BST-430 FINAL PROJECT REPORT (OPTION 1)

2018 FALL

Mengran Li, Zhirou Zhou.

Instructor: Zhengwu Zhang.

CONTENTS

1. INTRODUCTION

Option 1 of this final project is about the relationship between brain connectivity and different kinds of traits of human.

The brain connectivity is captured and traced via human brain connectomes, which is defined as the collection of white matter fiber tracts connecting different regions of the brain. Due to its biological properties and functions, it is closely related to many physiological feature and behavior trace of human. Based on this correlation, many studies were interest in human brain connectomes, trying to figure out whether and how they vary for different groups of individuals, sorted by their traits and substance exposures.

Human brain connectomes are divided into two categories, functional connectomes and structural connectomes. Early studies mainly focused on functional connectomes instead of structural connectomes, because of the difficulty of recovering reliable structural connectomes. Yet recent advances in noninvasive brain imaging and preprocessing offers us access to the data regarding brain imaging, by complicated tools that can routinely extract brain structural connectomes individually.

The sub-dataset used in this final project comes from Human Connectome Project (HCP). The HCP collected high quality imaging data of human brain and various categories of traits from about 1,200 healthy adults, which enable researchers to compare the brain circuits, behavior and genetics of human at an individual level. Its latest release contains different traits, structural MRI (sMRI) and diffusion MRI (dMRI) data, which can be accessed by Connectome DB.

In terms of this final project, we started our works by three datasets: 60*7 PC scores from seven different structural connectome features, 60 PC scores from the functional connectome and 175 traits that measure domains including cognition, substance use, motor, sensory and emotion. First, we merged them into one data frame, each row represents the data of one subject. To explore the data, we plot first three structural connectome PC scores and trait scores of specific subjects using three-dimensional scatterplot. And we chose two of the traits to study whether and how they relate to human brain connectomes, by hypothesis tests and prediction models, to explore whether human brain connectivity can predict cognition, emotion and so on.

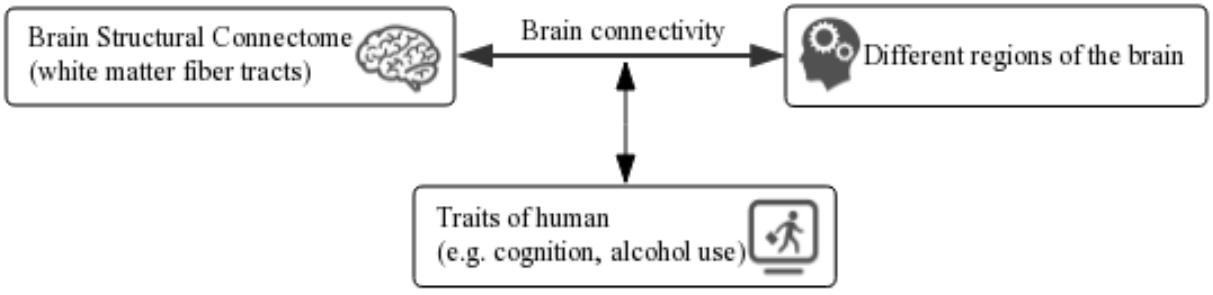


Figure 1: Background of the project

2. METHODOLOGY

2.1 Three-Way ANOVA

Analysis of variance (ANOVA) is used to analyze the differences among group means in a sample. In the typical application of ANOVA, the null hypothesis is that all groups are random samples from the same population. Rejecting the null hypothesis is taken to mean that the differences in observed effects between treatment groups are unlikely to be due to random chance.

The analysis of variance can be presented in terms of a linear model, which makes the following assumptions about the probability distribution of the responses:

1) Independence of observations – this is an assumption of the model that simplifies the statistical analysis.
2) Normality – the distributions of the residuals are normal.
3) Equality (or 'homogeneity') of variances, called homoscedasticity — the variance of data in groups should be the same.

In this project, we have 3 PC-scores as three factors and one trait as variable, so we use the three-way ANOVA.

2.2 Linear Discriminant Analysis (LDA)

As classification methods, logistic regression involves directly modeling Pr $(Y = k|X = x)$ using the logistic function. While linear discriminant analysis (LDA), as an alternative approach, models the distribution of the predictors X separately in each of the response classes, and then use Bayes' theorem to transform these into estimates for Pr $(Y = k|X = x)$.

When there are more than one predictor, the LDA classifier assumes that the observations in the kth class are drawn from a multivariate Normal distribution.

2.3 K-Nearest Neighbors (KNN)

K-Nearest Neighbors classifier is also one of the approach that attempt to estimate the conditional distribution of Y given X, and then classify a given observation to the class with highest estimated probability.

Given a positive integer K and a test observation x_0, the KNN classifier first identifies the K points in the training data that are closest to x_0, represented by $\mathcal{N}_0$. It then estimates the conditional probability for class j as the fraction of points in $\mathcal{N}_0$ whose response values equal j.

$$\Pr(Y = j|X = x_0) = \frac{1}{K}\sum_{i\in\mathcal{N}_0} I(y_i = j).$$

Finally, KNN applies Bayes rule and classifies the test observation x_0 to the class with the largest probability.

2.4 Linear Regression

Given a data set $\{y_i, x_{i1}, \dots, x_{ip}\}_{i=1}^{n}$ of n statistical units, a linear regression model assumes that the relationship between the dependent variable y and the p-vector of regressors x is linear. This relationship is modeled through a disturbance term or error variable ε — an unobserved random variable that adds "noise" to the linear relationship between the dependent variable and regressors. Thus the model takes the form

$$y_i = \beta_0 1 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \varepsilon_i = X_i^T \beta + \varepsilon_i, \qquad i = 1, \dots, n$$

so that $X_i^T \beta$ is the inner product between vectors X_i and β.

The following are the major assumptions made by standard linear regression models with standard estimation techniques:

1. Weak exogeneity. This essentially means that the predictor variables x can be treated as fixed values, rather than random variables.
2. Linearity. This means that the mean of the response variable is a linear combination of the parameters (regression coefficients) and the predictor variables.
3. Constant variance (a.k.a. homoscedasticity). This means that different values of the response variable have the same variance in their errors, regardless of the values of the predictor variables.
4. Independence of errors. This assumes that the errors of the response variables are uncorrelated with each other.
5. Lack of perfect multicollinearity in the predictors.

2.5 Random Forest

Random forests or random decision forests are an ensemble learning method for classification, regression and other tasks that operates by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees.

The random forest uses the bootstrap repeated sampling method, also known as the self-help method, which is a kind of uniform sampling from a given data set, and the effect is better when the sample is small. The actual operation extracts a certain number of samples from the original sample, allows repeated sampling; calculates a given statistic according to the extracted samples; repeats the above steps multiple times to obtain multiple calculated statistical results.

Compared to linear model, random forest regression has two advantages:

1. Random forest regression can effectively analyze nonlinear, collinear and interactive data;
2. It does not require the formal assumption of the model.

3. RESULT

3.1 Merge all Data into One Data Frame

We merged the dataset of structural connectomes, functional connectomes and traits into one big data frame, whose dimension is 1206*656, and each row represents data for one subject.

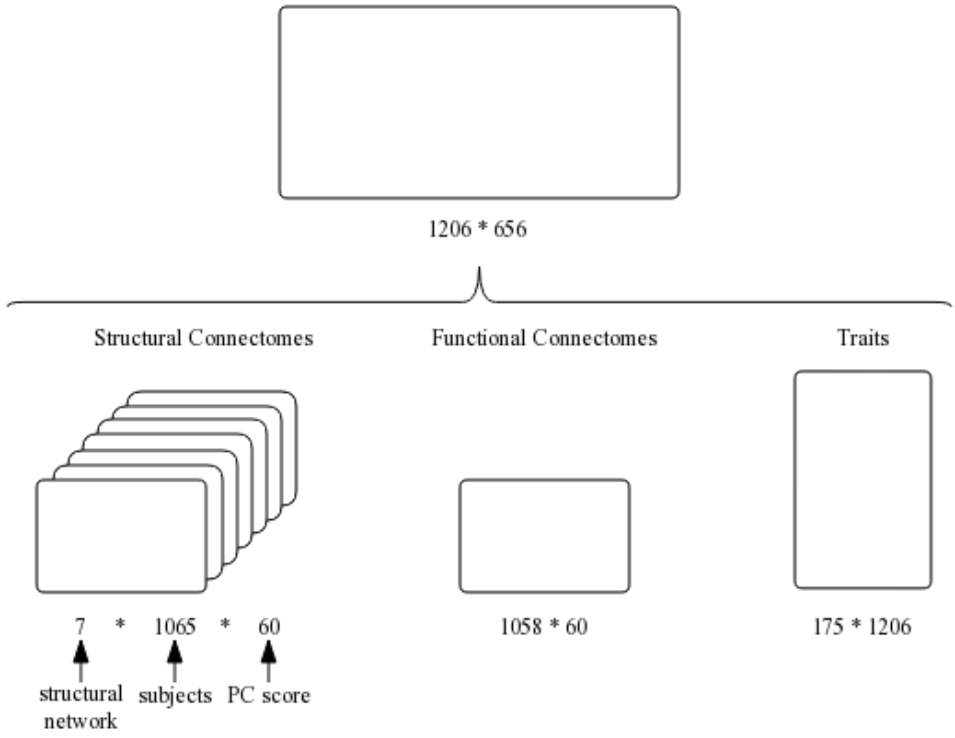


Figure 2: The process of merging the three data sets.

3.2 Data Visualization

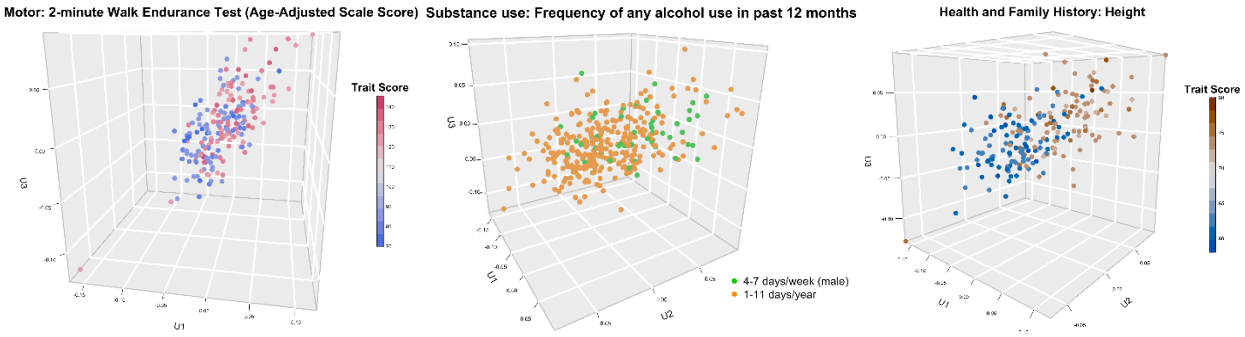
As the first step to explore the data, we visualized the data using PC scores and trait score. We sorted all subjects according to one specific trait, then selected the first and last 100 subjects. As for the connectome PC scores, we chose the second structural network in the structural connectome dataset, which is called the connected surface area (CSA) and plot its first 3 PC scores with the traits score. The traits score is reflected by the color of subjects.

The six traits that we chose to plot are shown in figure 3 and figure 4. And their details are in table 1 and table 2.

The figure 3(a) contains the trait of 2-minute walk endurance test. It belongs to the domain of motor and it is a continuous variable. From the plot, we can infer that there is a positive correlation between the first 3 PC scores and the trait score.

The second one is based on the trait of the frequency of any alcohol use in past 12 months which is from the domain of substance use. We selected the subjects whose frequencies are 4-7 days per week and 1-11 days per year as high trait score group and low trait score group. The plot shows that there is a negative correlation between the PC scores and trait score.

Figure 3(c) includes the trait called height, from the domain of health and family history. It is shown in the plot that there might be a positive relationship.



(a) (b) (c)

Figure 3: Plot of first three PC scores with high and low value trait score of one specific trait.

Table 1: Details of chosen traits.

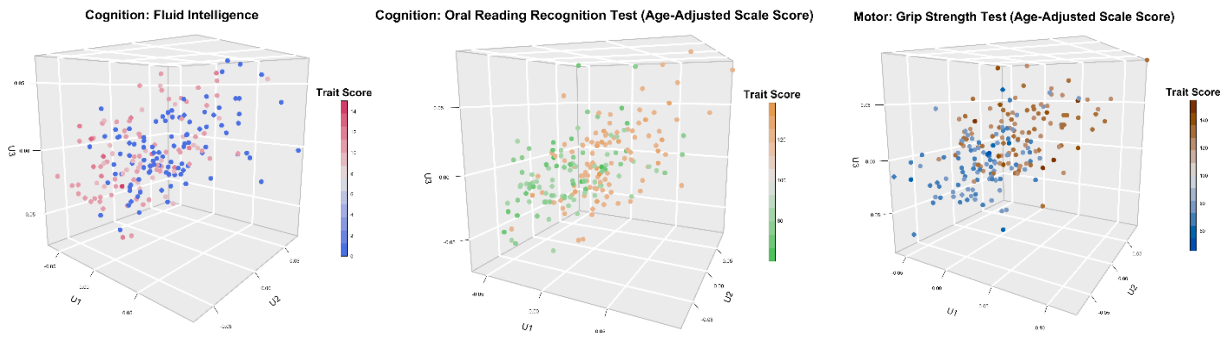
Index	Name	Category	Type	Correlation
49	2-minute Walk Endurance Test: Age-Adjusted Scale Score	Motor	Continuous	Positive
58	Frequency of any alcohol use in past 12 months	Substance Use	Ordinal	Negative
172	Height	Health and Family History	Continuous	Positive

Figure 4(a) is about the trait of 'Fluid Intelligence: Total Skipped Items'. It belongs to the domain of cognition and it is a continuous variable. We could infer from the plot that there is a negative correlation between the first 3 PC scores and the trait score.

Figure 4(b) is based on the trait of Oral Reading Recognition Test which is also from the domain of cognition. The plot shows that there is a positive correlation between the PC scores and trait score.

Figure 4(c) includes the trait of Grip Strength Test, from the domain of motor. It is shown in the plot that there might be a positive relationship.

From these scatterplots, we can observe that there is evident separation between the two groups of subjects. Indicates that the brain connectomes, which represents the brain connectivity, are different for these two groups.



(a) (b) (c)

Figure 4: Plot of first three PC scores with high and low value trait score of one specific trait.

Table 2: Details of chosen traits.

Index	Name	Category	Type	Correlation
2	Fluid Intelligence: Total Skipped Items	Cognition	Continuous	Negative
5	Oral Reading Recognition Test: Age-Adjusted Scale Score	Cognition	Continuous	Positive
52	Grip Strength Test: Age-Adjusted Scale Score	Motor	Continuous	Positive

3.3 Study the Relationship

3.3.1 Frequency of Any Alcohol Use in Past 12 Months

The first trait we want to focus on is the second one, frequency of any alcohol use in past 12 months. And the first step is to do the hypothesis test, so that we can check if the first 3 PC scores has a significant influence on the groups with low and high traits value.

Since we have three PC scores as three independent variables, we selected the 3-way ANOVA to perform the test. The model has several assumptions about the data and we need to verify them before the test.

Assumptions:

1. Errors are normally distributed;
2. Dependent variable and independent variables exhibit equal level of variance;
3. Outliers are removed

We used Henze-Zirkler's MVN test and Q-Q plot to verify the first assumption. And Bartlett test of homogeneity of variances are used to verify the second assumption.

As the result of the hypothesis test on this trait, the null hypothesis are accepted. And as shown in the Figure 5, the lines formed by points is roughly straight, indicating that the variables are normally distributed.

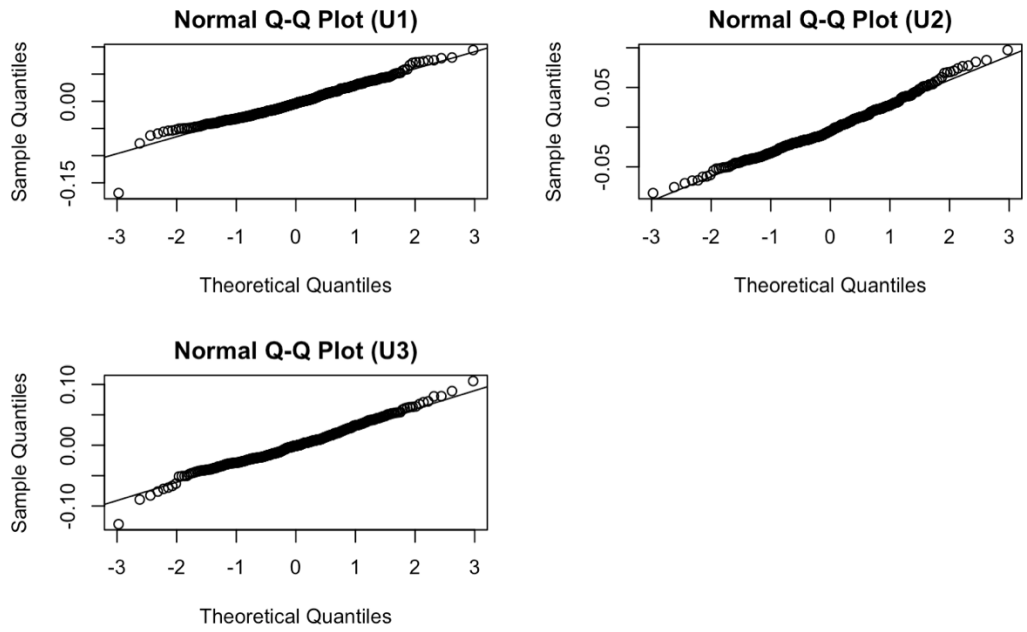


Figure 5: Normal Q-Q plot of three independent variables.

In terms of the outliers, the model identified eight observations as outliers (as shown in Figure 6). We removed them and use the new data set to perform the hypothesis test.

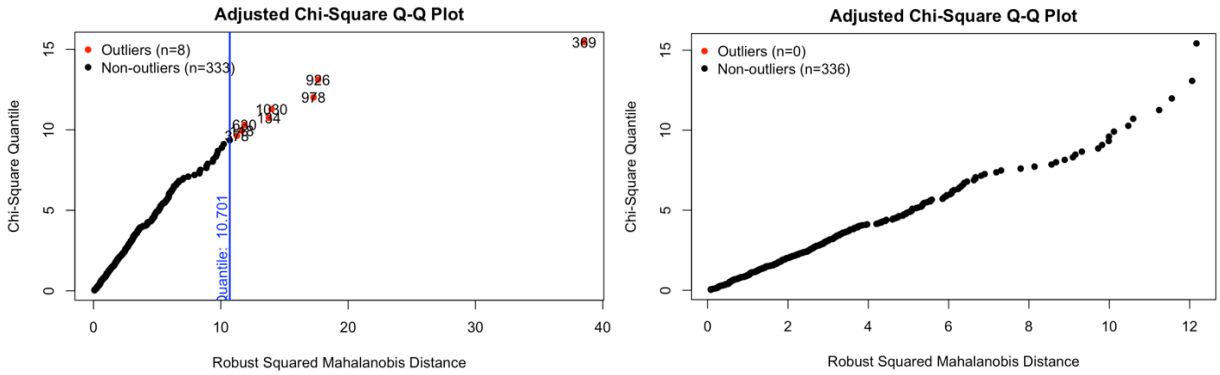


Figure 6: Adjusted Chi-Square Q-Q plot. Eight outliers are removed.

The result of 3-way ANOVA is in Figure 7. There is significant difference of the mean of first 3 PC scores between the two groups. The first PC score and the interaction of the first 3 PC scores have a significant influence to different groups.

Based on this result of hypothesis test, we can try to use the PC scores to predict the trait score.

```
             Df Sum Sq Mean Sq F value   Pr(>F)
U1            1   3.16   3.164  26.563 4.41e-07 ***
U2            1   0.25   0.250   2.099   0.1483
U3            1   0.08   0.076   0.635   0.4261
U1:U2         1   0.01   0.011   0.094   0.7594
U1:U3         1   0.13   0.130   1.092   0.2967
U2:U3         1   0.01   0.006   0.047   0.8293
U1:U2:U3      1   0.56   0.556   4.665   0.0315 *
Residuals   328  39.07   0.119
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Analysis of Variance Table

Model 1: group ~ U1 * U2 * U3
Model 2: group ~ U1 + U2 + U3 + U1:U2 + U1:U3 + U2:U3
  Res.Df    RSS Df Sum of Sq      F  Pr(>F)
1    328 39.067
2    329 39.623 -1  -0.55564 4.6651 0.03151 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 7: The result of 3-way ANOVA.

To build the model of prediction, we generated two sub-datasets (Table 3). Each one contains subjects with less trait scores. The third dataset is the only one can pass the test of normal distribution, so we did a LDA. While as for the other datasets, we performed KNN.

Table 3: Details of the datasets.

Dataset	Trait Score Contained	Dimension
whole dataset	1, 2, 3, 4, 5, 6	1011*31
sub dataset1	1, 2, 5, 6	718*31
sub dataset2	1, 6	341*31

Table 4 is a summary of the three models. The RMSE of predicting the sub-dataset1 (trait score = 1, 2, 5, 6) is relatively higher than predicting the whole dataset and the classification accuracy of sub dataset2 (trait score = 1, 6) is pretty high, indicating that the performance of prediction gets better as the trait score gets more extreme. Thus the classification is more effective when predicting whether a person drinks regularly or does not drink often.

Table 4: Comparison of model accuracy

Dataset	Trait Score Contained	Model	Classification accuracy	RMSE
whole dataset	1, 2, 3, 4, 5, 6	KNN	25.74%	2.157
sub dataset1	1, 2, 5, 6	KNN	42.65%	2.091
sub dataset2	1, 6	LDA	88.24%	-

3.3.2 Height

The second trait we studied is Height. The same as the first one, we checked if the first 3 PC scores has a significant influence on the groups with low and high traits value using 3-way ANOVA.

First, we used Henze-Zirkler's MVN test and Q-Q plot to verify the first assumption. And Bartlett test of homogeneity of variances are used to verify the second assumption.

As the result of the hypothesis test on this trait, the null hypothesis are accepted. And as shown in the Figure 8, the lines formed by points is roughly straight, indicating that the variables are normally distributed.

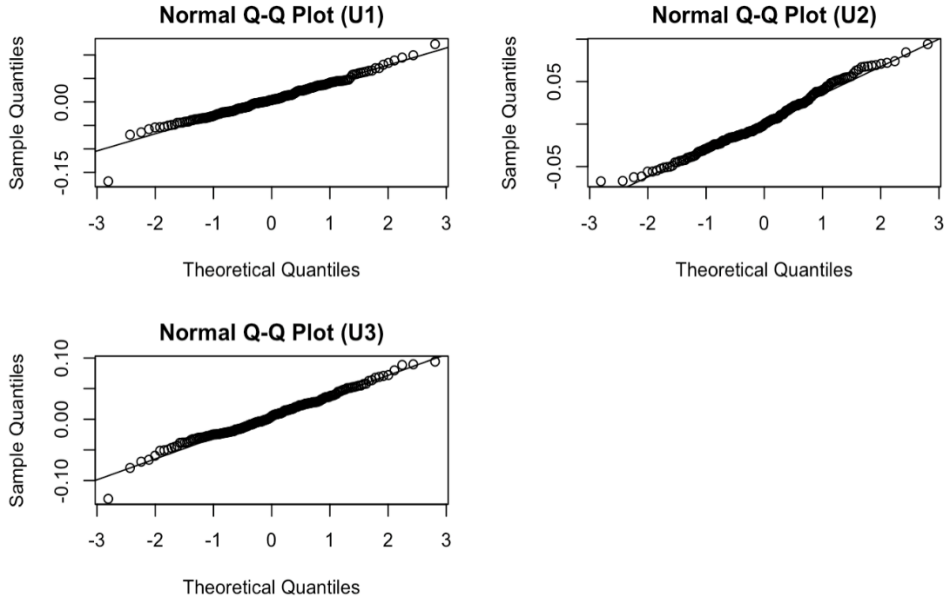


Figure 8: Normal Q-Q plot of three independent variables.

Perform the hypothesis test.

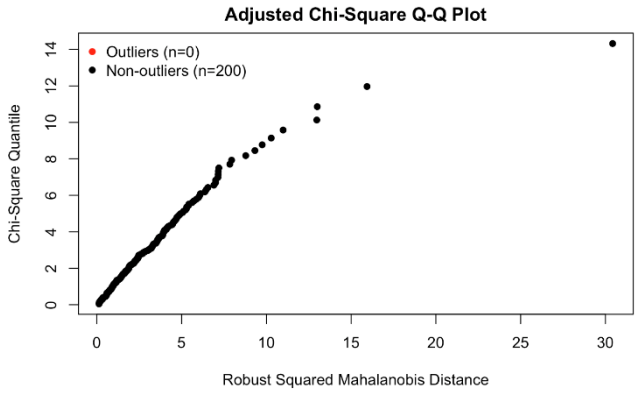


Figure 9: Chi-Square Q-Q plot.

The result of 3-way ANOVA is in Figure 10. There is significant difference of the mean of first 3 PC scores between the two groups. The first and the second PC scores and the interaction of the first 3 PC scores have a significant influence to different groups.

Based on this result of hypothesis test, we can try to use the PC scores to predict the trait score.

```
             Df Sum Sq Mean Sq F value   Pr(>F)
U1            1 12.430  12.430  86.908  < 2e-16 ***
U2            1  6.281   6.281  43.918 3.37e-10 ***
U3            1  0.454   0.454   3.173 0.076462 .
U1:U2         1  0.118   0.118   0.826 0.364559
U1:U3         1  0.408   0.408   2.852 0.092894 .
U2:U3         1  0.872   0.872   6.097 0.014413 *
U1:U2:U3      1  1.976   1.976  13.818 0.000264 ***
Residuals   192 27.461   0.143
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Analysis of Variance Table

Model 1: group ~ U1 * U2 * U3
Model 2: group ~ U1 + U2 + U3 + U1:U2 + U1:U3 + U2:U3
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1    192 27.461
2    193 29.437 -1   -1.9763 13.818 0.0002642 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 10: The result of 3-way ANOVA.

First, we used linear regression model to predict.

We built two models which the first one with the whole data and the second one using the first 100 and the last 100 data. The result are as follows. In each case, we used 3 PC-scores and 10 PC-scores to build model, and compared their goodness of regression.

Table 5: Regression result using the whole data.

PC-scores	Adjusted R-squared	p-value	RMSE
3	23.37%	< 2.2e-16	11.72
10	28.70%	< 2.2e-16	11.06

Table 6: Regression result using the first 100 and last 100 data.

PC-scores	Adjusted R-squared	p-value	RMSE
3	32.99%	3.69e-14	26.46
10	39.09%	3.85e-14	24.72

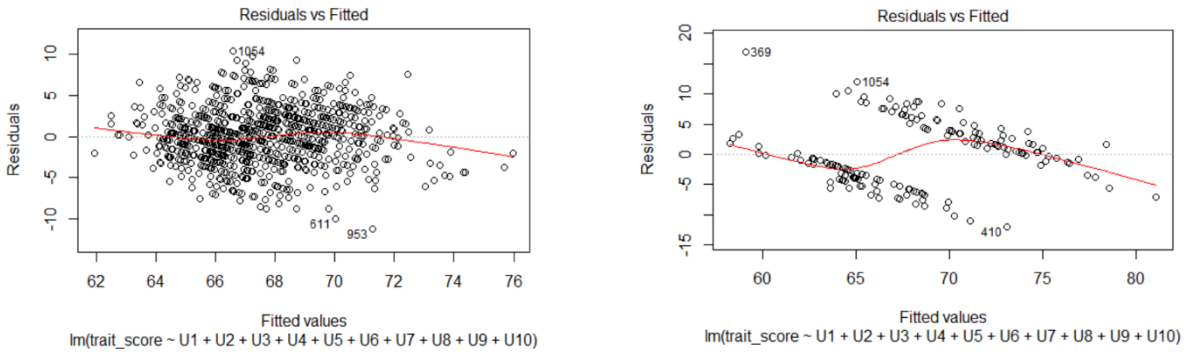


Figure 11: The residuals of regression models.

From the result, we can find that, model built from 10 PC-scores is much better than the 3 PC-scores one. On the other hand, though the R-squared is improved by using the two-group data, the RMSE is at the same time get larger. So we prefer the model built from the whole 10-dimension data.

Then we use random forest regression method to do the predict. The result are as follows.

Table 7: Regression result using the whole data.

PC-scores	% Var explained	Mean of squared residuals
3	16.54	13.22
10	25.37	11.82

Table 8: Regression result using the first 100 and last 100 data.

PC-scores	% Var explained	Mean of squared residuals
3	26.45	35.12
10	34.94	31.06

We can easily notice that the result is quite similar to the linear model. Though the %Var explained is improved by using the two-group data, the RMSE is at the same time get larger. So we still choose the model built from the whole data. And the same, model built from10 PC-scores is better.

Then we prepared the linear model and the random forest model both built from the whole data.

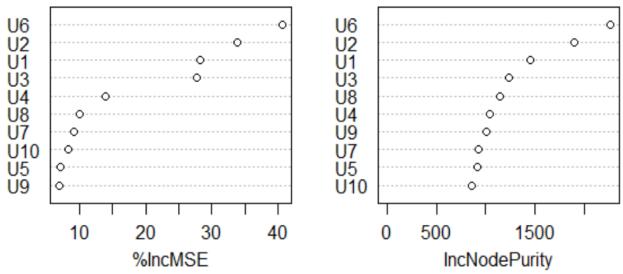


Figure 12: The efficient factors in the random forest regression models.

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  67.4865     0.1155 584.429  < 2e-16 ***
U1            8.6923    12.5197   0.694   0.4877
U2           43.2139     9.1203   4.738 2.53e-06 ***
U3           35.6382     6.1631   5.783 1.04e-08 ***
U4           10.5792     7.8532   1.347   0.1783
U5           -7.3983     5.7785  -1.280   0.2008
U6           33.6539     5.9357   5.670 1.97e-08 ***
U7          -12.1582     6.5504  -1.856   0.0638 .
U8           26.8483     6.6291   4.050 5.59e-05 ***
U9            8.9367     4.0607   2.201   0.0280 *
U10          -5.0743     5.1988  -0.976   0.3293
```

Figure 13: The coefficients of the linear regression models.

We can see that U6, U2, U1, U3 contribute a lot to the value of the trait in random forest model, while U2, U3, U6 and U8 are most powerful factors in the linear regression model. They are almost similar. And this result suits the result of the hypothesis test, too. (Figure 13)

3.3.3 Grip Strength Test

The last trait we studied is Grip Strength Test . We checked if the first 3 PC scores has a significant influence on the groups with low and high traits value using 3-way ANOVA. The result of this trait is quite similar to the second trait.

First, we used Henze-Zirkler's MVN test and Q-Q plot to verify the first assumption. And Bartlett test of homogeneity of variances are used to verify the second assumption.

As the result of the hypothesis test on this trait, the null hypothesis are accepted. And as shown in the Figure 14, the lines formed by points is roughly straight, indicating that the variables are normally distributed.

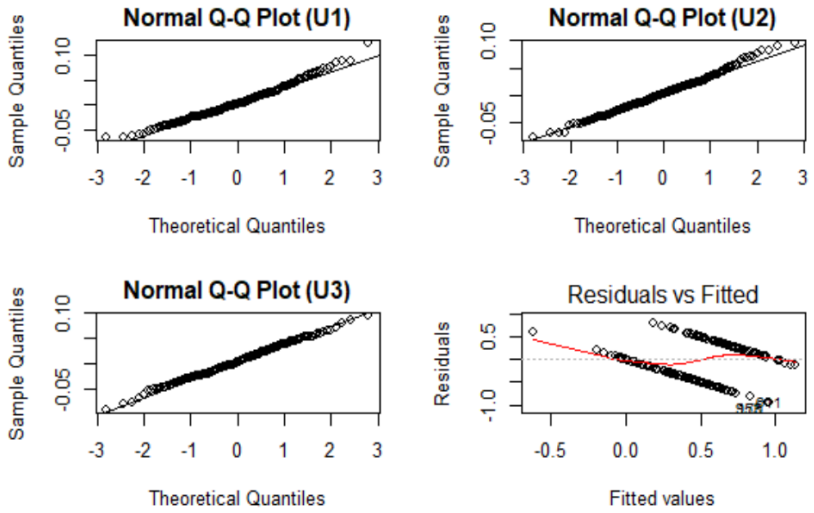


Figure 14: Normal Q-Q plot of three independent variables.

Perform the hypothesis test.

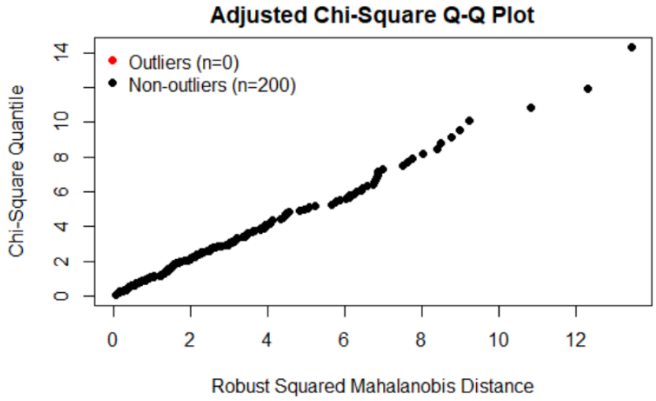


Figure 15: Chi-Square Q-Q plot.

The result of 3-way ANOVA is in Figure 15. There is significant difference of the mean of first 3 PC scores between the two groups. These three PC scores and the interaction of the first and third PC scores have a significant influence to different groups.

Based on this result of hypothesis test, we can try to use the PC scores to predict the trait score.

```
             Df Sum Sq Mean Sq F value   Pr(>F)
U1            1  10.27  10.271  60.019 5.29e-13 ***
U2            1   1.30   1.304   7.618  0.00634 **
U3            1   4.10   4.100  23.956 2.08e-06 ***
U1:U2         1   0.11   0.113   0.657  0.41849
U1:U3         1   1.01   1.005   5.875  0.01628 *
U2:U3         1   0.00   0.000   0.001  0.97792
U1:U2:U3      1   0.35   0.348   2.036  0.15525
Residuals   192  32.86   0.171
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Analysis of Variance Table

Model 1: group ~ U1 + U2 + U3 + U1:U3
Model 2: group ~ U1 + U2 + U3
  Res.Df    RSS Df Sum of Sq      F  Pr(>F)
1    195 33.316
2    196 34.325 -1   -1.0095 5.9087 0.01597 *
```

Figure 16: The result of 3-way ANOVA.

The same as the trait "Height", firstly, we used linear regression model to predict.

We also built two models which the first one with the whole data and the second one using the first 100 and the last 100 data. In each case, we used 3 PC-scores and 10 PC-scores to build model, and compared their goodness of regression. The result are as follows.

Table 9: Regression result using the whole data.

PC-scores	Adjusted R-squared	p-value	RMSE
3	17.83%	< 2.2e-16	353.3
10	25.72%	< 2.2e-16	324.6

Table 10: Regression result using the first 100 and last 100 data.

PC-scores	Adjusted R-squared	p-value	RMSE
3	32.39%	7.37e-14	833.8
10	41.87%	1.48e-15	664.6

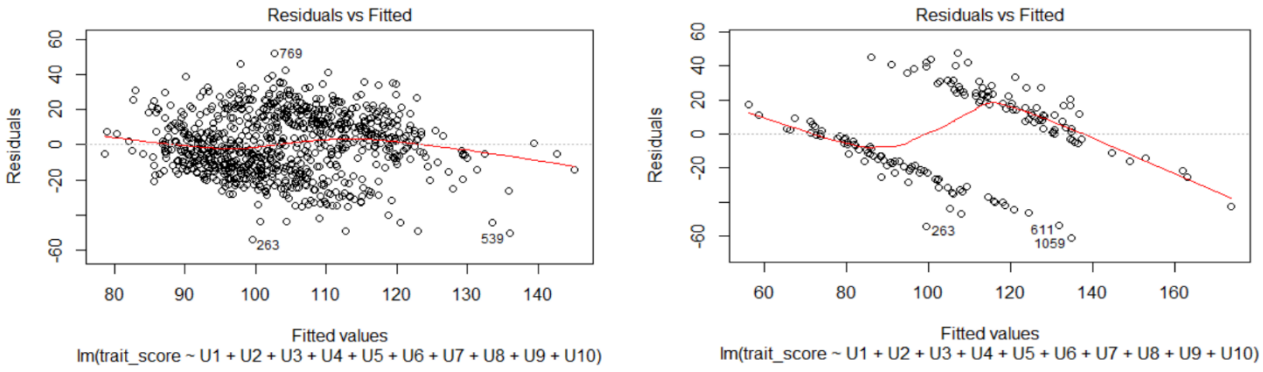


Figure 17: The residuals of regression models.

The same situation came out. We found that, model built from 10 PC-scores is much better than the 3 PC-scores one. On the other hand, though the R-squared is improved by using the two-group data, the RMSE is at the same time get larger. So we prefer the model built from the whole 10-dimension data.

Then we use random forest regression method to do the predict. The result are as follows.

Table 11: Regression result using the whole data.

PC-scores	% Var explained	Mean of squared residuals
3	12.58	354.5
10	22.51	314.2

Table 12: Regression result using the first 100 and last 100 data.

PC-scores	% Var explained	Mean of squared residuals
3	24.62	785.1
10	34.78	679.2

We can easily notice that the result is quite similar to the linear model. Though the %Var explained is improved by using the two-group data, the RMSE is at the same time get larger. So we still choose the model built from the whole data. And the same, model built from10 PC-scores is better.

4. DISCUSSION AND CONCLUSIONS

1. The performance of predicting the trait of ‘Frequency of Alcohol Use’ gets better as the trait score gets more extreme.
2. The classification is more effective when predicting whether a person drinks regularly or does not drink often.
3. Although the R-squire of the model from the small sample is greater, its RMSE gets larger at the same time. We still prefer the model with smaller RMSE.
4. The human brain connectivity can partially predict the traits of human (substance use, motor and health and family history).

5. REFERENCES

Girard, Gabriel, Kevin Whittingstall, Rachid Deriche, and Maxime Descoteaux. 2014. “Towards Quantitative Connectivity Analysis: Reducing Tractography Biases.” NeuroImage 98: 266–78.

Jones, D. K., T. R. Knosche, and R. Turner. 2013. “White matter integrity, fiber count, and other fallacies: the do’s and don’ts of diffusion MRI.” *Neuroimage* 73 (June): 239-54.

Maier-Hein, Klaus H, Peter F Neher, Jean-Christophe Houde, Marc-Alexandre Côté, Eleftherios Garyfallidis, Jidan Zhong, Maxime Chamberland, et al. 2017. “The Challenge of Mapping the Human Connectome Based on Diffusion Tractography.” Nature Communications 8 (1). Nature Publishing Group: 1349.

Park, Hae-Jeong, and Karl Friston. 2013. “Structural and Functional Brain Networks: From Connections to Cognition.” *Science* 342 (6158). American Association for the Advancement of Science: 1238411.

Price, Cathy J. 2012. “A Review and Synthesis of the First 20 Years of PET and fMRI Studies of Heard Speech, Spoken Language and Reading.” Neuroimage 62 (2). Elsevier: 816–47.

Van Essen, David C, Stephen M Smith, Deanna M Barch, Timothy EJ Behrens, Essa Yacoub, Kamil Ugurbil, Wu-Minn HCP Consortium, and others. 2013. “The WU-Minn Human Connectome Project: An Overview.” Neuroimage 80. Elsevier: 62–79.

Van Essen, David C, Kamil Ugurbil, E Auerbach, D Barch, TEJ Behrens, R Bucholz, Acer Chang, et al. 2012. “The Human Connectome Project: A Data Acquisition Perspective.” Neuroimage 62 (4). Elsevier: 2222–31.

Zhang, Zhengwu, Maxime Descoteaux, Jingwen Zhang, Gabriel Girard, Maxime Chamberland, David Dunson, Anuj Srivastava, and Hongtu Zhu. 2018. “Mapping Population-Based Structural Connectomes.” Neuroimage, to Appear.