Lecture 4: Selection v. shrinkage

Last time

We examined some basic facts about regression and found that among all linear, unbiased estimates, OLS produces the smallest variance

We then relaxed the condition of unbiasedness and considered mean squared error as criterion for comparing estimators

We then introduce a new kind of estimate that acts by *shrinking* the OLS coefficients and studied its MSE

Today

There was something a bit fishy about our derivation on Friday; in particular, how are we supposed to choose the shrinkage parameter *in practice*?

Today we will talk about a practical application of the shrinkage idea, introducing James-Stein estimates and their extensions

We then motivate the shrinkage idea from the standpoint of a *penalized* least squares criterion

Finally, we introduce a slightly different penalty and derive yet another estimator, this one providing a mix of shrinkage and selection

Regression revisited

Recall the basic assumptions for the normal linear model

- We have a response y and (possible) predictor variables $x_1, x_2, \ldots, x_p$
- We hypothesize a *linear model* for the response

$$y = \beta_1^* x_1 + \beta_2^* x_2 + \cdots \beta_p^* x_p + \epsilon$$

where ϵ has a normal distribution with mean 0 and variance σ^2

Regression revisited

Then, we collect data...

- We have a *n* observations $y_1, \ldots, y_n$ and each response y_i is associated with the predictors $x_{i1}, x_{i2}, \ldots, x_{ip}$
- Then, according to our linear model

$$y_i = \beta_1^* x_{i1} + \beta_2^* x_{i2} + \cdots \beta_p^* x_{ip} + \epsilon_i$$

and we assume the ϵ_i are independent, each having a normal distribution with mean 0 and variance σ^2

Regression revisited

To estimate the coefficients $\beta_1^*, \beta_2^*, \ldots, \beta_p^*$ we turn to OLS, ordinary least squares (this is also maximum likelihood under our normal linear model)

- We want to choose $\beta_1, \beta_2, \ldots, \beta_p$ to minimize the OLS criterion

$$\sum_{i=1}^{n} [y_i - \beta_1 x_{i1} - \beta_2 x_{i2} - \cdots - \beta_p x_{ip}]^2$$

- You recall that this is just the sum of squared errors that we incur if we predict y_i with $\beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip}$

Orthogonal (well, orthonormal) regression

Computing the OLS estimates was easy; for example, we can differentiate this expression with respect to β_1 and solve to find

$$\widehat{\beta}_1 = \sum_i y_i x_{i1}$$

From here we found that $\mathsf{E}\widehat{\beta}_1 = \beta_1^*$ and that $\mathsf{var}(\widehat{\beta}_1) = \sigma^2$

Regression revisited

In general, whether they come from an orthogonal regression or not, OLS estimates have a lot to recommend them

- They are unbiased (repeated sampling)
- Gauss-Markov Theorem: Among all “linear” unbiased estimates, they have the smallest variance (this property is often called BLUE for “best unbiased linear estimate”)

Last time, we saw that in terms of mean squared error (MSE), we could do better if we give up the idea of unbiasedness

Shrinkage estimators for orthogonal regression

We will replace our OLS estimates $\widehat{\beta}_k$ with something slightly smaller

$$\widetilde{\beta}_k = \frac{1}{1+\lambda}\widehat{\beta}_k$$

If λ is zero, we get our OLS estimates back; if λ gets really really big, things crush to zero

The new estimate $\widetilde{\beta}_k$ is clearly biased

$$\mathsf{E}\widetilde{\beta}_k = \frac{1}{1+\lambda}\mathsf{E}\widehat{\beta}_k = \frac{1}{1+\lambda}\beta^*_k$$

Mean squared error

Last time we saw that the MSE for $\widetilde{\beta}_k$ could be written as

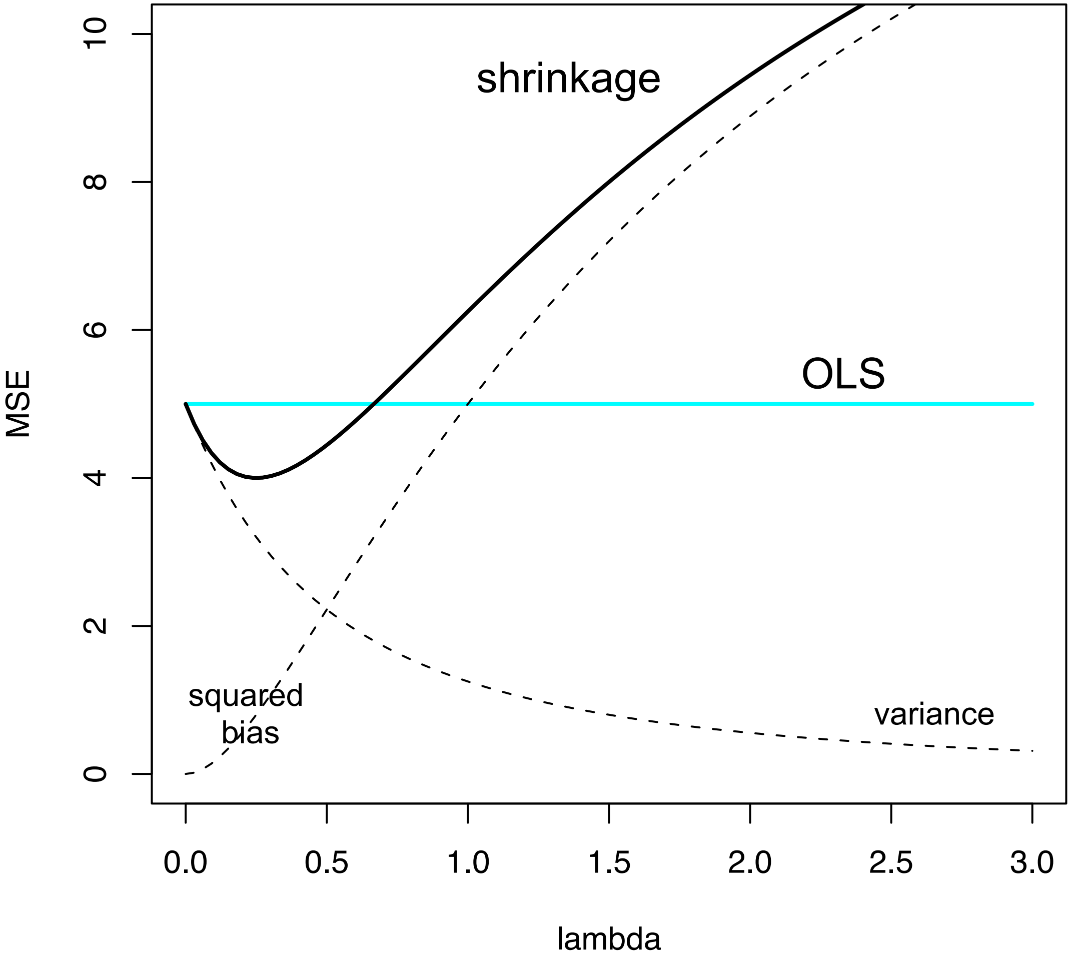
$$p\sigma^2 \left(\frac{1}{1+\lambda}\right)^2 + \left(\frac{\lambda}{1+\lambda}\right)^2 \sum_{k=1}^{p} {\beta_k^*}^2$$

The first term is the variance component; it is largest when λ is zero; the second is the squared bias it grows with λ

Shrinkage

In principle, with the right choice of λ we can get an estimator with a better MSE

The estimate is not unbiased, but what we pay for in bias, we make up for in variance



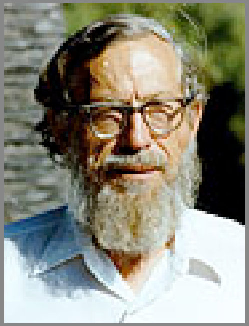
What’s missing?

The variance-bias decomposition and the accompanying plot suggest we can do better than OLS in terms of MSE if we can select a “good value” of λ ; the optimal being

$$\lambda = \frac{p\sigma^2}{\sum \beta_k^{*2}}$$

Do we ever really have access to this information?

Suppose we know σ^2



In the early 1960’s, Charles Stein (a statistician at Stanford University) working with a graduate student Willard James came up with the following specification

$$\widetilde{\beta}_k = \left(1 - \frac{(p-2)\sigma^2}{\sum_{k=1}^{p} \widehat{\beta}_k^2}\right) \widehat{\beta}_k$$

James-Stein estimators

With this definition, it is possible to show that the mean squared error is given by

$$\sum_{k=1}^{p} \mathsf{E}(\widetilde{\beta}_k - \beta_k^*)^2 = 2\sigma^2$$

which we can compare with the MSE for the OLS estimate

$$\sum_{k=1}^{p} \mathsf{E}(\widehat{\beta}_k - \beta_k^*)^2 = \sum_{k=1}^{p} \mathsf{var}\widehat{\beta}_k = p\sigma^2$$

and see that there is an improvement if $p \geq 3$

James-Stein estimators

This looks great! But are there still problems lurking here?

$$\widetilde{\beta}_k = \left(1 - \frac{(p-2)\sigma^2}{\sum_{k=1}^{p} \widehat{\beta}_k^2}\right) \widehat{\beta}_k$$

What happens if the quantity in parentheses is negative?
Is that even possible?

Let’s put on our probability hats*...

Suppose that all of the “true” parameters β_k^* are zero; then each term $\widehat{\beta}_k/\sigma$ has a standard normal and hence

$$\sum_{k=1}^{p} \frac{\widehat{\beta}_k^2}{\sigma^2}$$

has a chi-square distribution; so, what’s the chance that

$$\sum_{k=1}^{p} \frac{\widehat{\beta}_k^2}{\sigma^2} < p - 2$$

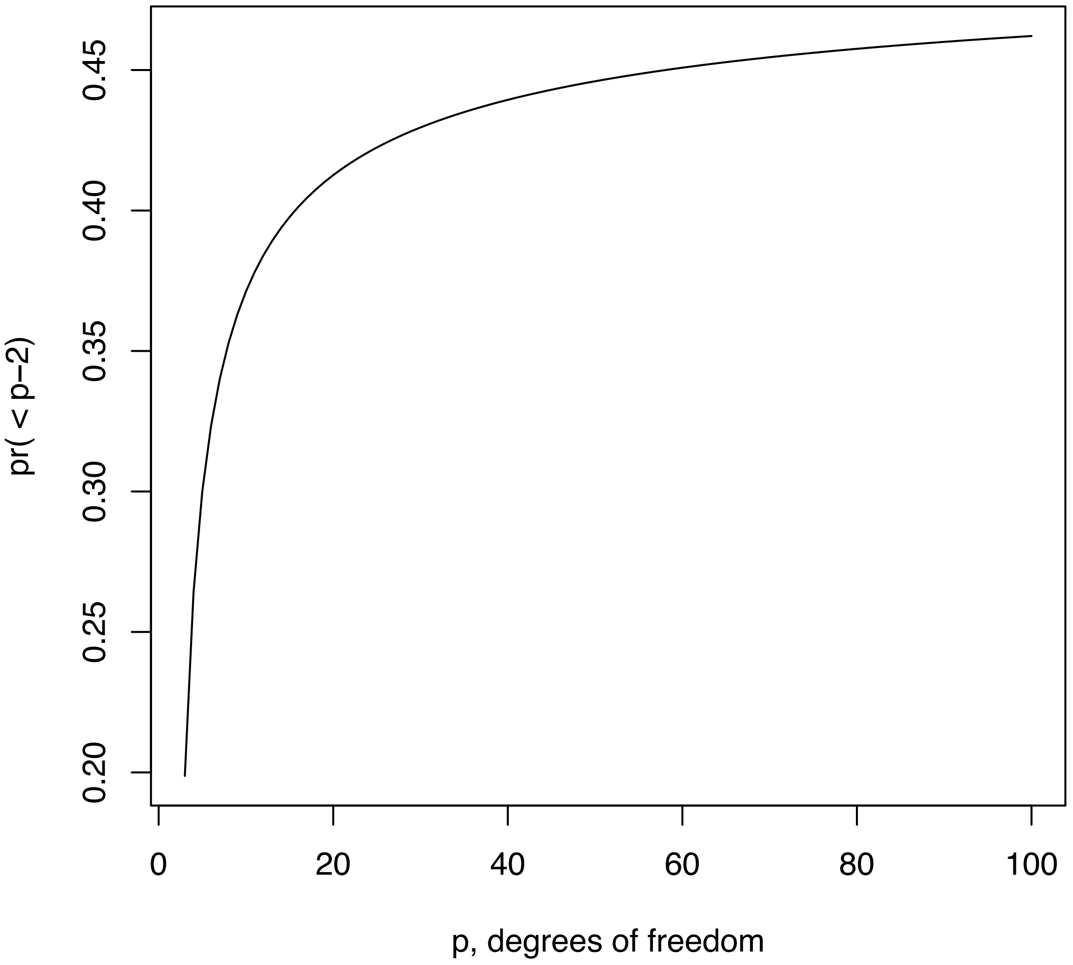
*Which are distinct from our statistics hats which look like ^ ... okay, a small joke

With our probability hats*...

The plot covers values of *p* from 3 to 100; recall that *p* represents the number of terms in our model

So, when our model is full of "small" coefficients (relative to the noise level), the chance of a negative shrinkage factor is not small

What should we do?



The first paper in your packet (at last!)

In the late 1960's, Stanley Sclove worked out a simple and elegant proposal

His proposal was to simply shrink the coefficients to zero if you get something negative; that is,

$$\widetilde{\beta}_k = \left(1 - \frac{(p-2)\sigma^2}{\sum_{k=1}^{p} \widehat{\beta}_k^2}\right)^+ \widehat{\beta}_k$$

where $(x)^+ = \max(x, 0)$

The first paper in your packet (at last!)

If σ^2 is unknown, he proposes that we use the criterion

$$\widetilde{\beta}_k = \left(1 - c\frac{RSS}{\sum_{k=1}^{p} \widehat{\beta}_k^2}\right)^+ \widehat{\beta}_k$$

for some value of *c;* have we seen this before?

Have we seen this before?

What if I rewrite this a bit, setting

$$F = \frac{\left(\sum_k \widehat{\beta}_k^2\right)/p}{RSS/(n-p)}$$

and then expressing Sclove’s estimate as

$$\widetilde{\beta}_k = \left(1 - \frac{c}{F}\right)^+ \widehat{\beta}_k$$

Have we seen this before?

Written this way, we set the coefficients to zero unless $F > c$

This gives us the (I think beautiful) result that we set all the coefficients to zero if we fail an *F*-test with significance level set by the value of *c*

If we pass this test, then we shrink the coefficients by an amount determined by how strongly the *F*-statistic protests against the null hypothesis

Let’s summarize a bit...

OLS estimates are great, but we can improve on them

We then considered shrinkage estimates and showed that if we were willing to give up a little in terms of bias, we could do better in terms of MSE

We then looked at making these things practical and saw variants of James-Stein estimators

Finally, we found a preliminary-testing procedure that either kills coefficients or keeps them (and shrinks them); this is kind of like model selection, except that it kills all the coefficients (unlike the keep-or-kill rules we saw with AIC and BIC)

But is any of this useful?

So far, we’ve only looked at orthogonal (or, worse yet, orthonormal) regression; you might wonder where this comes up

Aside from orthogonal polynomials from R, it comes up in a large number of places

In many signal-processing applications, we decompose our response vector into orthogonal pieces; Fourier analysis is one example, a wavelet decomposition is another

But is any of this useful?

We know that simple model selection via AIC and BIC can be applied to regular regressions

What about this shrinkage stuff?

A motivation

Last time we simply pulled the idea of shrinkage out of a hat; there is a fairly direct motivation in terms of a penalized least squares criterion

Suppose instead of OLS, we consider the following loss function

$$\sum_{i=1}^{n}[y_i - \beta_1 x_{i1} - \cdots - \beta_p x_{ip}]^2 + \lambda \sum_{k=1}^{p} \beta_k^2$$

A motivation

Now, if we have an orthogonal regression, we saw that we could write the OLS criterion as

$$\sum_i [y_i - \beta_1 x_{i1} - \cdots - \beta_p x_{ip}]^2 =$$

$$\sum_i y_i^2 + \beta_1^2 + \cdots + \beta_p^2 - 2\beta_1 \sum_i y_i x_{i1} - \cdots - 2\beta_p \sum_i y_i x_{ip}$$

and so adding the extra quadratic term

$$\lambda \sum_{k=1}^{p} \beta_k^2$$

doesn’t really complicate matters

A motivation

If we now minimize with respect to $\beta_1, \ldots, \beta_p$ we get exactly our shrinkage estimates

$$\widetilde{\beta}_k = \frac{1}{1+\lambda}\widehat{\beta}_k$$

A motivation

What happens when we don’t have an orthogonal model?

We recall that the ordinary OLS criterion gives rise to the normal equations

$$\boldsymbol{X}^t \boldsymbol{X} \boldsymbol{\beta} = \boldsymbol{X}^t \boldsymbol{y}$$

where $\boldsymbol{X}$ is the design matrix having , $[\boldsymbol{X}]_{ij} = x_{ij}$, $\boldsymbol{y}$ is the vector of response $\boldsymbol{y} = (y_1, \ldots, y_n)$, and $\boldsymbol{\beta}$ is the vector of regression coefficients $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)$

A motivation

If we add the penalty to our OLS criterion we now get

$$\sum_{i=1}^{n}[y_i - \beta_1 x_{i1} - \cdots - \beta_p x_{ip}]^2 + \lambda \sum_{k=1}^{p} \beta_k^2$$

which gives rise to a new set of equations

$$(\boldsymbol{X}^t\boldsymbol{X} + \lambda\boldsymbol{I})\boldsymbol{\beta} = \boldsymbol{X}^t\boldsymbol{y}$$

where $\boldsymbol{I}$ is the *pxp* identity matirx

Ridge regression

By adding this diagonal term, we arrive at something called ridge regression; why does this name make sense?

The second and third papers in your packet show the same kind of bias-variance decomposition as we derived for orthogonal regression, but for an ordinary set of predictors

What reasons might we have to think about adding a ridge like this?

Ridge regression

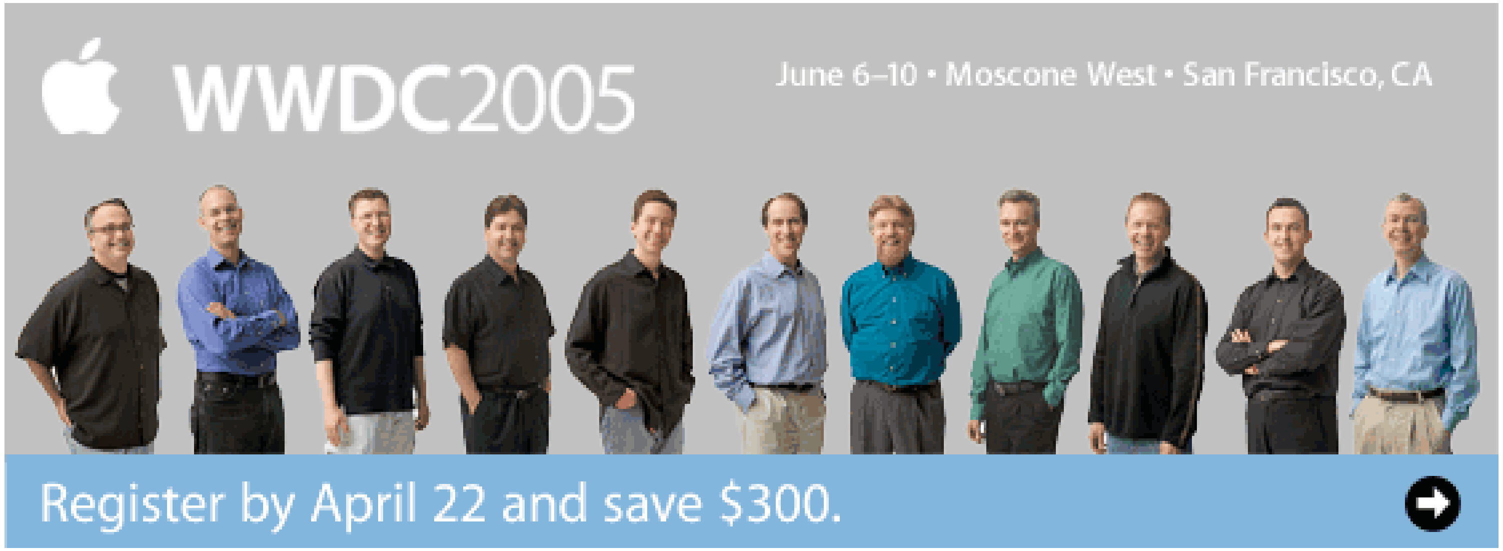
In R there is a function called `lm.ridge` that works just like `lm` but it slips in this extra penalty

Next time, you will get a homework assignment to play with ridge regression and compare your results to ordinary regression in a couple of different contexts

Speaking of computing...

Since we have spent a week + 1 day doing mainly math in what should be a computing class, I did want to bring up one observation

What's wrong with this picture?


WWDC2005
June 6–10 • Moscone West • San Francisco, CA
Register by April 22 and save $300.

Hint

It’s the same thing that’s wrong with this list

R Project Contributors

The current R is the result of a collaborative effort with contributions from all over the world. R was initially written by Robert Gentleman and Ross Ihaka —also known as "R & R" of the Statistics Department of the University of Auckland. Since mid-1997 there has been a core group with write access to the R source, currently consisting of

- Douglas Bates
- John Chambers
- Peter Dalgaard
- Robert Gentleman
- Kurt Hornik
- Stefano Iacus
- Ross Ihaka
- Friedrich Leisch
- Thomas Lumley
- Martin Maechler
- Duncan Murdoch
- Paul Murrell
- Martyn Plummer
- Brian Ripley
- Duncan Temple Lang
- Luke Tierney