DATA2020HW5

Zhirui Li

1. Mixed-Effect Models

(a) This is a random intercept model where the measured outcome $y_{i,j}$ is correlating with a person. The distribution of α_i suggests that the intercept for each person i is different according to a normal distribution with a mean $\alpha_0 + u_i\alpha$ and a variance σ^2_{alpha} for the response variable $y_{i,j}$. The above model is equivalent to:

$$y_{i,j} = \alpha_0 + u_i\alpha + x_{i,j}\beta + subject_i + \epsilon_{i,j} \ \ where\, subject_i \sim N(0,\, \sigma^2_\alpha)\, and\, \epsilon_{i,j} \sim N(0,\, \sigma^2_\gamma)$$

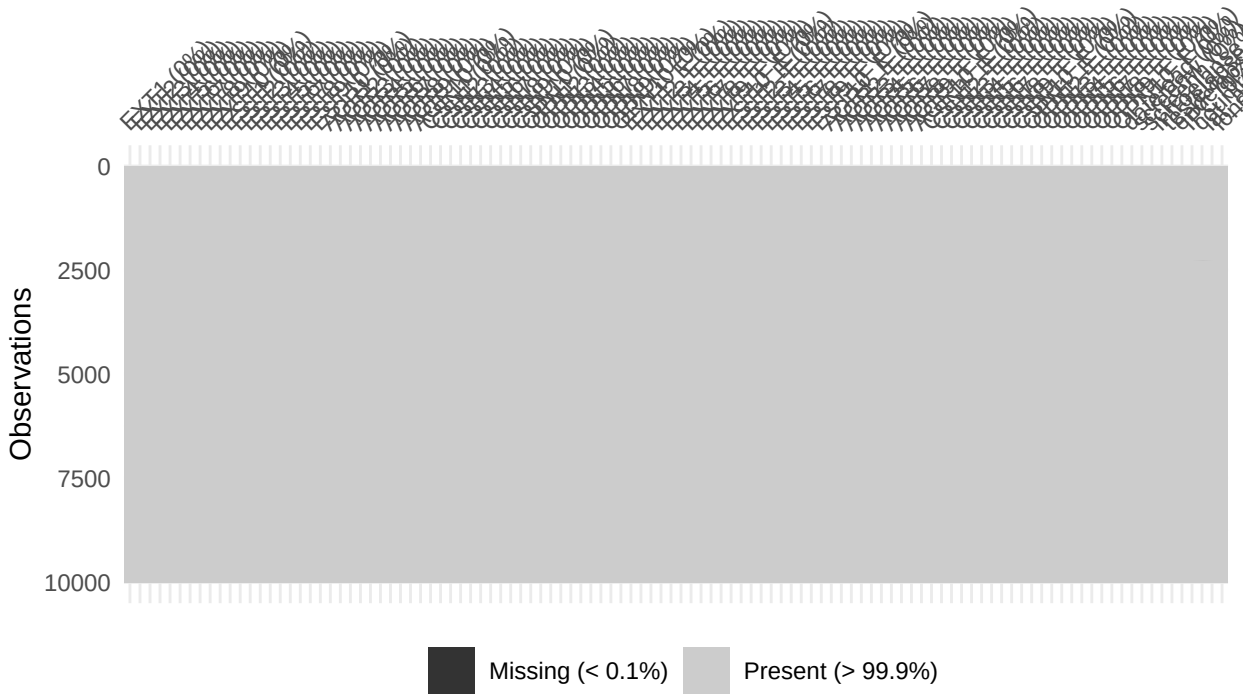
This is because instead of using α_i, we can just put the mean of α_i into the equation. Then, $subject_i$ will account for the variance of α_i. Thus, it becomes a linear regression model with random intercept accounts for different person i.

(b) The standard deviation in $y_{i,j}$ for a fixed person i is $\epsilon_{i,j}$. The standard deviation in $y_{i,j}$ for a random person i is $\epsilon_{i,j} + \sigma_\alpha$. This model allows the intercept to differ depending on person but the effect of the covariate is the same β for all persons(subjects).

2. Factor Analysis

This questions uses the data set called five_personality.csv. This data set is a subset of responses from an online personality survey that is based on the Five Personality Model. Consider only the first 50 variables of this data set and take a look to the five pers codebook.txt for variable descriptions. You can take the test yourself here https://openpsychometrics.org/tests/IPIP-BFFM/.

(a) Perform a factor analysis with 5 factors with no rotation. What is the total variance explained?



From above graph, we can see that there are only a few missing data.

```
##       EXT1            EXT2            EXT3            EXT4           EXT5
##  Min.   :0.000   Min.   :0.000   Min.   :0.000   Min.   :0.00   Min.   :0.000
##  1st Qu.:1.000   1st Qu.:2.000   1st Qu.:2.000   1st Qu.:2.00   1st Qu.:2.000
##  Median :3.000   Median :3.000   Median :3.000   Median :3.00   Median :3.000
##  Mean   :2.642   Mean   :2.791   Mean   :3.265   Mean   :3.17   Mean   :3.271
##  3rd Qu.:4.000   3rd Qu.:4.000   3rd Qu.:4.000   3rd Qu.:4.00   3rd Qu.:4.000
##  Max.   :5.000   Max.   :5.000   Max.   :5.000   Max.   :5.00   Max.   :5.000
##  NA's   :16      NA's   :16      NA's   :16      NA's   :16     NA's   :16
##       EXT6            EXT7            EXT8            EXT9
##  Min.   :0.000   Min.   :0.000   Min.   :0.000   Min.   :0.000
##  1st Qu.:1.000   1st Qu.:2.000   1st Qu.:2.000   1st Qu.:2.000
##  Median :2.000   Median :3.000   Median :4.000   Median :3.000
##  Mean   :2.405   Mean   :2.768   Mean   :3.443   Mean   :2.955
##  3rd Qu.:3.000   3rd Qu.:4.000   3rd Qu.:5.000   3rd Qu.:4.000
##  Max.   :5.000   Max.   :5.000   Max.   :5.000   Max.   :5.000
##  NA's   :16      NA's   :16      NA's   :16      NA's   :16
##      EXT10            EST1            EST2            EST3            EST4
##  Min.   :0.000   Min.   :0.00   Min.   :0.00   Min.   :0.000   Min.   :0.000
##  1st Qu.:3.000   1st Qu.:2.00   1st Qu.:2.00   1st Qu.:3.000   1st Qu.:2.000
##  Median :4.000   Median :4.00   Median :3.00   Median :4.000   Median :3.000
##  Mean   :3.573   Mean   :3.31   Mean   :3.16   Mean   :3.854   Mean   :2.642
##  3rd Qu.:5.000   3rd Qu.:4.00   3rd Qu.:4.00   3rd Qu.:5.000   3rd Qu.:4.000
##  Max.   :5.000   Max.   :5.00   Max.   :5.00   Max.   :5.000   Max.   :5.000
##  NA's   :16      NA's   :16     NA's   :16     NA's   :16      NA's   :16
##       EST5            EST6            EST7            EST8
##  Min.   :0.000   Min.   :0.000   Min.   :0.000   Min.   :0.000
##  1st Qu.:2.000   1st Qu.:2.000   1st Qu.:2.000   1st Qu.:2.000
##  Median :3.000   Median :3.000   Median :3.000   Median :3.000
##  Mean   :2.861   Mean   :2.861   Mean   :3.059   Mean   :2.695
##  3rd Qu.:4.000   3rd Qu.:4.000   3rd Qu.:4.000   3rd Qu.:4.000
```

```
##  Max.   :5.000   Max.   :5.000   Max.   :5.000   Max.   :5.000
##  NA's   :16      NA's   :16      NA's   :16      NA's   :16
##       EST9            EST10            AGR1            AGR2            AGR3
##  Min.   :0.000   Min.   :0.000   Min.   :0.000   Min.   :0.000   Min.   :0.00
##  1st Qu.:2.000   1st Qu.:2.000   1st Qu.:1.000   1st Qu.:3.000   1st Qu.:1.00
##  Median :3.000   Median :3.000   Median :2.000   Median :4.000   Median :2.00
##  Mean   :3.085   Mean   :2.786   Mean   :2.268   Mean   :3.808   Mean   :2.26
##  3rd Qu.:4.000   3rd Qu.:4.000   3rd Qu.:3.000   3rd Qu.:5.000   3rd Qu.:3.00
##  Max.   :5.000   Max.   :5.000   Max.   :5.000   Max.   :5.000   Max.   :5.00
##  NA's   :16      NA's   :16      NA's   :16      NA's   :16      NA's   :16
##       AGR4            AGR5            AGR6            AGR7
##  Min.   :0.000   Min.   :0.000   Min.   :0.000   Min.   :0.000
##  1st Qu.:3.000   1st Qu.:1.000   1st Qu.:3.000   1st Qu.:1.000
##  Median :4.000   Median :2.000   Median :4.000   Median :2.000
##  Mean   :3.923   Mean   :2.281   Mean   :3.742   Mean   :2.213
##  3rd Qu.:5.000   3rd Qu.:3.000   3rd Qu.:5.000   3rd Qu.:3.000
##  Max.   :5.000   Max.   :5.000   Max.   :5.000   Max.   :5.000
##  NA's   :16      NA's   :16      NA's   :16      NA's   :16
##       AGR8            AGR9           AGR10           CSN1            CSN2
##  Min.   :0.000   Min.   :0.000   Min.   :0.00   Min.   :0.00   Min.   :0.000
##  1st Qu.:3.000   1st Qu.:3.000   1st Qu.:3.00   1st Qu.:2.00   1st Qu.:2.000
##  Median :4.000   Median :4.000   Median :4.00   Median :3.00   Median :3.000
##  Mean   :3.671   Mean   :3.781   Mean   :3.59   Mean   :3.28   Mean   :2.938
##  3rd Qu.:4.000   3rd Qu.:5.000   3rd Qu.:4.00   3rd Qu.:4.00   3rd Qu.:4.000
##  Max.   :5.000   Max.   :5.000   Max.   :5.00   Max.   :5.00   Max.   :5.000
##  NA's   :16      NA's   :16      NA's   :16     NA's   :16     NA's   :16
##       CSN3            CSN4            CSN5            CSN6
##  Min.   :0.000   Min.   :0.000   Min.   :0.000   Min.   :0.000
##  1st Qu.:3.000   1st Qu.:2.000   1st Qu.:2.000   1st Qu.:2.000
##  Median :4.000   Median :2.000   Median :2.000   Median :3.000
##  Mean   :3.974   Mean   :2.634   Mean   :2.626   Mean   :2.848
##  3rd Qu.:5.000   3rd Qu.:4.000   3rd Qu.:4.000   3rd Qu.:4.000
##  Max.   :5.000   Max.   :5.000   Max.   :5.000   Max.   :5.000
##  NA's   :16      NA's   :16      NA's   :16      NA's   :16
##       CSN7            CSN8            CSN9           CSN10
##  Min.   :0.000   Min.   :0.000   Min.   :0.000   Min.   :0.000
##  1st Qu.:3.000   1st Qu.:2.000   1st Qu.:2.000   1st Qu.:3.000
##  Median :4.000   Median :3.000   Median :3.000   Median :4.000
##  Mean   :3.673   Mean   :2.495   Mean   :3.191   Mean   :3.589
##  3rd Qu.:5.000   3rd Qu.:3.000   3rd Qu.:4.000   3rd Qu.:4.000
##  Max.   :5.000   Max.   :5.000   Max.   :5.000   Max.   :5.000
##  NA's   :16      NA's   :16      NA's   :16      NA's   :16
##       OPN1            OPN2            OPN3            OPN4
##  Min.   :0.000   Min.   :0.000   Min.   :0.000   Min.   :0.000
##  1st Qu.:3.000   1st Qu.:1.000   1st Qu.:3.000   1st Qu.:1.000
##  Median :4.000   Median :2.000   Median :4.000   Median :2.000
##  Mean   :3.655   Mean   :2.094   Mean   :3.994   Mean   :2.006
##  3rd Qu.:5.000   3rd Qu.:3.000   3rd Qu.:5.000   3rd Qu.:3.000
##  Max.   :5.000   Max.   :5.000   Max.   :5.000   Max.   :5.000
##  NA's   :16      NA's   :16      NA's   :16      NA's   :16
##       OPN5            OPN6            OPN7            OPN8            OPN9
##  Min.   :0.000   Min.   :0.000   Min.   :0.000   Min.   :0.00   Min.   :0.000
##  1st Qu.:3.000   1st Qu.:1.000   1st Qu.:3.000   1st Qu.:2.00   1st Qu.:4.000
##  Median :4.000   Median :2.000   Median :4.000   Median :3.00   Median :4.000
```

```
##  Mean   :3.784   Mean   :1.897   Mean   :3.953   Mean   :3.19   Mean   :4.133
##  3rd Qu.:5.000   3rd Qu.:2.000   3rd Qu.:5.000   3rd Qu.:4.00   3rd Qu.:5.000
##  Max.   :5.000   Max.   :5.000   Max.   :5.000   Max.   :5.00   Max.   :5.000
##  NA's   :16      NA's   :16      NA's   :16      NA's   :16     NA's   :16
##      OPN10
##  Min.   :0.000
##  1st Qu.:3.000
##  Median :4.000
##  Mean   :3.963
##  3rd Qu.:5.000
##  Max.   :5.000
##  NA's   :16
```

We convert all the variables to numerical.

```
##
## Loadings:
##       MR1    MR2    MR3    MR4    MR5
## EXT1  -0.531  0.220 -0.312         0.236
## EXT2   0.521 -0.241  0.351        -0.150
## EXT3  -0.660  0.113 -0.142 -0.123
## EXT4   0.568         0.388        -0.194
## EXT5  -0.635  0.281 -0.192         0.197
## EXT6   0.519 -0.179  0.190
## EXT7  -0.590  0.242 -0.261         0.206
## EXT8   0.374         0.399        -0.155
## EXT9  -0.468  0.187 -0.303  0.131  0.243
## EXT10  0.563         0.355        -0.134
## EST1   0.382  0.552  0.120 -0.121  0.170
## EST2  -0.307 -0.302                -0.125
## EST3   0.301  0.533  0.203         0.104
## EST4  -0.255 -0.241
## EST5   0.310  0.396                0.168
## EST6   0.382  0.571                0.229
## EST7   0.368  0.589                0.226
## EST8   0.400  0.594                0.237
## EST9   0.423  0.454                0.358
## EST10  0.486  0.450         0.127
## AGR1   0.219 -0.204 -0.166  0.152  0.306
## AGR2  -0.456  0.363        -0.101 -0.157
## AGR3   0.204        -0.275  0.256  0.266
## AGR4  -0.282  0.487  0.330 -0.251 -0.350
## AGR5   0.328 -0.318 -0.119  0.294  0.355
## AGR6  -0.123  0.437  0.254 -0.259 -0.208
## AGR7   0.473 -0.304         0.261  0.302
## AGR8  -0.320  0.338  0.245 -0.191 -0.162
## AGR9  -0.268  0.499  0.287 -0.208 -0.249
## AGR10 -0.458  0.236  0.150
## CSN1  -0.263 -0.137  0.478         0.333
## CSN2   0.128  0.286 -0.305  0.202 -0.254
## CSN3  -0.172         0.427  0.133  0.166
## CSN4   0.365  0.379 -0.308  0.133 -0.133
## CSN5  -0.253 -0.123  0.410 -0.177  0.358
## CSN6   0.222  0.299 -0.344  0.168 -0.222
## CSN7                 0.465         0.314
```

```
## CSN8   0.302  0.203 -0.301  0.136
## CSN9  -0.213         0.445 -0.152  0.376
## CSN10 -0.212         0.407  0.161  0.240
## OPN1  -0.175         0.159  0.529
## OPN2   0.239               -0.441  0.207
## OPN3  -0.106  0.300  0.118  0.458 -0.104
## OPN4   0.174               -0.372  0.261
## OPN5  -0.337  0.112  0.149  0.516
## OPN6   0.200 -0.104        -0.369  0.158
## OPN7  -0.272         0.254  0.409
## OPN8          0.136         0.549
## OPN9          0.263  0.290  0.285
## OPN10 -0.298  0.216  0.124  0.578
##
##                  MR1   MR2   MR3   MR4   MR5
## SS loadings    6.469 4.545 3.241 2.910 2.217
## Proportion Var 0.129 0.091 0.065 0.058 0.044
## Cumulative Var 0.129 0.220 0.285 0.343 0.388
```

Above are the loading scores when we perform factor analysis for the data set five_personality.csv with first 50 features.

Here SS loadings represent the variance, or the eigenvalue for each factor. The first factor has an eigenvalue of 6.469. The proportion of variance explained by this factor is $6.469/50 = 0.129$. Here, 50 refers to the total features presented in the data set.

Since we don't used any rotation here, it is hard to understand what is each factor capturing because almost all features have a loading score for each factor.

```
##      EXT1      EXT2      EXT3      EXT4      EXT5      EXT6      EXT7      EXT8
## 0.5162660 0.5186539 0.5077056 0.4723880 0.4407954 0.6516533 0.4792585 0.6678173
##      EXT9     EXT10      EST1      EST2      EST3      EST4      EST5      EST6
## 0.5774516 0.5301869 0.4916706 0.7894292 0.5665140 0.8741385 0.7135418 0.4669658
##      EST7      EST8      EST9     EST10      AGR1      AGR2      AGR3      AGR4
## 0.4620781 0.4231093 0.4862060 0.5425051 0.7658329 0.6208330 0.7419775 0.3884229
##      AGR5      AGR6      AGR7      AGR8      AGR9     AGR10      CSN1      CSN2
## 0.5648680 0.6189763 0.5231555 0.6607344 0.4915399 0.7080107 0.5726198 0.7036798
##      CSN3      CSN4      CSN5      CSN6      CSN7      CSN8      CSN9     CSN10
## 0.7396395 0.5931826 0.5937852 0.6650561 0.6747241 0.7500744 0.5912745 0.7058013
##      OPN1      OPN2      OPN3      OPN4      OPN5      OPN6      OPN7      OPN8
## 0.6565651 0.6943049 0.6645931 0.7565696 0.5753247 0.7879038 0.6927907 0.6651293
##      OPN9     OPN10
## 0.7572185 0.5149843
```

Above table represents the remaining variance for each feature that hasn't been explained by the factors we created.

```
##      EXT1      EXT2      EXT3      EXT4      EXT5      EXT6      EXT7      EXT8
## 0.4837340 0.4813461 0.4922944 0.5276120 0.5592046 0.3483467 0.5207415 0.3321827
##      EXT9     EXT10      EST1      EST2      EST3      EST4      EST5      EST6
## 0.4225484 0.4698131 0.5083294 0.2105708 0.4334860 0.1258615 0.2864582 0.5330342
##      EST7      EST8      EST9     EST10      AGR1      AGR2      AGR3      AGR4
## 0.5379219 0.5768907 0.5137940 0.4574949 0.2341671 0.3791670 0.2580225 0.6115771
##      AGR5      AGR6      AGR7      AGR8      AGR9     AGR10      CSN1      CSN2
## 0.4351320 0.3810237 0.4768445 0.3392656 0.5084601 0.2919893 0.4273802 0.2963202
##      CSN3      CSN4      CSN5      CSN6      CSN7      CSN8      CSN9     CSN10
## 0.2603605 0.4068174 0.4062148 0.3349439 0.3252759 0.2499256 0.4087255 0.2941987
```

```
##      OPN1      OPN2      OPN3      OPN4      OPN5      OPN6      OPN7      OPN8
## 0.3434349 0.3056951 0.3354069 0.2434304 0.4246753 0.2120962 0.3072093 0.3348707
##      OPN9     OPN10
## 0.2427815 0.4850157
```

Above table represents the variance for each feature that has been explained by the factors we created. We can see that our model doesn't explain features such as EST4, AGR1, OPN4, OPN9, etc. well.

```
sum(apply(factor_analysis$loadings^2, 1, sum))/50
```

```
## [1] 0.3876418
```

The total percentage of variance explained by our factor analysis model is about 38.76%, which is not really good and we probably should use more factors in our model.

(b) Now perform the factor analysis with 5 factors and with the varimax rotation (remember to not scale the data). Comment on the differences and determine whether you would consider adding more factors.

```
##
## Loadings:
##       MR1    MR2    MR5    MR4    MR3
## EXT1   0.690
## EXT2  -0.680        -0.128
## EXT3   0.610 -0.202  0.247         0.132
## EXT4  -0.708  0.146
## EXT5   0.702         0.203  0.111  0.111
## EXT6  -0.528  0.103 -0.144 -0.195
## EXT7   0.701         0.149
## EXT8  -0.560                       0.109
## EXT9   0.628                0.163
## EXT10 -0.649  0.208
## EST1  -0.109  0.690  0.123
## EST2        -0.436         0.101
## EST3  -0.129  0.614  0.195
## EST4   0.132 -0.295                0.139
## EST5          0.527
## EST6          0.723
## EST7          0.724               -0.114
## EST8          0.747               -0.133
## EST9          0.698 -0.156
## EST10 -0.235  0.599         0.102 -0.183
## AGR1                -0.478
## AGR2   0.329         0.501  0.138
## AGR3   0.112  0.240 -0.399        -0.144
## AGR4          0.126  0.765
## AGR5  -0.132        -0.642
## AGR6          0.215  0.576
## AGR7  -0.298  0.105 -0.612
## AGR8   0.134         0.543         0.137
## AGR9          0.178  0.674
## AGR10  0.290         0.383  0.172  0.171
## CSN1                        0.103  0.639
## CSN2          0.171         0.169 -0.485
## CSN3                        0.261  0.424
## CSN4          0.406               -0.485
## CSN5                             0.625
## CSN6          0.238         0.104 -0.517
```

```
## CSN7                                0.556
## CSN8          0.253 -0.125        -0.410
## CSN9                                0.630
## CSN10                        0.273  0.466
## OPN1                         0.582
## OPN2          0.239         -0.495
## OPN3          0.142  0.101  0.547
## OPN4          0.156 -0.107 -0.439  0.121
## OPN5   0.201                 0.596  0.164
## OPN6                        -0.431
## OPN7         -0.101          0.495  0.220
## OPN8                         0.567
## OPN9  -0.119  0.174  0.174  0.403
## OPN10  0.177                 0.671
##
##                   MR1   MR2   MR5   MR4   MR3
## SS loadings     4.775 4.644 3.607 3.208 3.148
## Proportion Var  0.095 0.093 0.072 0.064 0.063
## Cumulative Var  0.095 0.188 0.261 0.325 0.388
```

Above are the loading scores when we perform factor analysis with varimax rotation for the data set five_personality.csv with first 50 features.

Here SS loadings represent the variance, or the eigenvalue of each factor. The first factor has an eigenvalue of 4.775. The proportion of variance explained by this factor is $4.775/50 = 0.095$. Here, 50 refers to the total features presented in the data set.

Since we apply varimax rotation to the factor analysis, it is much easier to interpret what each factor is capturing.

```
##      EXT1      EXT2      EXT3      EXT4      EXT5      EXT6      EXT7      EXT8
## 0.5162660 0.5186539 0.5077056 0.4723880 0.4407954 0.6516533 0.4792585 0.6678173
##      EXT9     EXT10      EST1      EST2      EST3      EST4      EST5      EST6
## 0.5774516 0.5301869 0.4916706 0.7894292 0.5665140 0.8741385 0.7135418 0.4669658
##      EST7      EST8      EST9     EST10      AGR1      AGR2      AGR3      AGR4
## 0.4620781 0.4231093 0.4862060 0.5425051 0.7658329 0.6208330 0.7419775 0.3884229
##      AGR5      AGR6      AGR7      AGR8      AGR9     AGR10      CSN1      CSN2
## 0.5648680 0.6189763 0.5231555 0.6607344 0.4915399 0.7080107 0.5726198 0.7036798
##      CSN3      CSN4      CSN5      CSN6      CSN7      CSN8      CSN9     CSN10
## 0.7396395 0.5931826 0.5937852 0.6650561 0.6747241 0.7500744 0.5912745 0.7058013
##      OPN1      OPN2      OPN3      OPN4      OPN5      OPN6      OPN7      OPN8
## 0.6565651 0.6943049 0.6645931 0.7565696 0.5753247 0.7879038 0.6927907 0.6651293
##      OPN9     OPN10
## 0.7572185 0.5149843
```

Above table represents the remaining variance for each feature that hasn't been explained by the factors we created using five factors and varimax rotation.

```
##      EXT1      EXT2      EXT3      EXT4      EXT5      EXT6      EXT7      EXT8
## 0.4837340 0.4813461 0.4922944 0.5276120 0.5592046 0.3483467 0.5207415 0.3321827
##      EXT9     EXT10      EST1      EST2      EST3      EST4      EST5      EST6
## 0.4225484 0.4698131 0.5083294 0.2105708 0.4334860 0.1258615 0.2864582 0.5330342
##      EST7      EST8      EST9     EST10      AGR1      AGR2      AGR3      AGR4
## 0.5379219 0.5768907 0.5137940 0.4574949 0.2341671 0.3791670 0.2580225 0.6115771
##      AGR5      AGR6      AGR7      AGR8      AGR9     AGR10      CSN1      CSN2
## 0.4351320 0.3810237 0.4768445 0.3392656 0.5084601 0.2919893 0.4273802 0.2963202
##      CSN3      CSN4      CSN5      CSN6      CSN7      CSN8      CSN9     CSN10
```

```
## 0.2603605 0.4068174 0.4062148 0.3349439 0.3252759 0.2499256 0.4087255 0.2941987
##      OPN1      OPN2      OPN3      OPN4      OPN5      OPN6      OPN7      OPN8
## 0.3434349 0.3056951 0.3354069 0.2434304 0.4246753 0.2120962 0.3072093 0.3348707
##      OPN9     OPN10
## 0.2427815 0.4850157
```

Above table represents the variance for each feature that has been explained by the factors we created in the factor analysis model with varimax rotation. We can see that our model doesn't explain features such as EST4, AGR1, OPN4, OPN9, etc. well.

```
sum(apply(factor_analysis_vari$loadings^2, 1, sum))/50
```

```
## [1] 0.3876418
```

Model with varimax rotation doesn't change the communalities or total variance explained; it just going to maximize each λ_{ij} in the model we created for each variable in our original data set. A varimax rotation will try to find high loadings to lead to more interpretable factors.

(c) Given your preferred model with five factors, look at the factor loadings matrix and interpret the factors. What would you rename these factors?

```
##
## Loadings:
##       MR1    MR2    MR5    MR4    MR3
## EXT1   0.690
## EXT2  -0.680        -0.128
## EXT3   0.610 -0.202  0.247         0.132
## EXT4  -0.708  0.146
## EXT5   0.702         0.203  0.111  0.111
## EXT6  -0.528  0.103 -0.144 -0.195
## EXT7   0.701         0.149
## EXT8  -0.560                       0.109
## EXT9   0.628                0.163
## EXT10 -0.649  0.208
## EST1  -0.109  0.690  0.123
## EST2        -0.436         0.101
## EST3  -0.129  0.614  0.195
## EST4   0.132 -0.295                0.139
## EST5          0.527
## EST6          0.723
## EST7          0.724               -0.114
## EST8          0.747               -0.133
## EST9          0.698 -0.156
## EST10 -0.235  0.599         0.102 -0.183
## AGR1                -0.478
## AGR2   0.329         0.501  0.138
## AGR3   0.112  0.240 -0.399        -0.144
## AGR4          0.126  0.765
## AGR5  -0.132        -0.642
## AGR6          0.215  0.576
## AGR7  -0.298  0.105 -0.612
## AGR8   0.134         0.543         0.137
## AGR9          0.178  0.674
## AGR10  0.290         0.383  0.172  0.171
## CSN1                        0.103  0.639
## CSN2          0.171         0.169 -0.485
## CSN3                        0.261  0.424
```

```
## CSN4          0.406               -0.485
## CSN5                               0.625
## CSN6          0.238         0.104 -0.517
## CSN7                               0.556
## CSN8          0.253 -0.125        -0.410
## CSN9                               0.630
## CSN10                        0.273  0.466
## OPN1                         0.582
## OPN2          0.239         -0.495
## OPN3          0.142  0.101  0.547
## OPN4          0.156 -0.107 -0.439  0.121
## OPN5   0.201                 0.596  0.164
## OPN6                         -0.431
## OPN7         -0.101          0.495  0.220
## OPN8                         0.567
## OPN9  -0.119  0.174  0.174  0.403
## OPN10  0.177                 0.671
##
##                   MR1   MR2   MR5   MR4   MR3
## SS loadings     4.775 4.644 3.607 3.208 3.148
## Proportion Var 0.095 0.093 0.072 0.064 0.063
## Cumulative Var 0.095 0.188 0.261 0.325 0.388
```

We will use the model with varimax rotation because it is much easier to interpret the result.

The first factor captures 9.5% of variance of the original data set. The second factor captures 9.3% of variance of the original data set. The third factor captures 7.2% of variance of the original data set. The fourth factor captures 6.4% of variance of the original data set. The fifth factor captures 6.3% of variance of the original data set. The cumulative variance captured is around 38.8%.

For the first factor, we can see that it has large positive loadings and large negative loadings for features EXT1 to EXT10. This means that features EXT1 to EXT10 highly influence factor 1. I will rename the first factor as whether or not you are communicative because all the questions are related to whether or not you like to communicate with others.

For the second factor, we can see that it has large positive loadings for features ARG1 to EST10. This means that features EST1 to EST10 highly influence factor 2. I will rename the second factor as whether or not you are anxious because all the questions are related to whether or not you have anxiety.

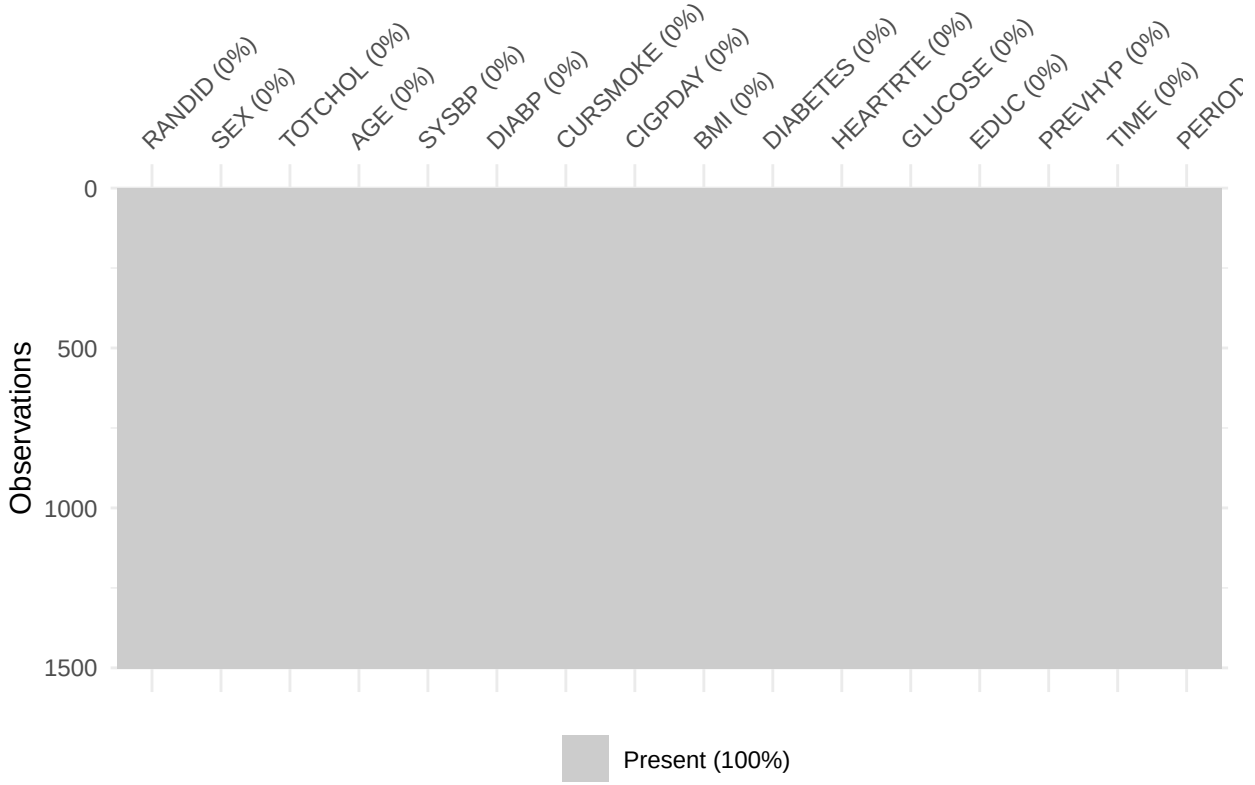
For the third factor, we can see that it has large positive loadings and large negative loadings for features ARG1 to ARG10. This means that features ARG1 to ARG10 highly influence factor 3. I will rename the third factor as whether or not you are sympathetic because all the questions are related to whether or not you have sympathy.

For the fourth factor, we can see that it has large positive loadings and large negative loadings for features OPN1 TO OPN10. This means that features OPN1 to OPN10 highly influence factor 4. I will rename the fourth factor as whether or not you are curious or imaginative because all the questions are related to whether or not you have curiosity and imagination.

For the fifth factor, we can see that it has large positive loadings and large negative loadings for features CSN1 to CSN10. This means that features CSN1 to CSN10 highly influence factor 5. I will rename the fifth factor as whether or not you are careful and organized because all the questions are related to whether or not you follow a schedule or like order etc.

3. Longitudinal Data Application

This problem will use the data set framingham multi.csv, which can be found in the Data folder on Canvas. This data is a subset of data from the Framingham Heart Study (https://www. framinghamheartstudy.org/) and contains health information for patients over time. For our teaching purposes, some methods were employed to ensure an anonymous dataset that protects patient confidentiality. The variables are given below.



From above graph, we can see that there are no missing data.

```
##      RANDID     SEX        TOTCHOL           AGE            SYSBP
##  6238   :   3   1:702   Min.   :135.0   Min.   :33.00   Min.   : 92.5
##  11263  :   3   2:798   1st Qu.:209.0   1st Qu.:47.00   1st Qu.:120.0
##  14367  :   3           Median :238.0   Median :54.00   Median :131.5
##  16365  :   3           Mean   :239.8   Mean   :54.11   Mean   :134.7
##  23727  :   3           3rd Qu.:266.0   3rd Qu.:61.00   3rd Qu.:147.1
##  34689  :   3           Max.   :625.0   Max.   :79.00   Max.   :246.0
##  (Other):1482
##      DIABP        CURSMOKE    CIGPDAY           BMI        DIABETES
##  Min.   : 52.00   0:876    Min.   : 0.000   Min.   :15.16   0:1445
##  1st Qu.: 75.00   1:624    1st Qu.: 0.000   1st Qu.:23.09   1:  55
##  Median : 82.00            Median : 0.000   Median :25.16
##  Mean   : 82.73            Mean   : 8.187   Mean   :25.61
##  3rd Qu.: 90.00            3rd Qu.:20.000   3rd Qu.:27.78
##  Max.   :127.00            Max.   :90.000   Max.   :52.94
##
##     HEARTRTE         GLUCOSE       EDUC    PREVHYP      TIME      PERIOD
##  Min.   : 50.00   Min.   : 45.00   1:582   0:819   Min.   :   0   1:500
##  1st Qu.: 68.00   1st Qu.: 73.00   2:486   1:681   1st Qu.:   0   2:500
##  Median : 75.00   Median : 80.00   3:249           Median :2177   3:500
##  Mean   : 76.76   Mean   : 84.47   4:183           Mean   :2177
```

```
##  3rd Qu.: 85.00   3rd Qu.: 91.00                    3rd Qu.:4312
##  Max.   :220.00   Max.   :420.00                    Max.   :4607
##
```

We convert all the categorical variables to factor and print out the summary table.

```
##      RANDID     SEX         TOTCHOL              AGE               SYSBP
##  6238   :   3   1:702   Min.   :-2.33695   Min.   :-2.18332   Min.   :-1.9890
##  11263  :   3   2:798   1st Qu.:-0.68613   1st Qu.:-0.73559   1st Qu.:-0.6931
##  14367  :   3           Median :-0.03919   Median :-0.01172   Median :-0.1512
##  16365  :   3           Mean   : 0.00000   Mean   : 0.00000   Mean   : 0.0000
##  23727  :   3           3rd Qu.: 0.58544   3rd Qu.: 0.71215   3rd Qu.: 0.5852
##  34689  :   3           Max.   : 8.59414   Max.   : 2.57351   Max.   : 5.2446
##  (Other):1482
##      DIABP           CURSMOKE    CIGPDAY             BMI           DIABETES
##  Min.   :-2.72423   0:876    Min.   :-0.6509   Min.   :-2.6847   0:1445
##  1st Qu.:-0.68511   1:624    1st Qu.:-0.6509   1st Qu.:-0.6478   1:  55
##  Median :-0.06451            Median :-0.6509   Median :-0.1174
##  Mean   : 0.00000            Mean   : 0.0000   Mean   : 0.0000
##  3rd Qu.: 0.64475            3rd Qu.: 0.9392   3rd Qu.: 0.5569
##  Max.   : 3.92506            Max.   : 6.5047   Max.   : 7.0196
##
##     HEARTRTE          GLUCOSE        EDUC    PREVHYP      TIME
##  Min.   :-2.0343   Min.   :-1.7312   1:582   0:819   Min.   :-1.2232039
##  1st Qu.:-0.6662   1st Qu.:-0.5032   2:486   1:681   1st Qu.:-1.2232039
##  Median :-0.1341   Median :-0.1962   3:249           Median :-0.0000214
##  Mean   : 0.0000   Mean   : 0.0000   4:183           Mean   : 0.0000000
##  3rd Qu.: 0.6259   3rd Qu.: 0.2863                   3rd Qu.: 1.1995628
##  Max.   :10.8870   Max.   :14.7156                   Max.   : 1.3653134
##
##  PERIOD
##  1:500
##  2:500
##  3:500
##
##
##
##
```

We standardize all the continuous variables to help convergence for optimization and print out the summary table.

```
length(unique(ramingham$RANDID))
```

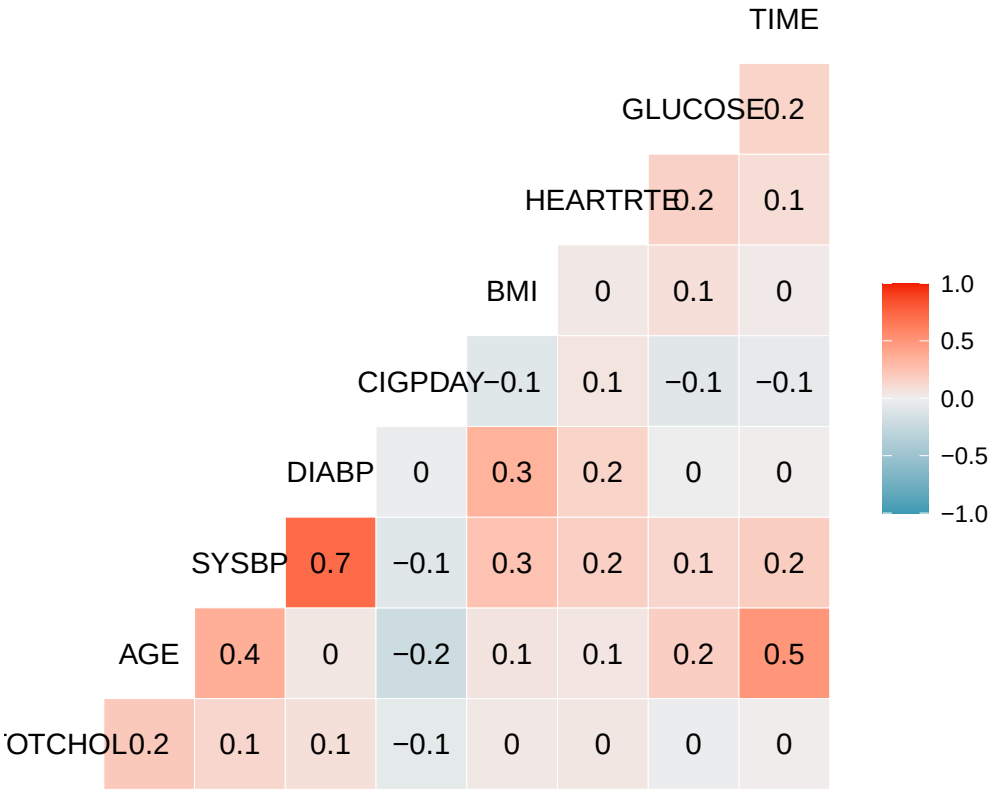
```
## [1] 500
```

There are total 500 different participants (500 subjects).

```
## [1] 1200   16
```

```
## [1] 300  16
```

We split the whole data set into train and test set. The dimension for the training set is (1200, 16). The dimension for the testing set is (300, 16).



From above correlation table, we can see that DIABP and SYSBP are highly correlated. TIME and AGE also have a noticeable correlation.

```
## Generalized linear mixed model fit by maximum likelihood (Laplace
##   Approximation) [glmerMod]
##  Family: binomial  ( logit )
## Formula: PREVHYP ~ SEX + AGE + EDUC + TOTCHOL + SYSBP + DIABP + CURSMOKE +
##     CIGPDAY + BMI + DIABETES + HEARTRTE + GLUCOSE + PERIOD +
##     TIME + (1 | RANDID)
##    Data: ramingham.train
## Control: glmerControl(optimizer = "optimx", optCtrl = list(method = "nlminb"),
##     nAGQ = 9)
##
##      AIC      BIC   logLik deviance df.resid
##    769.8    866.5   -365.9    731.8     1181
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -5.0330 -0.0009  0.0000  0.0014 14.5313
##
## Random effects:
##  Groups Name        Variance Std.Dev.
##  RANDID (Intercept) 256      16
## Number of obs: 1200, groups:  RANDID, 494
##
## Fixed effects:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -2.84237   19.22793  -0.148   0.8825
## SEX2        -3.29774    1.48757  -2.217   0.0266 *
```

```
## AGE          3.23401    0.82652   3.913 9.12e-05 ***
## EDUC2        1.20342    1.64500   0.732   0.4644
## EDUC3        1.32511    2.50862   0.528   0.5973
## EDUC4       -0.93519    2.16192  -0.433   0.6653
## TOTCHOL      0.55068    0.63992   0.861   0.3895
## SYSBP        7.28108    1.00958   7.212 5.51e-13 ***
## DIABP        5.35919    0.85834   6.244 4.27e-10 ***
## CURSMOKE1   -1.62301    1.40592  -1.154   0.2483
## CIGPDAY     -0.08188    0.76031  -0.108   0.9142
## BMI          3.39305    0.67488   5.028 4.97e-07 ***
## DIABETES1    5.58217    4.17549   1.337   0.1813
## HEARTRTE     2.13271    0.41476   5.142 2.72e-07 ***
## GLUCOSE      0.18383    0.45492   0.404   0.6861
## PERIOD2      2.23091   19.12255   0.117   0.9071
## PERIOD3      3.21638   38.06699   0.084   0.9327
## TIME         4.99379   15.43347   0.324   0.7463
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We build the first model based on all the available covariates in our data set. We also use random intercept for the covariate RANDID. This means that for each participant, we allow for a different intercept for the model to predict prevalent hypertensive.

The first model tells us that the only significant covariates at 5% significance level are: SEX, AGE, SYSBP, DIABP, BMI, and HEARTRTE.

```
## Generalized linear mixed model fit by maximum likelihood (Laplace
##   Approximation) [glmerMod]
##  Family: binomial  ( logit )
## Formula: PREVHYP ~ SEX + AGE + SYSBP + DIABP + BMI + HEARTRTE + (1 | RANDID)
##    Data: ramingham.train
## Control: glmerControl(optimizer = "optimx", optCtrl = list(method = "nlminb"),
##     nAGQ = 9)
##
##      AIC      BIC   logLik deviance df.resid
##    867.8    908.5   -425.9    851.8     1192
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -4.3378 -0.1617 -0.0108  0.1687  8.7214
##
## Random effects:
##  Groups Name        Variance Std.Dev.
##  RANDID (Intercept) 7.61     2.759
## Number of obs: 1200, groups:  RANDID, 494
##
## Fixed effects:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -0.0008954  0.2661462  -0.003 0.997316
## SEX2        -0.9129401  0.5909598  -1.545 0.122384
## AGE          1.8217897  1.1274408   1.616 0.106124
## SYSBP        2.5749593  0.7242108   3.556 0.000377 ***
## DIABP        1.4160392  0.5010354   2.826 0.004710 **
## BMI          0.7738579  0.4019819   1.925 0.054216 .
## HEARTRTE     0.5615099  0.2973592   1.888 0.058983 .
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##          (Intr) SEX2   AGE    SYSBP  DIABP  BMI
## SEX2     -0.346
## AGE      -0.135 -0.768
## SYSBP    -0.079 -0.732  0.881
## DIABP    -0.164 -0.639  0.884  0.685
## BMI      -0.146 -0.657  0.864  0.807  0.717
## HEARTRTE -0.082 -0.694  0.841  0.783  0.729  0.763
```

After we delete all the insignificant covariates and rerun the model, the AIC and BIC increase. This indicates that we have deleted too much predictors.

```
## Likelihood ratio test
##
## Model 1: PREVHYP ~ SEX + AGE + EDUC + TOTCHOL + SYSBP + DIABP + CURSMOKE +
##     CIGPDAY + BMI + DIABETES + HEARTRTE + GLUCOSE + PERIOD +
##     TIME + (1 | RANDID)
## Model 2: PREVHYP ~ SEX + AGE + SYSBP + DIABP + BMI + HEARTRTE + (1 | RANDID)
##   #Df  LogLik  Df  Chisq Pr(>Chisq)
## 1  19 -365.90
## 2   8 -425.88 -11 119.97  < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The likelihood-ratio test also tells us that we should stick with the full model because of the super low p-value.

```
## Generalized linear mixed model fit by maximum likelihood (Laplace
##   Approximation) [glmerMod]
##  Family: binomial  ( logit )
## Formula: PREVHYP ~ SEX + AGE + TOTCHOL + SYSBP + DIABP + CURSMOKE + BMI +
##     DIABETES + HEARTRTE + GLUCOSE + TIME + (1 | RANDID)
##    Data: ramingham.train
## Control: glmerControl(optimizer = "optimx", optCtrl = list(method = "nlminb"),
##     nAGQ = 9)
##
##      AIC      BIC   logLik deviance df.resid
##    759.9    826.1   -366.9    733.9     1187
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -4.6132 -0.0011  0.0000  0.0018 12.5266
##
## Random effects:
##  Groups Name        Variance Std.Dev.
##  RANDID (Intercept) 251.6    15.86
## Number of obs: 1200, groups:  RANDID, 494
##
## Fixed effects:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -0.62344    0.50643  -1.231 0.218307
## SEX2        -3.18211    1.15615  -2.752 0.005917 **
## AGE          3.12294    0.76730   4.070 4.70e-05 ***
## TOTCHOL      0.56836    0.39250   1.448 0.147605
```

```
## SYSBP        7.00993    0.37068  18.911  < 2e-16 ***
## DIABP        5.60772    0.40123  13.976  < 2e-16 ***
## CURSMOKE1   -1.68232    0.50965  -3.301 0.000964 ***
## BMI          3.18885    0.43150   7.390 1.47e-13 ***
## DIABETES1    6.15762    3.62482   1.699 0.089369 .
## HEARTRTE     2.14668    0.32617   6.581 4.66e-11 ***
## GLUCOSE      0.08912    0.37313   0.239 0.811231
## TIME         6.30853    0.36994  17.053  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##           (Intr) SEX2   AGE    TOTCHO SYSBP  DIABP  CURSMO BMI    DIABET HEARTR
## SEX2      -0.409
## AGE        0.170 -0.271
## TOTCHOL    0.153 -0.131  0.131
## SYSBP     -0.054  0.096 -0.333  0.161
## DIABP      0.049 -0.080  0.166 -0.039 -0.714
## CURSMOKE1 -0.222 -0.120  0.185  0.084 -0.091  0.067
## BMI       -0.080  0.062  0.046 -0.067  0.070 -0.204 -0.021
## DIABETES1  0.027 -0.163 -0.005 -0.058  0.032  0.002  0.003  0.015
## HEARTRTE  -0.033 -0.176  0.260  0.071  0.232 -0.051 -0.053  0.191  0.004
## GLUCOSE   -0.106  0.189 -0.253  0.059  0.036  0.051 -0.101 -0.023 -0.429 -0.175
## TIME      -0.170  0.123 -0.640 -0.059 -0.099 -0.124 -0.180  0.159 -0.075 -0.030
##           GLUCOS
## SEX2
## AGE
## TOTCHOL
## SYSBP
## DIABP
## CURSMOKE1
## BMI
## DIABETES1
## HEARTRTE
## GLUCOSE
## TIME       0.178
```

For model 3, we delete EDUC, CIGPDAY, and PERIOD from our model. We delete EDUC because education seems not to relate to whether you have hypertensive. We delete CIGPDAY and PERIOD because CIGPDAY and CURSMOKE are really similar; TIME and PERIOD are really similar as well.

In the third model, all the covariates are significant besides DIABETES.

```
## Likelihood ratio test
##
## Model 1: PREVHYP ~ SEX + AGE + EDUC + TOTCHOL + SYSBP + DIABP + CURSMOKE +
##     CIGPDAY + BMI + DIABETES + HEARTRTE + GLUCOSE + PERIOD +
##     TIME + (1 | RANDID)
## Model 2: PREVHYP ~ SEX + AGE + TOTCHOL + SYSBP + DIABP + CURSMOKE + BMI +
##     DIABETES + HEARTRTE + GLUCOSE + TIME + (1 | RANDID)
##   #Df  LogLik Df  Chisq Pr(>Chisq)
## 1  19 -365.90
## 2  13 -366.94 -6 2.0841     0.9118
```

The likelihood-ratio test also tells us that we should stick with the reduced model because of the high p-value.

```
## Generalized linear mixed model fit by maximum likelihood (Laplace
##   Approximation) [glmerMod]
##  Family: binomial  ( logit )
## Formula: PREVHYP ~ SEX + AGE + TOTCHOL + SYSBP + DIABP + CURSMOKE + BMI +
##     HEARTRTE + GLUCOSE + TIME + (1 | RANDID)
##    Data: ramingham.train
## Control: glmerControl(optimizer = "optimx", optCtrl = list(method = "nlminb"),
##     nAGQ = 9)
##
##      AIC      BIC   logLik deviance df.resid
##    760.0    821.1   -368.0    736.0     1188
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -4.5743 -0.0011  0.0000  0.0017 11.4212
##
## Random effects:
##  Groups Name        Variance Std.Dev.
##  RANDID (Intercept) 250.7    15.83
## Number of obs: 1200, groups:  RANDID, 494
##
## Fixed effects:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -0.4837     1.5693  -0.308  0.75788
## SEX2         -3.1977     2.1495  -1.488  0.13684
## AGE           3.1781     1.2848   2.474  0.01337 *
## TOTCHOL       0.6222     0.7197   0.864  0.38733
## SYSBP         7.0352     1.5256   4.611 4.00e-06 ***
## DIABP         5.5336     1.1785   4.695 2.66e-06 ***
## CURSMOKE1    -1.5611     1.3117  -1.190  0.23401
## BMI           3.1598     1.0729   2.945  0.00323 **
## HEARTRTE      2.1073     0.7692   2.740  0.00615 **
## GLUCOSE       0.4506     0.5606   0.804  0.42151
## TIME          6.2411     1.1130   5.608 2.05e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##           (Intr) SEX2   AGE    TOTCHO SYSBP  DIABP  CURSMO BMI    HEARTR GLUCOS
## SEX2      -0.678
## AGE        0.036 -0.144
## TOTCHOL    0.106 -0.221 -0.080
## SYSBP      0.049 -0.154 -0.119  0.182
## DIABP     -0.056 -0.041  0.254  0.060 -0.078
## CURSMOKE1 -0.374  0.093  0.091 -0.070 -0.108 -0.082
## BMI       -0.072  0.083  0.042 -0.191  0.065  0.025  0.011
## HEARTRTE   0.092 -0.187  0.101  0.293  0.134  0.191 -0.119  0.067
## GLUCOSE   -0.020 -0.021 -0.024  0.115  0.063  0.199 -0.017  0.053 -0.110
## TIME      -0.027 -0.022 -0.447  0.188  0.453  0.467 -0.243  0.133  0.180  0.154
## optimizer (optimx) convergence code: 1 (none)
## unable to evaluate scaled gradient
## Model failed to converge: degenerate  Hessian with 2 negative eigenvalues
```

The fourth model deletes covariate DIABETES. The BIC for the fourth model has slightly decreased. This

time, the significant covariates are: AGE, SYSBP, DIABP, BMI, HEARTRTE, and TIME.

```
## Likelihood ratio test
##
## Model 1: PREVHYP ~ SEX + AGE + TOTCHOL + SYSBP + DIABP + CURSMOKE + BMI +
##     DIABETES + HEARTRTE + GLUCOSE + TIME + (1 | RANDID)
## Model 2: PREVHYP ~ SEX + AGE + TOTCHOL + SYSBP + DIABP + CURSMOKE + BMI +
##     HEARTRTE + GLUCOSE + TIME + (1 | RANDID)
##   #Df  LogLik Df Chisq Pr(>Chisq)
## 1  13 -366.94
## 2  12 -368.00 -1 2.108     0.1465
```

The likelihood-ratio test tells us that we should stick with the reduced model, which is model 4.

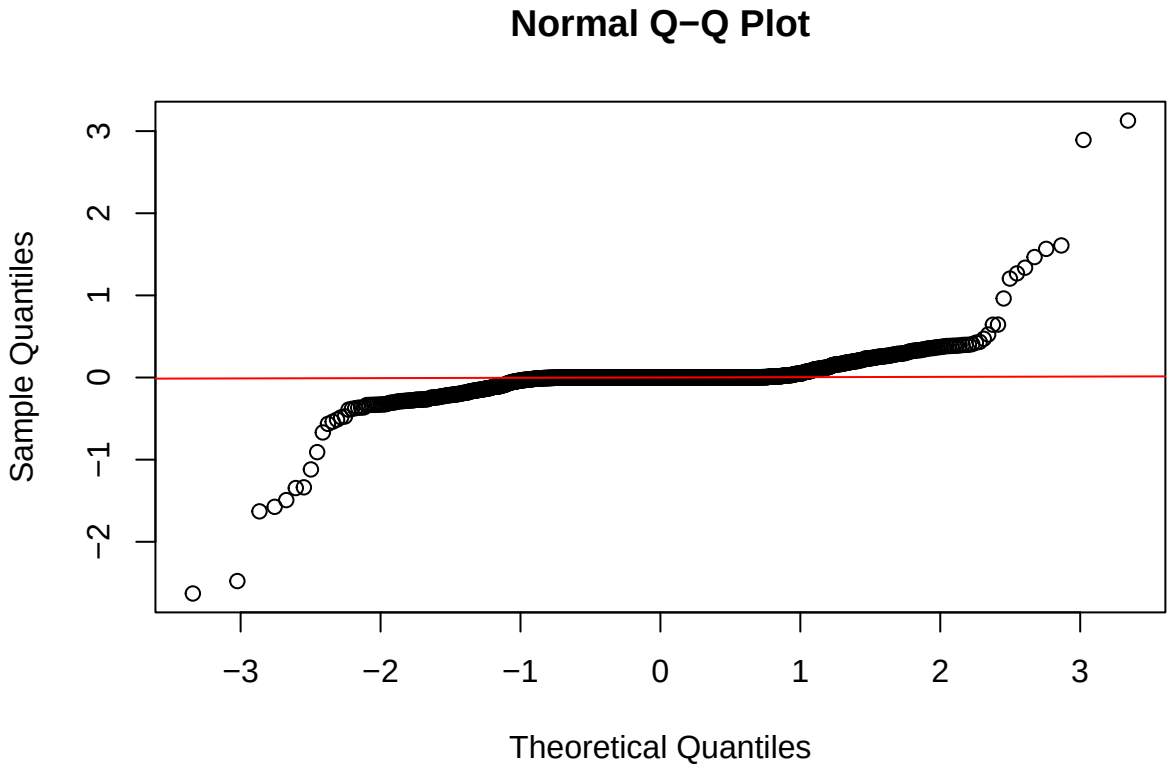
```
## Generalized linear mixed model fit by maximum likelihood (Laplace
##   Approximation) [glmerMod]
##  Family: binomial  ( logit )
## Formula: PREVHYP ~ AGE + SYSBP + DIABP + BMI + HEARTRTE + TIME + (1 |
##     RANDID)
##    Data: ramingham.train
## Control: glmerControl(optimizer = "optimx", optCtrl = list(method = "nlminb"),
##     nAGQ = 9)
##
##      AIC      BIC   logLik deviance df.resid
##    761.1    801.8   -372.5    745.1     1192
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -5.5426 -0.0015  0.0000  0.0021 11.5040
##
## Random effects:
##  Groups Name        Variance Std.Dev.
##  RANDID (Intercept) 247.7    15.74
## Number of obs: 1200, groups:  RANDID, 494
##
## Fixed effects:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -2.7288     0.1150  -23.73   <2e-16 ***
## AGE           3.3219     0.1127   29.47   <2e-16 ***
## SYSBP         6.5782     0.1146   57.38   <2e-16 ***
## DIABP         5.6821     0.1181   48.13   <2e-16 ***
## BMI           3.4921     0.1159   30.12   <2e-16 ***
## HEARTRTE      1.9944     0.1181   16.88   <2e-16 ***
## TIME          6.1155     0.1150   53.17   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##          (Intr) AGE    SYSBP  DIABP  BMI    HEARTR
## AGE       0.000
## SYSBP     0.014  0.011
## DIABP     0.005  0.014 -0.018
## BMI       0.004  0.009 -0.002 -0.005
## HEARTRTE  0.006  0.015  0.015 -0.242  0.017
## TIME     -0.004 -0.043  0.032  0.104  0.018 -0.242
```

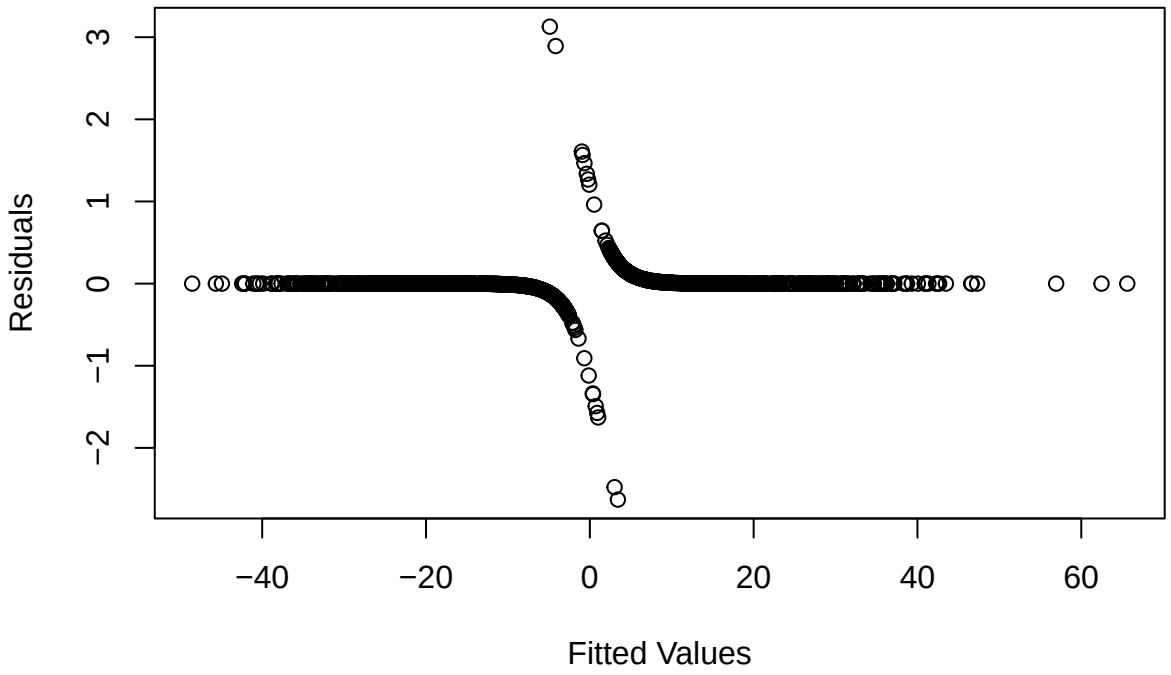
```
## optimizer (optimx) convergence code: 1 (none)
```

From the summary table, we can see that all the predictors(fixed effects) for model 5 are significant. The BIC also improves a lot compared to model 4.

```
## Likelihood ratio test
##
## Model 1: PREVHYP ~ SEX + AGE + TOTCHOL + SYSBP + DIABP + CURSMOKE + BMI +
##     HEARTRTE + GLUCOSE + TIME + (1 | RANDID)
## Model 2: PREVHYP ~ AGE + SYSBP + DIABP + BMI + HEARTRTE + TIME + (1 |
##     RANDID)
##   #Df  LogLik Df  Chisq Pr(>Chisq)
## 1  12 -368.00
## 2   8 -372.54 -4 9.0878    0.05894 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The likelihood-ratio test tells us that at 5% significance level, we fail to reject the null and should stick with the reduced model, which is model 5. Moreover, model 5 has far more less predictors, which is easier to interpret.





Above are the normal-QQ plot and residuals vs. fitted values plot for the model 5. Since model 5 is a logistic regression model, it is not really useful to look at the diagnostic plots to evaluate normality assumptions. Moreover, logistic regression doesn't have normality assumption on residuals.

```
##    grpvar        term   grp       condval    condsd
## 1  RANDID (Intercept)  6238 -1.143168e-01 14.908161
## 2  RANDID (Intercept) 11263  3.797126e-06 15.737845
## 3  RANDID (Intercept) 14367  8.457452e+00  5.197427
## 4  RANDID (Intercept) 16365  1.413561e-01 14.731632
## 5  RANDID (Intercept) 23727  1.484780e+01  4.059479
## 6  RANDID (Intercept) 34689  1.539704e-01 14.650983
## 7  RANDID (Intercept) 36459 -5.570308e-02 15.315062
## 8  RANDID (Intercept) 40435 -1.171695e+00 10.684411
## 9  RANDID (Intercept) 43770 -2.178810e+00  8.850270
## 10 RANDID (Intercept) 45464  1.739648e+00  9.522545
```

Above table shows 10 intercept values for different subjects(participants). For example, if you are a participant with identification number 6238, then your intercept for your model predicting response PREVHYP is -1.143168e-01. This intercept value is different for all the participants, but the coefficients for all the other covariates in the model are the same for all participants.

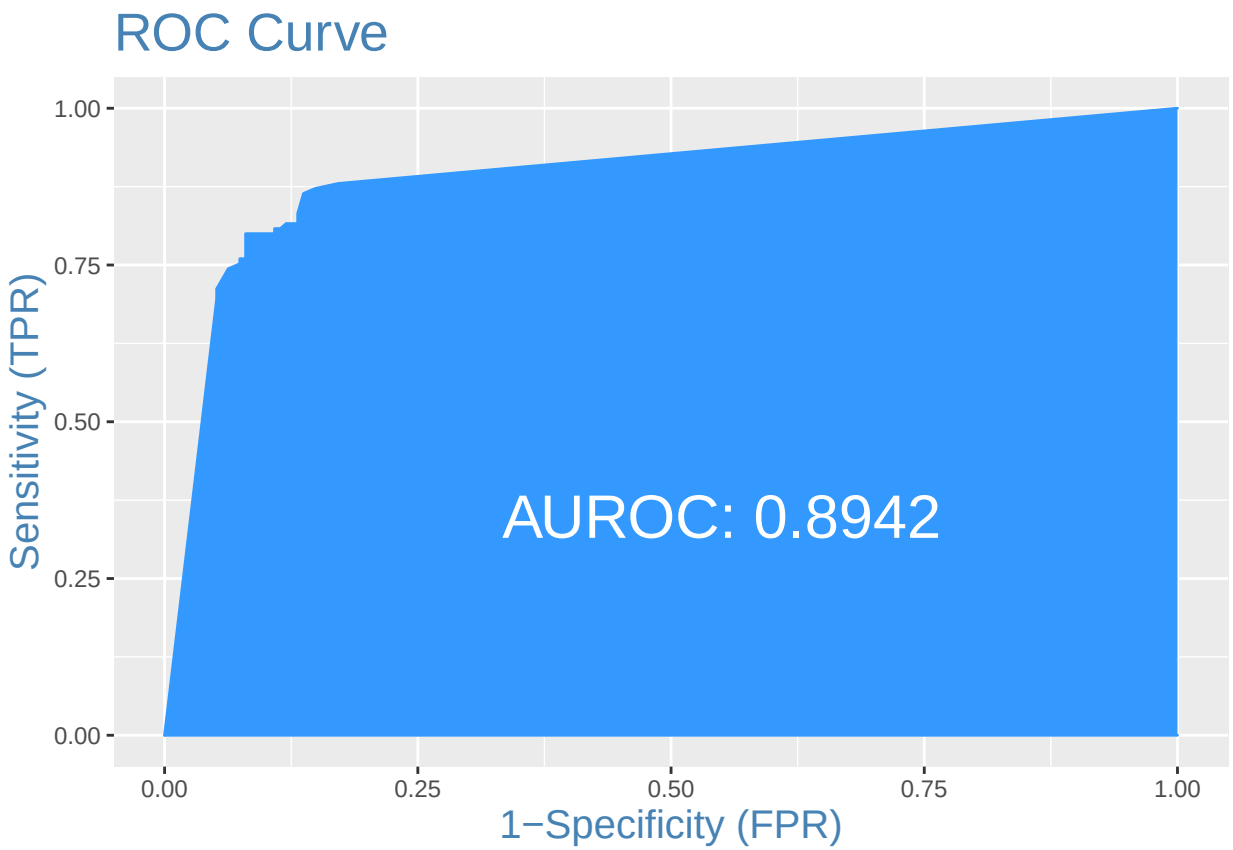
```
##
## pred   0   1
##    0 156  25
##    1  19 100
```

Above table shows the prediction result for the model 5 on the test data set.

```
sum(pred == ramingham.test$PREVHYP)/nrow(ramingham.test)
```

```
## [1] 0.8533333
```

The accuracy of model 5 on test set is 85.33%.



The AUROC score for model 5 is 89.42%, which is pretty high and indicating that our model has done a great job.

Code Appendix:

```
knitr::opts_chunk$set(echo = TRUE)
library(dplyr)
library(tidyr)
library(readr)
library(ggplot2)
library(ggfortify)
library(GGally) # For ggcor()
library(Hmisc)
library(naniar) # For vis_miss() function to visualize missing data
library(repr) # For adjusting plot sizes
options(repr.plot.width=10, repr.plot.height=8)
library(glmnet) # For ridge and LASSO
library(lme4) # For multi-level modeling
library(psych)
library(optimx)
library(InformationValue)
library(lmtest)
five <- read_csv("~/Desktop/five_personality.csv")
vis_miss(five, warn_large_data = FALSE)
five_50 = five[,1:50]
five_50 = na.omit(five_50)
```

```
five_50 <- five_50 %>%
    mutate(across(.cols = names(five_50), .fns = as.numeric))
summary(five_50)
factor_analysis <- fa(five_50, nfactors=5, rotate = "none")
factor_analysis$loadings
factor_analysis$uniquenesses
apply(factor_analysis$loadings^2, 1, sum)
sum(apply(factor_analysis$loadings^2, 1, sum))/50
factor_analysis_vari <- fa(five_50, nfactors=5, rotate = "varimax")
factor_analysis_vari$loadings
factor_analysis_vari$uniquenesses
apply(factor_analysis_vari$loadings^2, 1, sum)
sum(apply(factor_analysis_vari$loadings^2, 1, sum))/50
factor_analysis_vari$loadings
ramingham <- read_csv("~/Desktop/framingham_multi.csv")
vis_miss(ramingham)
ramingham <- ramingham %>%
    mutate(across(.cols=c(RANDID, SEX, EDUC, CURSMOKE, DIABETES, PREVHYP, PERIOD), .fns = as.factor))
summary(ramingham)
ramingham <- ramingham %>%
    mutate(across(.cols=c(TOTCHOL, AGE, SYSBP, DIABP, CIGPDAY, BMI, HEARTRTE, GLUCOSE, TIME), .fns = ~
summary(ramingham)
length(unique(ramingham$RANDID))
# Split into test and train sets
set.seed(1)
samp.size = floor(0.8*nrow(ramingham))
train.ind = sample(nrow(ramingham), size = samp.size)
ramingham.train = ramingham[train.ind,]
ramingham.test = ramingham[-train.ind,]
dim(ramingham.train)
dim(ramingham.test)
ggcorr(ramingham, label = TRUE)
lm1 = glmer(PREVHYP ~  SEX + AGE + EDUC + TOTCHOL + SYSBP + DIABP + CURSMOKE + CIGPDAY + BMI + DIABETES
summary(lm1)
lm2 = glmer(PREVHYP ~ SEX + AGE + SYSBP + DIABP + BMI + HEARTRTE + (1|RANDID), data=ramingham.train, fa
summary(lm2)
lrtest(lm1, lm2)
lm3 = glmer(PREVHYP ~  SEX + AGE + TOTCHOL + SYSBP + DIABP + CURSMOKE + BMI + DIABETES + HEARTRTE + GLU
summary(lm3)
lrtest(lm1, lm3)
lm4 = glmer(PREVHYP ~  SEX + AGE + TOTCHOL + SYSBP + DIABP + CURSMOKE + BMI + HEARTRTE + GLUCOSE+ TIME
summary(lm4)
lrtest(lm3, lm4)
lm5 = glmer(PREVHYP ~ AGE + SYSBP + DIABP + BMI + HEARTRTE + TIME + (1|RANDID), data=ramingham.train, fa
summary(lm5)
lrtest(lm4, lm5)
qqnorm(resid(lm5))
qqline(resid(lm5), col = "red")
plot(predict(lm5), residuals(lm5),xlab="Fitted Values",ylab="Residuals")
ranef_randid = as.data.frame(ranef(lm5))
ranef_randid[1:10,]
prob <- predict(lm5, newdata = ramingham.test, type = "response", allow.new.levels = TRUE)
pred <- ifelse(prob>0.5, 1, 0)
```

```
table(pred, ramingham.test$PREVHYP)
sum(pred == ramingham.test$PREVHYP)/nrow(ramingham.test)
plotROC(ramingham.test$PREVHYP, prob, Show.labels=F)
```