

## Supervised Learning: Regression, Part II

Noah Simon & Ali Shojaie

Aug 1-3, 2022  
Summer Institute in Statistics for Big Data  
University of Washington

## Linear Models in High Dimensions

- ▶ When  $p$  is large, least squares regression will lead to very low training error but terrible test error.

## Linear Models in High Dimensions

- ▶ When  $p$  is large, least squares regression will lead to very low training error but terrible test error.
- ▶ We will now see some approaches for fitting linear models in high dimensions,  $p \gg n$ .
- ▶ These approaches also work well when  $p \approx n$  or  $n > p$ .

## Motivating example

- We would like to build a model to predict survival time for breast cancer patients using a number of clinical measurements (tumor stage, tumor grade, tumor size, patient age, etc.) as well as some biomarkers.

## Motivating example

- ▶ We would like to build a model to predict survival time for breast cancer patients using a number of clinical measurements (tumor stage, tumor grade, tumor size, patient age, etc.) as well as some biomarkers.
- ▶ For instance, these biomarkers could be:

## Motivating example

- ▶ We would like to build a model to predict survival time for breast cancer patients using a number of clinical measurements (tumor stage, tumor grade, tumor size, patient age, etc.) as well as some biomarkers.
- ▶ For instance, these biomarkers could be:
  - ▶ the expression levels of genes measured using a microarray.

## Motivating example

- ▶ We would like to build a model to predict survival time for breast cancer patients using a number of clinical measurements (tumor stage, tumor grade, tumor size, patient age, etc.) as well as some biomarkers.
- ▶ For instance, these biomarkers could be:
  - ▶ the expression levels of genes measured using a microarray.
  - ▶ protein levels.

## Motivating example

- ▶ We would like to build a model to predict survival time for breast cancer patients using a number of clinical measurements (tumor stage, tumor grade, tumor size, patient age, etc.) as well as some biomarkers.
- ▶ For instance, these biomarkers could be:
  - ▶ the expression levels of genes measured using a microarray.
  - ▶ protein levels.
  - ▶ mutations in genes potentially implicated in breast cancer.



## Motivating example

- ▶ We would like to build a model to predict survival time for breast cancer patients using a number of clinical measurements (tumor stage, tumor grade, tumor size, patient age, etc.) as well as some biomarkers.
- ▶ For instance, these biomarkers could be:
  - ▶ the expression levels of genes measured using a microarray.
  - ▶ protein levels.
  - ▶ mutations in genes potentially implicated in breast cancer.
- ▶ How can we develop a model with low test error in this setting?

## Remember

- ▶ We have  $n$  training observations.
- ▶ Our goal is to get a model that will perform well on future test observations.
- ▶ We'll incur some bias in order to reduce variance.

## Variable Pre-Selection

The simplest approach for fitting a model in high dimensions:

1. Choose a small set of variables, say the  $q$  variables that are most correlated with the response, where  $q < n$  and  $q < p$ .
2. Use least squares to fit a model predicting  $y$  using only these  $q$  variables.

This approach is simple and straightforward.

## Variable Pre-Selection in R

```
xtr <- matrix(rnorm(100*100),ncol=100)
beta <- c(rep(1,10),rep(0,90))
ytr <- xtr%%beta + rnorm(100)
cors <- cor(xtr,ytr)
whichers <- which(abs(cors)>.2)
mod <- lm(ytr~xtr[,whichers])
print(summary(mod))
```

## How Many Variable to Use?

- We need a way to choose  $q$ , the number of variables used in the regression model.

## How Many Variable to Use?

- ▶ We need a way to choose  $q$ , the number of variables used in the regression model.
- ▶ We want  $q$  that minimizes the test error.

## How Many Variable to Use?

- ▶ We need a way to choose  $q$ , the number of variables used in the regression model.
- ▶ We want  $q$  that minimizes the test error.
- ▶ For a range of values of  $q$ , we can perform the validation set approach, leave-one-out cross-validation, or  $K$ -fold cross-validation in order to estimate the test error.

## How Many Variable to Use?

- ▶ We need a way to choose  $q$ , the number of variables used in the regression model.
- ▶ We want  $q$  that minimizes the test error.
- ▶ For a range of values of  $q$ , we can perform the validation set approach, leave-one-out cross-validation, or  $K$ -fold cross-validation in order to estimate the test error.
- ▶ Then choose the value of  $q$  for which the estimated test error is smallest.



## Estimating the Test Error For a Given $q$

This is the **right** way to estimate the test error using the validation set approach:

1. Split the observations into a training set and a validation set.
2. Using the training set only:
  - a. Identify the  $q$  variables most associated with the response.
  - b. Use least squares to fit a model predicting  $y$  using those  $q$  variables.
  - c. Let  $\hat{\beta}_1, \dots, \hat{\beta}_q$  denote the resulting coefficient estimates.
3. Use  $\hat{\beta}_1, \dots, \hat{\beta}_q$  obtained on training set to predict response on validation set, and compute the validation set MSE.

## Estimating the Test Error For a Given $q$

This is the **wrong** way to estimate the test error using the validation set approach:

1. Identify the  $q$  variables most associated with the response on the full data set.
2. Split the observations into a training set and a validation set.
3. Using the training set only:
  - a. Use least squares to fit a model predicting  $y$  using those  $q$  variables.
  - b. Let  $\hat{\beta}_1, \dots, \hat{\beta}_q$  denote the resulting coefficient estimates.
4. Use  $\hat{\beta}_1, \dots, \hat{\beta}_q$  obtained on training set to predict response on validation set, and compute the validation set MSE.

## Frequently Asked Questions

- ▶ **Q:** Does it really matter how you estimate the test error?  
**A:** Yes.

## Frequently Asked Questions

- ▶ **Q:** Does it really matter how you estimate the test error?  
**A:** Yes.
- ▶ **Q:** Would anyone make such a silly mistake?  
**A:** Yes.

## A Better Approach

- ▶ The variable pre-selection approach is simple and easy to implement — all you need is a way to calculate correlations, and software to fit a linear model using least squares.

## A Better Approach

- ▶ The variable pre-selection approach is simple and easy to implement — all you need is a way to calculate correlations, and software to fit a linear model using least squares.
- ▶ But it might not work well: just because a bunch of variables are correlated with the response doesn't mean that when used together in a linear model, they will predict the response well.

## A Better Approach

- ▶ The variable pre-selection approach is simple and easy to implement — all you need is a way to calculate correlations, and software to fit a linear model using least squares.
- ▶ But it might not work well: just because a bunch of variables are correlated with the response doesn't mean that when used together in a linear model, they will predict the response well.
- ▶ What we really want to do: pick the  $q$  variables that best predict the response.

## A Better Approach

- ▶ The variable pre-selection approach is simple and easy to implement — all you need is a way to calculate correlations, and software to fit a linear model using least squares.
- ▶ But it might not work well: just because a bunch of variables are correlated with the response doesn't mean that when used together in a linear model, they will predict the response well.
- ▶ What we really want to do: pick the  $q$  variables that best predict the response.
- ▶ Many methods have been developed to achieve this over the past 10-20 years! We cover few of them in this module.



## Best Subset Selection

- Ideally, we would like to consider all possible models using a subset of the  $p$  predictors.

## Best Subset Selection

- ▶ Ideally, we would like to consider all possible models using a subset of the  $p$  predictors.
- ▶ In other words, we'd like to consider all  $2^p$  possible models.

## Best Subset Selection

- ▶ Ideally, we would like to consider all possible models using a subset of the  $p$  predictors.
- ▶ In other words, we'd like to consider all  $2^p$  possible models.
- ▶ This is called **best subset selection**.

## Best Subset Selection

- ▶ Ideally, we would like to consider all possible models using a subset of the  $p$  predictors.
- ▶ In other words, we'd like to consider all  $2^p$  possible models.
- ▶ This is called **best subset selection**.
- ▶ Unfortunately, this is computationally intractable:
  - ▶ When  $p = 3$ ,  $2^p = 8$ .
  - ▶ When  $p = 6$ ,  $2^p = 64$ .
  - ▶ When  $p = 250$ , there are  $2^{250} \approx 10^{80}$  possible models. According to [www.universetoday.com](http://www.universetoday.com), this is around the number of atoms in the known universe.
  - ▶ **Not feasible to consider so many models!**

## Best Subset Selection

- ▶ Ideally, we would like to consider all possible models using a subset of the  $p$  predictors.
- ▶ In other words, we'd like to consider all  $2^p$  possible models.
- ▶ This is called **best subset selection**.
- ▶ Unfortunately, this is computationally intractable:
  - ▶ When  $p = 3$ ,  $2^p = 8$ .
  - ▶ When  $p = 6$ ,  $2^p = 64$ .
  - ▶ When  $p = 250$ , there are  $2^{250} \approx 10^{80}$  possible models. According to [www.universetoday.com](http://www.universetoday.com), this is around the number of atoms in the known universe.
  - ▶ **Not feasible to consider so many models!**

## Ridge Regression and the Lasso

- ▶ Best subset selection does a discrete search through model space, considering subsets of the predictors, and fitting each of the resulting models using least squares. Model complexity is controlled by using subsets of the predictors.

## Ridge Regression and the Lasso

- ▶ Best subset selection does a discrete search through model space, considering subsets of the predictors, and fitting each of the resulting models using least squares. Model complexity is controlled by using subsets of the predictors.
- ▶ Ridge regression and the lasso instead control model complexity by using an alternative to least squares, by shrinking the regression coefficients.

## Ridge Regression and the Lasso

- ▶ Best subset selection does a discrete search through model space, considering subsets of the predictors, and fitting each of the resulting models using least squares. Model complexity is controlled by using subsets of the predictors.
- ▶ Ridge regression and the lasso instead control model complexity by using an alternative to least squares, by shrinking the regression coefficients.
- ▶ This is known as regularization or penalization.



## Crazy Coefficients

- ▶ When  $p > n$ , some of the variables are highly correlated.

## Crazy Coefficients

- ▶ When  $p > n$ , some of the variables are **highly correlated**.
- ▶ Why does correlation matter?
  - ▶ Suppose that  $X_1$  and  $X_2$  are highly correlated with each other... assume  $X_1 = X_2$  for the sake of argument.
  - ▶ And suppose that the least squares model is

$$\hat{y} = X_1 - 2X_2 + 3X_3.$$

- ▶ Then this is **also** a least squares model:

$$\hat{y} = 100000001X_1 - 100000002X_2 + 3X_3.$$

## Crazy Coefficients

- ▶ When  $p > n$ , some of the variables are **highly correlated**.
- ▶ Why does correlation matter?
  - ▶ Suppose that  $X_1$  and  $X_2$  are highly correlated with each other... assume  $X_1 = X_2$  for the sake of argument.
  - ▶ And suppose that the least squares model is

$$\hat{y} = X_1 - 2X_2 + 3X_3.$$

- ▶ Then this is **also** a least squares model:

$$\hat{y} = 100000001X_1 - 100000002X_2 + 3X_3.$$

- ▶ **Bottom Line:** When there are too many variables, the least squares coefficients can get crazy!
- ▶ This craziness is **directly responsible for poor test error**.
- ▶ It amounts to **too much model complexity**.

## A Solution: Don't Let the Coefficients Get Too Crazy

- Recall that least squares involves finding  $\beta$  that minimizes

$$\|\mathbf{y} - \mathbf{X}\beta\|^2.$$

## A Solution: Don't Let the Coefficients Get Too Crazy

- Recall that least squares involves finding  $\beta$  that minimizes

$$\|\mathbf{y} - \mathbf{X}\beta\|^2.$$

- Ridge regression involves finding  $\beta$  that minimizes

$$\|\mathbf{y} - \mathbf{X}\beta\|^2 + \lambda \sum_j \beta_j^2.$$

## A Solution: Don't Let the Coefficients Get Too Crazy

- Recall that least squares involves finding  $\beta$  that minimizes

$$\|\mathbf{y} - \mathbf{X}\beta\|^2.$$

- Ridge regression involves finding  $\beta$  that minimizes

$$\|\mathbf{y} - \mathbf{X}\beta\|^2 + \lambda \sum_j \beta_j^2.$$

- Equivalently, find  $\beta$  that minimizes

$$\|\mathbf{y} - \mathbf{X}\beta\|^2$$

subject to the constraint that

$$\sum_{j=1}^p \beta_j^2 \leq s.$$

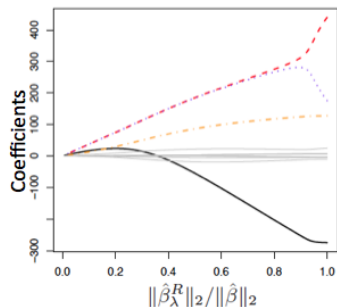
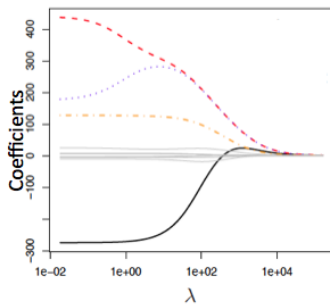
## Ridge Regression

- ▶ Ridge regression coefficient estimates minimize

$$\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2 + \lambda \sum_j \beta_j^2.$$

- ▶ Here  $\lambda$  is a nonnegative **tuning parameter** that shrinks the coefficient estimates.
- ▶ When  $\lambda = 0$ , then ridge regression is just the same as least squares.
- ▶ As  $\lambda$  increases, then  $\sum_{j=1}^p (\hat{\beta}_{\lambda,j}^R)^2$  decreases — i.e. coefficients become shrunken towards zero.
- ▶ When  $\lambda = \infty$ ,  $\hat{\boldsymbol{\beta}}_{\lambda}^R = \mathbf{0}$ .

# Ridge Regression As $\lambda$ Varies





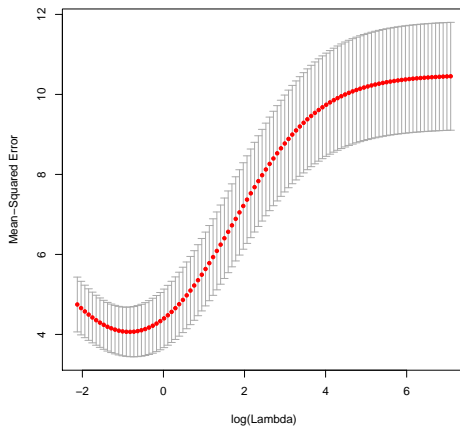
## Ridge Regression In Practice

- ▶ Perform ridge regression for a very fine grid of  $\lambda$  values.
- ▶ Use cross-validation or the validation set approach to select the optimal value of  $\lambda$  — that is, the best level of model complexity.
- ▶ Perform ridge on the full data set, using that value of  $\lambda$ .

## Example in R

```
xtr <- matrix(rnorm(100*100),ncol=100)
beta <- c(rep(1,10),rep(0,90))
ytr <- xtr%%beta + rnorm(100)
library(glmnet)
cv.out <- cv.glmnet(xtr,ytr,alpha=0,nfolds=5)
print(cv.out$cvm)
plot(cv.out)
cat("CV Errors", cv.out$cvm,fill=TRUE)
cat("Lambda with smallest CV Error",
cv.out$lambda[which.min(cv.out$cvm)],fill=TRUE)
cat("Coefficients", as.numeric(coef(cv.out)),fill=TRUE)
cat("Number of Zero Coefficients",
sum(abs(coef(cv.out))<1e-8),fill=TRUE)
```

## R Output



## Drawbacks of Ridge

- Ridge regression is a simple idea and has a number of attractive properties: for instance, you can continuously control model complexity through the tuning parameter  $\lambda$ .

## Drawbacks of Ridge

- ▶ Ridge regression is a simple idea and has a number of attractive properties: for instance, you can continuously control model complexity through the tuning parameter  $\lambda$ .
- ▶ But it suffers in terms of model interpretability, since the final model contains **all  $p$  variables, no matter what**.

## Drawbacks of Ridge

- ▶ Ridge regression is a simple idea and has a number of attractive properties: for instance, you can continuously control model complexity through the tuning parameter  $\lambda$ .
- ▶ But it suffers in terms of model interpretability, since the final model contains **all  $p$  variables, no matter what**.
- ▶ Often want a simpler model involving a subset of the features.

## Drawbacks of Ridge

- ▶ Ridge regression is a simple idea and has a number of attractive properties: for instance, you can continuously control model complexity through the tuning parameter  $\lambda$ .
- ▶ But it suffers in terms of model interpretability, since the final model contains **all  $p$  variables, no matter what**.
- ▶ Often want a simpler model involving a subset of the features.
- ▶ **The lasso** involves performing a little tweak to ridge regression so that the resulting model contains **mostly zeros**.

## Drawbacks of Ridge

- ▶ Ridge regression is a simple idea and has a number of attractive properties: for instance, you can continuously control model complexity through the tuning parameter  $\lambda$ .
- ▶ But it suffers in terms of model interpretability, since the final model contains **all  $p$  variables, no matter what**.
- ▶ Often want a simpler model involving a subset of the features.
- ▶ **The lasso** involves performing a little tweak to ridge regression so that the resulting model contains **mostly zeros**.
- ▶ In other words, the resulting model is **sparse**. We say that the lasso performs **feature selection**.



## Drawbacks of Ridge

- ▶ Ridge regression is a simple idea and has a number of attractive properties: for instance, you can continuously control model complexity through the tuning parameter  $\lambda$ .
- ▶ But it suffers in terms of model interpretability, since the final model contains **all  $p$  variables, no matter what**.
- ▶ Often want a simpler model involving a subset of the features.
- ▶ **The lasso** involves performing a little tweak to ridge regression so that the resulting model contains **mostly zeros**.
- ▶ In other words, the resulting model is **sparse**. We say that the lasso performs **feature selection**.
- ▶ The lasso is a very active area of research interest in the statistical community!

## The Lasso

- The lasso involves finding  $\beta$  that minimizes

$$\|\mathbf{y} - \mathbf{X}\beta\|^2 + \lambda \sum_j |\beta_j|.$$

## The Lasso

- The lasso involves finding  $\beta$  that minimizes

$$\|\mathbf{y} - \mathbf{X}\beta\|^2 + \lambda \sum_j |\beta_j|.$$

- Equivalently, find  $\beta$  that minimizes

$$\|\mathbf{y} - \mathbf{X}\beta\|^2$$

subject to the constraint that

$$\sum_{j=1}^p |\beta_j| \leq s.$$

## The Lasso

- The lasso involves finding  $\beta$  that minimizes

$$\|\mathbf{y} - \mathbf{X}\beta\|^2 + \lambda \sum_j |\beta_j|.$$

- Equivalently, find  $\beta$  that minimizes

$$\|\mathbf{y} - \mathbf{X}\beta\|^2$$

subject to the constraint that

$$\sum_{j=1}^p |\beta_j| \leq s.$$

- So lasso is just like ridge, except that  $\beta_j^2$  has been replaced with  $|\beta_j|$ .

# The Lasso

- ▶ Lasso is a lot like ridge:

## The Lasso

- ▶ Lasso is a lot like ridge:
  - ▶  $\lambda$  is a nonnegative tuning parameter that controls model complexity.

## The Lasso

- ▶ Lasso is a lot like ridge:
  - ▶  $\lambda$  is a nonnegative tuning parameter that controls model complexity.
  - ▶ When  $\lambda = 0$ , we get least squares.

## The Lasso

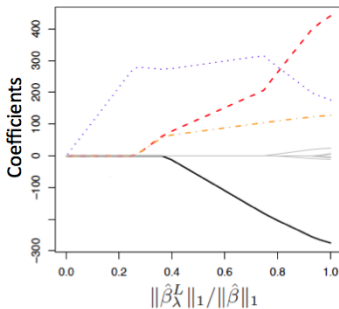
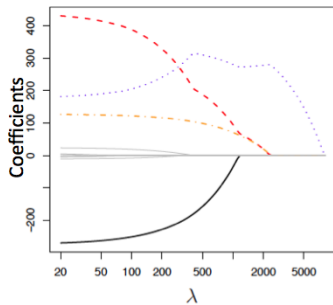
- ▶ Lasso is a lot like ridge:
  - ▶  $\lambda$  is a nonnegative tuning parameter that controls model complexity.
  - ▶ When  $\lambda = 0$ , we get least squares.
  - ▶ When  $\lambda$  is very large, we get  $\hat{\beta}_{\lambda}^L = 0$ .



## The Lasso

- ▶ Lasso is a lot like ridge:
  - ▶  $\lambda$  is a nonnegative tuning parameter that controls model complexity.
  - ▶ When  $\lambda = 0$ , we get least squares.
  - ▶ When  $\lambda$  is very large, we get  $\hat{\beta}_{\lambda}^L = 0$ .
- ▶ But unlike ridge, lasso will give some coefficients exactly equal to zero for intermediate values of  $\lambda$ !

# Lasso As $\lambda$ Varies



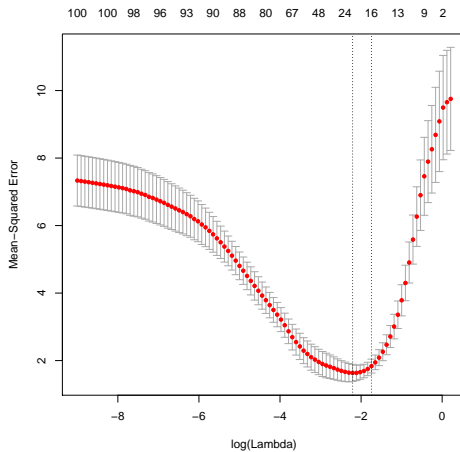
## Lasso In Practice

- ▶ Perform lasso for a very fine grid of  $\lambda$  values.
- ▶ Use cross-validation or the validation set approach to select the optimal value of  $\lambda$  — that is, the best level of model complexity.
- ▶ Perform the lasso on the full data set, using that value of  $\lambda$ .

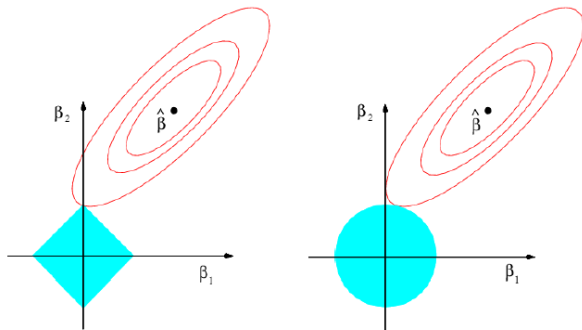
## Example in R

```
xtr <- matrix(rnorm(100*100),ncol=100)
beta <- c(rep(1,10),rep(0,90))
ytr <- xtr%%beta + rnorm(100)
library(glmnet)
cv.out <- cv.glmnet(xtr,ytr,alpha=1,nfolds=5)
print(cv.out$cvm)
plot(cv.out)
cat("CV Errors", cv.out$cvm,fill=TRUE)
cat("Lambda with smallest CV Error",
cv.out$lambda[which.min(cv.out$cvm)],fill=TRUE)
cat("Coefficients", as.numeric(coef(cv.out)),fill=TRUE)
cat("Number of Zero Coefficients",sum(abs(coef(cv.out))<1e-8),
fill=TRUE)
```

## R Output



## Ridge and Lasso: A Geometric Interpretation



Let's Try It Out in R!

# Chapter 6 R Lab, Part 2

[www.statlearning.com](http://www.statlearning.com)

## Pros/Cons of Each Approach

Approach	Simplicity?*	Sparsity?**	Predictions?***
Pre-Selection	Good	Yes	So-So
Ridge	Medium	No	Great
Lasso	Bad	Yes	Great

\* How simple is this model-fitting procedure? If you were stranded on a desert island with pretty limited statistical software, could you fit this model?

\*\* Does this approach perform feature selection, i.e. is the resulting model sparse?

\*\*\* How good are the predictions resulting from this model?



## No “Best” Approach

- ▶ There is no “best” approach to regression in high dimensions.

## No “Best” Approach

- ▶ There is no “best” approach to regression in high dimensions.
- ▶ Some approaches will work better than others. For instance:
  - ▶ Lasso will work well if it’s really true that just a few features are associated with the response.
  - ▶ Ridge will do better if all of the features are associated with the response.

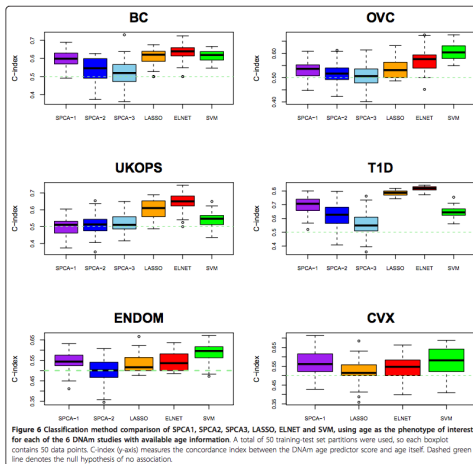
## No “Best” Approach

- ▶ There is no “best” approach to regression in high dimensions.
- ▶ Some approaches will work better than others. For instance:
  - ▶ Lasso will work well if it’s really true that just a few features are associated with the response.
  - ▶ Ridge will do better if all of the features are associated with the response.
- ▶ If somebody tells you that one approach is “best”... then they are mistaken. Politely contradict them.
- ▶ While no approach is “best”, some approaches are wrong (e.g.: there is a wrong way to do cross-validation)!

## Predicting Age Using DNA Methylation Data

- ▶ Comparison on 6 data sets
- ▶ SPC: A method based on dimension reduction (not discussed here).
- ▶ Elastic Net: A hybrid between ridge and lasso.
- ▶ SVM: We'll see it next lecture in the classification context.
- ▶ Citation: Zhuang et al., BMC Bioinformatics, 2012

# Didn't I Tell You? No Best Method!



High C-index indicates a low test error.

## Bottom Line

Much more important than what model you fit is how you fit it.

- ▶ Was cross-validation performed properly?
- ▶ Did you select a model (or level of model complexity) based on an estimate of test error?

## Discussion Questions

A collaborator comes to you and says:

*I really don't like this **LASSO** thing; I tried it on my data and it the resulting model only explained 15% of the variability in my data... Then I tried **variable pre-selection**, and I was able to get it to explain 95%! Why would anyone ever use the **LASSO**???*

What do you think is happening?

## Discussion Questions

What if instead they said:

*I really don't like this variable pre-selection thing; I tried it on my data and it the resulting model only explained 15% of the variability in my data... Then I tried the LASSO, and I was able to get it to explain 95%! Why would anyone ever use variable pre-selection???*



## Discussion Questions

Finally, what if they said:

*I really love the **LASSO**. I was originally just using **standard linear regression** and the resulting model only explained 15% of the variability in my data... Then I tried the **LASSO**, and I was able to get it to explain 95%!*

What do you think is happening here?

## Discussion Questions

A collaborator came to me and said:

“I am reviewing a paper where the authors claim to be able to predict the flu, by looking at serum gene expression values 3 weeks before symptom onset. This seems impossible, but I can't find an obvious error in the paper”

## Discussion Questions

Looking at the paper, the authors had used the following pipeline:

- ▶ Took banked blood from 100 patients (50 subsequently diagnosed with flu, 50 were not).
- ▶ They separately looked at the correlation of expression of each gene with flu-status, and selected the 70 top genes
- ▶ They split into a training and test set.
- ▶ On the training set they ran 5-fold cross validation to come up with an optimal aggregation of kernel-SVM, logistic regression, and boosted classification trees
- ▶ They evaluated this on the test set, and found almost perfect classification.

## Discussion Questions

What is going on???

Did they go on to create an enormously successful biotech company?