24 Lecture 24: April 4

Last time

- Diagnosing nonlinearity (JF chapter 12)
- Data transformation (JF chapter 4)

Today

- Collinearity (JF chapter 13, RD 8.3.2)
- Principal component analysis (JF 13.1.1, RD 8.3.4)
- Biased estimation:
  - Ridge Regression
  - Lasso Regression

Additional reference

- "A First Course in Linear Model Theory" by Nalini Ravishanker and Kipak K. Dey
- Lecture notes by Cedric Ginestet

Collinearity

In linear model

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$
$$\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I}_n)$$

Collinearity (or multicollinearity) exists when there is "near-dependency" between the columns of the design matrix $\mathbf{X}$.

- Two or more columns.
- In other words, high correlation between explanatory variables.
- the data/model pair is ill-conditioned when $\mathbf{X}^T\mathbf{X}$ is nearly singular.

Perfect collinearity leads to rank-deficiency in $\mathbf{X}$ such that $\mathbf{X}^T\mathbf{X}$ is singular. In the case of perfect collinearity, two or more columns are linear-dependent.

An example of perfect collinearity

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \beta_4 X_{i4} + \beta_5 X_{i5} + \epsilon_i$$

Consider the case, where

- Y_i represents the amount of sales.

- $X_{i1}, X_{i2}, ..., X_{i4}$ are categorical that represent the quarter in which the sample is collected: $X_{ij} = \mathbf{1}(\text{sample } i \text{ collected in quarter } j)$.
- X_{i5} represents expense spent in advertising.

The dummy variable trap $X_{i4} = 1 - X_{i1} - X_{i2} - X_{i3}$. Recall that we need $m - 1$ dummy variables for m categories.

An example of high correlation between predictors

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \epsilon_i$$

Consider the case, where

- Y_i represents the salary of individual i.
- X_{i1} represents the age of individual i.
- X_{i2} represents the experience of individual i.

How to interpret β_1?

We expect high correlation between age and experience.

Problems caused by multicollinearity

1. large standard errors of the regression coefficients
    - small associated t-statistics
    - conclusion that truly useful explanatory variables are insignificant in explaining the regression
2. the sign of regression coefficients may be the opposite of what a mechanistic understanding of the problem would suggest
3. deleting a column of the predictor matrix will cause large changes in the coefficient estimates for other variables

However, multicollinearity does **not** greatly affect the **predicted values**.

Signs and detections of multicollinearity

Some signs for multicollinearity:

1. Simple correlation between a pair of predictors exceeds 0.9 or R^2.
2. High value of the multiple correlation coefficient with some high partial correlations between the explanatory variables.
3. Large F-statistics with some small t-statistics for individual regression coefficients

Some approaches for detecting multicollinearity:

1. Pairwise correlations among the explanatory variables
2. Variance inflation factor
3. Condition number

Variance inflation factor

For a multiple linear regression with k explanatory variables. We can regress X_j on the $(k-1)$ other explanatory variables and denote R_j as the coefficient of determination.

Then the variance inflation factor (VIF) is defined as

$$\text{VIF}_j = \frac{1}{1 - R_j^2}$$

- $\text{VIF}_j \in [1, +\infty)$
- A suggested threshold is 10
- May use the averaged $\overline{\text{VIF}} = \sum_{j=1}^{k} \text{VIF}_j \Big/ k$.

Condition index and condition number

We first scale the design matrix $\mathbf{X}$ into column-equilibrated predictor matrix $\mathbf{X}_E$ such that $\{X_E\}_{ij} = X_{ij}/\sqrt{\mathbf{X}_j^T \mathbf{X}_j}$.

Let $\mathbf{X}_E = \mathbf{U}\mathbf{D}\mathbf{V}^T$ be the singular-value decomposition (SVD) of the $n \times p$ matrix $\mathbf{X}_E$ where $\mathbf{U}^T\mathbf{U} = \mathbf{V}^T\mathbf{V} = \mathbf{I}_p$ and $\mathbf{D} = diag(d_1, d_2, ..., d_p)$ is a diagonal matrix with $d_j \geqslant 0$.

The j^{th} condition index is defined as

$$\eta(\mathbf{X}_E) = d_{\max}/d_j, \;\; j = 1, 2, ..., p$$

The condition number is defined as

$$C = d_{\max}/d_{\min}$$

$C \geqslant 1$, $d_{\max} = \max\limits_{1 \leqslant j \leqslant p} d_j$ and $d_{\min} = \min\limits_{1 \leqslant j \leqslant p} d_j$

Some properties of the condition number

- Large condition number indicates evidence of multicollinearity
- Typical cutoff values, 10, 15 to 30.

Some problems with the condition number

- practitioners have different opinions of whether $\mathbf{X}$ should be centered around their means for SVD.
  - centering may remove nonessential ill conditioning, e.g. $Cor(X, X^2)$
  - centering may mask the role of the constant term in any underlying near-dependencies
- the degree of multicollinearity with dummy variables may be influenced by the choice of reference category
- condition number is affected by the scale of the $\mathbf{X}$ measurements
  - By scaling down any column of $\mathbf{X}$, the condition number can be made arbitrarily large
  - Known as *artificial ill-conditioning*
  - The condition number of the scaled matrix $\mathbf{X}_E$ is also referred to as the *scaled condition number*

Recall that $\mathbf{X}_E = \mathbf{UDV}^T$ is the singular-value decomposition (SVD) of $\mathbf{X}_E$, where $\mathbf{U}^T\mathbf{U} = \mathbf{V}^T\mathbf{V} = \mathbf{I}_p$ and $\mathbf{D} = diag(d_1, d_2, ..., d_p)$ is a diagonal matrix with $d_j \geqslant 0$.

Then

$$\begin{aligned}\mathbf{X}_E^T\mathbf{X}_E &= \mathbf{VDU}^T\mathbf{UDV}^T \\ &= \mathbf{VD}^2\mathbf{V}^T\end{aligned}$$

is the spectral decomposition of the Gramian matrix $\mathbf{X}_E^T\mathbf{X}_E$ with $\{d_j^2\}$ being the eigenvalues and $\mathbf{V}$ being the corresponding eigen vector matrix. This relationship links the condition numbers to the eigen values of the Gramian matrix.

Variance decomposition method

The variance-covariance matrix of the coefficient

$$\begin{aligned}Cov(\hat{\boldsymbol{\beta}}) &= \sigma^2(\mathbf{X}_E^T\mathbf{X}_E)^{-1} \\ &= \sigma^2\mathbf{VD}^{-2}\mathbf{V}^T\end{aligned}$$

Its j^{th} diagonal element is the estimated variance of the j^{th} coefficient, $\hat{\beta}_j$. Then

$$Var(\hat{\beta}_j) = \sigma^2 \sum_{h=1}^{p} \frac{v_{jh}^2}{d_h^2}$$

- Let $q_{jh} = \frac{v_{jh}^2}{d_h^2}$ and $q_j = \sum_{h=1}^{p} q_{jh}$.

Condition Index	Proportions of variance			
	$Var(\hat{\beta}_1)$	$Var(\hat{\beta}_2)$	...	$Var(\hat{\beta}_3)$
η_1	π_{11}	π_{12}	...	π_{1p}
η_2	π_{21}	π_{22}	...	π_{2p}
$\vdots$	$\vdots$	$\vdots$		$\vdots$
η_p	π_{p1}	π_{p2}	...	π_{pp}

Table 1: Table of condition index and proportions of variance

- The variance decomposition proportion is $\pi_{jh} = q_{jh}/q_j$.
- π_{jh} denotes the proportion of the variance of the j^{th} regression coefficient associated with the h^{th} component of its decomposition.
- The variance decomposition proportion matrix is $\mathbf{\Pi} = \{\pi_{jh}\}$.

In practice, it is suggested to combine condition index and proportions of variance for multicollinearity diagnostic. Identify multicollinearity if

- Two or more elements in the j^{th} row of matrix $\mathbf{\Pi}$ are relatively large
- And its associated condition index η_j is large too

Principal Components

The method of principal components, introduced by Karl Pearson (1901) and Harold Hotelling (1933), provides a useful representation of the correlational structure of a set of variables. Some advantages of the principal component analysis include

- more unified
- linear transformation of the original predictors into a new set of orthogonal predictors
- the new orthogonal predictors are called principal components

Principal components regression is an approach that inspects the sample data $(\mathbf{Y}, \mathbf{X})$ for directions of variability and uses this information to reduce the dimensionality of the estimation problem. The procedure is based on the observation that every linear regression model can be restated in terms of a set of orthogonal predictor variables, which are constructed as linear combinations of the original variables. The new orthogonal variables are called the principal components of the original variables.

Let $\mathbf{X}^T\mathbf{X} = \mathbf{Q}\mathbf{\Delta}\mathbf{Q}^T$ denote the spectral decomposition of $\mathbf{X}^T\mathbf{X}$, where $\mathbf{\Delta} = diag\{\lambda_1, \ldots, \lambda_p\}$ is a diagonal matrix consisting of the (real) eigenvalues of $\mathbf{X}^T\mathbf{X}$, with $\lambda_1 \geqslant \cdots \geqslant \lambda_p$ and $\mathbf{Q} = (\mathbf{q_1}, \ldots, \mathbf{q_p})$ denotes the matrix whose columns are the orthogonal eigenvectors of $\mathbf{X}^T\mathbf{X}$ corresponding to the ordered eigenvalues. Consider the transformation

$$\mathbf{Y} = \mathbf{X}\mathbf{Q}\mathbf{Q}^T\boldsymbol{\beta} + \boldsymbol{\epsilon} = \mathbf{Z}\boldsymbol{\theta} + \boldsymbol{\epsilon},$$

where $\mathbf{Z} = \mathbf{XQ}$, and $\theta = \mathbf{Q}^T\beta$.
The elements of θ are known as the regression parameters of the principal components. The matrix $\mathbf{Z} = \{\mathbf{z_1}, \ldots, \mathbf{z_p}\}$ is called the matrix of principal components of $\mathbf{X}^T\mathbf{X}$. $\mathbf{z}_j = \mathbf{Xq}_j$ is the jth principal component of $\mathbf{X}^T\mathbf{X}$ and $\mathbf{z}_j^T\mathbf{z}_j = \lambda_j$, the jth largest eigenvalue of $\mathbf{X}^T\mathbf{X}$.

Principal components regression consists of deleting one or more of the variables $\mathbf{z}_j$ (which correspond to small values of λ_j), and using OLS estimation on the resulting reduced regression model.

Derivation under standardized predictors, JF 13.1.1

Consider the vectors of standardized predictors, $\mathbf{x}_1^*, \mathbf{x}_2^*, \ldots, \mathbf{x}_p^*$ (obtained by subtracting the mean and divided by standard deviation of the original predictor vectors). Because the principal components are linear combinations of the original predictors, we write the first principal component as

$$\begin{aligned}\mathbf{w}_1 &= A_{11}\mathbf{x}_1^* + A_{21}\mathbf{x}_2^* + \cdots + A_{p1}\mathbf{x}_p^* \\ &= \mathbf{X}^*\mathbf{a}_1\end{aligned}$$

The variance of the first component becomes

$$\begin{aligned}S_{w_1}^2 &= \frac{1}{n-1}\mathbf{w}_1^T\mathbf{w}_1 \\ &= \frac{1}{n-1}\mathbf{a}_1^T\mathbf{X}^{*T}\mathbf{X}^*\mathbf{a}_1 \\ &= \mathbf{a}_1^T\mathbf{R}_{XX}\mathbf{a}_1\end{aligned}$$

where $\mathbf{R}_{XX} = \frac{1}{n-1}\mathbf{X}^{*T}\mathbf{X}^*$. We want to maximize $S_{w_1}^2$ under the normalizing constraint $\mathbf{a}_1^T\mathbf{a}_1 = 1$ (otherwise $S_{w_1}^2$ can be arbitrarily large by inflating $\mathbf{a}_1$). Consider

$$F_1 \equiv \mathbf{a}^T\mathbf{R}_{XX}\mathbf{a}_1 - L_1(\mathbf{a}_1^T\mathbf{a}_1 - 1)$$

where L_1 is a Lagrange multiplier. By differentiating this equation with respect to $\mathbf{a}_1$ and L_1,

$$\begin{aligned}\frac{\partial F_1}{\partial \mathbf{a}_1} &= 2\mathbf{R}_{XX}\mathbf{a}_1 - 2L_1\mathbf{a}_1 \\ \frac{\partial F_1}{\partial L_1} &= -(\mathbf{a}_1^T\mathbf{a}_1 - 1)\end{aligned}$$

Setting the partial derivatives to 0 produces

$$\begin{aligned}(\mathbf{R}_{XX} - L_1\mathbf{I}_p)\mathbf{a}_1 &= \mathbf{0} \\ \mathbf{a}_1^T\mathbf{a}_1 &= 1\end{aligned}$$

From the first equation, we see that L_1 is an eigenvalue of $\mathbf{R}_{XX}$ such that $\mathbf{R}_{XX}\mathbf{a}_1 = L_1\mathbf{a}_1$ such that

$$S_{w_1}^2 = \mathbf{a}_1^T\mathbf{R}_{XX}\mathbf{a}_1 = L_1\mathbf{a}_1^T\mathbf{a}_1 = L_1$$

To maximize $S_{w_1}^2$, we only need to pick the largest eigenvalue of $\mathbf{R}_{XX}$.

Ridge Regression

Ridge regression and the Lasso regression are two forms of regularized regression. These methods can be used to alleviate the consequences of multicollinearity.

1. When variables are highly correlated, a large coefficient in one variable may be alleviated by a large coefficient in another variable, which is negatively correlated to the former.
2. Regularization imposes an upper threshold on the values taken by the coefficients, thereby producing a more parsimonious solution, and a set of coefficients with smaller variance.

Constrained optimization

Ridge regression is motivated by a constrained minimization problem, which can be formulated as

$$\hat{\boldsymbol{\beta}}^{ridge} = \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^p} \sum_{i=1}^{n}(y_i - \mathbf{x}_i^T\boldsymbol{\beta})^2$$

$$\text{subject to } ||\boldsymbol{\beta}||_2^2 = \sum_{j=1}^{p}\beta_j^2 \leqslant t$$

for $t \geqslant 0$.

Use a Lagrange multiplier, we can rewrite the formula as

$$\hat{\boldsymbol{\beta}}^{ridge} = \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^p}\{\sum_{i=1}^{n}(y_i - \mathbf{x}_i^T\boldsymbol{\beta})^2 + \lambda\sum_{j=1}^{p}\beta_j^2\}$$

for $\lambda \geqslant 0$ and where there is a one-to-one correspondence between t and λ. λ is an arbitrary constant usually referred to as the "ridge constant".

Analytical solutions

The ridge-regression estimator has analytical solution

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

This is obtained by differentiating the objective function with respect to $\boldsymbol{\beta}$ and set it to 0:

$$\begin{aligned}&\frac{\partial}{\partial\boldsymbol{\beta}}\{(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta})^T(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta}) + \lambda\boldsymbol{\beta}^T\boldsymbol{\beta}\}\\ =&2(\mathbf{X}^T\mathbf{X})\boldsymbol{\beta} - 2\mathbf{X}^T\mathbf{Y} + 2\lambda\boldsymbol{\beta}\\ =&0\end{aligned}$$

Therefore,

$$(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})\boldsymbol{\beta} = \mathbf{X}^T\mathbf{Y}$$

Since we are adding a positive constant to the diagonal of $\mathbf{X}^T\mathbf{X}$, we are , in general, producing an invertible matrix, $\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I}$ even if $\mathbf{X}^T\mathbf{X}$ is singular. Historically, this particular aspect of ridge regression was the main motivation behind the adoption of this particular extension of OLS theory.

The ridge regression estimator is related to the classical OLS estimator, $\hat{\boldsymbol{\beta}}^{OLS}$, in the following manner

$$\hat{\boldsymbol{\beta}}^{ridge} = \left[\mathbf{I} + \lambda(\mathbf{X}^T\mathbf{X})^{-1}\right]^{-1} \hat{\boldsymbol{\beta}}^{OLS},$$

assuming $\mathbf{X}^T\mathbf{X}$ is non-singular. This relationship can be verified by applying the definition of $\hat{\boldsymbol{\beta}}^{OLS}$,

$$\begin{aligned}\hat{\boldsymbol{\beta}}^{ridge} &= \left[\mathbf{I} + \lambda(\mathbf{X}^T\mathbf{X})^{-1}\right]^{-1} (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y} \\ &= (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T\mathbf{Y}\end{aligned}$$

using the fact $\mathbf{B}^{-1}\mathbf{A}^{-1} = (\mathbf{AB})^{-1}$.

Moreover, when $\mathbf{X}$ is composed of orthonormal variables, such that $\mathbf{X}^T\mathbf{X} = \mathbf{I}_p$, it then follows that

$$\hat{\boldsymbol{\beta}}^{ridge} = \frac{1}{1+\lambda}\hat{\boldsymbol{\beta}}^{OLS}$$

Bias and variance of ridge estimator

Ridge estimation produces a biased estimator of the true parameter $\boldsymbol{\beta}$. With the definition of $\hat{\boldsymbol{\beta}}^{ridge}$ and the model assumption $\mathbf{E}(\mathbf{Y}|\mathbf{X}) = \mathbf{X}\boldsymbol{\beta}$, we obtain,

$$\begin{aligned}\mathbf{E}\left(\hat{\boldsymbol{\beta}}^{ridge}|\mathbf{X}\right) &= (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T\mathbf{X}\boldsymbol{\beta} \\ &= (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I} - \lambda\mathbf{I})\boldsymbol{\beta} \\ &= \boldsymbol{\beta} - \lambda(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\boldsymbol{\beta}\end{aligned}$$

where the bias of the ridge estimator is proportional to λ. The variance of the ridge estimator is

$$\mathbf{Var}\left(\hat{\boldsymbol{\beta}}^{ridge}|\mathbf{X}\right) = \sigma^2(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}(\mathbf{X}^T\mathbf{X})(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}.$$

When λ increases, the inverted term $(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}$ is increasingly dominated by $\lambda\mathbf{I}$. The variance of the ridge estimator, therefore, is a decreasing function of λ. This result is intuitively reasonable because the estimator itself is driven toward $\mathbf{0}$.

Variance-bias tradeoff

The mean-squared error of an estimator can be decomposed into the sum of its squared bias and sampling variance.

$$\begin{aligned}MSE(\hat{\theta}) &= \mathbf{E}\left((\hat{\theta} - \theta)^2\right) = \mathbf{E}(\hat{\theta}^2) + \theta^2 - 2\theta\mathbf{E}(\hat{\theta}) \\ Bias^2(\hat{\theta}) &= \left[\mathbf{E}(\hat{\theta}) - \theta\right]^2 = \mathbf{E}^2(\hat{\theta}) + \theta^2 - 2\theta\mathbf{E}(\hat{\theta}) \\ \mathrm{Var}(\hat{\theta}) &= \mathbf{E}(\hat{\theta}^2) - \mathbf{E}^2(\hat{\theta})\end{aligned}$$

Therefore

$$MSE(\hat{\theta}) = Bias^2(\hat{\theta}) + \text{Var}(\hat{\theta})$$

The essential idea here is to trade a small amount of bias in the coefficient estimates for a large reduction in coefficient sampling variance. Hoerl and Kennard (1970) prove that it is always possible to choose a positive value of the ridge constant λ so that the mean-squared error of the ridge estimator is less than the mean-squared error of the least-squares estimator. These ideas are illustrated heuristically in Figure 24.1

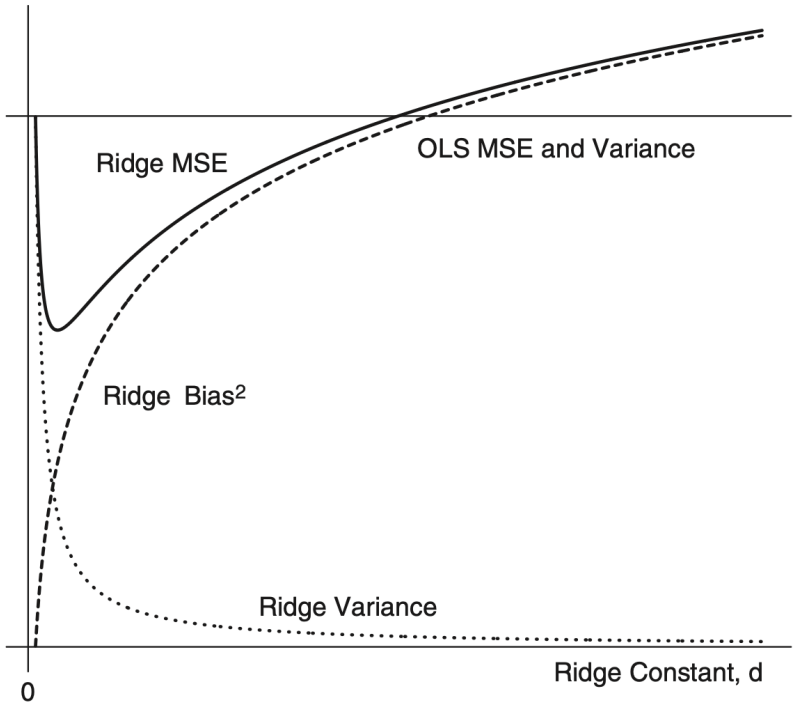

Figure 24.1: Trade-off of bias and against variance for the ridge-regression estimator. The horizontal line gives the variance of the least-squares (OLS) estimator; because the OLS estimator is unbiased, its variance and mean-squared error are the same. The broken line shows the squared bias of the ridge estimator as an increasing function of the ridge constant d (i.e. λ in our notes). The dotted line shows the variance of the ridge estimator. The mean-squared error (MSE) of the ridge estimator, given by the heavier solid line, is the sum of its variance and squared bias. For some values of d, the MSE error of the ridge estimator is below the variance of the OLS estimator. JF Figure 13.9.

Lasso regression

We have seen that ridge regression essentially re-scales the OLS estimates. The lasso, by contrast, tries to produce a *sparse* solution, in the sense that several of the slope parameters will be set to zero.

Constrained optimization

Different from the L_2 penalty for ridge regression, the Lasso regression employs L_1-penalty.

$$\hat{\boldsymbol{\beta}}^{lasso} = \underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\arg\min} \sum_{i=1}^{n}(y_i - \mathbf{x}_i^T\boldsymbol{\beta})^2$$
$$\text{subject to } ||\boldsymbol{\beta}||_1 = \sum_{j=1}^{p}|\boldsymbol{\beta}_j| \leqslant t$$

for $t \geqslant 0$; which can again be re-formulated using the Lagrangian for the L_1-penalty,

$$\hat{\boldsymbol{\beta}}^{lasso} = \underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\arg\min}\{\sum_{i=1}^{n}(y_i - \mathbf{x}_i^T\boldsymbol{\beta})^2 + \lambda\sum_{j=1}^{p}|\beta_j|\}$$

where $\lambda > 0$ and, as before, there exists a one-to-one correspondence between t and λ.

Parameter estimation

Contrary to ridge regression, the Lasso does not have a closed-form solution. The L_1-penalty makes the solution non-linear in y_i's. The above constrained minimization is a quadratic programming problem, for which many solvers exist.

Choice of Hyperparameters

Regularization parameter

The choice of λ in both ridge and lasso regressions is more of an art than a science. This parameter can be constructed as a complexity parameter, since as λ increases, less and less effective parameters are likely to be included in both ridge and lasso regressions. Therefore, one can adopt a model selection perspective and compare different choices of λ using cross-validation or an information criterion. That is, the value of λ should be chosen adaptively, in order to minimize an estimate of the expected prediction error (as in cross-validation), for instance, which is well approximated by AIC. We will discuss model selection in more detail later.

Bayesian perspective

The penalty terms in ridge and lasso regression can also be justified, using a Bayesian framework, whereby these terms arise as aresult of the specification of a particular prior distribution on the vector of slope parameters.

1. The use of an L_2-penalty in multiple regression is analogous to the choice of a Normal prior on the β_j's, in Bayesian statistics.

$$y_i \overset{iid}{\sim} \mathcal{N}(\beta_0 + \mathbf{x}_i^T\boldsymbol{\beta}, \sigma^2), \quad i = 1, \ldots, n$$
$$\beta_j \overset{iid}{\sim} \mathcal{N}(0, \tau^2), \quad j = 1, \ldots, p$$

2. Similarly, the use of an L_1-penalty in multiple regression is analogous to the choice of a Laplace prior on the β_j's, such that

$$\beta_j \overset{iid}{\sim} Laplace(0, \tau^2), \quad j = 1, \ldots, p$$

In both cases, the value of the hyperparameter, τ^2, will be inversely proportional to the choice of the particular value for λ. For ridge regression, λ is exactly equal to the shrinkage parameter of the hierarchical model, $\lambda = \sigma^2/\tau^2$.