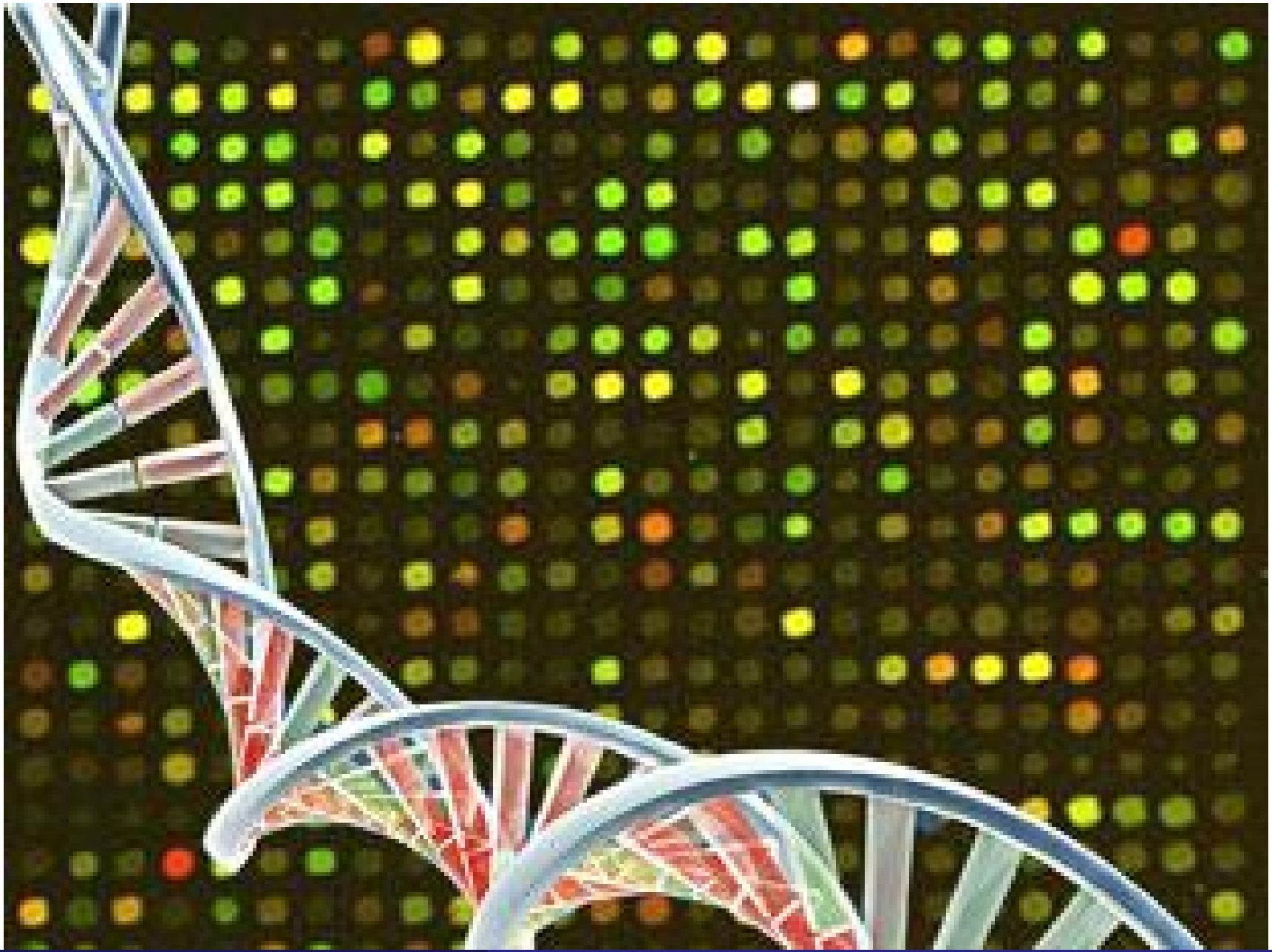
Gene Expression Analysis

W. Zachary Horton and Matthew Oehler

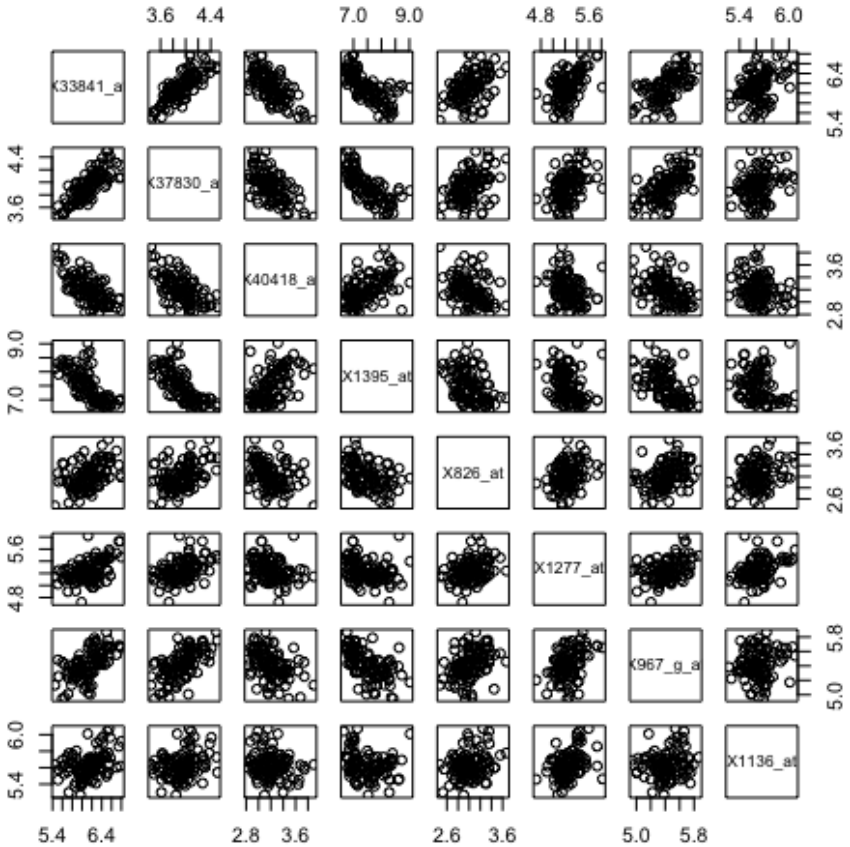
Brigham Young University

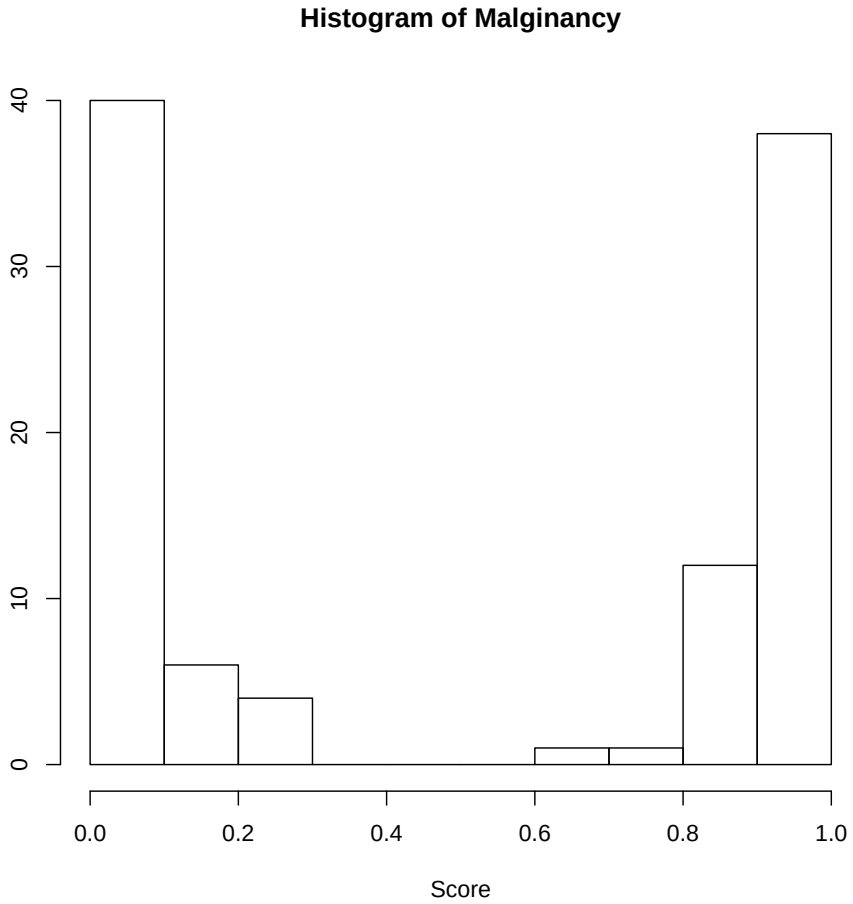
February 16, 2018

Problem Introduction



Cancer Patient Data





Goals

Goals of the Analysis:

- Use the cancer patient data to determine which genes are associated with highly malignant tumors

Foreseeable Problems:

- We have more variables than observations, which may cause problems with common statistical methods used to approach this kind of problem.

Model

LASSO Regression

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

$\mathbf{y}$ = response vector $(nx1)$

$\mathbf{X}$ = model matrix (nxp)

$\boldsymbol{\beta}$ = model coefficients $(px1)$

$\boldsymbol{\epsilon}$ = errors $(nx1)$

Estimated $\hat{\boldsymbol{\beta}}$ minimizes:

$$\min_{\boldsymbol{\beta}} \quad \sum_{i=1}^{n}(\mathbf{y} - \mathbf{x}_i'\boldsymbol{\beta})^2 + \lambda \sum_{j=1}^{p} |\beta_j|$$

where λ is a penalty parameter.

Model Bootstrap

- Use bootstrap to obtain confidence intervals
- Keep same λ over each iteration

Centered 95% Confidence Interval Formula:

$$(2\hat{\beta} - \hat{\beta}^{0.975}_{\text{boot}}, 2\hat{\beta} - \hat{\beta}^{0.025}_{\text{boot}})$$

where $\hat{\beta}$ is the estimated coefficient and $\hat{\beta}^{t}_{\text{boot}}$ is the t percentile of the bootstrap estimates.

Model Choice

Advantages of Regression

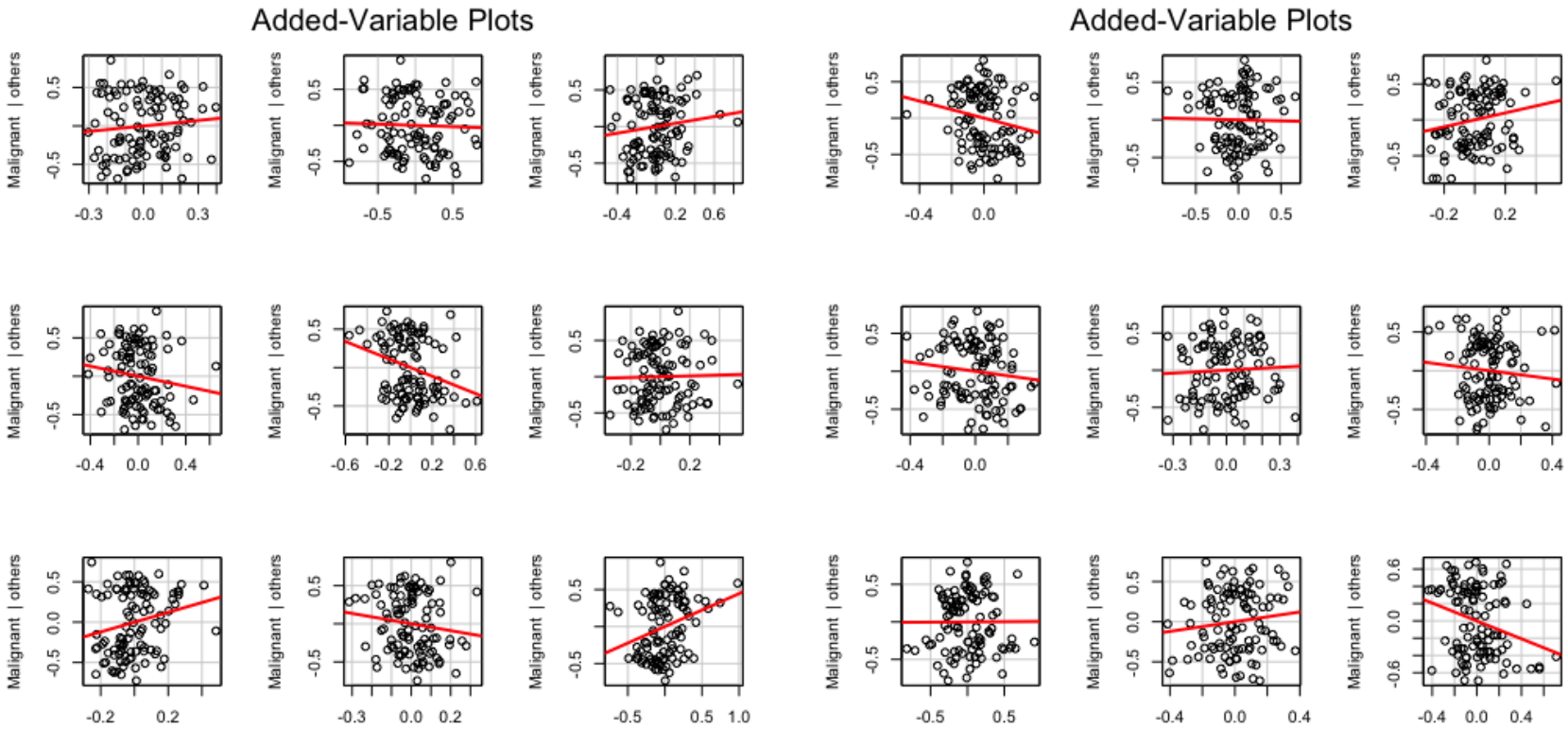
- Estimated coefficients show the effect on tumor malignancy
- Confidence intervals show significance

Why choose LASSO?

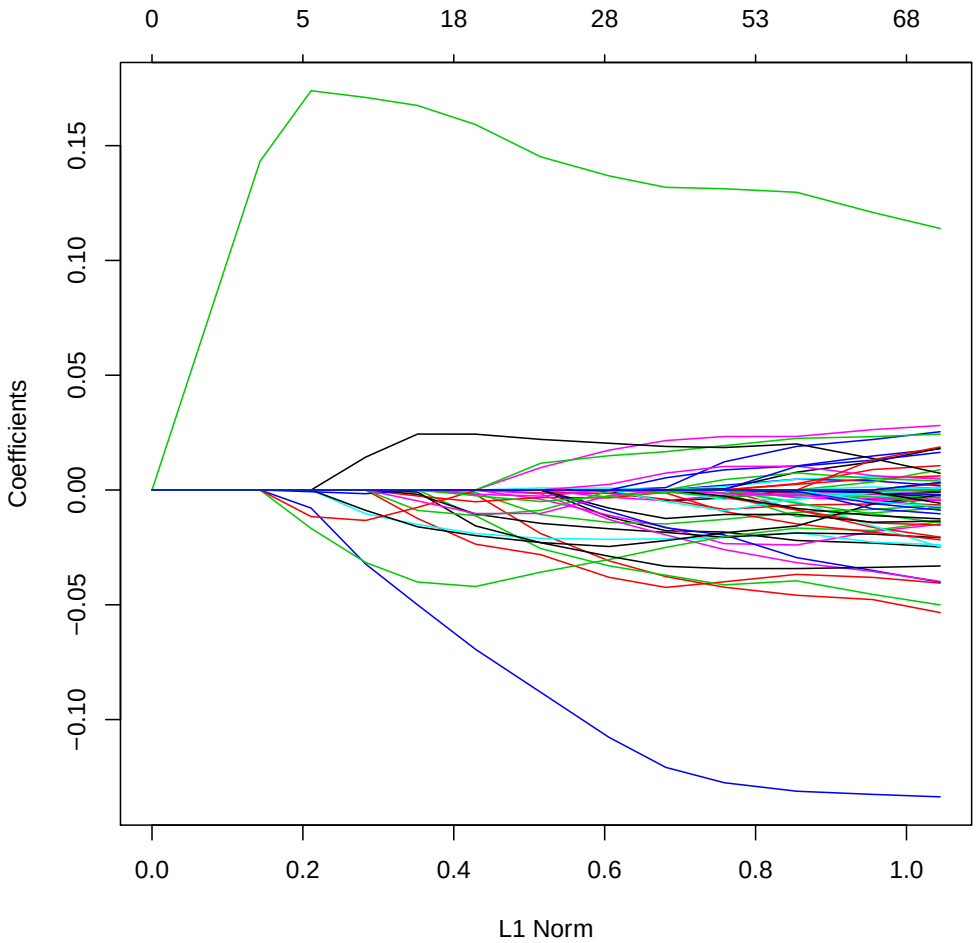
- Too many variables
- Implicit variable selection

Model Assumptions

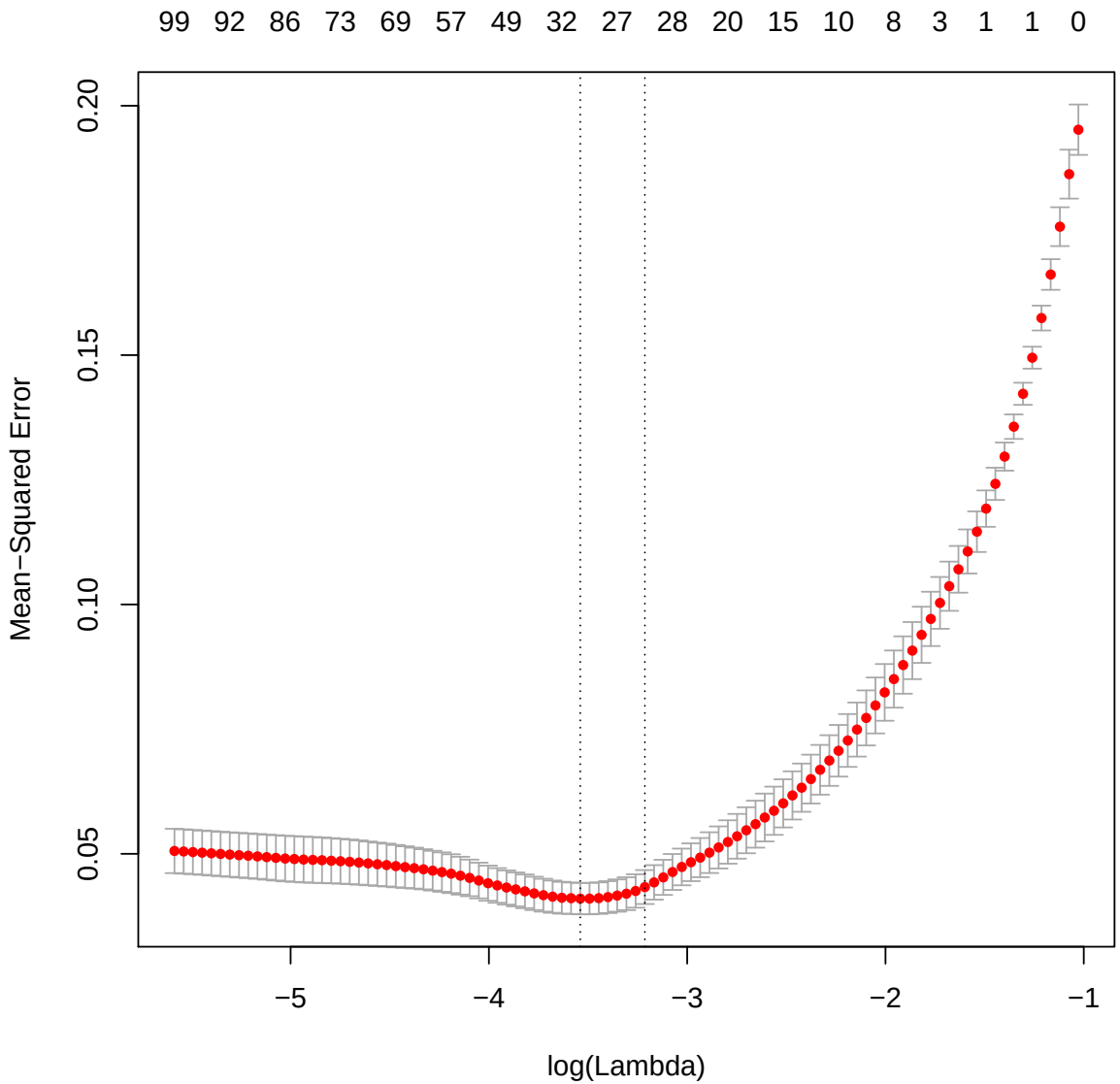
Linearity is the only assumption.
Hard to assess due to excess of covariates.
Random sample will suffice.



LASSO Model


0
5
18
28
53
68
Coefficients
0.15
0.10
0.05
0.00
−0.05
−0.10
0.0
0.2
0.4
0.6
0.8
1.0
L1 Norm

Shrinkage Parameter



Model Performance

- Shrinkage Parameters:

Lambda Values

	Result
Minimum	0.03
1 Std. Error	0.04

- How well did the model do?

Performance Assessment

	Result
MSE	0.04
R-squared	0.89

Results

- Reduced down to 29 genes
- 3 genes were significant
- Variables were standardized

Table of Significant Genes

	CI Lower	Estimate	CI Upper
(Intercept)	0.474	0.511	0.543
X33921_at	-0.063	-0.031	-0.005
X37639_at	0.085	0.139	0.206
X38087_s_at	-0.180	-0.101	-0.065

Results

- Reduced down to 29 genes
- 3 genes were significant
- Variables were standardized

Table of Significant Genes

	CI Lower	Estimate	CI Upper
(Intercept)	0.474	0.511	0.543
X33921_at	-0.063	-0.031	-0.005
X37639_at	0.085	0.139	0.206
X38087_s_at	-0.180	-0.101	-0.065

RESULT: Activating and inhibiting these genes in the right way can significantly influence tumor malignancy

Conclusions

Conclusions

- We found 3 genes that significantly affected malignancy levels
- Shortcomings:
  - Handling collinear variables
- Next steps:
  - Exploring interactions between different genes

Distribution of Work

Problem Statement and Understanding	Matt
Describe the method/model(s) that are used ...	Zach
Model Justification and Performance Evaluation	Matt
Results	Zach
Conclusions	Joint
Code	Simultaneous