UNIVERSITÀ DEGLI STUDI DI PADOVA

Breast Cancer Survival Prediction

Caltran Lorenzo

Careddu Luca

Francario Felice

Mariethoz Jean Zacharie

Statistical Methods for High Dimensional Data
(2023 / 2024)

Project Presentation

Problem formulation and aim of the work

- ***Dataset:*** *"Molecular Taxonomy of Breast Cancer International Consortium"* (METABRIC), containing records of patients who underwent breast cancer surgery
- ***Problem:*** What is the probability of survival given these data?
- ***Method:*** We use both Survival analysis and binary classification methods to solve the problem and use Shrinkage penalization terms to get the most important variables

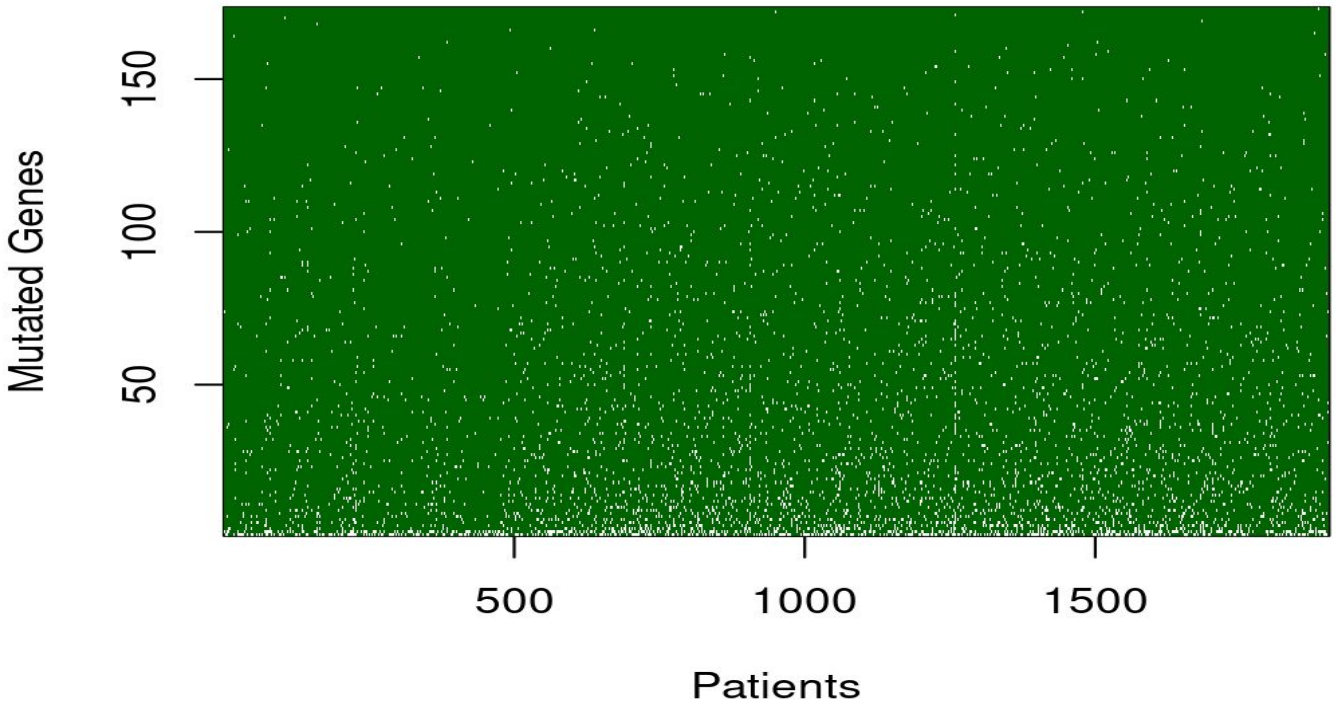
METABRIC

- It contains 1904 samples of 693 variables

The variables

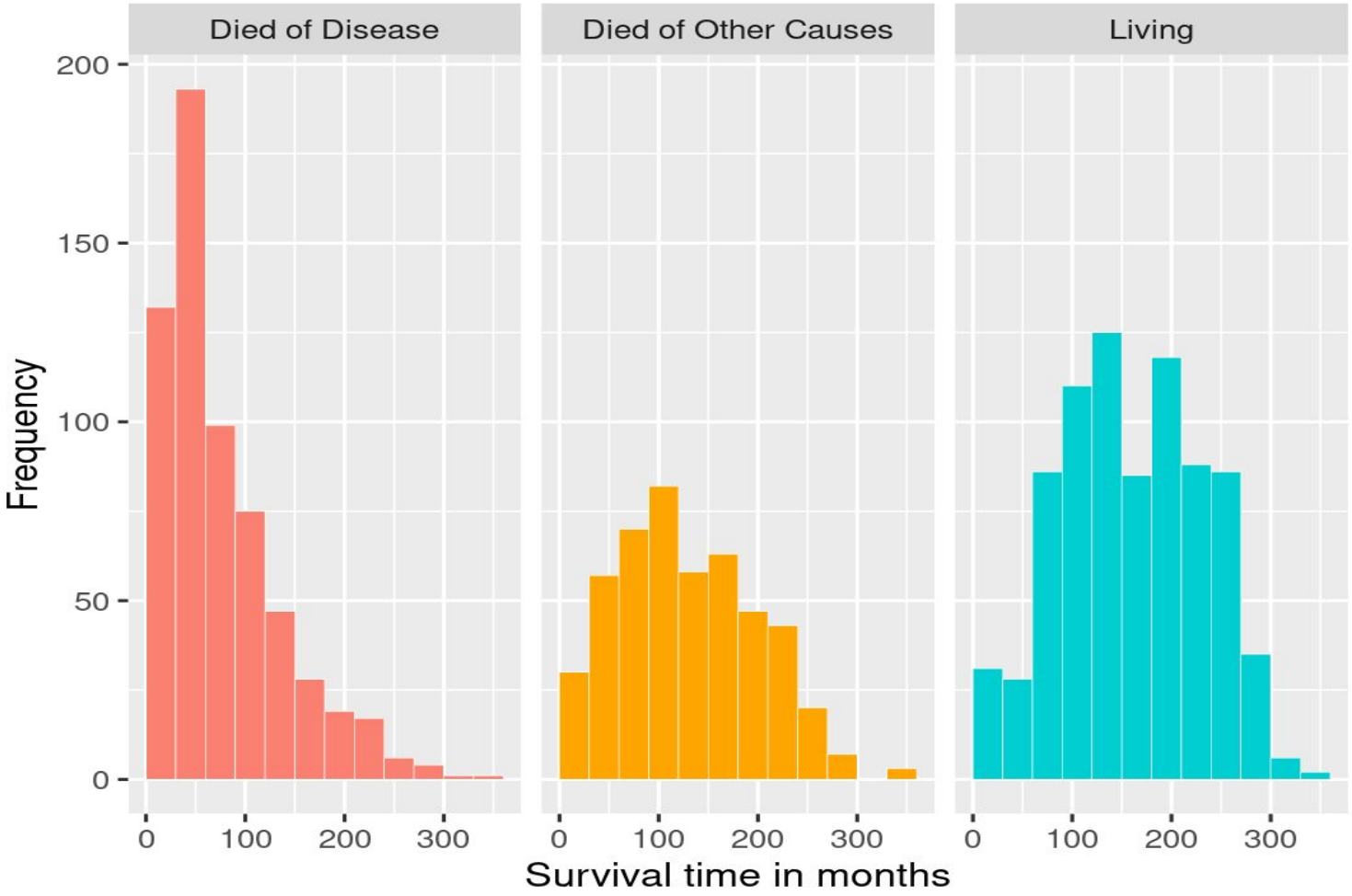
31	**Clinical**
489	**Genes**
173	**Mutated Genes**

Sparsity in the mutated genes variables

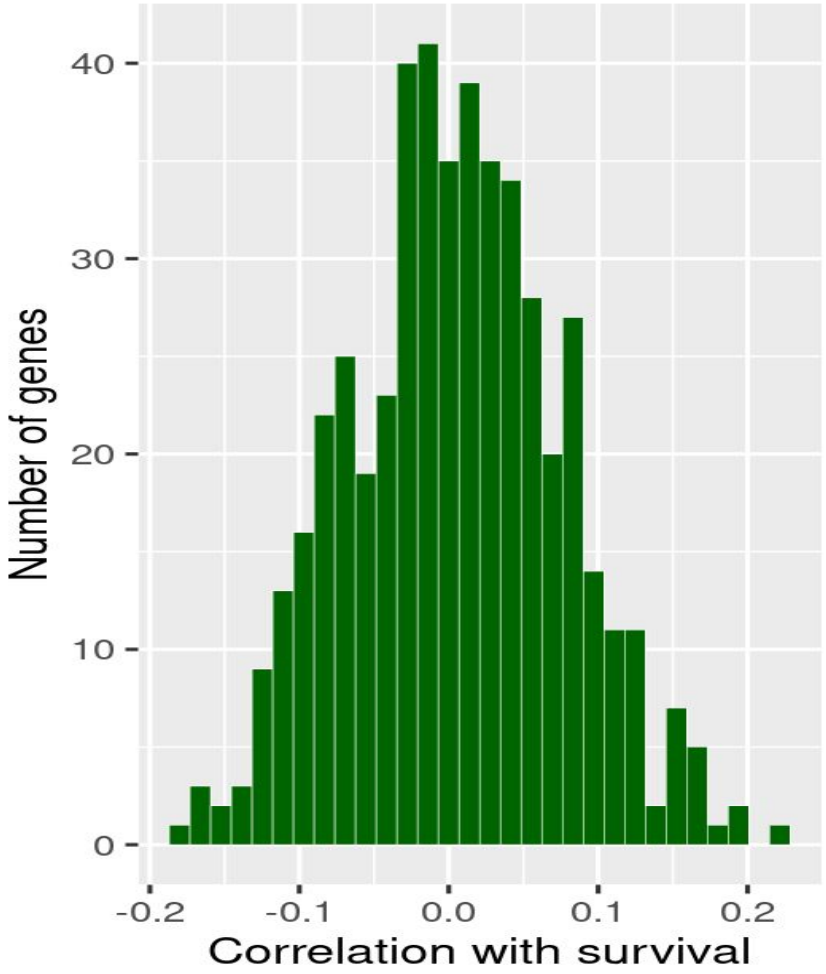


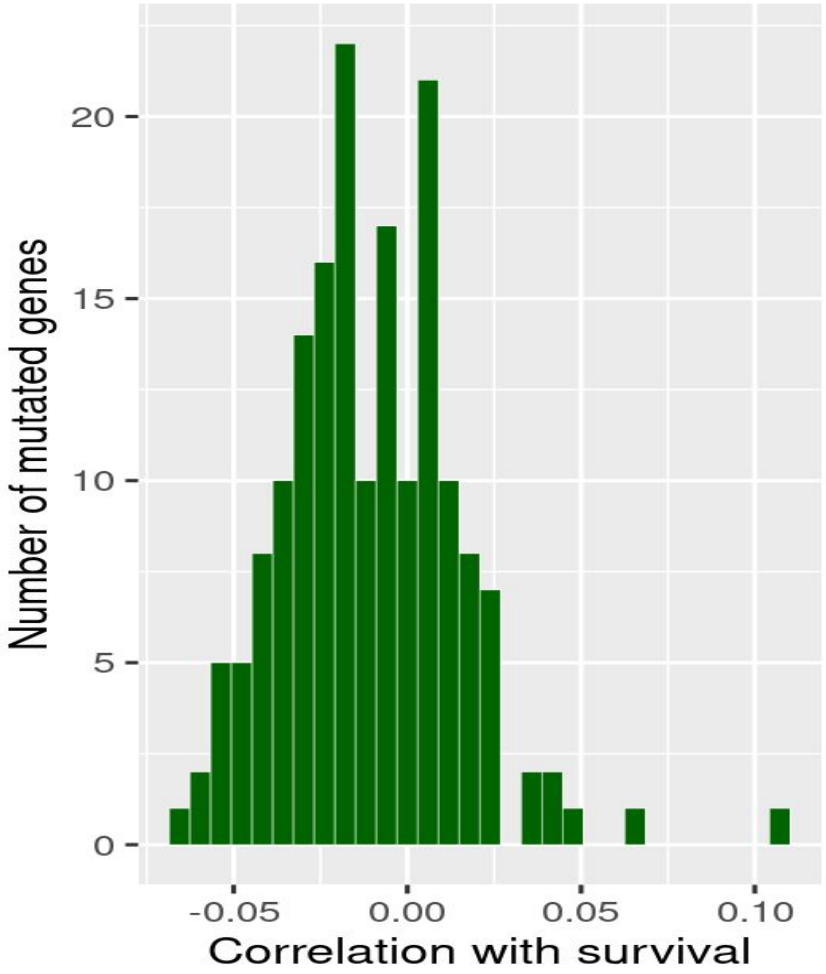
- Therefore, whenever possible, we use sparse matrix formats to speed up cross-validations

Frequencies: Survival



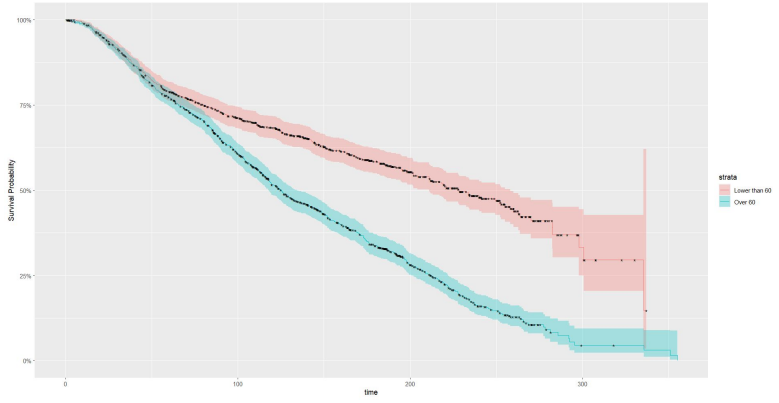
Frequencies: Correlations

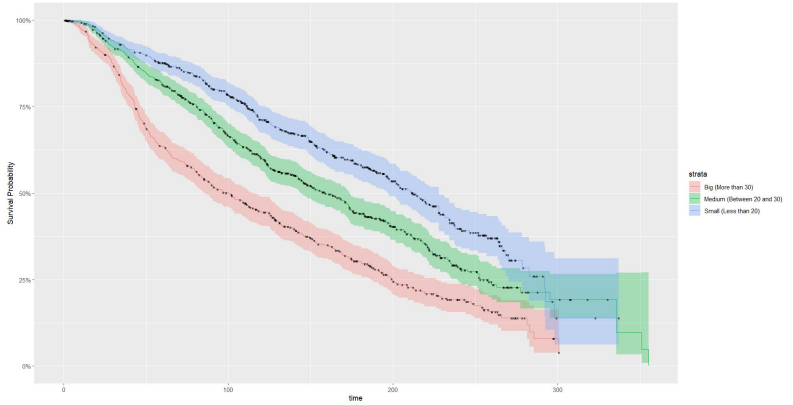


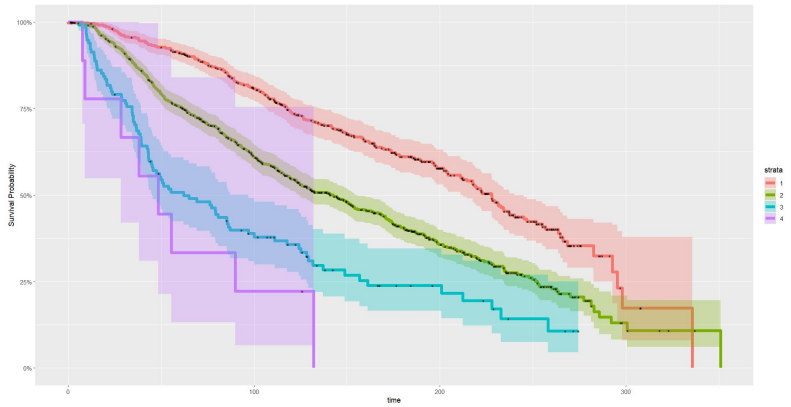


Survival Analysis

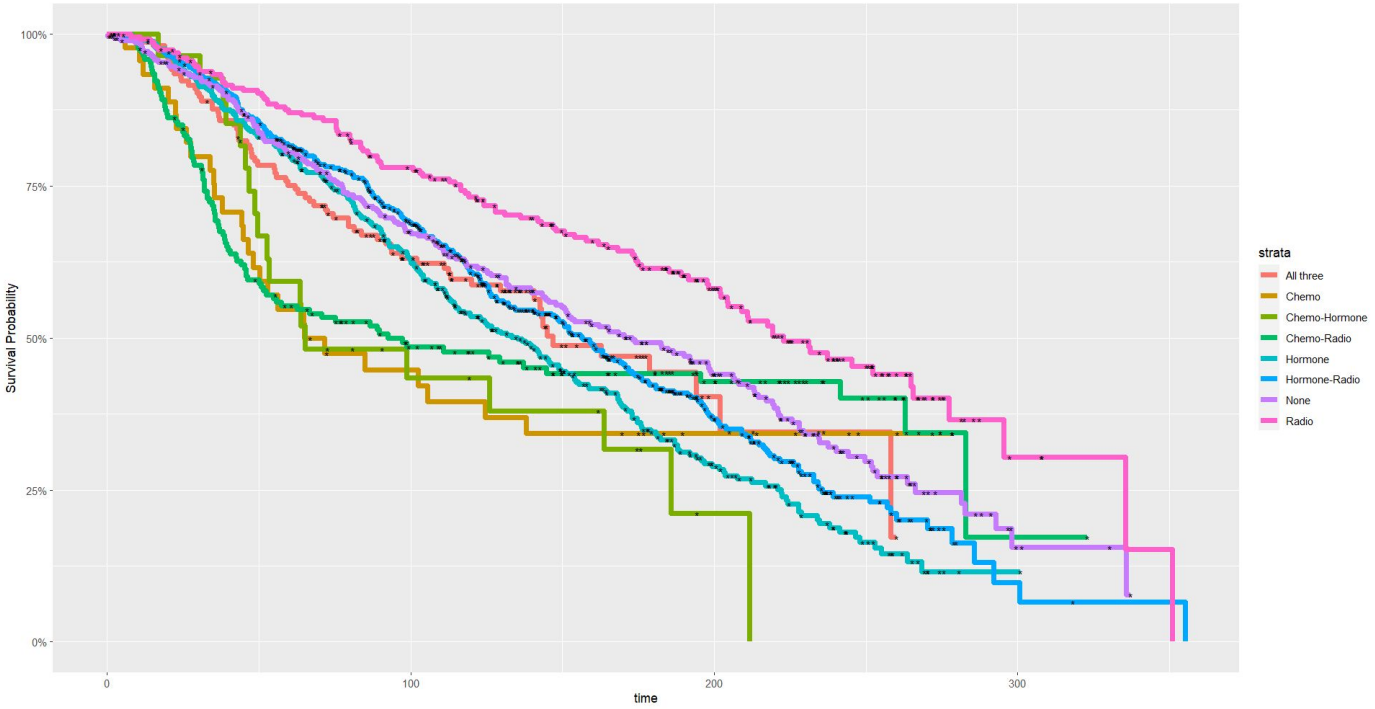
In this first part we want to investigate how the 29 clinical variables affect the survival probability. To do that we observe at the Kaplan-Meier curve for the features “Age at diagnosis”, “Tumor size” and “Cancer stage”



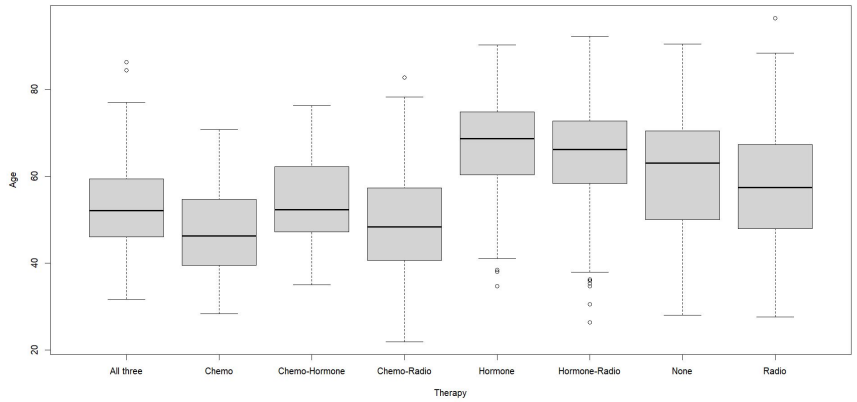


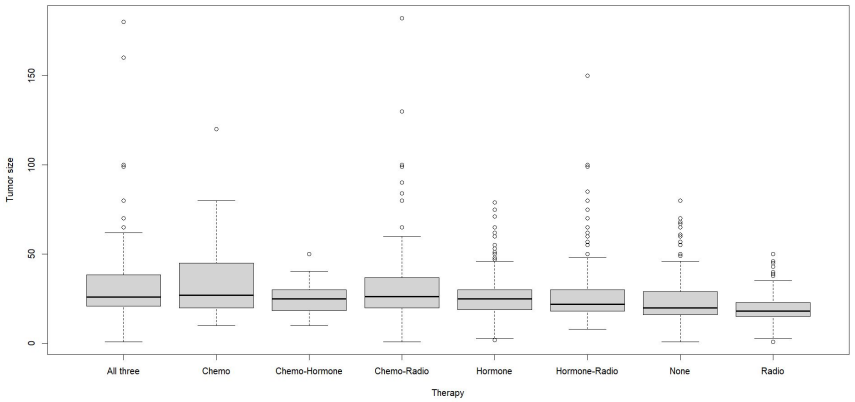


From this plot it appears that Chemotherapy and the combination of Chemotherapy and Radiotherapy are the most effective.
We can also observe that some of the most aggressive forms of therapy seem to guarantee a lower survival probability, which is an unexpected result.



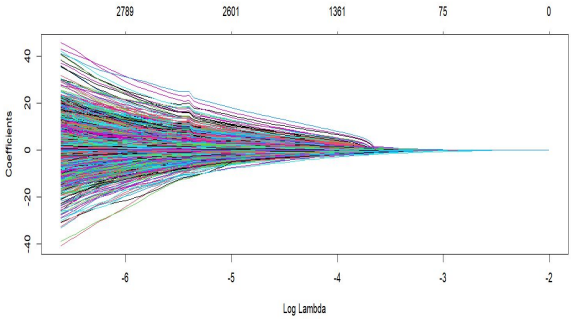
The reason why this happens is because the most aggressive types of therapy are used for patients more at risk and that affect negatively the survival probability.



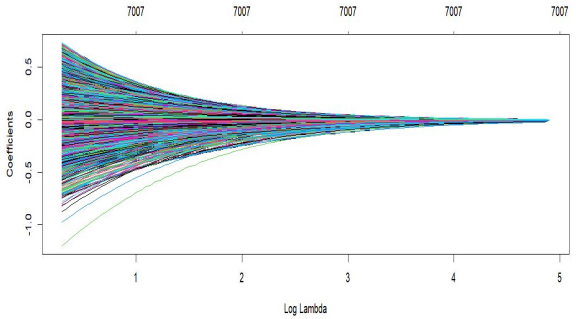


Survival Models

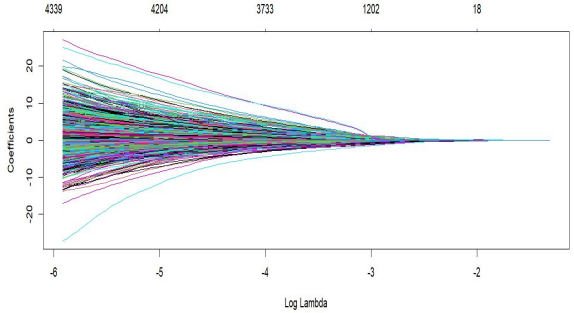
In this section we extract the genomic part of the dataset to test Cox Hazard Proportional model with different types of regularization: Lasso, Ridge and Elastic Net. For each type of regularization we get a series of models with different values of regularization parameters and degrees of freedom



Lasso

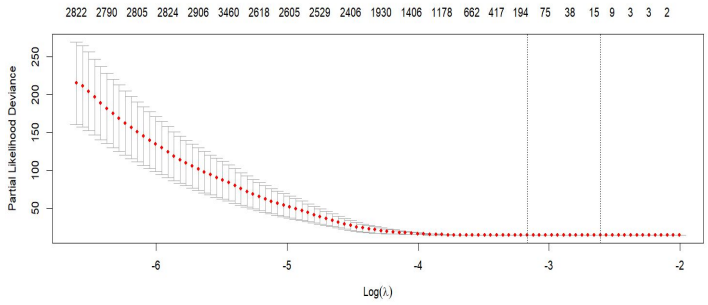


Ridge

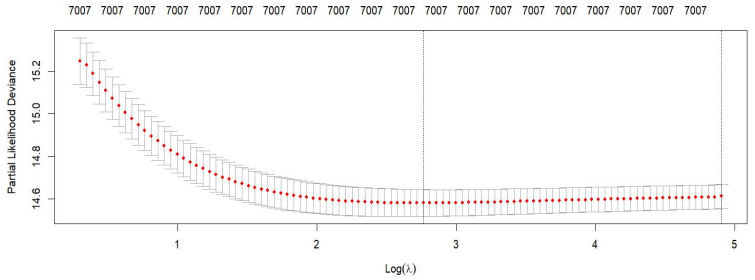


Elastic Net (alpha = 0.5)

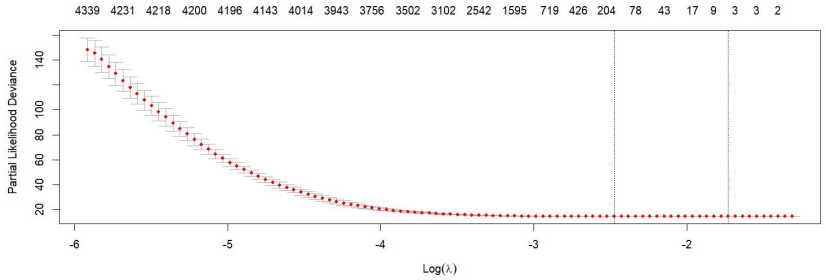
We then selected the model with regularization parameter that results in the minimum partial likelihood deviance by using 5-fold Cross Validation



Lasso



Ridge



Elastic Net (alpha = 0.5)

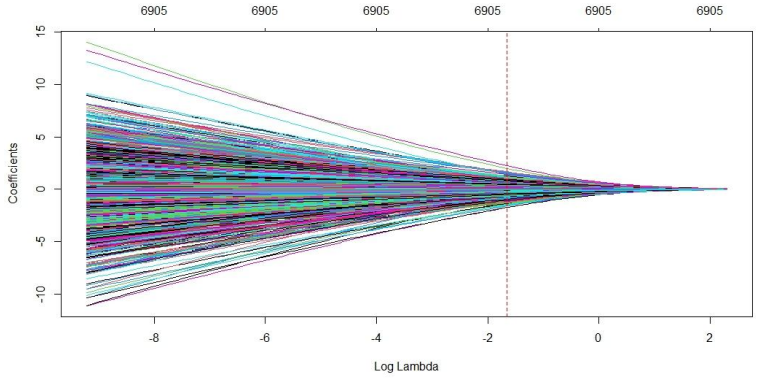
To evaluate the performance of the models we use the Harrell's concordance index. The concordance index assesses the ability of the model to correctly order pairs of individuals with respect to their survival times

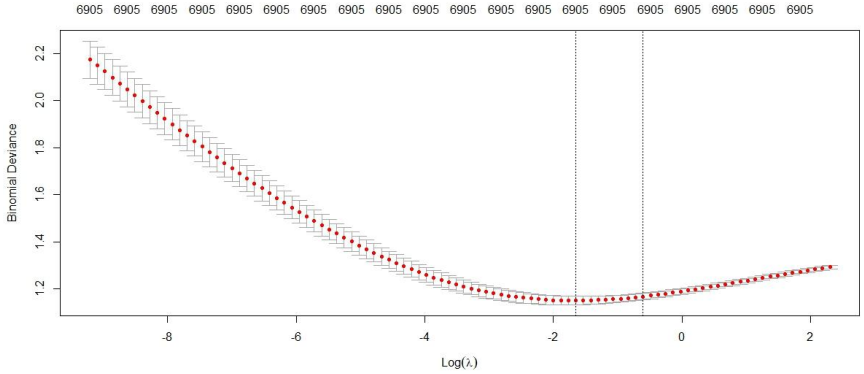
Type of regularization	C-index
Lasso	0.5974989
Elastic Net	0.5992224
Ridge	0.6355165

According to this metric, it appears that the best model is the Cox Hazard proportional model with Ridge regularization.

Logistic Regression and SVM’S

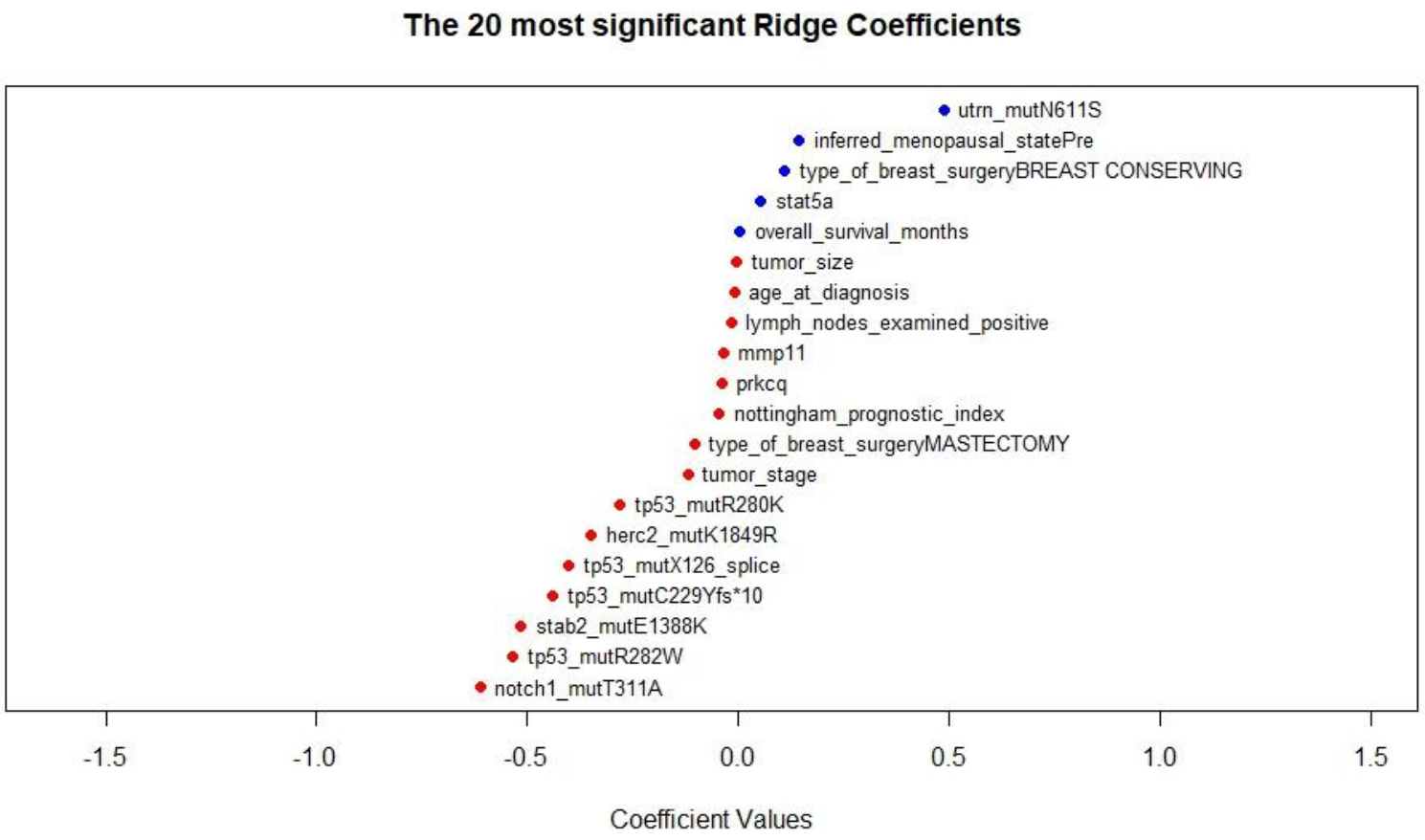
Ridge Regression



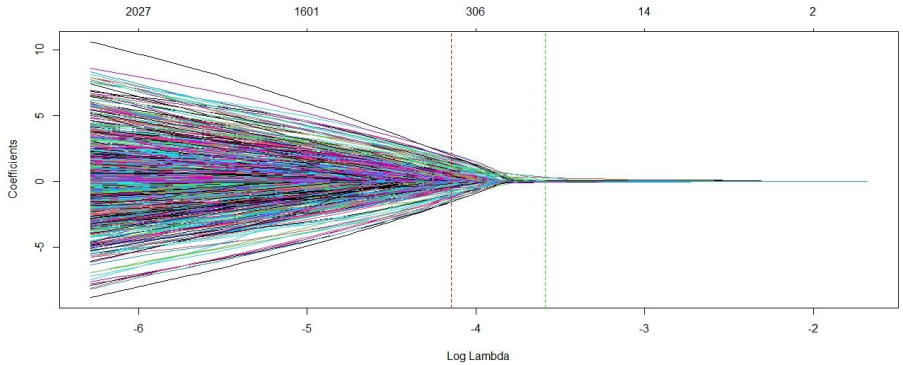


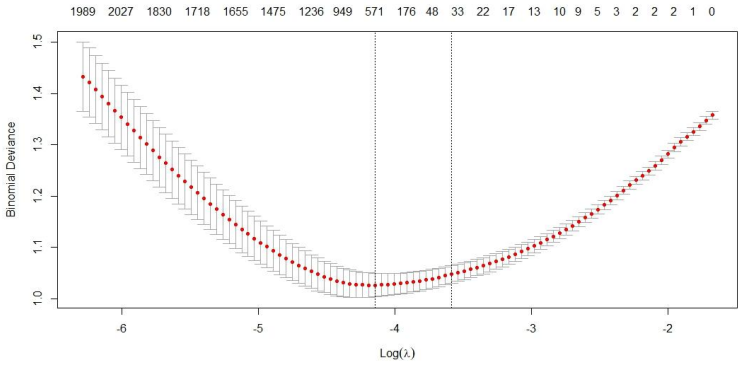
- Accuracy=70.87%

Ridge Regression



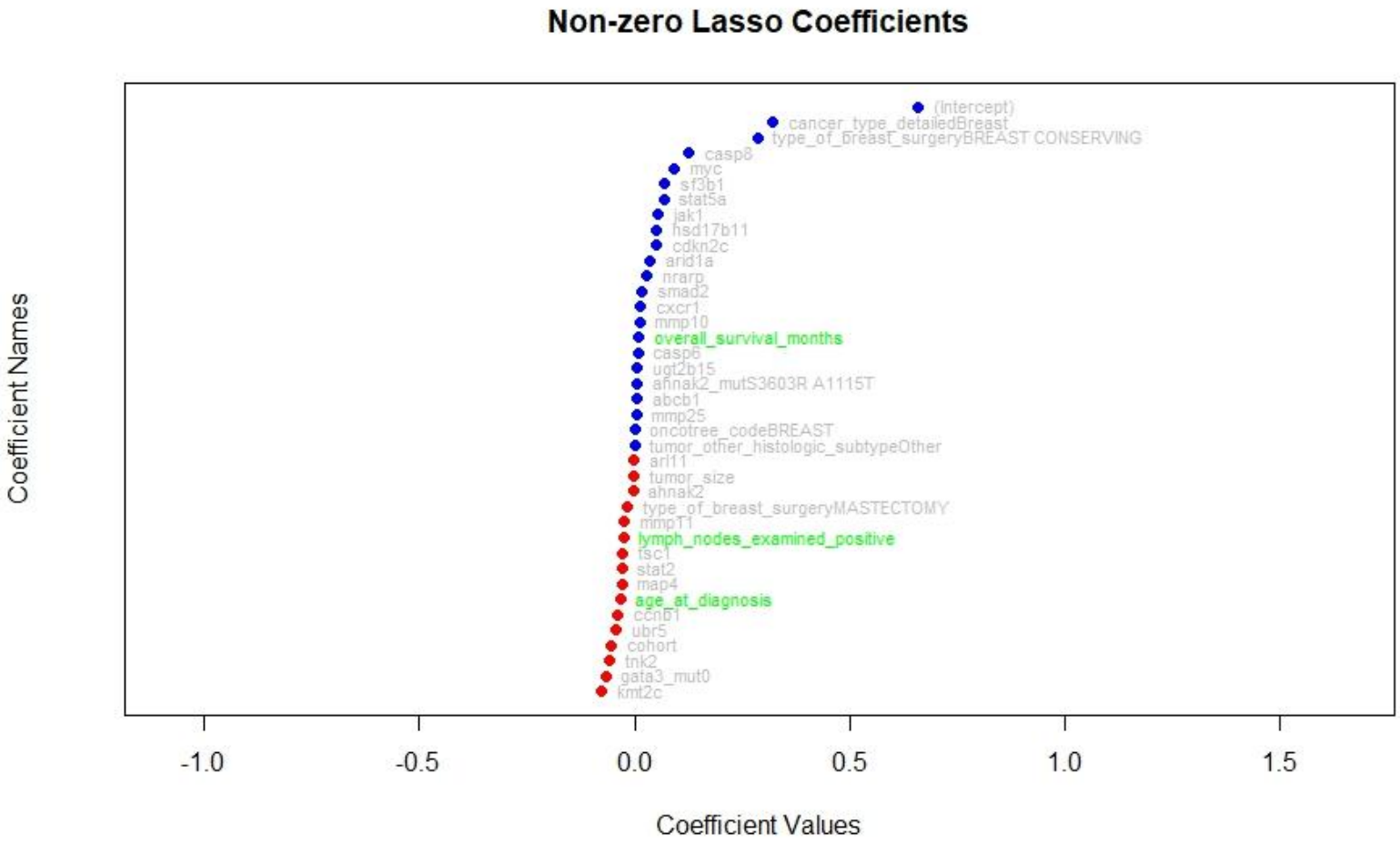
Lasso



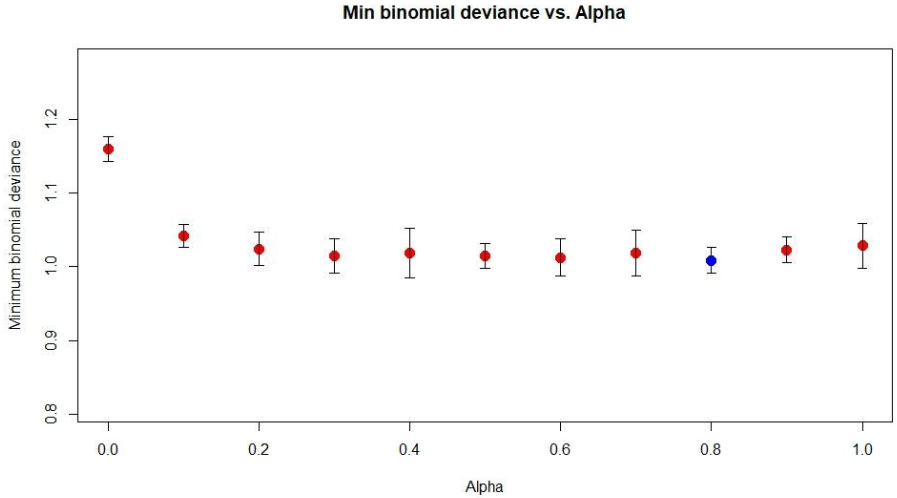


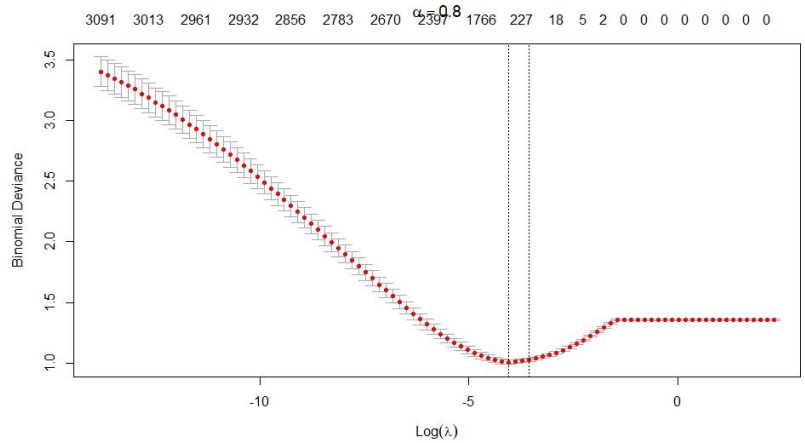
- Accuracy=75.85%
- 38 non-zero coefficients

Lasso


Non-zero Lasso Coefficients
(Intercept)
cancer_type_detailedBreast
type_of_breast_surgeryBREAST CONSERVING
casp8
myc
sf3b1
stat5a
jak1
hsd17b11
cdkn2c
arid1a
nrarp
smad2
cxcr1
mmp10
overall_survival_months
casp6
ugt2b15
ahnak2_mutS3603R A1115T
abcb1
mmp25
oncotree_codeBREAST
tumor_other_histologic_subtypeOther
arl11
tumor_size
ahnak2
type_of_breast_surgeryMASTECTOMY
mmp11
lymph_nodes_examined_positive
tsc1
stat2
map4
age_at_diagnosis
ccnb1
ubr5
cohort
tnk2
gata3_mut0
kmt2c
Coefficient Names
-1.0
-0.5
0.0
0.5
1.0
1.5
Coefficient Values

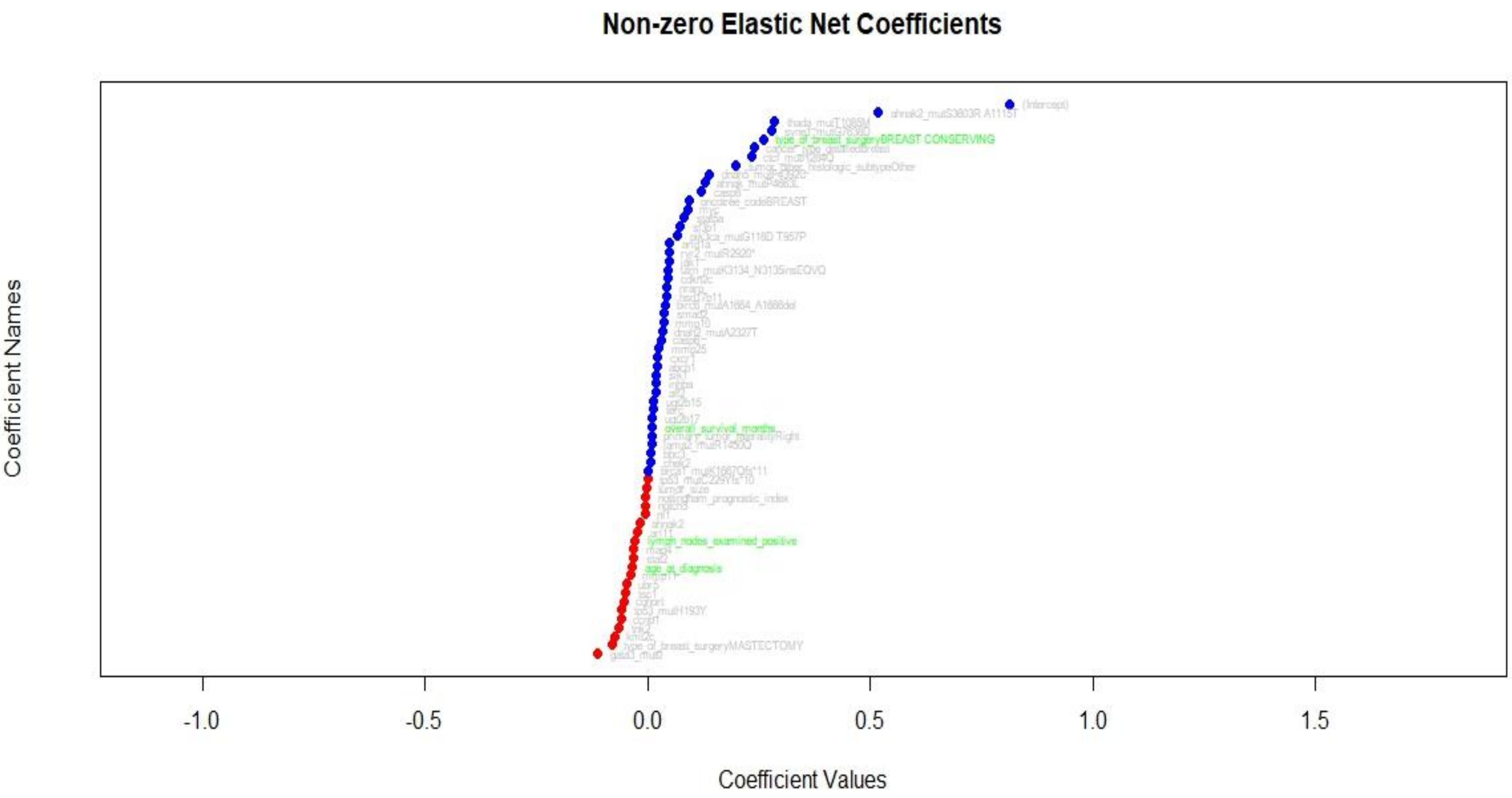
Elastic Net



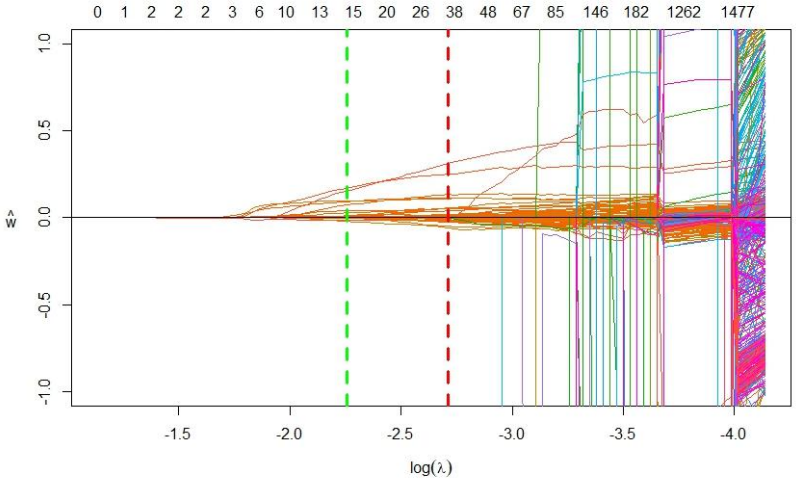


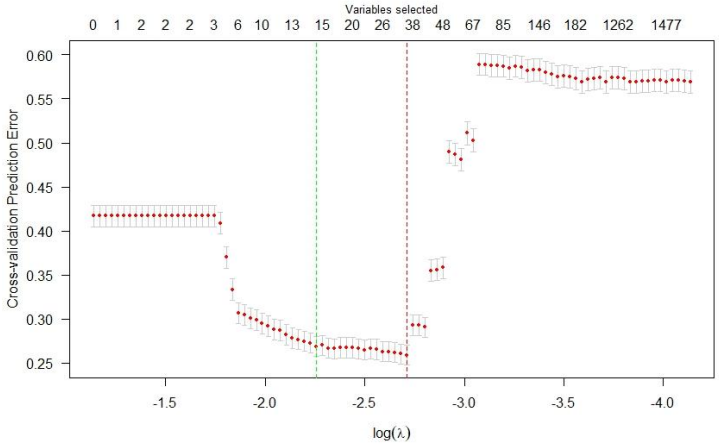
- Accuracy=76.38%
- 64 non-zero coefficients

Elastic Net


Non-zero Elastic Net Coefficients
Coefficient Names
Coefficient Values
-1.0
-0.5
0.0
0.5
1.0
1.5

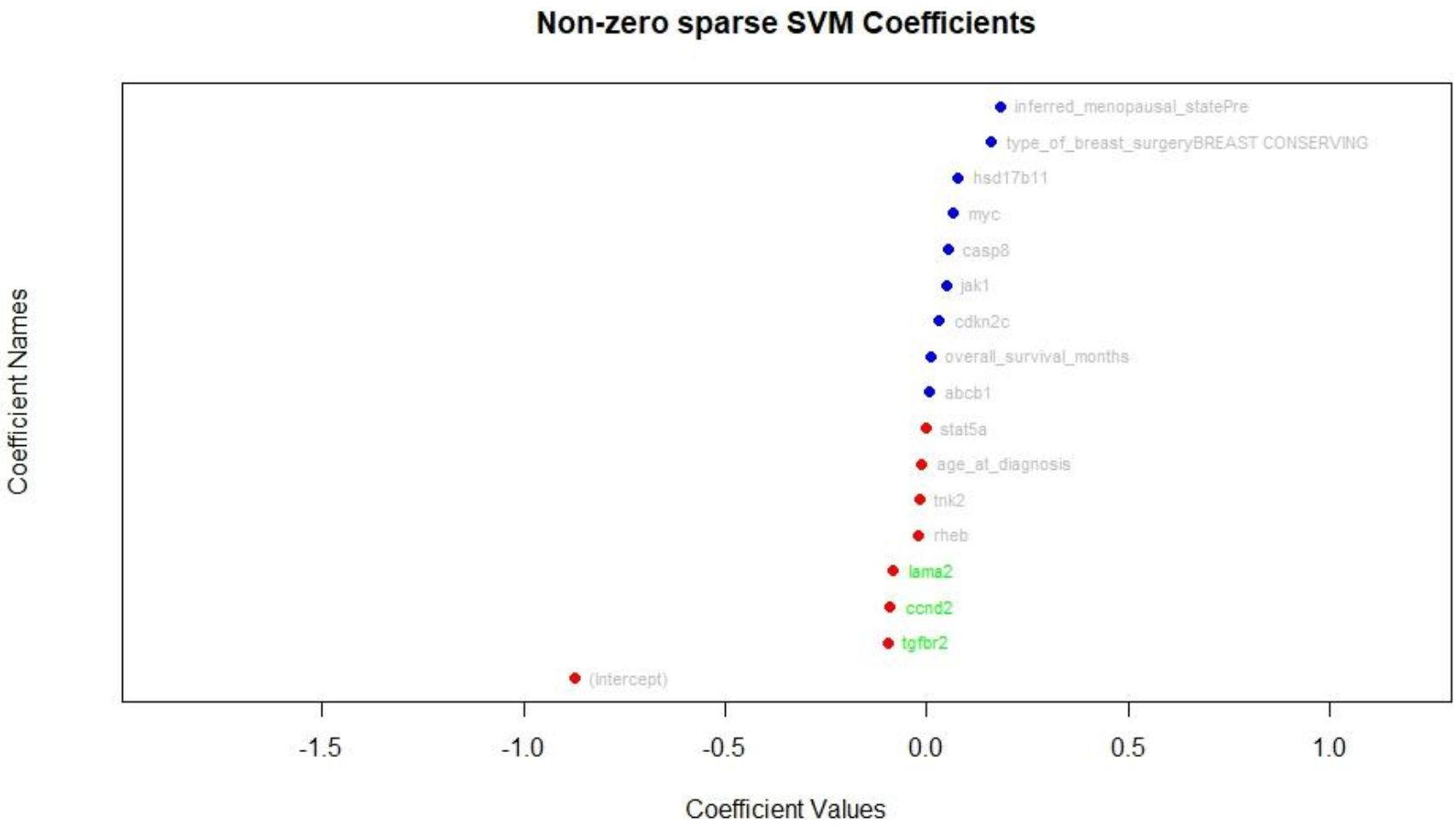
Sparse SVM



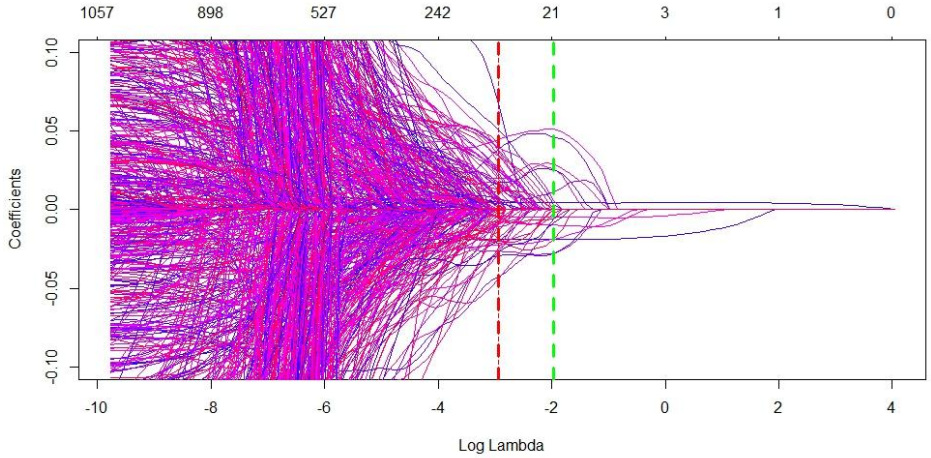


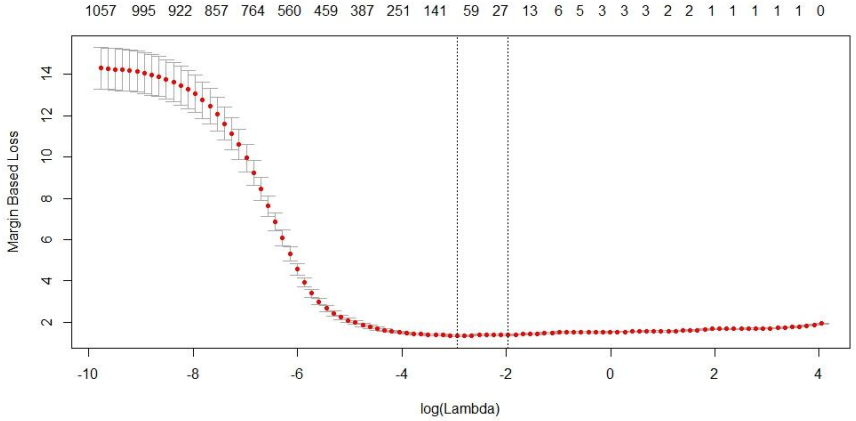
- Accuracy=69.55%
- 17 non-zero coefficients

Sparse SVM


Non-zero sparse SVM Coefficients
Coefficient Names
inferred_menopausal_statePre
type_of_breast_surgeryBREAST CONSERVING
hsd17b11
myc
casp8
jak1
cdkn2c
overall_survival_months
abcb1
stat5a
age_at_diagnosis
tnk2
rheb
lama2
ccnd2
tgfbr2
(Intercept)
-1.5
-1.0
-0.5
0.0
0.5
1.0
Coefficient Values

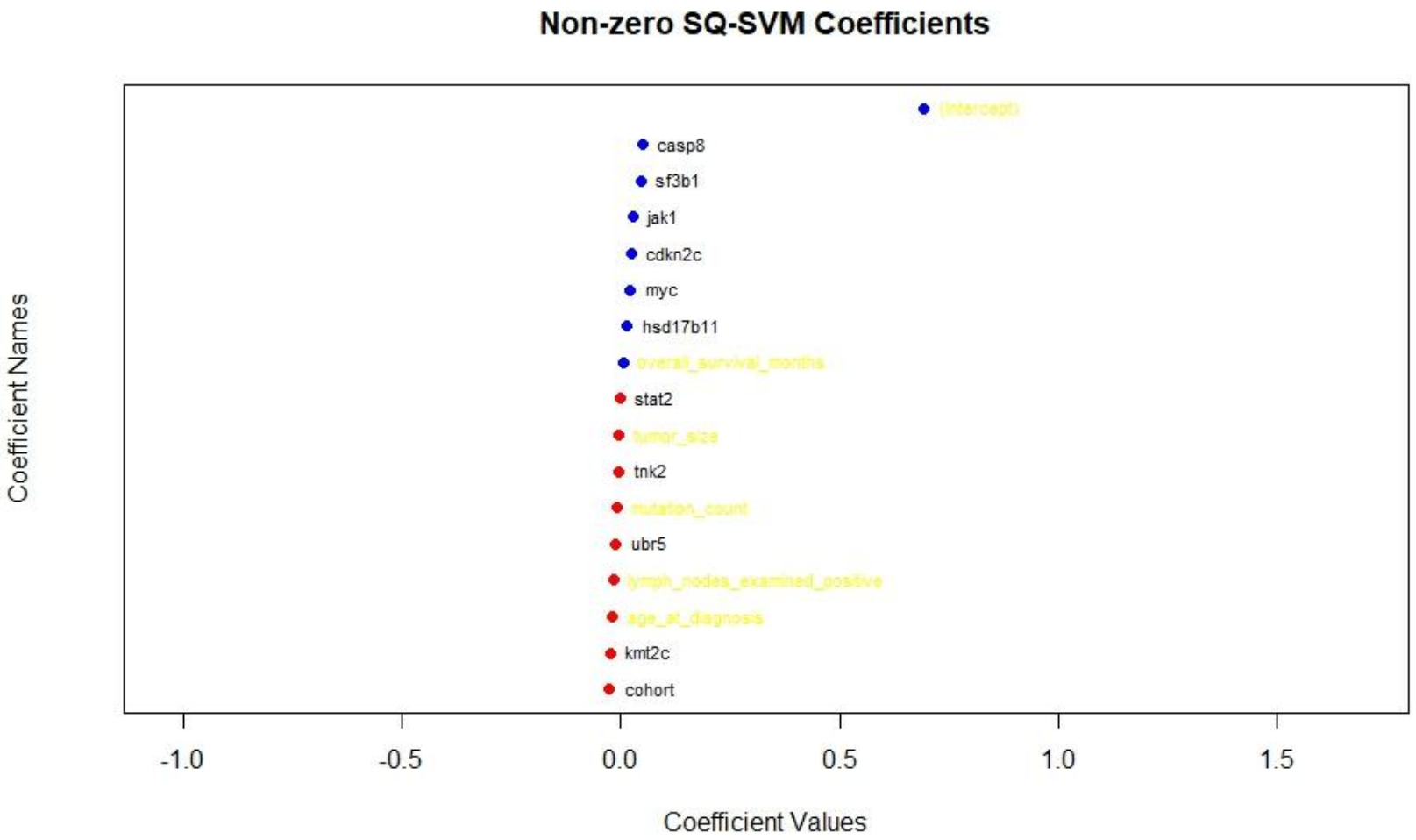
Squared Hinge Loss





- Accuracy=74.8%
- 21 non-zero coefficients

Squared Hinge Loss



Group Lasso

We use three different groupings to create our models:

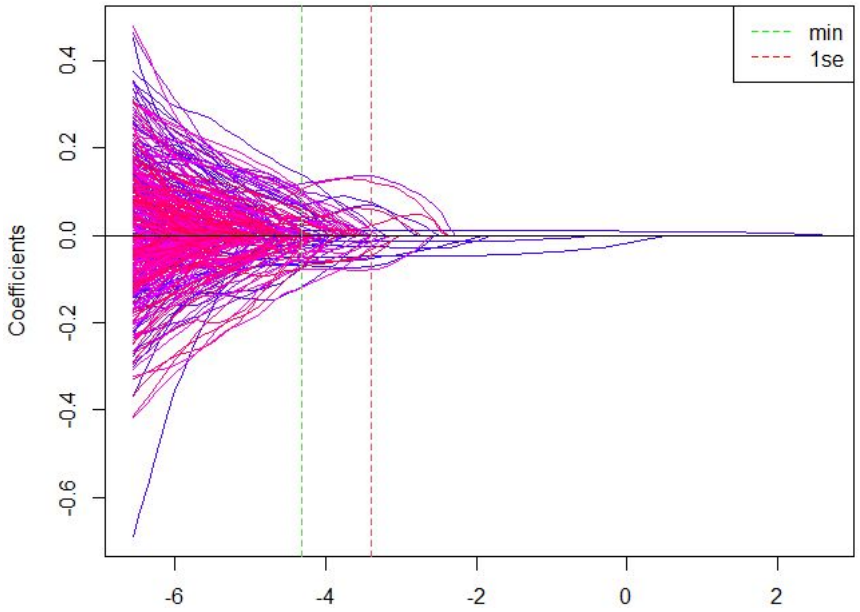
- **Grouping 1:** We consider each variable as a group. For the continuous variable the groups will contain only the variable itself. For the categorical variable, the groups will contain the corresponding indicator variables

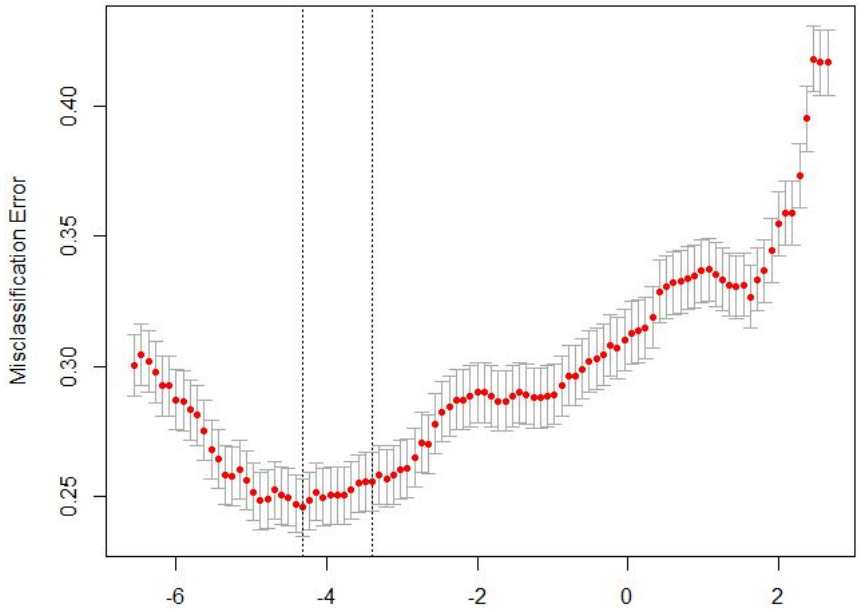
- **Grouping 2:** Same groups as grouping 1 + every variable representing the m-RNA z-scores of a gene is grouped together

- **Grouping 3:** Same groups as grouping 1 + every variable representing a mutated gene is grouped with the variable representing the m-RNA z-scores of the same gene

Group Lasso

Grouping 1:

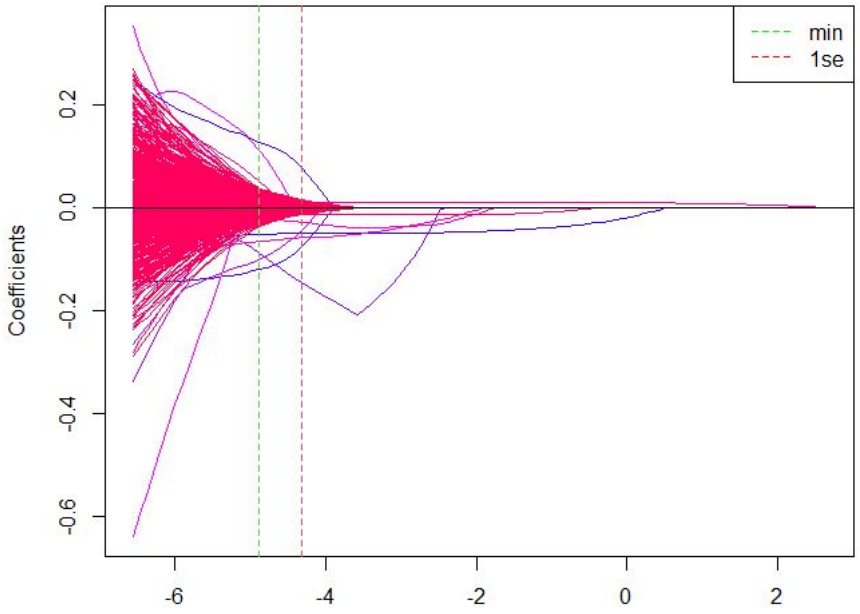


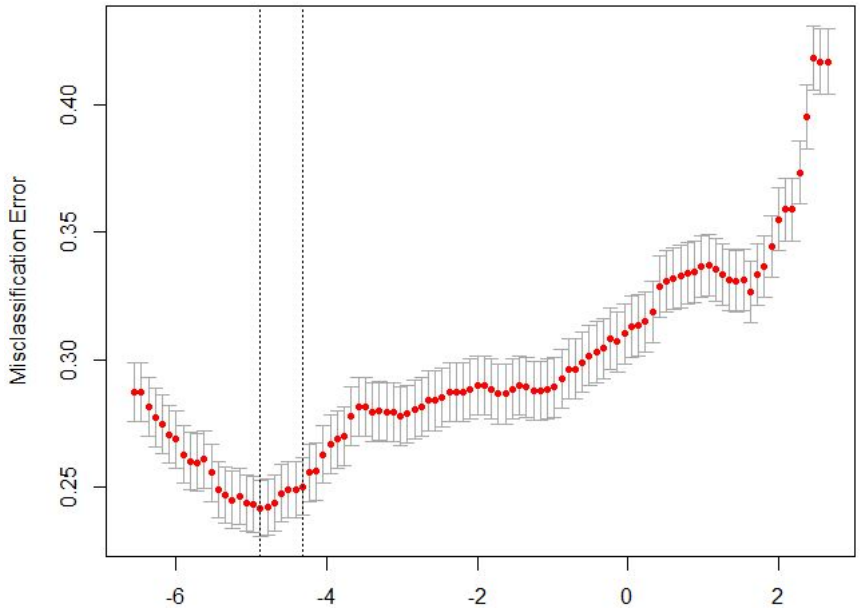


- Accuracy = 74.54%
- Cross-Validation error: 0.2458909
- Non-zero coefficients: 82

Group Lasso

Grouping 2:

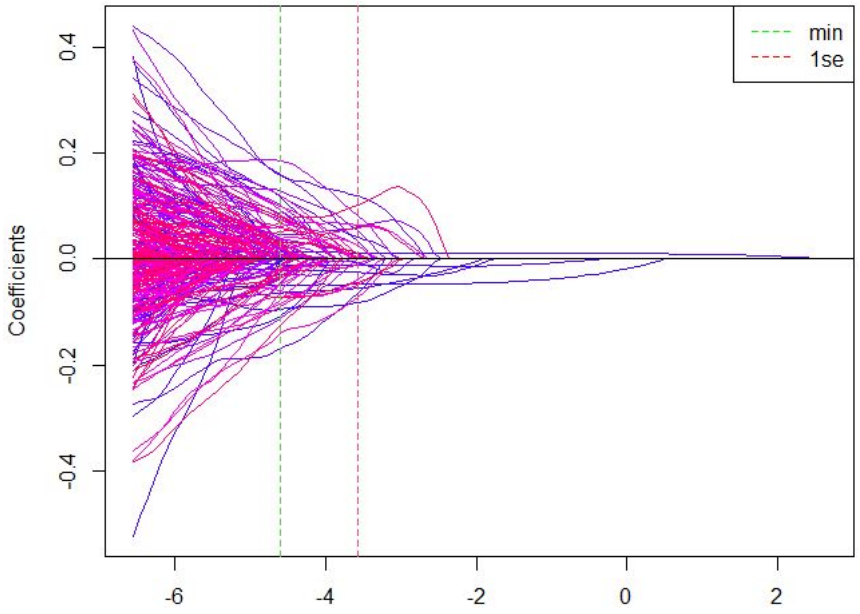


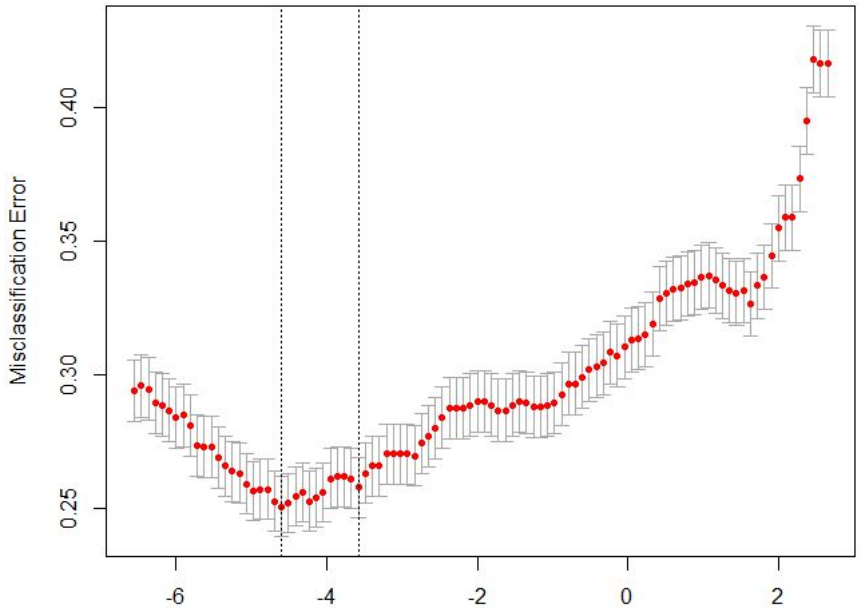


- Accuracy = 76.64%
- Cross-Validation error: 0.2419461
- Non-zero coefficients: 502

Group Lasso

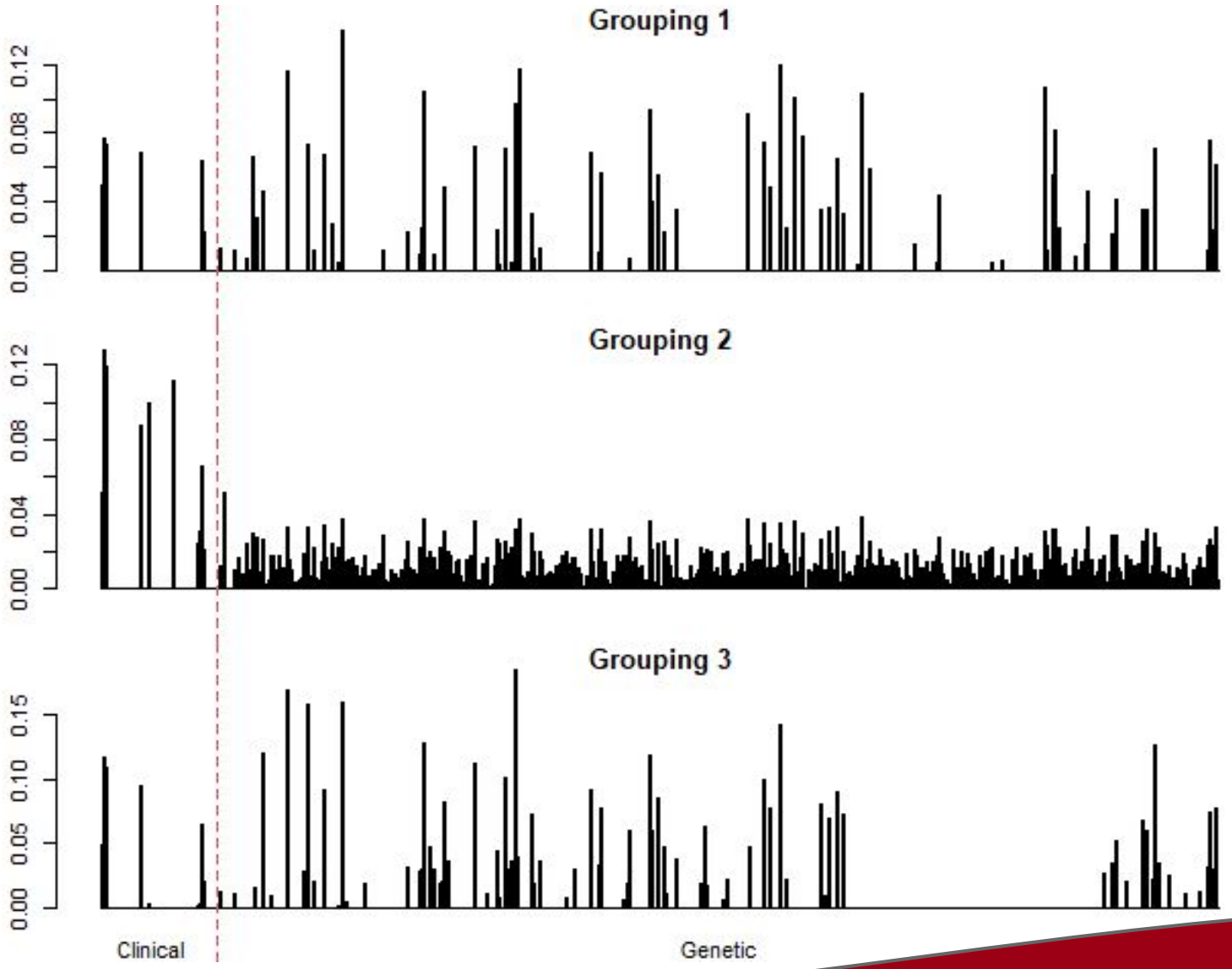
Grouping 3:





- Accuracy = 75.59%
- Cross-Validation error: 0.2504931
- Non-zero coefficients: 92

Difference of the coefficients



Difference of the coefficients

UNIVERSITÀ DEGLI STUDI DI PADOVA

First 10 biggest coefficients:

Grouping 1	Grouping 2	Grouping 3
stat5a	type_of_breast_surgeryBREAST CONSERVING	casp6
tsc1	type_of_breast_surgeryMASTECTOMY	ccnb1
casp8	hormone_therapy	stat5a
ccnb1	neoplasm_histologic_grade	myc
sf3b1	cohort	tsc1
nrarp	lymph_nodes_examined_positive	nrarp
aff2	radio_therapy	hsd17b4
arid1a	age_at_diagnosis	mlh1
casp6	aff2	mmp10
mmp10	casp8	type_of_breast_surgeryBREAST CONSERVING

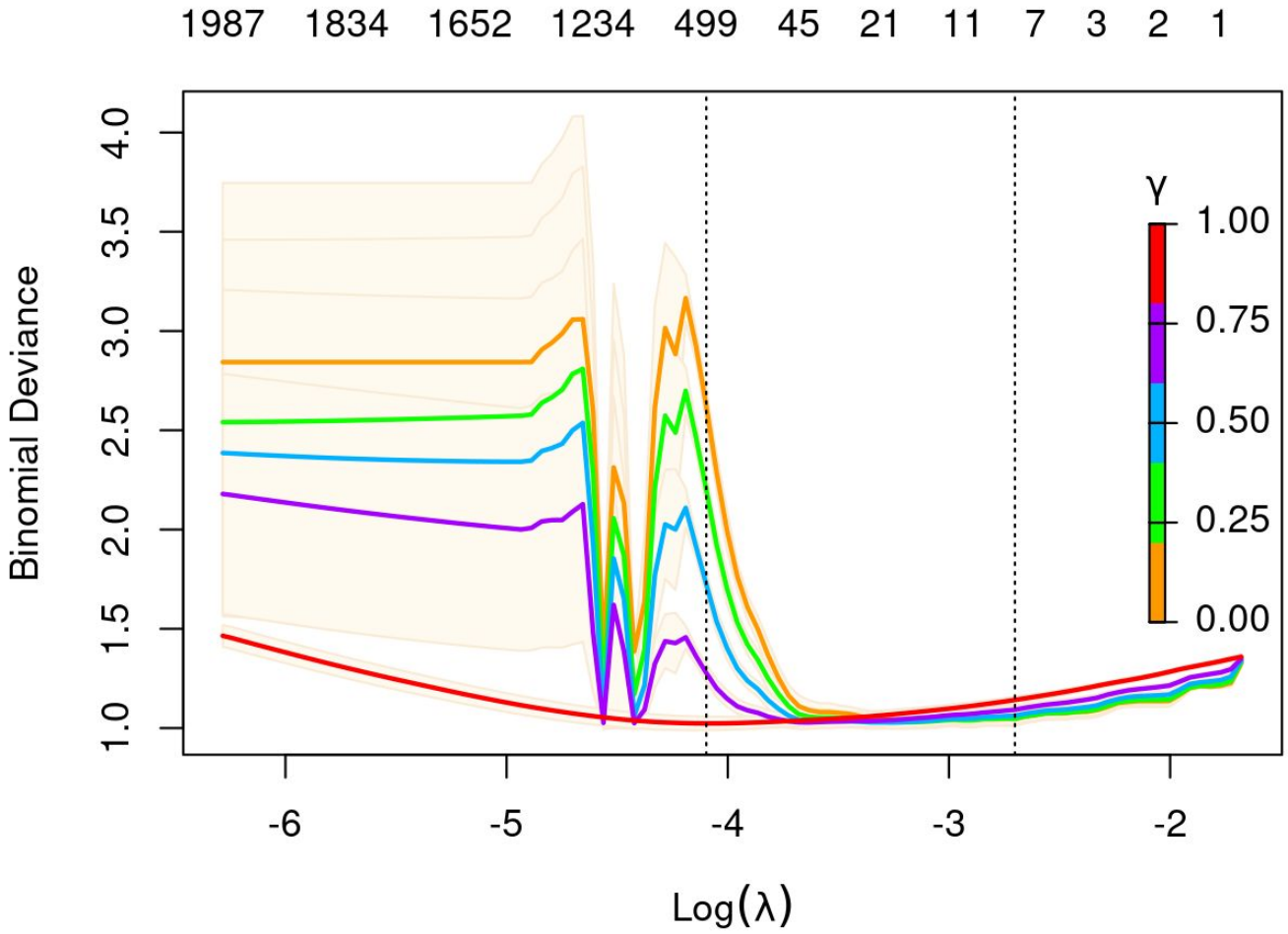
Adaptive Lasso

We applied the two-stage approach using as weights the initial estimates obtained by:

- Ridge
- Lasso
- Elastic Net with alpha = 0.5

Initial Estimates	**Accuracy**	**Non-zero coef**	**C-V Error**
Ridge	43.25%	4137	0.2991453
Lasso	74.54%	245	0.1900066
Elastic Net	76.64%	77	0.2163051

BONUS: Relaxed Lasso



BONUS: Relaxed Lasso (continued)

	Results
Non-zero coeff min (Vanilla Lasso)	499
Non-zero coeff 1SE	10
Accuracy (1SE)	0.7559055

	1SE non-zero coefficients:
1	age_at_diagnosis
2	type_of_breast_surgeryBREAST CONSERVING
3	type_of_breast_surgeryMASTECTOMY
4	overall_survival_months
5	myc
6	jak1
7	casp8
8	kmt2c
9	cdkn2c
10	hsd17b11

Thanks for your attention