Modern regression 2: The lasso

Ryan Tibshirani
Data Mining: 36-462/36-662

March 21 2013

Optional reading: ISL 6.2.2, ESL 3.4.2, 3.4.3

Reminder: ridge regression and variable selection

Recall our setup: given a response vector $y \in \mathbb{R}^n$, and a matrix $X \in \mathbb{R}^{n \times p}$ of predictor variables (predictors on the columns)

Last time we saw that ridge regression, $\lambda \sum_{j=1}^{p} (\beta_j - c)^2$ same idea

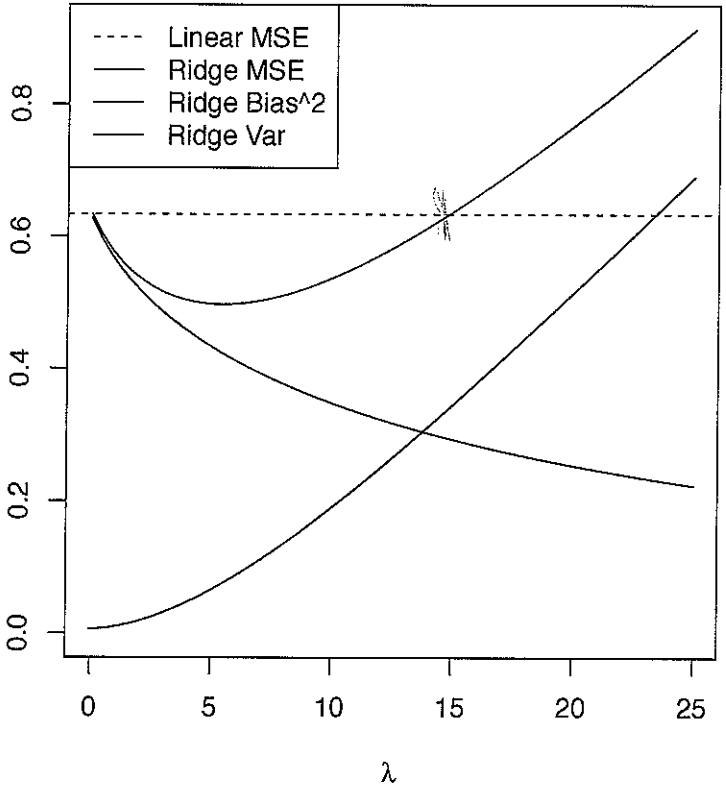
$$\hat{\beta}^{\text{ridge}} = \underset{\beta \in \mathbb{R}^p}{\text{argmin}} \ \|y - X\beta\|_2^2 + \lambda \|\beta\|_2^2$$

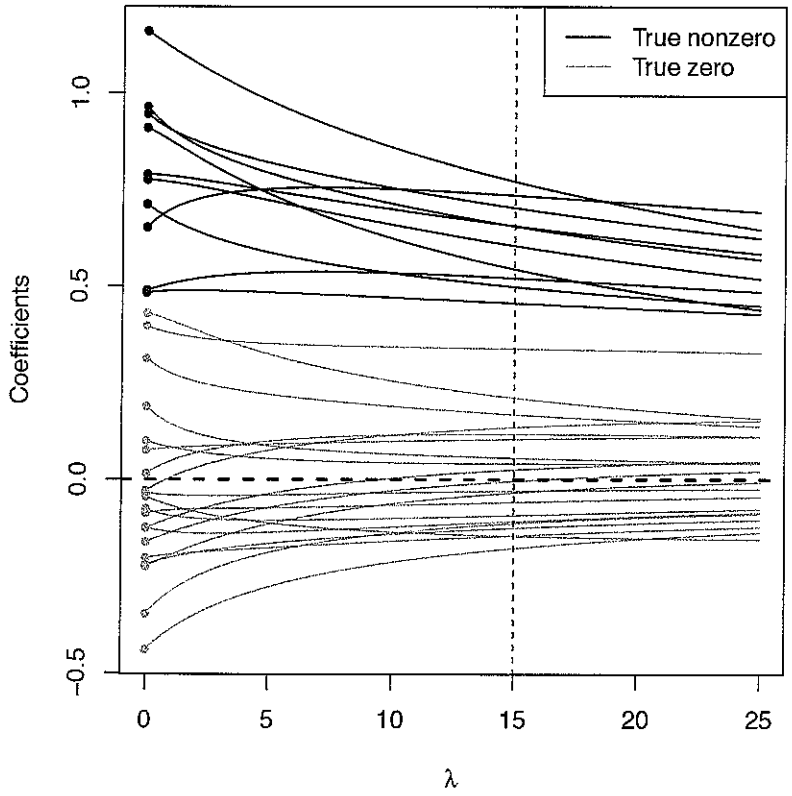
can have better prediction error than linear regression in a variety of scenarios, depending on the choice of λ. It worked best when there was a subset of the true coefficients that are small or zero

But it will never sets coefficients to zero exactly, and therefore cannot perform variable selection in the linear model. While this didn't seem to hurt its prediction ability, it is not desirable for the purposes of interpretation (especially if the number of variables p is large)

$\hat{\beta}_j = 0$

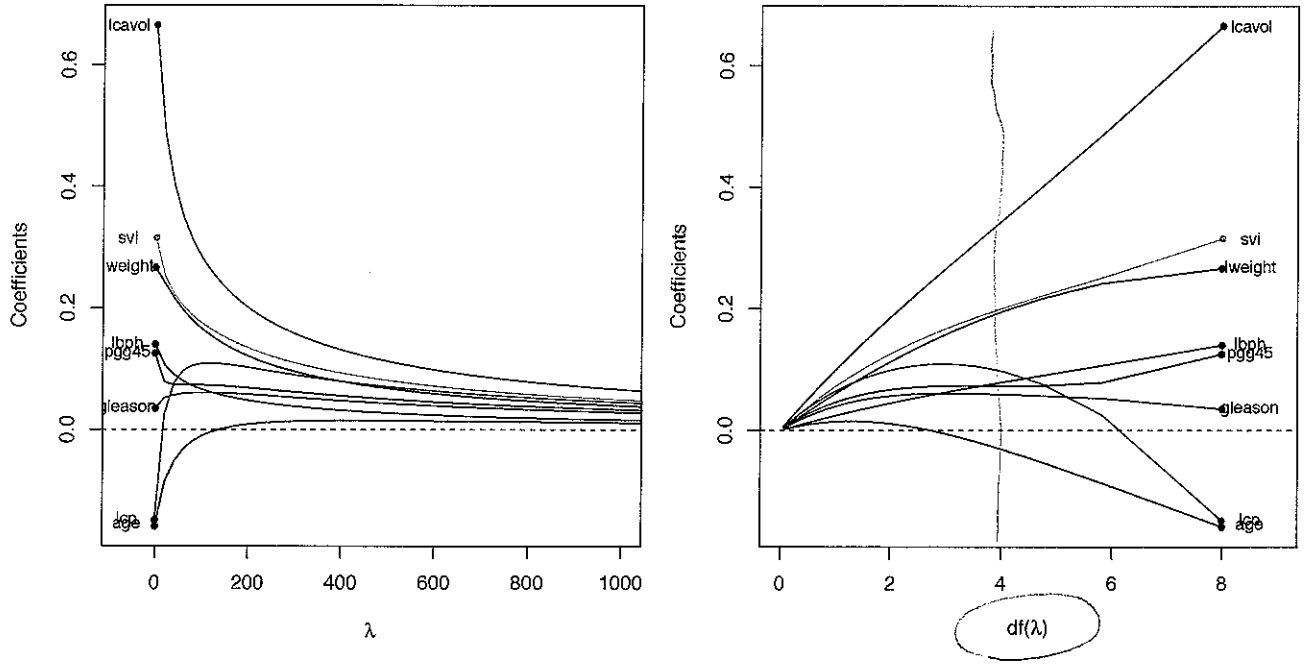
Recall our example: $n = 50$, $p = 30$; true coefficients: 10 are nonzero and pretty big, 20 are zero





Example: prostate data

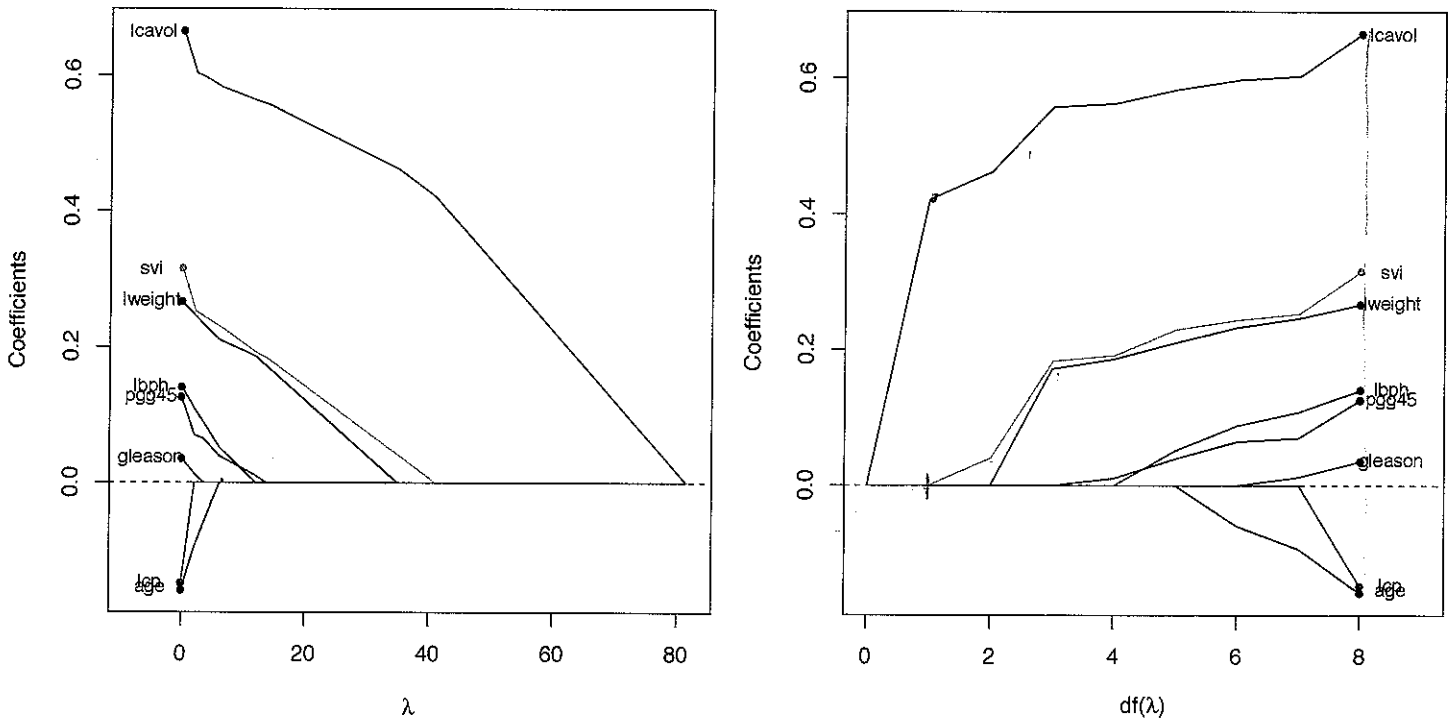
Recall the prostate data example: we are interested in the level of prostate-specific antigen (PSA), elevated in men who have prostate cancer. We have measurements of PSA on $n = 97$ men with prostate cancer, and $p = 8$ clinical predictors. Ridge coefficients:



What if the people who gave this data want us to derive a linear model using only a few of the 8 predictor variables to predict the level of PSA?

Now the lasso coefficient paths:

$\hat{\beta}(\lambda)$, $\hat{y}(\lambda) = X\hat{\beta}(\lambda)$ change with λ



We might report the first 3 coefficients to enter the model: lcavol (the log cancer volume), svi (seminal vesicle invasion), and lweight (the log prostate weight)

How would we choose 3 (i.e., how would we choose λ?) We'll talk about this later

The lasso

The lasso[1] estimate is defined as

$$\hat{\beta}^{\text{lasso}} = \underset{\beta \in \mathbb{R}^p}{\operatorname{argmin}} \ \|y - X\beta\|_2^2 + \lambda \sum_{j=1}^{p} |\beta_j|$$

$$= \underset{\beta \in \mathbb{R}^p}{\operatorname{argmin}} \ \underbrace{\|y - X\beta\|_2^2}_{\text{Loss}} + \lambda \underbrace{\|\beta\|_1}_{\text{Penalty}}$$

The only difference between the lasso problem and ridge regression is that the latter uses a (squared) ℓ_2 penalty $\|\beta\|_2^2$, while the former uses an ℓ_1 penalty $\|\beta\|_1$. But even though these problems look similar, their solutions behave very differently

Note the name "lasso" is actually an acronym for: Least Absolute Selection and Shrinkage Operator

[1] Tibshirani (1996), "Regression Shrinkage and Selection via the Lasso"

$$\hat{\beta}^{\text{lasso}} = \underset{\beta \in \mathbb{R}^p}{\operatorname{argmin}} \ \|y - X\beta\|_2^2 + \lambda \|\beta\|_1$$

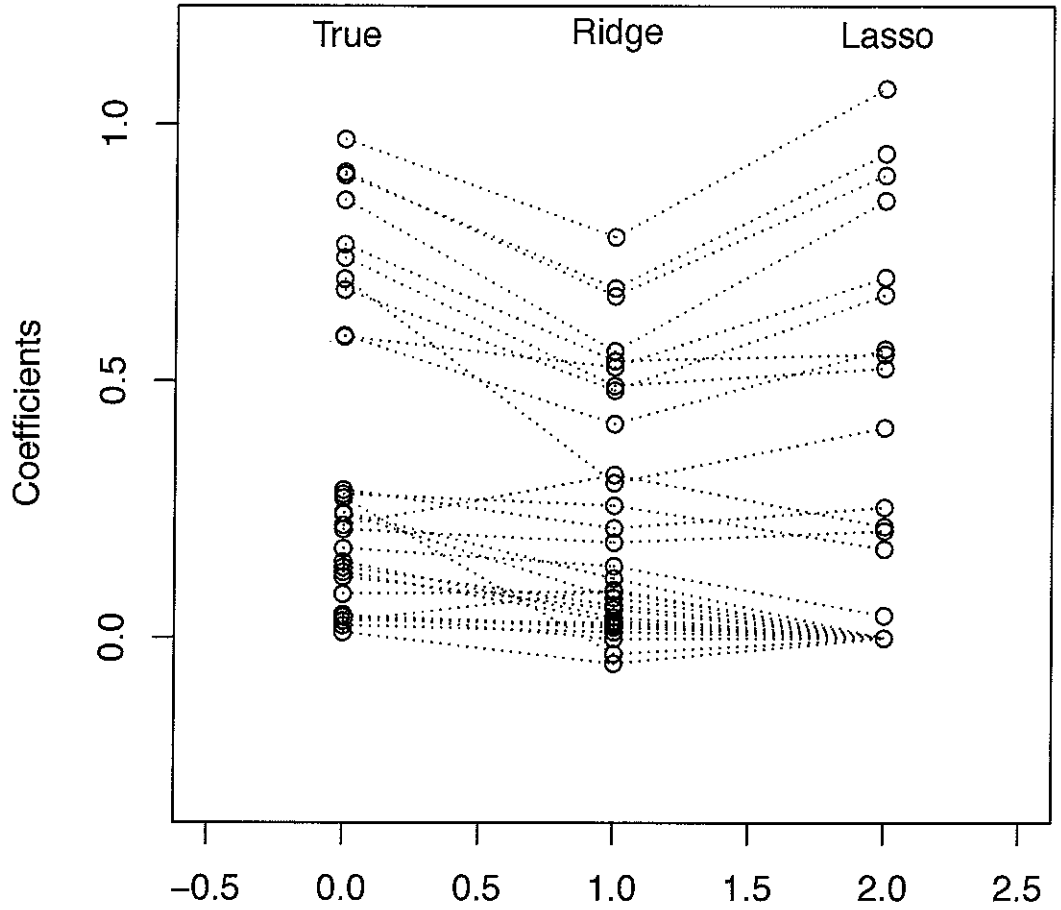
The tuning parameter λ controls the strength of the penalty, and (like ridge regression) we get $\hat{\beta}^{\text{lasso}} =$ the linear regression estimate when $\lambda = 0$, and $\hat{\beta}^{\text{lasso}} = 0$ when $\lambda = \infty$

For λ in between these two extremes, we are balancing two ideas: fitting a linear model of y on X, and shrinking the coefficients. But the nature of the ℓ_1 penalty causes some coefficients to be shrunken to zero exactly

This is what makes the lasso substantially different from ridge regression: it is able to perform variable selection in the linear model. As λ increases, more coefficients are set to zero (less variables are selected), and among the nonzero coefficients, more shrinkage is employed

Example: visual representation of lasso coefficients

Our running example from last time: $n = 50$, $p = 30$, $\sigma^2 = 1$, 10 large true coefficients, 20 small. Here is a visual representation of lasso vs. ridge coefficients (with the same degrees of freedom):



Important details

When including an intercept term in the model, we usually leave this coefficient unpenalized, just as we do with ridge regression. Hence the lasso problem with intercept is

$$\hat{\beta}_0, \hat{\beta}^{\text{lasso}} = \underset{\beta_0 \in \mathbb{R},\, \beta \in \mathbb{R}^p}{\operatorname{argmin}} \; \|y - \beta_0 \mathbb{1} - X\beta\|_2^2 + \lambda \|\beta\|_1$$

As we've seen before, if we center the columns of X, then the intercept estimate turns out to be $\hat{\beta}_0 = \bar{y}$. Therefore we typically center y, X and don't include an intercept them

As with ridge regression, the penalty term $\|\beta\|_1 = \sum_{j=1}^p |\beta_j|$ is not fair is the predictor variables are not on the same scale. Hence, if we know that the variables are not on the same scale to begin with, we scale the columns of X (to have sample variance 1), and then we solve the lasso problem

Bias and variance of the lasso

Although we can't write down explicit formulas for the bias and variance of the lasso estimate (e.g., when the true model is linear), we know the general trend. Recall that

$$\hat{\beta}^{\text{lasso}} = \underset{\beta \in \mathbb{R}^p}{\operatorname{argmin}} \ \|y - X\beta\|_2^2 + \lambda \|\beta\|_1$$

Generally speaking:

- The bias increases as λ (amount of shrinkage) increases
- The variance decreases as λ (amount of shrinkage) increases

What is the bias at ~~at~~ $\lambda = 0$? The variance at $\lambda = \infty$?

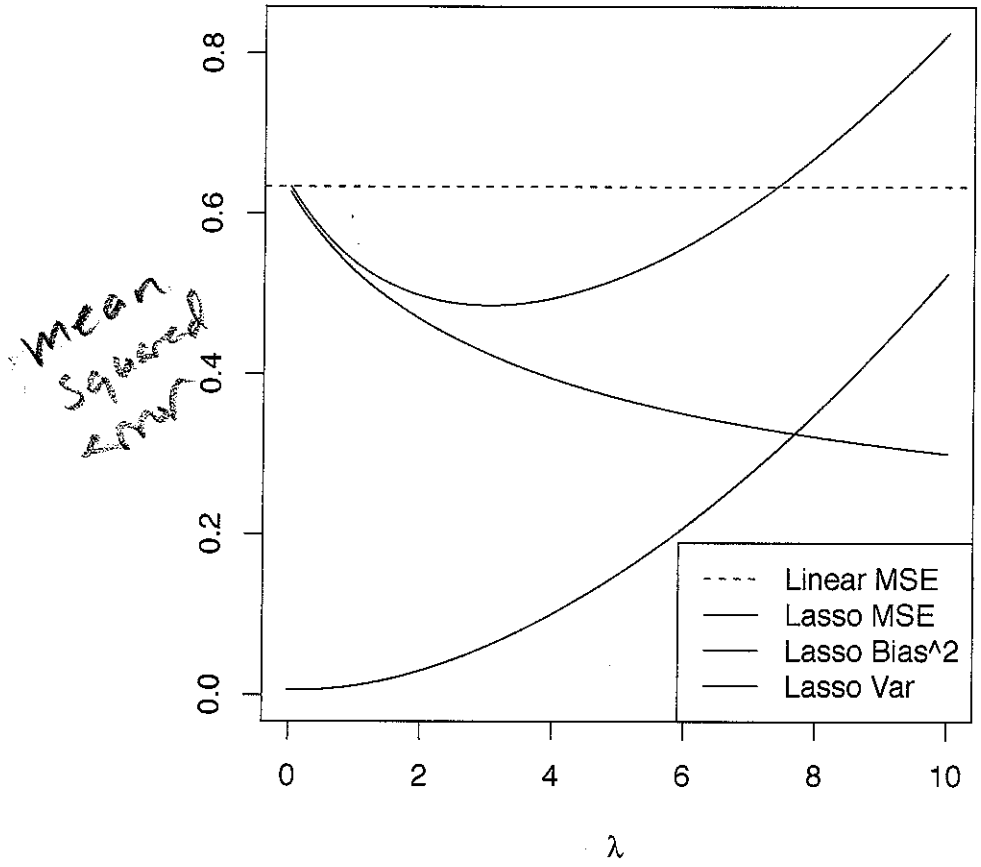
0 0

In terms of prediction error (or mean squared error), the lasso performs comparably to ridge regression

$\hat{\beta}^{\text{ridge}}$ = see Homework, $\hat{\beta}^{\text{lasso}}$? = no explicit formula

Example: subset of small coefficients

Example: $n = 50$, $p = 30$; true coefficients: 10 large, 20 small

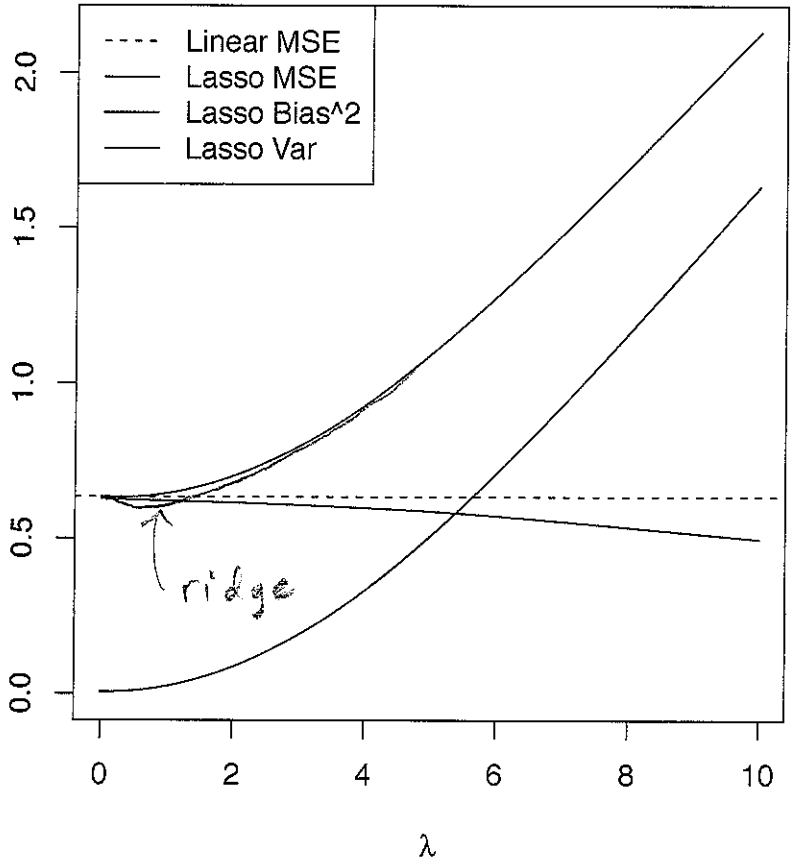


The lasso: see the function `lars` in the package `lars`

glmnet package

Example: all moderate coefficients

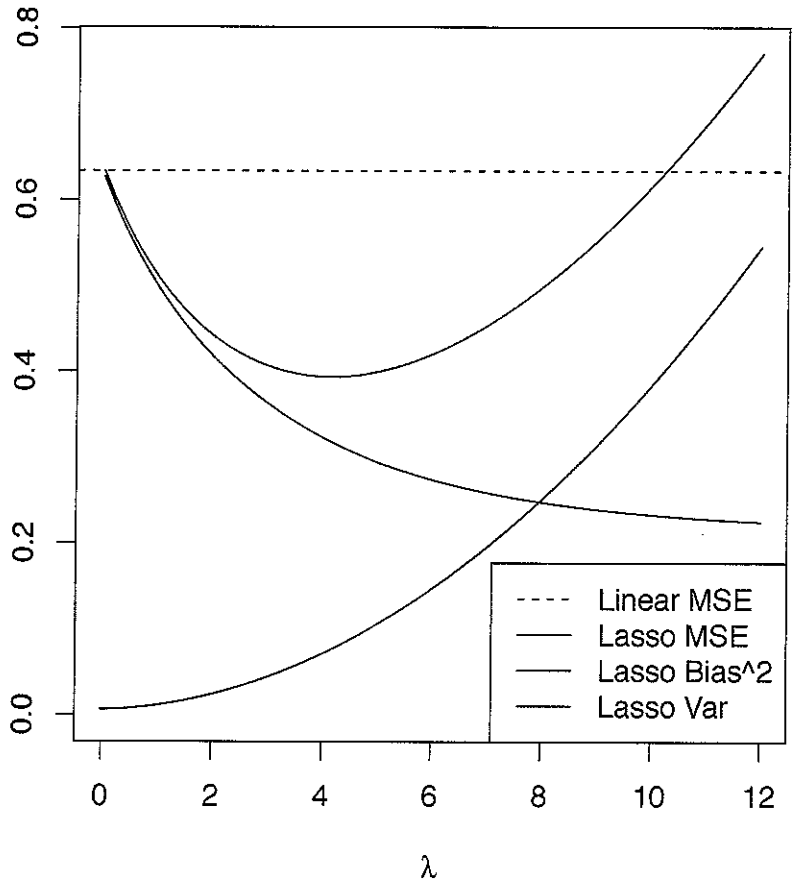
Example: $n = 50$, $p = 30$; true coefficients: 30 moderately large



Note that here, as opposed to ridge regression the variance doesn't decrease fast enough to make the lasso favorable for small λ

Example: subset of zero coefficients

Example: $n = 50$, $p = 30$; true coefficients: 10 large, 20 zero

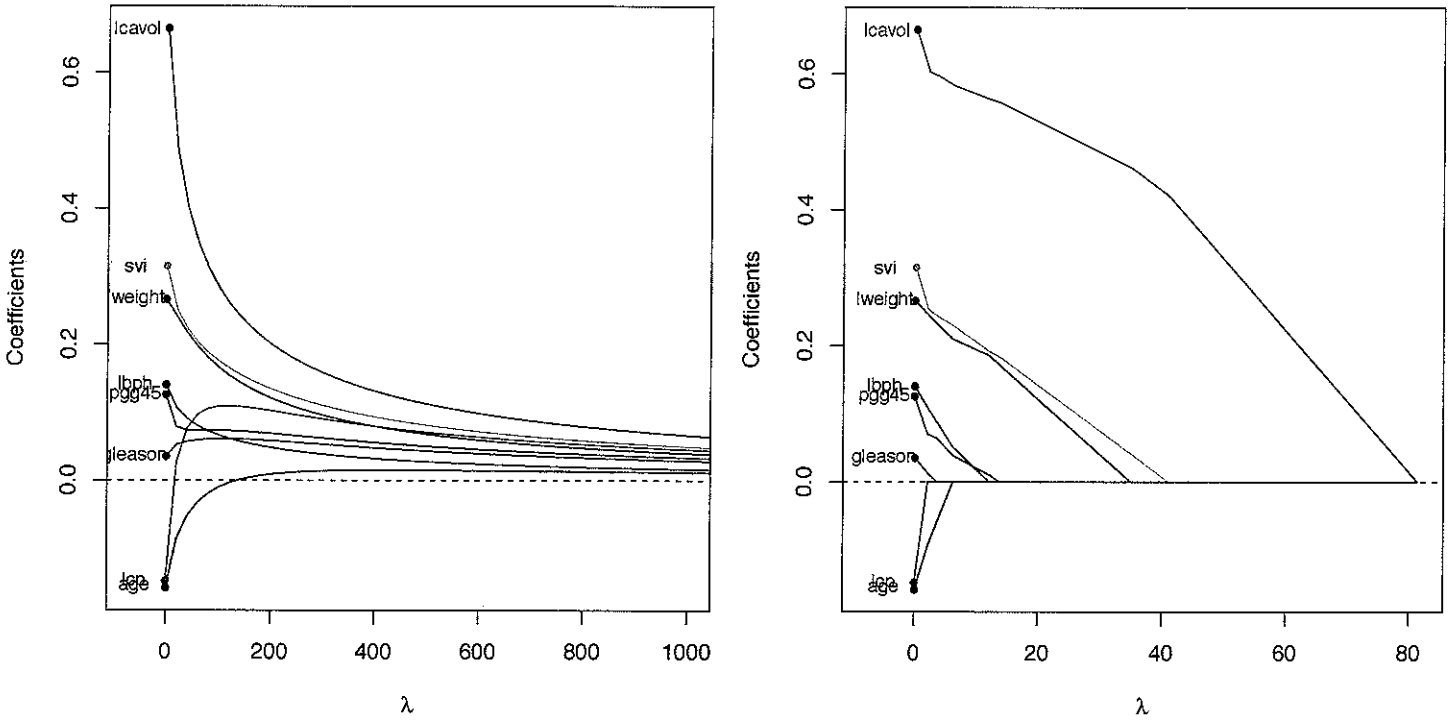


Advantage in interpretation

On top the fact that the lasso is competitive with ridge regression in terms of this prediction error, it has a big advantage with respect to interpretation. This is exactly because it sets coefficients exactly to zero, i.e., it performs variable selection in the linear model

sparsity
sparse $\hat{\beta}$ means lots of $\hat{\beta}_j = 0$,

Recall the prostate cancer data example:



Example: murder data

Example: we study the murder rate (per 100K people) of $n = 2215$ communities in the U.S.[2] We have $p = 101$ attributes measured each community, such as

```
[1] "racePctHisp"      "agePct12t21"      "agePct12t29"
[4] "agePct16t24"      "agePct65up"       "numbUrban"
[7] "pctUrban"         "medIncome"        "pctWWage"
...
```

Our goal is to predict the murder rate as a linear function of these attributes. For the purposes of interpretation, it would be helpful to have a linear model involving only a small subset of these attributes. (Note: interpretation here is *not causal*)

[2] Data from UCI Machine Learning Repository, `http://archive.ics.uci.edu/ml/datasets/Communities+and+Crime+Unnormalized`

With ridge regression, regardless of the choice of $\lambda < \infty$, we never get zero coefficient estimates. For $\lambda = 25,000$, which corresponds to approximately 5 degrees of freedom, we get estimates:

racePctHisp	agePct12t21	agePct12t29
0.0841354923	0.0076226029	0.2992145264
agePct16t24	agePct65up	numbUrban
-0.2803165408	0.0115873137	0.0154487020
pctUrban	medIncome	pctWWage
-0.0155148961	-0.0105604035	-0.0228670567

...

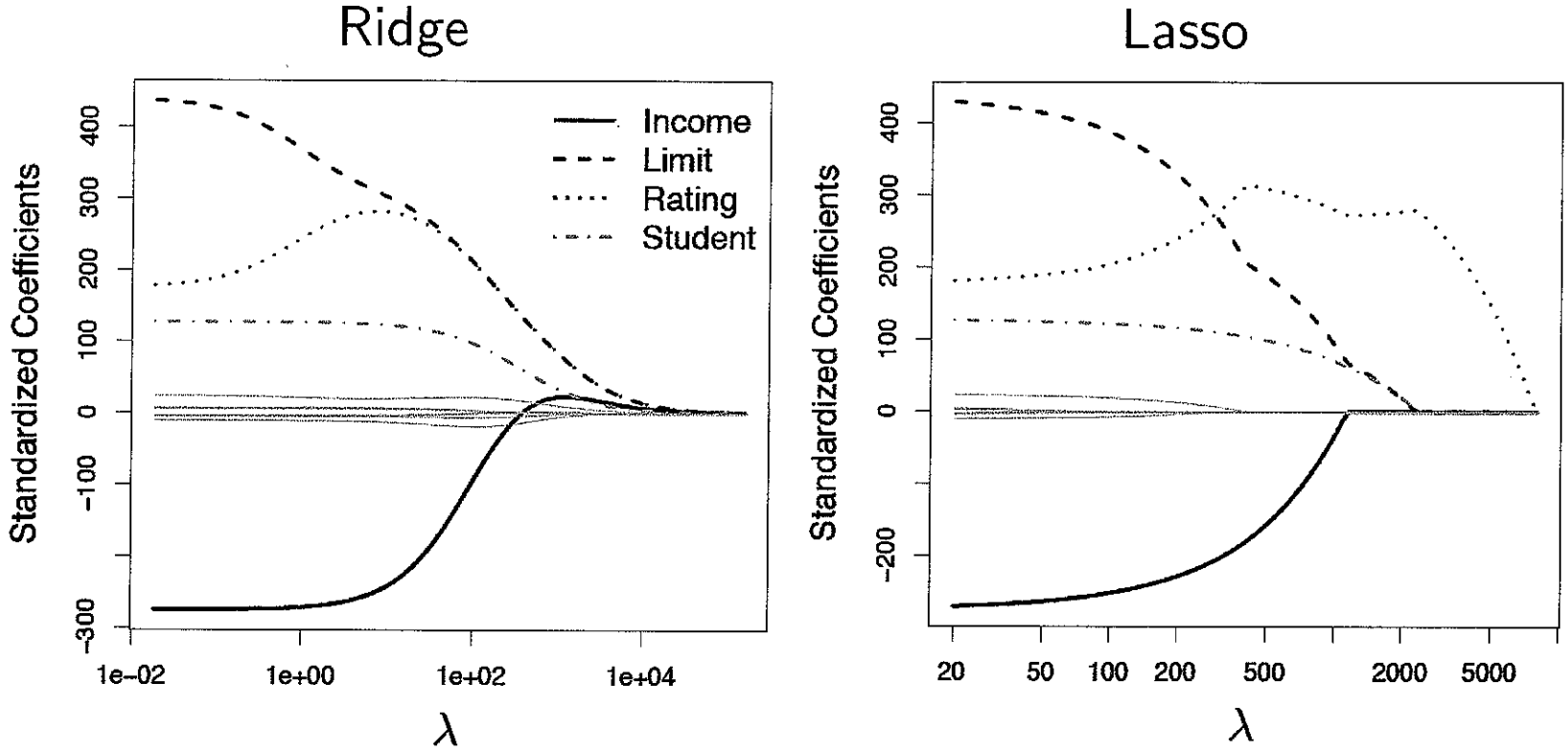
With the lasso, for about the same degrees of freedom, we get:

agePct12t29	agePct16t24	NumKidsBornNeverMar
0.7113530	-1.8185387	-0.6835089
PctImmigRec10	OwnOccLowQuart	
1.3825129	1.0234245	

and all other coefficient estimates are zero. That is, we get exactly 5 nonzero coeffcients out of $p = 101$ total

Example: credit data

Example from ISL sections 6.6.1 and 6.6.2: response is average credit debt, predictors are income, limit (credit limit), rating (credit rating), student (indicator), and others



min Loss + $\lambda \|\beta\|_2^2$
min Loss + $\lambda \|\beta\|_1$

Constrained form

It can be helpful to think of our two problems constrained form:

$$\hat{\beta}^{\text{ridge}} = \underset{\beta \in \mathbb{R}^p}{\text{argmin}} \ \|y - X\beta\|_2^2 \quad \text{subject to } \|\beta\|_2^2 \leq t \qquad \sum \beta_j^2 \leq t$$

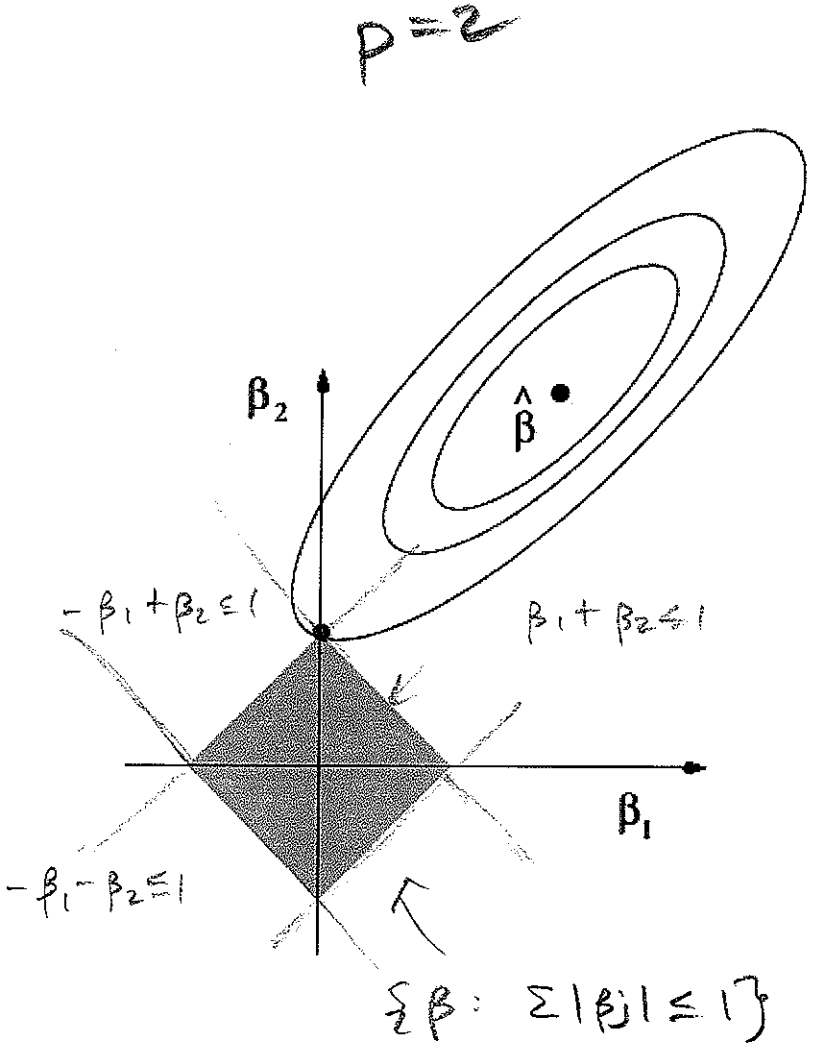
$$\hat{\beta}^{\text{lasso}} = \underset{\beta \in \mathbb{R}^p}{\text{argmin}} \ \|y - X\beta\|_2^2 \quad \text{subject to } \|\beta\|_1 \leq t$$

$\hat{\beta}$ LS. solution, consider $t \geq \|\hat{\beta}\|_1$

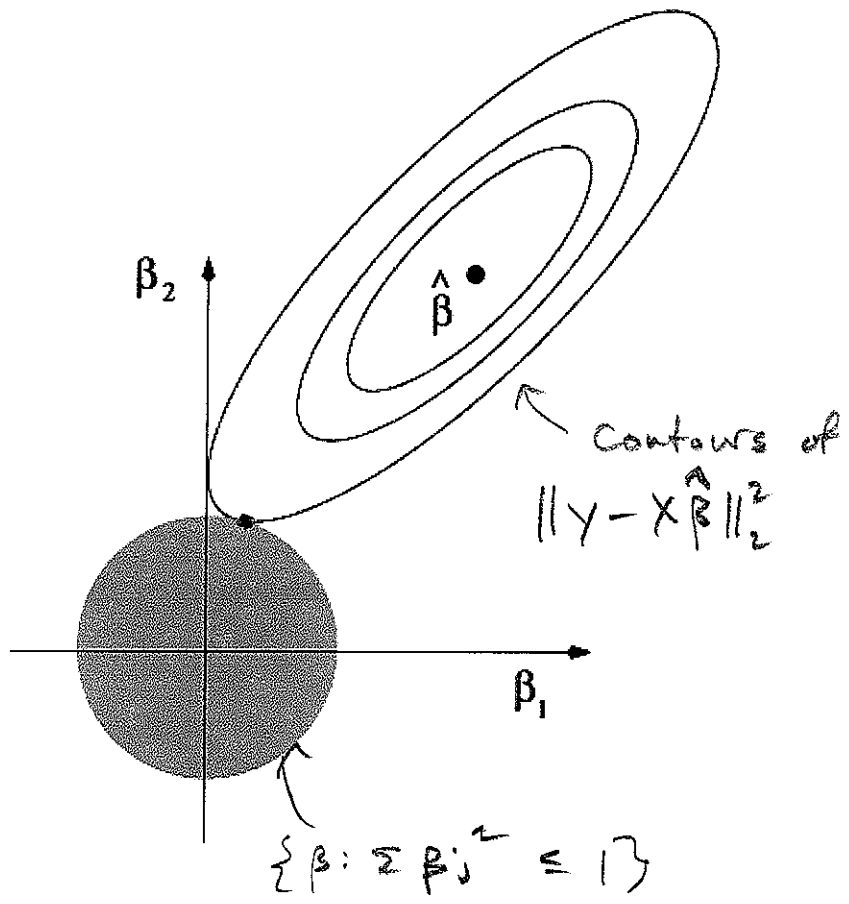
Now t is the tuning parameter (before it was λ). For any λ and corresponding solution in the previous formulation (sometimes called penalized form), there is a value of t such that the above constrained form has this same solution

In comparison, the usual linear regression estimate solves the unconstrained least squares problem; these estimates constrain the coefficient vector to lie in some geometric shape centered around the origin. This generally reduces the variance because it keeps the estimate close to zero. But which shape we choose really matters!

Why does the lasso give zero coefficients?



(From page 71 of ESL)



minimize $\|y - X\beta\|$

s.t. $\sum \beta_j^2 \le 1$

this means

What is degrees of freedom?

Broadly speaking, the degrees of freedom of an estimate describes its effective number of parameters

More precisely, given data $y \in \mathbb{R}^n$ from the model

$$y_i = \mu_i + \epsilon_i, \quad i = 1, \ldots n$$

where $\mathrm{E}[\epsilon_i] = 0$, $\mathrm{Var}(\epsilon_i) = \sigma^2$, $\mathrm{Cov}(\epsilon_i, \epsilon_j) = 0$, suppose that we estimate y by $\hat{y}$. The degrees of freedom of the estimate $\hat{y}$ is

$$\mathrm{df}(\hat{y}) = \frac{1}{\sigma^2} \sum_{i=1}^{n} \mathrm{Cov}(\hat{y}_i, y_i)$$

The higher the correlation between the ith fitted value and the ith data point, the more adaptive the estimate, and so the higher its degrees of freedom

Let $X \in \mathbb{R}^{n \times p}$ be a fixed matrix of predictors[3]

- For linear regression, $\hat{y} = X\hat{\beta}^{\text{linear}}$, we have $\text{df}(\hat{y}) = p$
- For ridge regression, $\hat{y} = X\hat{\beta}^{\text{ridge}}$, where

$$\hat{\beta}^{\text{ridge}} = \underset{\beta \in \mathbb{R}^p}{\text{argmin}} \ \|y - X\beta\|_2^2 + \lambda\|\beta\|_2^2$$

  we have $\text{df}(\hat{y}) = \text{trace}\Big(X(X^TX + \lambda I)^{-1}X^T\Big)$

- For the lasso, $\hat{y} = X\hat{\beta}^{\text{lasso}}$, where

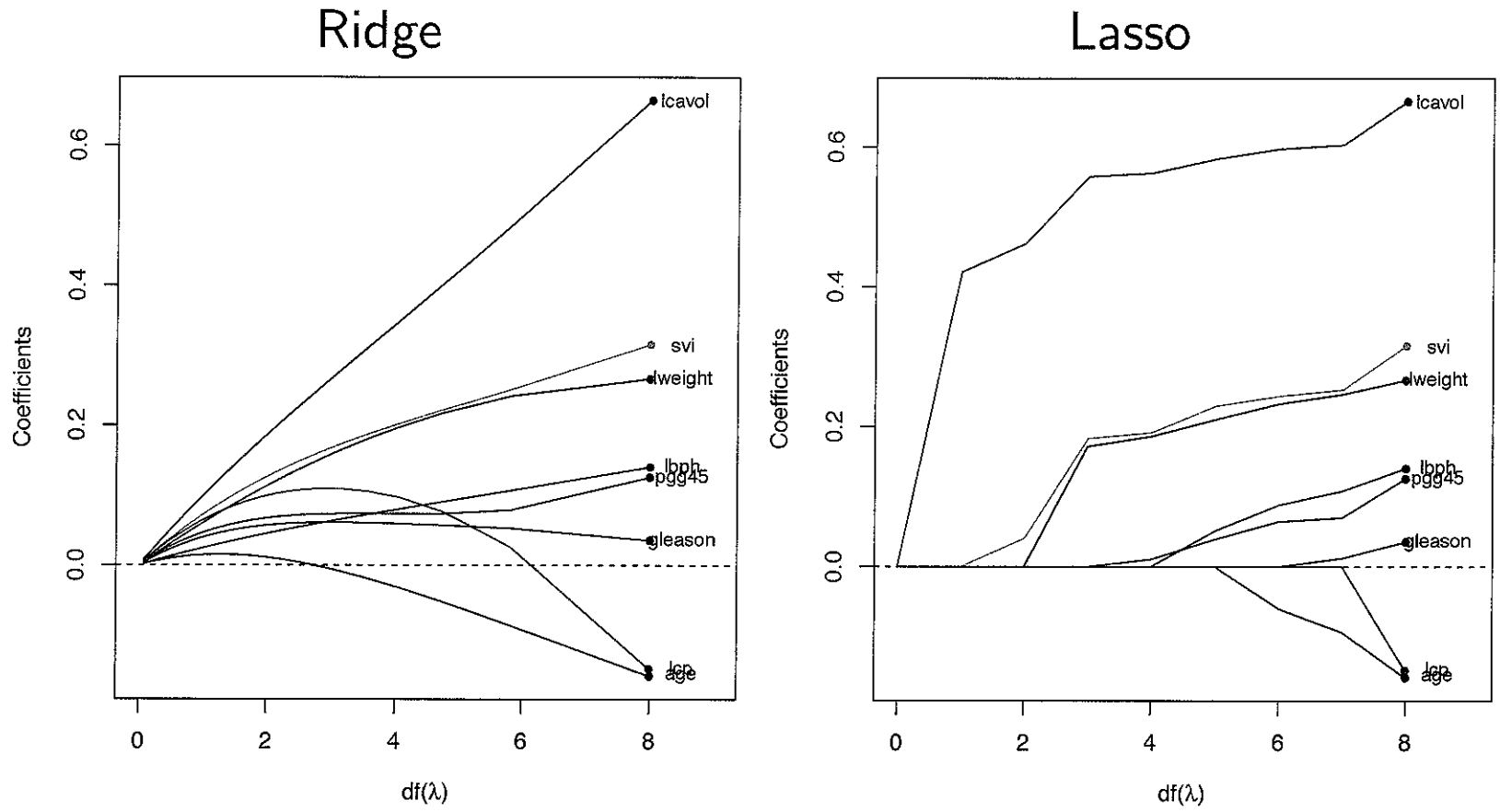
$$\hat{\beta}^{\text{lasso}} = \underset{\beta \in \mathbb{R}^p}{\text{argmin}} \ \|y - X\beta\|_2^2 + \lambda\|\beta\|_1$$

  we have $\text{df}(\hat{y}) = \text{E}[\text{number of nonzero coefficients in } \hat{\beta}^{\text{lasso}}]$

[3]For simplicity, we assume that the predictors are linearly independent; the case for dependent predictors is similar

One usage of degrees of freedom is to put two different estimates on equal footing

E.g., comparing ridge and lasso for the prostate cancer data set



Recap: the lasso

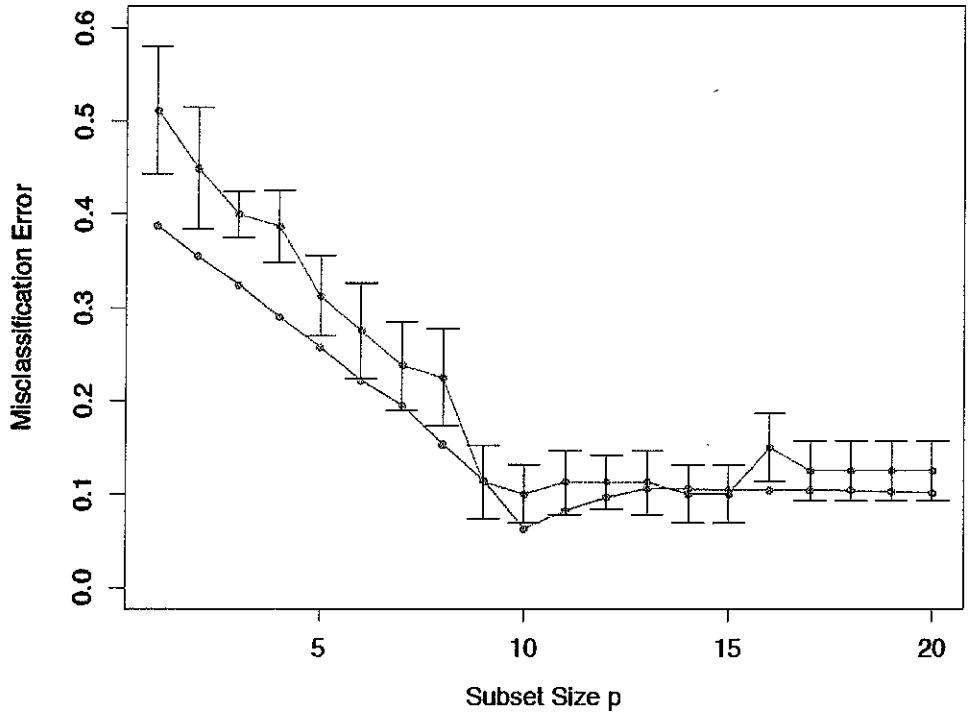
In this lecture we learned a variable selection method in the linear model setting: the lasso. The lasso uses a penalty like ridge regression, except the penalty is the ℓ_1 norm of the coefficient vector, which causes the estimates of some coefficients to be exactly zero. This is in constrast to ridge regression which never sets coefficients to zero

The tuning parameter λ controls the strength of the ℓ_1 penalty. The lasso estimates are generally biased, but have good mean squared error (comparable to ridge regression). On top of this, the fact that it sets coefficients to zero can be a big advantage for the sake of interpretation

We defined the concept of degrees of freedom, which measures the effective number of parameters used by a estimator. This allows us to compare estimators with different tuning parameters

Next time: model selection and validation

Cross-validation can be used to estimate the prediction error curve



(From ESL page 44)