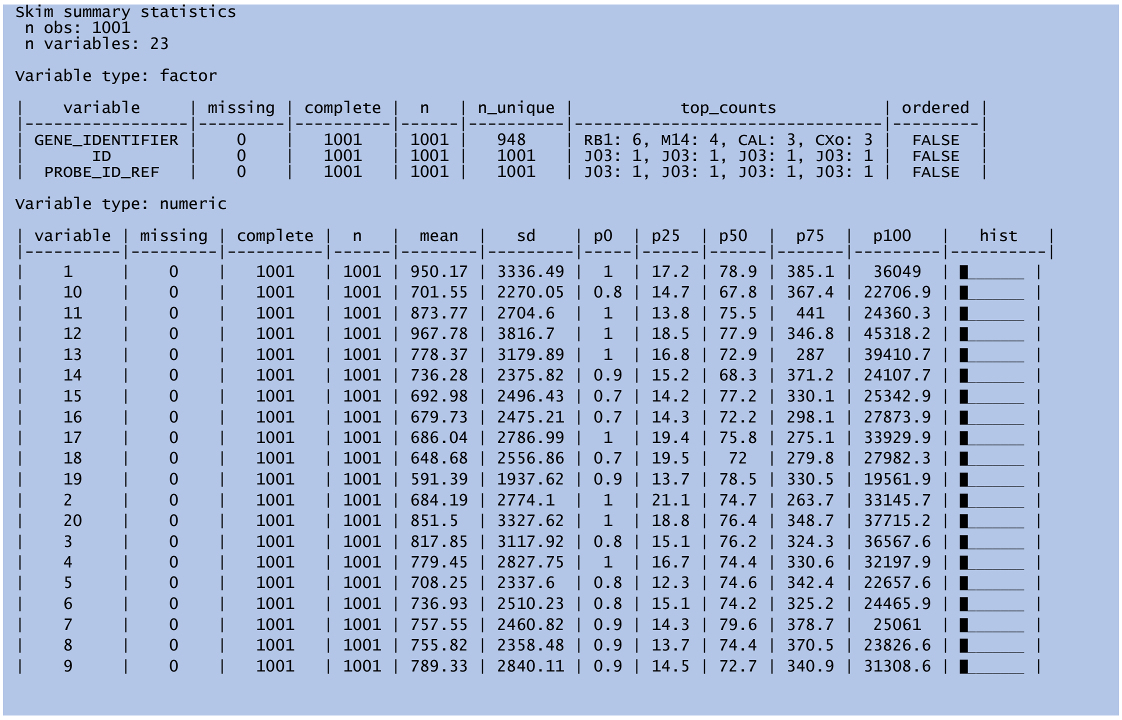
Introduction

The importance of genetic predisposition, inflammation, and auto-immune mechanisms in the development of pulmonary arterial hypertension (PAH) is becoming increasingly clear. It is hypothesised that the development of PAH requires first a genetic susceptibility followed by one or several secondary trigger factors such as a viral infection or drug exposure. The individual's genetic susceptibility and the interaction of the genotype with the promoting factor(s) remain areas of active research. In this project, genes that discriminate between hypertensive and healthy individuals will be identified by analysing gene expression profiles from peripheral blood mononuclear cells.

Exploratory Data Analysis

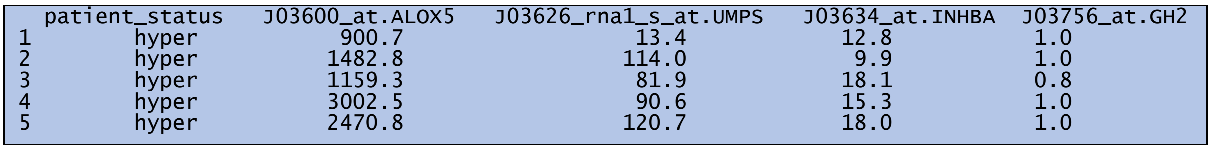
By looking at the data summary, our data comprises of 1001 genes (rows), 20 individuals (columns) and 3 genes identification columns. Samples are comprised of 14 hypertensive and 6 healthy individuals (imbalanced classes). The dataset has numeric variables and factor variables with no missing data. To create a unique ID for genes, genes identifier and prop ID identifier are combined. The data is cleaned such that samples represent rows and genes represent columns. After data cleaning and manipulation, our data consists of 1002 variables (1001 genes) and 20 samples (individuals). The variable of interest is patient status which indicates whether patient is healthy (value 1) or with PAH (value 0). Also, genes have different ranges and part of the genes are highly correlated (Figure 1 and 3). Our goal is to specify genes which have impact on patient status (healthy or hypertensive).

Figure 1: Summary of data before manipulation



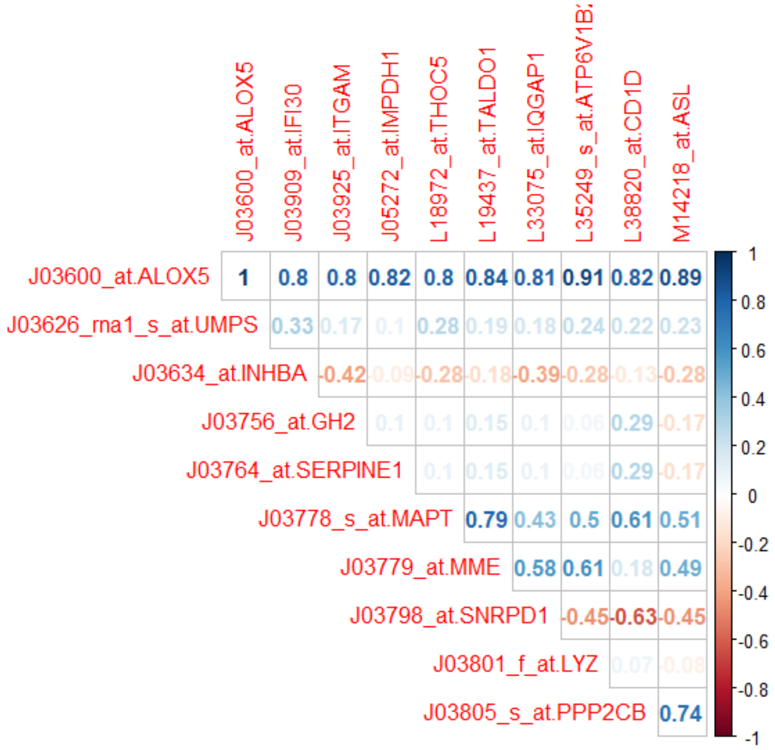
```
Skim summary statistics
 n obs: 1001 
 n variables: 23 

Variable type: factor

|    variable     | missing | complete |  n   | n_unique |            top_counts             | ordered |
|-----------------|---------|----------|------|----------|-----------------------------------|---------|
| GENE_IDENTIFIER |    0    |   1001   | 1001 |   948    | RB1: 6, M14: 4, CAL: 3, CXo: 3 |  FALSE  |
|       ID        |    0    |   1001   | 1001 |   1001   | J03: 1, J03: 1, J03: 1, J03: 1 |  FALSE  |
|  PROBE_ID_REF   |    0    |   1001   | 1001 |   1001   | J03: 1, J03: 1, J03: 1, J03: 1 |  FALSE  |

Variable type: numeric

| variable | missing | complete |  n   |  mean  |   sd    | p0  | p25  | p50  |  p75  |  p100   |   hist   |
|----------|---------|----------|------|--------|---------|-----|------|------|-------|---------|----------|
|    1     |    0    |   1001   | 1001 | 950.17 | 3336.49 |  1  | 17.2 | 78.9 | 385.1 |  36049  | ▇▁▁▁▁▁▁▁ |
|    10    |    0    |   1001   | 1001 | 701.55 | 2270.05 | 0.8 | 14.7 | 67.8 | 367.4 | 22706.9 | ▇▁▁▁▁▁▁▁ |
|    11    |    0    |   1001   | 1001 | 873.77 | 2704.6  |  1  | 13.8 | 75.5 |  441  | 24360.3 | ▇▁▁▁▁▁▁▁ |
|    12    |    0    |   1001   | 1001 | 967.78 | 3816.7  |  1  | 18.5 | 77.9 | 346.8 | 45318.2 | ▇▁▁▁▁▁▁▁ |
|    13    |    0    |   1001   | 1001 | 778.37 | 3179.89 |  1  | 16.8 | 72.9 |  287  | 39410.7 | ▇▁▁▁▁▁▁▁ |
|    14    |    0    |   1001   | 1001 | 736.28 | 2375.82 | 0.9 | 15.2 | 68.3 | 371.2 | 24107.7 | ▇▁▁▁▁▁▁▁ |
|    15    |    0    |   1001   | 1001 | 692.98 | 2496.43 | 0.7 | 14.2 | 77.2 | 330.1 | 25342.9 | ▇▁▁▁▁▁▁▁ |
|    16    |    0    |   1001   | 1001 | 679.73 | 2475.21 | 0.7 | 14.3 | 72.2 | 298.1 | 27873.9 | ▇▁▁▁▁▁▁▁ |
|    17    |    0    |   1001   | 1001 | 686.04 | 2786.99 |  1  | 19.4 | 75.8 | 275.1 | 33929.9 | ▇▁▁▁▁▁▁▁ |
|    18    |    0    |   1001   | 1001 | 648.68 | 2556.86 | 0.7 | 19.5 |  72  | 279.8 | 27982.3 | ▇▁▁▁▁▁▁▁ |
|    19    |    0    |   1001   | 1001 | 591.39 | 1937.62 | 0.9 | 13.7 | 78.5 | 330.5 | 19561.9 | ▇▁▁▁▁▁▁▁ |
|    2     |    0    |   1001   | 1001 | 684.19 | 2774.1  |  1  | 21.1 | 74.7 | 263.7 | 33145.7 | ▇▁▁▁▁▁▁▁ |
|    20    |    0    |   1001   | 1001 | 851.5  | 3327.62 |  1  | 18.8 | 76.4 | 348.7 | 37715.2 | ▇▁▁▁▁▁▁▁ |
|    3     |    0    |   1001   | 1001 | 817.85 | 3117.92 | 0.8 | 15.1 | 76.2 | 324.3 | 36567.6 | ▇▁▁▁▁▁▁▁ |
|    4     |    0    |   1001   | 1001 | 779.45 | 2827.75 |  1  | 16.7 | 74.4 | 330.6 | 32197.9 | ▇▁▁▁▁▁▁▁ |
|    5     |    0    |   1001   | 1001 | 708.25 | 2337.6  | 0.8 | 12.3 | 74.6 | 342.4 | 22657.6 | ▇▁▁▁▁▁▁▁ |
|    6     |    0    |   1001   | 1001 | 736.93 | 2510.23 | 0.8 | 15.1 | 74.2 | 325.2 | 24465.9 | ▇▁▁▁▁▁▁▁ |
|    7     |    0    |   1001   | 1001 | 757.55 | 2460.82 | 0.9 | 14.3 | 79.6 | 378.7 |  25061  | ▇▁▁▁▁▁▁▁ |
|    8     |    0    |   1001   | 1001 | 755.82 | 2358.48 | 0.9 | 13.7 | 74.4 | 370.5 | 23826.6 | ▇▁▁▁▁▁▁▁ |
|    9     |    0    |   1001   | 1001 | 789.33 | 2840.11 | 0.9 | 14.5 | 72.7 | 340.9 | 31308.6 | ▇▁▁▁▁▁▁▁ |
```

Figure 2: First 5 observations and columns of data after manipulation



```
  patient_status   J03600_at.ALOX5   J03626_rna1_s_at.UMPS   J03634_at.INHBA   J03756_at.GH2
1          hyper             900.7                    13.4              12.8             1.0
2          hyper            1482.8                   114.0               9.9             1.0
3          hyper            1159.3                    81.9              18.1             0.8
4          hyper            3002.5                    90.6              15.3             1.0
5          hyper            2470.8                   120.7              18.0             1.0
```

Figure 3: Correlation between 10 of the genes



Methodology

The variable range of the genes solved by subtracting the mean and dividing by the standard error (standardized,). Our data set is very small, so we don't have enough observations to put aside as a test data. Although the number of samples is too small, we split the data to training data (70%) and test data (30%) to give an outlook of the model performance. In the training dataset, there are 15 samples, among which 10 are hypertensive individuals and 5 are healthy individuals. The remaining 5 samples are used to test prediction accuracy (1 healthy and 4 hypertensive).

Ridge regression, Lasso and Elastic net are all applied to the training data. Model fitting is carried out on the training data. Five-fold cross validation is used for tuning or regularisation parameters selection, lambda for ridge, lasso and elastic net, while alpha for elastic net model. To assess model performance, we apply internal bootstrap validation. The model is fitted to the training data and the AUC is estimated. This process is repeated 100 times on a bootstrap sample from the training data and every time the AUC is estimated. Then the overall AUC is calculated. We used BootVAlidation package to obtain the validated measure of predictive accuracy (AUC). To examine if cross validation works well for the high dimensional data (P>>N), we draw 1000 realizations of the data and get the CV error for each iteration.

Results:

Figure 4 shows ridge model result. The plot of the coefficients' value as a function of the log of lambda. Each line represents one of the coefficients from the model (1001genes).
For larger values of lambda, the values of the coefficients go to zero. When the logarithm of lambda is 6, all the coefficients are approximately 0. As lambda gets smaller the coefficients grow away from zero and when lambda is zero the coefficients are unregularized. At the top of the plot, the number 1001 does not change at all no matter the value of lambda. This is expected in ridge regression as the coefficients never take the value 0. So, the number of predictors for ridge regression is equal to the number of variables we have 1001.

Figure 5 and figure 6 show the same plots for lasso regression and elastic net. But in lasso and elastic net the number of parameters decreases as the logarithm of lambda increases. These results are also expected since lasso regression performs variable selection by making some coefficients' values 0.
Using the cross validation to find optimal lambda for ridge regression, figure 4 shows the misclassification error explained for each value of lambda with the number of parameters equal to the number of genes 1001. The two vertical lines represent the value of lambda at the minimum misclassification error (λ=8.06) and the value within one standard error of the minimum misclassification error (λ=71.75). It can be noticed that the difference between them is high. Same result is shown in figure 5 and figure 6 for lasso regression and elastic net respectively. But the number of parameters is not constant as ridge, it is decreasing for a larger value of lambda.
Moreover, for the three models, figure 4,5 and 6 also display the coefficients' values as a function of fraction deviance explained (R^2). It also shows the relationship between the fraction deviance explained and the number of predictors. Each coloured line describes a single predictor and its coefficient value. Toward the end of the path, R^2 doesn't change much, but the coefficients are growing larger indicating likely overfitting of the model. The key is to find the optimal tuning parameters, so the selected model explains a large amount of patient status variability.

Table 1 shows the results from each model. Tuning parameters best values are provided. All our lasso and ridge results are focused on the lambda value that is associated with one standard error of the minimum misclassification error. It also shows performance comparison between the three models in terms of cross validation error and test error. It can be noticed for the three models the wide range of the 95% confidence interval, particularly with small lower bound, for cross validation and test accuracy. In terms of CV accuracy, the best model is elastic net, ridge comes after and lasso model has the lowest accuracy. But in terms of test accuracy all models have the same accuracy 80%. Ridge regression selected all the genes (1001) as it can't perform variable selection but rather shrinks parameter values. While 2 genes are selected by lasso, 24 more genes are selected by elastic net. If we check the correlation matrix between all the selected genes, we notice some high correlation between the selected genes, the highest is 0.89 (see figure 7).
Table 1 also provides bootstrap validation to the 5-fold cross validation for the three models. We used BootValidation package that provides a good method to assess the model performance by applying internal bootstrap validation. From the results, model with the best area under the curve is ridge, elastic comes after and the last one is lasso.
Figure 9 provides CV error using 1000 different realizations of our response variable for the three models. We can see that the CV error has high variance especially for ridge and lasso model. While elastic net has less variance. Same results in details are shown in table 2.

Table 1: Model performance and parameters of Ridge, Lasso and Elastic net

Model	Tuning parameters		Number of variable selected
	Lambda_min - lambda_1se	Alpha(α)	
Ridge	8.059337-71.74543	0	1001
Lasso	0.1196786-0.3828838	1	2
Elastic net	0.138098	0.5591837	26

Model	CV AUC	Bootstrap Validated AUC
Ridge	1	0.992
Lasso	1	0.976
Elastic net	1	0.972

Model	CV Accuracy	95% CI of the CV Accuracy	5-fold CV Error	Test Accuracy	95% CI of the Test Accuracy	Test Error
Ridge	0.8	(0.5191,0.9567)	3/15	0.8	(0.2836,0.9949)	1/5
Lasso	0.667	(0.3838, 0.8818)	5/15	0.8	(0.2836,0.9949)	1/5
Elastic net	1	(0.782,1)	0/5	0.8	(0.2836,0.9949)	1/5

Table 2: Shows CV error iterations for ridge,lasso and elastic model

Ridge CV	0/20	1/20	2/20	3/20	4/20	5/20	6/20
Frequency	364	107	7	31	14	-	477
Lasso CV error	0/20	1/20	2/20	3/20	4/20	5/20	6/20
Frequency	277	107	285	49	-	179	153
Elastic CV error	0/20	1/20	2/20	3/20	4/20	5/20	6/20
Frequency	980	15	4	-	1	-	-

Discussion

- We have two sample size concerns:
  - The training data is too small thus the models get unstable (change much if a few training instances are exchanged). This instability is checked by iterated CV. The predictions are changed in each iteration (see figure 9).
  - The test data is also small (<100).
- From our models, we noticed that CV have high variance. It counts errors in 15 sample and divide it by 15 to get the error estimate. So, the error estimates can only change by 1/15 increments. Because our sample is too small, 1/15 is quite large, which creates irreducible element of variability. This high variance is a problem. The result of a single CV can accidently have CV error 0, while the real error can be higher.
- Also, with small test sample size, the result will be subject to high variance. This can be noticed by observing the 95% confidence interval (CI) (see table 1). Our confidence interval spans the range from “worse than guessing” to “perfect”. The width of the confidence interval for an individual study depends to a large extent on the sample size. This is an indicator that the sample size of the test data is very small that hardy anything can be concluded from the test results.
- The main goal of this project is to specify genes or combinations of genes which have impact on patient status. From a biologist’s point of view, there is a grouping property among genes. Genes sharing the same biological pathway, are forming a group and can be highly correlated. The ideal model should have gene selection property in high dimensional data. Also, it should include the whole group into the model once one gene from the group is selected (group selection).
  - Ridge model has good accuracy. It shrinks the coefficients of correlated variables towards each other. It always tends to keep all the predictors in the model, so it cannot perform variable selection.
  - Unlike ridge, Lasso model is good for variable selection, but it has limiting features. When the number of variables significantly exceed the sample size (1001 >>20), lasso can select maximum 20 variables out of 1001. Also, it cannot perform group selection. For a group of variables with high pairwise correlation (like our data as shown in figure 3) lasso selects one of the variables and not necessary the most important one.

- Elastic net model performs well in a high dimensional data, when the number of variables is much larger than the sample size. Also, it has group selection capabilities. To tune parameters, CV chooses a model with minimal prediction error. Since elastic net can select any model that is selected by lasso, it can find model that predicts better than lasso. This can be noticed from the lasso and elastic models result in table 1.

- Considering the required model properties for this kind of data and our three models results, elastic net is the most suitable one. Providing all the above details about the model validity and given the confidence intervals for the performance, will not reduce the uncertainty as it is fundamentally caused by too few samples but will allow us to judge what can be concluded from the data. Mostly more samples are needed to have certain results.

Figure 4: Ridge Regression

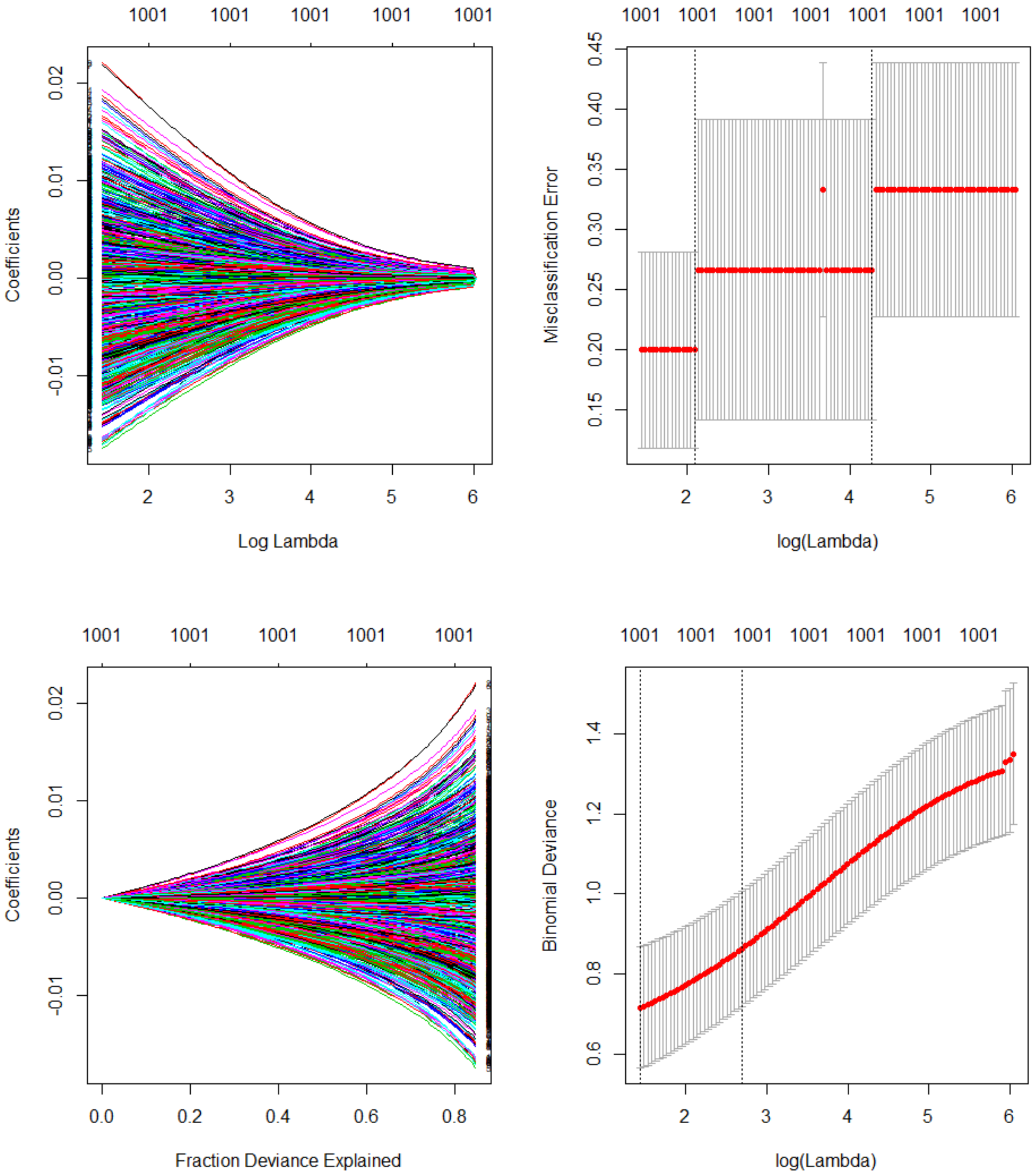


Figure 5: Lasso Logistic Regression

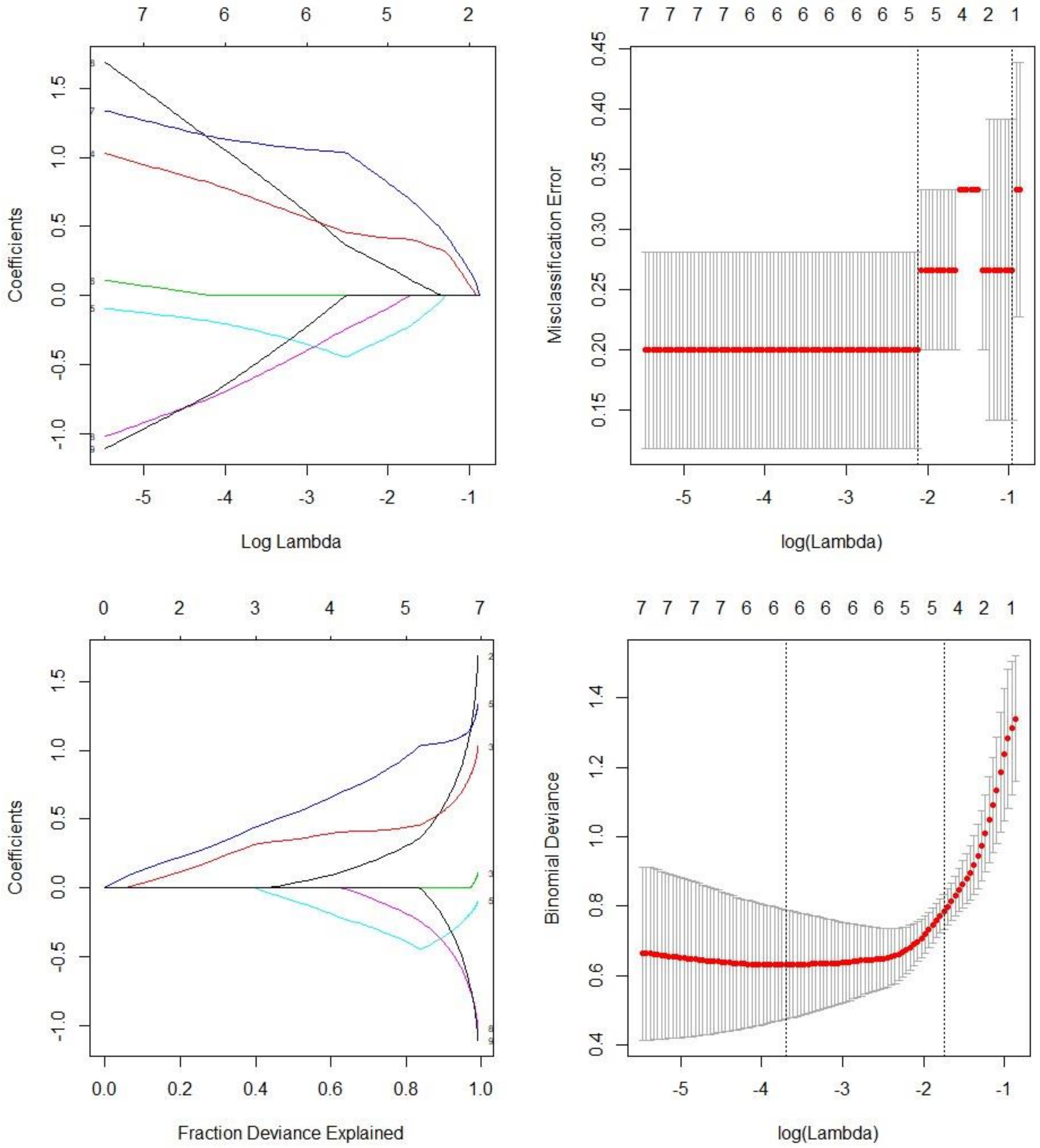


Figure 6 : Elastic Net Regression

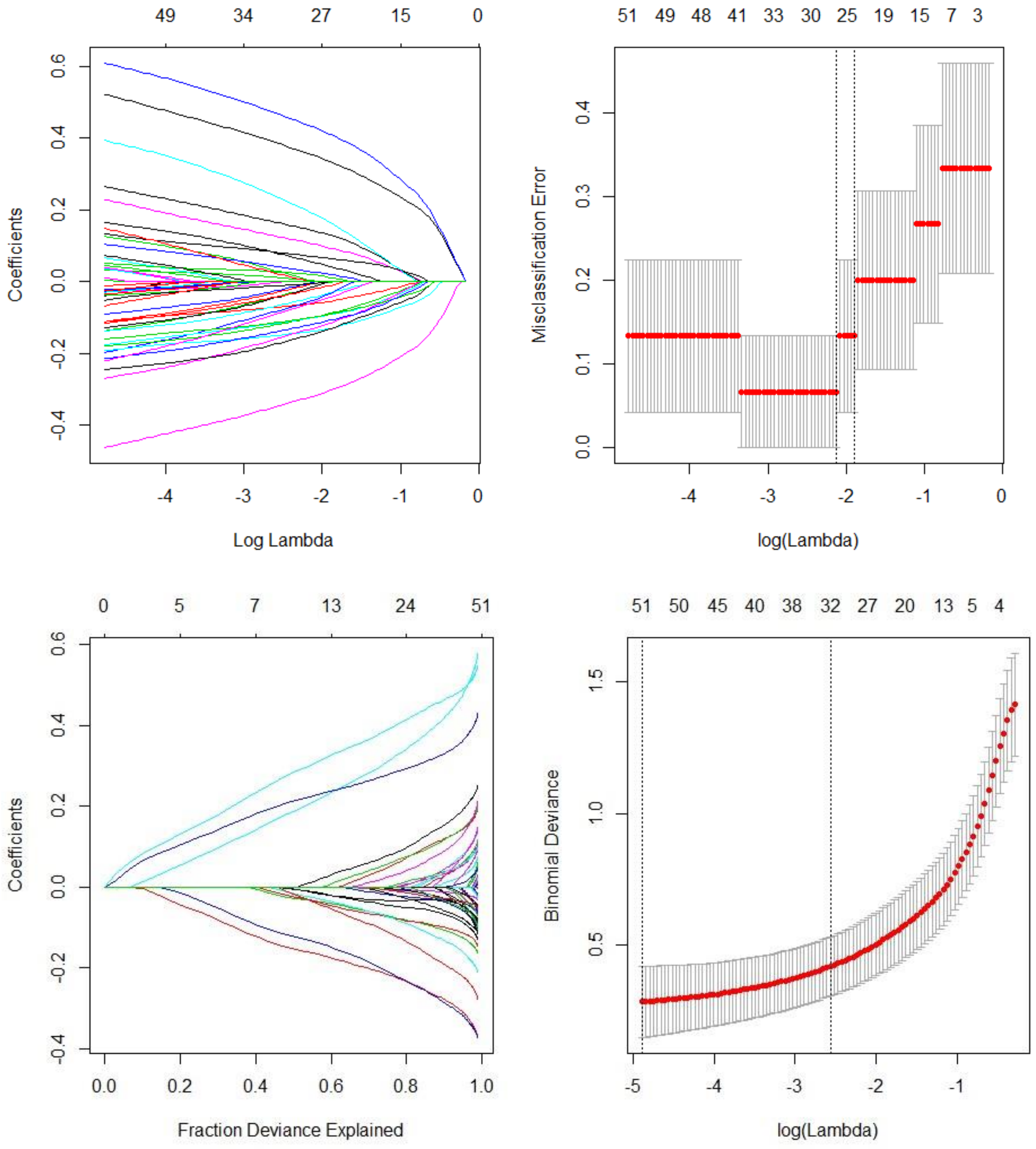


Figure 8: Correlation between selected predictors by lasso and elastic models.

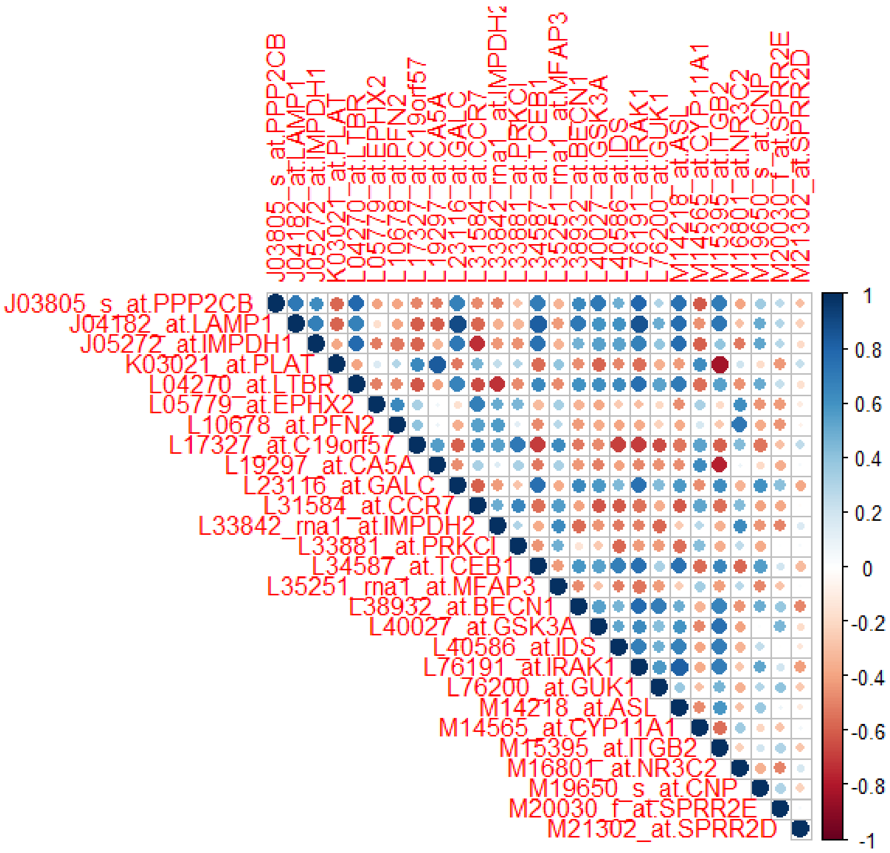
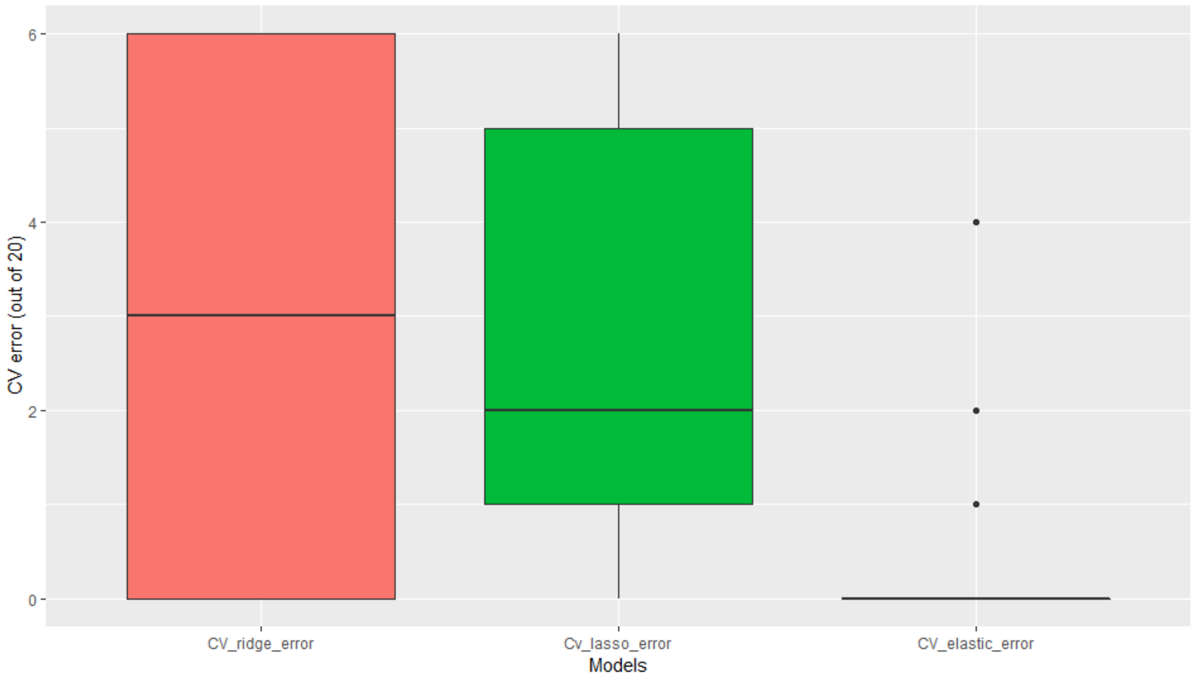


Figure 9: CV Error for 1000 realizations for the three models


CV error (out of 20)
CV_ridge_error
Cv_lasso_error
CV_elastic_error
Models