STAT406- Methods of Statistical Learning Lecture 5

Matias Salibian-Barrera

UBC - Sep / Dec 2018

http://xkcd.com/292/

Comparing models

- For Gaussian errors we have

$$\text{AIC} = n \log\left(\frac{\text{RSS}}{n}\right) + 2p + \text{constant}$$

where

$$\text{RSS} = \sum_{i=1}^{n} r_i^2 ,$$

the **constant** depends on n, not on p

Comparing models

- However, many times we find

$$\text{AIC} = \frac{1}{n}\frac{1}{\hat{\sigma}^2}\left(\text{RSS} + 2\,p\,\hat{\sigma}^2\right) + \text{constant}$$

(e.g. [JWHT13])

Where does this expression come from?

Comparing models

- Regularity assumptions are needed
  - This is an asymptotic approximation, n should be large
  - One of the models should include truth
  - $\theta_1 \neq \theta_2 \ \Rightarrow f(y, \theta_1) \neq f(y, \theta_2)$
  - Standard large-sample MLE assumptions to obtain asymptotic normality

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 “prediction taming”
- 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

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

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

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

Ridge Regression

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

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

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

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

Ridge Regression

- And also equivalent to solving

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

subject to

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

for some $C > 0$

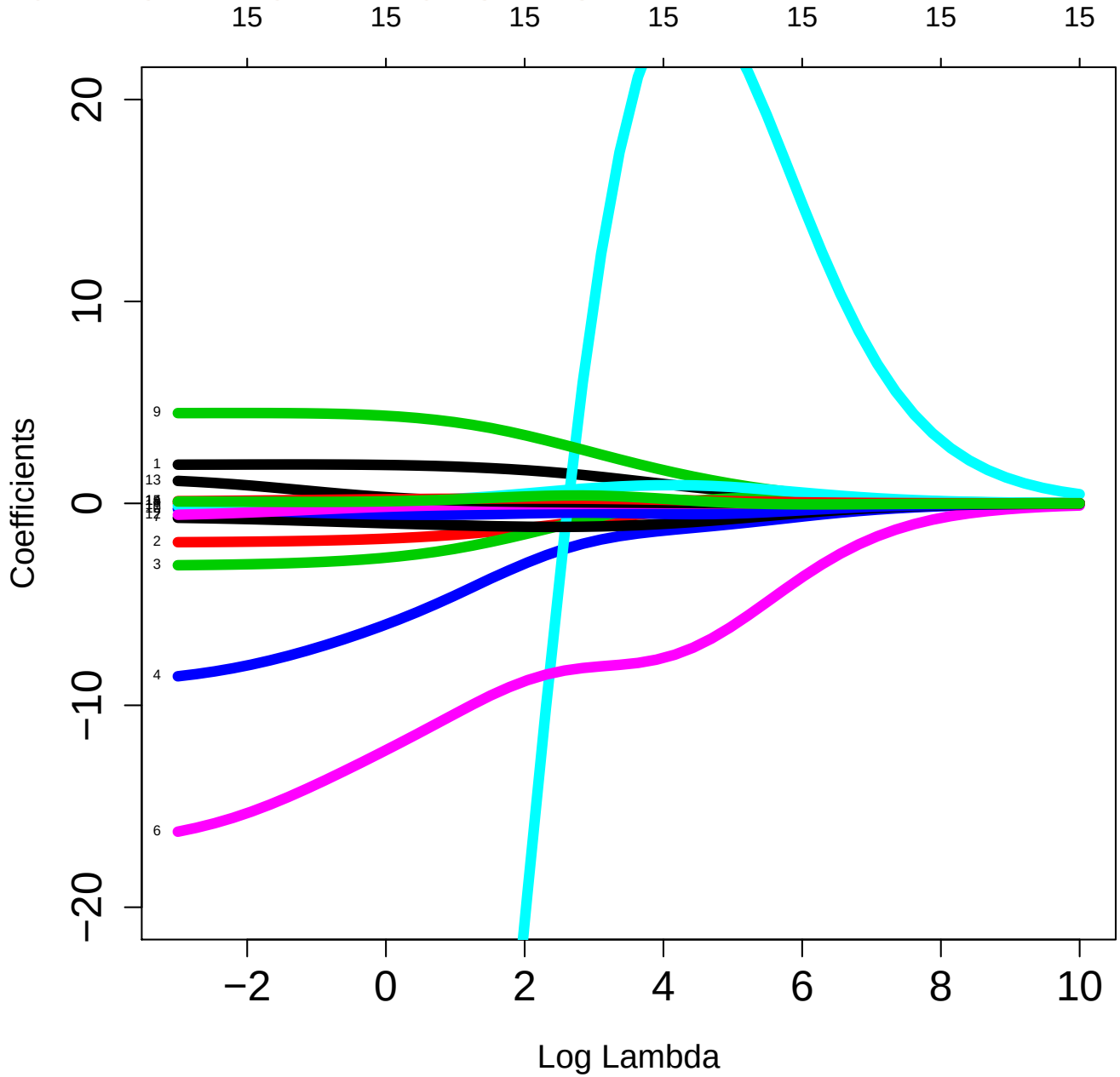
Bias / variance trade-off

- 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”?

Effective degrees of freedom

- How many “effective” parameters are we using?
- In linear regression, we have p parameters
- A more general definition is as follows. For a fitting method producing $\hat{y}_1$, $\hat{y}_2$, $\ldots$, $\hat{y}_n$,

$$\text{edf} = \frac{1}{\sigma^2}\sum_{i=1}^{n}\text{cov}\left(\hat{y}_i, y_i\right)$$

Efron, B. (1986). How biased is the apparent error rate of a prediction rule? Journal of the American Statistical Association, 81(394):461-470.

Effective degrees of freedom

- It is easy to see that for least squares predictors, we have

$$\hat{\mathbf{y}} = \mathbf{H}\,\mathbf{y}$$

with

$$\mathbf{H} = \mathbf{X}\,(\mathbf{X}'\mathbf{X})^{-1}\,\mathbf{X}'$$

and

$$\text{edf} = \frac{1}{\sigma^2}\sum_{i=1}^{n}\text{cov}\,(\hat{y}_i, y_i) = \text{trace}\,(\mathbf{H}) = p$$

Effective degrees of freedom

- More in general, for any linear predictor

$$\hat{\mathbf{y}} = \mathbf{S}\,\mathbf{y}$$

we have

$$\text{edf} = \text{trace}(\mathbf{S}) = \sum_{i=1}^{n} \mathbf{S}_{i,i}$$

Effective degrees of freedom

- The ridge regression fit satisfies

$$\hat{\mathbf{y}} = \mathbf{S}_\lambda \mathbf{y}$$

where

$$\mathbf{S}_\lambda = \mathbf{X} \left(\mathbf{X}'\mathbf{X} + \lambda \mathbf{I}_p\right)^{-1} \mathbf{X}'$$

$$\text{trace}\left(\mathbf{S}\right) = ?$$

Effective degrees of freedom

- Using the singular value decomposition (SVD) of $\mathbf{X}$

$$\mathbf{X} = \mathbf{U}\mathbf{\Lambda}\mathbf{V}'$$

where $\mathbf{U} \in \mathbb{R}^{n \times p}$, $\mathbf{V} \in \mathbb{R}^{p \times p}$ with

$$\mathbf{U}'\mathbf{U} = \mathbf{I}_p = \mathbf{V}'\mathbf{V}$$

and

$$\mathbf{\Lambda} = \text{diag}\left(d_1, \ldots, d_p\right) ,$$

we have

$$\text{trace}\left(\mathbf{S}\right) = \sum_{i=1}^{p} \left(\frac{d_i^2}{d_i^2 + \lambda} \right)$$

Effective degrees of freedom

- For example, in the Air Pollution data example, if we use

$$\lambda = \exp(6)$$

we get

$$\text{edf} = 9.9$$

How do we select λ?

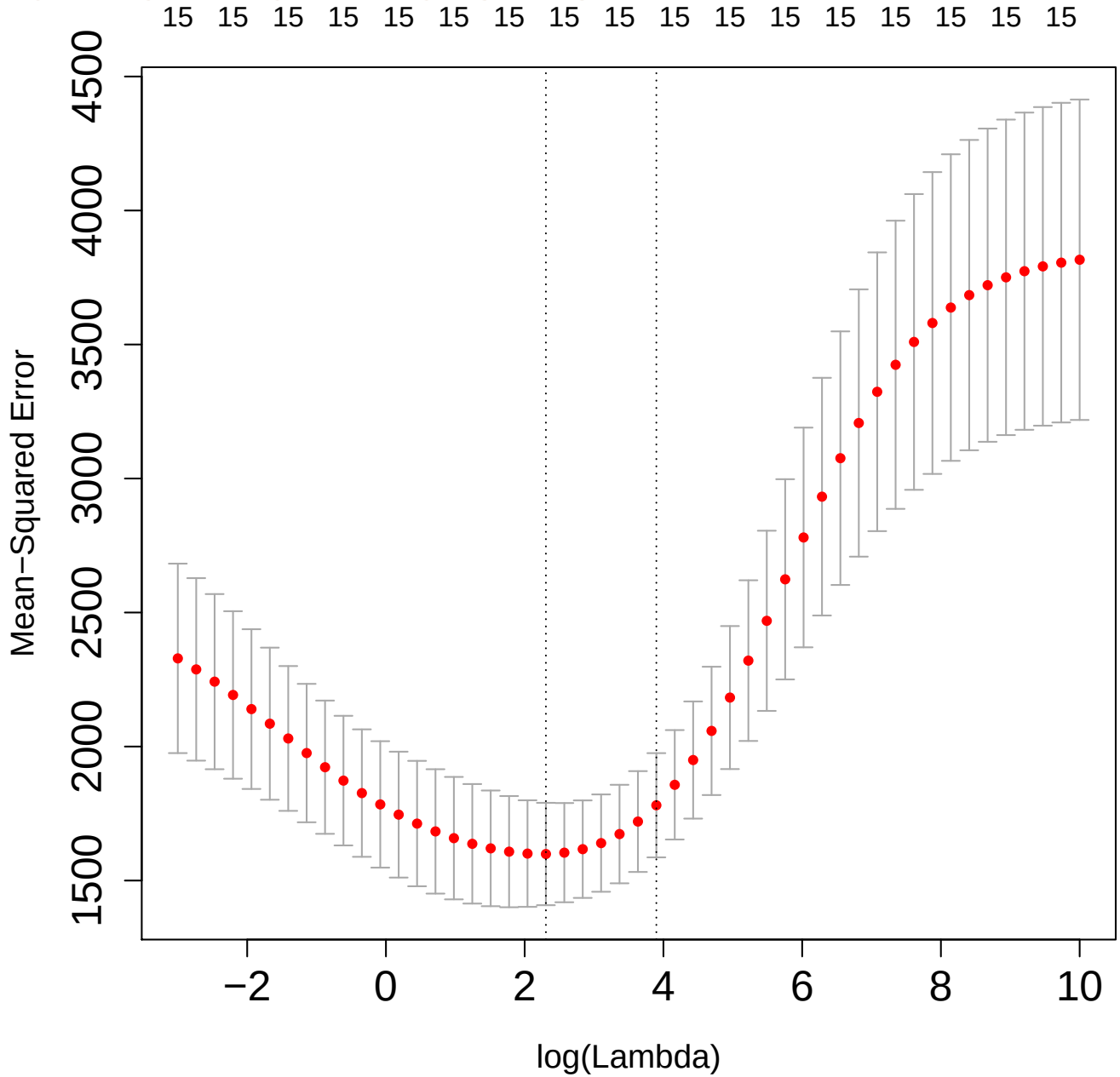
How can we select λ?

How do we select λ?

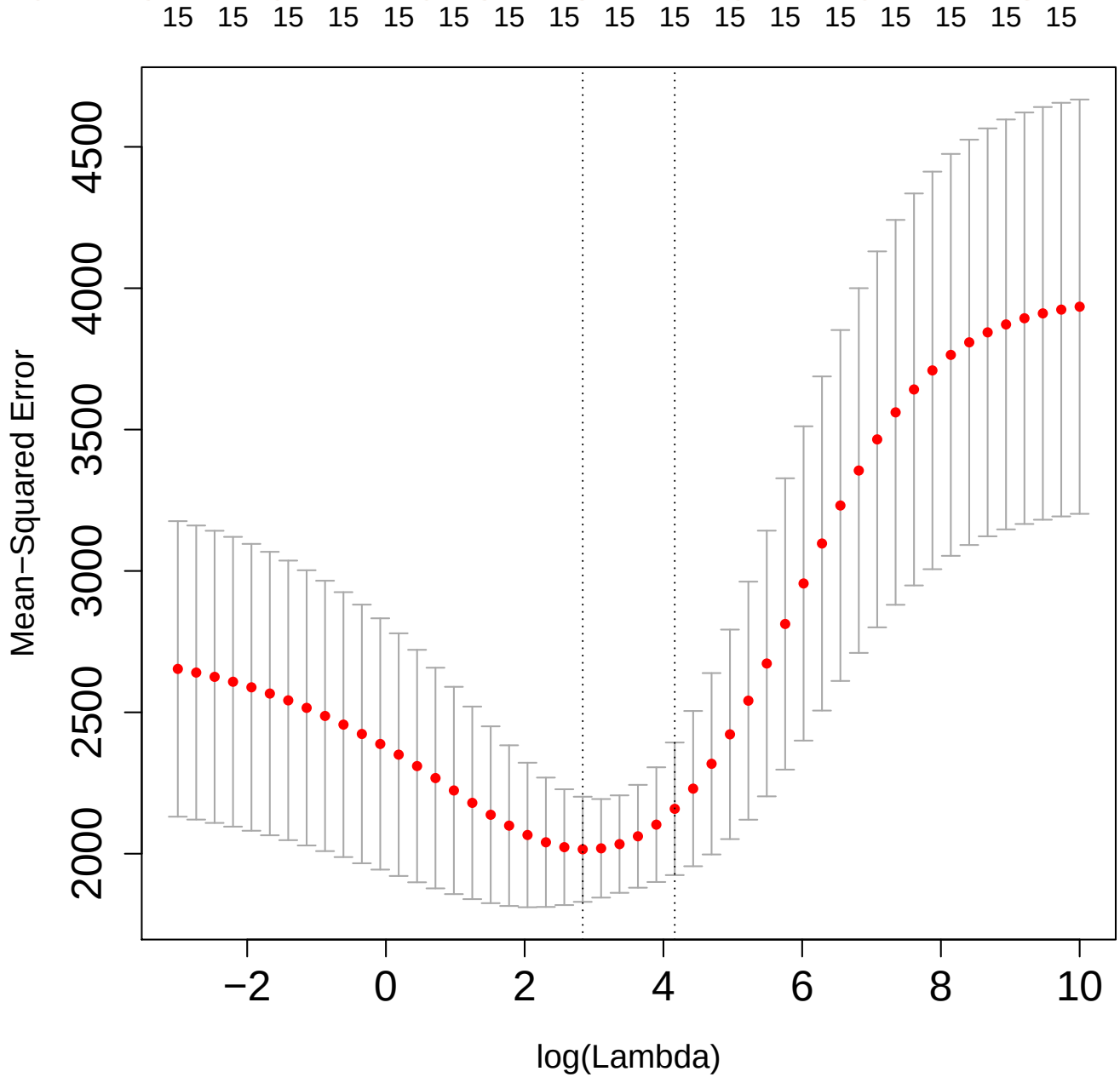
CV!

```
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

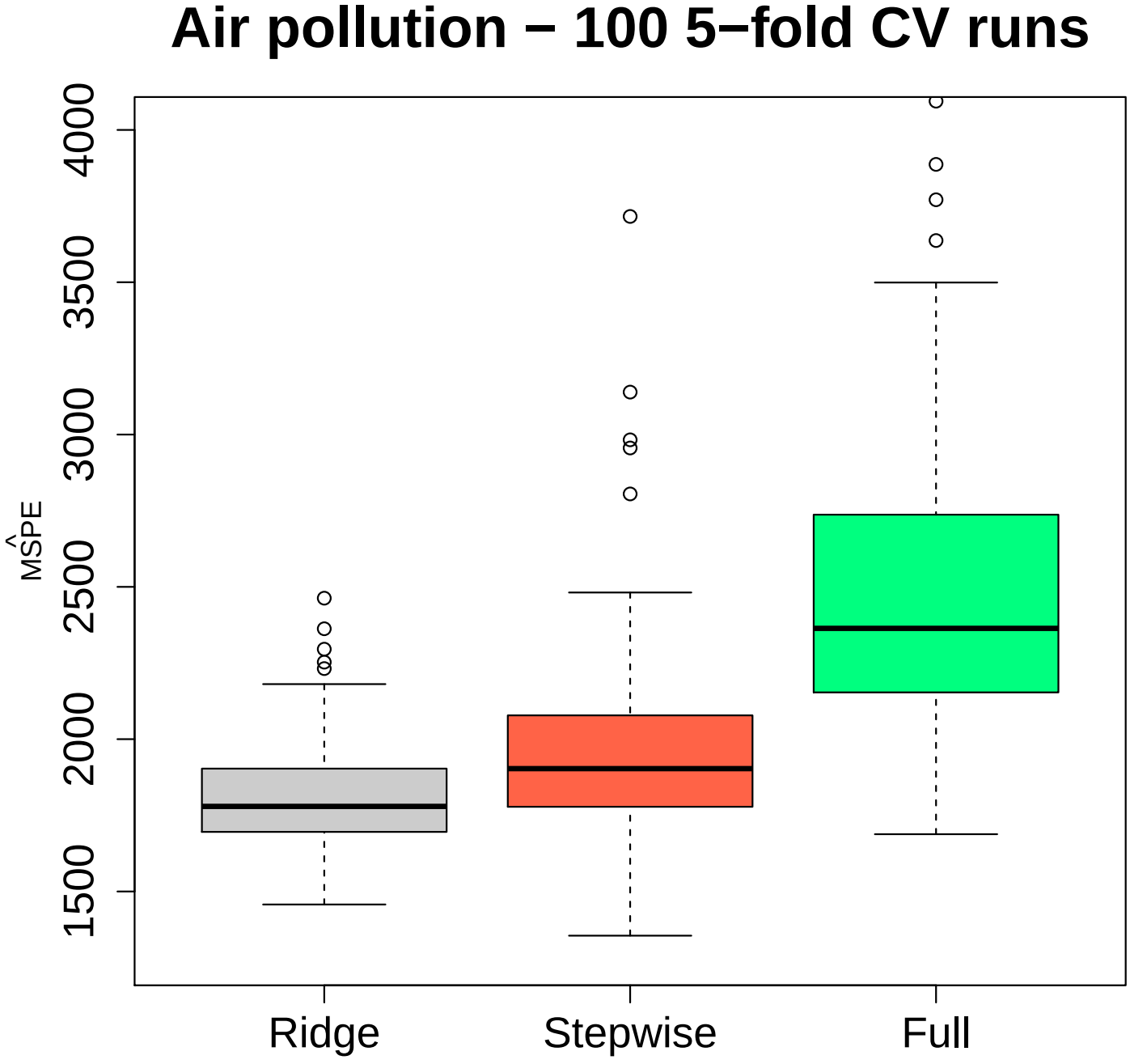
Cross validation selects

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

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

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
$\hat{MSPE}$
1500
2000
2500
3000
3500
4000
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.