Dietary Patterns and Mental Distress
Project Report

by

Yunhui Liu

Supervised by: Xingye Qiao

Department of Mathematical Sciences
Statistics
GD Harpur

Binghamton University

Binghamton, New York

04/2020

Contents

1. Data Cleaning and Visualization

1.1 Background

The data comes from an online questionnaire provided by Professor Lina Begdache. The questionnaire asks questions mainly about participants' mental distress and dietary patterns, and some other information including gender, age, region and education, to study the relationship between mental distress and dietary pattern.

There are 2636 participants, 6 questions asking about the mental status and 21 other questions. Therefore, in the original dataset, there are 2636 observations, 21 predictors and 6 responses

1.2 Dataset cleaning

Observations with N/A and invalid value are deleted at first.

Then check all variables and make some corrections.

Gender

There is only one "Other" observation, for better analysis of data, remove that observation

---

Region

- There are two "Option 8" observations and delete them.
- There are repeated levels such as: "Middle East/North Africa" and "Africa", "Australia" and "Australia /New Zealand". Therefore, combine those observations into "Africa" and "Australia".
- After combining those levels, there are still too few observations for "South America" and "Australia", so combine "Australia" with "Europe" and "South America" with "North America", based on economics and geography

---

Education

- There are some levels with too few observations, so combine below levels into one level:
  - "Less than high school" and "High school"
  - "Master's degree" and "Graduate"
  - "Doctoral Degree", "Professional Degree (MD, JD, PharmD, ...)" and "Professional"

`Dietary pattern`

- There are many repeated levels, repeated `"Mediterranean Diet"` due to spelling and repeated `"Vegan diet"` due to different way to call it. Combine repeated level
- Some `"Sophomore"` observations are deleted
- For `"Asian diet"`, `"Caribbean diet"`, `"Korean diet"` and `"Vegan diet"`, they are removed due to only one observation for each level

`Fish oil`

There are 88% observations do not eat fish oil weekly, delete this variable

`Other`
For the frequency of consuming different type of food, combine some repeated levels

After the above cleaning, there are 2314 observations left.

1.3 Convert categorical variables into numerical variables

To simplified the analysis, some categorical variables are converted into numerical variables.

- Transform `"Education"` into numeric 1 $\sim$ 4, according to different education levels.
- For the frequency of consuming different type of food, convert frequency between 0 and X into number, and `"More than X times"` into number $X+1$,
- For 6 variables describe how often negative mentality happens, convert `"All the time"`...`"None of the time"` into 1 to 5, and get a new response `"mental health score"` from the average of the original 6 responses.

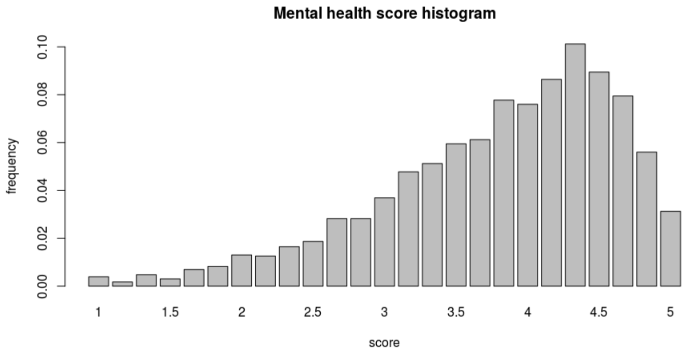
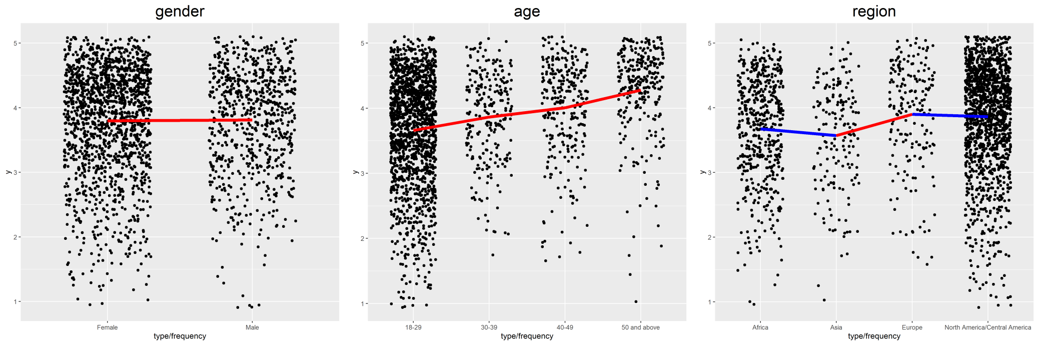
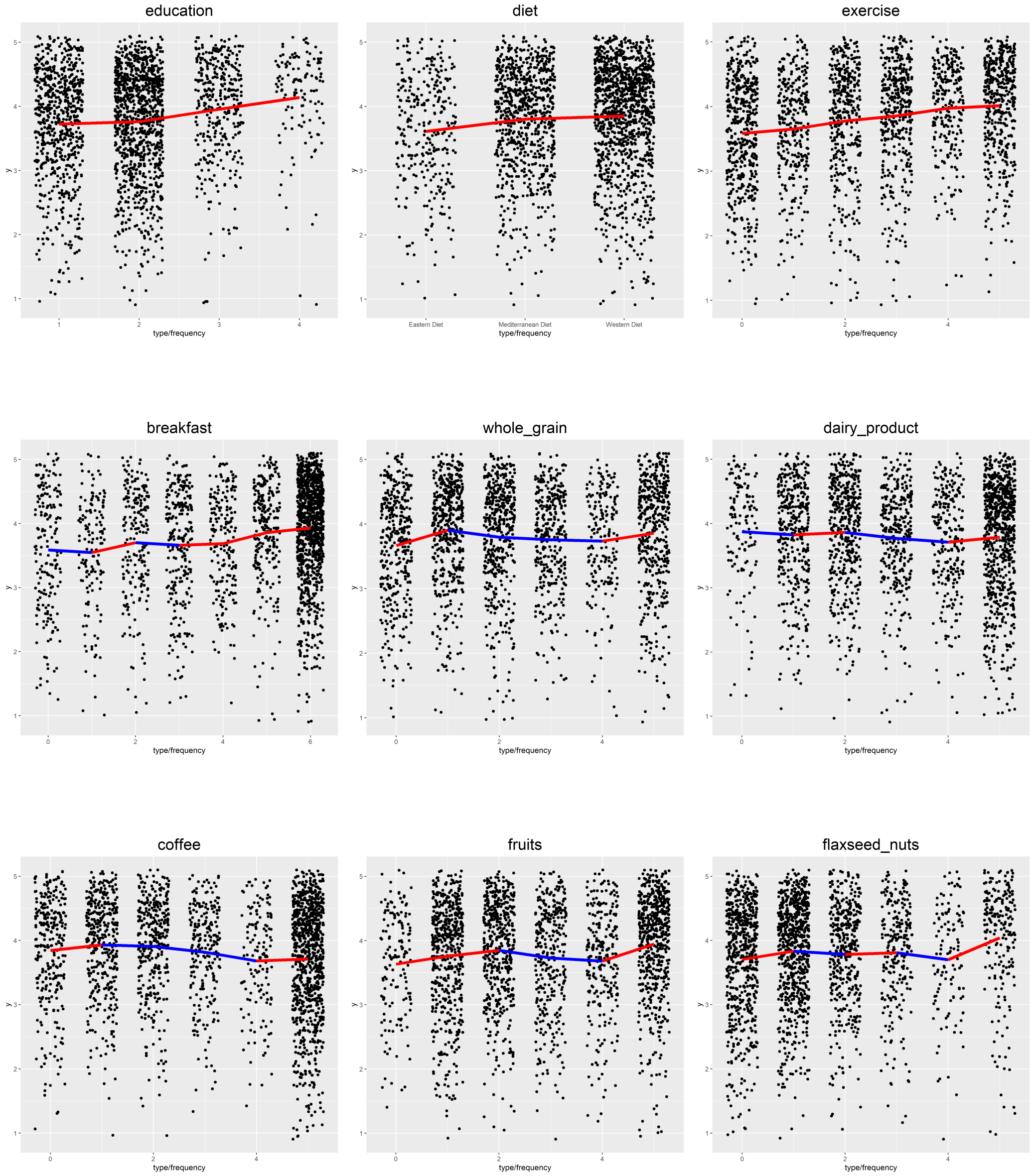


Figure 1.1: Mental Health Score Distribution

1.4 Visualization

The scatter plots show the distribution and mean value of mental health score for each variables. The mean values are connected by lines, indicating increasing by red lines and decreasing by blue lines, or difference among levels.




education
diet
exercise
breakfast
whole_grain
dairy_product
coffee
fruits
flaxseed_nuts
y
1
2
3
4
5
type/frequency
0
1
2
3
4
6
Eastern Diet
Mediterranean Diet
Western Diet

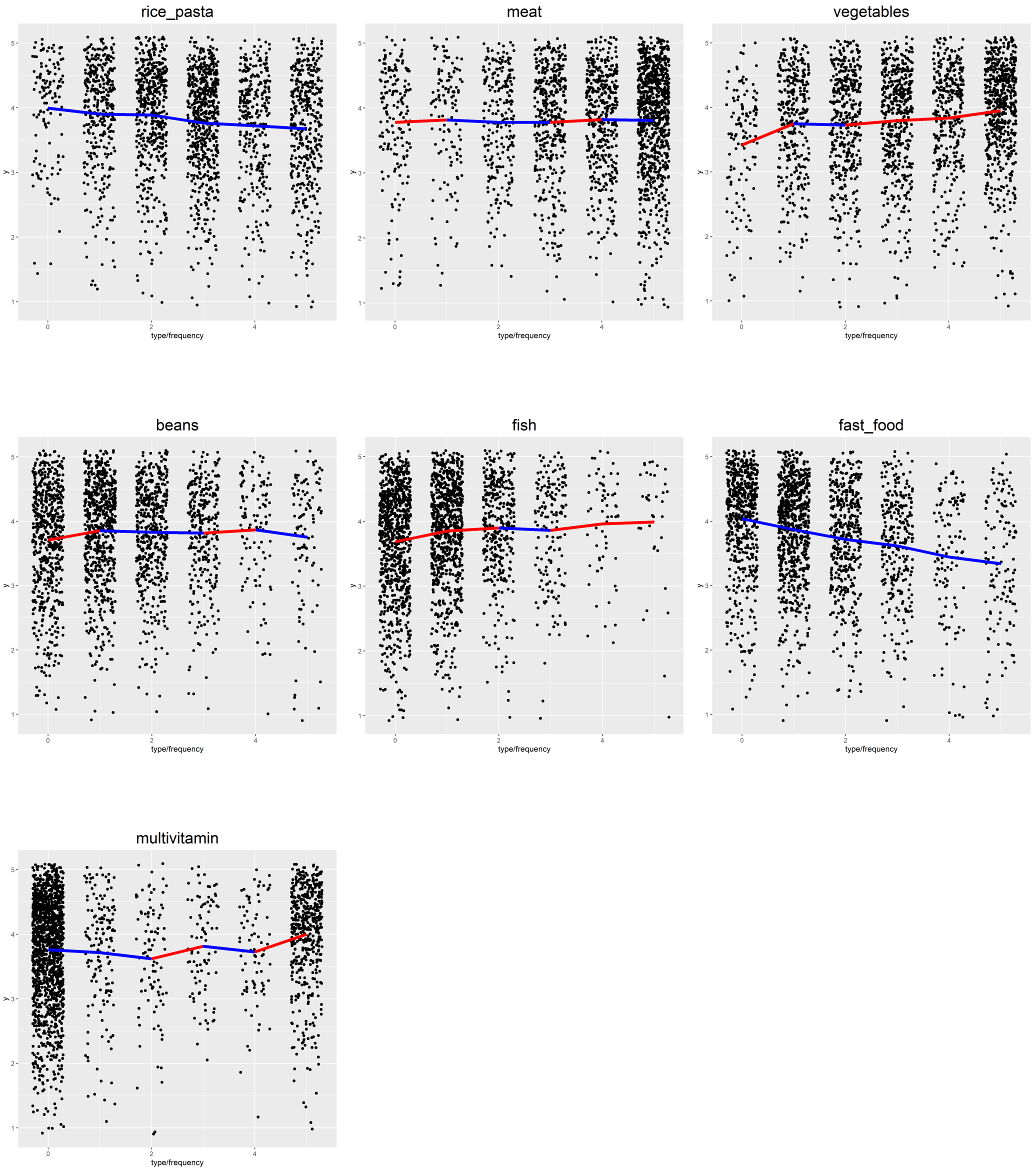


Figure 1.2: Scatter Plot and Mean Value of Variables

After visualization, it is found that:

- There are some variables have large increasing only within one interval, such as:

  - `"Flaxseed/nuts", "Fruits"`, when $x \geq 4$
  - `"Vegetables", "Whole grain"`, when $x \leq 1$

  These variables are suitable to do segmented regression. Segmented regression treats the variable as two variables: one is the interval of increase, the other is the interval of no obvious increase.

- There are 71% (1498) observations are 18∼29 years old, this group should be analyzed individually.

1.5 Summary of Variables

Name of variable	**Type**	**Description**
`y` (Response)	Numerical	Mental health score, from 1 to 5
`Gender`	Categorical	Female or Male
`Age`	Categorical	18-29, 30-39, 40-49 or 50 and above
`Region`	Categorical	Regions participant comes from
`Education`	Numerical	From 1 to 4, indicating education level
`Diet`	Categorical	Different diet type
`Exercise`	Numerical	Times of exercise per week
`Breakfast`	Numerical	Times of breakfast per week
`Whole grain`	Numerical	Times of eating whole grain per week
`Dairy product`	Numerical	Times of eating dairy product per week
`Coffee`	Numerical	Times of drinking coffee per week
`Fruit`	Numerical	Times of eating fruit per week
`Flaxseed/nuts`	Numerical	Times of eating flaxseed/nuts per week
`Rice/pasta`	Numerical	Times of eating rice/pasta per week
`Meat`	Numerical	Times of eating meat per week
`Vegetables`	Numerical	Times of eating vegetables per week
`Beans`	Numerical	Times of eating beans per week
`Fish`	Numerical	Times of eating fish per week
`Fast food`	Numerical	Times of eating fast food per week
`Multivitamin`	Numerical	Times of eating multivitamin per week

2. Analysis

2.1 Setup

Because there are enough observations, the whole dataset is split randomly into 50% training dataset and 50% test dataset. In the following analysis, we will select variables using the training data and test the significance in the test data.

Our analysis will be done by 3 ways: analysis on the whole data, analysis on the data separated by age, and analysis on the data separated by gender.

2.2 Terminologies

- Linear regression: Linear regression is a linear approach to modeling the relationship between a response and explanatory variables. The model takes the form:

$$y_i = \sum_{i=0}^{p} \beta_j x_{ij} + \varepsilon_i, i = 1...n$$

β_j is the coefficient of predictor, ε_i is the error term, i is the index of observation and j is the index of predictor. The β_j is estimated by ordinary least squares.

- Backward AIC: The Akaike information criterion (AIC) is an estimator of out-of-sample prediction error and thereby relative quality of statistical models for a given set of data. Backward means starting with all candidate variables, testing the deletion of each variable using a chosen model fit criterion, deleting the variable whose loss gives the most statistically insignificant deterioration of the model fit, and repeating this process until no further variables can be deleted without a statistically significant loss of fit.
- Cross validation: Cross validation is any of various similar model validation techniques for assessing how the results of a statistical analysis will generalize to an independent data set.
- LASSO: LASSO (Least Absolute Shrinkage and Selection Operator) is a regression analysis method that performs both variable selection and regularization in order to enhance the prediction accuracy and interpretability of the statistical model it produces.
- Segmented regression: A method in regression analysis in which the independent variable is partitioned into intervals and a separate line segment is fit to each interval.
- P-value: P-value is the probability of obtaining test results at least as extreme as the results actually observed during the test, assuming that the null hypothesis is correct. The smaller p-value is, the less possible for the coefficient

to be zero. In the analysis, if the variable's p-value ≤ 0.05, the variable is considered to be significant.

- Significance: In statistical hypothesis testing, a result has statistical significance when it is very unlikely to have occurred given the null hypothesis. In the analysis, null hypothesis H_0 is that $\beta_p = 0$. A variable is significant indicating the rejection of the null hypothesis and a linear relationship between the response and the predictor.

2.3 Statistical Analysis Methods

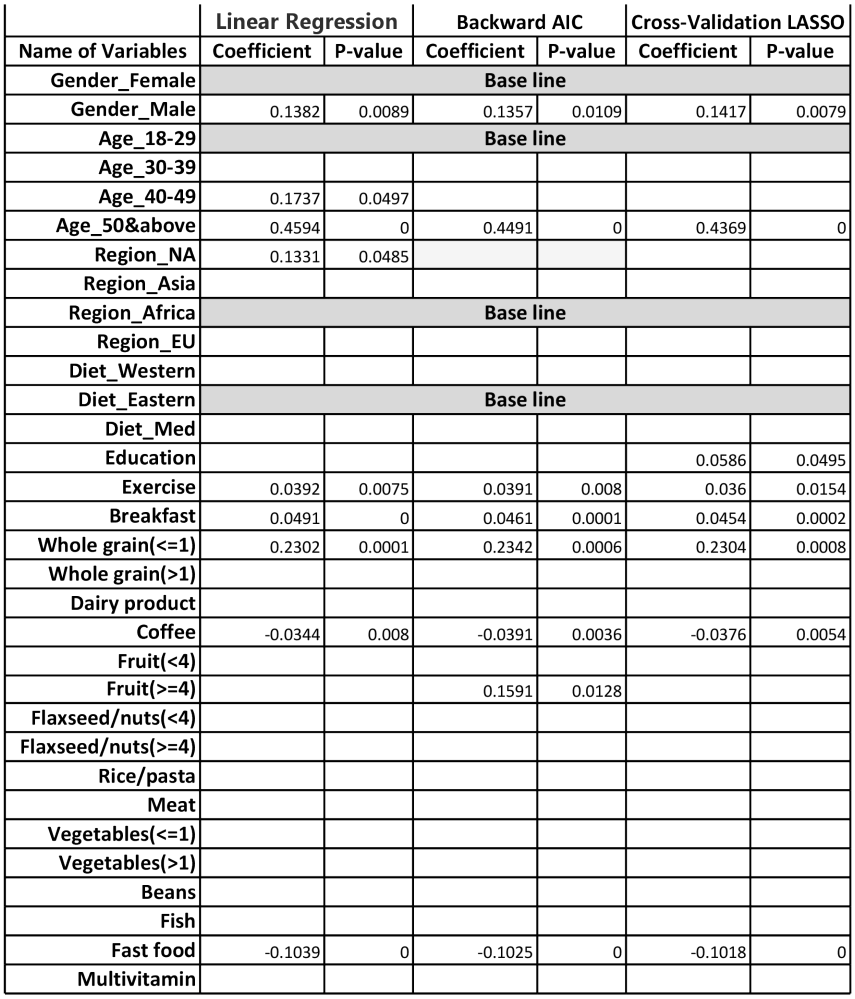
- Linear Regression: Fit a linear regression model on training dataset, use the obtained significant variables to fit another model on test set to check the significance.
- Backward AIC: Based on Akaike Information Criterion, fit a linear regression model and conduct best subset selection in backward direction on training data. Then use the test set to test the significance of the selected variables.
- Cross Validation LASSO: Select the variables by LASSO, whose parameter is chosen by cross validation. Fit a linear regression model using the selected variables on the training set. Then use the test set to test the significance of the selected variables.

2.4 Result

- Interpretation
    - The tables only display the significant variables.
    - Base line means this level of a variable is treated as reference level for comparing.
    - N/A indicates the variable is not available in respond analysis.
    - Different column represents the different analysis method.
    - Positive sign of coefficient means the increase of the variable will increase the mental health, and vice the verse.
    - Variables like "Whole grain($\leq$1)" are the variables using segmented regression. For example, in the following table, "Whole grain($\leq$1)" is significant and the coefficient is positive. The significance means when eating whole grain once per week, it increases mental status significantly compared with not eating whole grain. However, "Whole grain(>1)" is not significant. The non-significance represents that for people eat whole grain for more than once, the times of whole grain consumption is not significant for mental status.

The result of analysis on whole data is summarized in table.

Table 1: the result of analysis on whole data



	Linear Regression		Backward AIC		Cross-Validation LASSO	
Name of Variables	Coefficient	P-value	Coefficient	P-value	Coefficient	P-value
Gender_Female	Base line					
Gender_Male	0.1382	0.0089	0.1357	0.0109	0.1417	0.0079
Age_18-29	Base line					
Age_30-39						
Age_40-49	0.1737	0.0497				
Age_50&above	0.4594	0	0.4491	0	0.4369	0
Region_NA	0.1331	0.0485				
Region_Asia						
Region_Africa	Base line					
Region_EU						
Diet_Western						
Diet_Eastern	Base line					
Diet_Med						
Education					0.0586	0.0495
Exercise	0.0392	0.0075	0.0391	0.008	0.036	0.0154
Breakfast	0.0491	0	0.0461	0.0001	0.0454	0.0002
Whole grain(<=1)	0.2302	0.0001	0.2342	0.0006	0.2304	0.0008
Whole grain(>1)						
Dairy product						
Coffee	-0.0344	0.008	-0.0391	0.0036	-0.0376	0.0054
Fruit(<4)						
Fruit(>=4)			0.1591	0.0128		
Flaxseed/nuts(<4)						
Flaxseed/nuts(>=4)						
Rice/pasta						
Meat						
Vegetables(<=1)						
Vegetables(>1)						
Beans						
Fish						
Fast food	-0.1039	0	-0.1025	0	-0.1018	0
Multivitamin						

From the table, all three methods have agreement on significance of following variables or levels:

- Positive: Male, Age over 50, exercise, breakfast, whole grain($\leq$1)
- Negative: Coffee, fast food

The results suggest that:

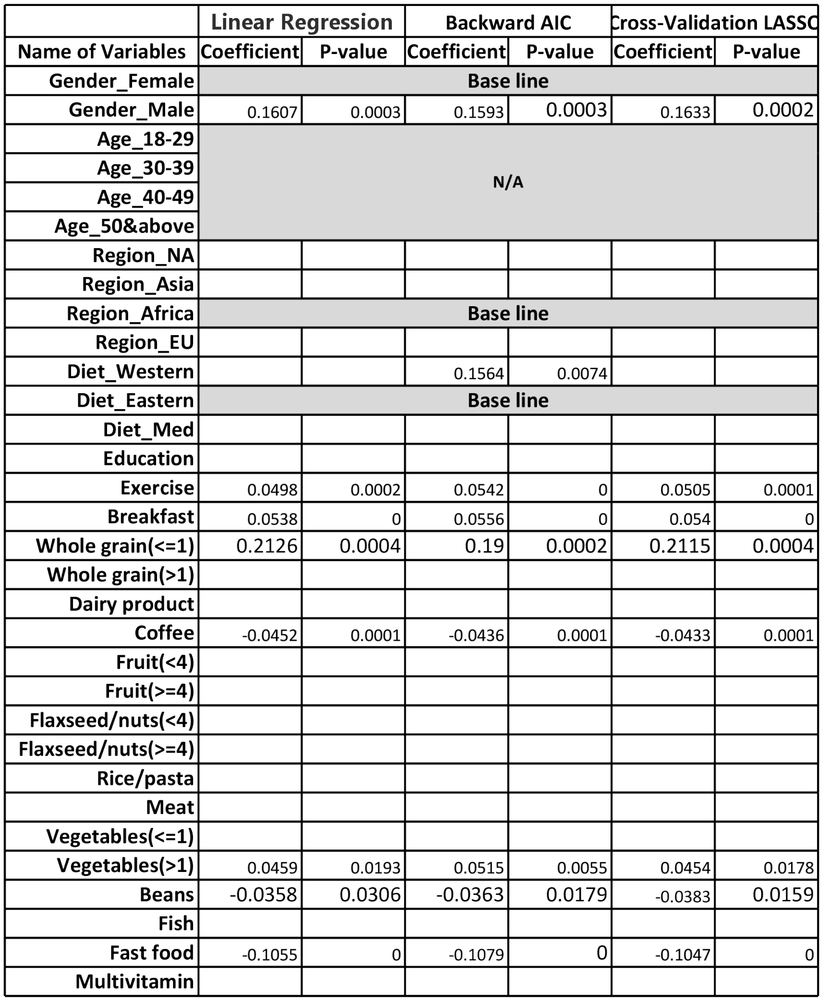
- Matured and male observations are likely to have higher mental health score.
- People with good mentality usually have stable exercise and breakfast, which is understandable. Eating whole grain could also be a sign of good mental status.
- More consumption of coffee and fast food relates with low mental health score. It makes sense, because busy or stressful people drinks coffee and people with low salaries eat fast food more often.

2.5 Analysis on Young and Matured Observations

To find any difference between the young group and the matured group, split the whole dataset by different age, then conduct analysis using the same methods as above on the two data subsets.

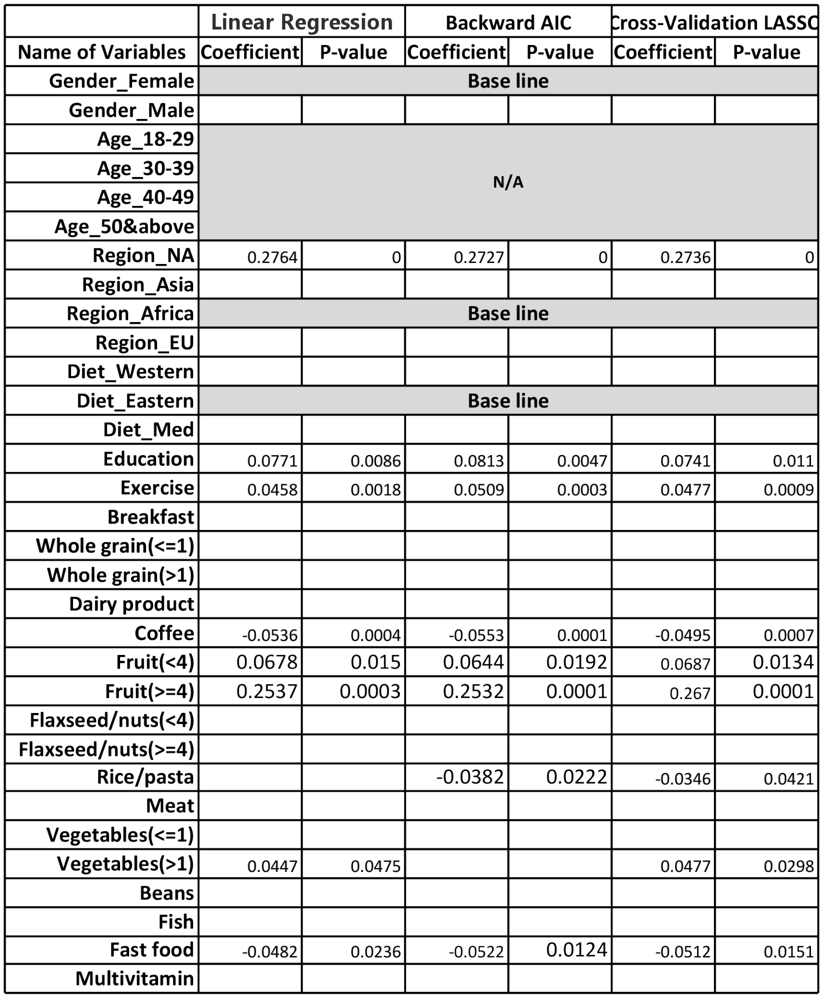
The result of analysis is summarized in the following tables.

Table 2: the result of analysis on young group



	Linear Regression		Backward AIC		Cross-Validation LASSO	
Name of Variables	Coefficient	P-value	Coefficient	P-value	Coefficient	P-value
Gender_Female	Base line					
Gender_Male	0.1607	0.0003	0.1593	0.0003	0.1633	0.0002
Age_18-29	N/A					
Age_30-39						
Age_40-49						
Age_50&above						
Region_NA						
Region_Asia						
Region_Africa	Base line					
Region_EU						
Diet_Western			0.1564	0.0074		
Diet_Eastern	Base line					
Diet_Med						
Education						
Exercise	0.0498	0.0002	0.0542	0	0.0505	0.0001
Breakfast	0.0538	0	0.0556	0	0.054	0
Whole grain(<=1)	0.2126	0.0004	0.19	0.0002	0.2115	0.0004
Whole grain(>1)						
Dairy product						
Coffee	-0.0452	0.0001	-0.0436	0.0001	-0.0433	0.0001
Fruit(<4)						
Fruit(>=4)						
Flaxseed/nuts(<4)						
Flaxseed/nuts(>=4)						
Rice/pasta						
Meat						
Vegetables(<=1)						
Vegetables(>1)	0.0459	0.0193	0.0515	0.0055	0.0454	0.0178
Beans	-0.0358	0.0306	-0.0363	0.0179	-0.0383	0.0159
Fish						
Fast food	-0.1055	0	-0.1079	0	-0.1047	0
Multivitamin						

Table 3: the result of analysis on matured group



	Linear Regression		Backward AIC		Cross-Validation LASSO	
Name of Variables	Coefficient	P-value	Coefficient	P-value	Coefficient	P-value
Gender_Female	Base line					
Gender_Male						
Age_18-29	N/A					
Age_30-39						
Age_40-49						
Age_50&above						
Region_NA	0.2764	0	0.2727	0	0.2736	0
Region_Asia						
Region_Africa	Base line					
Region_EU						
Diet_Western						
Diet_Eastern	Base line					
Diet_Med						
Education	0.0771	0.0086	0.0813	0.0047	0.0741	0.011
Exercise	0.0458	0.0018	0.0509	0.0003	0.0477	0.0009
Breakfast						
Whole grain(<=1)						
Whole grain(>1)						
Dairy product						
Coffee	-0.0536	0.0004	-0.0553	0.0001	-0.0495	0.0007
Fruit(<4)	0.0678	0.015	0.0644	0.0192	0.0687	0.0134
Fruit(>=4)	0.2537	0.0003	0.2532	0.0001	0.267	0.0001
Flaxseed/nuts(<4)						
Flaxseed/nuts(>=4)						
Rice/pasta			-0.0382	0.0222	-0.0346	0.0421
Meat						
Vegetables(<=1)						
Vegetables(>1)	0.0447	0.0475			0.0477	0.0298
Beans						
Fish						
Fast food	-0.0482	0.0236	-0.0522	0.0124	-0.0512	0.0151
Multivitamin						

From the table, all three methods have agreement on the significance of following variables or levels. Highlighted variable means that variables only significant in the relative data:

For young group:

- Positive: Male, exercise, breakfast, whole grain($\leq$1), vegetables($\geq$1)
- Negative: Coffee, beans , fast food

For matured group:

- Positive: Region North America, education, exercise, fruit(<4), fruit($\geq$4)
- Negative: Coffee, fast food

The results show that:

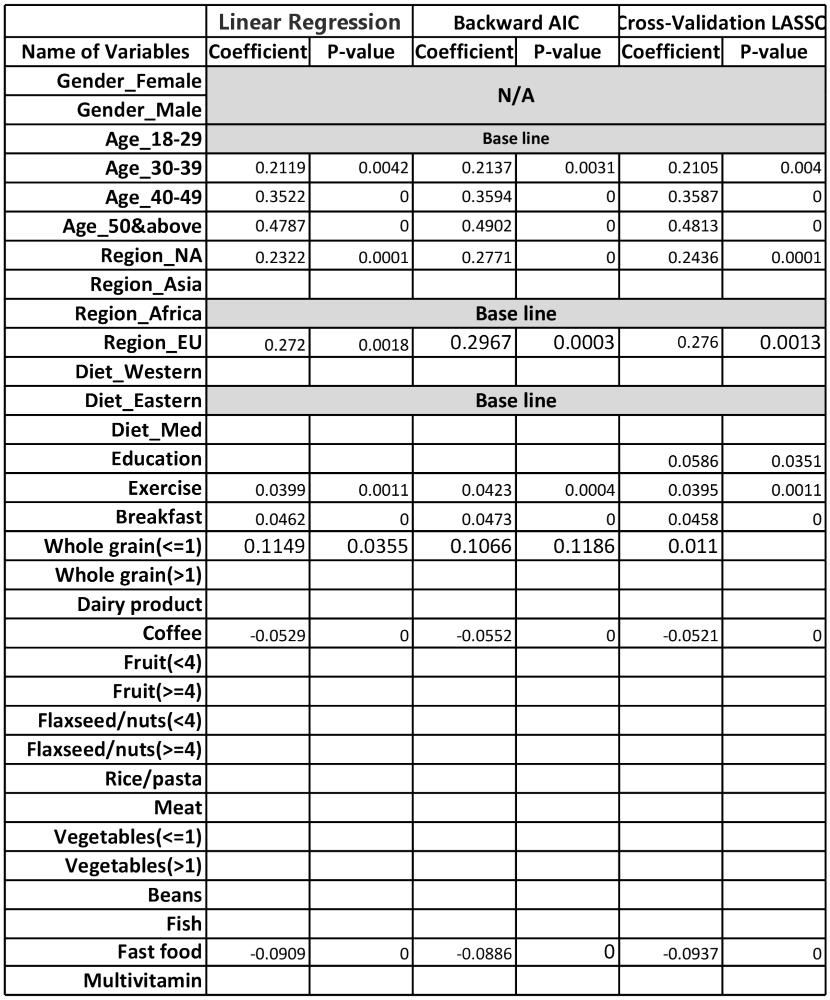
- In young group
  - Different gender are more likely to have different mental health status. Possibilities are that the experience of young females and males could be very different.
  - Breakfast and dietary patterns including eating whole grain, vegetables and beans could also be the indicators of the score for young people.
- In matured group
  - People who lives in North America usually has higher mental health score than those who lives in Africa. Probably it is due to the economic difference.
  - Because education is the most important indicator of income, it makes sense that people with different education have very different mental status.

2.6 Analysis on Female and Male

To find any difference between different gender, split the whole dataset by the gender, then conduct analysis using the same methods as above on the two dataset.

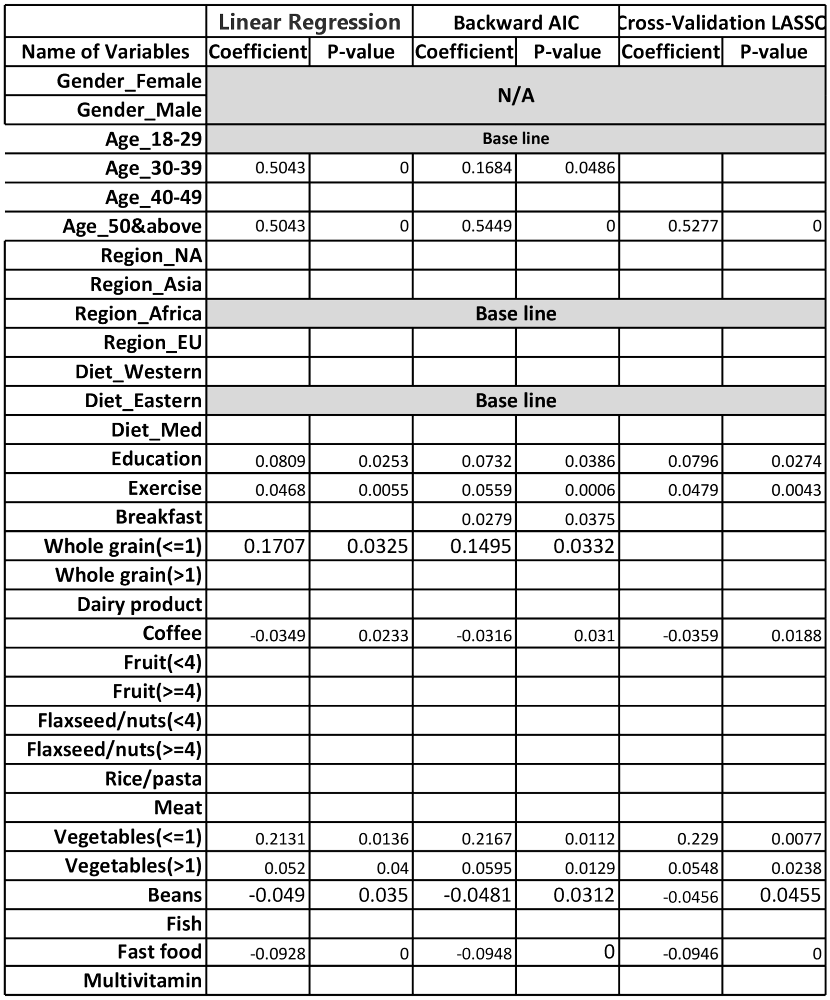
The result is summarized in the following tables.

Table 4: the result of analysis on female



	Linear Regression		Backward AIC		Cross-Validation LASSO	
Name of Variables	Coefficient	P-value	Coefficient	P-value	Coefficient	P-value
Gender_Female	N/A					
Gender_Male						
Age_18-29	Base line					
Age_30-39	0.2119	0.0042	0.2137	0.0031	0.2105	0.004
Age_40-49	0.3522	0	0.3594	0	0.3587	0
Age_50&above	0.4787	0	0.4902	0	0.4813	0
Region_NA	0.2322	0.0001	0.2771	0	0.2436	0.0001
Region_Asia						
Region_Africa	Base line					
Region_EU	0.272	0.0018	0.2967	0.0003	0.276	0.0013
Diet_Western						
Diet_Eastern	Base line					
Diet_Med						
Education					0.0586	0.0351
Exercise	0.0399	0.0011	0.0423	0.0004	0.0395	0.0011
Breakfast	0.0462	0	0.0473	0	0.0458	0
Whole grain(<=1)	0.1149	0.0355	0.1066	0.1186	0.011	
Whole grain(>1)						
Dairy product						
Coffee	-0.0529	0	-0.0552	0	-0.0521	0
Fruit(<4)						
Fruit(>=4)						
Flaxseed/nuts(<4)						
Flaxseed/nuts(>=4)						
Rice/pasta						
Meat						
Vegetables(<=1)						
Vegetables(>1)						
Beans						
Fish						
Fast food	-0.0909	0	-0.0886	0	-0.0937	0
Multivitamin						

Table 5: the result of analysis on male



	Linear Regression		Backward AIC		Cross-Validation LASSO	
Name of Variables	Coefficient	P-value	Coefficient	P-value	Coefficient	P-value
Gender_Female	N/A					
Gender_Male						
Age_18-29	Base line					
Age_30-39	0.5043	0	0.1684	0.0486		
Age_40-49						
Age_50&above	0.5043	0	0.5449	0	0.5277	0
Region_NA						
Region_Asia						
Region_Africa	Base line					
Region_EU						
Diet_Western						
Diet_Eastern	Base line					
Diet_Med						
Education	0.0809	0.0253	0.0732	0.0386	0.0796	0.0274
Exercise	0.0468	0.0055	0.0559	0.0006	0.0479	0.0043
Breakfast			0.0279	0.0375		
Whole grain(<=1)	0.1707	0.0325	0.1495	0.0332		
Whole grain(>1)						
Dairy product						
Coffee	-0.0349	0.0233	-0.0316	0.031	-0.0359	0.0188
Fruit(<4)						
Fruit(>=4)						
Flaxseed/nuts(<4)						
Flaxseed/nuts(>=4)						
Rice/pasta						
Meat						
Vegetables(<=1)	0.2131	0.0136	0.2167	0.0112	0.229	0.0077
Vegetables(>1)	0.052	0.04	0.0595	0.0129	0.0548	0.0238
Beans	-0.049	0.035	-0.0481	0.0312	-0.0456	0.0455
Fish						
Fast food	-0.0928	0	-0.0948	0	-0.0946	0
Multivitamin						

From the table, all three methods have agreement on significance of following variables or levels. Highlighted variable means that variables only significant in the relative data:

For female:

- Positive: Age, region North America, region Europe, exercise, breakfast, whole grain($\leq$1)
- Negative: Coffee, fast food

For male:

- Positive: Age, education, exercise, whole grain($\leq$1), vegetables($\leq$1), vegetables($>$1)
- Negative: Coffee, beans, fast food

The results indicate that:

- In female group
  - Females in North America and Europe have higher mental health score than females in Africa. The reason could be that gender equality is a problem in Africa.
  - How often females eating breakfast could also be the indicators of the mental health score.
- In male group
  - Males usually provide more income in families, and education often determine the income. Therefore, education is significant among males.
  - Males with different mental status are more likely to have difference in dietary patterns including vegetables and beans

3. Conclusions

- From the result of the analysis on the whole data set, we know that there is difference between different gender and age. People have more exercise, breakfast, and few whole grain usually have better mental status. Those who have more coffee and fast food are more likely to have lower mental health score. Age, gender, exercise, coffee and fast food showed significance in every tests and are considered to be very important predictors.
- Besides some difference in dietary pattern, the differences between young people and older people are that, gender is only significant factor in the young group, and regions and education are only significant in the older group.
- Besides some difference in dietary pattern, the main differences between females and males are that, region is only significant in females, and education is only significant in males.