Stat 601 Final Project

TEAM 7

***Abstract*—In this article, we use linear regression and GMC method to choose predictors that explain the response variables well. From the diagnostic of linear regression of "response", problem with normality, homogeneity and collinearity is showed. Box-Cox transformation is used. To overcome the effect of collinearity, stepwise regression and lasso regression are used for variable selection. We also use polynomial regression to investigate the relationship between the response variables and the polynomial terms of the predictor variables. PCA and PLS are also applied but they have limited effect on variable selection.**

For the part to optimize the simplified GMC function with two penalty we use two methods. One is to set the optimized beta with values less than a threshold to zero, and then do optimization iteratively. Another way we choose is similar to forward selection. Every time we add a variable, we calculate the corresponding β for each variable subset that maximize the simplified GMC function with two penalty, and then select the one with the largest GMC(Y—g(x)). The second way performs better with less variables selected and relatively large GMC values. For the part to optimize GMC with lasso penalty, we do the similar forward procedure as above. And choose the four most important variables, their linear combination has the GMC value 0.11.

We also found that there are two variables appear in both linear model and GMC variable selections. They are "PYY" and "KCNJ10". We consider them to be relatively more important.

***Keywords*—Stepwise, Lasso, Polynomial, GMC.**

I. Introduction

This Article reports the analysis of a medical clinic data. We consider to find some genes that may have some relationship with ovarian carcinoma and the reasonable model for these genes.

There are two response variables we will consider. The main response variable is named "response", which is the value of "daystolastfollowup". Another one is "TP53", which is related to ovarian cancer.

We choose two subsets of the data("Set_a.csv" and "Set_i'.csv') to analyze. In this paper, we simply call this two datasets as "subset A" and "subset B". Each subsets contains all these 567 observation with 201 genes(including "TP53"). And there are 567 observation with 391 different genes in the combined subset A&B.

II. Pre-processing

In order to find the proper transformation of the response variable("response"), we first build a linear regression with all the 201 explanatory variables in each subset.

TABLE I: three kinds of points for subset A and subset B.

	Outlier	Leverage Points	Influential Points
subset A	1(495)	0	29
subset B	1(495)	1(310)	25
A&B	1(495)	0	39

TABLE II: VIF for A and subset B.

	subset A	subset B	combined subset A&B
VIF>5	28	30	381
VIF>10	2	0	199

A. Diagnostic

- Outliers, Leverage Points and Influential Points

With the criteria for outliers(using $outlierTest()$),high leverage points($h_i > \frac{2p}{n}$) and influential points($D_i > \frac{4}{n-p}$), there is only one outlier and high leverage point, while there are above 20 influential points in each set.

- Multicollinearity

Through VIF, high multicollinearity is shown in both two datasets(See TABLE II). The combined dataset has much more serious collinearity. Therefore, remedies are needed to overcome the effects of collinearity.

- Normality

From the Q-Q plot of the linear regression "$response \sim .$" in each subset(see Fig. 1, Fig. 2 and Fig. 3), we consider that the normality of the error term is not very good, the distribution of the error seem like right skewed.

- Homogeneity

From the residual plot of the linear regression "$response \sim .$" in each subset(see Fig. 1, Fig. 2 and Fig. 3), we observed that the residual is biased, the number of the residuals that are positive is larger than the number of the residuals that are negative. And the variance of the residual become larger when the fitted value is larger, hence, we consider the error term of this full linear regression model is heteroscedastic.

B. Transformation

Since the diagnostic plots using the original response variable show that the homogeneity and the normality

TABLE III: Lasso regression of $TP53$.

	subset A	subset B	combined A&B
Mallow's Cp	84	51	94
k-fold CV	16	17	7

are not good. Thus transformation is needed. Since the response variables are positive, we consider to use log transformation(see Fig. 4, Fig. 5 and Fig. 6) and square root transformation(see Fig.7, Fig. 8 and Fig. 9). We also consider Box-Cox transformation(see Fig.10, Fig. 11 and Fig. 12) due to the lack of normality.

From these diagnostic plots, we observe that after using Box-Cox transformation and square root transformation, the residual are around zero, and the homogeneity and the normality both become much better than the original one. Compared this two transformation we found that the residual plot using Box-Cox transformation is a little bit better than the one using square root transformation.

Therefore, the transformation of the response variable $Box-Cox(response)$ is chosen.

III. Linear Regression Models

A. Stepwise Regression

Due to the large number of variables and collinearity, variable selection is needed. The most common way is to use stepwise regression.

We use forward, backward,stepwise method together with AIC and BIC criteria to do variable selection. We get 6 corresponding models for each response and variables combination.

In order to choose the best linear model for each response and variables combination, we use k-fold cross validation and choose the one with the smallest overall mean square error(Overall MS) among the six model.

Since all the models using AIC retain so many variables while using BIC the model can get one digit number which is exactly what we want. Therefore, we decide to first choose models using BIC and then compare Overall MS among three models(forward, backward, both) to choose the final model in each subset.

The results of variables which are selected in three subsets are below:

Subset A: $response \sim PYY + FNDC4$
Subset B: $response \sim KCNJ10$
Subset A&B: $response \sim FNDC4 + CRTAM + KCNJ10 + CXorf9$

Subset A: $TP53 \sim PALMD + GYPB + CRTAM + SLIT3 + PCAF + RHOBTB3 + RAB36 + ATP5G2 + TM9SF4 + RPL7A + HSPB2 + PYY + LOC390688 + GPR126 + C9orf78 + POLL + VAPB$
Subset B: $TP53 \sim ODC1 + FOSB + ANKRD46 + NUBP1 + PYY + VWA1 + CHSY1 + SNRPE + BRWD1 + AMIGO2 + MAP7D3 + KNTC1 + KIAA0753 + KIF5C + NDUFA5 + ATP10A + ITIH2 + HERC2$
Subset A&B: $TP53 \sim PYY + CRTAM + SLIT3 + PCAF + NIF3L1 + SURF2 + HSPB2 + SLC9A8 + POP5 + LOC390688 + TCF15 + CLK4 + C9orf78 + POLL + LBA1 + MED24 + RELB + FOSB + ANKRD46 + MIF + FLJ10815 + CHSY1 + MAP7D3 + CDKN2A + KIF5C + GCAT + ATP10A + HERC2 + WDR79 + ACVRL1 + TM9SF4 + RNF144A$

Remark: The response has been transformed by Box-Cox in stepwise regression.

B. Polynomial Regression

When we do GMC variable selection which is in the next section, we find that although nearly all predictor variable has a small $GMC(Y|X_i)$, some of that after some transformations such as x^2 or e^x will have a much larger $GMC(Y|X_i)$. In that case, we consider that there are some predictor variables, whose polynomial term may has relation with response variable, therefore, we decide to do polynomial regression.

Since the number of observations is only 567, we cannot do more than 2 degrees of polynomial regression to guarantee that $(X'X)$ is non-singular. Therefore, we only do polynomial regression with degree 2 in subset A and subset B .

Since the models using AIC retain so many variables, we use cross-validation to choose the final model from the three BIC model in each subset.

The results of variables which are selected in two subsets are below:

Subset A: $response \sim SPINK1\_1 + HOXC10\_2 + HABP2\_2 + SBNO2\_1 + SPRY4\_1 + FNDC4\_1 + ANXA7\_2 + CRTAM\_1$
Subset B: $response \sim KCNJ10\_1 + HOXC10\_2 + IFNA2\_2 + CEL\_1$

Subset A: $TP53 \sim GYPB\_1 + FAM59A\_2 + SLC5A1\_2 + RAI16\_2 + FTHP1\_1 + CYP1B1\_2 + SLIT3\_1 + PCAF\_1 + GAS2\_1 + RHOBTB3\_1 + SS18L2\_2 + UBP1\_2 + TM9SF4\_1 + FAHD2A\_2 + COL5A1\_1 + COX8A\_2 + PYY\_1 + DOCK9\_2 + C9orf78\_1 + NPFFR1\_1 + C8orf33\_2 + TRIM36\_2 + VAPB\_1 + RELB\_1 + OSBPL8\_2 + SS18L2\_1 + HSPB2\_1 + RAB36\_1 + PALMD\_1 + IL6ST\_1$
Subset B: $TP53 \sim RNF144A\_1 + FOSB\_1 + ANKRD46\_1 + MYBPH\_1 + ABCA3\_2 + TARDBP\_2 + NUBP1\_1 + TAF1C\_2 + CHSY1\_1 + BRWD1\_1 + CEL\_1 + KNTC1\_1 + KIAA0753\_1 + SRR\_2 + NDUFA5\_1 + ATP10A\_1 + HERC2\_1 + PYY\_1 + LIF\_2 + R3HDM1\_1 + WDR79\_2$

Remark: The response has been transformed by Box-Cox in stepwise regression.

TABLE IV: Lasso regression of $response$.

	subset A		subset B		combined subset A&B	
	$response$	$Box-Cox(response)$	$response$	$Box-Cox(response)$	$response$	$Box-Cox(response)$
Mallow's Cp	3	0	7	13	9	1
k-fold CV	0	0	1	1	0	1

C. Lasso Regression

Since high multicollinearity is shown in both two datasets, we decide to use ridge or lasso regression to overcome the effect of the collinearity. A ridge solution can be hard to interpret because it is not sparse (no β are set exactly to 0). But through lasso, some of the coefficients may be shrunk exactly to zero. Therefore, we choose lasso regression instead of ridge regression.

There are two way to choose the optimal λ argument. One is choosing the argument that minimize Mallow's Cp, the other one is using k-fold cross validation. From the TABLE III and TABLE IV, we observe that the number of variables which are selected by k-fold cross validation method is much smaller than the number of variables which are selected by Mallow's Cp. In order to obtain the appropriate number of variables, we choose to use Mallow's Cp when choosing the variables for $response$, and use k-fold cross validation to choose the variables for $TP53$.

• Lasso regression of "TP53"

In subset A, the lasso regression of "$TP53$" selects 16 variable. In subset B, the lasso regression of "$TP53$" selects 17 variable. And in combined subset A&B, the lasso regression of "$TP53$" selects 7 variable.

The results of variables which are selected in three subsets are below:

Subset A: $TP53 \sim FARS2 + C20orf12 + RBMX + NR5A2 + RPLP0 + RHOBTB3 + RAB36 + ATP5G2 + TM9SF4 + RNF40 + LOC390688 + NPFFR1 + C9orf78 + POLL + DNMBP + ARHGEF2$
Subset B: $TP53 \sim PITPNA + ANKRD46 + MIF + NUBP1 + SCAMP1 + PYY + CRELD1 + JARID2 + WDR79 + KIAA0753 + KIF5C + APEX1 + ATP10A + KBTBD4 + KIF1C + HERC2 + E2F5$
Subset A&B: $TP53 \sim ATP5G2 + TM9SF4 + NPFFR1 + POLL + WDR79 + KIAA0753 + HERC2$

• Lasso regression of "response"

Since on one hand, the diagnostic plot using the original "$response$"is not too bad, on the other hand, the number of the variables which are selected by Lasso regression of "$Box-Cox(response)$" is either too small or too large. Therefore, we decided to use the original "$response$" to select the variables. (Detail are showed in Appendix)

Therefore, the results of variables which are selected in three subsets are below:

Subset A: $response \sim SPINK1 + PYY + CEL$
Subset B: $response \sim SPINK1 + PYY + CEL + MMP16 + HSD11B2 + KCNJ10 + ITIH2$
Subset A&B: $response \sim SPINK1 + PYY + CEL + CRTAM + FAHD2A + MMP16 + HSD11B2 + KCNJ10 + ITIH2$

Remark: The response variable isn't transformed in the final model of lasso regression.

• Comments on Lasso regression

From the result above, we observe that, the genes PYY, CEL and SPINK1 are three very important genes, which are selected in subset A, subset B and combined subset A&B. Besides, gene KCNJ10 is also a very important gene, which are in subset B and combined subset A&B. And when we use k-fold cross validation, the lasso regression only selects gene KCNJ10.

Since TP53 isn't selected in any subset, we consider that maybe there are some variables which are more important than TP53. We then check whether these four important genes variables also are selected from the lasso regression of "$TP53$". We observe that PYY is also in the model in subset B. Therefore, we consider that gene PYY may be a gene that is more important than TP53.

D. Other We Tried

Since PCA and PLS can also overcome the effect of the collinearity, we also try to use PCA and PLS to generate some components which are obtained from the data X. However, from the summary of the pcr() and plsr() in R, we find that in order to explain 85% above variance of X, at least 85 components should be involved into the model. In that case, it is impossible to delete any predictor variables by using PCA or PLS. Therefore, PCA or PLS regression is not suitable in this project.

E. Summary for linear regression model

We observe that the variable selection using AIC or BIC is not consistent in the single subset and combined subset. This is mainly due to the collinearity, which cannot be fully overcome by AIC or BIC.When using Lasso regression, all the variables in the final model of subset A or subset B are still in the final model of combined subset. This is because Lasso regression can overcome the effect of collinearity. Since from the result of VIF, we know that the combined subset has extremely serious collinearity, the Lasso regression is more suitable in this study, and gives more consistent result.There are

some variables which are selected in three models, such as PYY, SPINK1, CEL and KCNJ10.

IV. GMC Variable Selections

In GMC variable selection, we first consider to choose some of the variables which has large $GMC(Y|X_i)$. However, even the largest $GMC(Y|X_i)$ is still lower than 0.1. Therefore, variables can't be selected simply by comparing their GMC, instead, we need to find the GMC between response variable and the combination of predictor variables.

In order to obtain large $GMC(Y|g(X))$ with few variables, we use GMC and Lasso penalty. To simplify, we just use the linear combination of the predictor variables. Hence, $g(x_1, x_2, ..., x_p) = g(X\beta)$.

- $g(x) = X\beta$, $Y' = g^{-1}(Y) = Y$
- $g(x) = e^{X\beta}$, $Y' = g^{-1}(Y) = log(Y)$
- $g(x) = (X\beta)^2$, $Y' = g^{-1}(Y) = Y^{\frac{1}{2}}$
- $g(x) = (X\beta)^3$, $Y' = g^{-1}(Y) = Y^{\frac{1}{3}}$
- $g(x) = sin(X\beta)$, $Y' = g^{-1}(Y) = arcsin(Y)$

When the dependent variable"$response$" and predictor variables are not standardized, the GMC is always close to 1. Since the "$response$" and predictor variables are not in the same scale, we decided to scale the response variable and predictor variables without centering, so that all the variable are in the same scale and all positive.

A. Simplified GMC+Lasso

In this section, we assume that $Y = g(x_1, x_2, ..., x_p) + e$ and fitted model $g(.)$ and error term e is independent of each other. In this simplified case, we want to select the variables and coefficients which maximize $\frac{var(g(x)}{var(g(x))+var(e)} - \lambda_1|cov(g(x), e)| - \lambda_2 \sum_{i=1} |\beta_i| - -(1)$.

We use the linear regression 's ordinary least squares estimator as the initial value of β. The response variable is transformed by the inverse function of $g(x)$.

As for the choice of λ, our initial consideration is to make the 3 items to the similar magnitude. We first choose $lambda_2$ seq(0,01,0.1,by=0.01), since initial $\sum_{i=1} |\beta_i|$ is around 100, and choose $lambda_1$ seq(0,1,1,by=0.1), since the initial $|cov(g(x), e)|$ is approximately among 0.1 to 5. However, the coefficients don't have apparently change after reaching the max number of iteration(we set the max number of iteration to be 10000). Therefore, we try a wider range of λ, which does not make big difference, either

From TABLE V and TABLE VI, we observe that in subset A, $g(x) = X\beta$ gives the largest GMC among five, which is 0.34, while $g(x) = sin(X\beta)$ gives the smallest GMC, which is 0.23. In subset B, $g(x) = (X\beta)^2$ gives the largest GMC which is 0.33, while $g(x) = e^{X\beta}$ gives the smallest GMC, which is 0.28. In subset A&B, $g(x) = X\beta$ gives the largest GMC which is 0.61, while $g(x) = e^{X\beta}$ gives the smallest GMC, which is 0.47.

We first tried to drop some variables according to the coefficients from maximizing the function(1). However, after 1 single loop, we found that the number of coefficients less than 0.05 is no more than 80(see TABLE V and TABLE VI) , and the value of GMC decrease fast, and thus could not select the proper numbers of variables. So we decide to do it iteratively. Every time we set the optimized beta with values less than a threshold to zero, and then do optimization again. However, as the numbers of variables drops, the GMC value decrease a lot. So we did not adopt it as our final method.

After calculating the corresponding GMC to the maximized function (1), we found that although $g(x) = X\beta$ and $g(x) = X\beta^2$ give larger GMC, they are hard to drop variables. $g(x) = (X\beta)^3$ have relatively smaller GMC compared with these two, but it can drop more variables than $g(x) = X\beta$ and $g(x) = e^{X\beta}$.

Since the λ_2 has limited effect on the coefficients and GMC. We decide to use the way similar to forward stepwise:

- Fix λ according to our prior experiment. Start with all coefficients equal to zero.
- Find one predictor X_i most correlated with y (largest $GMC(Y|X_i)$, and add it into the model.
- Add one predictor $X_j, j \neq i$, optimize coefficients by maximizing the function above(function (1)). Choose the predictor X_j which has largest $GMC(Y|\beta_0 + \beta_1 X_i + \beta_2 X_j)$
- Until: GMC increase less than 0.01 when added any one more predictor.

In the Appendix, we simply select ten variables for $g(x) = X\beta$, $g(x) = X\beta^2$ and $g(x) = (X\beta)^3$ (The other two are not showed since their behavior are not good in the previous experiment). By comparing the GMC, we observe that $g(x) = X\beta$ and $g(x) = (X\beta)^3$ are better than $g(x) = X\beta^2$. Besides, most variables in three models are the same, and there are some variables which are also selected in the previous lasso regression or stepwise regression. For each response variable in each subset, we choose the one with the largest GMC among three transformations.

The results of variables which are selected in three subsets are below:

Subset A: $repsonse \sim (NRG1 + C1orf27 + PTPN3 + GPR6 + FARS2 + CDH22 + NPFFR1 + C21orf59 + LMO7 + VAPB)^3$
Subset B: $response \sim (KCNJ10 + GPR27 + PYY + CLEC4E + ORM1 + SNRPE + AQP5 + WDR79 + C18orf1 + CRISP1)^3$
Subset A&B: $response \sim KCNJ10 + CYB5R2 + C1orf27 + ESRRG + SLC16A5 + ACCN3 + FZR1 + GADD45G + NPY + MAP3K6$

Subset A: $TP53 \sim (POLL + FARS2 + ARHGEF2 + RELB + PIGG + RCAN3 + NCAPG + MAPKAP1 + TMEM2 + TSPAN1)^2$
Subset B: $TP53 \sim (KIAA0753 + KBTBD4 + GPR27 + KLRB1 + HSD11B2 + ZNF343 + AQP5 + HDGF + ZNF587 + CCT6A)^3$
Subset A&B: $TP53 \sim KCNJ10 + CYB5R2 + C1orf27 + ESRRG + SLC16A5 + ACCN3 + FZR1 + GADD45G + NPY + MAP3K6$

B. General GMC+Lasso

For the situation that error term e and fitted model $g(.)$ are not independent, we use $GMC(Y|g(X\beta)) - \lambda \sum_{i=1} |\beta_i|$ instead.

We first use simulation method to choose the optimal variable combinations. Simulate 200 times, each time randomly select 20 variables from the 201 predictor variables. The simulation result is not very good, the values of $\frac{Var(g(x))}{Var(g(x))+Var(e)}$ could be larger than 0.3 sometimes, but the values of $GMC(Y|g(x))$ exceed 0.1 rarely.

Therefore, we also use forward selection here. The function we optimize is $GMC(Y|g(X\beta)) - \lambda \sum_{i=1} |\beta_i|$. Since the $g(x) = X\beta$ gives relatively large GMC values, we simply use $g(x) = X\beta$ here. We calculated all the GMC value with a sequence of λ (see TABLE V and TABLE VI, which is in Appendix), and find the impact from the value of λ is limited so we set it to be 1 here for simplicity.

We only use "response" in subset B as an example. The first 4 variables we choose are **PYY, KCNJ10, MBIP, FLG**. The corresponding GMC value is 0.1074.

Compared with previous selection, for "response" in subset B, we use less variables but obtain larger GMC, this is probably because the $g(x)$ and the error term are not independent, so the large value of function(1) may not lead to large value of GMC.

C. Comments on GMC variable selection

In this section, we combine GMC and lasso to do variable selection, however, it seems that this method didn't work very well. From Lasso regression ,we knows that the large λ will make most of the coefficients to zero. However, when we optimize the two functions, even the λ is up to 500 or more, there is nearly no change of the variable coefficients, few of them have the small coefficient (less than 0.01 in the regression of "$response$"). Here are some reasons we surmise:

First, due to limited time of iteration,most of the optimizations can't converge, which means that the result we get is not the optimal solution. To be specific, there are 201 arguments we need to optimize in the function, so it may need a super large number of iteration in order to find the optimal solution. To solve this problem, we may need to drop some variables in the beginning.

Second, due to the initial value which is set intuitively, even when the optimization converge, the results are local optimal solution, which is far different with the global optimal solution. Since there are 201 coefficients, there may be a large number of local optimal. Therefore, we may need to try different value of initial parameter.

Third, the penalty applied to GMC is aimed at constraining the number of non-zero coefficients. Its effect may be influenced due to the different range of GMC term and the penalty, term, which are (0,1) and (0, ∞) respectively. According to the first term in the lasso regression, which is residual sum of square, the penalty's effect might be improved by transforming the GMC term into a value between zero to infinity.

D. Summary for GMC variable selection

In this part, we found that using forward selection may have a much better result compared with the simulation method or dropping variable whose coefficient less than a certain value. By optimizing general GMC, we may select some variable such as PYY and KCNJ10, which is also in three linear model.

V. Future Work

First, the computation speed seems to be the most problem for us. With improvement of it, we would be able to apply GMC criterion in stepwise(both direction) variable selection, which might perform better than forward method. Second, There are much more choice for $g(x)$. Since we choose the very common five. There might be some function that fit this issue better and we can dig them out in future. Last but not least, biological background might be a helper when deciding the relative importance of different genes. If we had some prior knowledge, we can use method like Bayes or weighted linear regression to get a better choice of genes.

APPENDIX

```
# ########################
# #     DIAGNOSTIC       #
# ########################
# Outliers, high leverage points and influential points
m1<-lm(response~.,data=data1Y)
m2<-lm(response~.,data=data2Y)
m3<-lm(response~.,data=alldataY)
outlierTest(m1)
# rstudent unadjusted p-value Bonferonni p
#495     5.91           7.99e-09     4.53e-06
outlierTest(m2)
#rstudent unadjusted p-value Bonferonni p
#495     5.83           1.21e-08     6.85e-06
outlierTest(m3)
#rstudent unadjusted p-value Bonferonni p
#495     4.11           6.16e-05        0.035

hcr<-2*p/n
c(1:n)[hat(model.matrix(m1))>hcr]
#integer(0)----no high leverage point in m1
c(1:n)[hat(model.matrix(m2))>hcr]
#310----one high leverage point in m2
c(1:n)[hat(model.matrix(m3))>2*391/n]
#integer(0)----no high leverage point in m3

p<-ncol(data1Y)-1;n<-nrow(data1Y)
cr<-4/(n-p)
#[1] 0.0109----criterion for Cook's distance
c(1:n)[cooks.distance(m1)>cr]
#[1]  27  29  57  62  76 116 123 159 166 195 220 228 293 310
      314 381 384 393 411 419 421 423 425 495 496 498 518
     544 561
sum(cooks.distance(m1)>cr)
#29----number of influential points in m1
c(1:n)[cooks.distance(m2)>cr]
#[1]  46  56  57  84 116 117 119 120 166 170 220 228 268 282
      310 313 314 356 403 421 492 495 533 548 561
sum(cooks.distance(m2)>cr)
#25----number of influential points in m2
sum(cooks.distance(m3)>4/(n-391))
#39----number of influential points in m3


# ########################
# #    TRANSFORMATION    #
# ########################
#
# Use powerTransform() to compute lambda for Boc-Cox
    transformation
require(alr3)
lambda1<-powerTransform(lm(response~.,data=data1Y))$lambda
#0.382---subset A response
lambda2<-powerTransform(lm(response~.,data=data2Y))$lambda
#0.364---subset B response
lambda3<-powerTransform(lm(response~.,data=alldataY))$lambda
#0.385---combined subset A & B response


# ########################
# #      STEPWISE        #
# ########################
#
################### SUBSET A of RESPONSE ###################
# AIC forward data1Y
Call:
  lm(formula=response ~ PYY + FNDC4 + SPINK1 + RAB36 + CRTAM
     + C3orf64 + ADAMTS5 + FAHD2A + HYAL2 + PYCR1 + SBNO2 +
       COX8A + GEMIN8 + RHOBTB3 + SPRY4 + ATP6V0E2 + NRG1 +
       SETMAR + NEK4 + RCAN3 + HIST1H2BH + TSPAN1 + WWC2 +
       ARFIP1 + FTHP1 + ARHGEF2 + HSPB2 + TCF15 + MED24 +
       DCC + CYB5R2 + PAEP, data = data1Y)

Residual standard error: 11.11 on 534 degrees of freedom
Multiple R-squared:  0.1958,  Adjusted R-squared:  0.1476
F-statistic: 4.063 on 32 and 534 DF,  p-value: 4.909e-12

# BIC forward data1Y
Call:
  lm(formula = response ~ PYY + FNDC4, data = data1Y)

Residual standard error: 11.91 on 564 degrees of freedom
Multiple R-squared:  0.02296,  Adjusted R-squared:  0.01949
F-statistic: 6.626 on 2 and 564 DF,  p-value: 0.001432

# AIC backward data1Y
Call:
  lm(formula = response ~ PIK3R2 + PYCR1 + NRG1 + C3orf64 +
       RAI16 + FLG + ARFIP1 + CRTAM + RCAN3 + SPINK1 + DCC +
       PCAF + GAS2 + HSPA12A + TGM1 + GEMIN8 + SNTB2 + NEK4
     + SURF2 + HIST1H2BH + MTERFD1 + TSPAN1 + MMP7 + RHOBTB3
     + ADAMTS5 + FAHD2A + PML + COX8A + ANKRD15 + HSPB2 +
       SLC9A8 + BTBD7 + SETMAR + SBNO2 + PYY + SFRS11 +
       ERCC6L + WDR57 + ATP6V0E2 + TCF15 + CEL + SPRY4 +
       PAEP + KLRF1 + FNDC4 + CADM3 + WWC2 + KIAA0143 +
       CYB5R2 + RELB + ARHGEF2 + LMO7, data = data1Y)

Residual standard error: 10.95 on 514 degrees of freedom
Multiple R-squared:  0.2473,  Adjusted R-squared:  0.1712
F-statistic: 3.248 on 52 and 514 DF,  p-value: 7.182e-12

# BIC backward data1Y
Call:
  lm(formula = response ~ PYY + FNDC4, data = data1Y)

Residual standard error: 11.91 on 564 degrees of freedom
Multiple R-squared:  0.02296,  Adjusted R-squared:  0.01949
F-statistic: 6.626 on 2 and 564 DF,  p-value: 0.001432

# AIC both data1Y
Call:
  lm(formula = response ~ PYCR1 + NRG1 + C3orf64 + RAI16 +
       ARFIP1 + CRTAM + RCAN3 + SPINK1 + DCC + GAS2 +
       HSPA12A + TGM1 + GEMIN8 + SNTB2 + NEK4 + SURF2 +
       HIST1H2BH + MTERFD1 + TSPAN1 + MMP7 + RHOBTB3 +
       ADAMTS5 + FAHD2A + PML + COX8A + ANKRD15 + HSPB2 +
       SLC9A8 + BTBD7 + SETMAR + SBNO2 + PYY + SFRS11 +
       ERCC6L + WDR57 + ATP6V0E2 + TCF15 + CEL + SPRY4 +
       PAEP + KLRF1 + FNDC4 + CADM3 + WWC2 + KIAA0143 +
       CYB5R2 + RELB + ARHGEF2 + LMO7 + RAB36,
       data = data1Y)

Residual standard error: 10.96 on 516 degrees of freedom
Multiple R-squared:  0.2436,  Adjusted R-squared:  0.1703
F-statistic: 3.324 on 50 and 516 DF,  p-value: 5.225e-12

# BIC both data1Y
Call:
  lm(formula = response ~ PYY + FNDC4, data = data1Y)

Residual standard error: 11.91 on 564 degrees of freedom
Multiple R-squared:  0.02296,  Adjusted R-squared:  0.01949
F-statistic: 6.626 on 2 and 564 DF,  p-value: 0.001432

################### SUBSET B of RESPONSE ###################

# AIC forward data2Y
Call:
  lm(formula = response ~ KCNJ10 + PYY + FNDC4 + CXorf9 +
       B3GAT3 + MMP16 + SPINK1 + SRR + CRB1 + FGF5 +
       CRISP1 + NUDT2 + ASTE1 + RACGAP1 + APOBEC3B +
       HOXC10 + SCAMP1 + TRIM58 + ENC1 + APEX1 + P2RX1 +
       LY75 + AGTR1 + TLR5 + POLS + GSTT1, data = data2Y)

Residual standard error: 9.97 on 540 degrees of freedom
Multiple R-squared:  0.1715,  Adjusted R-squared:  0.1316
F-statistic:   4.3 on 26 and 540 DF,  p-value: 2.94e-11

# BIC forward data2Y
Call:
  lm(formula = response ~ KCNJ10, data = data2Y)
Residual standard error: 10.54 on 565 degrees of freedom
Multiple R-squared:  0.0304,  Adjusted R-squared:  0.02868
F-statistic: 17.71 on 1 and 565 DF,  p-value: 2.988e-05

# AIC backward data2Y
Call:
  lm(formula = response ~ MMP16 + SLC25A20 + AGTR1 + ODC1 +
       COL4A5 + P2RX1 + CRB1 + PITPNA + GPR56 + PTPN12 +
       PLA2G4A + TAPBPL + NUBP1 + HSD11B2 + SPINK1 + CXorf9
     + SCAMP1 + KCNJ10 + ACCN3 + PYY + FLJ10815 + RACGAP1 +
       CHSY1 + DSC2 + ENC1 + HOXC10 + C18orf1 + CRELD1 +
       YPEL1 + APOBEC3B + JARID2 + E2F6 + CADM3 + NUDT9 +
       RPL23AP13 + CCDC90A + KIAA0753 + NUDT2 + MPI + SRR +
       TRIM58 + ASTE1 + APEX1 + NDUFA5 + STON1 + B3GAT3 +
       PUS7 + POLS + FNDC4 + FLG + LY75 + CRISP1,
       data = data2Y)
```

```
Residual standard error: 9.708 on 514 degrees of freedom
Multiple R-squared:  0.2523,  Adjusted R-squared:  0.1766
F-statistic: 3.335 on 52 and 514 DF,  p-value: 2.159e-12

# BIC backward data2Y
Call:
  lm(formula = response ~ KCNJ10, data = data2Y)
Residual standard error: 10.54 on 565 degrees of freedom
Multiple R-squared:  0.0304,  Adjusted R-squared:  0.02868
F-statistic: 17.71 on 1 and 565 DF,  p-value: 2.988e-05

# AIC both data2Y
Call:
  lm(formula = response ~ MMP16 + SLC25A20 + AGTR1 + ODC1 +
       COL4A5 + P2RX1 + CRB1 + PITPNA + GPR56 + PTPN12 +
       PLA2G4A + TAPBPL + NUBP1 + HSD11B2 + SPINK1 +
       CXorf9 + SCAMP1 + KCNJ10 + ACCN3 + PYY + FLJ10815 +
       RACGAP1 + CHSY1 + DSC2 + ENC1 + HOXC10 + C18orf1 +
       CRELD1 + YPEL1 + APOBEC3B + JARID2 + E2F6 + CADM3 +
       NUDT9 + RPL23AP13 + CCDC90A + KIAA0753 + NUDT2 +
       MPI + SRR + TRIM58 + ASTE1 + APEX1 + NDUFA5 + STON1 +
       B3GAT3 + PUS7 + POLS + FNDC4 + FLG + LY75 + CRISP1,
       data = data2Y)

Residual standard error: 9.708 on 514 degrees of freedom
Multiple R-squared:  0.2523,  Adjusted R-squared:  0.1766
F-statistic: 3.335 on 52 and 514 DF,  p-value: 2.159e-12

# BIC both data2Y
Call:
  lm(formula = response ~ KCNJ10, data = data2Y)
Residual standard error: 10.54 on 565 degrees of freedom
Multiple R-squared:  0.0304,  Adjusted R-squared:  0.02868
F-statistic: 17.71 on 1 and 565 DF,  p-value: 2.988e-05

################## SUBSET A&B of RESPONSE ##################

# AIC forward alldataY
Call:
  lm(formula = response ~ KCNJ10 + PYY + FNDC4 + SETMAR +
       CRTAM + CXorf9 + B3GAT3 + SPINK1 + FAHD2A + MMP16 +
       SCAMP1 + ARHGEF2 + RAB36 + MED24 + ADAMTS5 + CRB1 +
       SRR + TRIM58 + FGF5 + NUDT2 + C3orf64 + GPR56 +
       SLC9A8 + PAEP + CYB5R2 + COX8A + JARID2 + IFNA2 +
       NRG1 + GAS2 + HIST1H2BH + ASTE1 + ANXA7 + RPL23AP13 +
       AMELY + BTBD7 + TMEM2 + PCAF + FLJ10815 + TFG + WWC2
     + HSPB2 + LGALS2 + GPR6 + MTL5 + RACGAP1 + TLR5 + TCF15
     + P2RX1 + DBP + HSD11B2 + ITIH2 + ARFIP1 + FTSJ3 +
       GEMIN8 + MPI + DCC + NEK4 + HCRTR2 + PYCR1 + SBNO2 +
       FTHP1 + KIAA0753 + LY75 + ITPR3 + HERC2 + APOBEC3B +
       UBB, data = alldataY)

Residual standard error: 10.4 on 498 degrees of freedom
Multiple R-squared:  0.3593,  Adjusted R-squared:  0.2718
F-statistic: 4.107 on 68 and 498 DF,  p-value: < 2.2e-16

# BIC froward alldataY
Call:
  lm(formula = response ~ KCNJ10, data = alldataY)
Residual standard error: 12.01 on 565 degrees of freedom
Multiple R-squared:  0.03058,  Adjusted R-squared:  0.02886
F-statistic: 17.82 on 1 and 565 DF,  p-value: 2.827e-05

# AIC backward alldataY
Call:
  lm(formula = response ~ FLG + SPINK1 + HABP2 + ORM1 +
       FNDC4 + OR7A17 + SLC17A4 + FARS2 + LRP6 + NRG1 +
       ERBB4 + MTA1 + FAM59A + C3orf64 + SLC5A1 + OSBPL8 +
       PKP4 + LILRA5 + TSPAN13 + SPG20 + ARFIP1 + CRTAM +
       FXYD6 + RCAN3 + DCC + NCAPG + WDR32 + NIF3L1 + AMOT +
       HYAL2 + RPLP0 + DDIT3 + GEMIN8 + RBM6 + FAM120A +
       MTERFD1 + TSPAN1 + AMELY + FUT5 + SS18L2 + SHANK1 +
       AHNAK + UBP1 + TSC22D4 + ADAMTS5 + TM9SF4 + FAHD2A +
       MTL5 + UBIAD1 + CA6 + PML + LGALS2 + COX8A + GRAMD1C
     + PTPN3 + HSPB2 + SLC9A8 + BTBD7 + ADAMTS3 + SPSB3 +
       SETMAR + SBNO2 + SFRS11 + ERCC6L + SMAD7 + PIK3C2B +
       PACS2 + BFSP1 + WDR57 + C5orf23 + DOCK9 + TCF15 +
       CLK4 + NPR3 + POLL + NOTCH1 + MEIS2 + IL6ST + GPR6 +
       SPRY4 + VAPB + PAEP + SNW1 + PIGG + C14orf104 +
       C21orf59 + LTB + ANXA7 + RELB + ARHGEF2 + LMO7 +
       MMP16 + SLC25A20 + AGTR1 + ODC1 + RPL8 + ANP32A +
       SLC39A9 + P2RX1 + POU2F1 + CRB1 + FOSB + GPR56 +
       CDC25A + PTPN12 + SERPINB7 + PTDSS2 + FBXW2 + GALNT1
     + SLC25A31 + ABCA3 + BIN3 + TAPBPL + AK1 + NUBP1 +
       HSD11B2 + GADD45G + ITGA9 + TAF1C + HBB + XYLT2 +
       CXorf9 + ITPR3 + ANAPC2 + FAM83E + KCNJ10 + ACCN3 +
       FGF5 + UFM1 + PUS7L + SLC16A5 + PITPNM1 + PSAT1 +
       HDGF + RACGAP1 + EXOSC8 + SNRPE + DSC2 + COPG +
       CRELD1 + YPEL1 + CROP + APOBEC3B + JARID2 + EDA +
       PCDH1 + GPBP1L1 + PRPF39 + NUDT9 + RPL23AP13 + CFHR2
     + HIST1H2BN + CCDC90A + RDBP + KIAA0753 + ASTE1 +
       TRIM38 + HD + APEX1 + HCRTR2 + NDUFA5 + ZNF518 +
       KLRB1 + NPY + PRPF4B + ATF4 + CLN5 + PUS7 + ITIH2
     + PITPNC1 + ACTL6A + POLS + HERC2 + TWF1 + LOC130074 +
       PUM1 + C17orf62 + LY75 + TINAGL1 + CRISP1,
       data = alldataY)

Residual standard error: 9.558 on 386 degrees of freedom
Multiple R-squared:  0.5808,  Adjusted R-squared:  0.3854
F-statistic: 2.972 on 180 and 386 DF,  p-value: < 2.2e-16

# BIC backward alldataY
Call:
  lm(formula = response ~ CRTAM + CXorf9 + KCNJ10, data =
       alldataY)

Residual standard error: 11.88 on 563 degrees of freedom
Multiple R-squared:  0.05578,  Adjusted R-squared:  0.05075
F-statistic: 11.09 on 3 and 563 DF,  p-value: 4.436e-07

# AIC both alldataY
Call:
  lm(formula = response ~ FLG + SPINK1 + HABP2 + FNDC4 +
       OR7A17 + SLC17A4 + FARS2 + LRP6 + NRG1 + ERBB4 +
       FAM59A + C3orf64 + SLC5A1 + OSBPL8 + PKP4 + TSPAN13
     + SPG20 + ARFIP1 + CRTAM + FXYD6 + RCAN3 + DCC +
       NCAPG + WDR32 + NIF3L1 + NOX5 + HYAL2 + RPLP0 +
       DDIT3 + GEMIN8 + CDH22 + RBM6 + FAM120A + MTERFD1 +
       TSPAN1 + AMELY + FUT5 + SS18L2 + SHANK1 + AHNAK +
       UBP1 + TSC22D4 + ADAMTS5 + TM9SF4 + ACVRL1 + GLI3 +
       FAHD2A + MTL5 + PML + LGALS2 + COX8A + GRAMD1C +
       PTPN3 + HSPB2 + SLC9A8 + BTBD7 + ADAMTS3 + SPSB3 +
       SETMAR + SBNO2 + SFRS11 + ERCC6L + SMAD7 + PIK3C2B +
       BFSP1 + WDR57 + C5orf23 + DOCK9 + CLK4 + NPR3 +
       POLL + NOTCH1 + OGT + MEIS2 + IL6ST + GPR6 + SPRY4 +
       VAPB + PAEP + SNW1 + DNMBP + PIGG + C14orf104 +
       C21orf59 + LTB + ANXA7 + RELB + ARHGEF2 + LMO7 +
       MMP16 + SLC25A20 + AGTR1 + ODC1 + RPL8 + ANP32A +
       SLC39A9 + P2RX1 + POU2F1 + CRB1 + FOSB + GPR56 +
       CDC25A + PTPN12 + PTDSS2 + FBXW2 + GALNT1 + SLC25A31
     + ABCA3 + BIN3 + TAPBPL + AK1 + NUBP1 + HSD11B2 +
       GADD45G + ITGA9 + TAF1C + HBB + XYLT2 + CXorf9 +
       ITPR3 + ANAPC2 + FAM83E + KCNJ10 + ACCN3 + FGF5 +
       UFM1 + PUS7L + FLJ10815 + SLC16A5 + PITPNM1 +
       PSAT1 + HDGF + RACGAP1 + EXOSC8 + SNRPE + DSC2 +
       COPG + CRELD1 + YPEL1 + CROP + APOBEC3B + JARID2 +
       EDA + PCDH1 + GPBP1L1 + PRPF39 + NUDT9 + RPL23AP13 +
       CFHR2 + HIST1H2BN + CCDC90A + RDBP + KIAA0753 +
       ASTE1 + TRIM38 + HD + APEX1 + HCRTR2 + NDUFA5 +
       ZNF518 + KLRB1 + NPY + ATF4 + CLN5 + PUS7 + ITIH2 +
       PITPNC1 + ACTL6A + POLS + HERC2 + TWF1 + LOC130074 +
       PUM1 + C17orf62 + LY75 + TINAGL1 + S100A11 + C18orf1
     + HIST1H2BH + IFNA2 + KIF1C + FTHP1 + PYY + NUDT2 +
       LILRA5, data = alldataY)

Residual standard error: 9.483 on 381 degrees of freedom
Multiple R-squared:  0.5927,  Adjusted R-squared:  0.3949
F-statistic: 2.997 on 185 and 381 DF,  p-value: < 2.2e-16

# BIC both alldataY
Call:
  lm(formula = response ~ FNDC4 + CRTAM + KCNJ10 + CXorf9,
       data = alldataY)

Residual standard error: 11.81 on 562 degrees of freedom
Multiple R-squared:  0.0677,  Adjusted R-squared:  0.06106
F-statistic:  10.2 on 4 and 562 DF,  p-value: 5.588e-08

#################### SUBSET A of TP53 ####################

# AIC forward data1TP
Call:
  lm(formula = TP53 ~ TM9SF4 + ATP5G2 + POLL + LOC390688 +
       TSPAN13 + PCAF + RFXAP + CRTAM + PYY + NR5A2 +
       RAB36 + MTL5 + ACVRL1 + RBMX + RELB + CLK4 + GYPB +
       CTNNBIP1 + IL6ST + PKP4 + AMOT + C20orf12 + GNAZ +
       PTPN2 + POP5 + C9orf78 + RPL7A + SLC9A8 + RHOB +
       RPLP0 + PALMD + GPR126 + CDS2 + NPFFR1 + HSPB2 +
```

```
      COL5A1 + RCAN3 + FTHP1 + FARS2 + TMEM2 + SLIT3 +
      RHOBTB3 + VAPB + BTBD7 + MTA1 + GAS2 + GPR6 +
      SLC17A4 + CCL15 + AMELY + SPSB3 + NIF3L1 +
      KIAA0143 + HSPA12A + DBP + HYAL2 + MFAP2 + CYB5R2 +
      PCYT2 + RBM6 + MED24 + CHD1, data = data1TP)

Residual standard error: 0.8673 on 504 degrees of freedom
Multiple R-squared:  0.3969,  Adjusted R-squared:  0.3227
F-statistic:  5.35 on 62 and 504 DF,  p-value: < 2.2e-16

# BIC forward data1TP
Call:
  lm(formula = TP53 ~ TM9SF4 + ATP5G2 + POLL + LOC390688 +
      TSPAN13 + PCAF + RFXAP + CRTAM, data = data1TP)

Residual standard error: 0.9816 on 558 degrees of freedom
Multiple R-squared:  0.1447,  Adjusted R-squared:  0.1324
F-statistic:  11.8 on 8 and 558 DF,  p-value: 1.38e-15

# AIC backward data1TP
Call:
  lm(formula = TP53 ~ NOC4L + PALMD + OR7A17 + SLC17A4 +
      TMEM2 + C20orf12 + GYPB + PKP4 + SNX6 + TSPAN13 +
      FTHP1 + SPG20 + GNAZ + FLG + CRTAM + RCAN3 + SLIT3
    + RBMX + PCAF + WDR32 + NIF3L1 + HSPA12A + HYAL2 +
      BTC + RBM6 + PCYT2 + MFAP2 + KIAA0947 + RHOBTB3 +
      RAB36 + SS18L2 + UBP1 + ATP5G2 + TM9SF4 + ACVRL1 +
      RPL7A + IARS + MTL5 + COL5A1 + OGDHL + CTNNBIP1 +
      ANKRD15 + HSPB2 + SPSB3 + ATP8B3 + POP5 + SETMAR +
      PYY + ERCC6L + LOC390688 + GPR126 + DOCK9 + NPFFR1 +
      TCF15 + CDS2 + CEL + CLK4 + C9orf78 + POLL + NOTCH1
    + OGT + IL6ST + GPR6 + VAPB + ZNF391 + MED24 +
      C14orf104 + LTB + ANXA7 + CYB5R2, data = data1TP)

Residual standard error: 0.8619 on 496 degrees of freedom
Multiple R-squared:  0.4138,  Adjusted R-squared:  0.331
F-statistic: 5.001 on 70 and 496 DF,  p-value: < 2.2e-16

# BIC backward data1TP
Call:
  lm(formula = TP53 ~ PALMD + GYPB + CRTAM + SLIT3 + PCAF +
      RHOBTB3 + RAB36 + ATP5G2 + TM9SF4 + RPL7A + HSPB2 +
      PYY + LOC390688 + GPR126 + C9orf78 + POLL + VAPB,
      data = data1TP)

Residual standard error: 0.9444 on 549 degrees of freedom
Multiple R-squared:  0.2209,  Adjusted R-squared:  0.1968
F-statistic: 9.157 on 17 and 549 DF,  p-value: < 2.2e-16

# AIC both data1TP
Call:
  lm(formula = TP53 ~ NOC4L + PALMD + OR7A17 + SLC17A4 +
      TMEM2 + C20orf12 + GYPB + PKP4 + SNX6 + FTHP1 +
      SPG20 + GNAZ + CRTAM + RCAN3 + SLIT3 + PCAF +
      NIF3L1 + HSPA12A + HYAL2 + BTC + RBM6 + PCYT2 +
      MFAP2 + KIAA0947 + RHOBTB3 + RAB36 + SS18L2 + UBP1 +
      ATP5G2 + TM9SF4 + ACVRL1 + RPL7A + IARS + MTL5 +
      COL5A1 + OGDHL + CTNNBIP1 + ANKRD15 + HSPB2 + SPSB3
    + POP5 + SETMAR + PYY + ERCC6L + LOC390688 + GPR126 +
      DOCK9 + NPFFR1 + TCF15 + CDS2 + CEL + CLK4 + C9orf78
    + POLL + NOTCH1 + IL6ST + GPR6 + VAPB + ZNF391 + MED24
    + C14orf104 + ANXA7 + CYB5R2 + MTA1 + SLC9A8,
      data = data1TP)

Residual standard error: 0.8623 on 501 degrees of freedom
Multiple R-squared:  0.4073,  Adjusted R-squared:  0.3304
F-statistic: 5.297 on 65 and 501 DF,  p-value: < 2.2e-16

# BIC both data1TP
Call:
  lm(formula = TP53 ~ PALMD + GYPB + CRTAM + SLIT3 + PCAF +
      RHOBTB3 + RAB36 + ATP5G2 + TM9SF4 + RPL7A + HSPB2 +
      PYY + LOC390688 + GPR126 + C9orf78 + POLL + VAPB,
      data = data1TP)

Residual standard error: 0.9444 on 549 degrees of freedom
Multiple R-squared:  0.2209,  Adjusted R-squared:  0.1968
F-statistic: 9.157 on 17 and 549 DF,  p-value: < 2.2e-16

##################### SUBSET A of TP53 ######################

# AIC forward data2TP
Call:
  lm(formula = TP53 ~ KIAA0753 + KIF5C + HERC2 + ATP10A +
      KNTC1 + NUBP1 + MIF + PYY + RNF144A + CHSY1 +
      BRWD1 + ANKRD46 + NDUFA5 + ITIH2 + MAP7D3 + R3HDM1
    + FOSB + XPA + POU2F1 + VWA1 + AMIGO2 + SNRPE + MMP16
    + TWF1 + CROP + GADD45G + SCAMP1 + TWSG1 + BAG5 +
      RUNDC3B + LIF + JARID1A + SLFN12 + HLA.G + ITPR3 +
      MC3R + BMP4, data = data2TP)

Residual standard error: 0.8683 on 529 degrees of freedom
Multiple R-squared:  0.3655,  Adjusted R-squared:  0.3211
F-statistic: 8.236 on 37 and 529 DF,  p-value: < 2.2e-16

# BIC forward data2TP
Call:
  lm(formula = TP53 ~ KIAA0753 + KIF5C + HERC2 + ATP10A +
      KNTC1 + NUBP1 + MIF, data = data2TP)

Residual standard error: 0.9594 on 559 degrees of freedom
Multiple R-squared:  0.1815,  Adjusted R-squared:  0.1712
F-statistic:  17.7 on 7 and 559 DF,  p-value: < 2.2e-16

# AIC backward data2TP
Call:
  lm(formula = TP53 ~ AGTR1 + ODC1 + RNF144A + IFNA2 +
      SLC39A9 + POU2F1 + FOSB + GPR56 + CDC25A +
      ANKRD46 + PTPN12 + MIF + TWSG1 + TROAP + PLA2G4A +
      NUBP1 + ISG20 + ITGA9 + SPINK1 + XYLT2 + CXorf9 +
      ITPR3 + KCNJ10 + PYY + ZNF587 + JARID1A + RUNDC3B
    + UFM1 + PUS7L + FLJ10815 + ALX3 + TBL1X + IFNAR2 +
      PITPNM1 + VWA1 + CHSY1 + HLA.G + SNRPE + ENC1 +
      HOXC10 + LIF + BRWD1 + YPEL1 + BMP4 + AMIGO2 +
      MAP7D3 + CDKN2A + SLFN12 + R3HDM1 + MC3R + E2F6 +
      ANXA6 + NUDT9 + WDR79 + KNTC1 + KIAA0753 + KIAA0020
    + MPI + KIF5C + NDUFA5 + GCAT + ATP10A + ZNF518 +
      B3GAT3 + KLRB1 + KBTBD4 + FBXO3 + SRP72 + BAG5 +
      FLNB + ITIH2 + POLS + KIF1C + FNDC4 + HERC2 + GLDC
    + PPAT + PUM1 + E2F5 + PLXNA3 + MBIP, data = data2TP)

Residual standard error: 0.8428 on 485 degrees of freedom
Multiple R-squared:  0.452,  Adjusted R-squared:  0.3604
F-statistic: 4.938 on 81 and 485 DF,  p-value: < 2.2e-16

# BIC backward data2TP
Call:
  lm(formula = TP53 ~ ODC1 + FOSB + ANKRD46 + NUBP1 + PYY +
      VWA1 + CHSY1 + SNRPE + BRWD1 + AMIGO2 + MAP7D3 +
      KNTC1 + KIAA0753 + KIF5C + NDUFA5 + ATP10A + ITIH2 +
      HERC2, data = data2TP)

Residual standard error: 0.9098 on 548 degrees of freedom
Multiple R-squared:  0.2784,  Adjusted R-squared:  0.2547
F-statistic: 11.74 on 18 and 548 DF,  p-value: < 2.2e-16

# AIC both data2TP
Call:
  lm(formula = TP53 ~ AGTR1 + ODC1 + RNF144A + IFNA2 +
      SLC39A9 + POU2F1 + FOSB + GPR56 + CDC25A + ANKRD46 +
      PTPN12 + MIF + TWSG1 + TROAP + PLA2G4A + NUBP1 +
      ISG20 + ITGA9 + SPINK1 + XYLT2 + CXorf9 + ITPR3 +
      KCNJ10 + PYY + ZNF587 + JARID1A + RUNDC3B + UFM1 +
      PUS7L + FLJ10815 + ALX3 + TBL1X + IFNAR2 + PITPNM1 +
      VWA1 + CHSY1 + HLA.G + SNRPE + ENC1 + HOXC10 + LIF +
      BRWD1 + YPEL1 + BMP4 + AMIGO2 + MAP7D3 + CDKN2A +
      SLFN12 + R3HDM1 + MC3R + E2F6 + ANXA6 + NUDT9 + WDR79
    + KNTC1 + KIAA0753 + KIAA0020 + MPI + KIF5C + NDUFA5 +
      GCAT + ATP10A + ZNF518 + B3GAT3 + KLRB1 + KBTBD4 +
      FBXO3 + SRP72 + BAG5 + FLNB + ITIH2 + POLS + KIF1C +
      FNDC4 + HERC2 + GLDC + PPAT + PUM1 + E2F5 + PLXNA3 +
      MBIP, data = data2TP)

Residual standard error: 0.8428 on 485 degrees of freedom
Multiple R-squared:  0.452,  Adjusted R-squared:  0.3604
F-statistic: 4.938 on 81 and 485 DF,  p-value: < 2.2e-16

# BIC both data2TP
Call:
  lm(formula = TP53 ~ ODC1 + FOSB + ANKRD46 + NUBP1 + PYY +
      VWA1 + CHSY1 + SNRPE + BRWD1 + AMIGO2 + MAP7D3 +
      KNTC1 + KIAA0753 + KIF5C + NDUFA5 + ATP10A + ITIH2 +
      HERC2, data = data2TP)

Residual standard error: 0.9098 on 548 degrees of freedom
Multiple R-squared:  0.2784,  Adjusted R-squared:  0.2547
F-statistic: 11.74 on 18 and 548 DF,  p-value: < 2.2e-16
```

```
#################### SUBSET A&B of TP53 ####################

# AIC forward alldataTP
Call:
  lm(formula = TP53 ~ KIAA0753 + SFRS11 + HERC2 + ATP10A +
       KIF5C + HSPB2 + C9orf78 + PYY + SLIT3 + PCAF + NUBP1
     + CDC25A + UBP1 + MAP7D3 + MIF + MAPKAP1 + SS18L2 +
       SLC9A8 + CRTAM + R3HDM1 + CHSY1 + MFAP2 + ITIH2 +
       MTL5 + CYB5R2 + LIF + CTNNBIP1 + ANKRD46 + POP5 +
       TROAP + ZDHHC3 + SLC17A4 + HYAL2 + CRELD1 + PFKFB2 +
       GCAT + NIF3L1 + ACVRL1 + FLJ10815 + MTERFD1 + FOSB +
       SLFN12 + C14orf104 + POLL + AQP5 + CLK4 + RNF40 +
       ODC1 + NSMAF + FAM83E + FRMD4B + RPL7A + GYPB +
       TBL1X + SNX6 + ANXA7 + LGALS2 + OR7A17 + PTPN12 +
       KBTBD4 + FBXO3 + HCRTR2 + CXorf9 + E2F6 + XYLT2 +
       RELB + TCF15 + CDKN2A + OR1G1 + LBA1 + PCYT2 +
       HLA.G + DIP2C + PLXNA3 + OSBPL8 + JARID1A + MBIP +
       ZNF391 + XPA + RPL23AP13 + COX8A + LRP6 + KCNJ10 +
       WDR79 + TLR5 + MED24 + PSAT1 + CA9 + WDR57 + NDUFA5
     + YPEL1 + SLC25A20 + TWSG1 + RNF44 + OGDHL + BTC +
       BET1L + TRIO + CPT1B + AK1 + CDS2 + MTA1 + GNAZ +
       RHOB + PITPNM1 + LMO7, data = alldataTP)

Residual standard error: 0.7271 on 460 degrees of freedom
Multiple R-squared:  0.6131,  Adjusted R-squared:  0.5239
F-statistic: 6.876 on 106 and 460 DF,  p-value: < 2.2e-16

# BIC forward alldataTP
Call:
  lm(formula = TP53 ~ KIAA0753 + SFRS11 + HERC2 + ATP10A +
       KIF5C + HSPB2 + C9orf78 + PYY + SLIT3 + PCAF +
       NUBP1 + CDC25A + UBP1 + MAP7D3 + MIF + MAPKAP1 +
       SS18L2 + SLC9A8 + CRTAM + R3HDM1 + CHSY1 + MFAP2 +
       ITIH2, data = alldataTP)

Residual standard error: 0.8672 on 543 degrees of freedom
Multiple R-squared:  0.3503,  Adjusted R-squared:  0.3228
F-statistic: 12.73 on 23 and 543 DF,  p-value: < 2.2e-16

# AIC backward alldataTP
Call:
  lm(formula = TP53 ~ TRIM36 + HOXC10 + PYY + CADM3 + OR7A17
     + SLC17A4 + TMEM2 + LRP6 + GYPB + CENPQ + MTA1 + FAM59A
     + PTH2R + RAI16 + TSPAN13 + GNAZ + NUCKS1 + CRTAM +
       RCAN3 + SLIT3 + NCAPG + ZBTB33 + RBMX + PCAF + WDR32
     + NIF3L1 + NOX5 + HYAL2 + TGM1 + DDIT3 + SNTB2 + DBP +
       BTC + SURF2 + FAM120A + PCYT2 + FOXN2 + MFAP2 + HINT1
     + FRMD4B + TSC22D4 + ACVRL1 + RPL7A + C8orf33 + COL5A1
     + CA6 + PML + LGALS2 + CTNNBIP1 + PTPN3 + ANKRD15 +
       HSPB2 + SLC9A8 + ADAMTS3 + RHOB + POP5 + ESRRG +
       AHNAK2 + PACS2 + LOC390688 + WDR57 + DOCK9 + ATP6V0E2
     + TCF15 + CDS2 + CLK4 + C9orf78 + POLL + NOTCH1 + MEIS2
     + LBA1 + FASTKD1 + KLRF1 + TGFBRAP1 + SNW1 + PARC +
       DNMBP + MED24 + PIGG + C14orf104 + C21orf59 + ANXA7
     + UBB + RELB + LMO7 + P2RX1 + ZNF343 + NSMAF + FOSB +
       GPR56 + POU6F1 + ANKRD46 + PTPN12 + MIF + MYBPH +
       GALNT1 + CCT6A + TSPYL4 + TROAP
     + TARDBP + BIN3 + AK1 + RNF44 + TAF1C + HBB + GH1 +
       CXorf9 + MRE11A + COL4A2 + ITPR3 + ANAPC2 + SCAMP1 +
       FAM83E + TFG + JARID1A + FLJ10815 + TBL1X + PITPNM1 +
       PSAT1 + VWA1 + HDGF + CHSY1 + CPT1B + DSC2 + SLC25A13
     + CRELD1 + LIF + YPEL1 + MAP7D3 + CD6 + CDKN2A + EDA +
       MC3R + E2F6 + PCDH1 + NUDT9 + WDR79 + HIST1H2BN +
       CLSTN1 + KIAA0753 + OR1G1 + NUDT2 + KIF5C + HD +
       HCRTR2 + NDUFA5 + GCAT + ATP10A + ZNF518 + STON1 +
       KLRB1 + KBTBD4 + FBXO3 + XPA + BAG5 + FLNB + GPR27 +
       PUS7 + HERC2 + GLDC + E2F5 + C17orf62 + MBIP,
       data = alldataTP)

Residual standard error: 0.6947 on 403 degrees of freedom
Multiple R-squared:  0.6906,  Adjusted R-squared:  0.5655
F-statistic: 5.519 on 163 and 403 DF,  p-value: < 2.2e-16

# BIC backward alldataTP
Call:
  lm(formula = TP53 ~ PYY + CRTAM + SLIT3 + PCAF + SURF2
     + MFAP2 + CTNNBIP1 + HSPB2 + SLC9A8 + CLK4 + C9orf78
     + POLL + LBA1 + C14orf104 + RELB + ANKRD46 + MIF +
       CHSY1 + MAP7D3 + CDKN2A + KIAA0753 + KIF5C + ATP10A
     + XPA + HERC2, data = alldataTP)

Residual standard error: 0.8708 on 541 degrees of freedom
Multiple R-squared:  0.3473,  Adjusted R-squared:  0.3171
F-statistic: 11.51 on 25 and 541 DF,  p-value: < 2.2e-16

# AIC both alldataTP
Call:
  lm(formula = TP53 ~ HOXC10 + PYY + CADM3 + OR7A17 +
       SLC17A4 + TMEM2 + LRP6 + GYPB + CENPQ + MTA1 +
       FAM59A + GRB2 + PTH2R + RAI16 + TSPAN13 + GNAZ +
       NUCKS1 + CRTAM + RCAN3 + SLIT3 + NCAPG + ZBTB33
     + RBMX + PCAF + WDR32 + NIF3L1 + NOX5 + HYAL2 +
       TGM1 + DDIT3 + DBP + BTC + SURF2 + FAM120A +
       PCYT2 + FOXN2 + MFAP2 + HINT1 + FRMD4B + TSC22D4
     + ACVRL1 + C8orf33 + COL5A1 + CA6 + LGALS2 +
       CTNNBIP1 + PTPN3 + ANKRD15 + HSPB2 + SLC9A8 +
       RHOB + POP5 + ESRRG + PACS2 + LOC390688 + WDR57
     + DIP2C + DOCK9 + ATP6V0E2 + TCF15 + CDS2 + CLK4 +
       C9orf78 + POLL + NOTCH1 + MEIS2 + LBA1 + FASTKD1
     + KLRF1 + TGFBRAP1 + SNW1 + PARC + DNMBP + MED24 +
       C14orf104 + C21orf59 + ANXA7 + UBB + RELB + LMO7
     + P2RX1 + ZNF343 + NSMAF + FOSB + GPR56 + POU6F1 +
       ANKRD46 + PTPN12 + MIF + MYBPH + GALNT1 + CCT6A +
       TSPYL4 + TROAP + TARDBP + BIN3 + AK1 + RNF44 +
       TAF1C + HBB + GH1 + CXorf9 + MRE11A + COL4A2 +
       ITPR3 + ANAPC2 + SCAMP1 + FAM83E + TFG + JARID1A
     + FLJ10815 + SLC16A5 + TBL1X + PITPNM1 + PSAT1 +
       VWA1 + CHSY1 + CPT1B + DSC2 + SLC25A13 + CRELD1
     + LIF + YPEL1 + MAP7D3 + CDKN2A + EDA + MC3R +
       PCDH1 + NUDT9 + WDR79 + HIST1H2BN + CLSTN1 +
       KIAA0753 + OR1G1 + NUDT2 + KIF5C + HD + HCRTR2 +
       NDUFA5 + GCAT + ATP10A + ZNF518 + STON1 + KLRB1 +
       KBTBD4 + FBXO3 + XPA + BAG5 + FLNB + GPR27 +
       PUS7 + HERC2 + GLDC + E2F5 + C17orf62 + TWF1 +
       UFM1 + FAHD2A + COPG + SERPINB7 + FARS2 + TINAGL1,
       data = alldataTP)

Residual standard error: 0.6913 on 404 degrees of freedom
Multiple R-squared:  0.6928,  Adjusted R-squared:  0.5697
F-statistic: 5.625 on 162 and 404 DF,  p-value: < 2.2e-16

# BIC both alldataTP
Call:
  lm(formula = TP53 ~ PYY + CRTAM + SLIT3 + PCAF + NIF3L1 +
       SURF2 + HSPB2 + SLC9A8 + POP5 + LOC390688 + TCF15 +
       CLK4 + C9orf78 + POLL + LBA1 + MED24 + RELB + FOSB +
       ANKRD46 + MIF + FLJ10815 + CHSY1 + MAP7D3 + CDKN2A +
       KIF5C + GCAT + ATP10A + HERC2 + WDR79 + ACVRL1 +
       TM9SF4 + RNF144A, data = alldataTP)

Residual standard error: 0.8306 on 534 degrees of freedom
Multiple R-squared:  0.4139,  Adjusted R-squared:  0.3787
F-statistic: 11.78 on 32 and 534 DF,  p-value: < 2.2e-16

#################### CROSS VALIDATION ####################

result1
#[1] 132 143 132 143 132 143
result2
#[1] 107 112 109 112 109 112
result3
#[1] 130 145 148 142 147 141
result4
#[1] 0.863 0.968 0.867 0.917 0.870 0.917
result5
#[1] 0.821 0.931 0.877 0.846 0.877 0.846
result6
#[1] 0.687 0.797 0.730 0.818 0.745 0.760


# ########################
# #        LASSO         #
# ########################
#
################### SUBSET A of RESPONSE ###################
# response and Box-Cox(response)
num<-which.min(la2$Cp)
coef(la1)[num,][coef(la1)[num,]!=0]
#SPINK1    PYY     CEL
#30.74    56.13    6.73
num<-which.min(cv.lasso.1$cv)
coef(la1)[num,][coef(la1)[num,]!=0]
#named numeric(0)

num<-which.min(la1.trans$Cp)
coef(la1.trans)[num,][coef(la1.trans)[num,]!=0]
#named numeric(0)
num<-which.min(cv.lasso.1.trans$cv)
```

```
coef(la1.trans)[num,][coef(la1.trans)[num,]!=0]
#named numeric(0)

################### SUBSET B of RESPONSE ###################

num<-which.min(la2$Cp)
coef(la2)[num,][coef(la2)[num,]!=0]
#MMP16  HSD11B2  SPINK1  KCNJ10  PYY   CEL  ITIH2
#-291.8 36.5     22.2    623.2   59.0  18.1 26.3
num<-which.min(cv.lasso.2$cv)
coef(la2)[num,][coef(la2)[num,]!=0]
#KCNJ10
#274

num<-which.min(la2.trans$Cp)
coef(la2.trans)[num,][coef(la2.trans)[num,]!=0]
#MMP16  CRB1  TNFRSF11B  HSD11B2  SPINK1  CXorf9  KCNJ10
#-2.46  0.33  0.09       0.40     0.28    -0.53   9.39
# PYY   RACGAP1  CEL  B3GAT3  FNDC4  FLG
#0.70   -0.22    0.18 0.74    -1.55  -0.16
num<-which.min(cv.lasso.2.trans$cv)
coef(la2.trans)[num,][coef(la2.trans)[num,]!=0]
#KCNJ10
#5.55

################### SUBSET A&B of RESPONSE ###################

num<-which.min(laall$Cp)
coef(laall)[num,][coef(laall)[num,]!=0]
#SPINK1  PYY   CEL   CRTAM   FAHD2A  MMP16  HSD11B2  KCNJ10
#22.13  59.38 19.17  12.22   -9.61  -301.19 39.31    623.30
#ITIH2
#28.59

num<-which.min(cv.lasso.all$cv)
coef(laall)[num,][coef(laall)[num,]!=0]
#named numeric(0)

num<-which.min(laall.trans$Cp)
coef(laall.trans)[num,][coef(laall.trans)[num,]!=0]
#KCNJ10
#6.3
num<-which.min(cv.lasso.all.trans$cv)
coef(laall.trans)[num,][coef(laall.trans)[num,]!=0]
#KCNJ10
#6.3

#################### SUBSET A of TP53 ####################

num<-which.min(summary(la1.TP)$Cp)
sum(coef(la1.TP)[num,]!=0)
#84 --- too many
num<-which.min(cv.lasso.1$cv)
coef(la1.TP)[num,][coef(la1.TP)[num,]!=0]
#FARS2  C20orf12  RBMX  NR5A2  RPLP0  RHOBTB3  RAB36
#0.04   0.004     0.04  -0.02  0.02   0.005    0.02
#ATP5G2  TM9SF4  RNF40  LOC390688  NPFFR1  C9orf78  POLL
#0.11    0.12    0.005  -0.15      -0.04   0.001    0.16
#DNMBP  ARHGEF2
#0.003  0.06

#################### SUBSET B of TP53 ####################

num<-which.min(summary(la2.TP)$Cp)
coef(la2.TP)[num,][coef(la2.TP)[num,]!=0]
#MMP16  RNF144A  ORM1  POU2F1  FOSB  CDC25A  ANKRD46
#-0.23  0.06     -0.04 0.06    -0.04 -0.14   0.08
#PTPN12  MIF  MYBPH  NUBP1  HSD11B2  ITPR3  SCAMP1
#-0.02   0.09 -0.13  0.09   0.03     0.03   0.06
#PYY    RUNDC3B  VWA1   CHSY1  SNRPE  DSC2  CA9
#-0.10  -0.04    -0.06  -0.10  -0.05  0.01  0.01
#HOXC10  CRELD1  BRWD1  CROP  MAP7D3  CDKN2A  JARID2
#-0.02   0.02    -0.14  -0.02 0.06    -0.03   0.03
#SLFN12  R3HDM1  CEL  WDR79  KNTC1  KIAA0753  KIF5C
#0.02  -0.16   -0.02  0.18  -0.08  0.22   -0.08
#APEX1  NDUFA5  ATP10A  B3GAT3  KLRB1  KBTBD4  XPA
#0.05  -0.12   -0.28   -0.02   0.02   0.03    0.08
#BAG5  FLNB  ITIH2  POLS  KIF1C  HERC2  TWF1
#0.04  0.05  -0.07  -0.004  0.10   0.18   -0.03
#GSTT1  E2F5
#0.02   0.07

num<-which.min(cv.lasso.2.TP$cv)
coef(la2.TP)[num,][coef(la2.TP)[num,]!=0]
#PITPNA  ANKRD46  MIF  NUBP1  SCAMP1  PYY
#0.02    0.03     0.02 0.01   0.04    -0.01
#CRELD1  JARID2  WDR79  KIAA0753  KIF5C  APEX1
#0.01    0.0005  0.15   0.17      -0.04  0.05
#ATP10A  KBTBD4  KIF1C  HERC2  E2F5
#-0.10   0.04    0.08   0.09   0.05

################### SUBSET A&B of TP53 ###################

num<-which.min(summary(laall.TP)$Cp)
coef(la2.TP)[num,][coef(la2.TP)[num,]!=0]
#MMP16  AGTR1  ODC1  RNF144A  ORM1  SLC39A9  POU2F1
#-0.31  -0.03  0.03  0.09  -0.06  0.03  0.23
#ZNF343  NSMAF  FOSB   CDC25A  ANKRD46  PTPN12  MIF
#0.06    0.005  -0.06  -0.19   0.11     -0.11   0.14
#MYBPH  TWSG1  TSPYL4  PLA2G4A  NUBP1  HSD11B2  GADD45G
#-0.19  0.04   0.01    -0.01    0.10   0.06     -0.04
#ITGA9  SPINK1  XYLT2  ITPR3  SCAMP1  KCNJ10  PYY
#0.02   0.01    -0.08  0.10   0.08    -0.11   -0.13
#ZNF587  JARID1A  RUNDC3B  PUS7L  FLJ10815  SLC16A5  IFNAR2
#-0.02   -0.07    -0.09    -0.03  -0.05     0.003    0.05
#PITPNM1  VWA1   CHSY1  HLA.G  SNRPE  DSC2  CA9
#-0.08    -0.12  -0.18  -0.03  -0.16  0.01  0.03
#HOXC10  COPG  CRELD1  LIF   BRWD1  YPEL1  BMP4
#-0.04   0.02  0.03   -0.04 -0.21  0.04   -0.05
#AMIGO2  CROP  MAP7D3  CDKN2A  SLFN12  BET1L  R3HDM1
#0.02   -0.04  0.11    -0.04   0.07    -0.02  -0.21
#ZDHHC3  MC3R  CEL   ANXA6  PRPF39  NUDT9  WDR79
#0.02    -0.14 -0.01 0.02   -0.02   -0.001 0.25
#KNTC1  KIAA0753  TLR5  CLCN6  NUDT2  KIF5C  ASTE1
#-0.13  0.24      -0.01 -0.03  -0.02  -0.10  -0.03
#APEX1  NDUFA5  GCAT  ATP10A  ZNF518  B3GAT3  KLRB1
#0.04  -0.17    -0.03 -0.32   0.05    -0.06   0.04
#KBTBD4  NPY  FBXO3  XPA   BAG5  FLNB  ITIH2
#0.04  -0.02 -0.01  0.09  0.07  0.07  -0.13
#POLS   KIF1C  FNDC4  HERC2  TWF1  GLDC  GSTT1
#-0.04  0.09   0.01   0.22   -0.07 -0.02 0.01
#E2F5  PLXNA3  MBIP
#0.09  0.01    -0.03

num<-which.min(cv.lasso.all.TP$cv)
coef(laall.TP)[num,][coef(laall.TP)[num,]!=0]
#ATP5G2  TM9SF4  NPFFR1  POLL  WDR79  KIAA0753  HERC2
# 0.08    0.05   -0.05   0.04  0.10   0.18      0.01

# ########################
# #      POLYNOMIAL      #
# ########################
#
################### SUBSET A of RESPONSE ###################
# BIC forward data1Y
Call:
  lm(formula = response ~ HOXC10_2 + LGALS2_2 + TRIM36_2 +
       FAHD2A_1, data = data1Y.poly)

Residual standard error: 11.7 on 562 degrees of freedom
Multiple R-squared:  0.0575,  Adjusted R-squared:  0.0508
F-statistic: 8.57 on 4 and 562 DF,  p-value: 1.03e-06

# BIC backward data1Y
Call:
  lm(formula = response ~ SPINK1_1 + HOXC10_2 + HABP2_2 +
       SBNO2_1 + SPRY4_1 + FNDC4_1 + ANXA7_2,
       data = data1Y.poly)

Residual standard error: 11.6 on 559 degrees of freedom
Multiple R-squared:  0.0789,  Adjusted R-squared:  0.0674
F-statistic: 6.84 on 7 and 559 DF,  p-value: 8.19e-08

# BIC both data1Y
Call:
  lm(formula = response ~ SPINK1_1 + HOXC10_2 + HABP2_2 +
       SBNO2_1 + SPRY4_1 + FNDC4_1 + ANXA7_2 + CRTAM_1,
       data = data1Y.poly)

Residual standard error: 11.6 on 558 degrees of freedom
Multiple R-squared:  0.0911,  Adjusted R-squared:  0.0781
F-statistic: 6.99 on 8 and 558 DF,  p-value: 8.26e-09

################### SUBSET B of RESPONSE ###################

#BIC forward data2Y
Call:
  lm(formula = response ~ KCNJ10_1 + HOXC10_2 + IFNA2_2 +
```

```
      CEL_1, data = data2Y.poly)

Residual standard error: 10.3 on 562 degrees of freedom
Multiple R-squared:  0.0719,  Adjusted R-squared:  0.0653
F-statistic: 10.9 on 4 and 562 DF,  p-value: 1.67e-08

#BIC backward data2Y
Call:
  lm(formula = response ~ IFNA2_2 + HABP2_2 + KCNJ10_1 +
      HOXC10_2 + HCRTR2_2, data = data2Y.poly)
Residual standard error: 10.3 on 561 degrees of freedom
Multiple R-squared:  0.0801,  Adjusted R-squared:  0.0719
F-statistic: 9.76 on 5 and 561 DF,  p-value: 5.87e-09

#BIC both data2Y
Call:
  lm(formula = response ~ IFNA2_2 + HABP2_2 + KCNJ10_1 +
      HOXC10_2 + HCRTR2_2, data = data2Y.poly)

Residual standard error: 10.3 on 561 degrees of freedom
Multiple R-squared:  0.0801,  Adjusted R-squared:  0.0719
F-statistic: 9.76 on 5 and 561 DF,  p-value: 5.87e-09

#################### SUBSET A of TP53 ####################
# BIC forwarrd data1TP
Call:
  lm(formula = TP53 ~ TM9SF4_1 + ATP5G2_1 + DOCK9_2 +
      PTH2R_2 + RELB_1 + RAI16_2 + FAHD2A_2 + RHOBTB3_1
    + TSPAN13_1 + C8orf33_2 + PYY_1 + SLIT3_1 + GPR126_2
    + HSPB2_1 + FTHP1_1 + SLC5A1_2 + LOC390688_1 +
      GYPB_1 + PCAF_1 + TRIM36_2 + RFXAP_1,
      data = data1TP.poly)

Residual standard error: 0.899 on 545 degrees of freedom
Multiple R-squared:  0.299,  Adjusted R-squared:  0.272
F-statistic: 11.1 on 21 and 545 DF,  p-value: <2e-16

# BIC backward data1TP
Call:
  lm(formula = TP53 ~ GYPB_1 + MTA1_1 + FAM59A_2 + SLC5A1_2
    + RAI16_2 + SNX6_1 + GNAZ_1 + CYP1B1_2 + PCAF_1 +
      MAPKAP1_1 + GAS2_1 + RHOBTB3_1 + RAB36_1 + SS18L2_1 +
      SS18L2_2 + UBP1_2 + TM9SF4_1 + FAHD2A_2 + COL5A1_1 +
      COX8A_2 + PYY_1 + AHNAK2_2 + GPR126_2 + DOCK9_2 +
      C9orf78_1 + VAPB_2 + PIGG_2, data = data1TP.poly)

Residual standard error: 0.889 on 539 degrees of freedom
Multiple R-squared:  0.322,  Adjusted R-squared:  0.288
F-statistic: 9.47 on 27 and 539 DF,  p-value: <2e-16

# BIC both data1TP
Call:
  lm(formula = TP53 ~ GYPB_1 + FAM59A_2 + SLC5A1_2 + RAI16_2
    + FTHP1_1 + CYP1B1_2 + SLIT3_1 + PCAF_1 + GAS2_1 +
      RHOBTB3_1 + SS18L2_2 + UBP1_2 + TM9SF4_1 + FAHD2A_2 +
      COL5A1_1 + COX8A_2 + PYY_1 + DOCK9_2 + C9orf78_1 +
      NPFFR1_1 + C8orf33_2 + TRIM36_2 + VAPB_1 + RELB_1 +
      OSBPL8_2 + SS18L2_1 + HSPB2_1 + RAB36_1 + PALMD_1 +
      IL6ST_1, data = data1TP.poly)

Residual standard error: 0.859 on 536 degrees of freedom
Multiple R-squared:  0.37,  Adjusted R-squared:  0.335
F-statistic: 10.5 on 30 and 536 DF,  p-value: <2e-16


#################### SUBSET B of TP53 ####################

# BIC forward data2TP
Call:
  lm(formula = TP53 ~ KIAA0753_1 + KIF5C_1 + HERC2_1 +
      ATP10A_1 + KNTC1_1 + NUBP1_1 + FLJ10815_2 +
      MIF_1 + RNF7_2, data = data2TP.poly)

Residual standard error: 0.948 on 557 degrees of freedom
Multiple R-squared:  0.203,  Adjusted R-squared:  0.19
F-statistic: 15.8 on 9 and 557 DF,  p-value: <2e-16

# BIC backward data2TP
Call:
  lm(formula = TP53 ~ SLC25A20_2 + AGTR1_2 + RNF144A_1 +
      FOSB_1 + ANKRD46_1 + MYBPH_1 + ABCA3_2 + TARDBP_2
    + NUBP1_1 + TAF1C_2 + PUS7L_2 + CHSY1_1 + BRWD1_1 +
      ZDHHC3_2 + CEL_1 + KNTC1_1 + KIAA0753_1 + SRR_2 +
      NDUFA5_1 + ATP10A_1 + HERC2_1, data = data2TP.poly)

Residual standard error: 0.897 on 545 degrees of freedom
Multiple R-squared:  0.303,  Adjusted R-squared:  0.276
F-statistic: 11.3 on 21 and 545 DF,  p-value: <2e-16

# BIC both data2TP
Call:
  lm(formula = TP53 ~ RNF144A_1 + FOSB_1 + ANKRD46_1 +
      MYBPH_1 + ABCA3_2 + TARDBP_2 + NUBP1_1 + TAF1C_2 +
      CHSY1_1 + BRWD1_1 + CEL_1 + KNTC1_1 + KIAA0753_1 +
      SRR_2 + NDUFA5_1 + ATP10A_1 + HERC2_1 + PYY_1 +
      LIF_2 + R3HDM1_1 + WDR79_2, data = data2TP.poly)

Residual standard error: 0.892 on 545 degrees of freedom
Multiple R-squared:  0.31,  Adjusted R-squared:  0.284
F-statistic: 11.7 on 21 and 545 DF,  p-value: <2e-16

#################### CROSS VALIDATION ####################

result1
#[1] 141 138 137
result2
#[1] 108 109 109
result3
#[1] 0.829 0.832 0.781
result4
#[1] 0.921 0.838 0.824

# ########################
# #        GMC           #
# ########################
#
################## No TRANSFORMATION ##################

# repsonse in subset A
colnames(x)[index]
#[1]  "NRG1"  "C1orf27"  "PTPN3"  "C21orf59"  "MMP7"  "
     AHNAK2"  "PKN2"  "RPL7A"  "PTPN2"  "DCC"
gmc
#[1]  0.05342191

# response in subset B
colnames(x)[index]
#[1]  "KCNJ10"  "GPR27"  "PYY"  "ATF4"  "NDUFA5"  "GADD45G"
      "APEX1"  "HCRTR2"  "ISG20"  "TAPBPL"
gmc
#[1]  0.09739211

# response in combined subset A&B
colnames(x)[index]
#[1]  "KCNJ10"  "CYB5R2"  "C1orf27"  "ESRRG"  "SLC16A5"  "
     ACCN3"  "FZR1"  "GADD45G"  "NPY"  "MAP3K6"
gmc
#[1]  0.11469

# TP53 in subset A
colnames(x)[index]
#[1]  "POLL"  "FARS2"  "ARHGEF2"  "SFRS11"  "PTPN2"  "CENPQ"
      "TRIM36"  "HOXD3"  "KIAA0143"  "RAB36"
gmc
#[1]  0.1429933

# TP53 in subset B
colnames(x)[index]
#[1]  "KIAA0753"  "KBTBD4"  "GPR27"  "KIF5C"  "KIF1C"  "
     KLRB1"  "WDR79"  "PITPNM1"  "PITPNC1"  "AGTR1"
gmc
#[1]  0.2173538


# TP53 in combined subset A&B
colnames(x)[index]
#[1]  "KCNJ10"  "CYB5R2"  "C1orf27"  "ESRRG"  "SLC16A5"  "
     ACCN3"  "FZR1"  "GADD45G"  "NPY"  "MAP3K6"
gmc
#[1]  0.11469

################## SQUARE TRANSFORMATION ##################

# repsonse in subset A
colnames(x)[index]
#[1]  "NRG1"  "C1orf27"  "PTPN3"  "PCAF"  "SMAD7"  "MMP7"  "
     CA6"  "C21orf59"  "KIAA0947"  "GRB2"
gmc
```

```
#[1]  0.05886893

# repsonse in subset B
colnames(x)[index]
#[1]  "IFNA2"  "ASTE1"  "FGF5"  "PRPF39"  "KCNJ10"  "XPA"  "
     HERC2"  "AGTR1"  "ZDHHC3"  "KIF5C"
gmc
#[1]  0.08715772

# repsonse in combined subset A&B
colnames(x)[index]
#[1]  "IFNA2"  "FRMD4B"  "AMOT"  "RHOBTB3"  "RPL8"  "PITPNM1
     "  "HLA.G"  "KCNJ10"  "GPR77"  "COPG"
gmc
#[1]  0.08699893

# TP53 in subset A
colnames(x)[index]
#[1]  "POLL"  "FARS2"  "ARHGEF2"  "RELB"  "PIGG"  "RCAN3"  "
     NCAPG"  "MAPKAP1"  "TMEM2"  "TSPAN1"
gmc
#[1]  0.1958418

# TP53 in subset B
colnames(x)[index]
#[1]  "KIAA0753"  "KBTBD4"  "GPR27"  "PCDH1"  "PUM1"  "
     PLXNA3"  "FAM83E"  "AQP5"  "CCDC90A"  "S100A11"
gmc
#[1]  0.1724622

# TP53 in combined subset A&B
colnames(x)[index]
#[1]  "IFNA2"  "FRMD4B"  "AMOT"  "RHOBTB3"  "RPL8"  "PITPNM1
     "  "HLA.G"  "KCNJ10"  "GPR77"  "COPG"
gmc
#[1]  0.08699893

################## CUBIC TRANSFORMATION ##################

# response in subset A
colnames(x)[index]
#[1]  "NRG1"  "C1orf27"  "PTPN3"  "GPR6"  "FARS2"  "CDH22"
     "NPFFR1"  "C21orf59"  "LMO7"  "VAPB"
gmc
#[1]  0.07306184

# response in subset B
colnames(x)[index]
#[1]  "KCNJ10"  "GPR27"  "PYY"  "CLEC4E"  "ORM1"  "SNRPE"  "
     AQP5"  "WDR79"  "C18orf1"  "CRISP1"
gmc
#[1]  0.1101837

# response in subset A&B
colnames(x)[index]
#[1]  "KCNJ10"  "CYB5R2"  "RPL8"  "GSTT1"  "CPT1B"  "AHNAK"
     "TSPYL4"  "TARDBP"  "SLC9A8"  "MAP3K6"
gmc
#[1]  0.1045698

# TP53 in subset A
colnames(x)[index]
#[1]  "POLL"  "FARS2"  "ARHGEF2"  "SFRS11"  "PTPN2"  "
     ADAMTS5"  "GRB2"  "SURF2"  "FNDC4"  "MED24"
gmc
#[1]  0.1558251

# TP53 in suubset B
colnames(x)[index]
#[1]  "KIAA0753"  "KBTBD4"  "GPR27"  "KLRB1"  "HSD11B2"  "
     ZNF343"  "AQP5"  "HDGF"  "ZNF587"  "CCT6A"
#  gmc.re
#[1]  0.2173554

# TP53 in subset A&B
colnames(x)[index]
#[1]  "KCNJ10"  "CYB5R2"  "RPL8"  "GSTT1"  "CPT1B"  "AHNAK"
     "TSPYL4"  "TARDBP"  "SLC9A8"  "MAP3K6"
gmc
#[1]  0.1045698
```

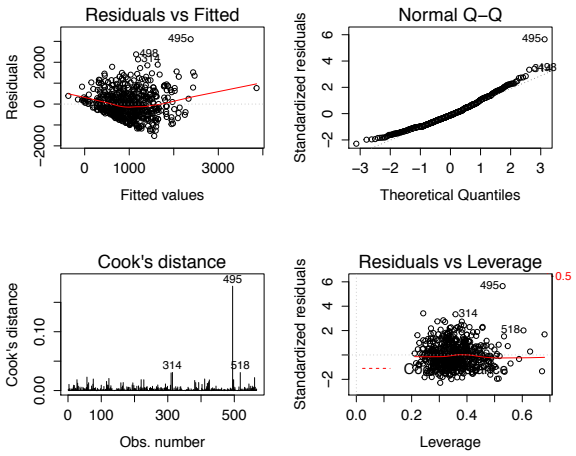

Fig. 1: diagnostic plot of "$response \sim .$" in subset A

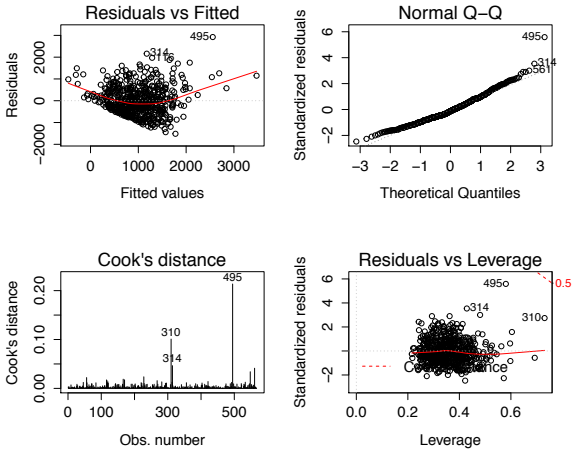

Fig. 2: diagnostic plot of "$response \sim .$" in subset B

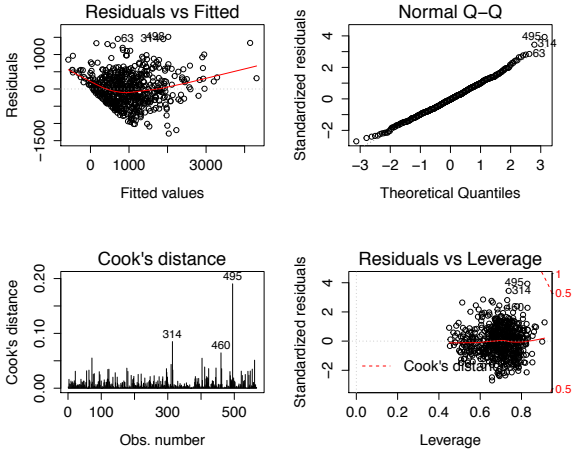

Fig. 3: diagnostic plot of "$response \sim .$" in subset A&B

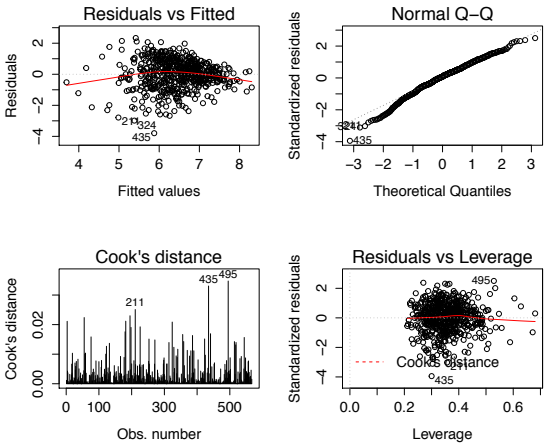

Fig. 4: diagnostic plot of *"log(response)~."* in set A

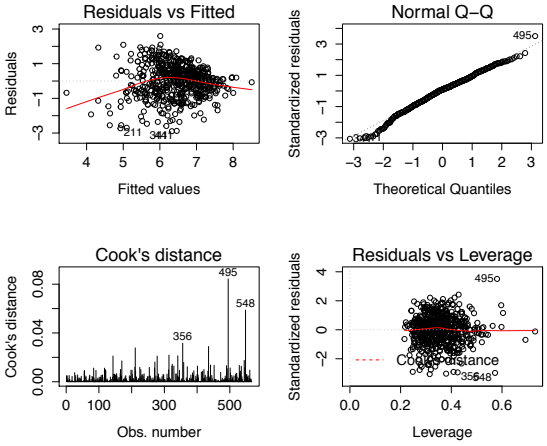

Fig. 5: diagnostic plot of *"log(response)~."* in set B

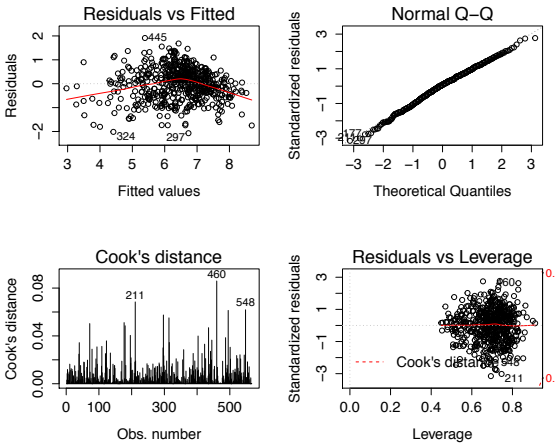

Fig. 6: diagnostic plot of *"log(response)~."* in A&B

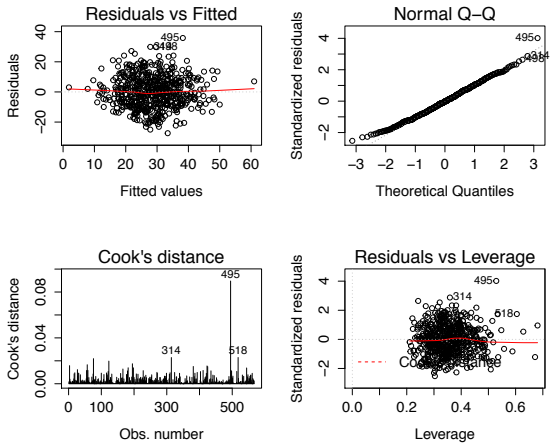

Fig. 7: diagnostic plot of "$\sqrt{response}$~." in set A

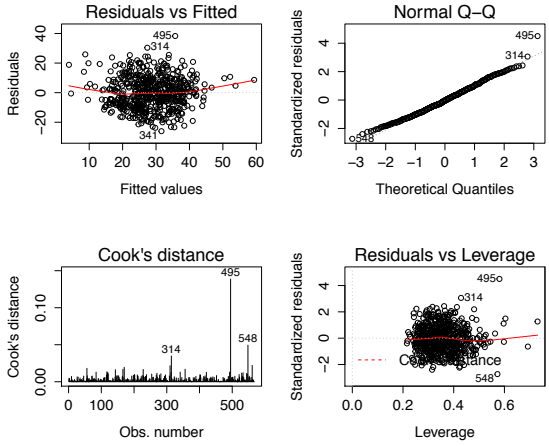

Fig. 8: diagnostic plot of "$\sqrt{response}$~." in set B

Fig. 9:diagnostic plot of "$\sqrt{response}$~." in A&B

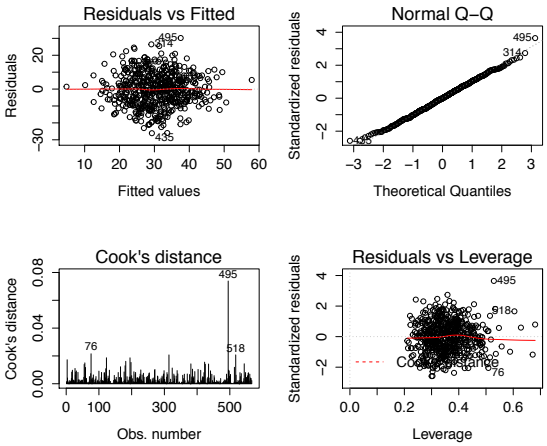

Fig. 10:diagnostic plot of "$\frac{response^{\lambda}-1}{\lambda}\sim.$" in set A

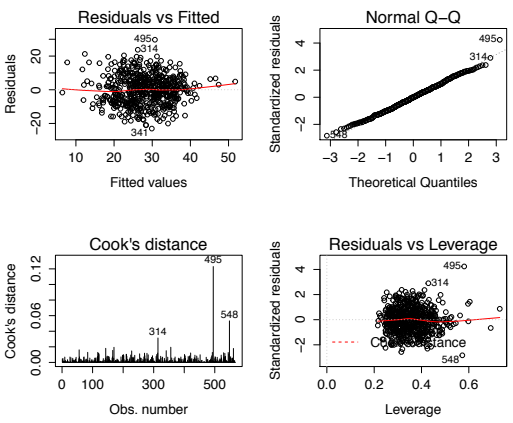

Fig. 11:diagnostic plot of "$\frac{response^{\lambda}-1}{\lambda}\sim.$" in set B

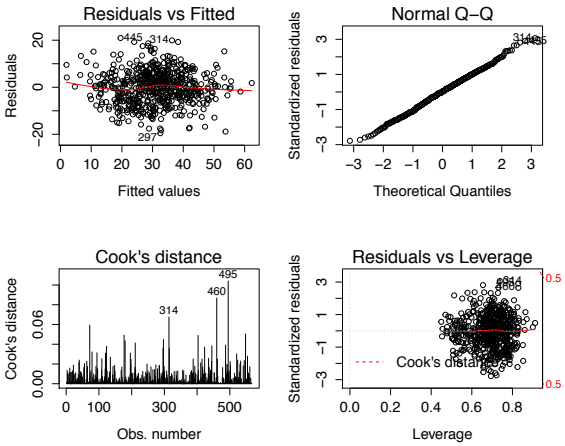

Fig. 12:diagnostic plot of "$\frac{response^{\lambda}-1}{\lambda}\sim.$" in A&B

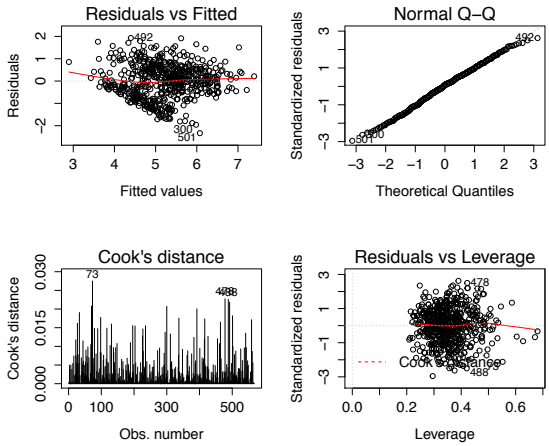

Fig. 13: diagnostic plot of *"TP53~."* in subset A

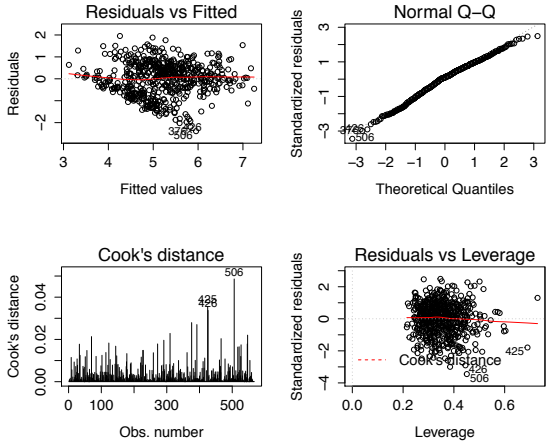

Fig. 14: diagnostic plot of *"TP53~."* in subset B

Fig. 15: diagnostic plot of *"TP53~."* in subset A&B

TABLE V: GMC of "response" for each $g(x)$ function

		$g(x)=x$	$g(x)=e^x$	$g(x)=(x)^2$	$g(x)=(x)^3$	$g(x)=sin(x)$
Subset A	GMC	0.34	0.27	0.32	0.302	0.23
	$\#\lvert\beta_i\rvert < 0.01$	2	2	2	12	5
	$\Delta < 0.01$	16	21	26	37	17
Subset B	GMC	0.32	0.28	0.33	0.31	0.31
	$\#\lvert\beta_i\rvert < 0.01$	6	1	2	3	3
	$\Delta < 0.01$	27	10	32	31	47
Set A&B	GMC	0.61	0.47	0.60	0.58	0.50
	$\#\lvert\beta_i\rvert < 0.01$	5	2	8	9	5
	$\Delta < 0.01$	21	19	33	52	19

Remark: $\Delta < \mathbf{0.01}$ ***means the number of variables drop while GMC change less than*** $\mathbf{0.01}$

TABLE VI: GMC of "response" for each $g(x)$ function

		$g(x)=x$	$g(x)=e^x$	$g(x)=(x)^2$	$g(x)=(x)^3$	$g(x)=sin(x)$
Subset A	GMC	0.48	0.49	0.48	0.47	0.33
	$\#\lvert\beta_i\rvert < 0.01$	7	7	18	31	5
	GMC	0.45	0.45	0.39	0.29	0.36
	$\#\lvert\beta_i\rvert < 0.05$	39	53	98	133	40
Subset B	GMC	0.49	0.49	0.48	0.48	0.45
	$\#\lvert\beta_i\rvert < 0.01$	12	12	33	33	7
	GMC	0.47	0.48	0.38	0.31	0.44
	$\#\lvert\beta_i\rvert < 0.05$	49	58	99	129	35
Set A&B	GMC	0.75	0.75	0.75	0.74	0.63
	$\#\lvert\beta_i\rvert < 0.01$	17	15	8	51	22
	GMC	0.70	0.70	0.63	0.45	0.63
	$\#\lvert\beta_i\rvert < 0.05$	84	88	72	207	72