# Draft report

Yuhan Hu

2020/2/29

## Contents

# Summary

In this study ,we are interested in association between changes in health risk factors and BLR scores as well as association between frequency of BLR platform usage and change in BLR score. We decided to use LASSO shrinkage to select predictor for the first question and represent frequency of BLR platform usage by ratio of sum of accumulaiton of points to times of platform usage for second question. By fitting linear regression model for both research question, we conclude that on average, change in physical activity score will result in the greatest improvement in BLR score, and frequency of platform usage is not associated with improvement of BLR score.


# Introduction

BestLifeRewarded Innvations are running platform where user can evaluate their health status by completing survey or earn reward points by participating in activities. Candidates are from employees from 3 companies.

By completing survery, BLR score, a score represents candidates' overall health status, will be computed based on health risk factors such as gastrointestinal score, cancer score, mental health score, diet score, physical activity score, financial health score, medication score, alcohol score, stress score, smoking score, heart score, respiratory score, diabetes score, arthritis pain score, Social financial relationship score, sleep score and BMI.

For this analysis, we are interested in which health risk factors is associated with a higher rate of improvement in BLR score, or in another word, how does changes in those given health risk factors affect BLR score rate of change.

Users can earn points by participating in platform activities, and those points can be spend on the platform for reward. Here we are interested in association between platform usage frequency and improvement in BLR score. Is a higher frequency of usage of the platform associated with a general improvement in health as measured by the risk factors.

Statistical analysis will be discussed in method section. Answer to the research question is in result section. Potential question underlying this analysis and possible direction for future research are in discussion section.

# Method

## Brief data management

User can alter data entries after created, time of last change is labelled as finished date. If a data entry has been altered several days after created, it is unlikely to reflect a concise health state of the user, therefore we remove all data points whose created date is not the same as finished date.Also, recall the research question, we are interested in chronological changes between data points of each user.Therefore, we remove user with only one valid entry. Detailed list of removed entries can be found in supplement file.

In addition, we did not use original data for analysis. Since we are interested in how changes on health-risk factors affect change on BLR score. For each user, we subtract values in the first entry by values in the last entry(chronological order), then divide the result by time difference between this two entry.

For the second research question, we decide to use ratio of sum of accumulaiton of points to times of platform usage to model frequency of platform usage.

## Model 1

Changes in BLR score is the variable of interest, thus the response variable. All other variables are potential predictor of interest. Since we do not have additional information about covariates, there is no clue for

eliminating non-significant features, therefore we decide to use LASSO shrinkage method to choose features automatically.

As for tunning parameter $\lambda$ of LASSO shrinkage, we chose $\lambda_{1se}$(second vertical dotted line), that gives a model such that error is within one standard error of the minimum.

By LASSO shrinkage, changes in gastrointestinal score, cancer score, mental health score, diet score, physical activity score, financial health score, medication score, alcohol score, stress score and smoking score are included in the model, while changes in heart score, respiratory score, diabetes score, arthritis pain score, Social financial relationship score, sleep score and BMI are removed from the model.

Since we do not have additional information about possible interaction between potential variables, no interaction terms are added into the model.

Therefore, our model is linear regression model given below.

$Y = \beta_0 + \sum \beta_i X_i + \epsilon$ where $\sum \beta_i X_i$ is effect of those predictors and $\epsilon$ is the error term.

Detailed comparision about simple linear regression and linear mixed model is discussed in Discussion section.

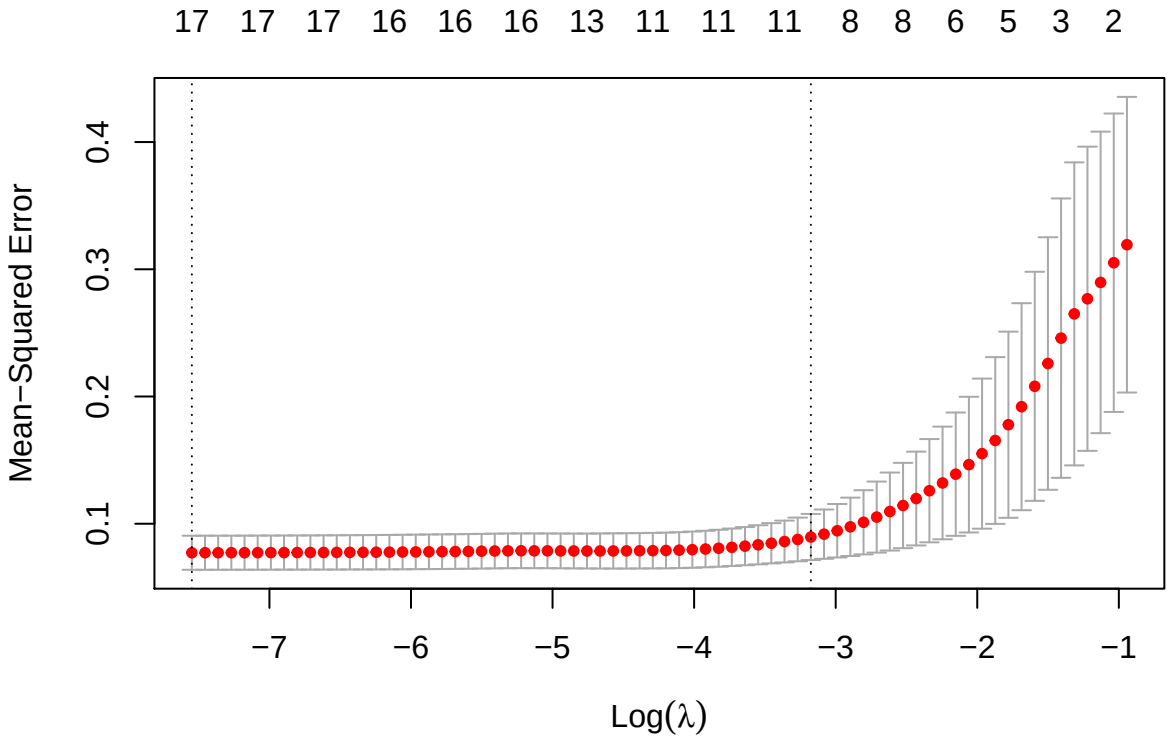



Table 1: Coefficient of health-risk factors by LASSO

| | Coefficient |
|---|---|
| (Intercept) | 0.008 |
| GastrointestinalScore_change | -0.113 |
| DiabetesScore_change | -0.007 |
| CancerScore_change | -0.030 |
| MentalHealthScore_change | -0.519 |
| DietScore_change | 1.323 |

| | Coefficient |
|---|---|
| PhysicalActivityScore_change | 2.132 |
| FinancialHealthScore_change | 0.872 |
| MedicationScore_change | 0.364 |
| AlcoholScore_change | 0.180 |
| StressScore_change | 0.762 |
| SmokingScore_change | 0.794 |

### Model 2

As for the second research question, initially we have several potential predictors such as sum of points earning, sum of points spending, times of platform usage and ratio of sum of points earning to times of platform usage.

We choose linear regression model for this research question.

$Y = \beta_0 + \sum \beta_i X_i + \epsilon$ where $\sum \beta_i X_i$ is effect of those predictors and $\epsilon$ is the error term.

Linear regression models has been fitted for each predictors individually, and only ratio of sum of points earning to times of platform usage returns a significant coefficient.

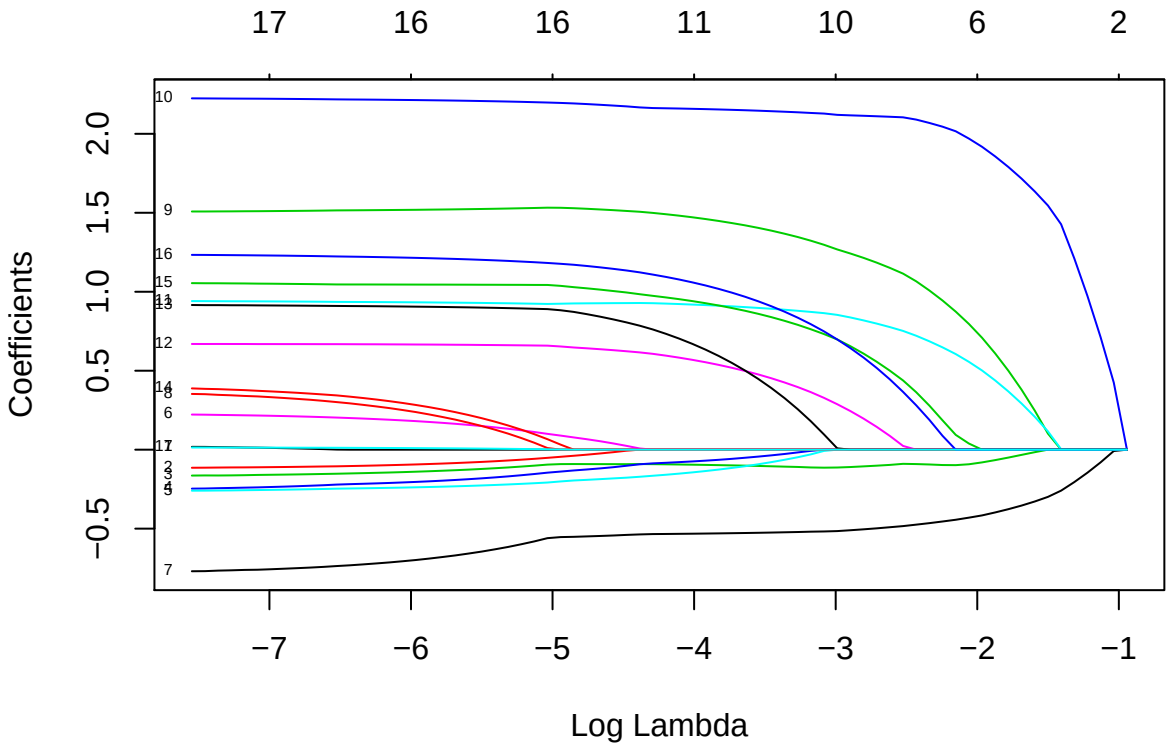
Table 2: coefficient of models 2

| | coefficient |
|---|---|
| (Intercept) | 2.6412179 |
| points_per_use | -0.0000026 |

## Result

### Research Question 1

Refer to Table, changes in physical activity scores has the greatest coefficient, which indicate that this characteristics is associated with the higher improvement rate of BLR score.Changes in cancer score, gastrointestinal score and mentalhealth score has negative impact on changes of BLR score. i.e. the larger these 3 score, the smaller the BLR score, which indicates a worst health condition. As for changes in alcohol score, medication score, smoking score, stress score, financial health score, diet score and physical activity score, change in BLR score increases with changes in these 7 health risk factors. Both negative effect predictors and positive predictors are given in increasing order of effect on BLR score change.(i.e. in increasing order of absolute value of corresponding coefficients.) Order of predictors' effect on BLR score change may differ with alternative value for tunning parameter $\lambda$, but the order is roughly consistent.




### Research Question 2

As mentioned above, most potential predictors do not return a significant coefficient, which indicates that those predictors do not affect changes in BLR scores. Moreover, value of the only significant coefficient, coefficient of ratio of sum of points earning to times of platform usage, is close to 0, and the corresponding p-value is at the boundary between significant and non-significant. Therefore, the answer to the second research question is negative. Higher frequency of platform usage does not associate with a general improvement in BLR score.

## Discussion

In conclusion, changes in physical activity scores impact BLR score rate of improvement the most, and a higher frequency of platform usage is not associated with a higher BLR score rate of improvement.

In addition, there are several aspects worth mentioning about this analysis.

### Model 1

1. We are interested in linear mixed model initially since intuitively, health status are supposed to be affected by individual differences. However, after built several mixed model as attempt, the result shows that the random effect explain nearly none of variability in data. (i.e. $\sigma^2/\ (\sigma^2 + \tau^2) = 0$ with 3 significant digits) Therefore, we prefer simpler model, which is linear regression model.(**APPENDIX mixed model section**)

2. Residual plot of both models are not perfectly normal. Since we are interested in inference in this analysis, heavily tailed error term may not be a significant problem. However, in case of prediction, this non-normal distribution of error term requires concern.

3. Process of predictor selection was done by LASSO algorithm automatically. However, in analysis, there might be variables that are supposed to be in the model regardless of any measure of significance. Due to lack of information, we didn't know if there were such variable.

4. Variance inflcation factor(VIF) for some varibales in this model is moderately high, i.e. VIFs for changes in gastrointestinal scores and mental health scores are greater than 3. Here we decide to ignore this problem, however, exactly how large a VIF has to be before it causes issues is a subject of debate, if others prefer a more strict standard, they need to concern about multicolinarity here. (**Appendix VIF section**)

### Model 2

1. The research question is ambiguous, **'frequency of usage of the platform'** is not clearly defined. Though we tried to fit several models with various representation of **'frequency of usage of the platform'** and obtained consistent conclusion, if collaborator's idea of **'frequency of usage of the platform'** was not in the given set of representations, result from current analysis may become invalid.

2. Based on our analysis, higher frequency of platform usage is not associated with higher rate of BLR score improvement, which is counter-intuitive. A potential explanation for this result could be that frequency of platform usage is associated with rate of improvement of those predictors, but these impacts canceled out and thus has no impact on BLR score. Analysis about those potential associations require further investigation.

## Appendix

### VIF

Table 3: VIF table

| | VIF |
|---|---|
| GastrointestinalScore_change | 3.233 |
| CancerScore_change | 1.247 |
| MentalHealthScore_change | 3.020 |
| DietScore_change | 1.342 |
| PhysicalActivityScore_change | 1.541 |
| FinancialHealthScore_change | 1.369 |
| MedicationScore_change | 1.125 |
| AlcoholScore_change | 1.079 |
| StressScore_change | 1.765 |
| SmokingScore_change | 1.072 |

### mixed model

Table 4: mixed model

| | MLE | Std.Error | DF | t-value | p-value |
|---|---|---|---|---|---|
| (Intercept) | 0.142 | 0.015 | 29488 | 9.457 | 0 |
| Heart_change | -0.084 | 0.010 | 29488 | -8.067 | 0 |
| Respiratory_change | -0.113 | 0.014 | 29488 | -7.970 | 0 |
| Gastrointestinal_change | -0.203 | 0.014 | 29488 | -14.692 | 0 |
| Diabetes_change | -0.170 | 0.018 | 29488 | -9.428 | 0 |
| Cancer_change | -0.428 | 0.015 | 29488 | -27.889 | 0 |
| MentalHealth_change | -0.275 | 0.012 | 29488 | -22.813 | 0 |
| SocioFinancial_change | -0.283 | 0.017 | 29488 | -16.640 | 0 |
| Diet_change | 1.419 | 0.022 | 29488 | 65.840 | 0 |
| PhysicalActivity_change | 1.416 | 0.019 | 29488 | 73.208 | 0 |
| FinancialHealth_change | 1.267 | 0.017 | 29488 | 76.318 | 0 |
| Medication_change | 0.325 | 0.018 | 29488 | 18.578 | 0 |
| Alcohol_change | 1.226 | 0.031 | 29488 | 39.369 | 0 |
| Sleep_change | 1.182 | 0.040 | 29488 | 29.879 | 0 |
| Stress_change | 1.199 | 0.030 | 29488 | 40.625 | 0 |
| Smoking_change | 0.812 | 0.013 | 29488 | 64.353 | 0 |
| $\sigma$ | 0.000 | NA | NA | NA | NA |
| $\tau$ | 2.926 | NA | NA | NA | NA |