Regularized estimation of LM and GLM

Part 1: Ridge regression

Pedro Delicado
Departament d'Estadística i Investigació Operativa
Universitat Politècnica de Catalunya

References:

Section 3.4 in Hastie, Tibshirani, and Friedman (2009)

Introduction

- In the multiple linear regression model (with n observations and p predictors, p possibly greater than n) we consider the penalized least squares coefficients estimator where the penalization is given by the L_1 or the L_2 norms of the estimator.
- This procedures leads to ridge regression (L_2 penalization) and to lasso (least absolute shrinkage and selection operator) estimation (L_1 penalization).
- In the pathway, we will learn:
  - Linear estimators of a regression function.
  - Effective number of parameters (or effective degrees of freedom) of a regression estimator.
  - Tuning parameters choice based on leave-one-out cross-validation, k-fold cross-validation or generalized cross-validation.
  - Efficient computation of leave-one-out cross-validation for linear estimators.

Multiple linear regression model

- Consider that n pairs $(\boldsymbol{x}_i, y_i)$, $i = 1, \ldots, n$ of data, $y_i \in \mathbb{R}$ and $\boldsymbol{x}_i \in \mathbb{R}^p$, are observed from the multiple linear regression model

$$y_i = \beta_0 + \sum_{j=1}^{p} x_{ij}\beta_j + \varepsilon_i,$$

where $\varepsilon_1, \ldots, \varepsilon_n$ are i.i.d. r.v. with zero mean and variance σ^2, and $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_p)^\mathsf{T} \in \mathbb{R}^{p+1}$ is a vector of unknown coefficients.

- Fitting the model consists in providing estimators for $\boldsymbol{\beta}$ and σ^2, jointly with information about the sampling distribution of these estimators (standard errors, hypothesis testing, among others).

Ordinary Least Squares (OLS)

- Ordinary Least Squares (OLS) estimator:

$$\hat{\boldsymbol{\beta}}_{\text{OLS}} = \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2$$

- In matrix notation: $\hat{\boldsymbol{\beta}}_{\text{OLS}} = (\boldsymbol{X}^{\mathsf{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathsf{T}}\boldsymbol{y}$.
- $\hat{\boldsymbol{\beta}}_{\text{OLS}}$ is an unbiased estimator of $\boldsymbol{\beta}$.

1

- **Gauss–Markov Theorem.** For any $\boldsymbol{a} \in \mathbb{R}^{p+1}$, the OLS estimator of the linear combination $\boldsymbol{a}^\mathsf{T}\boldsymbol{\beta}$, namely $\boldsymbol{a}^\mathsf{T}\hat{\boldsymbol{\beta}}_{\text{OLS}}$, is unbiased and it has the lowest variance among the linear unbiased estimates of $\boldsymbol{a}^\mathsf{T}\boldsymbol{\beta}$.
- In particular, following the Bayes rule, the prediction for a new observation $\boldsymbol{x}$, is $\hat{y} = \boldsymbol{x}^\mathsf{T}\boldsymbol{\beta}$.
- So its best unbiased estimator is $\hat{y}_{\text{OLS}} = \boldsymbol{x}^\mathsf{T}\hat{\boldsymbol{\beta}}_{\text{OLS}}$.

[1] Slides with colored background are optional material.

Multicolinearity and bad conditioned matrices

- The computation of $\hat{\boldsymbol{\beta}}_{\text{OLS}}$ is numerically unstable when $\boldsymbol{X}^{\mathsf{T}}\boldsymbol{X}$ is close to be singular.
- Condition number of a symmetric matrix $\boldsymbol{A}$: $\kappa(\boldsymbol{A}) = \sqrt{\frac{\gamma_{\text{max}}}{\gamma_{\text{min}}}}$, where γ_{max} and γ_{min} are, respectively, the largest and lowest eigenvalue absolute values of $\boldsymbol{A}$.
- $\boldsymbol{A}$ is not invertible if and only if $\kappa(\boldsymbol{A}) = \infty$.
- A large value of $\kappa(\boldsymbol{A})$ (in practice, larger than 30), indicates that numerical problems may appear when inverting $\boldsymbol{A}$.
- In these cases we say that $\boldsymbol{A}$ is bad conditioned.
- If $\boldsymbol{X}^{\mathsf{T}}\boldsymbol{X}$ is bad conditioned: unstable computation of $\hat{\boldsymbol{\beta}}_{\text{OLS}}$.
- A large condition number indicates that $\boldsymbol{X}$ is close to be singular, that is, close that some columns of $\boldsymbol{X}$ can be written as linear combinations of the other.
- We talk about multicolinearity between columns of $\boldsymbol{X}$.

Regularized regression

- Beyond numerical problems, $\hat{\boldsymbol{\beta}}_{\text{OLS}}$ can not be computed when the rank of $\boldsymbol{X}$ is lower than the number of variables $(p+1)$ (this is an extreme case of multicolinearity).
- This is the case when $p+1 > n$ (or $p+1 \gg n$, as it can happen in applications with large scale data).
- In practical terms, what happens is that $\boldsymbol{y}$ can be written as a linear combination of the predictors using infinitely many coefficient vectors, for which the objective OLS objective function is equal to 0, its minimum. So there is no way to select *the best* among those coefficient vectors.
- Shrinkage (or regularized) methods: They add a penalty (depending on $\boldsymbol{\beta}$) to the objective function in such a way that the new optimum is attained at a unique vector $\hat{\boldsymbol{\beta}}$.
- The unbiasedness of OLS estimation is lost, but the new estimators may have lower Mean Square Error (and they are numerically stable).

Ridge regression

- The ridge coefficients minimize a penalized sum of squares residuals:

$$\begin{aligned}\hat{\boldsymbol{\beta}}_{\text{ridge}} &= \arg\min_{\boldsymbol{\beta}} \text{SSR}_{\text{pen}}(\boldsymbol{\beta}) \\ &= \arg\min_{\boldsymbol{\beta}} \left\{ \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \right\} \\ &= \arg\min_{\boldsymbol{\beta}} \left\{ (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^{\mathsf{T}} (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) + \lambda \|\boldsymbol{\beta}_{-0}\|_2^2 \right\}\end{aligned}$$

  where $\boldsymbol{\beta}_{-0} = (\beta_1, \ldots, \beta_p)^{\mathsf{T}}$.

- Here $\lambda \geq 0$ is a complexity parameter that controls the amount of shrinkage: the larger the value of λ, the greater the amount of shrinkage of $\boldsymbol{\beta}_{-0}$ toward zero.

- Ridge regression is a penalized least squares problem:

$$\hat{\boldsymbol{\beta}}_{\text{ridge}} = \arg\min_{\boldsymbol{\beta}} \left\{ \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \right\}$$

- Alternative expression, constrained least squares problem:

$$\hat{\boldsymbol{\beta}}_{\text{ridge}} = \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2$$

$$\text{subject to } \sum_{j=1}^{p} \beta_j^2 \leq t,$$

for $t \geq 0$. There is a one-to-one decreasing correspondence between parameters $\lambda \in [0, \infty)$ and $t \in (0, \|\hat{\boldsymbol{\beta}}_{-0,\text{OLS}}\|^2]$.

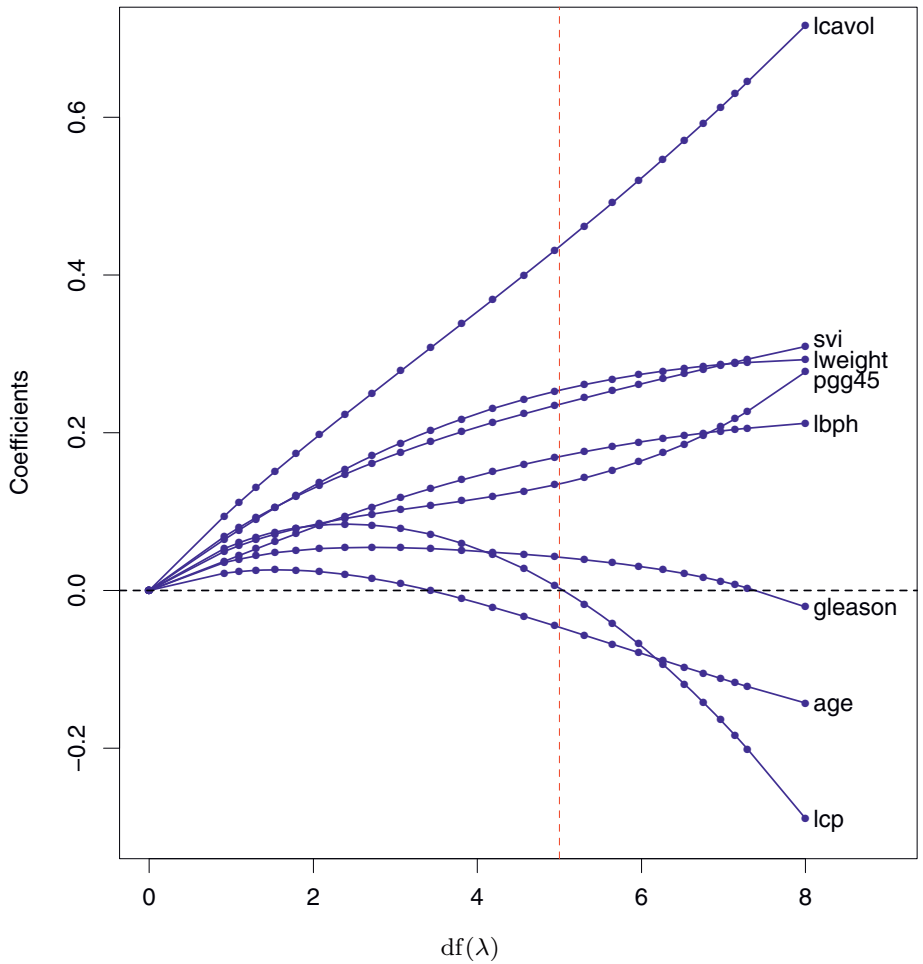
- Observe that changes in scale of the explanatory variables affect the constraint effects (or, equivalently, the effects of penalization term).
- For this reason, from now on we assume that the predictor variables have zero mean and unit variance (otherwise we center and standardize them in advance):

$$\sum_{i=1}^{n} x_{ij} = 0, \frac{1}{n}\sum_{i=1}^{n} x_{ij}^2 = 1, \quad j = 1, \ldots, p.$$

- Moreover, the response variable is assumed to have zero mean ($\sum_{i=1}^{n} y_i = 0$), that is, $\beta_0 = 0$.

Prostate cancer example. Ridge regression

- Goal: To examine the correlation between the level of log of prostate-specific antigen (lpsa) and a number of clinical measures in 97 men who were about to receive a radical prostatectomy.
- The predictor variables are
  - log cancer volume (lcavol),
  - log prostate weight (lweight),
  - age,
  - log of the amount of benign prostatic hyperplasia (lbph),
  - seminal vesicle invasion (svi),
  - log of capsular penetration (lcp),
  - Gleason score (gleason), and
  - percent of Gleason scores 4 or 5 (pgg45).



Source: Hastie, Tibshirani, and Friedman (2009)

Explicit solution for the ridge regression

- The ridge regression estimators are the solution of the penalized least squares problem

$$\min_{\boldsymbol{\beta}\in\mathbb{R}^p} \left\{ \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \right\}$$

that can be expressed as

$$\min_{\boldsymbol{\beta}\in\mathbb{R}^p} \Psi(\boldsymbol{\beta}) = (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^{\mathsf{T}}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) + \lambda\boldsymbol{\beta}^{\mathsf{T}}\boldsymbol{\beta},$$

that has an explicit solution, as we show now.

- Taking the gradient

$$\nabla\Psi(\boldsymbol{\beta}) = -2\boldsymbol{X}^{\mathsf{T}}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) + 2\lambda\boldsymbol{\beta},$$

and solving in $\boldsymbol{\beta}$ the equation $\nabla\Psi(\boldsymbol{\beta}) = \mathbf{0}$, we obtain

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

- Ridge regression estimator: $\hat{\boldsymbol{\beta}}_{\text{ridge}} = \left(\boldsymbol{X}^{\mathsf{T}}\boldsymbol{X} + \lambda \boldsymbol{I}_p\right)^{-1} \boldsymbol{X}^{\mathsf{T}}\boldsymbol{y}$.
- Therefore, for any $\boldsymbol{x} \in \mathbb{R}^p$, the corresponding predicted value is

$$\hat{y} = \boldsymbol{x}^{\mathsf{T}}\hat{\boldsymbol{\beta}}_{\text{ridge}} = \boldsymbol{x}^{\mathsf{T}}\left(\boldsymbol{X}^{\mathsf{T}}\boldsymbol{X} + \lambda \boldsymbol{I}_p\right)^{-1} \boldsymbol{X}^{\mathsf{T}}\boldsymbol{y}.$$

- The vector of fitted values is

$$\hat{\boldsymbol{y}} = \boldsymbol{X}\left(\boldsymbol{X}^{\mathsf{T}}\boldsymbol{X} + \lambda \boldsymbol{I}_p\right)^{-1} \boldsymbol{X}^{\mathsf{T}}\boldsymbol{y} = \boldsymbol{H}_\lambda \boldsymbol{y}.$$

- Compare with the OLS solution: $\hat{\boldsymbol{\beta}}_{\text{OLS}} = \left(\boldsymbol{X}^{\mathsf{T}}\boldsymbol{X}\right)^{-1} \boldsymbol{X}^{\mathsf{T}}\boldsymbol{y}$.

$$\hat{\boldsymbol{y}}_{\text{OLS}} = \boldsymbol{X}\left(\boldsymbol{X}^{\mathsf{T}}\boldsymbol{X}\right)^{-1} \boldsymbol{X}^{\mathsf{T}}\boldsymbol{y} = \boldsymbol{H}\boldsymbol{y},$$

  where $\boldsymbol{H} = \boldsymbol{X}\left(\boldsymbol{X}^{\mathsf{T}}\boldsymbol{X}\right)^{-1} \boldsymbol{X}^{\mathsf{T}}$ is called the hat matrix.
- $\lim_{\lambda \longrightarrow 0} \hat{\boldsymbol{\beta}}_{\text{ridge}} = \hat{\boldsymbol{\beta}}_{\text{OLS}}$, $\lim_{\lambda \longrightarrow \infty} \hat{\boldsymbol{\beta}}_{\text{ridge}} = \boldsymbol{0}$.

Practice:

- Prostate data: Ridge regression estimation and *coefficients path*.
- Use the R script `prostate.ridge.regression.R`.

Singular Value Decomposition of $\boldsymbol{X}$

- Let $\boldsymbol{X} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^{\mathsf{T}}$ be the Singular Value Decomposition of $\boldsymbol{X}$. That is:
  - $\boldsymbol{U}$, $n \times p$ orthonormal matrix whose columns span the $\boldsymbol{X}$ column space.
  - $\boldsymbol{D}$, $p \times p$ diagonal matrix with elements $d_1 \geq \ldots \geq d_p \geq 0$ in the diagonal, that are called singular values of $\boldsymbol{X}$.
  - $\boldsymbol{V}$, $p \times p$ orthonormal matrix whose columns span the row space of $\boldsymbol{X}$.
- Observe that $\boldsymbol{X}^{\mathsf{T}}\boldsymbol{X} = \boldsymbol{V}\boldsymbol{D}\boldsymbol{U}^{\mathsf{T}}\boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^{\mathsf{T}} = \boldsymbol{V}\boldsymbol{D}^2\boldsymbol{V}^{\mathsf{T}}$ and it follows that the eigenvalues of $\boldsymbol{X}^{\mathsf{T}}\boldsymbol{X}$ are the squared singular values of $\boldsymbol{X}$:

$$\gamma_j = d_j^2, \; j = 1, \ldots, p.$$

As we are assuming that the explanatory variables have zero mean, we have that $\boldsymbol{X}^{\mathsf{T}}\boldsymbol{X}$ is the sample covariance matrix. Then the columns of $\boldsymbol{V}$ are the principal components of $\boldsymbol{X}$. Moreover the columns of $\boldsymbol{UD}$ are the scores of the observed data in the principal components.

Numerical stability of ridge regression

- $\hat{\boldsymbol{\beta}}_{\text{ridge}} = \left(\boldsymbol{X}^{\mathsf{T}}\boldsymbol{X} + \lambda \boldsymbol{I}_p\right)^{-1} \boldsymbol{X}^{\mathsf{T}}\boldsymbol{y}$
- Let us compute the condition number of $\boldsymbol{X}^{\mathsf{T}}\boldsymbol{X} + \lambda \boldsymbol{I}_p$,

$$\boldsymbol{X}^{\mathsf{T}}\boldsymbol{X} + \lambda \boldsymbol{I}_p = \boldsymbol{V}\boldsymbol{D}^2\boldsymbol{V}^{\mathsf{T}} + \lambda \boldsymbol{V}\boldsymbol{V}^{\mathsf{T}} = \boldsymbol{V}\left(\boldsymbol{D}^2 + \lambda \boldsymbol{I}_p\right)\boldsymbol{V}^{\mathsf{T}}.$$

- $\left(\boldsymbol{D}^2 + \lambda \boldsymbol{I}_p\right)$ is a diagonal matrix whose elements in the diagonal are

$$d_j^2 + \lambda = \gamma_j + \lambda, \ j = 1, \ldots, p.$$

- Therefore the condition number of $\boldsymbol{X}^{\mathsf{T}}\boldsymbol{X} + \lambda \boldsymbol{I}_p$ is

$$\kappa\left(\boldsymbol{X}^{\mathsf{T}}\boldsymbol{X} + \lambda \boldsymbol{I}_p\right) = \sqrt{\frac{\gamma_1 + \lambda}{\gamma_p + \lambda}}$$

  lower than $\kappa\left(\boldsymbol{X}^{\mathsf{T}}\boldsymbol{X}\right) = \sqrt{\gamma_1/\gamma_p}$ for all $\lambda > 0$.

- By the way, $\left(\boldsymbol{X}^{\mathsf{T}}\boldsymbol{X} + \lambda \boldsymbol{I}_p\right)^{-1} = \boldsymbol{V}\left(\boldsymbol{D}^2 + \lambda \boldsymbol{I}_p\right)^{-1}\boldsymbol{V}^{\mathsf{T}}$, and $\left(\boldsymbol{D}^2 + \lambda \boldsymbol{I}_p\right)^{-1} = \text{Diagonal}(1/(d_j^2 + \lambda), j = 1, \ldots, p)$.

Variance of the ridge regression estimator

Remember that $\hat{\boldsymbol{\beta}}_{\text{ridge}} = \left(\boldsymbol{X}^{\mathsf{T}}\boldsymbol{X} + \lambda \boldsymbol{I}_p\right)^{-1} \boldsymbol{X}^{\mathsf{T}}\boldsymbol{y}$. Then

$$\begin{aligned}
\text{Var}(\hat{\boldsymbol{\beta}}_{\text{ridge}}) &= \left(\boldsymbol{X}^{\mathsf{T}}\boldsymbol{X} + \lambda \boldsymbol{I}_p\right)^{-1} \boldsymbol{X}^{\mathsf{T}}\text{Var}(\boldsymbol{y})\boldsymbol{X} \left(\boldsymbol{X}^{\mathsf{T}}\boldsymbol{X} + \lambda \boldsymbol{I}_p\right)^{-1} \\
&= \sigma^2 \left(\boldsymbol{X}^{\mathsf{T}}\boldsymbol{X} + \lambda \boldsymbol{I}_p\right)^{-1} \boldsymbol{X}^{\mathsf{T}}\boldsymbol{X} \left(\boldsymbol{X}^{\mathsf{T}}\boldsymbol{X} + \lambda \boldsymbol{I}_p\right)^{-1}.
\end{aligned}$$

From the s.v.d. of $\boldsymbol{X}$, $\boldsymbol{X} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^{\mathsf{T}}$, we have deduced that $\boldsymbol{X}^{\mathsf{T}}\boldsymbol{X} = \boldsymbol{V}\boldsymbol{D}^2\boldsymbol{V}^{\mathsf{T}}$ and that $\left(\boldsymbol{X}^{\mathsf{T}}\boldsymbol{X} + \lambda \boldsymbol{I}_p\right)^{-1} = \boldsymbol{V}\left(\boldsymbol{D}^2 + \lambda \boldsymbol{I}_p\right)^{-1}\boldsymbol{V}^{\mathsf{T}}$
Therefore,

$$\begin{aligned}
\text{Var}(\hat{\boldsymbol{\beta}}_{\text{ridge}}) &= \sigma^2 \boldsymbol{V}\left(\boldsymbol{D}^2 + \lambda \boldsymbol{I}_p\right)^{-1} \boldsymbol{D}^2 \left(\boldsymbol{D}^2 + \lambda \boldsymbol{I}_p\right)^{-1} \boldsymbol{V}^{\mathsf{T}} \\
&= \sigma^2 \boldsymbol{V}\,\text{Diagonal}\left(d_j^2/(d_j^2 + \lambda)^2, j = 1, \ldots, p\right) \boldsymbol{V}^{\mathsf{T}}.
\end{aligned}$$

Linear estimators of a regression function

- Let $(\boldsymbol{x}_i, y_i)$, $i = 1, \ldots, n$, be n i.i.d. observs. from the r.v. $(\boldsymbol{X}, Y)$.
- Let $m(\boldsymbol{x}) = \mathbb{E}(Y|\boldsymbol{X} = \boldsymbol{x})$ be the regression function of Y over $\boldsymbol{X}$.
- Let $\hat{m}(\boldsymbol{x})$ an estimator (parametric, non-parametric, ...) of the regression function $m(\boldsymbol{x})$.
- We say that $\hat{m}(\boldsymbol{x})$ is a linear estimator when for any fix $\boldsymbol{x}$, $\hat{m}(\boldsymbol{x})$ is a linear function of $y_1, \ldots, y_n$:

$$\hat{m}(\boldsymbol{x}) = \sum_{i=1}^{n} w_i(\boldsymbol{x}) y_i,$$

  where in fact $w_i(\boldsymbol{x}) = w_i(\boldsymbol{x}; \boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$.
- For the n observed values $\boldsymbol{x}_i$ of the explanatory variable, let

$$\hat{y}_i = \hat{m}(\boldsymbol{x}_i) = \sum_{j=1}^{n} w_j(\boldsymbol{x}_i) y_j$$

  be the fitted values.

- In matrix format,

$$\hat{\boldsymbol{y}} = \boldsymbol{W}\boldsymbol{y},$$

where the column vectors $\boldsymbol{y}$ and $\hat{\boldsymbol{y}}$ have elements y_i and $\hat{y}_i$, respectively, and the matrix $\boldsymbol{W}$ has generic (i, j) element

$$w_{ij} = w_j(\boldsymbol{x}_i).$$

- The matrix $\boldsymbol{W}$ is analogous to the hat matrix $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top$ in OLS estimation of the multiple linear regression:

$$\hat{\boldsymbol{y}}_{\text{OLS}} = \boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{y} = \boldsymbol{H}\boldsymbol{y}.$$

- Observe that ridge regression is a linear estimation method:

$$\hat{\boldsymbol{y}}_{\text{ridge}} = \boldsymbol{X}\left(\boldsymbol{X}^\top\boldsymbol{X} + \lambda\boldsymbol{I}_p\right)^{-1}\boldsymbol{X}^\top\boldsymbol{y} = \boldsymbol{H}_\lambda\boldsymbol{y}.$$

Effective number of parameters for linear estimators

- Consider the multiple linear regression with p regressors (including the constant term, if it appears in the model):

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \varepsilon,$$

  $\boldsymbol{X}$ being a $n \times p$ matrix, $\boldsymbol{\beta} \in \mathbb{R}^p$.

- It is known that

$$\text{Trace}(\boldsymbol{H}) = \text{Trace}(\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top) = \text{Trace}((\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{X}) = \text{Trace}(\boldsymbol{I}_p) = p,$$

  that is the number of parameters in the model.

- For a linear estimator with matrix $\boldsymbol{W}$ ($\hat{\boldsymbol{y}} = \boldsymbol{W}\boldsymbol{y}$) we define

$$\nu = \text{Trace}(\boldsymbol{W}) = \sum_{i=1}^{n} w_{ii},$$

  the sum of diagonal elements of $\boldsymbol{W}$.

- $\nu = \text{Trace}(\boldsymbol{W})$ is called the effective number of parameters of the linear estimator corresponding to matrix $\boldsymbol{W}$.
- In some books (and softwares) ν is called effective degrees of freedom (df) of the regression estimator. This is the terminology used by Hastie, Tibshirani, and Friedman (2009) and Hastie, Tibshirani, and Wainwright (2015), and related packages.
- The interpretation of ν as the effective number of parameters is valid for any linear estimator of the regression function (parametric, nonparametric, ...).
- Then we can compare the degree of complexity of two linear estimators of a regression function just comparing their effective numbers of parameters.
- Usually a good estimator of σ^2, the residual variance, is

$$\hat{\sigma}^2 = \frac{1}{n-\nu}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2.$$

Effective number of parameters in ridge regression

In the case of ridge regression $\nu = \nu(\lambda) = \mathrm{df}(\lambda)$ has an explicit expression:

$$\boldsymbol{W} = \boldsymbol{H}_\lambda = \boldsymbol{X}\left(\boldsymbol{X}^\mathsf{T}\boldsymbol{X} + \lambda \boldsymbol{I}_p\right)^{-1}\boldsymbol{X} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^\mathsf{T}\boldsymbol{V}\left(\boldsymbol{D}^2 + \lambda \boldsymbol{I}_p\right)^{-1}\boldsymbol{V}^\mathsf{T}\boldsymbol{V}\boldsymbol{D}\boldsymbol{U}^\mathsf{T} =$$

$$\boldsymbol{U}\boldsymbol{D}\left(\boldsymbol{D}^2 + \lambda \boldsymbol{I}_p\right)^{-1}\boldsymbol{D}\boldsymbol{U}^\mathsf{T} = \boldsymbol{U}\left(\mathrm{Diagonal}(d_j^2/(d_j^2+\lambda), j = 1, \ldots, p\right)\boldsymbol{U}^\mathsf{T}$$

$$\Rightarrow \nu(\lambda) = \mathrm{df}(\lambda) = \mathrm{Trace}(\boldsymbol{H}_\lambda) =$$

$$\mathrm{trace}(\boldsymbol{U}\left(\mathrm{Diagonal}(d_j^2/(d_j^2+\lambda), j = 1, \ldots, p\right)\boldsymbol{U}^\mathsf{T}) =$$

$$\mathrm{trace}(\left(\mathrm{Diagonal}(d_j^2/(d_j^2+\lambda), j = 1, \ldots, p\right)\boldsymbol{U}^\mathsf{T}\boldsymbol{U}) =$$

$$\mathrm{trace}(\left(\mathrm{Diagonal}(d_j^2/(d_j^2+\lambda), j = 1, \ldots, p)\right) = \sum_{j=1}^{p}\frac{d_j^2}{d_j^2+\lambda}.$$

$$\nu(\lambda) = \mathrm{df}(\lambda) = \sum_{j=1}^{p} \frac{d_j^2}{d_j^2 + \lambda}$$

- $\lim_{\lambda \longrightarrow \infty} \mathrm{df}(\lambda) = 0, \quad \lim_{\lambda \longrightarrow 0} \mathrm{df}(\lambda) = \mathrm{rank}(\boldsymbol{X})$.
- The effective number of parameters $\nu(\lambda) = \mathrm{df}(\lambda)$ is a decreasing function of penalizing parameter λ:
  - Small values of λ correspond to large numbers ν of effective parameters, close to the number of linearly independent explanatory variables (usually $\min\{n, p\}$), allowing complex and flexible estimators.
  - Large values of λ correspond to small numbers ν of effective parameters, that is, to regression estimators with low complexity and flexibility.

Practice:

- Prostate data: Effective number of parameters in ridge regression.
- Use the R script `prostate.ridge.regression.R`.

Effective degrees of freedom for non-linear estimators

- Let $\hat{m}(\boldsymbol{x})$ an estimator of the regression function $m(\boldsymbol{x})$ (a random function because it is based on $(Y_1, \ldots, Y_n)$). Let $\hat{Y}_i = \hat{m}(\boldsymbol{x}_i)$.
- The effective degrees of freedom of $\hat{m}(\boldsymbol{x})$ is defined as

$$\mathrm{df}(\hat{m}) = \frac{1}{\sigma^2} \sum_{i=1}^{n} \mathrm{Cov}(\hat{Y}_i, Y_i).$$

- Interpretation:
  - A very flexible regression estimator $\hat{m}(\boldsymbol{x})$ will be able to interpolate the observed data, and then

    $$\hat{Y}_i = Y_i,\ \mathrm{Var}(\hat{Y}_i) = \mathrm{Var}(Y_i) = \sigma^2,\ \mathrm{Cov}(\hat{Y}_i, Y_i)/\sigma^2 = \mathrm{Cor}(\hat{Y}_i, Y_i) = 1,$$

    so $\mathrm{df}(\hat{m}) = n$: $\hat{m}(\boldsymbol{x})$ has as many degrees of freedom as the number of observed data.
  - The constant function equal to the sample mean of $Y_1, \ldots, Y_n$ for all $\boldsymbol{x}$ has 1 degree of freedom.
  - A function that is constant in $\boldsymbol{x}$ has 0 degrees of freedom if this constant does not depend on the data.

Both definitions of df coincide in linear estimators

Assume that $\hat{m}(x)$ is a linear estimator with matrix $\boldsymbol{W}$. Assume also that $\mathbb{E}(Y) = 0$. Then

$$\mathrm{df}(\hat{m}) = \frac{1}{\sigma^2}\sum_{i=1}^{n}\mathrm{Cov}(\hat{Y}_i, Y_i) = \frac{1}{\sigma^2}\mathrm{Trace}\left(\mathrm{Cov}(\hat{\boldsymbol{Y}}, \boldsymbol{Y})\right) =$$

$$\frac{1}{\sigma^2}\mathrm{Trace}\left(\mathbb{E}(\hat{\boldsymbol{Y}}\boldsymbol{Y}^{\mathsf{T}})\right) = \frac{1}{\sigma^2}\mathrm{Trace}\left(\mathbb{E}(\boldsymbol{W}\boldsymbol{Y}\boldsymbol{Y}^{\mathsf{T}})\right) =$$

$$\frac{1}{\sigma^2}\mathrm{Trace}\left(\boldsymbol{W}\mathbb{E}(\boldsymbol{Y}\boldsymbol{Y}^{\mathsf{T}})\right) = \frac{1}{\sigma^2}\mathrm{Trace}\left(\boldsymbol{W}\sigma^2\boldsymbol{I}_p\right) = \mathrm{Trace}(\boldsymbol{W}).$$

Choosing the tuning parameter λ

- The tuning parameter λ can be chosen by cross-validation (CV), k-fold cross-validation (k-fold CV) or by generalized cross-validation (GCV).
- Given the expression of $\hat{\boldsymbol{\beta}}_{\text{ridge}}$ (linear in $\boldsymbol{y}$) CV and GCV are not computationally expensive.
- We will first introduce these concepts before talking about efficient computation.

- Predictive Mean Square Error (PMSE). It is the expected squared error made when predicting

$$Y = m(\boldsymbol{x}) + \varepsilon$$

by $\hat{m}(\boldsymbol{x})$, where $\boldsymbol{x}$ is an observation of the random variable $\boldsymbol{X}$, distributed as the observed explanatory variable, when $\boldsymbol{X}$ and ε are independent from the sample $\mathcal{Z} = \{(\boldsymbol{X}_i, Y_i) : i = 1, \ldots, n\}$ used to compute $\hat{m}$:

$$\text{PMSE}(\hat{m}) = \mathbb{E}_{\mathcal{Z},\mathbf{X},\varepsilon}\left[(Y - \hat{m}(\boldsymbol{X}))^2\right].$$

Prediction error in a validation set

- When the number of available data is large (as it usually happens in data mining or in Big Data problems) the sample is randomly divided in three sets:
  - The training set: it is used to fit the model.
  - The validation set: it is used to compute feasible versions of the PMSE for model selection and/or parameter tuning.
  - The test set: it is used to evaluate the generalization (or prediction) error of the final chosen model in independent data.
- Assuming that at least a validation set has been preserved, an estimation of PMSE is the Predictive Mean Squared Error in the validation set:

$$\text{PMSE}_{\text{val}}(\hat{m}) = \frac{1}{n_V}\sum_{i=1}^{n_V}(y_i^V - \hat{m}(\boldsymbol{x}_i^V))^2,$$

where $(\boldsymbol{x}_i^V, y_i^V)$, $i = 1, \ldots, n_V$, is the validation set and $\hat{m}(\boldsymbol{x})$ is the estimator computed using the training set.

Leave-one-out cross-validation

- When the sample size does not allow us to set a validation set aside, leave-one-out cross-validation is an attractive alternative:
  1. Remove the observation $(\boldsymbol{x}_i, y_i)$ from the sample and fit the regression using the other $(n-1)$ data. Let $\hat{m}_{(i)}(\boldsymbol{x})$ be the resulting estimator.
  2. Now use $\hat{m}_{(i)}(\boldsymbol{x}_i)$ to predict y_i.
  3. Repeat the previous steps for $i = 1, \ldots, n$.
  4. Compute

$$\text{PMSE}_{\text{CV}}(\hat{m}) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{m}_{(i)}(\boldsymbol{x}_i))^2.$$

- In ridge regression:

$$\lambda_{CV} = \arg\min_{\lambda \geq 0} \text{PMSE}_{\text{CV}}(\lambda) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \boldsymbol{x}_i^{\mathsf{T}}\hat{\boldsymbol{\beta}}^{(i)}_{\text{ridge},\lambda})^2.$$

Practice:

- Prostate data: Leave-one-out cross-validation in ridge regression.
- Use the R script `prostate.ridge.regression.R`.

k-fold cross validation

- $\text{PMSE}_{\text{CV}}(\hat{m})$ is an approximately unbiased estimator of $\text{PMSE}(\hat{m})$, but has a considerable variance.
- The variance can be reduced doing k-fold cross-validation: The sample is randomly divided in k subsets, each of them is removed by turns from the sample, the model is estimated with the other $(k-1)$ subsamples and the removed subsample is used to compute prediction errors.
- n-fold cross-validation is leave-one-out cross-validation.
- k-fold cross-validation has lower variance than leave-one-out cross-validation but larger bias.
- General recommendation: Use 5-fold or 10-fold cross-validation.

Efficient computation of PMSE_{CV}

- Consider a linear estimator of the regression function with matrix $\boldsymbol{W} = (w_{ij})_{i,j}$: $\hat{\boldsymbol{y}} = \boldsymbol{W}\boldsymbol{y}$.
- That is

$$\hat{y}_i = \sum_{j=1}^{n} w_{ij} y_j, \ \ i = 1, \ldots, n,$$

  where $w_{ij} = w_j(\boldsymbol{x}_i) = w_j(\boldsymbol{x}_i; \boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$.
- In these cases PMSE_{CV} can be calculated avoiding the computational cost of fitting n different regression models.
- For most linear estimators it can be proved that

$$\text{PMSE}_{\text{CV}} = \frac{1}{n} \sum_{i=1}^{n} \left(\frac{y_i - \hat{y}_i}{1 - w_{ii}} \right)^2.$$

Self-stable linear estimators

(See Fan, Li, Zhang, and Zou 2020, page 48)

- Let $\hat{m}(\boldsymbol{x})$ be a linear estimator of the regression function $m(\boldsymbol{x}) = \mathbb{E}(Y|\boldsymbol{X} = \boldsymbol{x})$ fitted on the data $(\boldsymbol{x}_i, y_i)$, $i = 1, \ldots, n$.
- Let $\boldsymbol{x}_0$ be a new covariate vector and $\hat{m}(\boldsymbol{x}_0)$ be its predicted value using the linear estimator.
- We augment the data set by including $(\boldsymbol{x}_0, \hat{m}(\boldsymbol{x}_0))$ as a new point, and refit the linear estimator on this augmented data set: $\hat{m}_{\mathrm{a}}(\boldsymbol{x})$.
- The linear estimator of the regression function is said to be **self-stable** if the fit based on the augmented data set is identical to the fit based on the original data regardless of x_0.

Self-stability and efficient computation of PMSE_{CV}

Theorem (Theorem 2.7, Fan, Li, Zhang, and Zou 2020)

For any linear smoother $\hat{\boldsymbol{Y}} = \boldsymbol{W}\boldsymbol{Y}$ *with the self-stability property, we have*

$$y_i - \hat{m}_{(i)}(\boldsymbol{x}_i) = \frac{y_i - \hat{y}_i}{1 - w_{ii}}$$

and, therefore, its leave-one-out cross validation error is

$$PMSE_{CV} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i - \hat{y}_i}{1 - w_{ii}}\right)^2.$$

Proof for the ridge regression estimation

- Let $\hat{\boldsymbol{\beta}}^{(i)}_{\text{ridge},\lambda}$ be the estimation of $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)$ when leaving out the i-th observation:

$$\hat{\boldsymbol{\beta}}^{(i)}_{\text{ridge},\lambda} = \arg\min_{\boldsymbol{\beta}} \left\{ \sum_{l=1, l\neq i}^{n} \left(y_l - \sum_{j=1}^{p} x_{lj}\beta_j \right)^2 + \lambda \|\boldsymbol{\beta}\|_2^2 \right\}$$

- Let us define

$$\tilde{y}_l^{(i)} = \begin{cases} y_l & \text{if} \quad l \neq i, \\ \hat{y}_i^{(i)} = \sum_{j=1}^{p} x_{ij} \hat{\beta}^{(i)}_{\text{ridge},\lambda,j} & \text{if} \quad l = i. \end{cases}$$

- It follows that for all $\boldsymbol{\beta} \in \mathbb{R}^p$,

$$\sum_{l=1}^{n} \left(\tilde{y}_l^{(i)} - \sum_{j=1}^{p} x_{lj} \beta_j \right)^2 + \lambda \|\boldsymbol{\beta}\|_2^2 =$$

$$\left\{ \sum_{l=1, l \neq i}^{n} \left(y_l - \sum_{j=1}^{p} x_{lj} \beta_j \right)^2 + \lambda \|\boldsymbol{\beta}\|_2^2 \right\} + \left(\sum_{j=1}^{p} x_{ij} (\hat{\beta}^{(i)}_{\text{ridge},\lambda,j} - \beta_j) \right)^2$$

- Observe that $\hat{\boldsymbol{\beta}}^{(i)}_{\text{ridge},\lambda}$ minimizes both terms in the right hand side. Then it is also

$$\hat{\boldsymbol{\beta}}^{(i)}_{\text{ridge},\lambda} = \arg\min_{\boldsymbol{\beta}} \left\{ \sum_{l=1}^{n} \left(\tilde{y}^{(i)}_l - \sum_{j=1}^{p} x_{lj}\beta_j \right)^2 + \lambda \|\boldsymbol{\beta}\|_2^2 \right\}$$

- This is the ridge regression estimator corresponding to a data set with matrix of explanatory variables $\boldsymbol{X}$ and vector of responses $\tilde{\boldsymbol{y}}^{(i)} = (\tilde{y}^{(i)}_1, \ldots, \tilde{y}^{(i)}_n)^{\mathsf{T}}$.
- Then

$$\hat{\boldsymbol{\beta}}^{(i)}_{\text{ridge},\lambda} = (\boldsymbol{X}^{\mathsf{T}}\boldsymbol{X} + \lambda \boldsymbol{I}_p)^{-1}\boldsymbol{X}^{\mathsf{T}}\tilde{\boldsymbol{y}}^{(i)},$$

$$\hat{\tilde{\boldsymbol{y}}}^{(i)} = \boldsymbol{X}\hat{\boldsymbol{\beta}}^{(i)}_{\text{ridge},\lambda} = \boldsymbol{X}(\boldsymbol{X}^{\mathsf{T}}\boldsymbol{X} + \lambda \boldsymbol{I}_p)^{-1}\boldsymbol{X}^{\mathsf{T}}\tilde{\boldsymbol{y}}^{(i)} = \boldsymbol{H}_\lambda \tilde{\boldsymbol{y}}^{(i)}.$$

- Observe that the i-th element of $\hat{\tilde{\boldsymbol{y}}}^{(i)}$ is just $\hat{y}_i^{(i)} = \sum_{j=1}^{p} x_{ij} \hat{\beta}^{(i)}_{\text{ridge},\lambda,j}$.
- Let $\boldsymbol{e}_i$ be the n-dimensional vector whose i-th element is 1 and the others are equal to 0.
- Then $\tilde{\boldsymbol{y}}^{(i)} = \boldsymbol{y} - (y_i - \hat{y}_i^{(i)})\boldsymbol{e}_i$ and, consequently,

$$\hat{\tilde{\boldsymbol{y}}}^{(i)} = \boldsymbol{H}_\lambda \tilde{\boldsymbol{y}}^{(i)} = \boldsymbol{H}_\lambda \left(\boldsymbol{y} - (y_i - \hat{y}_i^{(i)})\boldsymbol{e}_i \right) =$$

$$\boldsymbol{H}_\lambda \boldsymbol{y} - (y_i - \hat{y}_i^{(i)})\boldsymbol{H}_\lambda \boldsymbol{e}_i = \hat{\boldsymbol{y}} - (y_i - \hat{y}_i^{(i)})\boldsymbol{h}_i^\lambda,$$

where $\boldsymbol{h}_i^\lambda$ is the i-th column of $\boldsymbol{H}_\lambda$.

- Looking just at the i-th component, $\hat{y}_i^{(i)} = \hat{y}_i - (y_i - \hat{y}_i^{(i)})h_{ii}^{\lambda}$, where h_{ii}^{λ} is the element (i, i) of $\boldsymbol{H}_{\lambda}$, or the i-th element in the diagonal of $\boldsymbol{H}_{\lambda}$.
- Then $y_i - \hat{y}_i^{(i)} = y_i - \hat{y}_i + (y_i - \hat{y}_i^{(i)})h_{ii}^{\lambda}$ and we conclude that

$$y_i - \hat{y}_i^{(i)} = \frac{y_i - \hat{y}_i}{1 - h_{ii}^{\lambda}}.$$

So the loo-CV errors $(y_i - \hat{y}_i^{(i)})$ can be computed if we know the errors $(y_i - \hat{y}_i)$ when fitting the ridge regression with all the data, and the diagonal of $\boldsymbol{H}_{\lambda}$, and the proof concludes.

Practice:

- Prostate data: Efficient computation of PMSE_{CV} in ridge regression.
- Use the R script `prostate.ridge.regression.R`.

Generalized cross-validation

- For linear estimators of the regression function, a modification can be done in the measure of PMSE_{CV}.
- It is known as generalized cross-validation (GCV).
- It consists in replacing in the expression of PMSE_{CV} the values w_{ii}, coming from the diagonal of $\boldsymbol{W}$, by their average value:

$$\text{PMSE}_{\text{GCV}} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i - \hat{y}_i}{1 - \nu/n}\right)^2,$$

  $\nu = \text{Trace}(\boldsymbol{W}) = \sum_{i=1}^{n} w_{ii}$ is the effective number of parameters.
- In ridge regression, $\lambda_{\text{GCV}} = \arg\min_{\lambda} \text{PMSE}_{\text{GCV}}(\lambda)$.
- Manipulating the expression of PMSE_{GCV} it follows that

$$\text{PMSE}_{\text{GCV}} = \frac{n\hat{\sigma}_{\varepsilon}^2}{n - \nu},$$

  where $\hat{\sigma}_{\varepsilon}^2 = \frac{1}{n-\nu}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ estimates the residual variance.

Practice:

- Prostate data: PMSE_{GCV} in ridge regression.
- Use the R script

  `prostate.ridge.regression.R`.

Fan, J., R. Li, C.-H. Zhang, and H. Zou (2020).
Statistical foundations of data science.
Chapman and Hall/CRC.

Hastie, T., R. Tibshirani, and J. Friedman (2009).
The Elements of Statistical Learning (2nd ed.).
Springer.

Hastie, T., R. Tibshirani, and M. Wainwright (2015).
Statistical learning with sparsity: the lasso and generalizations.
CRC Press.