Part 2

1. Main finding:

Firstly, we screen the 200 variables using SEVI, RandomForest, XGBoost. The result turns out that SEVI is slightly weaker than the machine learning method when we choose few variables based on the same procedure. The difference of the GMC will be about 0.02 to 0.05.

Secondly, log function and exponential function are not suitable for our G function. Therefore, we try to check the G function on two families:

Notice that the second function guarantee the response is positive. The results show that the second family performs better than first one. But the difference is still very tiny.

In addition, there is a strong tread off between the GMC value and the number of variables we choose. In the result, we just find the model about 10 variables but that doesn’t mean these variables are sufficient.

The assumption to first take a linear combination of x then use g function to represent y might not very useful. Under this condition, when we increase the GMC, we will lose some accuracy of the model fit as the result shown by the mean absolute error.

After check the residual, we find that for response variable, it has a lot of extreme values so that the traditional model cannot fit it well. The residual plot for TP53 looks fine.

2. Details:

Below is our pseudo code:

For i in 4 databases (2 sets 2 responses)

For j in 3 screen methods (SEVI, Randomforest, XGBoost)

For n in 5:200 (choose subset variables)

For t in 2 families

For alpha in its domain

For lambda1, lambda2 in [10-5,1]

Estimate beta by optimize the target

End

End

End

End

End

End

We calculate the GMC based on each database, method, and family, which means the GMC is chosen by some specific alpha, lambda1, lambda2. When we change the dimension on variables, the best G function will change.

The GMC in the set K with response for different screening in first g family

|  |  |  |  |
| --- | --- | --- | --- |
| # of nonzero varaibles | SEVI | RandomForest | XGBoost |
| 5 | 0.145 | 0.174 | 0.164 |
| 10 | 0.154 | 0.185 | 0.174 |
| 20 | 0.165 | 0.229 | 0.214 |
| 50 | 0.243 | 0.270 | 0.303 |
| 100 | 0.318 | 0.371 | 0.380 |
| 150 | 0.451 | 0.451 | 0.451 |

The GMC in the set K with TP53 for different screening in first g family

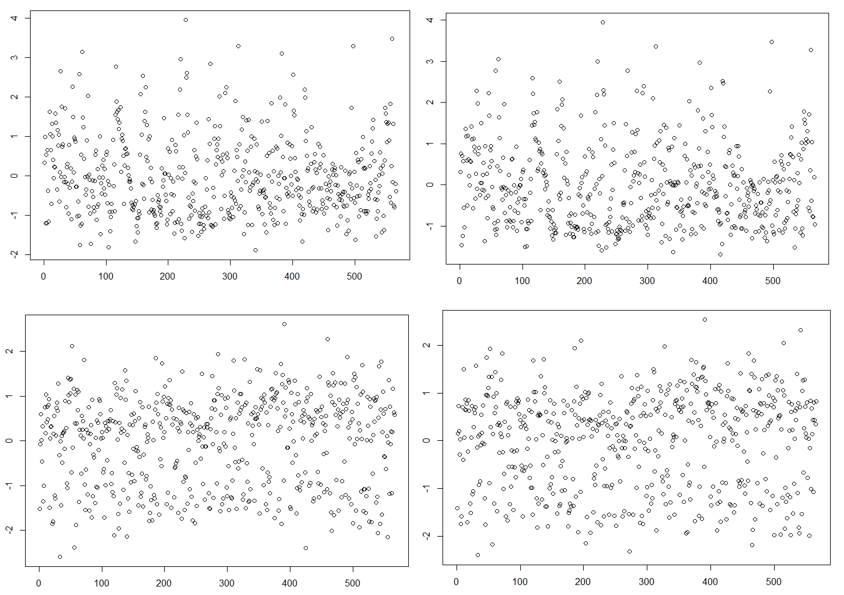
|  |  |  |  |
| --- | --- | --- | --- |
| # of nonzero varaibles | SEVI | RandomForest | XGBoost |
| 5 | 0.096 | 0.122 | 0.071 |
| 10 | 0.114 | 0.149 | 0.125 |
| 20 | 0.139 | 0.201 | 0.191 |
| 50 | 0.197 | 0.268 | 0.264 |
| 100 | 0.287 | 0.304 | 0.327 |
| 150 | 0.442 | 0.442 | 0.442 |

The GMC in different family using randomforest for screening

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Family/# of nonzero | Kset Y | Kset TP53 | Pset Y | Pset TP53 |
| 1 5  2 | 0.174  0.180 | 0.122  0.153 | 0.147  0.146 | 0.129  0.128 |
| 1 10  2 | 0.185  0.192 | 0.149  0.157 | 0.150  0.180 | 0.143  0.146 |
| 1 20  2 | 0.229  0.236 | 0.201  0.205 | 0.202  0.211 | 0.202  0.203 |
| 1 50  2 | 0.270  0.268 | 0.268  0.264 | 0.237  0.248 | 0.284  0.261 |
| 1 100  2 | 0.371  0.368 | 0.304  0.295 | 0.301  0.306 | 0.322  0.337 |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Database | Variables | G function | GMC | MAE |
| K set Y | 8 | X  G1(x,1.1)  G2(x,150) | 0.179  0.184  0.196 | 579  620  677 |
| P set Y | 9 | X  G1(x,0.5)  G2(x,20) | 0.162  0.182  0.184 | 589  717  645 |
| K set TP | 11 | X  G1(x,1.3)  G2(x,50) | 0.135  0.150  0.147 | 0.83  21  9 |
| P set TP | 7 | X  G1(x,2.9)  G2(x,90) | 0.132  0.140  0.141 | 0.84  24  13 |

|  |  |
| --- | --- |
| Database | Variables |
| K set Y | PYY, CYP3A43, CAP1, HOXC10, EPB41L5, FNDC4, S100A8, CAMK2A |
| P set Y | BTN3A1, PYY, BRP44L, NAG18, HOXC10, GATM, MSX1, FNDC4, SLC9A7 |
| K set TP | EFNB3, TCTN1, FLJ14154, TSC2, EFTUD1, RPS23, NIPSNAP1, SENP6, RBM12, COL4A3BP, C14ORF94 |
| P set TP | PLSCR3, CRNKL1, USP21, USP22, VDAC3, FAM50B, QARS |



The plot of the standard residual. Plot1 kset Yplot2 pset Y plot3 kset TP, plot4 pset T