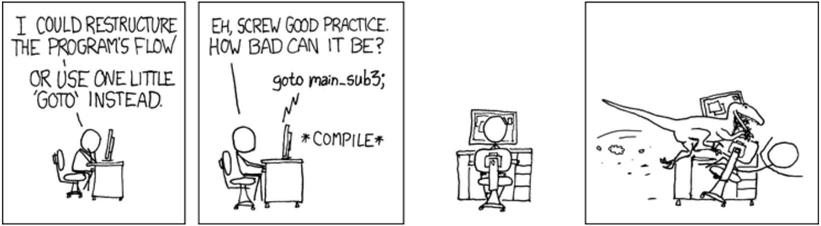
STAT406- Methods of Statistical Learning Lecture 5

Matias Salibian-Barrera

UBC - Sep / Dec 2017



http://xkcd.com/292/

Comparing models

- AIC suggests a submodel
- Prediction-wise the full model is better
- AIC can be highly variable

“Smoother” model selection

- Ridge regression
- Can be thought as a type of feature selection
- It is a member of a larger class called “shrinkage methods”
- However, its origins are rather different

Without loss of generality...

- If covariates are centered, $\sum_{i=1}^{n} \mathbf{x}_i = \mathbf{0}$

$$\arg\min_{\alpha, \boldsymbol{\beta}} \sum_{i=1}^{n} \left(y_i - \alpha - \boldsymbol{\beta}' \mathbf{x}_i \right)^2$$

satisfies

$$\hat{\alpha}_n = \frac{1}{n} \sum_{i=1}^{n} y_i = \bar{y}_n \, ,$$

and

$$\hat{\boldsymbol{\beta}}_{LS} = \left(\mathbf{X}' \mathbf{X} \right)^{-1} \mathbf{X}' \mathbf{Y} \, ,$$

Without loss of generality...

- We can always assume that

$$\sum_{i=1}^{n} \mathbf{x}_i = \mathbf{0}$$

and hence

$$\hat{\alpha}_n = \frac{1}{n} \sum_{i=1}^{n} y_i = \bar{y}_n \, ,$$

- In what follows, there is no intercept

Shrinkage methods

- When covariates are correlated, LS estimators can be highly variable

$$\hat{\beta}_{LS} = (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}' \mathbf{Y}$$

$$\mathrm{var}\left(\hat{\beta}_n\right) = \sigma^2 (\mathbf{X}'\mathbf{X})^{-1}$$

- When $\mathbf{X}'\mathbf{X}$ is close to singular...

Ridge Regression

- One way to "avoid" this problem is to add a "ridge" to $\mathbf{X}'\mathbf{X}$...

$$\hat{\beta}_{RR} = \left(\mathbf{X}'\mathbf{X} + \lambda\,\mathbf{I}_p\right)^{-1}\mathbf{X}'\mathbf{Y}$$

where $\lambda > 0$ and

$$\mathbf{I}_p = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ 0 & \cdots & \ddots & 0 \\ 0 & \cdots & \cdots & 1 \end{pmatrix}$$

Ridge Regression

- This is equivalent to solving

$$\min_{\boldsymbol{\beta}} \; \sum_{i=1}^{n} \left(y_i - \boldsymbol{\beta}' \mathbf{x}_i \right)^2 \, + \, \lambda \, \|\boldsymbol{\beta}\|_2^2$$

Ridge Regression

- And also equivalent to solving

$$\min_{\boldsymbol{\beta}} \ \sum_{i=1}^{n} \left(y_i - \boldsymbol{\beta}' \mathbf{x}_i \right)^2$$

subject to

$$\sum_{j=1}^{p} \beta_j^2 \ \leq \ C$$

for some $C > 0$

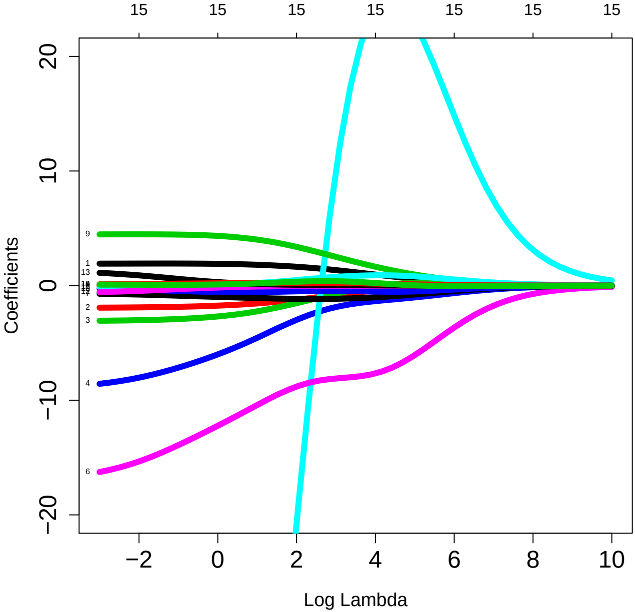
Bias / variance trade-off

- Ridge regression was originally proposed as a “hack” to “push” $\mathbf{X}'\mathbf{X}$ away from singularity
- It can also be thought as a way of reducing the variance of $\hat{\beta}_n$
- This may increase the bias of the estimator, but if the variance is reduced even more, we might gain overall in expected squared error performance...

Ridge regression

- We now have a sequence ("path") of estimators (one for each $\lambda > 0$)
- $\mathbf{X}'\mathbf{X} + \lambda \, \mathbf{I}_p$ is always non-singular for $\lambda > 0$ (why?)
- Why are they called "shrinkage methods"?

Air pollution data



Questions

- What does λ measure?
- How do I choose one among these infinitely many "solutions"?

How do we select λ?

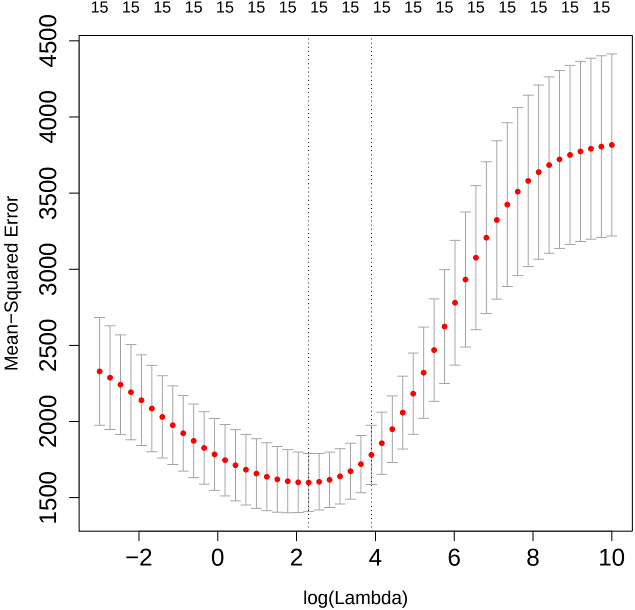
How can we select λ?

How do we select λ?

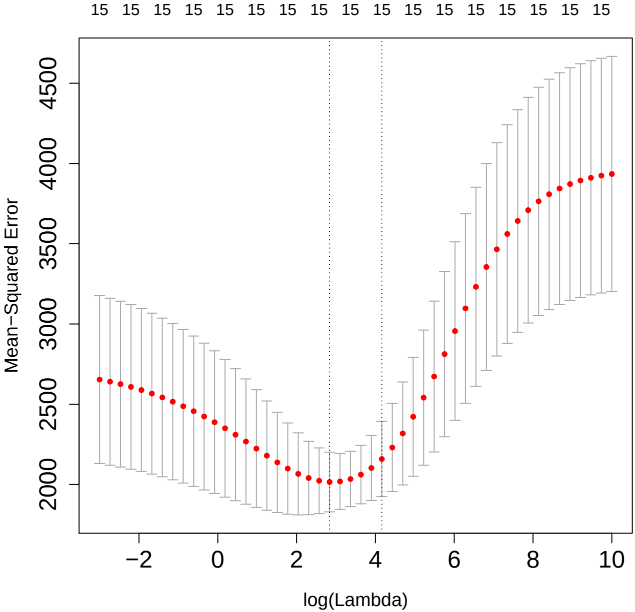
CV!

```
library(glmnet)
airp <- read.table('rutgers-lib-30861_CSV-1.csv'
      header=TRUE, sep=',')

y <- as.vector(airp$MORT)
xm <- as.matrix(airp[, -16])
lambdas <- exp( seq(-3, 10, length=50))

set.seed(123)
tmp <- cv.glmnet(x=xm, y=y, lambda=lambdas,
      nfolds=5, alpha=0,
      family='gaussian')
```

Air pollution data



Air pollution data - another run



Questions

- How are the standard errors estimated?
- Can we use AIC to compare these models?
- Why or why not?
  - If the answer is yes, how?
  - If the answer is no, why not?

CV

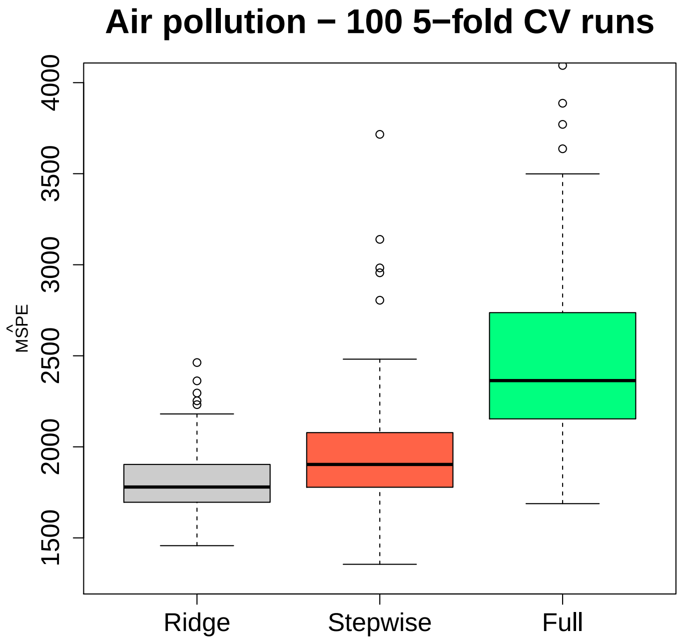
Cross validation selects

$$\lambda_{\text{op}} \approx \exp(3)$$

$$\text{edf} \approx 13$$

Stepwise selects

```
Call:
lm(formula = MORT ~ NONW + EDUC + JANT + SO.
+ PREC + JULT + POPN, data = airp)

Coefficients:
(Intercept)     NONW   EDUC       JANT
  1429.1866   5.2161  -16.9656  -1.8934

    SO.   PREC     JULT      POPN
0.2253  1.6485  -2.3006  -62.0118
```



Air pollution − 100 5−fold CV runs
4000
3500
3000
2500
2000
1500
$\widehat{MSPE}$
Ridge
Stepwise
Full

Sometimes...

- Selecting variables is not always necessary in terms of prediction accuracy.
- One such an example is discussed on Github (Lecture 4)...
- ...and revised in Lecture 5. Read it carefully.