STAT406- Methods of Statistical Learning Lecture 4

Matias Salibian-Barrera

UBC - Sep / Dec 2018

Proper use of CV

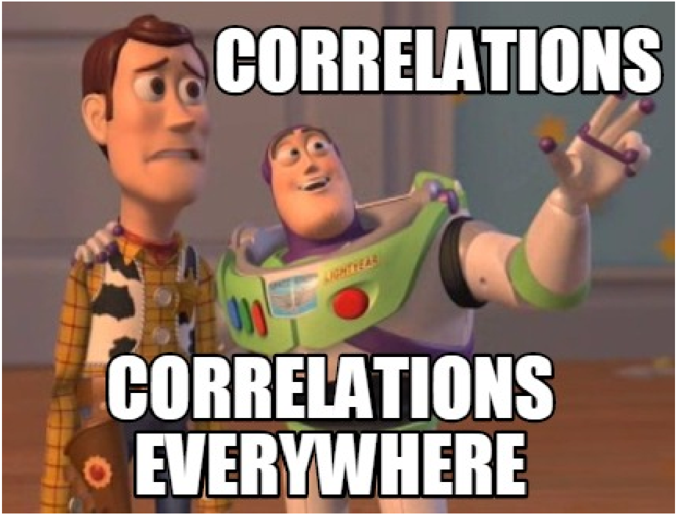
- An example of the importance and relevance of what we discussed in our last class:

  Ambroise, C. and McLachlan, G.J. Selection bias in gene extraction on the basis of microarray gene-expression data, PNAS, 2002, 99 (10), 6562-6566.

  `https://doi.org/10.1073/pnas.102102699`

Discussion points

- Why? Why would anybody want to **not** use all available features?
- “Somewhat obvious”: model parsimony, identify features that are relevant for the process under study.
- “Not so obvious?”: does prediction suffer if we use fewer variables? how much variability is induced by the feature selection step?


CORRELATIONS
CORRELATIONS EVERYWHERE

Model / feature selection

- Simple example:

```
set.seed(123)
x1 <- rnorm(506)
x2 <- rnorm(506, mean=2, sd=1)
x3 <- rexp(506, rate=1)
x4 <- x2 + rnorm(506, sd=.1)
x5 <- x1 + rnorm(506, sd=.1)
x6 <- x1 - x2 + rnorm(506, sd=.1)
x7 <- x1 + x3 + rnorm(506, sd=.1)
y <- x1*3 + x2/3 + rnorm(506, sd=2.2)
```

- Variables X_1 and X_2 are clearly important. But they are also highly correlated to X_4, X_5, X_6 and X_7.

Model / feature selection

- However, nothing is significant?

```
> summary(lm(y~., data=x))

Call:
lm(formula = y ~ ., data = x)
Residuals:
   Min     1Q Median     3Q    Max
-6.882 -1.474 -0.033  1.415  5.823

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.03457    0.23018   0.150   0.8807
x1           3.22612    1.68088   1.919   0.0555 .
x2           0.23867    1.39355   0.171   0.8641
x3          -0.35926    0.98680  -0.364   0.7160
x4          -0.69359    0.99025  -0.700   0.4840
x5           0.09271    0.91162   0.102   0.9190
x6          -0.73887    1.01114  -0.731   0.4653
x7           0.31651    0.98610   0.321   0.7484
---
Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1

Residual standard error: 2.148 on 498 degrees of freedom
Multiple R-squared: 0.6353,Adjusted R-squared: 0.6302
F-statistic: 123.9 on 7 and 498 DF,  p-value: < 2.2e-16
```

Model / feature selection

- But...

```
> summary(lm(y~x1+x2, data=x))

Call:
lm(formula = y ~ x1 + x2, data = x)

Residuals:
    Min      1Q  Median      3Q     Max
-6.9303 -1.5736 -0.0068  1.3840  5.9567

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.00733    0.20900   0.035  0.97204
x1           2.89168    0.09806  29.490  < 2e-16 ***
x2           0.27903    0.09249   3.017  0.00268 **
---
Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1

Residual standard error: 2.141 on 503 degrees of freedom
Multiple R-squared: 0.6343,Adjusted R-squared: 0.6328
F-statistic: 436.2 on 2 and 503 DF,  p-value: < 2.2e-16
```

Model / feature selection

- Even worse...

```
> summary(lm(y~x1+x2+x4, data=x))

Call:
lm(formula = y ~ x1 + x2 + x4, data = x)

Residuals:
    Min      1Q  Median      3Q     Max
-6.8064 -1.5229 -0.0308  1.4226  5.8861

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.0001127  0.2093588   0.001    1.000
x1           2.8964461  0.0983390  29.454   <2e-16 ***
x2           0.9740807  0.9917783   0.982    0.326
x4          -0.6934442  0.9851714  -0.704    0.482
---
Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1

Residual standard error: 2.142 on 502 degrees of freedom
Multiple R-squared: 0.6347,Adjusted R-squared: 0.6325
F-statistic: 290.7 on 3 and 502 DF,  p-value: < 2.2e-16
```

Model / feature selection

- If we use AIC

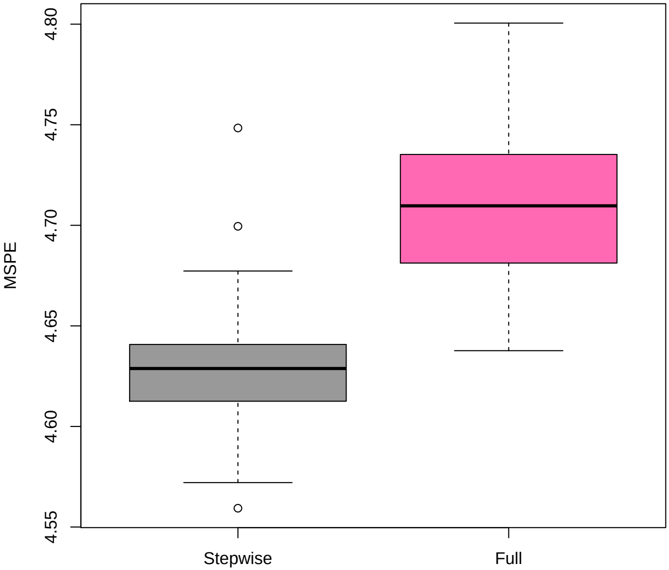
```
> st <- stepAIC(null,
     scope=list(lower=null, upper=full))
> st

Call:
lm(formula = y ~ x1 + x6, data = x)

Coefficients:
(Intercept)           x1           x6
  -0.000706     3.175239    -0.282906
```

Discussion points

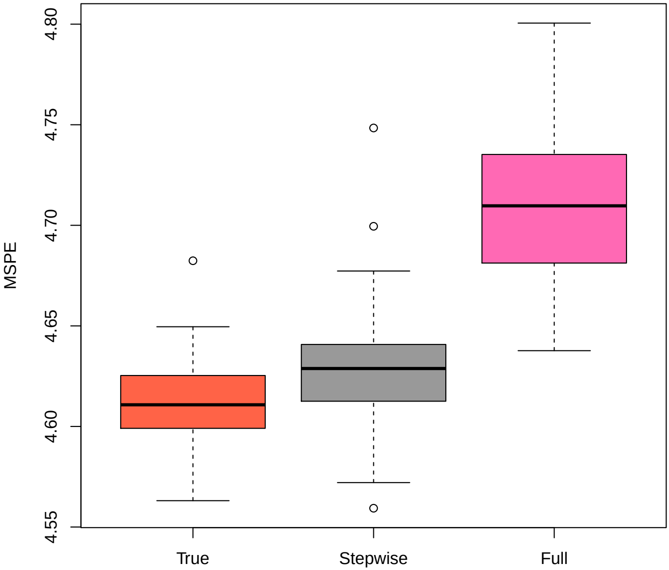
- Modeling problem (important variables may be missed)

- Prediction? Does stepwise give “the best” predicting model?

MSPEs



MSPEs



Discussion points

- Correlated covariates have become prevalent
- Researchers can (and do) collect data “blindly”
- Data are collected without a specific question in mind

Discussion points

- Correlated covariates:
- Mask each other when included simultaneously in a model

- May reduce prediction accuracy

Model / feature selection

One strategy:

- (1): Select models to be considered
- (2): Select a quantitative criterion to compare them (e.g. AIC, C_p, CV-based $\widehat{\text{MSPE}}$)
- (3): Choose a strategy to explore the models under consideration

Model / feature selection

For example:

- (1): Consider all possible models
- (2): Use AIC to compare them
- (3): Best subset search (2^p fits!)
- (3'): Stepwise search

- Is this strategy prediction-based?

AIC?

Why not compare models using residual sum of squares, or R^2?

LS vs MLE

Note that, if we assume that the error distribution is Gaussian, then a least squares fit for a linear regression model is the same as the MLE fit

... or is it?

Comparing models

- Comparing likelihoods / residuals isn't very useful
- More complex models have higher likelihoods (smaller residuals)
- The Akaike Information Criterion provides a way to compare models with different number of parameters
- There are many different ways to motivate it

Comparing models

- We can measure the “distance” between the true distribution of the data ($f_0(y)$) and our model $f(y, \theta)$

$$d\left(\theta, f_0\right) \, = \, E_0\left[-2\,\ell\left(y, \theta\right)\right] \, =$$

$$\int -2\,\ell\left(y, \theta\right)\, f_0(y)\, dy \, =$$

$$2\,\left[\mathcal{K}\left(\theta, f_0\right) \, - \, \int \log\left(f_0(y)\right) f_0(y)\, dy\,\right]$$

Comparing models

- Given our estimator $\hat{\theta}_n$ we could use

$$d(\hat{\theta}_n \, , f_0) \; = \; E_0 \left[-2 \, \ell(y, \theta) \right]_{\theta = \hat{\theta}_n}$$

  to see “how far” our model-based estimator is from the true distribution

- However, we can’t compute $d(\hat{\theta}_n, f_0)$
- Can we use $-2 \, \ell(y, \hat{\theta}_n)$ to estimate $d(\hat{\theta}_n, f_0)$?

Comparing models

- Yes, but this estimator is biased

$$E_0 \left[-2\,\ell(y, \hat{\theta}_n) \right] = E_0 \left\{ E_0 \left[-2\,\ell(y, \theta) \right]_{\theta = \hat{\theta}_n} \right\} - 2\,p + o(1)$$

Comparing models

- In other words

$$E_0\,[\mathrm{AIC}] \;\approx\; E_0\left[d(\hat{\theta}_n\,, f_0)\right]$$

where

$$\mathrm{AIC} \;=\; -2\ell(y, \hat{\theta}_n) + 2\,p$$

Comparing models

- For Gaussian errors we have

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

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

- Air pollution example in ℝ
- Synthetic data example

Sometimes...

- Selecting variables is not always necessary in terms of prediction accuracy.
- Doing so may in fact yield worse results.
- One such an example is discussed on Github.
- Read it carefully.

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

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

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

- 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 $C > 0$

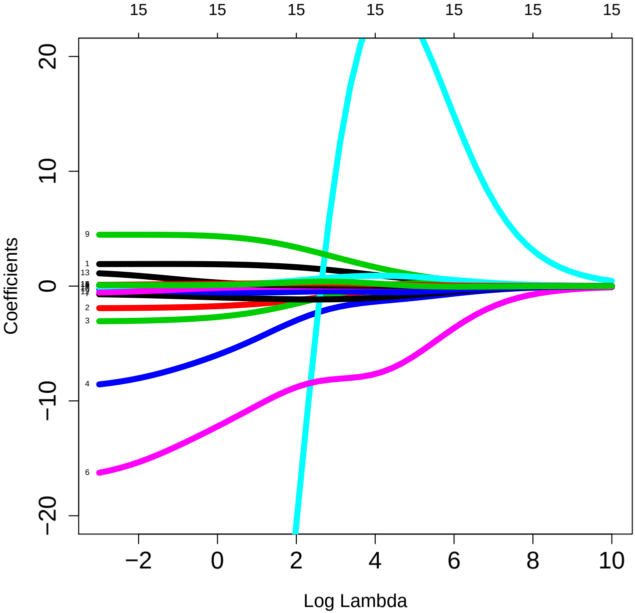
Bias / variance trade-off

- 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 $\hat{y}_1$, $\hat{y}_2$, $\ldots$, $\hat{y}_n$,

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

Effective degrees of freedom

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

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

with

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

and

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

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?

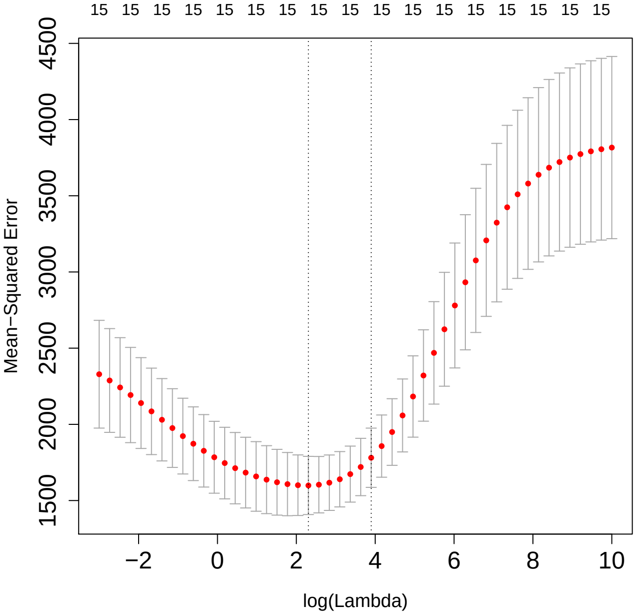
How can we select λ?

How do we select λ?

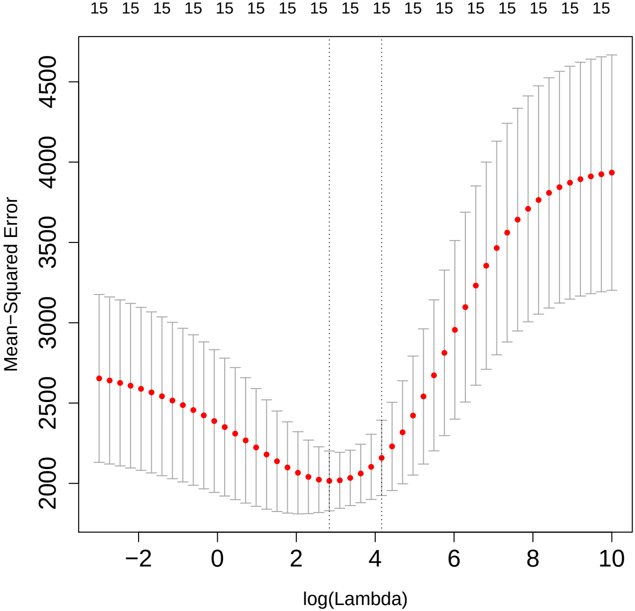
CV!

```
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

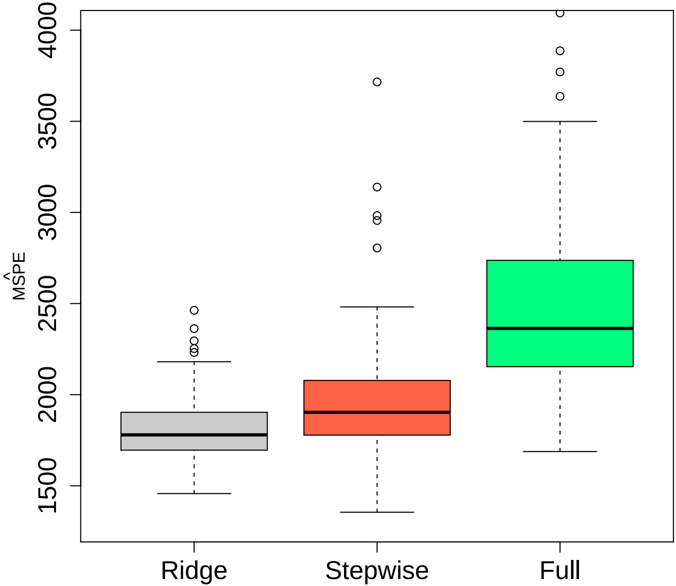
Cross validation selects

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

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

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