NTNU
Norwegian University of
Science and Technology

TMA4267 Linear Statistical Models V2014 (23)

Model selection [ISLR6.1]
Shrinkage [ISLR6.2]

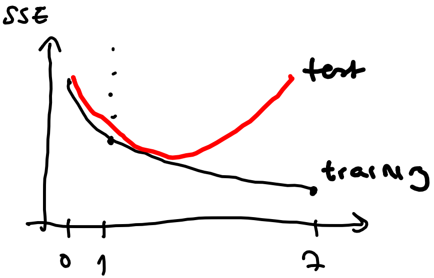
Mette Langaas

To be lectured: March 24, 2014
wiki.math.ntnu.no/emner/tma4267/2014v/start/

(Lecture 23) PART 7: Model selection [ISLR 6.1]
shrinkage [ISLR 6.2]
and dimension reduction [ISLR 6.3 + 10.2]

Last time: Model selection
- best subsets $(2^p + 1)$
- stepwise
  - forward
  - backward

$\left(\frac{p(p+1)}{2} + 1 \right)$



Mallows C_p (=AIC), R^2_{adj}, BIC } mimic the use of a test dataset

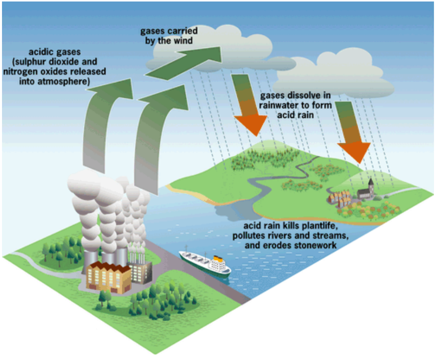
And rain:

All subsets: Mallows C_p: model with 5 predictors is the best: X_1, X_2, X_3, X_5, X_7

R^2_{adj}: 3 predictors preferred
$X1, X2, X3$

Acid rain

occurs when emissions of sulfur dioxide (SO2) and oxides of nitrogen (NOx) react in the atmosphere with water, oxygen, and oxidants to form various acidic compounds. These compounds then fall to the earth in either dry form (such as gas and particles) or wet form (such as rain, snow, and fog).



Source: http://myecoproject.org/get-involved/pollution/acid-rain/

Acid rain in Norwegian lakes

Measured pH in Norwegian lakes explained by content of

— x1: SO_4: sulfate (the salt of sulfuric acid),

— x2: $N0_3$: nitrate (the conjugate base of nitric acid),

— x3: Ca: calsium,

— x4: latent Al: aluminium,

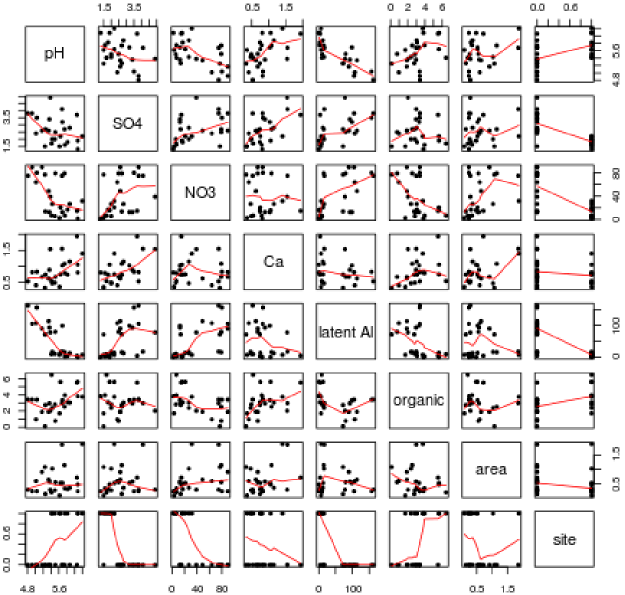
— x5: organic substance,

— x6: area of lake,

— x7: position of lake (Telemark or Trøndelag),

pH is a measure of the acidity of alkalinity of water, expressed in terms of its concentration of hydrogen ions. The pH scale ranges from 0 to 14. A pH of 7 is considered to be neutral. Substances with pH of less that 7 are acidic; substances with pH greater than 7 are basic.

Acid rain data



Acid rain (1). Best subset

```
regfit.full=regsubsets(y~.,data=ds)
sumreg <- summary(regfit.full)
Subset selection object
Call: regsubsets.formula(y ~ ., data = ds)
1 subsets of each size up to 7
Selection Algorithm: exhaustive
         x1  x2  x3  x4  x5  x6  x7
1  ( 1 ) " " " " " " "*" " " " " " "
2  ( 1 ) "*" " " "*" " " " " " " " "
3  ( 1 ) "*" "*" "*" " " " " " " " "
4  ( 1 ) "*" "*" "*" " " "*" " " " "
5  ( 1 ) "*" "*" "*" " " "*" " " "*"
6  ( 1 ) "*" "*" "*" "*" "*" " " "*"
7  ( 1 ) "*" "*" "*" "*" "*" "*" "*"
```

Acid rain (2)

```
# to mimic test set: R2adj and Cp
plot(1:7, sumreg$adjr2,type="l")
which.max(sumreg$adjr2) #5
plot(1:7, sumreg$cp,type="l")
which.min(sumreg$cp) #3
# so, model 3 or 5 is suggested for us
# model 3: x1+x2+x3
# model 5: x1+x2+x3+x5+x7

which.min(sumreg$bic) #3
```

Acid rain (3): Forward

```
# stepwise
regfitF=regsubsets(y~.,data=ds,method="forward")
sumregF <- summary(regfitF)
Selection Algorithm: forward
         x1  x2  x3  x4  x5  x6  x7
1  ( 1 ) " " " " " " "*" " " " " " "
2  ( 1 ) " " " " "*" "*" " " " " " "
3  ( 1 ) "*" "*" "*" " " " " " " " "
4  ( 1 ) "*" "*" "*" "*" " " " " " "
5  ( 1 ) "*" "*" "*" "*" "*" " " " "
6  ( 1 ) "*" "*" "*" "*" "*" " " "*"
7  ( 1 ) "*" "*" "*" "*" "*" "*" "*"
which.max(sumregF$adjr2)#5
which.min(sumregF$cp) #3
```

Acid rain (4): Backward

```
regfitB=regsubsets(y~.,data=ds,method="backward")
sumregB <- summary(regfitB)
Selection Algorithm: backward
         x1  x2  x3  x4  x5  x6  x7
1  ( 1 ) " " " " "*" " " " " " " " "
2  ( 1 ) "*" " " "*" " " " " " " " "
3  ( 1 ) "*" "*" "*" " " " " " " " "
4  ( 1 ) "*" "*" "*" " " "*" " " " "
5  ( 1 ) "*" "*" "*" " " "*" " " "*"
6  ( 1 ) "*" "*" "*" "*" "*" " " "*"
7  ( 1 ) "*" "*" "*" "*" "*" "*" "*"
which.max(sumregB$adjr)#5
# backward finds same as best subset
which.min(sumregB$cp) #3
```

Shrinkage methods [ISLR 6.2]

Model selection: use least squares (LS), but only fit a subset of the predictors.

Now: add penalty to LS criterion

pay a price for large coefficient.

Ridge regression (from 1970s) [RR]

$$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_{i1} - \cdots \beta_p x_{ip})^2$$

In LS estimation we minimize SSE.

$\hat{\beta}^R$ (ridge) is the result from a penalized minimization of

$$SSE^R(\lambda) = SSE + \lambda \cdot \sum_{j=1}^{p} \beta_j^2$$

where $\lambda \geq 0$ is a tuning parameter - to be determined separately.

$$SSE^R(\lambda) = (Y - X\beta)^T (Y - X\beta) + \lambda \beta^T \beta$$

(homework, see ch3)

$$\hat{\beta}^R = (X^T X + \lambda I)^{-1} X^T Y$$

When $\lambda = 0$ we get the ordinary LS solution.

$\lambda \to \infty$ all $\hat{\beta}^R = 0$

$$\frac{\|\hat{\beta}^R_\lambda\|_2}{\|\hat{\beta}\|_2} = \frac{\sqrt{\sum_{j=1}^{p} \hat{\beta}_j^{R\,2}}}{\sqrt{\sum_{j=1}^{p} \hat{\beta}_j^{2}}}$$

ridge

LS

1 when $\lambda = 0$

0 when $\lambda \to \infty$

used in plots, on x-axis – instead of λ.

Remark: we don't want to shrink the intercept, so that is not included in the penalty

Credit data

```
> names(credit)
 [1] "Income"    "Limit"     "Rating"    "Cards"     "Age"
 [7] "Gender"    "Student"   "Married"   "Ethnicity" "Balan
> dim(credit)
[1] 400  11
```

- Response: Balance, amount due at the end of the month (some have 0).
- Income
- Limit (credit limit)
- Rating (credit rating)
- Cards: number of credit cards (1-9).
- Education: number of years (5-20).
- Gender
- Student or not.
- Married or not.
- Ethicity three levels (African American, Asian, Caucasian), coded as factor.

Credit data: Ridge regression

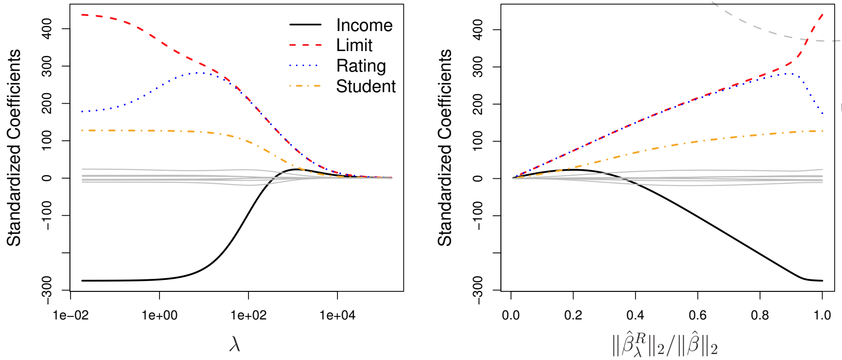


Figure 6.4 from An Introduction to Statistical Learning (2013)

Why use ridge instead of LS?

When λ increase the flexibility of the RR fit will decrease, which will result in increased bias but decreased variance (remember: bias-variance trade-off).

Simulation study (A) to show this:

$p = 45$ & $\beta_1, \beta_2, \ldots, \beta_{45}$ all $\neq 0$
$n = 50$

training set $n=50 \to$ fit $\hat{\beta}^R$ | test data: MSE | plot
report black: bias^2 | ↑ pink
green: variance

RR works best in situations where LS estimators have high variance (and many predictors are truly non-zero).

Computationally fast!

Simulated data (A): Ridge regression

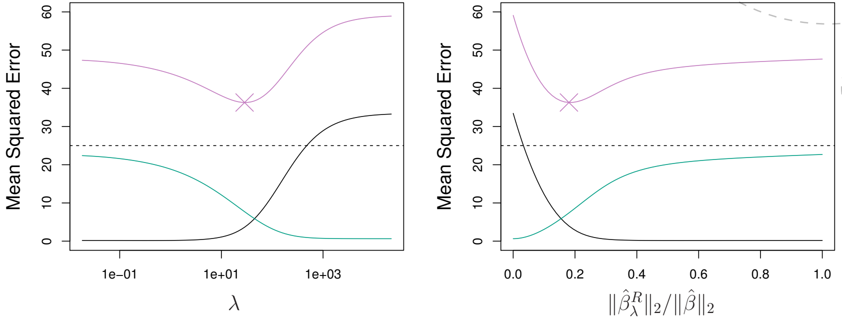


Figure 6.5 from An Introduction to Statistical Learning (2013)

Scale issue

Q: Let X_1 be a measurement in NOK.

Fitting $Y = \beta_0 + \overset{\frac{1}{1000}}{\beta_1} \overset{\cdot 1000}{X_1} + \varepsilon$ using LS give you $\hat{\beta}_1$.

Instead use X_1 to be KNOK (1000's of NOKs)

Before $X_1 = 1000$, now $X_1 = 1$.

What happens with the LS solution for the new coding?

Solution: no problem $\hat{\beta}_1$ is just scaled accordingly

In RR the scale matters, and $\hat{\beta}_1^R$ can change substantially if X_1 is changed. This is because of the $\sum \beta_j^2$-penalty. Solution: work with standardized predictors:

$$\tilde{x}_{ij} = \frac{x_{ij}}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)^2}}$$

So all predictors are on the same scale.

Question

You perform ridge regression on a problem where your third predictor, x3, is measured in dollars. You decide to refit the model after changing x3 to be measured in cents. Which of the following is true?

A $\hat{\beta}_3$ and $\hat{y}$ will remain the same.

B $\hat{\beta}_3$ will change, but $\hat{y}$ will remain the same.

C $\hat{\beta}_3$ will remain the same, but $\hat{y}$ will change.

D $\hat{\beta}_3$ and $\hat{y}$ will both change.

Vote at clicker.math.ntnu.no, classroom TMA4267.

Lasso regression

Problem with ridge: all p predictors are included in the fitted model. The $\hat{\beta}^R$ are shrunk, but not to zero unless $\lambda = \infty$. This makes interpretation hard.

Lasso:

$$SSE^L(\lambda) = SSE + \lambda \cdot \underbrace{\sum_{j=1}^{p} |\beta_j|}_{\|\beta\|_1 \ (L_1 \text{ norm})}$$

There is no closed form solution for $\hat{\beta}^L$ here, but the L_1-penalty will for large enough λ <u>force</u> some $\hat{\beta}_j$'s to be exactly equal to zero.

$\Rightarrow$ Lasso gives a sparse model, and does variable selection.

Credit data $\rightarrow$ lasso estimates

Credit data: Lasso regression

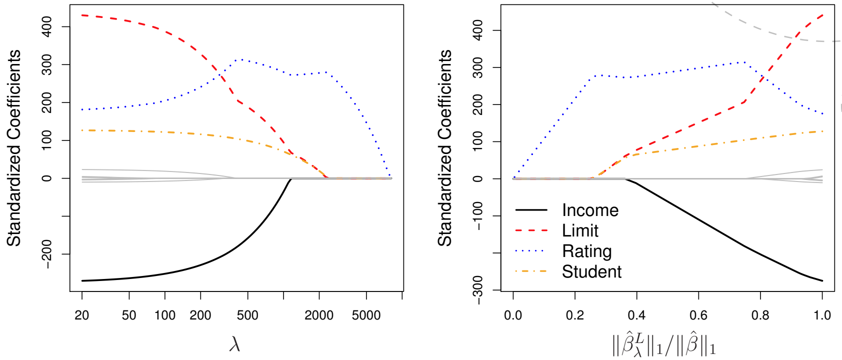


Figure 6.6 from An Introduction to Statistical Learning (2013)

Comparing ridge end lasso $\Rightarrow$ perform different for $0 < \lambda < \infty$.

Simulated data (A) : $n=50$, $p=45$ [Figure 6.8]

lasso = solid
dotted = ridge
} ridge is the best

data perfect for lasso

Simulated data (B) : $n=50$, $p=2$ end 43 noise covariates

$\Rightarrow$ lasso is the best

[Fig 6.9]

Neighter rigde or lasso dominates the other in all situation.

Simulated data (A): Ridge and lasso

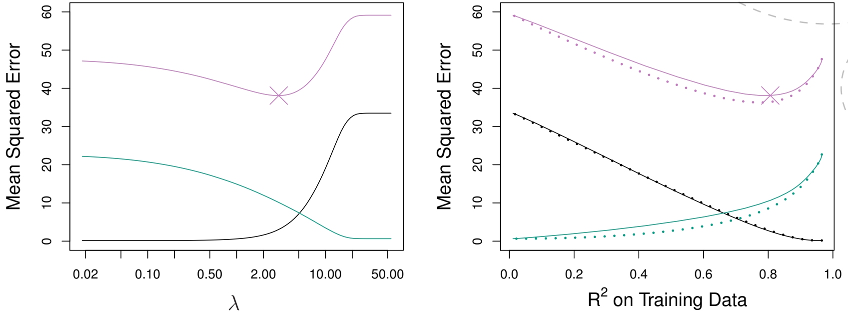


Figure 6.8 from An Introduction to Statistical Learning (2013)

Alternative formulation and graphics

RR:

$$\underset{\beta}{\text{minimize}} \{ SSE \} \quad \text{subject to} \quad \sum \beta_j^2 \leq s$$

Lasso:

$$\underset{\beta}{\text{minimize}} \{ SSE \} \quad \text{subject to} \quad \sum |\beta_j| \leq s$$

· where there is a one-to-one between λ end s.

penelty constraint

$P=2$: graphical display

Simulated data (B): Ridge and lasso

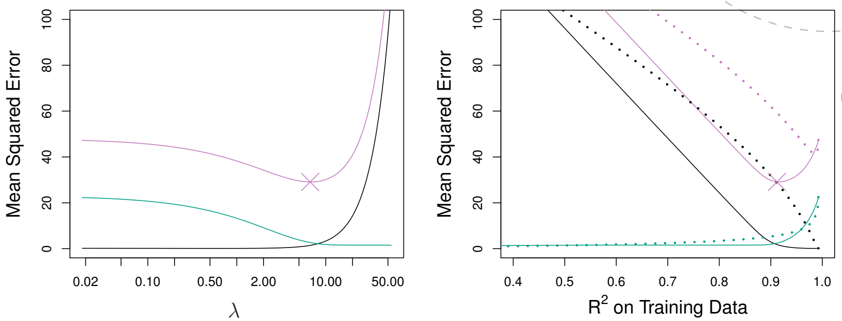


Figure 6.9 from An Introduction to Statistical Learning (2013)

Graphically: Lasso (left) and ridge (right)

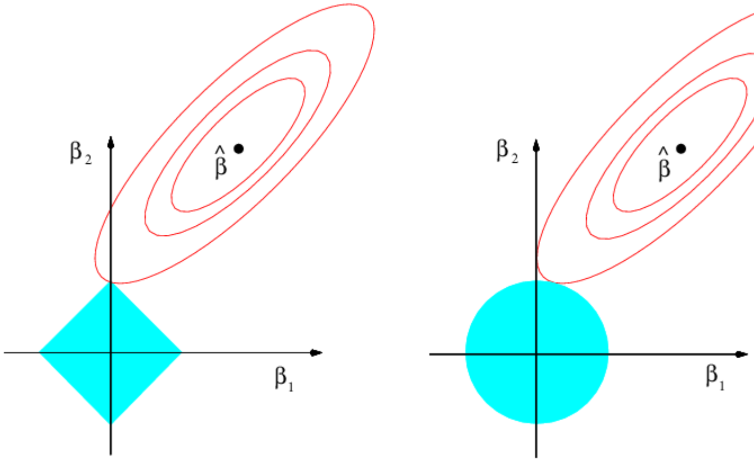


Figure 6.7 from An Introduction to Statistical Learning (2013)