BST-430 FINAL PROJECT PRESENTATION (OPTION 1)
2018 FALL

Mengran Li, Zhirou Zhou

Instructor: Zhengwu Zhang

CONTENTS

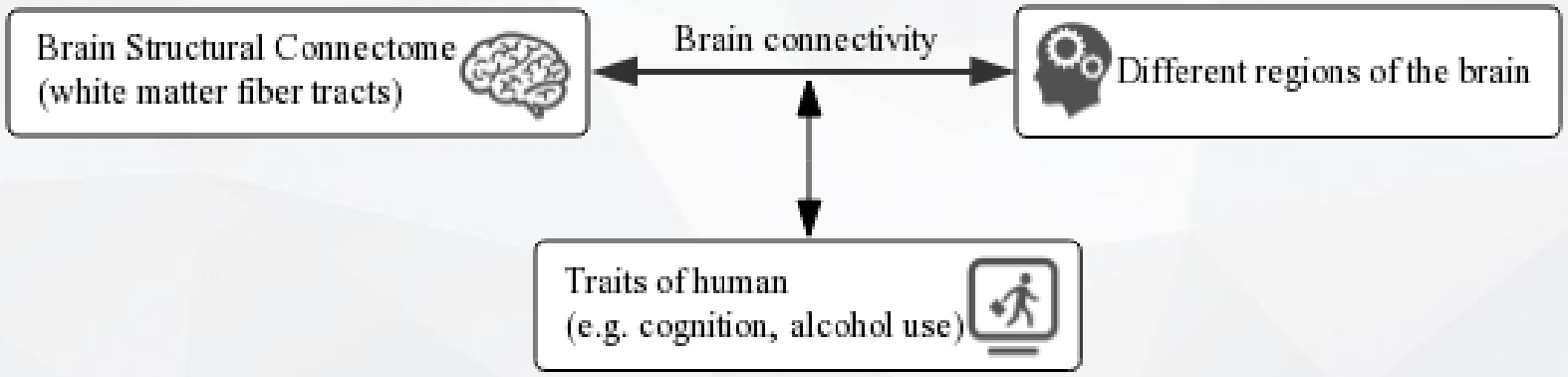
Introduction I

- relationship between brain connectivity and different kinds of traits of human

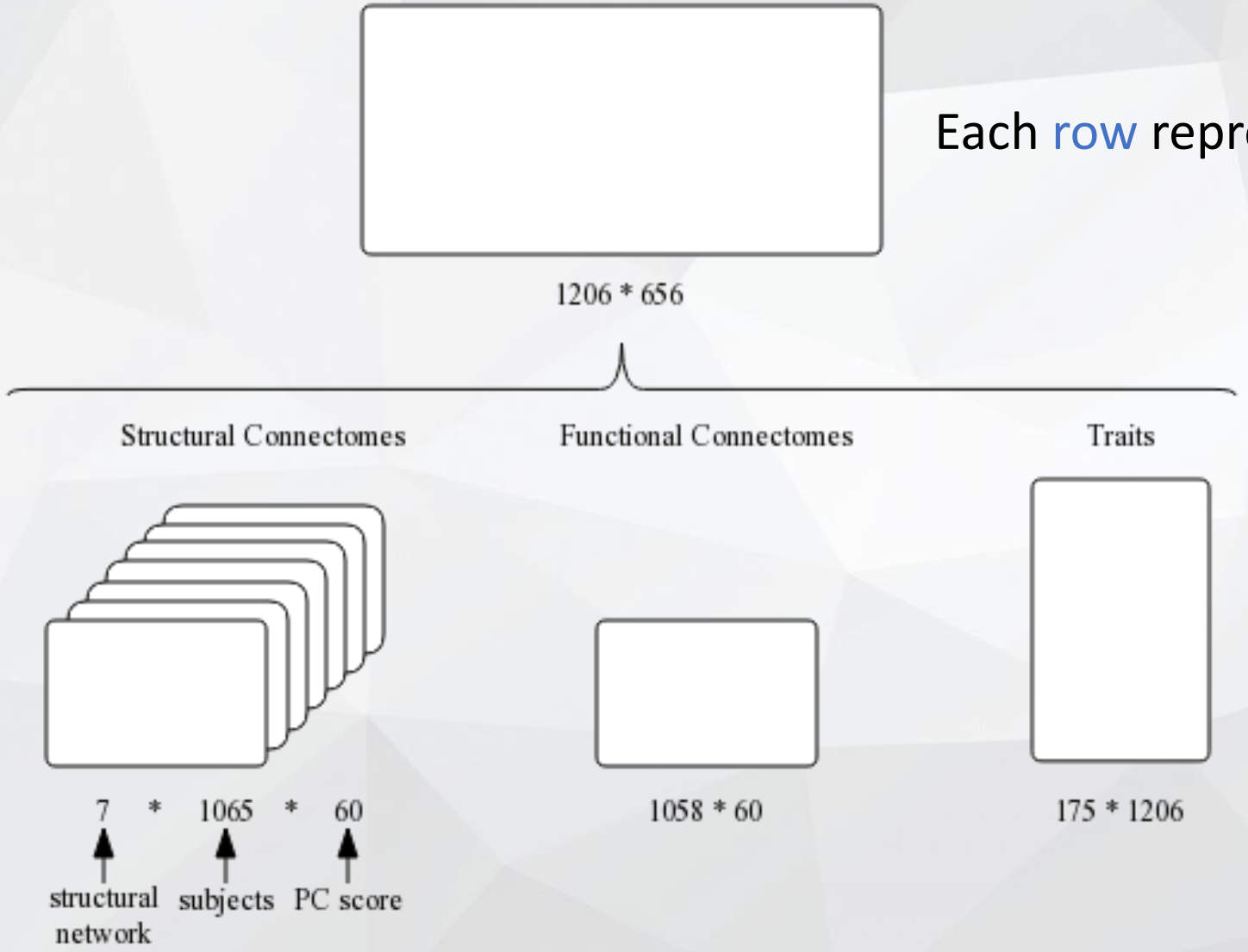
- human brain connectomes: collection of white matter fiber tracts connecting different regions of the brain
- functional connectomes and structural connectomes
- Human Connectome Project (HCP)

- explore whether and how human brain connectivity can predict the traits of human (cognition, emotion and so on)

Introduction II

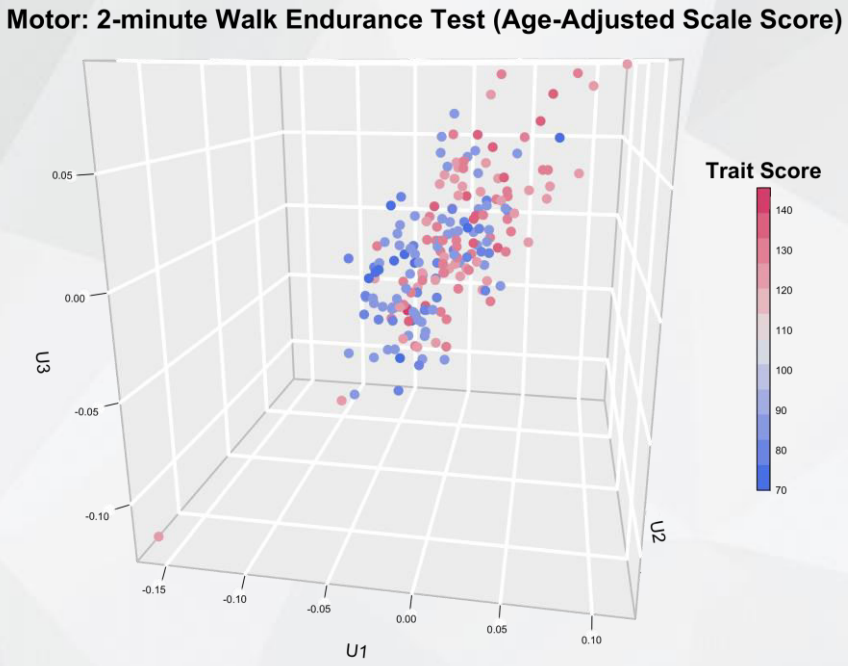


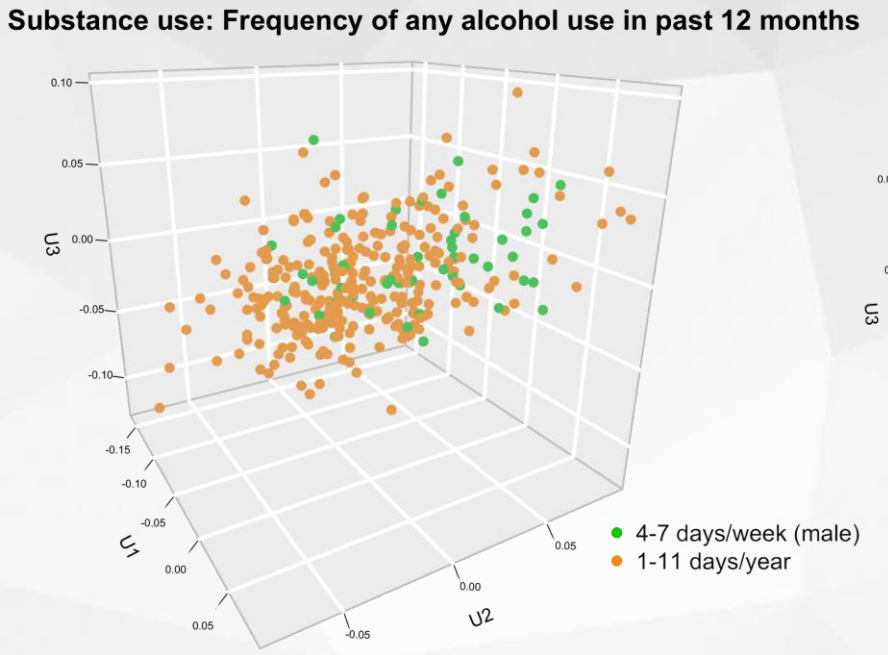
Result: merge all data into one data frame



Each row represents data for one subject

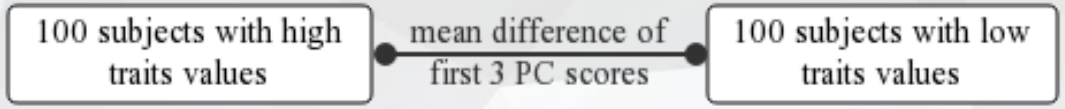
Result: plots





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

Result: hypothesis test



- Test the mean difference of first 3 PC scores between the two groups
- 3-way ANOVA
- Assumptions:
  - Errors are normally distributed (Henze-Zirkler's MVN test, Q-Q plots)
  - Dependent variable and independent variables exhibit equal level of variance (Bartlett Test of Homogeneity of Variances)
  - Outliers removed

Result: hypothesis test - Frequency of any alcohol use in past 12 months

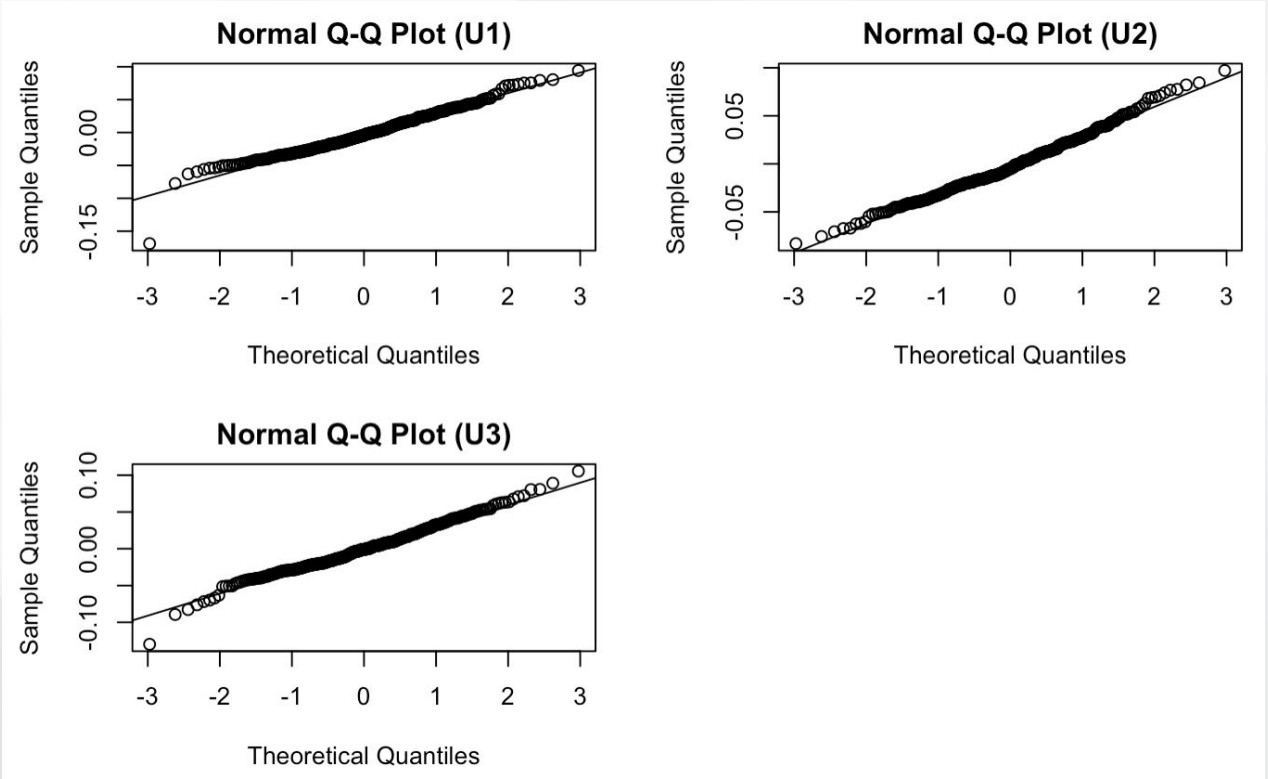
Test <fctr>	HZ <dbl>	p value <dbl>	MVN <fctr>
Henze-Zirkler	0.968079	0.1268342	YES

```
        Bartlett test of homogeneity of variances

data:  U1 + U2 + U3 by group
Bartlett's K-squared = 0.2133, df = 1, p-value = 0.6442
```

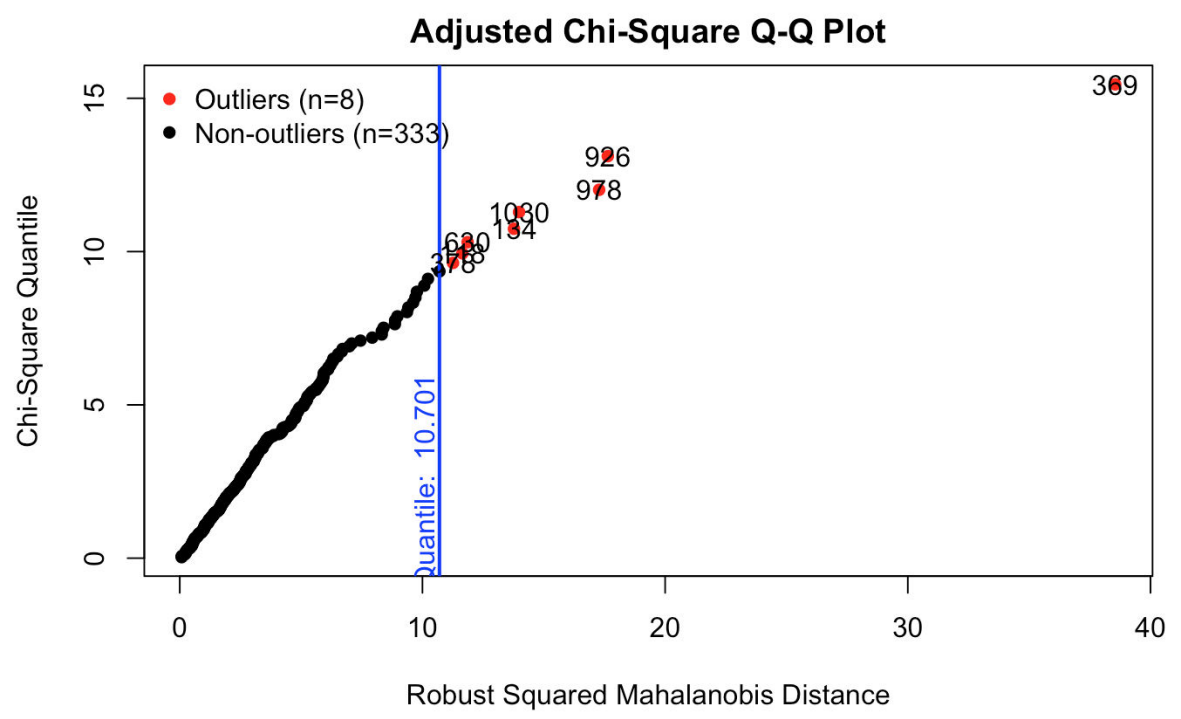
```
        Bartlett test of homogeneity of variances

data:  interaction(U1, U2) by group
Bartlett's K-squared = 0.63708, df = 1, p-value = 0.4248
```

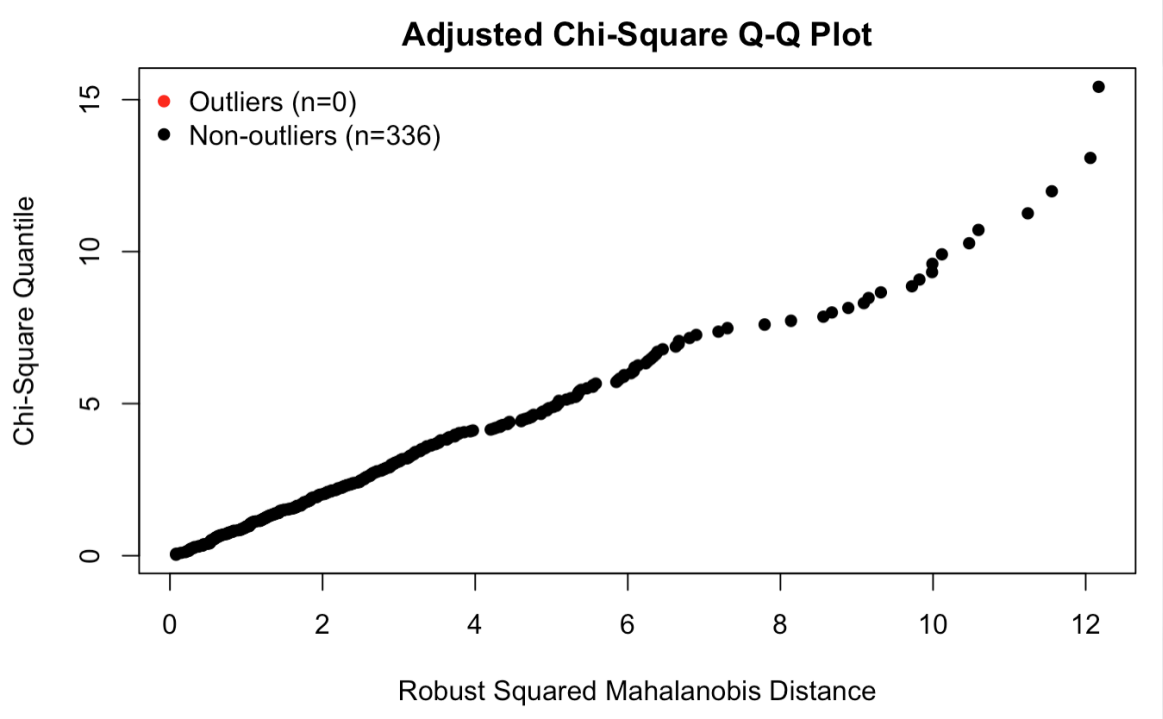
```
        Bartlett test of homogeneity of variances

data:  interaction(U2, U3) by group
Bartlett's K-squared = 0.72329, df = 1, p-value = 0.3951
```

```
        Bartlett test of homogeneity of variances

data:  interaction(U1, U3) by group
Bartlett's K-squared = 0.72505, df = 1, p-value = 0.3945
```



Result: hypothesis test - Frequency of any alcohol use in past 12 months





Result: hypothesis test - Frequency of any alcohol use in past 12 months

```
             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
```

There is significant difference between the mean of first 3 PC scores between the two groups

Result: Datasets - Frequency of any alcohol use in past 12 months

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

Result: LDA - Frequency of any alcohol use in past 12 months

	Predicted Group	
Actual Group	1	6
1	3	5
6	3	57

[1] 0.8823529 = Classification accuracy

Result: KNN - Frequency of any alcohol use in past 12 months

	Predicted Group					
Actual Group	1	2	3	4	5	6
1	1	0	0	0	3	4
2	3	1	1	4	7	13
3	2	2	5	1	8	15
4	3	2	4	2	16	7
5	2	7	4	2	21	17
6	0	5	1	2	15	22

[1] 0.2574257
[1] 2.159483

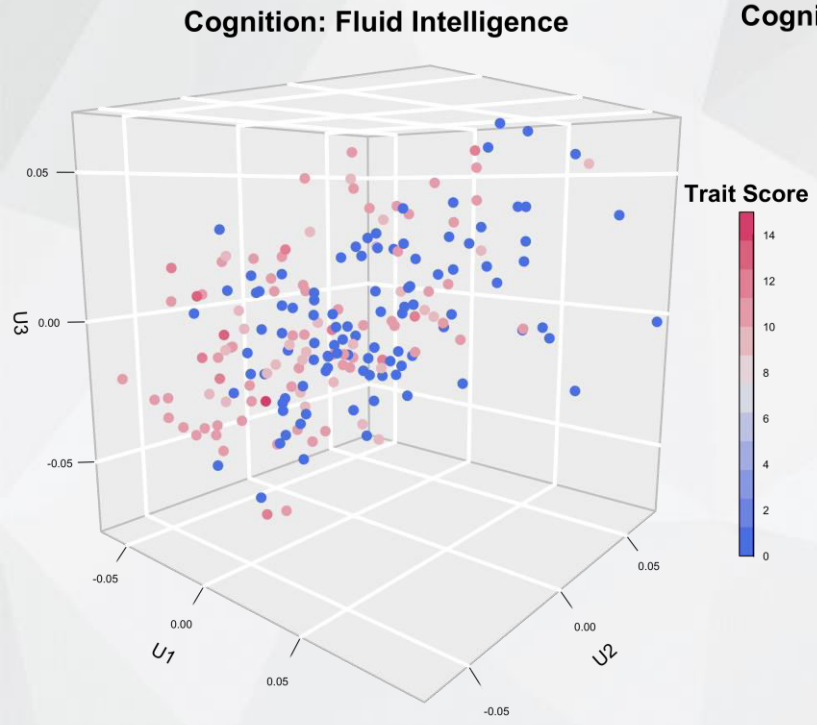
Classification accuracy
RMSE (root-mean-square error)

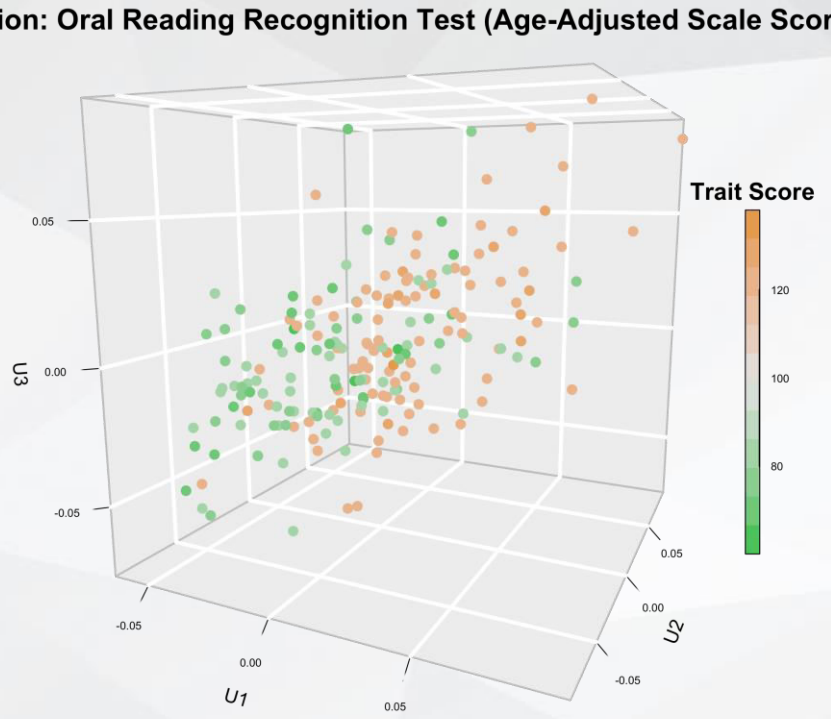
	Predicted Group			
Actual Group	1	2	5	6
1	0	0	7	4
2	0	3	13	10
5	0	2	23	23
6	0	4	19	35

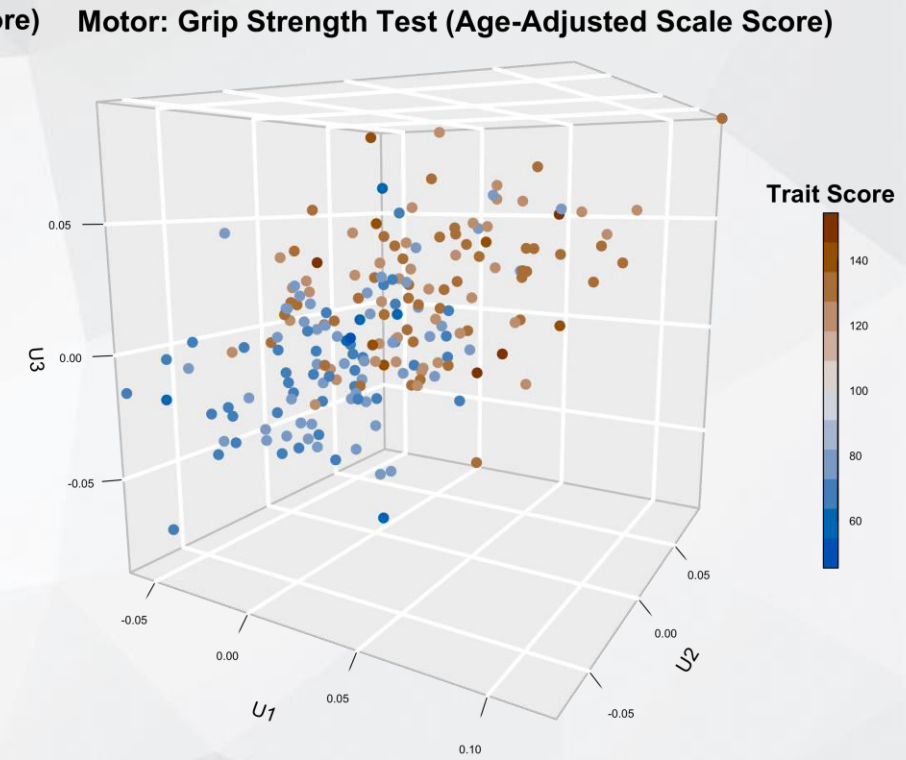
[1] 0.4265734
[1] 2.090605

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%	-

Result: plots







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

Result: hypothesis test - Height

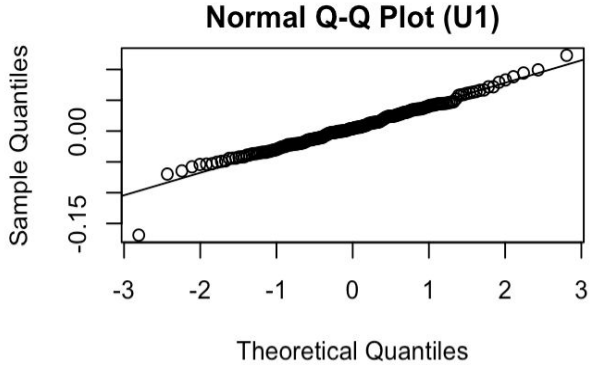
Test <fctr>	HZ <dbl>	p value <dbl>	MVN <fctr>
Henze-Zirkler	0.8876629	0.2008879	YES

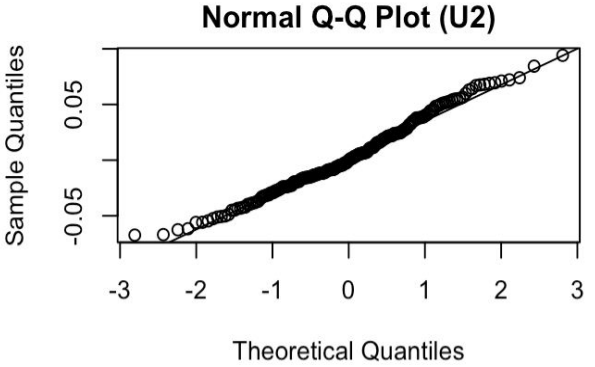
```
        Bartlett test of homogeneity of variances

data:  U1 + U2 + U3 by group
Bartlett's K-squared = 0.8247, df = 1, p-value = 0.3638
```

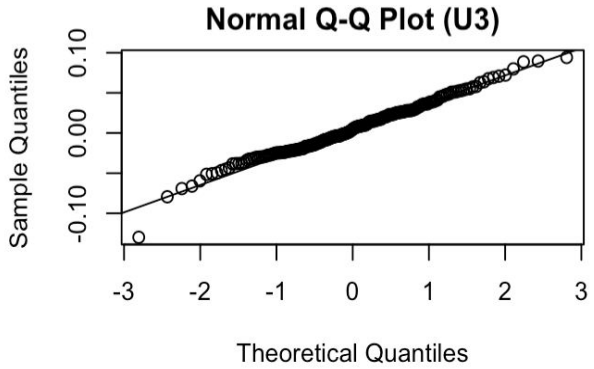
```
        Bartlett test of homogeneity of variances

data:  interaction(U1, U2) by group
Bartlett's K-squared = 0.97667, df = 1, p-value = 0.323
```

```
        Bartlett test of homogeneity of variances

data:  interaction(U1, U3) by group
Bartlett's K-squared = 0.46475, df = 1, p-value = 0.4954
```

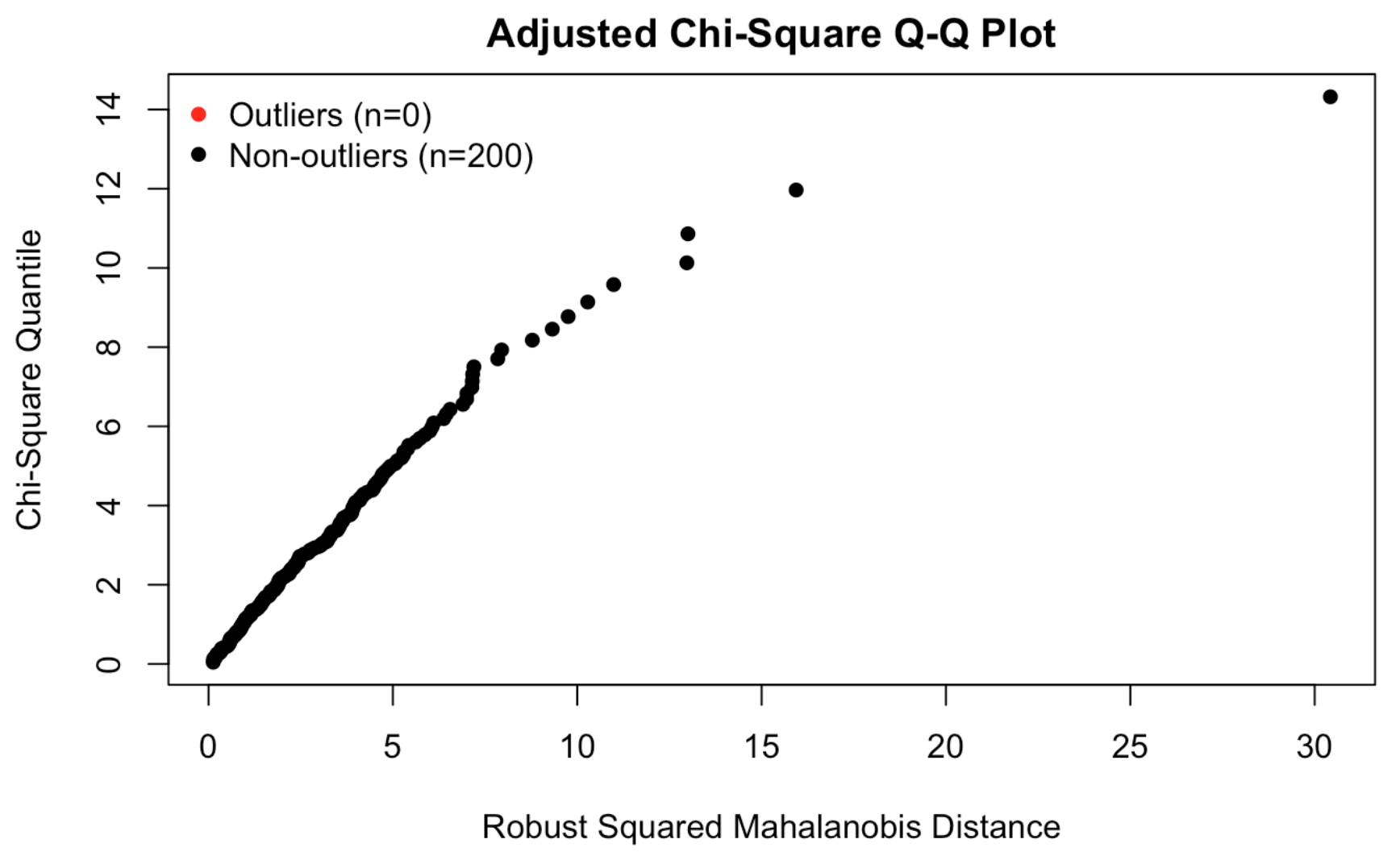
```
        Bartlett test of homogeneity of variances

data:  interaction(U2, U3) by group
Bartlett's K-squared = 0.47185, df = 1, p-value = 0.4921
```







Result: hypothesis test - Height



Result: hypothesis test - Height

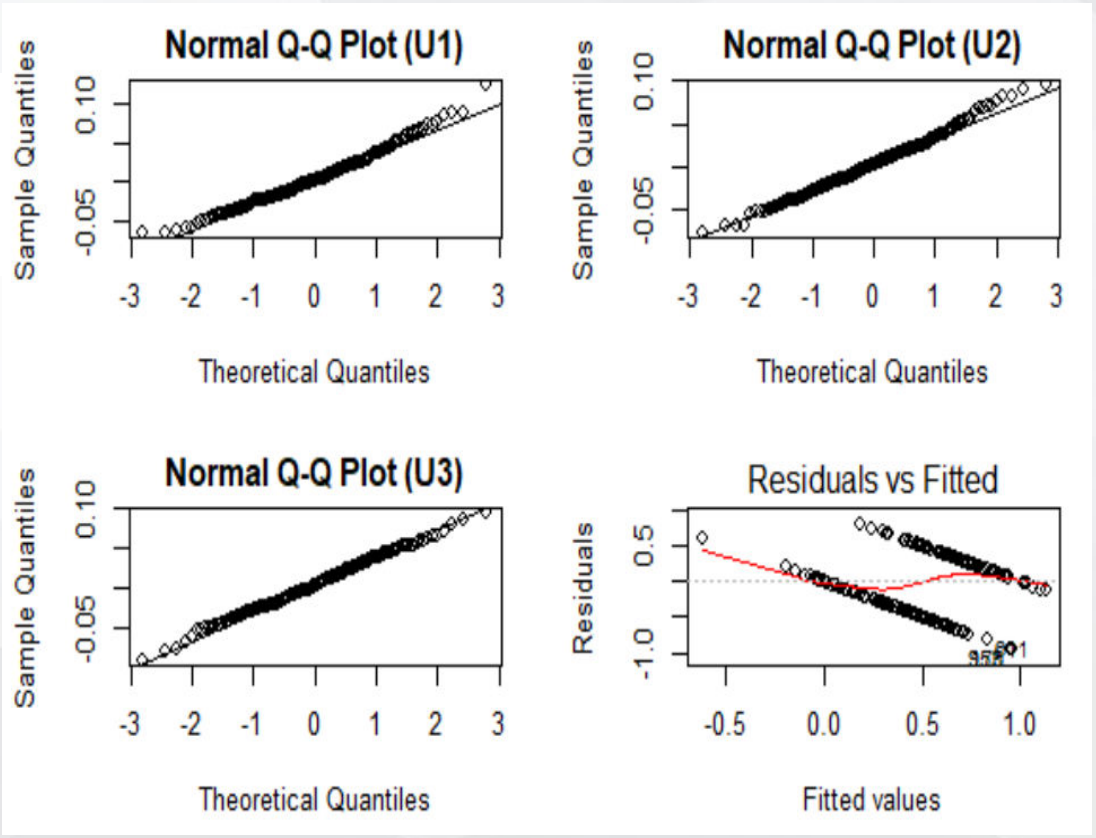
```
           Df Sum Sq Mean Sq F value   Pr(>F)
U1          1 12.430  12.430  86.908  < 2e-16 ***
U2          1  6.281   6.281  43.918 3.37e-10 ***
U3          1  0.454   0.454   3.173 0.076462 .
U1:U2       1  0.118   0.118   0.826 0.364559
U1:U3       1  0.408   0.408   2.852 0.092894 .
U2:U3       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
```

There is significant difference between the mean of first 3 PC scores between the two groups

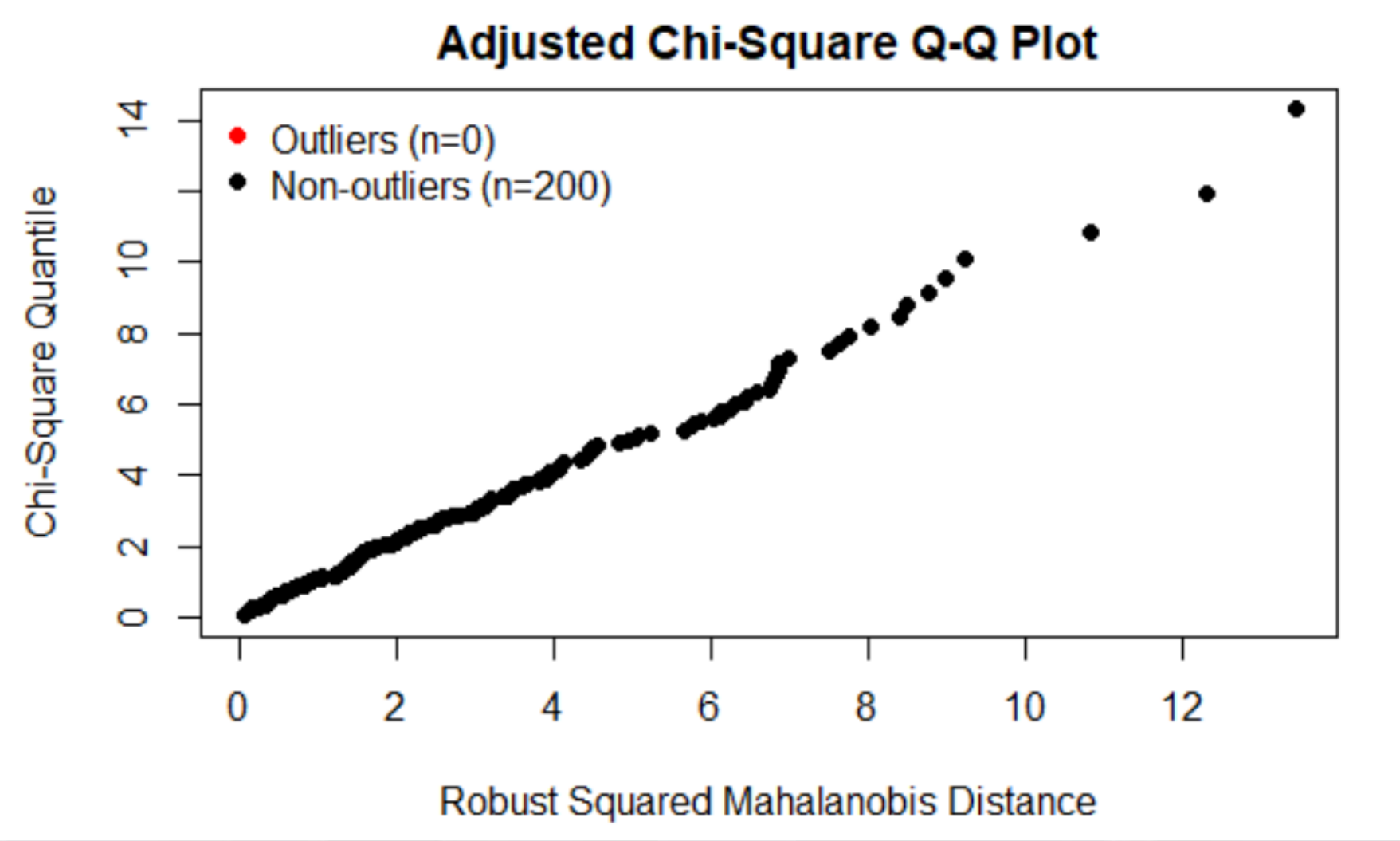
Result: hypothesis test - Grip Strength Test

Test <fctr>	HZ <dbl>	p value <dbl>	MVN <fctr>
Henze-Zirkler	0.8222594	0.3375921	YES

```
        Bartlett test of homogeneity of variances

data:  U1 + U2 + U3 by group
Bartlett's K-squared = 1.0292, df = 1, p-value = 0.3103

data:  interaction(U1, U2) by group
Bartlett's K-squared = 0.092498, df = 1, p-value = 0.761

data:  interaction(U1, U2) by group
Bartlett's K-squared = 0.092498, df = 1, p-value = 0.761

data:  interaction(U2, U3) by group
Bartlett's K-squared = 0.088031, df = 1, p-value = 0.7667
```



Result: hypothesis test - Grip Strength Test



Result: hypothesis test - Grip Strength Test

```
             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 *
```

There is significant difference between the mean of first 3 PC scores between the two groups

Result: Linear Regression - Height

- By using the whole data

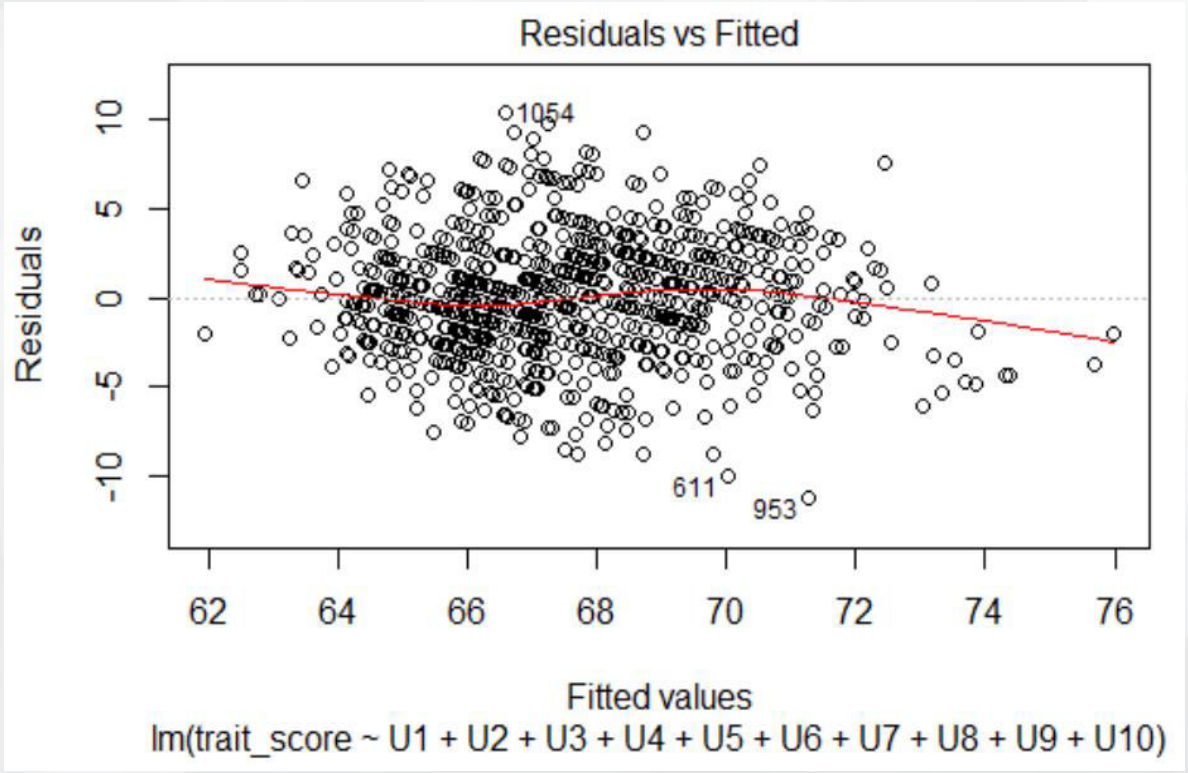
PC-scores	Adjusted R-squared	p-value	RMSE	RMSE/ (Mean of Traits from Test Data)
3	23.37%	< 2.2e-16	11.72	0.0508
10	28.70%	< 2.2e-16	11.06	0.0494

- By using the first 100 and last 100 data

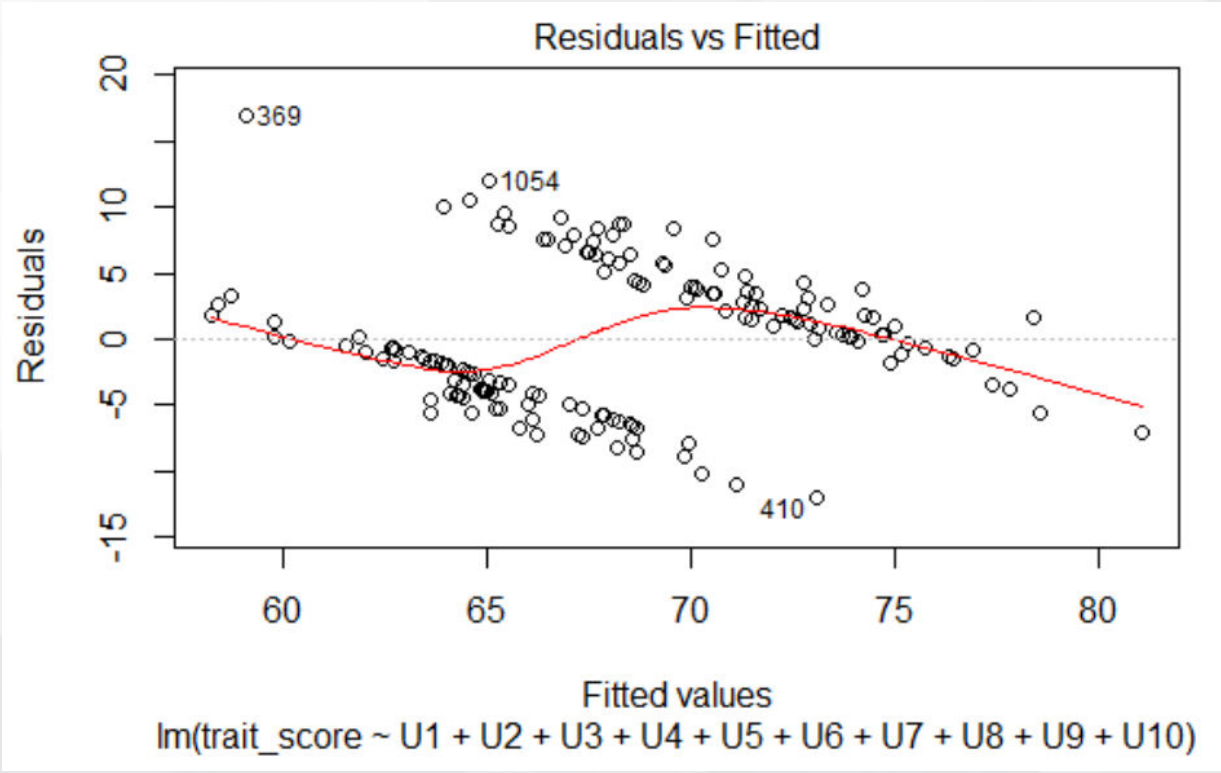
PC-scores	Adjusted R-squared	p-value	RMSE	RMSE/ (Mean of Traits from Test Data)
3	32.99%	3.69e-14	26.46	0.0776
10	39.09%	3.85e-14	24.72	0.0750

Result: Linear Regression - Height

- By using the whole data



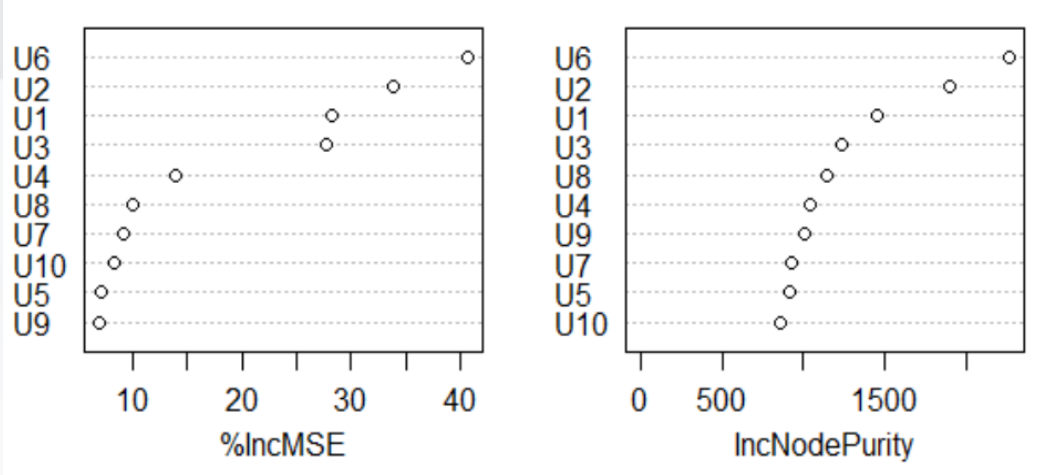
- By using the first 100 and last 100 data



Result: Random Forest Regression - Height

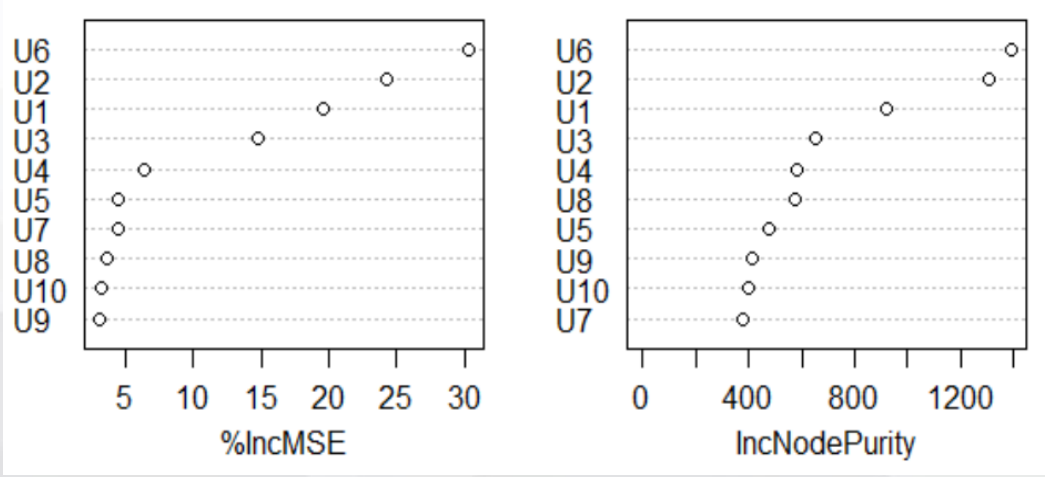
- By using the whole data

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



- By 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



Result: Linear Regression - Grip Strength Test

- By using the whole data

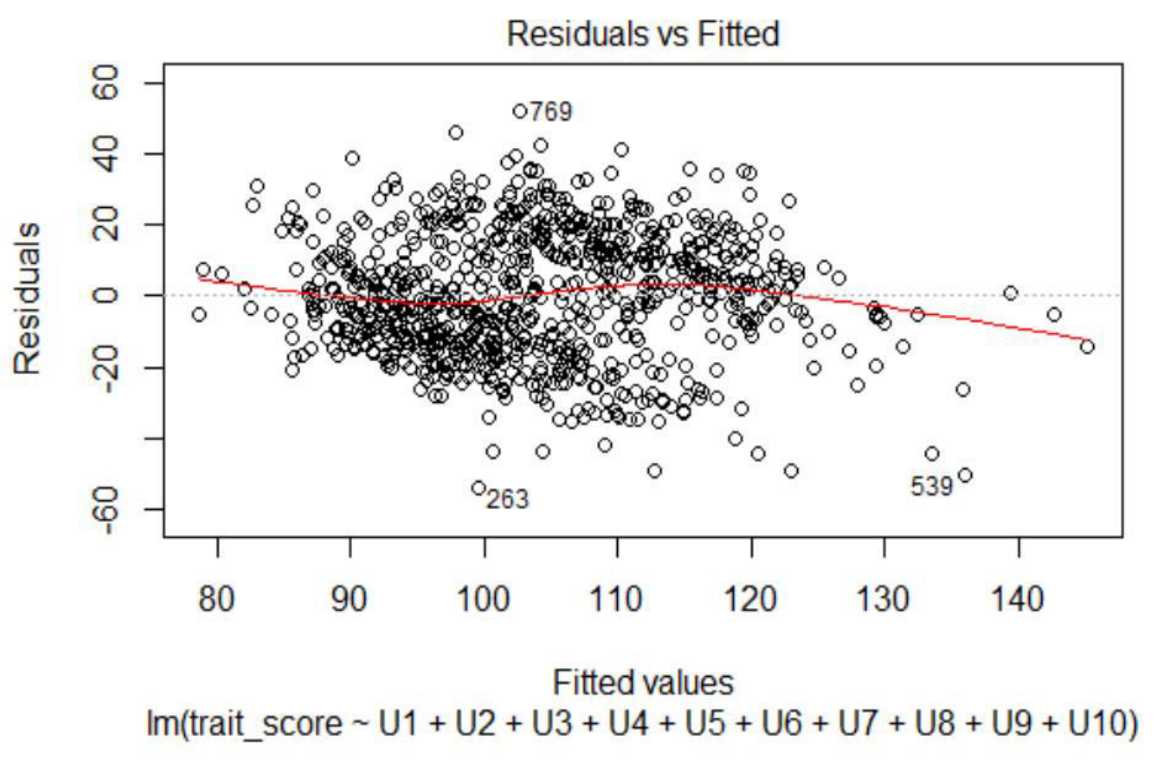
PC-scores	Adjusted R-squared	p-value	RMSE	RMSE/ (Mean of Traits from Test Data)
3	17.83%	< 2.2e-16	353.3	0.1836
10	25.72%	< 2.2e-16	324.6	0.1759

- By using the first 100 and last 100 data

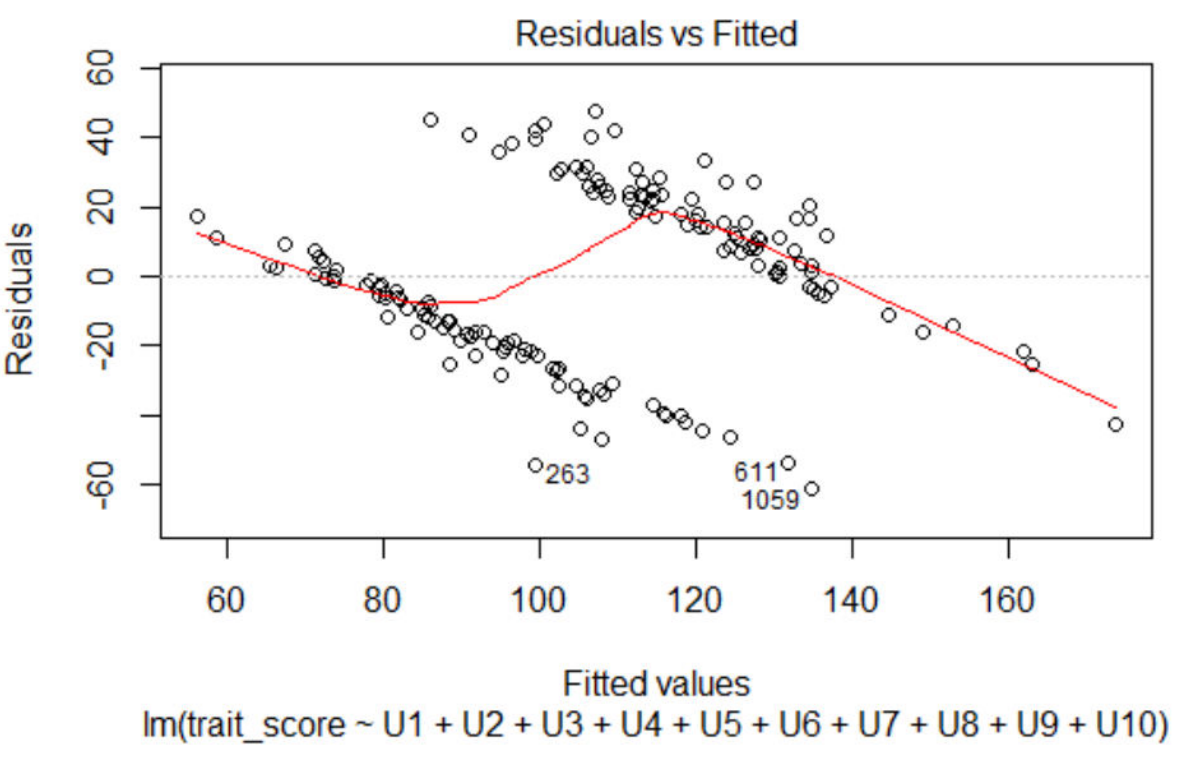
PC-scores	Adjusted R-squared	p-value	RMSE	RMSE/ (Mean of Traits from Test Data)
3	32.39%	7.37e-14	833.8	0.2943
10	41.87%	1.48e-15	664.6	0.2627

Result: Linear Regression - Grip Strength Test

- By using the whole data



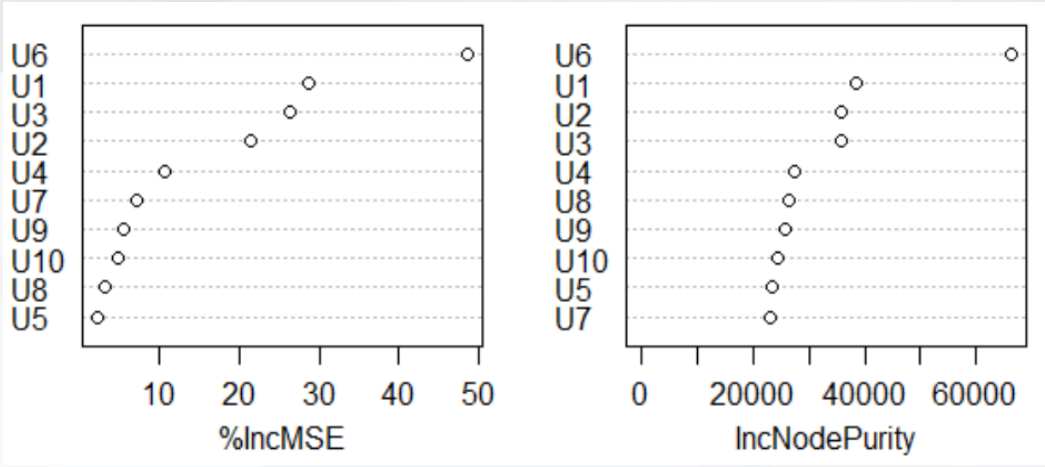
- By using the first 100 and last 100 data



Result: Random Forest Regression - Grip Strength Test

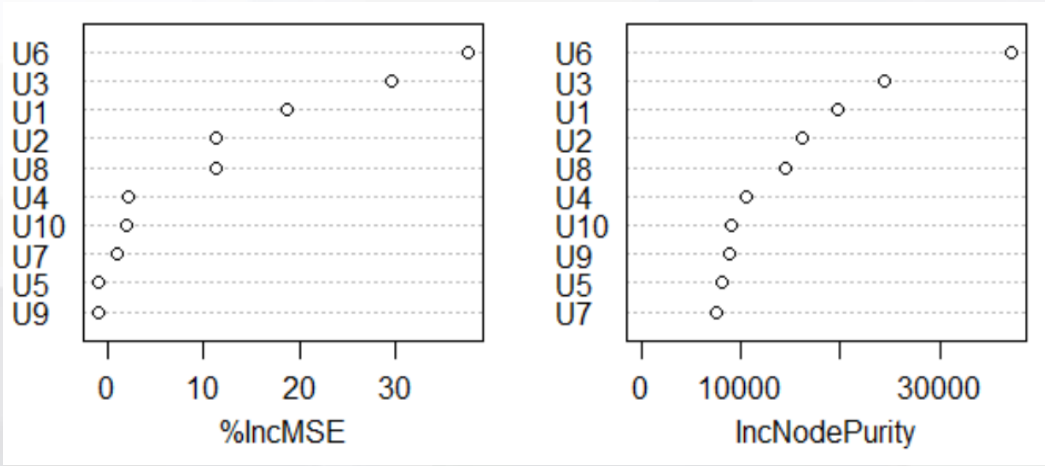
- By using the whole data

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



- By 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



Conclusion

- The RMSE of predicting the sub-dataset1 (trait score = 1, 2, 5, 6) is relatively higher than predicting the whole dataset
- 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.
- The human brain connectivity can partially predict the traits of human (substance use, motor and health and family history)

Q&A