

Math 455

Class 39

May 1

Midterm 2 has been graded.

Median ≈ 130 (of 174)

Standard Deviation ≈ 18 .

Sample Final Exam has been posted.

This is HW 12.

Requests for regrades must be in writing.

I will look at your whole test.

Your score may go up OR DOWN.

Last time: the picture on the front cover of the textbook is explained in Ch. 11.3 & 11.4.

Ridge regression & LASSO regression.

Confidence ellipses where SSR is constant on each ellipse are in Ch. 3. They come from the idea: $\hat{\beta}_1, \hat{\beta}_2$ are RVs with mean, variance & covariance; these numbers define the confidence ellipse.

Questions about this?!

Q1: This whole thing depends on an undetermined parameter λ (or k for the constraint).

How do we know what λ should be?

Q2: What is the effect of increasing / decreasing λ ?

Q3: What are the differences between ridge / LASSO and OLS regression?

Q4: Why these funny names?

Q5: How do we do this in R?

Another way of dealing with this: treat the symptom directly:

If the problem is "large and unstable coefficient estimates," let's penalize large coefficients.

Instead of minimizing $SSR = \sum_{i=1}^n e_i^2$

minimize: $SSR + \text{penalty term}$.

Two typical penalty terms:

"Ridge" and "LASSO"

Ridge: $\lambda (\hat{\beta}_1^2 + \dots + \hat{\beta}_p^2)$

LASSO: $\lambda (|\hat{\beta}_1| + \dots + |\hat{\beta}_p|)$.

Notice $\hat{\beta}_0$ is NOT (typically) included in either penalty term.

These penalty terms explain the picture on the front cover.

ICBST: penalty terms are equivalent to

"Constrained regression"

Another way of making the coefficient estimates smaller is to say, "let's minimize SSR

subject to the constraint

$$\hat{\beta}_1^2 + \dots + \hat{\beta}_p^2 \leq K \quad (\text{Ridge})$$

$$\text{or } |\hat{\beta}_1| + \dots + |\hat{\beta}_p| \leq K. \quad (\text{Lasso})$$

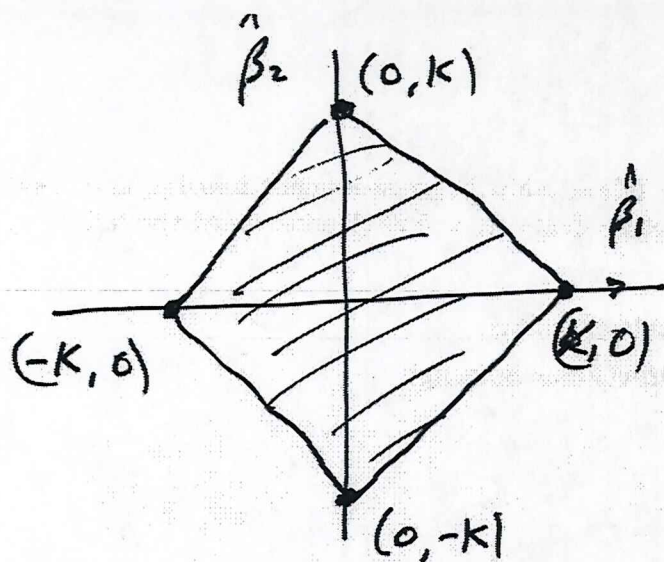
this is equivalent to the penalty term in the sense that for any choice of λ , there is a corresponding choice of K that gives the same result.

Explaining the picture: let $p=2$, so we can draw something in the $\hat{\beta}_1, \hat{\beta}_2$ plane.

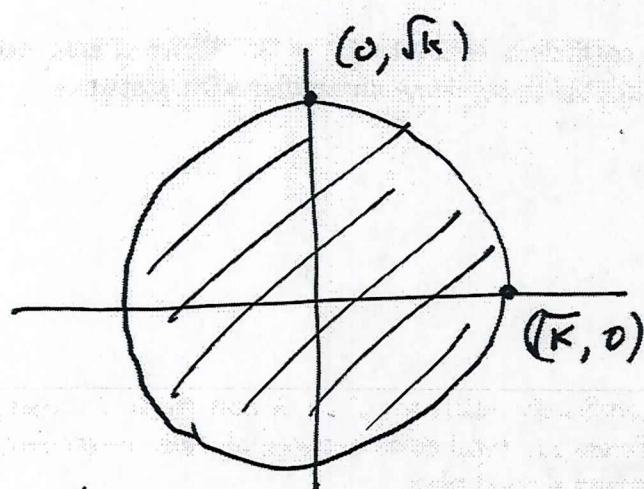
Let's draw the regions:

$$|\hat{\beta}_1| + |\hat{\beta}_2| \leq K$$

$$\hat{\beta}_1^2 + \hat{\beta}_2^2 \leq K.$$



Shaded region
is $|\hat{\beta}_1| + |\hat{\beta}_2| \leq K$.



Shaded region
is $\hat{\beta}_1^2 + \hat{\beta}_2^2 \leq K$.

What point ~~with~~ $(\hat{\beta}_1, \hat{\beta}_2)$ will be selected
by the constrained minimization procedure?

Picture on front cover shows unconstrained
minimization & confidence ellipses with increasing
SSR (constant on each ellipse).

Another difference: OLS estimators are unbiased but ridge & lasso estimates are NOT unbiased.

We are willing sometimes to ~~see~~ sacrifice unbiasedness for reduced variance, because of the

"Bias - Variance Tradeoff" (Wackerly Ch. 8)
(Faraway p. 177)

A2: If the penalty $\lambda = 0$, then we just have OLS regression. If λ is very large, all coefficient estimates must be approximately 0.

Increasing λ corresponds to decreasing K and vice versa.

Large K : the circle and diamond on the cover become large enough to include the OLS estimate and there is no difference between the ridge/lasso est. and the OLS estimate. Small K : the circle and diamond are small and ridge/lasso coeff. estimates

constraint region will be a corner, unless there is some strange coincidence, like the axes of the ellipse being at exactly 45° to the β_1, β_2 axes of the plane.

So: standard conceptual question about Lasso and ridge:

- Are Lasso and ridge variable selection procedures,
- Lasso is a variable selection procedure, but ridge is not.

A4, continued: the origin of the name ridge.

The efforts to explain this are not as convincing as for Lasso. One explanation is that the ridge coefficient estimates lie in the intersection of an ellipsoid and sphere where there is more stability to the estimates. This region is the "ridge".

A3: We saw that Lasso does variable selection but OLS does NOT and ridge does NOT.

A4: Lasso is an acronym:

Least Absolute Shrinkage and Selection Operator.

(The remark on p. 177 is wrong and noted in the errata.)

Q6: The remark suggests, as does the acronym, that LASSO is a variable selection procedure. Why is this so?

A6: This is the point of the picture on the front cover. Notice that in the picture, the estimate of $\hat{\beta}_1, \hat{\beta}_2$ for the Lasso is on the β_2 axis, so $\beta_1 = 0$. This means that the X_1 -variable corresponding to β_1 has been eliminated from the model.

The ridge procedure did NOT eliminate X_1 .

Q7: Why did this happen?

A7: The Lasso constraint region is a diamond: the corners of the diamond stick out. When we consider expanding concentric ellipses of constant SSR, the point we hit first in the

are all close to 0.

A1: We will generally choose λ by a procedure called "Cross-Validation".

~~the divide~~

This is a generalization of the idea of "training set" and "test set".

Instead of dividing into "training" and "test",
divide into m parts

train on $m-1$ parts, test on remaining part.

We can do this m times.

Round 1:

train	train	train	train	test
-------	-------	-------	-------	------

Round 2:

train	train	train	test	train
-------	-------	-------	------	-------

etc.

Idea: Choose a range of λ s, $\lambda_1, \dots, \lambda_m$.

Round 1: fit model to training set using penalty λ_1
Measure error $SSR + \lambda_1 (\quad)$ on TEST.

Round 2: fit model to training set with
penalty λ_2 . measure error $SSR + \lambda_2 ()$
on TEST.

etc.

Now choose the λ_i corresponding to the
minimal SSR on TEST.

Issues with this CV procedure are similar
to issues with Bootstrap.

Some data sets have structure that is not
preserved by "chopping up" procedures.