TEAM 5

STAT 602 FINAL PROJECT

Jing Tao, Yilun Chen, Yishan Cai

Part I Linear Regression

- **Methods Introduction :**

Firstly, we use all the 200 variables to fit the full model. We make diagnostic plots of the full model, based on which we remove outliers and do box-cox transformation to Y(response). After the treatment above, we obtain the new data to process in the following steps. First, do WLS fitting if heteroscedasticity is obvious. Secondly, conduct model selection with AIC and BIC criteria by three kinds of procedure: forward, backward and stepwise. Then, use PLS method in order to explain more variance with less variables. At last, use ridge regression.

- **Results & Presentations:**

Section 1------Data set k of response 1

(a) Remove two outliers the 496th and 310th detected by Cook's distance, one by one, from the full model.(Figure 1.1.1) Do box-cox transformation*((response ^lambda)-1)/lambda*, where *lambda*=0.3974732.(Figure 1.1.2) There are 6 model s from the combination of 2 kinds of criteria and 3 kinds of directions. We choose stepwise with BIC as the final model:

response ~ CYP3A43 + FNDC4 + CAP1 + KCNS3

(b)PLS: 107 components can explain 91.39 percent of variance. (Figure 1.1.3)

(c)Ridge regression with *lambda*=58378 by Cross validation.

Section 2------Data set k of response TP53

(a) Remove one outlier 310th detected by Cook's distance from the full mode and we need to do WLS fitting. (Figure 1.2.1) Do box-cox transformation *log(TP53)*, since *lambda*= 0.9407042. (Figure 1.2.2) In model selection, choose stepwise with BIC as the final model:

response ~ TCTN1 + HSD17B3 + PUS3 + RPH3AL + PYY +
PSTPIP1 + HDHD1A + ZNF556 + C14orf94 +
MAP2K3 + SFRS1 + CSRP1

(b) PLS: 104 components can explain 91.58 percent of variance. (Figure 1.2.3)

(c) Ridge regression with *lambda*= 1376 by Cross validation.

Section 3------ Data set n of response 1

(a) Remove two outliers 496th and 310th detected by Cook's distance, one by one, from

the full model. (Figure 1.3.1) Do box-cox transformation *((response^lambda)-1)/lambda*, where *lambda*= 0.3974732. (Figure 1.3.2) In model selection, choose stepwise with BIC as the final model:

response ~ FNDC4 + BRP44L + PYY

(b) PLS: 105 components can explain 91.70 percent of variance. (Figure 1.3.3)

(c) Ridge regression: *lambda* is almost infinite by Cross validation.

Section 4------ Data set n of response TP53

(a)No outlier is detected by Cook's distance but we need to do WLS fitting. (Figure 1.4.1) Do box-cox transformation is *log(TP53)*, since *lambda*= 0.9321922. (Figure 1.4.2) In model selection, choose stepwise with BIC as the final model:

response ~ PLSCR3 + MSH3 + FAM50B + TAF7L + TUFM

(b) PLS: 105 components can explain 90.67 percent of variance. (Figure 1.4.3)

(c) Ridge regression with *lambda*= 827.8 by Cross validation.

Part II Nonparametric Regression-GMC

- **Main findings:**

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:

$$g_1(x,\alpha) = x^{\alpha} \quad \alpha \in [0,3]$$

$$g_2(x,\alpha) = \begin{cases} \alpha(e^x - 1) + \alpha & x < 0 \\ x + \alpha & o.w. \end{cases} \quad \alpha \in [10{,}3000]$$

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 checking the residual plots, 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.

- **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
```

Note: 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.

- **Results & Presentations:**

The GMC in the set K with response for different screening in first **g** family is presented in Table 2.1.1. The GMC in the set K with TP53 for different screening in first **g** family is presented in Table 2.1.2. The GMC in different families using random forest for screening is presented in Table 2.1.3. The summary of 4 databases containing the number of selected variables, G function, GMC and MAE is presented in Table 2.1.4. The final results of selected variables corresponding to 4 databases are presented in Table 2.1.5. The residual plots corresponding to 4 databases are presented in Figure 2.1.6.

Part III Logistic Regression

- **Methods Introduction :**

This part contains four section. Each section corresponds to one response of one data set performing logistic regression using the variables selected by stepwise-BIC in part 1. Each section would give four figures. The first one is the summary of the OLS, the other three are summary of logistic regression with the response converted by 0.25, 0.5, 0.75 quantile, respectively.

- **Main findings:**

  (1) All the coefficients of logistic regression are much smaller than the OLS model. The sign of the coefficients of logistic models are same with the corresponding OLS model except for some insignificant coefficients.
  (2) The p-values are almost all greater than the OLS model.
  (3) The logistic model with response converted by 0.5 quantile have more significant coefficients than the other two logistic model.

- **Results & Presentations:**

Section 1------Data set k of response 1

The OLS model from BIC stepwise procedure is as Figure 3.1.1. Using the same variables, converting the response into 0-1 by 0.25 quantile, the logistic model is as Figure 3.1.2. Converting the response into 0-1 by 0.5 quantile, the logistic model is as Figure 3.1.3. Converting the response into 0-1 by 0.75 quantile, the logistic model is as Figure 3.1.4. The misclassification plot for the three thresholds are showed in Figure 3.1.5, Figure 3.1.6, Figure 3.1.7.

Section 2------Data set k of response TP53

The OLS model from BIC stepwise procedure is as Figure 3.2.1. Using the same variables, converting the response into 0-1 by 0.25 quantile, the logistic model is as Figure 3.2.2. Converting the response into 0-1 by 0.5 quantile, the logistic model is as Figure 3.2.3. Converting the response into 0-1 by 0.75 quantile, the logistic model is as Figure 3.2.4. The misclassification plot for the three thresholds are showed in Figure 3.2.5, Figure 3.2.6, Figure 3.2.7.

Section 3------ Data set n of response 1

The OLS model from BIC stepwise procedure is as Figure 3.3.1. Using the same variables, converting the response into 0-1 by 0.25 quantile, the logistic model is as Figure 3.3.2. Converting the response into 0-1 by 0.5 quantile, the logistic model is as Figure 3.3.3. Converting the response into 0-1 by 0.75 quantile, the logistic model is as Figure 3.3.4. The misclassification plot for the three thresholds are showed in Figure 3.3.5, Figure 3.3.6, Figure 3.3.7.

Section 4------ Data set n of response TP53

The OLS model from BIC stepwise procedure is as Figure 3.4.1. Using the same variables, converting the response into 0-1 by 0.25 quantile, the logistic model is as Figure 3.4.2. Converting the response into 0-1 by 0.5 quantile, the logistic model is as Figure 3.4.3. Converting the response into 0-1 by 0.75 quantile, the logistic model is as Figure 3.4.4. The misclassification plot for the three thresholds are showed in Figure 3.4.5, Figure 3.4.6, Figure 3.4.7.

Figure 1.1.1

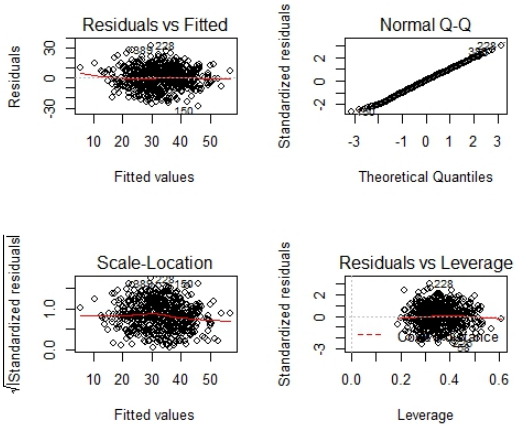

Figure 1.1.2

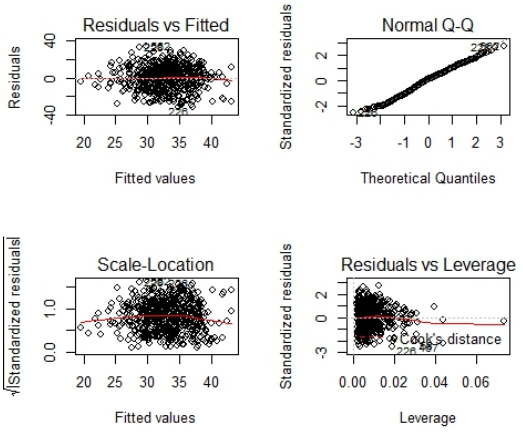

Figure 1.1.3

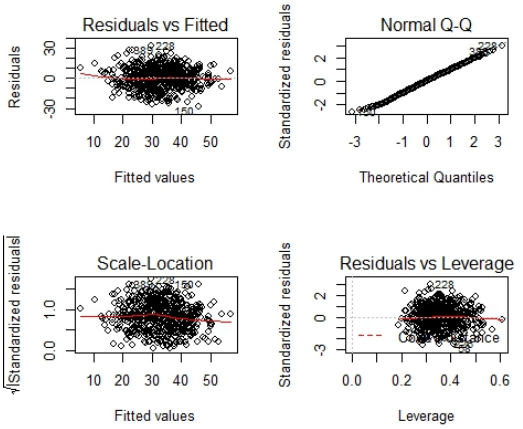

Figure 1.2.1

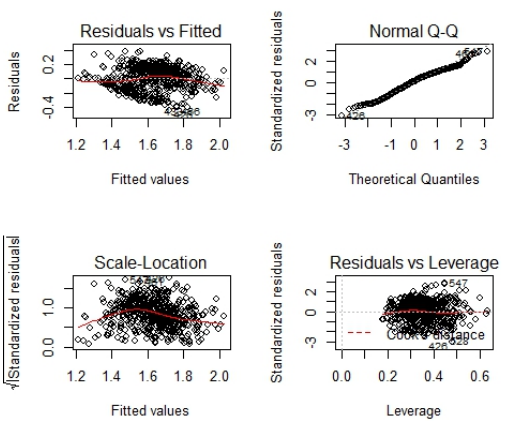

Figure 1.2.2

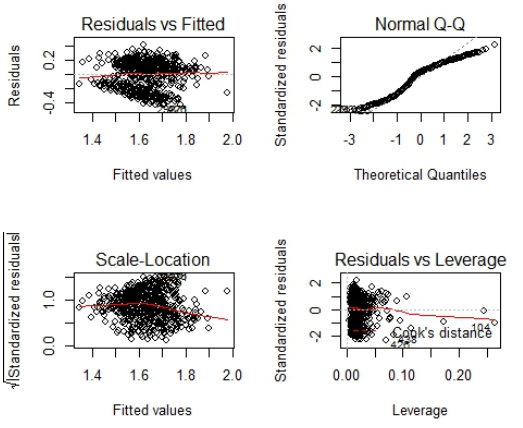

Figure 1.2.3

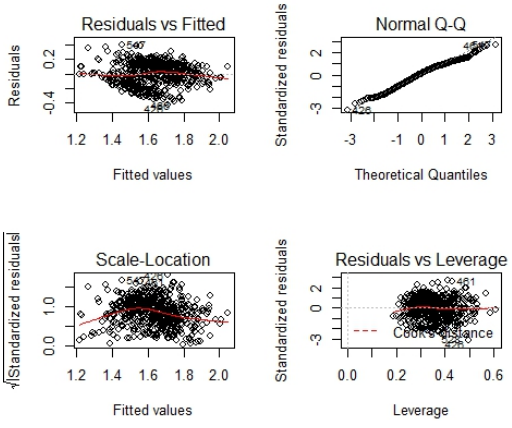

Figure 1.3.1

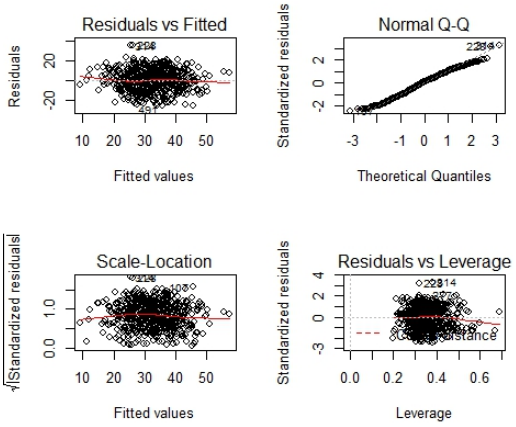

Figure 1.3.2

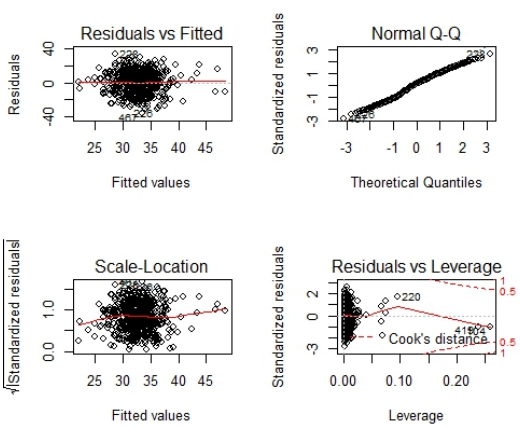

Figure 1.3.3

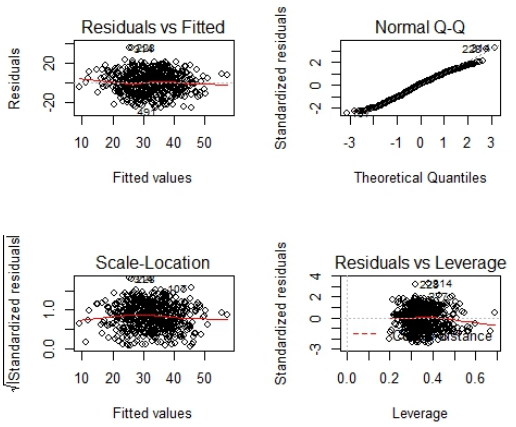

Figure 1.4.1

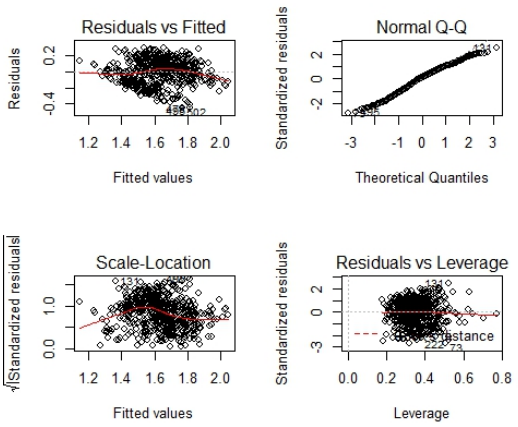

Figure 1.4.2

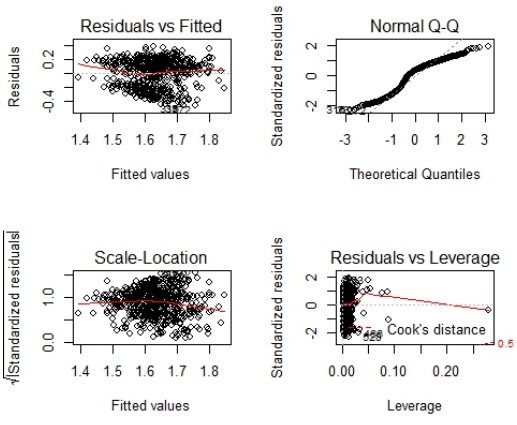

Figure 1.4.3

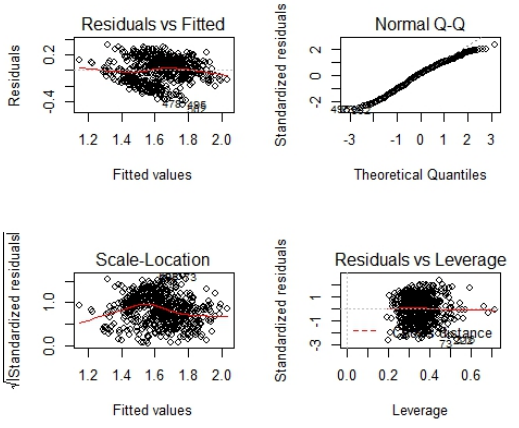

Table 2.1.1

# 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

Table 2.1.2

# 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

Table 2.1.3

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

Table 2.1.4

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

Table 2.1.5

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

Figure 2.1.6

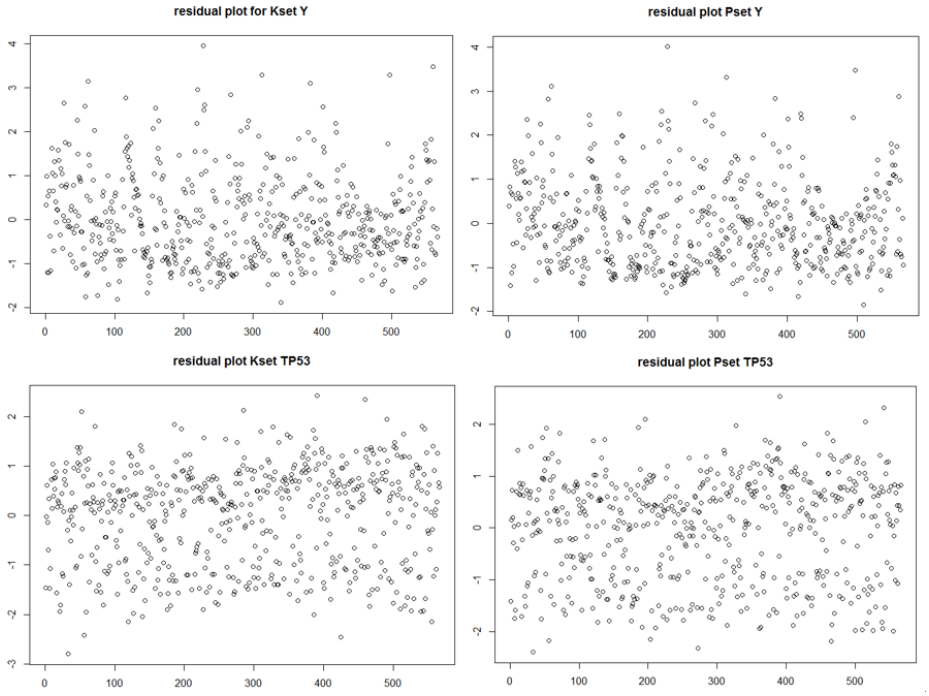

Figure 3.1.1

```
Call:
lm(formula = response ~ CYP3A43 + FNDC4 + CAP1 + KCNS3, data = s

Residuals:
    Min      1Q  Median      3Q     Max
-31.866  -8.602   1.296   8.708  34.547

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  25.7911    23.2965   1.107 0.268733
CYP3A43      20.3120     5.4063   3.757 0.000190 ***
FNDC4        -6.7553     1.9254  -3.509 0.000487 ***
CAP1         -3.4289     1.0801  -3.174 0.001584 **
KCNS3         1.5198     0.6038   2.517 0.012105 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 12.67 on 561 degrees of freedom
Multiple R-squared:  0.07753,   Adjusted R-squared:  0.07095
F-statistic: 11.79 on 4 and 561 DF,  p-value: 3.359e-09
```

Figure 3.1.2

```
Call:
glm(formula = response ~ ., family = binomial(link = "logit"),
    data = y.part1.0.25)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.3571  -0.7003   0.6694   0.7911   1.2430

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -3.9508     4.4516  -0.887  0.37481
CYP3A43       3.3476     1.0720   3.123  0.00179 **
FNDC4        -0.5651     0.3431  -1.647  0.09954 .
CAP1         -0.3844     0.1995  -1.927  0.05394 .
KCNS3         0.1882     0.1162   1.620  0.10517
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 637.66  on 565  degrees of freedom
Residual deviance: 615.78  on 561  degrees of freedom
AIC: 625.78

Number of Fisher Scoring iterations: 4
```

Figure 3.1.3

```
Call:
glm(formula = response ~ ., family = binomial(link = "logit"),
    data = y.part1.0.5)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-1.87708  -1.10925   0.03785   1.11644   1.76628

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -1.3326     3.8609  -0.345  0.72997
CYP3A43       3.1254     0.9108   3.432  0.00060 ***
FNDC4        -1.0349     0.3319  -3.118  0.00182 **
CAP1         -0.5270     0.1816  -2.901  0.00371 **
KCNS3         0.2772     0.1005   2.760  0.00579 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 784.64  on 565  degrees of freedom
Residual deviance: 743.69  on 561  degrees of freedom
AIC: 753.69

Number of Fisher Scoring iterations: 4
```

Figure 3.1.4

```
Call:
glm(formula = response ~ ., family = binomial(link = "logit"),
    data = y.part1.0.75)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.3374  -0.7863  -0.6349   0.6858   2.2324

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   0.8217     4.3599   0.188  0.85051
CYP3A43       2.8647     0.9993   2.867  0.00415 **
FNDC4        -1.3189     0.4207  -3.135  0.00172 **
CAP1         -0.6219     0.2122  -2.931  0.00338 **
KCNS3         0.2141     0.1105   1.938  0.05263 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 637.66  on 565  degrees of freedom
Residual deviance: 603.29  on 561  degrees of freedom
AIC: 613.29

Number of Fisher Scoring iterations: 4
```

Figure 3.1.5

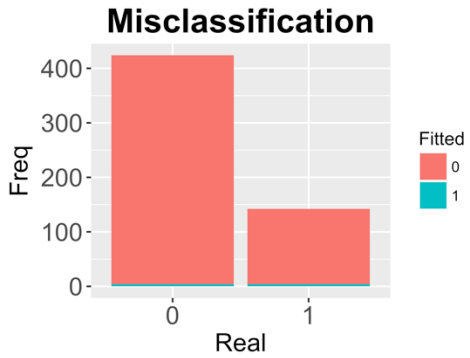

Figure 3.1.6

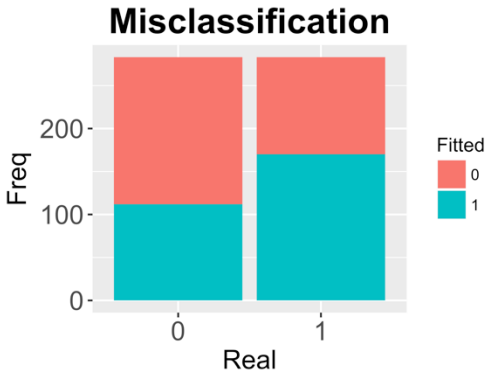

Figure 3.1.7

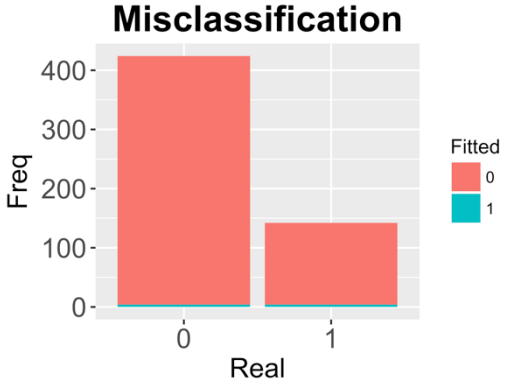

Figure 3.2.1

```
Call:
lm(formula = response ~ TCTN1 + HSD17B3 + PUS3 + RPH3AL + PYY +
    PSTPIP1 + HDHD1A + ZNF556 + C14orf94 + MAP2K3 + SFRS1 + CSRP1,
    data = tpk1.tr)

Residuals:
     Min       1Q   Median       3Q      Max
-0.43106 -0.15537  0.05343  0.14445  0.41827

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.055023   0.280696   0.196 0.844664
TCTN1        0.034267   0.011694   2.930 0.003526 **
HSD17B3     -0.134548   0.034610  -3.888 0.000114 ***
PUS3         0.039657   0.014595   2.717 0.006791 **
RPH3AL       0.053633   0.014069   3.812 0.000153 ***
PYY         -0.036527   0.012484  -2.926 0.003574 **
PSTPIP1      0.073098   0.020374   3.588 0.000363 ***
HDHD1A      -0.039247   0.009908  -3.961 8.44e-05 ***
ZNF556       0.054096   0.015813   3.421 0.000670 ***
C14orf94     0.029170   0.011130   2.621 0.009012 **
MAP2K3       0.058987   0.017871   3.301 0.001027 **
SFRS1        0.069304   0.018970   3.653 0.000283 ***
CSRP1        0.033557   0.012047   2.785 0.005529 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1904 on 554 degrees of freedom
Multiple R-squared:  0.1929,    Adjusted R-squared:  0.1755
F-statistic: 11.04 on 12 and 554 DF,  p-value: < 2.2e-16
```

Figure 3.2.2

```
Call:
glm(formula = response ~ ., family = binomial(link = "logit"),
    data = y.part2.0.25)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.2847  -0.1714   0.5860   0.7804   1.6286

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -5.9965     3.5997  -1.666 0.095748 .
TCTN1         0.3592     0.1536   2.339 0.019336 *
HSD17B3      -1.6097     0.4405  -3.654 0.000258 ***
PUS3          0.1995     0.1839   1.085 0.278119
RPH3AL        0.3371     0.1907   1.768 0.077115 .
PYY          -0.2364     0.1400  -1.689 0.091229 .
PSTPIP1       0.5729     0.2922   1.961 0.049924 *
HDHD1A       -0.3487     0.1308  -2.666 0.007672 **
ZNF556        0.4212     0.2207   1.908 0.056369 .
C14orf94      0.2331     0.1442   1.617 0.105975
MAP2K3        0.5238     0.2304   2.273 0.022998 *
SFRS1         0.1455     0.2425   0.600 0.548354
CSRP1         0.2548     0.1514   1.682 0.092481 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 638.24  on 566  degrees of freedom
Residual deviance: 587.64  on 554  degrees of freedom
AIC: 613.64

Number of Fisher Scoring iterations: 4
```

Figure 3.2.3

```
Call:
glm(formula = response ~ ., family = binomial(link = "logit"),
    data = y.part2.0.5)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.9691  -1.0234  -0.2542   1.0183   2.0691

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -19.4307     3.5162  -5.526 3.27e-08 ***
TCTN1         0.3247     0.1357   2.393 0.016701 *
HSD17B3      -1.0655     0.4072  -2.617 0.008870 **
PUS3          0.4782     0.1701   2.812 0.004930 **
RPH3AL        0.5021     0.1668   3.011 0.002603 **
PYY          -0.4530     0.1999  -2.266 0.023425 *
PSTPIP1       0.5735     0.2363   2.427 0.015211 *
HDHD1A       -0.4252     0.1191  -3.571 0.000355 ***
ZNF556        0.5939     0.1957   3.035 0.002409 **
C14orf94      0.3078     0.1323   2.326 0.020016 *
MAP2K3        0.6772     0.2099   3.226 0.001257 **
SFRS1         1.0652     0.2308   4.615 3.93e-06 ***
CSRP1         0.3407     0.1415   2.408 0.016023 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 786.03  on 566  degrees of freedom
Residual deviance: 682.44  on 554  degrees of freedom
AIC: 708.44

Number of Fisher Scoring iterations: 4
```

Figure 3.2.4

```
Call:
glm(formula = response ~ ., family = binomial(link = "logit"),
    data = y.part2.0.75)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.4103  -0.6761  -0.4161   0.1385   2.4906

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -32.3395     4.6360  -6.976 3.04e-12 ***
TCTN1         0.5071     0.1606   3.157 0.001595 **
HSD17B3      -0.7071     0.4718  -1.499 0.133914
PUS3          0.7859     0.2138   3.676 0.000237 ***
RPH3AL        0.6952     0.1849   3.760 0.000170 ***
PYY          -0.5900     0.3465  -1.703 0.088567 .
PSTPIP1       0.7716     0.2667   2.894 0.003807 **
HDHD1A       -0.6180     0.1502  -4.115 3.87e-05 ***
ZNF556        0.4197     0.2215   1.895 0.058144 .
C14orf94      0.7656     0.1678   4.561 5.09e-06 ***
MAP2K3        0.8481     0.2575   3.294 0.000989 ***
SFRS1         1.4743     0.2901   5.081 3.74e-07 ***
CSRP1         0.4901     0.1795   2.730 0.006325 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 638.24  on 566  degrees of freedom
Residual deviance: 499.30  on 554  degrees of freedom
AIC: 525.3

Number of Fisher Scoring iterations: 6
```

Figure 3.2.5

Figure 3.2.6

Figure 3.2.7

Figure 3.3.1

```
Call:
lm(formula = response ~ FNDC4 + BRP44L + PYY, data = setp2.tr)

Residuals:
    Min      1Q  Median      3Q     Max
-35.954  -9.871   1.513   9.331  33.443

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  75.8619    10.7543   7.054 5.13e-12 ***
FNDC4        -6.7683     1.9404  -3.488 0.000525 ***
BRP44L       -2.8033     0.8926  -3.141 0.001775 **
PYY           2.2891     0.8295   2.760 0.005976 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 12.83 on 562 degrees of freedom
Multiple R-squared:  0.05242,   Adjusted R-squared:  0.04736
F-statistic: 10.36 on 3 and 562 DF,  p-value: 1.201e-06
```

Figure 3.3.2

```
Call:
glm(formula = response ~ ., family = binomial(link = "logit"),
    data = y.part3.0.25)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.9860  -0.8603   0.7182   0.7820   1.0914

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   4.7389     1.9841   2.388   0.0169 *
FNDC4        -0.5589     0.3358  -1.664   0.0960 .
BRP44L       -0.2485     0.1614  -1.540   0.1236
PYY           0.2021     0.2066   0.978   0.3279
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 637.66  on 565  degrees of freedom
Residual deviance: 630.99  on 562  degrees of freedom
AIC: 638.99

Number of Fisher Scoring iterations: 4
```

Figure 3.3.3

```
Call:
glm(formula = response ~ ., family = binomial(link = "logit"),
    data = y.part3.0.5)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.8782  -1.1540  -0.1548   1.1437   1.5780

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   5.5681     1.8631   2.989  0.00280 **
FNDC4        -0.9867     0.3228  -3.057  0.00224 **
BRP44L       -0.3558     0.1441  -2.468  0.01359 *
PYY           0.4318     0.2084   2.072  0.03829 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 784.64  on 565  degrees of freedom
Residual deviance: 760.68  on 562  degrees of freedom
AIC: 768.68

Number of Fisher Scoring iterations: 4
```

Figure 3.3.4

```
Call:
glm(formula = response ~ ., family = binomial(link = "logit"),
    data = y.part3.0.75)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.1898  -0.7941  -0.6897   0.6728   2.0006

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   6.7427     2.1322   3.162  0.00157 **
FNDC4        -1.2632     0.4129  -3.059  0.00222 **
BRP44L       -0.4020     0.1664  -2.415  0.01572 *
PYY           0.2039     0.1326   1.538  0.12406
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 637.66  on 565  degrees of freedom
Residual deviance: 617.94  on 562  degrees of freedom
AIC: 625.94

Number of Fisher Scoring iterations: 4
```

Figure 3.3.5

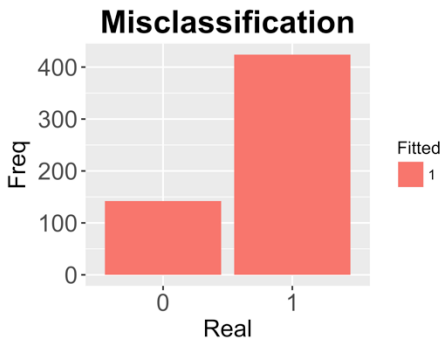

Figure 3.3.6

Figure 3.3.7

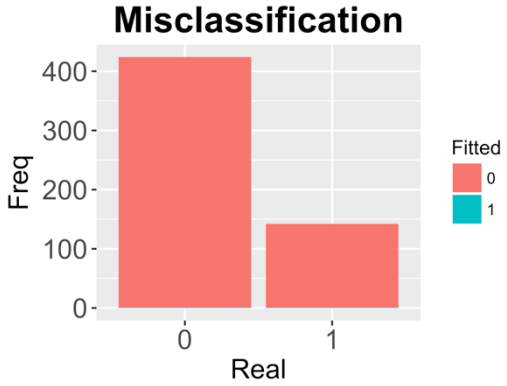

Figure 3.4.1

```
Call:
lm(formula = response ~ PLSCR3 + MSH3 + FAM50B + TAF7L + TUFM,
    data = tpp.tr)

Residuals:
     Min       1Q   Median       3Q      Max
-0.46034 -0.17997  0.06519  0.15373  0.38136

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.52085    0.22381   2.327 0.020307 *
PLSCR3       0.06532    0.01437   4.544 6.76e-06 ***
MSH3         0.06681    0.01790   3.733 0.000209 ***
FAM50B       0.03852    0.01008   3.821 0.000148 ***
TAF7L       -0.08549    0.02742  -3.118 0.001913 **
TUFM         0.04468    0.01736   2.574 0.010300 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1974 on 562 degrees of freedom
Multiple R-squared:  0.1227,    Adjusted R-squared:  0.1149
F-statistic: 15.71 on 5 and 562 DF,  p-value: 1.706e-14
```

Figure 3.4.2

```
Call:
glm(formula = response ~ ., family = binomial(link = "logit"),
    data = y.part4.0.25)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.1437   0.1550   0.6529   0.7799   1.4559

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.83160    2.69536  -0.309 0.757680
PLSCR3       0.37704    0.17605   2.142 0.032225 *
MSH3         0.20474    0.21754   0.941 0.346626
FAM50B       0.41837    0.12353   3.387 0.000707 ***
TAF7L       -0.94979    0.33143  -2.866 0.004161 **
TUFM        -0.03619    0.20551  -0.176 0.860230
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 638.81  on 567  degrees of freedom
Residual deviance: 611.95  on 562  degrees of freedom
AIC: 623.95

Number of Fisher Scoring iterations: 4
```

Figure 3.4.3

```
Call:
glm(formula = response ~ ., family = binomial(link = "logit"),
    data = y.part4.0.5)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.30281  -1.04744   0.08192   1.03961   2.19177

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -13.6216     2.8252  -4.821 1.43e-06 ***
PLSCR3        0.6732     0.1642   4.101 4.11e-05 ***
MSH3          0.9945     0.2054   4.841 1.29e-06 ***
FAM50B        0.4227     0.1137   3.719 0.000200 ***
TAF7L        -1.2823     0.3991  -3.213 0.001313 **
TUFM          0.6753     0.1974   3.421 0.000623 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 787.42  on 567  degrees of freedom
Residual deviance: 694.61  on 562  degrees of freedom
AIC: 706.61

Number of Fisher Scoring iterations: 4
```

Figure 3.4.4

```
Call:
glm(formula = response ~ ., family = binomial(link = "logit"),
    data = y.part4.0.75)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.9990  -0.7278  -0.4998   0.0467   2.5027

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -24.0736     3.4313  -7.016 2.28e-12 ***
PLSCR3        1.0345     0.1937   5.339 9.33e-08 ***
MSH3          1.0629     0.2335   4.551 5.33e-06 ***
FAM50B        0.4930     0.1314   3.753 0.000175 ***
TAF7L        -0.5865     0.4013  -1.461 0.143918
TUFM          1.0014     0.2410   4.155 3.26e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 638.81  on 567  degrees of freedom
Residual deviance: 537.19  on 562  degrees of freedom
AIC: 549.19

Number of Fisher Scoring iterations: 5
```

Figure 3.4.5

Figure 3.4.6

Figure 3.4.7