Penalised Likelihood

Alex Lewin

Department of Medical Statistics,
London School of Hygiene and Tropical Medicine

Spring 2022

Multiple Linear Regression

Standard notation in supervised learning:
n = number of observations
p = number of variables

Data: $\boldsymbol{y}$ vector of response variables (length n)

X matrix of predictor variables (dimension $n \times p$)

Write the model in matrix format:

$$\boldsymbol{y} = \boldsymbol{X}\beta + \boldsymbol{\epsilon}$$

Maximum Likelihood solution satisfies

$$(\boldsymbol{X}^T\boldsymbol{X})\beta = \boldsymbol{X}^T\boldsymbol{y}$$

(Ordinary Least Squares):

$$\hat{\beta}^{OLS} = (\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{y}$$

$p > n \implies \boldsymbol{X}^T \boldsymbol{X}$ is not full rank (so not invertible), so the OLS solution does not exist.

These are p equations to find p parameters (β):

$$\underset{p \times p}{(\boldsymbol{X}^T \boldsymbol{X})} \underset{p \times 1}{\beta} = \underset{p \times 1}{\boldsymbol{X}^T \boldsymbol{y}}$$

$\boldsymbol{X}^T \boldsymbol{X}$ has rank n (ie n independent equations). So n degrees of freedom to find p parameters (β).

So there is not a unique solution to the Maximum Likelihood equation.

Possible solutions to the problem:

- Subset selection (not backwards, as size of subset limited to $< n$)
- Bayesian approach, as yesterday and briefly later today
- Regularisation/shrinkage/penalisation approach (can be seen as Bayesian)

Ridge Penalisation

OLS solution minimises the cost function:

$$RSS = ||\boldsymbol{y} - X\boldsymbol{\beta}||^2$$

Ridge regression minimises a different cost function:

$$RSS + Penalty \quad = \quad ||\boldsymbol{y} - X\boldsymbol{\beta}||^2 + \lambda\boldsymbol{\beta}^2$$

Ridge solution:

$$\hat{\boldsymbol{\beta}}^{ridge} = (\boldsymbol{X}^T\boldsymbol{X} + \lambda\boldsymbol{I}_p)^{-1}\boldsymbol{X}^T\boldsymbol{y}$$

where $\boldsymbol{I}_p$ is the $p \times p$ identity matrix.

Ridge regression:

$$RSS + Penalty \quad = \quad ||\boldsymbol{y} - X\boldsymbol{\beta}||^2 + \lambda\boldsymbol{\beta}^2$$

- λ is a shrinkage parameter (to be set, see later)
- larger $\lambda \implies$ penalty term more important

Prostate Cancer - Small Example

Small data set of 67 men with prostate cancer, used as an example in Hastie, Tibshirani and Friedman 2001.

Outcome variable is log of prostate specific antigen (PSA), several predictor variables, given in table below.

Covariates
log cancer volume
log prostate weight
age
log benign prostatic hyperplasia (BPH)
seminal vesicle invasion (SVI)
log capsular penetration
Gleason score
% Gleason score above 4

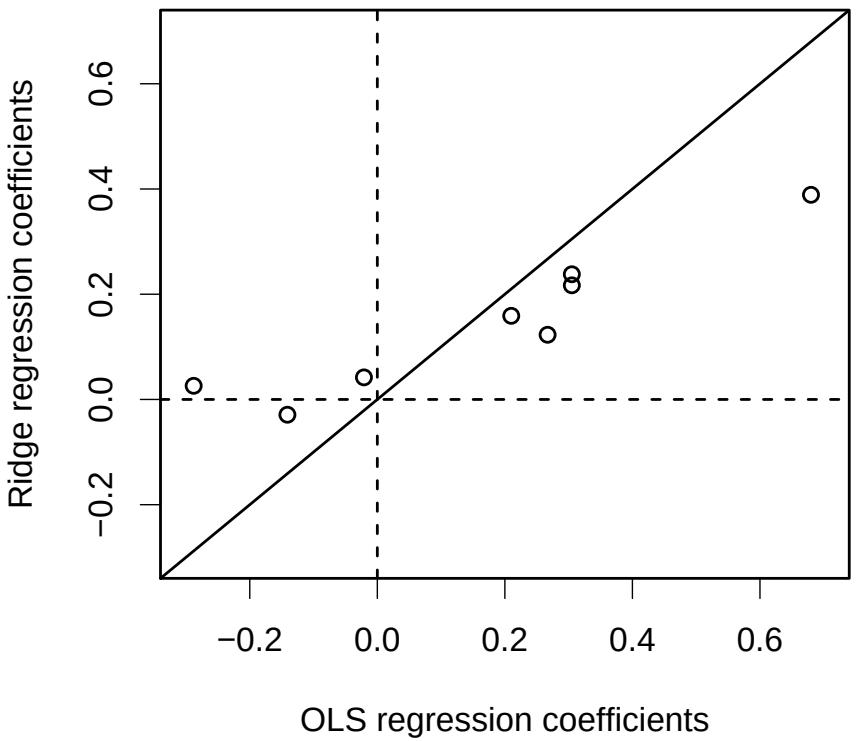
Aim to find out which covariates are associated with PSA - which may be causal.

Prediction of less importance here.

Ridge v. OLS

Example: Prostate Cancer data

- Response variable Y = log of PSA (prostate specific antigen)
- Covariates X: 8 predictors (age, cancer volume etc.)



- Ridge coefficients are shrunk towards zero
- i.e. $|\hat{\beta}_j^{Ridge}| < |\hat{\beta}_j^{OLS}|$ for most predictors

Lasso Regression

Ridge regression:
Find parameters estimates $\hat{\beta}_0, \cdots, \hat{\beta}_p$ which minimise

$$RSS + Penalty = \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{p} \beta_j^2$$

Lasso, minimise:

$$RSS + Penalty = \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j|$$

Leads to quite different results!

Lasso penalty $\lambda \sum_{j=1}^{p} |\beta_j|$ leads to some $\hat{\beta}_j = 0$.

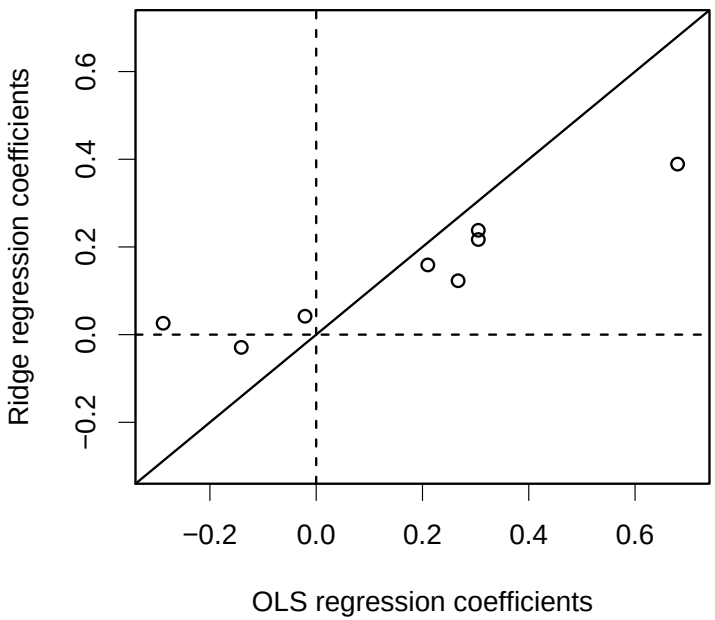
Automatic variable selection
(only a subset of the predictors appear in the model).

Lasso v. Ridge v. OLS

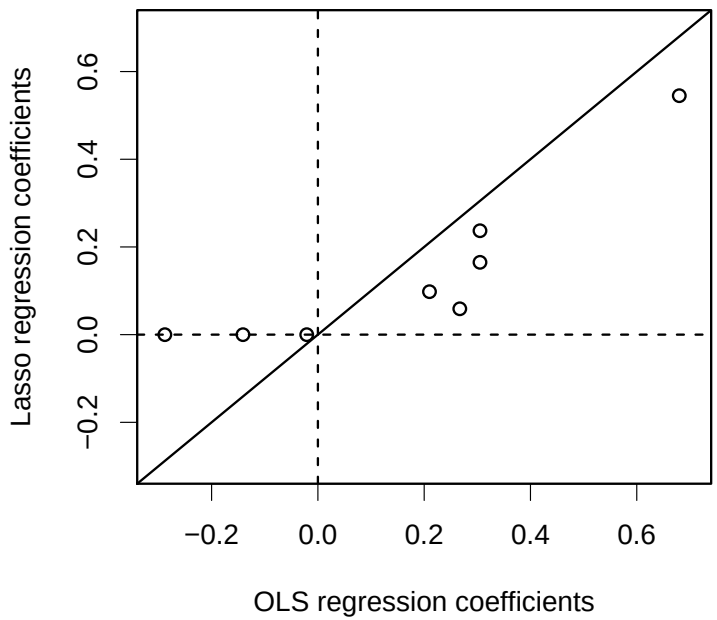
Example: Prostate Cancer data

- Response variable Y = log of PSA (prostate specific antigen)
- Covariates X: 8 predictors (age, cancer volume etc.)

Ridge regression



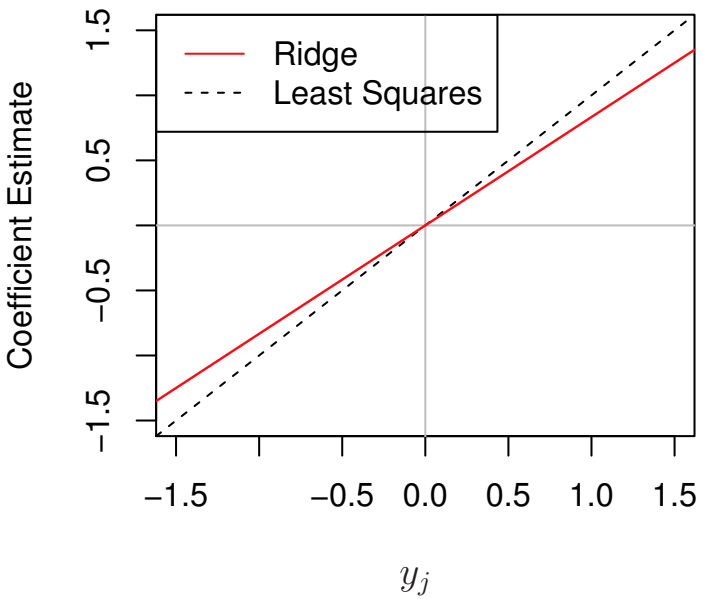
Lasso regression

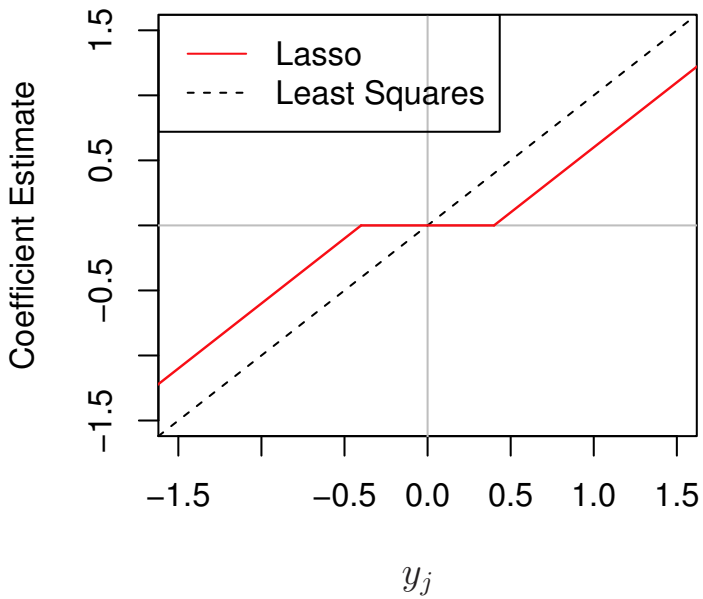


Lasso selects 5 variables out of 8 to put in the model.

Shrinking/thresholding

Simple example with Xs uncorrelated.





Ridge shrinks all coefficients **proportionally** towards zero (proportion depends on λ).

Lasso performs "soft-thresholding":
- coefficients are **shifted** towards zero (amount depends on λ),
- below certain threshold coefs are set to zero (threshold depends on λ).

Changing the tuning parameter λ

Ridge regression minimises the quantity

$$RSS + \lambda \sum_{j=1}^{p} \beta_j^2$$

- minimise RSS $\rightarrow$ fitting line to the data
- minimise $\sum_j \beta_j^2 \rightarrow$ make βs smaller

λ balances between these two objectives.

- λ smaller $\rightarrow$ better fit to data
- λ bigger $\rightarrow$ regression coefficients shrunk

So choice of λ crucial!

Choosing Lambda in Ridge or Lasso Regression

Validation Set method for choosing the best lambda:

1. Split data into Training and Test data sets (just once)
2. Train models for several different values of λ
3. Choose λ where Test MSE is minimised

Cross-Validation approach:

1. Split data into Training and Test data sets (lots of times)
2. Train models for several different values of λ (on each different Training set)
3. Calculate MSE averaged over the different Test data sets
4. Choose λ where MSE is minimised

Summary of fitting procedure

- Minimise penalised likelihood for grid of λ values $\longrightarrow$ estimated regression coefficients $\beta \longrightarrow$ predicted responses y
- Find λ which minimises the test prediction error (MSE for linear model)
- Point estimates of regression coefficients are those found for that λ value

Some notes:

- Usual to standardise variables (since penalty term makes regression coefficients 'compete')
- Value of λ is not important - only controls relative importance of likelihood and penalty terms
- May get very different values of λ for ridge and lasso on the same data set