Modern regression 1: Ridge regression

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

March 19 2013

Optional reading: ISL 6.2.1, ESL 3.4.1

Reminder: shortcomings of linear regression

Last time we talked about:

1. Predictive ability: recall that we can decompose prediction error into squared bias and variance. Linear regression has low bias (zero bias) but suffers from high variance. So it may be worth sacrificing some bias to achieve a lower variance
2. Interpretative ability: with a large number of predictors, it can be helpful to identify a smaller subset of important variables. Linear regression doesn't do this

Also: linear regression is not defined when $\underline{p > n}$ (Homework 4)

high-dimensional.

Setup: given fixed covariates $x_i \in \mathbb{R}^p$, $i = 1, \ldots n$, we observe

$$y_i = f(x_i) + \epsilon_i, \quad i = 1, \ldots n,$$

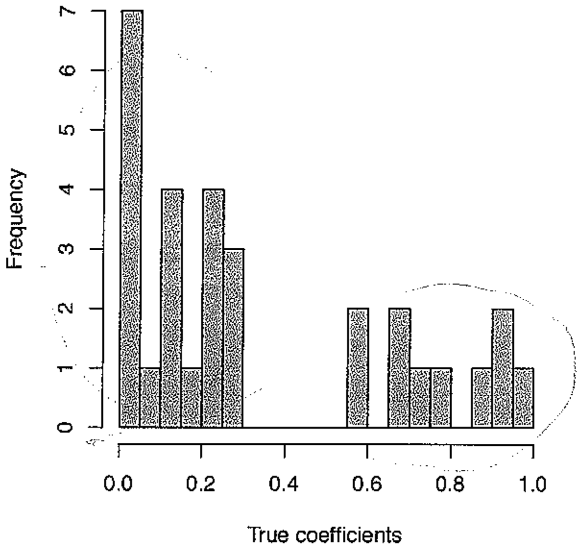
true coefficient vector

where $f : \mathbb{R}^p \to \mathbb{R}$ is unknown (think $f(x_i) = x_i^T \beta^*$ for a linear model) and $\epsilon_i \in \mathbb{R}$ with $\mathrm{E}[\epsilon_i] = 0, \mathrm{Var}(\epsilon_i) = \sigma^2, \mathrm{Cov}(\epsilon_i, \epsilon_j) = 0$

Example: subset of small coefficients

Recall our example: we have $n = 50$, $p = 30$, and $\sigma^2 = 1$. The true model is linear with 10 large coefficients (between 0.5 and 1) and 20 small ones (between 0 and 0.3). Histogram:

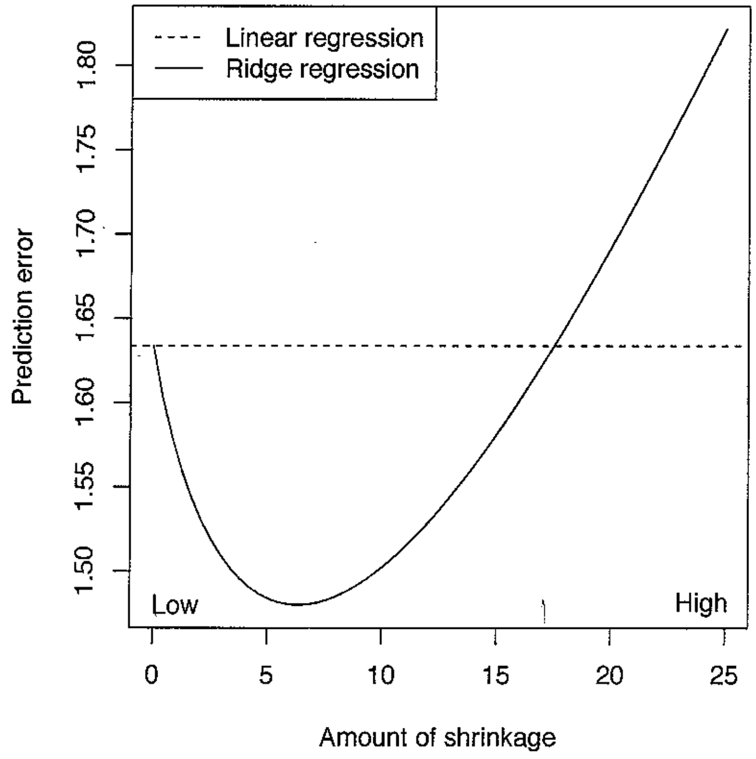
$$y_i = \underbrace{\beta_1 x_{i1} + \dots \beta_{10} x_{i,10}}_{\text{large}} + \underbrace{\beta_{11} x_{i,11} + \dots \beta_{30} x_{i,30}}_{\text{small}}$$



The linear regression fit:

Squared bias ≈ 0.006
Variance ≈ 0.627
Pred. error $\approx \underset{\substack{\uparrow \\ \sigma^2 + (\text{Bias})^2 + \text{Var}}}{1} + 0.006 + 0.627 \approx 1.633$

We reasoned that we can do better by shrinking the coefficients, to reduce variance



Linear regression:
Squared bias ≈ 0.006
Variance ≈ 0.627
Pred. error $\approx 1 + 0.006 + 0.627$
≈ 1.633

Ridge regression, at its best:
Squared bias ≈ 0.077
Variance ≈ 0.403
Pred. error $\approx 1 + 0.077 + 0.403$
≈ 1.48

Ridge regression

Ridge regression is like least squares but shrinks the estimated coefficients towards zero. Given a response vector $y \in \mathbb{R}^n$ and a predictor matrix $X \in \mathbb{R}^{n \times p}$, the ridge regression coefficients are defined as

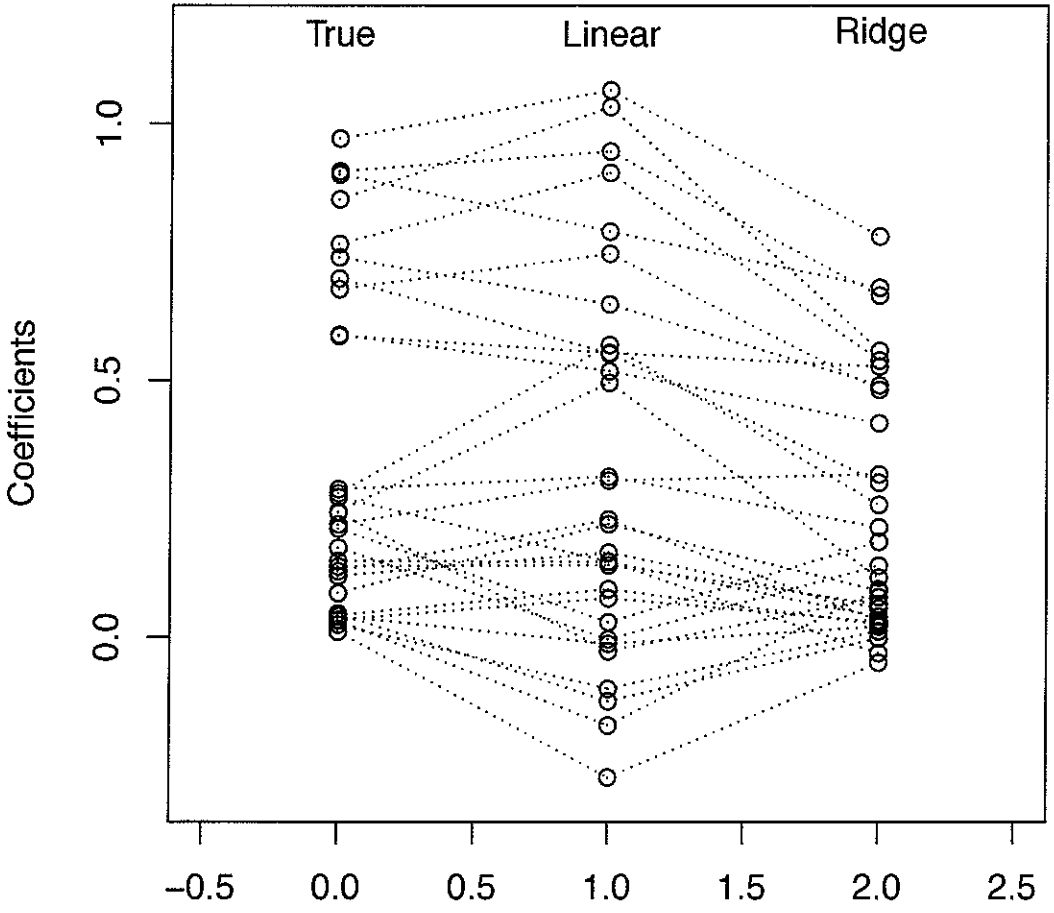
$$\begin{aligned}\hat{\beta}^{\text{ridge}} &= \underset{\beta \in \mathbb{R}^p}{\operatorname{argmin}} \sum_{i=1}^{n} (y_i - x_i^T \beta)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \\ &= \underset{\beta \in \mathbb{R}^p}{\operatorname{argmin}} \underbrace{\|y - X\beta\|_2^2}_{\text{Loss}} + \lambda \underbrace{\|\beta\|_2^2}_{\text{Penalty}}\end{aligned}$$

Here $\lambda \geq 0$ is a tuning parameter, which controls the strength of the penalty term. Note that:

- When $\lambda = 0$, we get the linear regression estimate
- When $\lambda = \infty$, we get $\hat{\beta}^{\text{ridge}} = 0$
- For λ in between, we are balancing two ideas: fitting a linear model of y on X, and shrinking the coefficients

Example: visual representation of ridge coefficients

Recall our last example ($n = 50$, $p = 30$, and $\sigma^2 = 1$; 10 large true coefficients, 20 small). Here is a visual representation of the ridge regression coefficients for $\lambda = 25$:



Important details

When including an intercept term in the regression, we usually leave this coefficient unpenalized. Otherwise we could add some constant amount c to the vector y, and this would not result in the same solution. Hence ridge regression with intercept solves

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

$$y - \bar{y}\,\mathbb{1}$$

If we center the columns of X, then the intercept estimate ends up just being $\boxed{\hat{\beta}_0 = \bar{y},}$ so we usually just assume that y, X have been centered and don't include an intercept

Also, the penalty term $\|\beta\|_2^2 = \sum_{j=1}^p \beta_j^2$ is unfair is the predictor variables are not on the same scale. (Why?) Therefore, if we know that the variables are not measured in the same units, we typically scale the columns of X (to have sample variance 1), and then we perform ridge regression

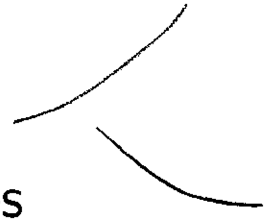
Bias and variance of ridge regression

The bias and variance are not quite as simple to write down for ridge regression as they were for linear regression, but closed-form expressions are still possible (Homework 4). Recall that

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

The general trend is:

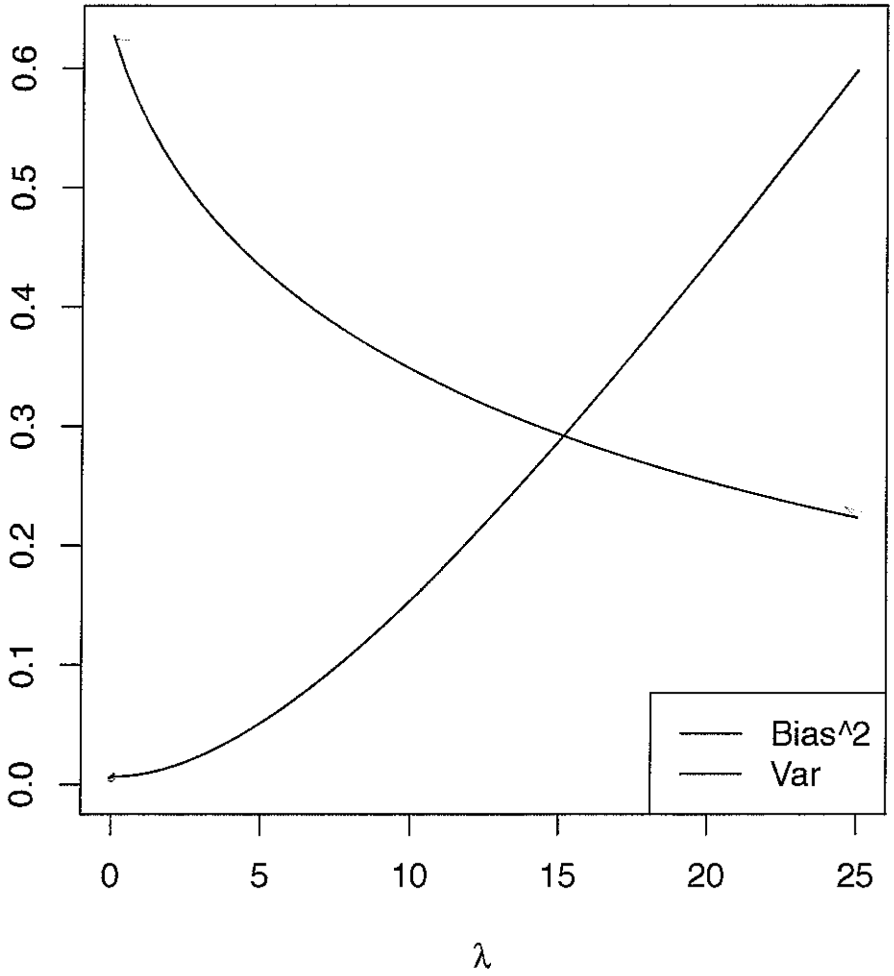
- The bias increases as λ (amount of shrinkage) increases
- The variance decreases as λ (amount of shrinkage) increases



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

Example: bias and variance of ridge regression

Bias and variance for our last example ($n = 50$, $p = 30$, $\sigma^2 = 1$; 10 large true coefficients, 20 small):

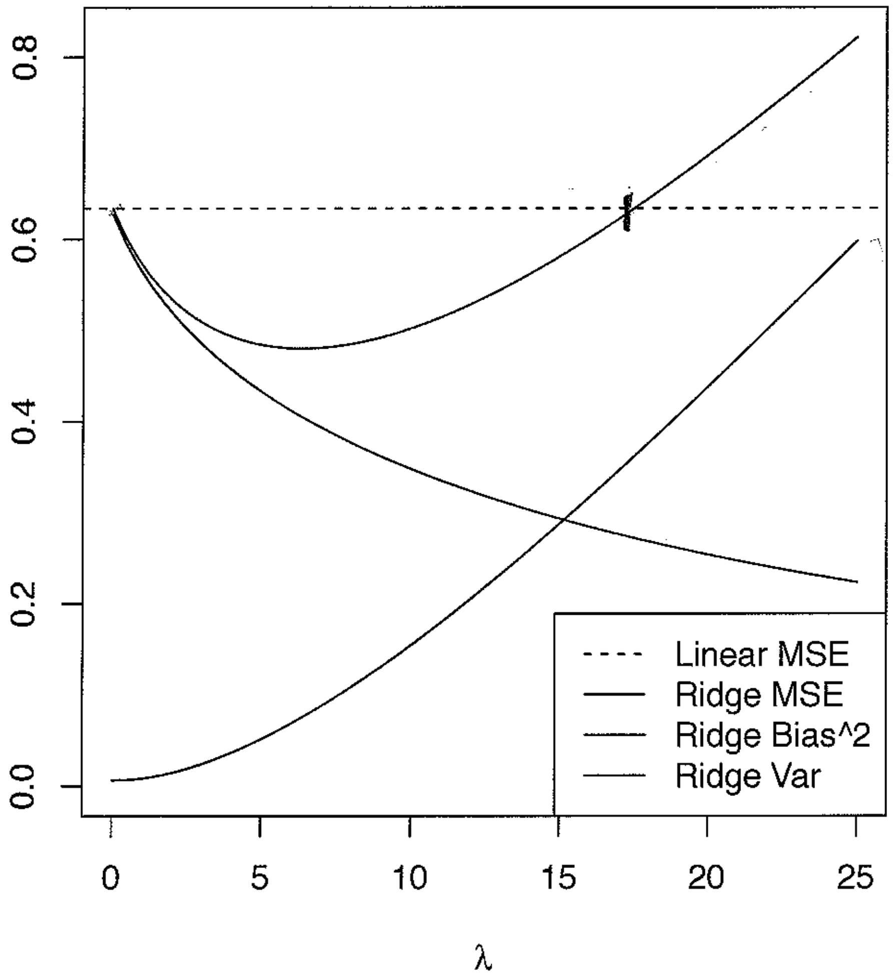


→

shrinkage

Mean squared error for our last example:

$=(\text{Bias})^2+\text{Var}$



Ridge regression in R: see the function `lm.ridge` in the package `MASS`, or the `glmnet` function and package

glmnet

What you may (should) be thinking now

Thought 1:

- "Yeah, OK, but this only works for some values of λ. So how would we choose λ in practice?"

This is actually quite a hard question. We'll talk about this in detail later → good way for purposes of prediction (for interpretation, harder)

Thought 2:

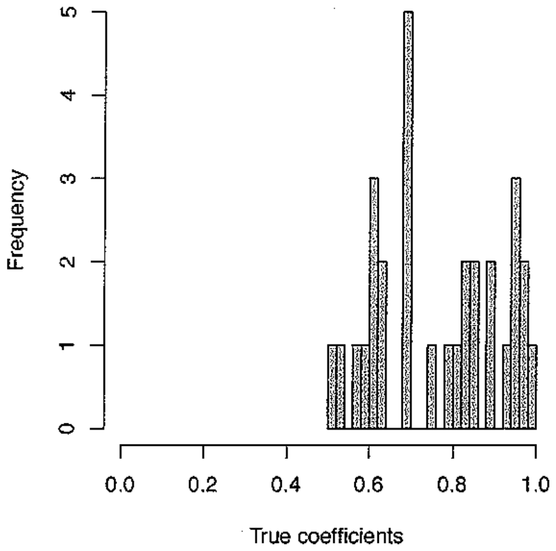
- "What happens when we none of the coefficients are small?"

In other words, if all the true coefficients are moderately large, is it still helpful to shrink the coefficient estimates? The answer is (perhaps surprisingly) still "yes". But the advantage of ridge regression here is less dramatic, and the corresponding range for good values of λ is smaller

James-Stein shrinkage

Example: moderate regression coefficients

Same setup as our last example: $n = 50$, $p = 30$, and $\sigma^2 = 1$. Except now the true coefficients are all moderately large (between 0.5 and 1). Histogram:

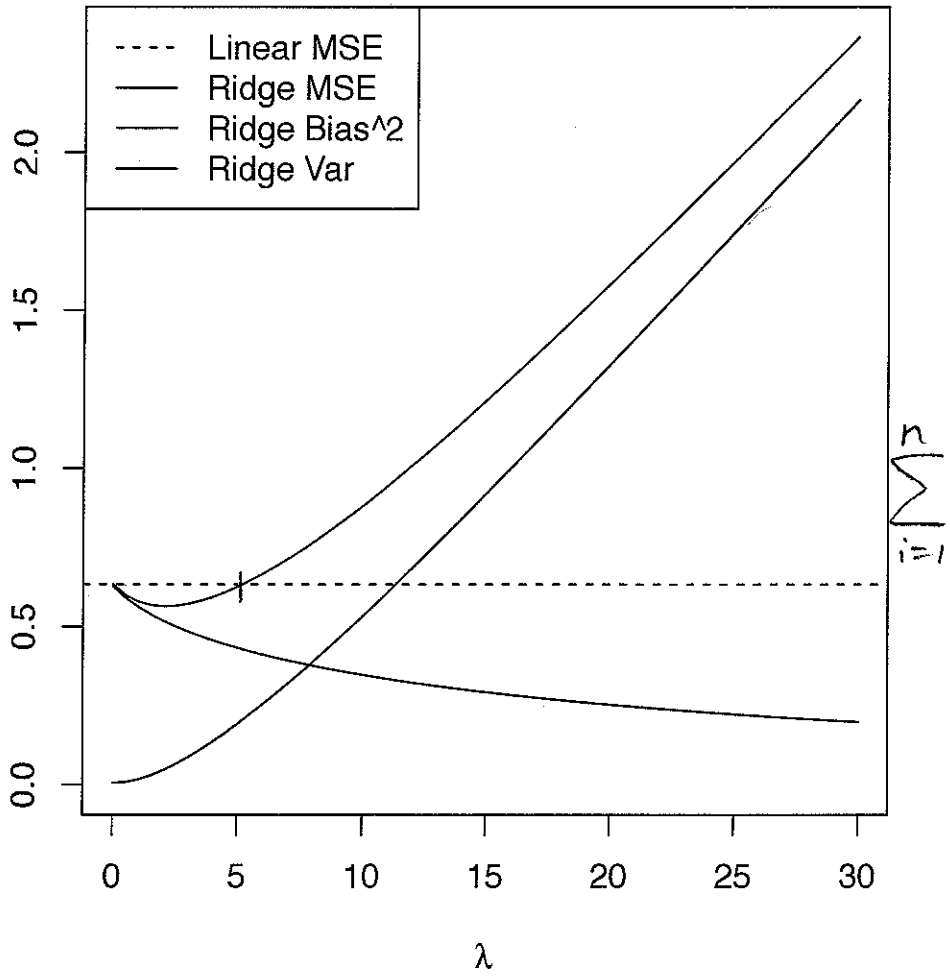


The linear regression fit:

Squared bias ≈ 0.006
Variance ≈ 0.628
Pred. error $\approx 1 + 0.006 + 0.628 \approx 1.634$

Why are these numbers essentially the same as those from the last example, even though the true coefficients changed?

Ridge regression can still outperform linear regression in terms of mean squared error:



$$\text{Bias}^2 = \sum_{i=1}^{n} \left(E[X_i^T \hat{\beta}] - X_i^T \beta^* \right)^2$$

Only works for λ less than ≈ 5, otherwise it is very biased. (Why?)

Variable selection

To the other extreme (of a subset of small coefficients), suppose that there is a group of true coefficients that are identically zero. This means that the mean response doesn't depend on these predictors at all; they are completely extraneous.

$$y_i = \beta_1^* x_{i1} + \dots \beta_{10}^* x_{i,10} + 0 \cdot x_{i,11} + \dots 0 \cdot x_{i,30}$$

The problem of picking out the relevant variables from a larger set is called variable selection. In the linear model setting, this means estimating some coefficients to be exactly zero. Aside from predictive accuracy, this can be very important for the purposes of model interpretation

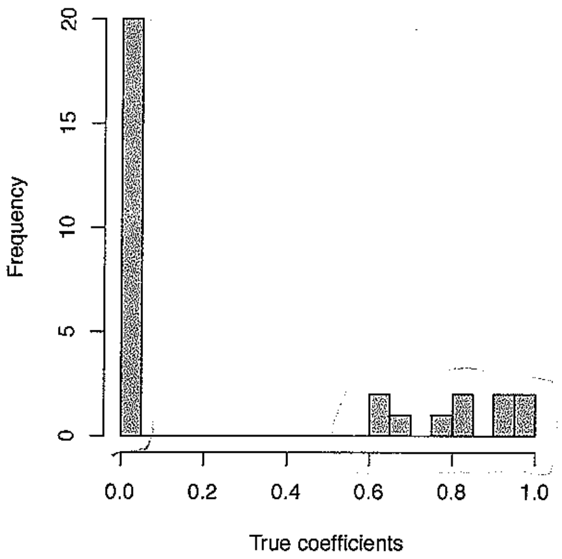
Thought 3:

- "How does ridge regression perform if a group of the true coefficients was exactly zero?"

The answer depends whether on we are interested in prediction or interpretation. We'll consider the former first

Example: subset of zero coefficients

Same general setup as our running example: $n = 50$, $p = 30$, and $\sigma^2 = 1$. Now, the true coefficients: 10 are large (between 0.5 and 1) and 20 are exactly 0. Histogram:

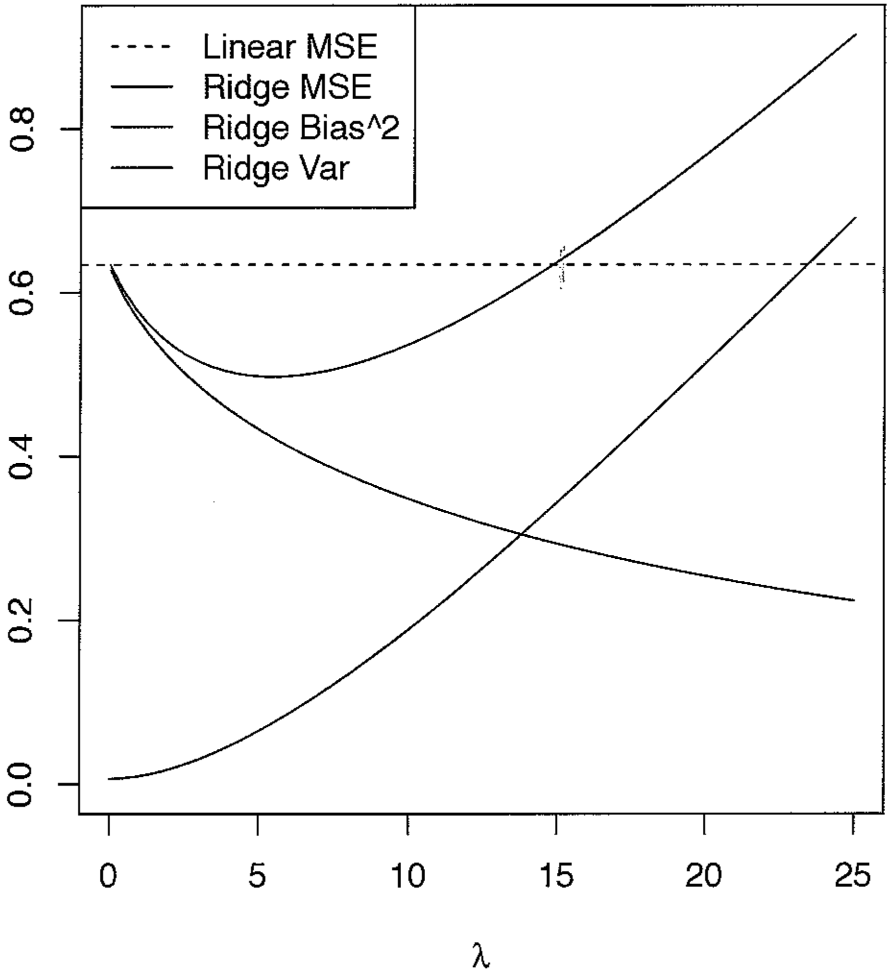


The linear regression fit:

Squared bias ≈ 0.006
Variance ≈ 0.627
Pred. error $\approx 1 + 0.006 + 0.627 \approx 1.633$

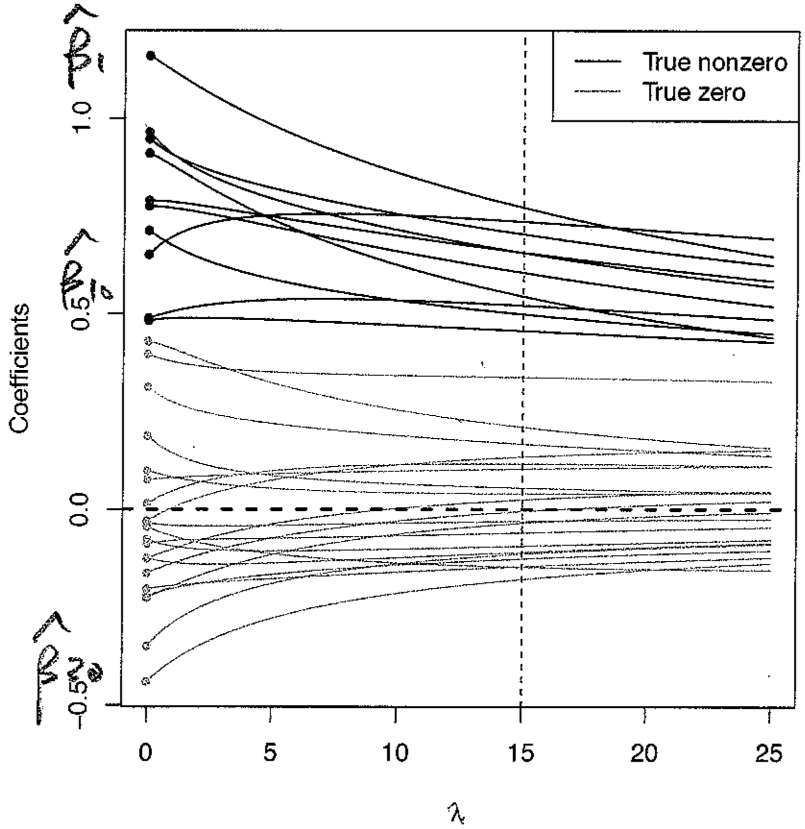
Note again that these numbers haven't changed

Ridge regression performs well in terms of mean-squared error:



Why is the bias not as large here for large λ?

Remember that as we vary λ we get different ridge regression coefficients, the larger the λ the more shrunken. Here we plot them again λ



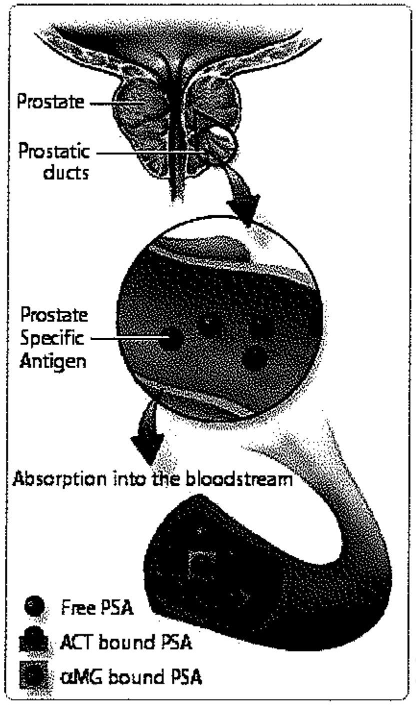
The red paths correspond to the true nonzero coefficients; the gray paths correspond to true zeros. The vertical dashed line at $\lambda = 15$ marks the point above which ridge regression's MSE starts losing to that of linear regression

An important thing to notice is that the gray coefficient paths are not exactly zero; they are shrunken, but still nonzero

Ridge regression doesn't perform variable selection

We can show that ridge regression doesn't set coefficients exactly to zero unless $\lambda = \infty$, in which case they're all zero. Hence ridge regression cannot perform variable selection, and even though it performs well in terms of prediction accuracy, it does poorly in terms of offering a clear interpretation

E.g., suppose that we are studying the level of prostate-specific antigen (PSA), which is often elevated in men who have prostate cancer. We look at $n = 97$ men with prostate cancer, and $p = 8$ clinical measurements.[1] We are interested in identifying a small number of predictors, say 2 or 3, that drive PSA

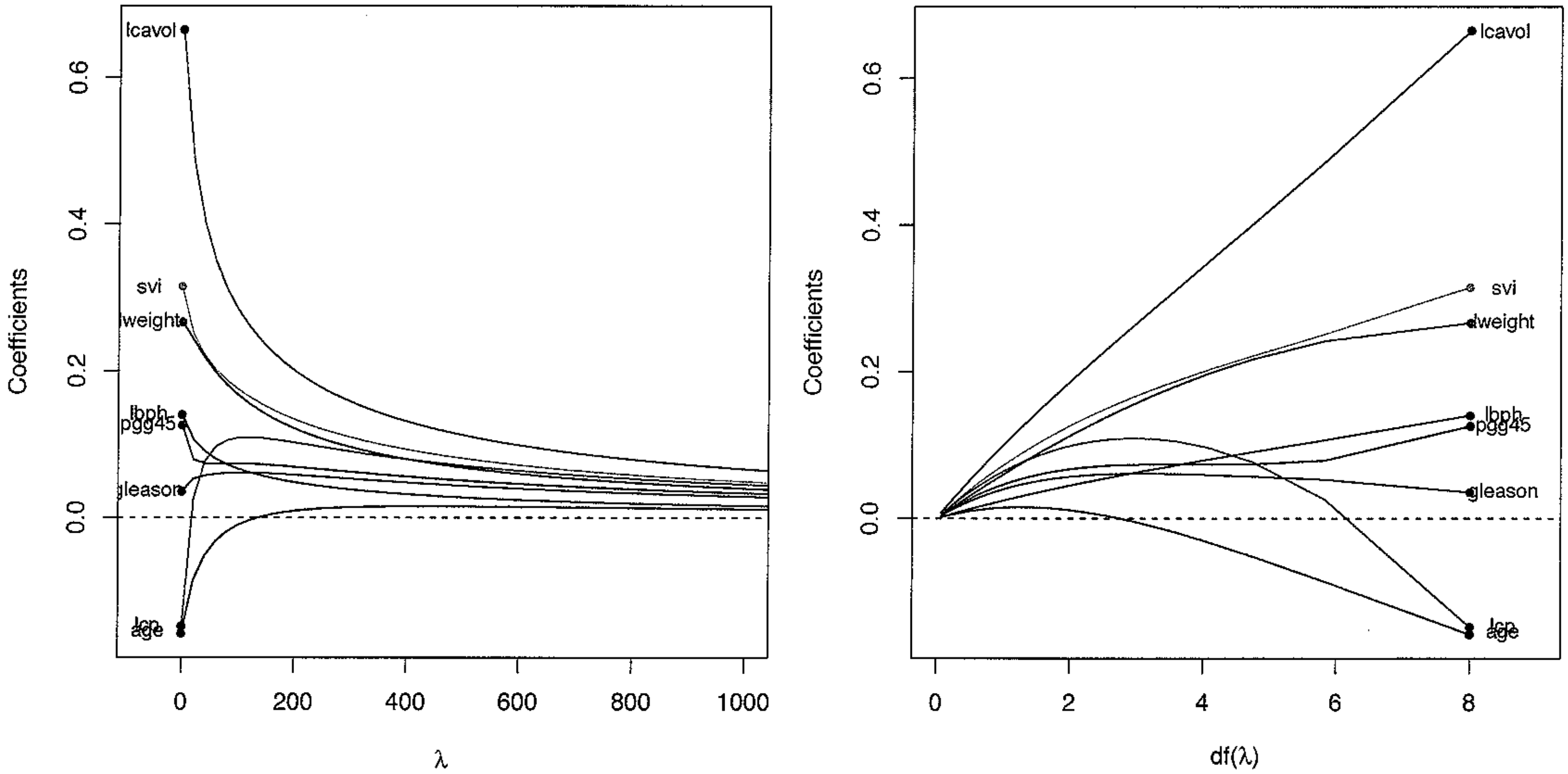


[2]

[1] Data from Stamey et al. (1989), "Prostate specific antigen in the diag..."
[2] Figure from `http://www.mens-hormonal-health.com/psa-score.html`

Example: ridge regression coefficients for prostate data

We perform ridge regression over a wide range of λ values (after centering and scaling). The resulting coefficient profiles:



This doesn't give us a clear answer to our question ...

loss + penalty

Recap: ridge regression

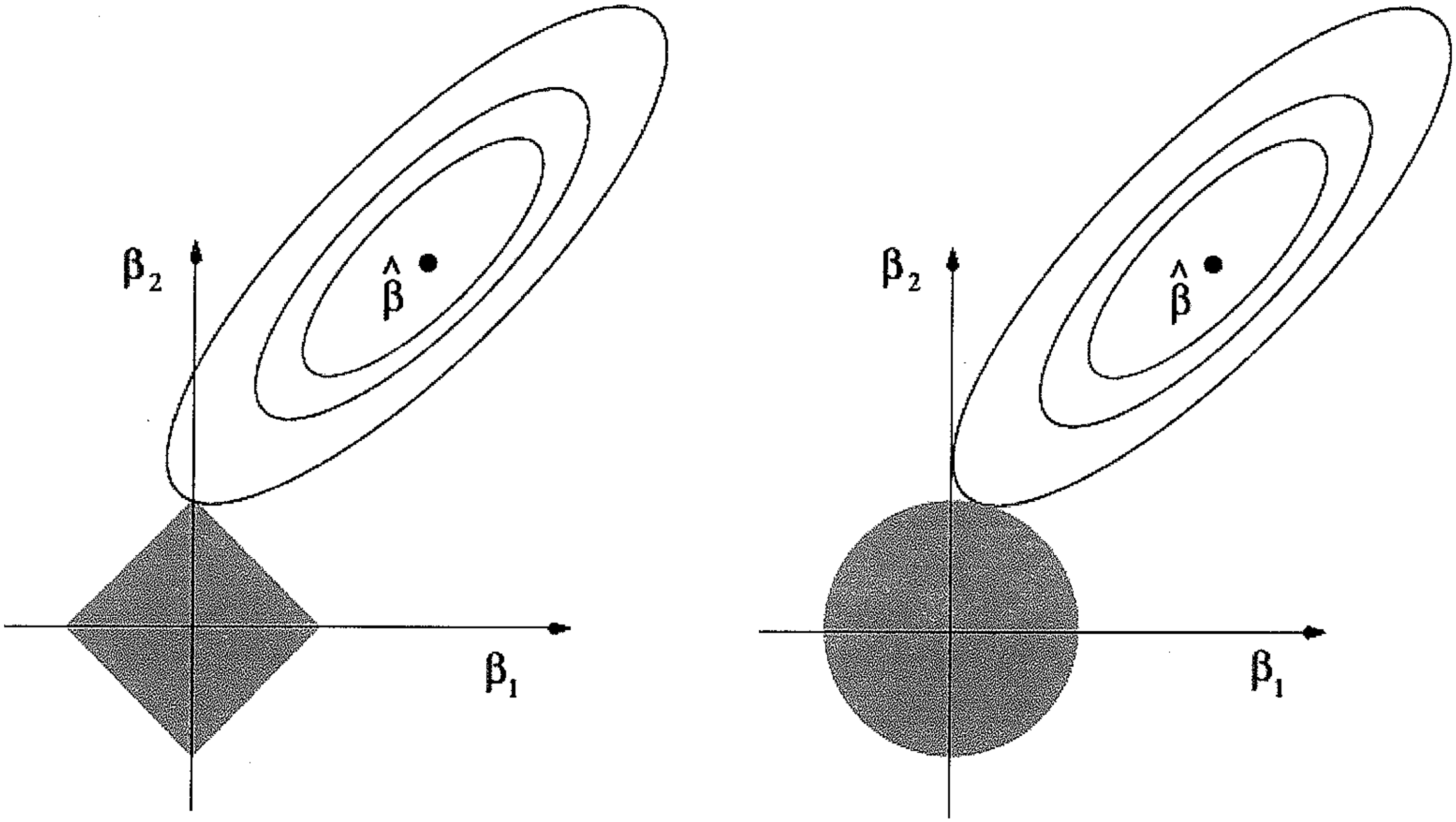
We learned ridge regression, which minimizes the usual regression criterion plus a penalty term on the squared ℓ_2 norm of the coefficient vector. As such, it shrinks the coefficients towards zero. This introduces some bias, but can greatly reduce the variance, resulting in a better mean-squared error

The amount of shrinkage is controlled by λ, the tuning parameter that multiplies the ridge penalty. Large λ means more shrinkage, and so we get different coefficient estimates for different values of λ. Choosing an appropriate value of λ is important, and also difficult. We'll return to this later

Ridge regression performs particularly well when there is a subset of true coefficients that are small or even zero. It doesn't do as well when all of the true coefficients are moderately large; however, in this case it can still outperform linear regression over a pretty narrow range of (small) λ values

Next time: the lasso

The lasso combines some of the shrinking advantages of ridge with variable selection



(From ESL page 71)