor of father and mother' years of education and the respondent's year of education, is also positively correlated to health condition in later life (Fig. 6). The more educated the family is, the more healthy respondents would be in later life.

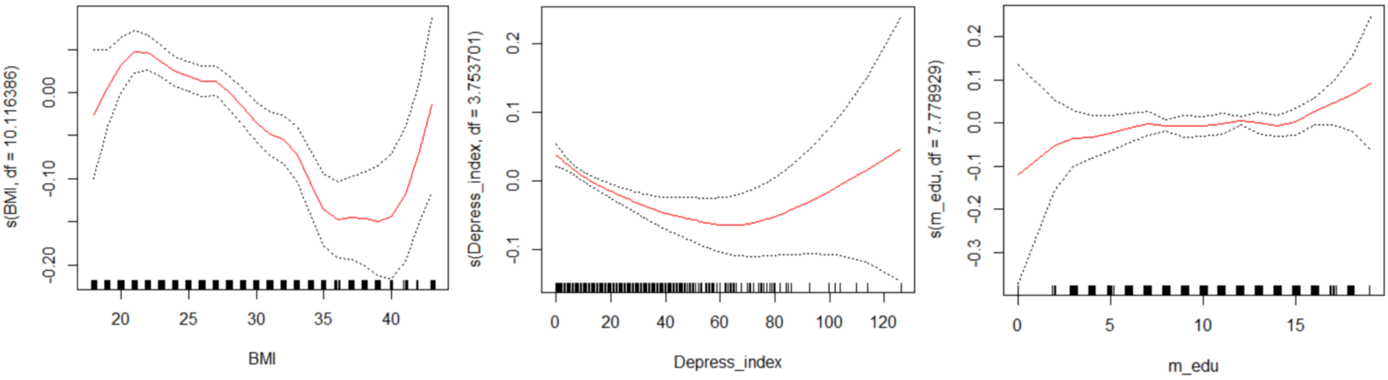

Figure 6. Partial Contribution of BMI (left), Depression Index (center) and Mother’s Education (right) to Health Index

In addition, the result shows that respondents with a strong neurotic personality would be less healthy in their later life. Lastly, though not significant, the frequency of accessing alcohol and cigarette would also influence the health index. Frequent consumption on these items would result in less healthy conditions in later life.

4.3 Random Forest

4.3.1 Random Forest Procedure

We apply random forest model to make the prediction and select important predictors since the random forest model has many advantages against other models. Firstly, it can handle categorical predictors well without creating many dummy variables, and it is simple and straightforward, especially when explaining to people. More importantly, since random forest does not allow to consider a majority of the available predictors and only work on a subset of predictors in each split, it solves the problem of little reduction in variance due to highly correlated trees predicted by bagging method by decorrelating the results. Thus, it makes the average of resulting trees less variable and hence more reliable.

We grow random forest with the default number of predictors for each split of the tree. Notice that by default, when building a random forest of regression trees, randomforest() function uses p/3 variables. Hence, the number of predictors for each split should be 8 in this model.

4.3.2 Random Forest Result

Mean Decrease Accuracy (%IncMSE) and Mean Decrease Gini (IncNodePurity) are criteria to evaluate the significance of predictors in random forest. Mean Decrease Accuracy shows how much the model accuracy decreases if we leave out that variable; Mean Decrease Gini is a measure of variable importance based on the Gini impurity index used for calculating the splits in trees.

According to the result of Mean Decrease Gini (image on the right), the frequency of getting depressed seems to be the most important factor related to the health index. Also, the depression level has a relatively strong impact on humans' health status. In short, the degree of depression is highly correlated to respondents' states of health. Besides, body mass index (BMI) at a young age, father's job, parents' income as well as IQ score, affect the health index in later life, to some extent.

There are some differences between results from Mean Decrease Gini (Image on the right) and Mean Decrease Accuracy (Image on the right). According to Mean Decrease Accuracy, a score in neuroticism on a personality test, as well as the average number of alcoholic drinks per day, is also somehow related to people's health status.

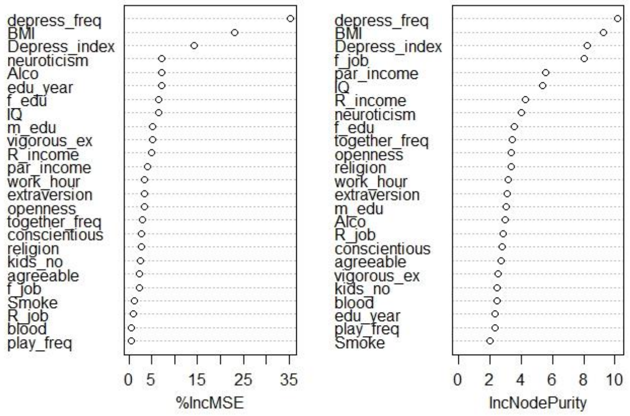

Figure 7. Importance Measures Attained from Random Forest Model

4.4 Stepwise regression

4.4.1 Stepwise selection procedure

This study uses bidirectional elimination, a combination of forward selection and backward elimination method, tests at each step for variables to be excluded or included. In each step, a variable is considered for subtraction from or addition to the set of explanatory variables based on a pre-specified criterion. For each possible subset size, this section calculates the estimated testing errors by techniques of Mallows's Cp, Bayesian information criterion, Adjusted R square. Mallows's Cp is used to assess the fit of a regression model that has been estimated using ordinary least squares. A small value of Cp means that the model is relatively precise. Bayesian information criterion is based on the likelihood function, and the model with the

lowest BIC is preferred. For the adjusted R square criteria, subsets with higher Adjusted R2 is better. From Figure 8 we see by techniques of Cp, BIC, and Adjusted R^2, best subset sizes 18, 10, 24 are obtained for each one.

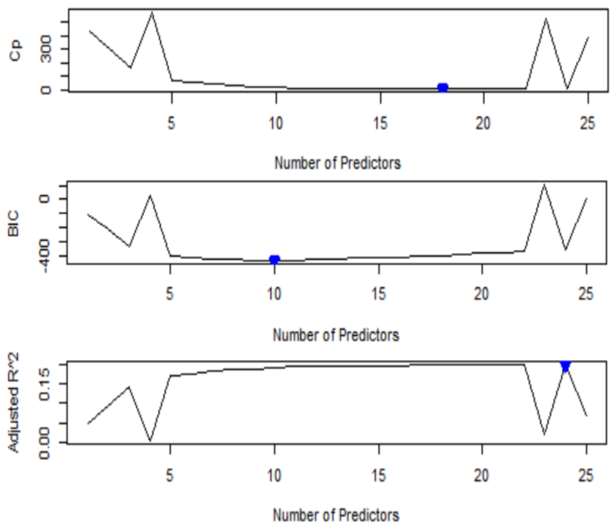

Figure 8. Cp (top), BIC (center) and Adjusted R^2 (bottom) as Function of Number of Predictors

4.4.2 Subset results for three techniques

Although stepwise regression is usually more suitable for samples than new out-of-sample data, one of its main problems is that it searches a large number of possible models, so it is prone to overfit the data. Hence, this section would try to explain the three models as a whole. After calculating the coefficients, three good models are shown in Figure 9. Three models all indicate that health index of respondents in old age is positively correlated with mother's level of education, respondents' IQ score, average number of alcoholic drinks per day when young, and is negatively correlated with depressing index, body mass index, frequency of feeling negative emotions, frequency of participating in vigorous physical exercise, summary score of neuroticism in their youth. Apart from these eight significant predictors, Cp and Adjusted R^2 also pick some other important variables: respondent's job, blood type, the summary score of extraversion, and the summary score of conscientiousness. Training MSE for Cp, BIC, and Adjusted R^2 models is approximately 0.037. Testing MSE for three models is about 0.050.

C_p model

```
  (Intercept)         R_job3         R_job4          m_edu 
 9.640903e-01   4.248204e-02  -2.331644e-02   4.542718e-03 
     R_income             IQ          Smoke           Alco 
 3.487916e-05   8.348171e-04  -8.224020e-03   9.790755e-03 
Depress_index            BMI  depress_freq1  depress_freq2 
-1.411974e-03  -9.801083e-03  -2.643929e-02  -1.047517e-01 
depress_freq3  depress_freq4   vigorous_ex4         blood1 
-2.695338e-01  -2.401159e-01  -2.911897e-02   3.254931e-02 
 extraversion    neuroticism  conscientious 
 2.166634e-03  -4.790942e-03   3.282788e-03 
```

BIC model

```
  (Intercept)          m_edu             IQ           Alco 
 1.0182036379   0.0046509064   0.0008248654   0.0104535860 
Depress_index            BMI  depress_freq2  depress_freq3 
-0.0015794641  -0.0097572601  -0.0982853289  -0.2635623661 
depress_freq4   vigorous_ex4    neuroticism 
-0.2422278735  -0.0302463557  -0.0056794957 
```

Adjusted R^2 model

```
  (Intercept)         R_job1         R_job2         R_job4 
 9.759777e-01  -2.753906e-02  -2.407426e-02  -5.031435e-02 
    work_hour          f_edu          m_edu       R_income 
 3.840043e-04   1.816716e-03   3.404397e-03   3.405921e-05 
       f_job6         f_job9             IQ          Smoke 
-1.593731e-02   2.017321e-02   7.858744e-04  -7.842580e-03 
         Alco  Depress_index            BMI  depress_freq1 
 9.748536e-03  -1.396016e-03  -9.821351e-03  -2.648779e-02 
depress_freq2  depress_freq3  depress_freq4   vigorous_ex4 
-1.043735e-01  -2.665610e-01  -2.395358e-01  -2.904639e-02 
     religion         blood1   extraversion    neuroticism 
-2.167010e-04   3.327692e-02   2.172737e-03  -4.774623e-03 
conscientious 
 3.374977e-03 
```

Figure 9. Coefficient Estimates of The Models With the Lowest Cp (top), Lowest BIC (center) and Largest Adjusted R^2

4.5 Comparison Between Models

From a relatively rigid model like Lasso and stepwise selection to more flexible ones like smoothing splines and random forest, four models discussed above cover a wide range of flexibility. Thus, it is reasonable for us to compare between different models and identify those having good performances. The training and testing mean square error (MSE) for different models are computed and shown below (Fig. 10). Lasso, stepwise have relatively the same performance while smoothing splines have worse performance. This may because the real relationship underlying is closer to a linear relationship rather than non-linear ones. Besides, the similarity in training and testing MSE for Lasso and stepwise echoes the fact that they are both linear regularization model but with different predictor selection methods and fitting goals. The small differences between coefficients of these two models could also support this point. For the case of random forest, it also has similar performance with two other more rigid models while outperforms smoothing splines. This suggests that the real relationship is more likely to be closer than the tree regression form rather than the splines polynomial form. Although it seems like random forest is also a suitable model, by changing the random seed in R code, the random forest result, along with its training and testing MSE varies considerably, suggesting the underlying high variance and non-robustness of this model. The relatively significant differences between training and testing MSE in smoothing splines and random forest may indicate that these models overfit the data while as we use real world data where the composition of variance and bias in testing MSE is unknown, it could not be certain. In conclusion, Lasso and stepwise election model may be two better models among these four.

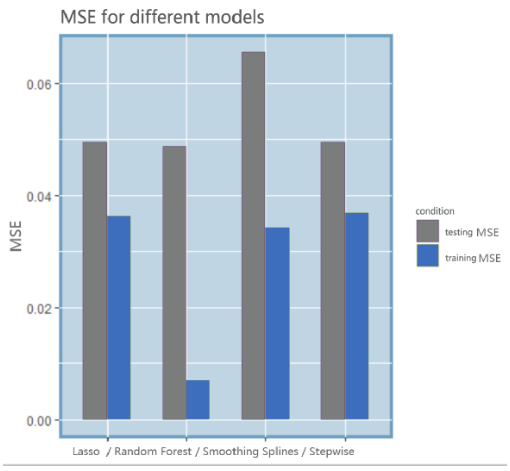

Figure 10. Training and Testing MSE of Four Models

5. Discussion

5.1 Conclusions

According to the four models we applied, "BMI", "depression index" and "depression frequency" seem to be the most important predictors for predicting the HUI score among all the selected predictors since their scores of significance are all ranked as top five by at least three models. Some other predictors, including "neuroticism score", "alcohol consumption", "job", and "income" are also considered as important by some of the models.

The outcome seems reasonable and there are many possible explanations for these results. Firstly, people with high BMI may have many health problems related to obesity, therefore their HUI scores can be lowered. Besides, depression may also lead to digestion problems, lack of sleep and some illnesses like headache, therefore it can seriously damage people's health. Since neuroticism also includes many negative emotions such as anger and anxiety, it can make people's endocrine system unstable, resulting in hormone adjustment disorder. Too much alcohol consumption brings pressure to our liver, thus it leads to lower HUI scores.

Job and income affect people's health more indirectly. For example, the dirty environment of some jobs may cause respiratory infections, and high income guarantees a healthy living condition and a good medical environment when people have some health problems.

5.2 Suggestions

The outcome of the models suggests that the future health index has a significant negative relation with the current BMI index, indicating that people should limit their BMI at a reasonable level to have a healthy body in the future. Besides, the results also show that whether a person feels depressed or not can greatly influence their present and future states of health. Therefore, learning to adjust the mood and sort out depressed feelings can be vital to one's well-being. Apart from depression, other negative emotions, including anger and anxiety should also be handled appropriately. Moreover, it is also essential for people to make a plan to limit their alcohol consumption to maintain their health.

5.3 Limitations

There are also some limitations to our study. First of all, our predictors are not enough and inconclusive, so many of them are not capable of summarily describing our desirable investigation aspect, such as lifestyle, comprehensively. There are only 25 selected predictors, and some of them, according to the results of our models, seem to have relatively weak relationships with the response.

Besides, the selected samples are all from the graduates of the same university and are of the same year, therefore they are very likely to be biased. The results of the model can merely suggest the relationship between health score and other factors of this crowd of people, so it may be a problem to generalize our findings to all people.

Lastly, the data are relatively old, so they may not be representative of the contemporary situation. Since the Wisconsin study was conducted years ago, many data are not capable of describing today's relationship between health and other factors since as science and technology develop, many factors may not be able to dramatically affect people's health state like before.

5.4 Improvements

According to the limitations of our research, we might improve our study in several aspects. The first one is the selection of predictors. Predictors can be chosen from a broader range to cover more related ones to health index. Secondly, our data can also be chosen from randomly selected people, but not from a specific crowd of people. This can be done by picking different datasets from more studies and integrating them together. Last but not least, we can find more recent studies or collect data ourselves to get the latest data, which will make the results more persuasive under current circumstances.

6. References

[1] T. Pudrovska and A. Anishkin (2013, April). *Early-life socioeconomic status and physical activity in later life: evidence from structural equation models*. Journal of aging and health, 25(3), 383–404.

[2] Northwestern University. (2012, March 2). *Lifestyle choices made in your 20s can impact your heart health in your 40s*. ScienceDaily. Accessed on: Nov. 30, 2019. [Online]. Available:

https://www.sciencedaily.com/releases/2012/03/120302132426.htm

[3] B. Arpino, J. Gumà, A. Julià (2018). *Early-life conditions and health at older ages: The mediating role of educational attainment, family and employment trajectories*. PLoS ONE. Accessed on: Nov. 30, 2019. [Online]. Available:

https://doi.org/10.1371/journal.pone.0195320

[4] Horsman, John et al. "*The Health Utilities Index (HUI): concepts, measurement properties and applications*." Health and quality of life outcomes vol. 1 54. 16 Oct. 2003. Accessed on: Nov. 30, 2019. [Online]. Available:

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC293474/