Math 455 Class 38 May 2

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?

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

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.

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

We are willing sometimes to 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

are all close to 0.

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

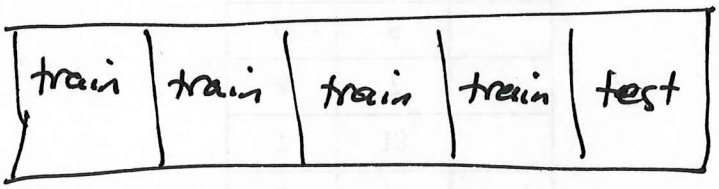
This is a generalization of the idea of "training set" and "test set".

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

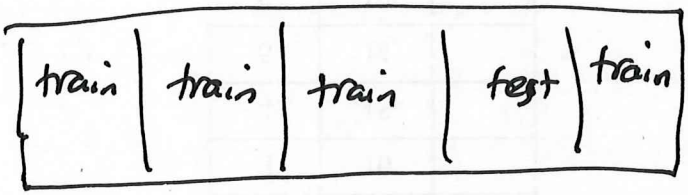
train on $m-1$ parts, test on remaining part.

We can do this m times.

Round 1:



Round 2:



etc.

Idea: Choose a range of $\lambda_0, \lambda_1, \ldots, \lambda_m$.

Round 1: fit model to training set using penalty λ_1

Measure error SSR + λ_1 ( ) on TEST.

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