# Chapter 5. Advanced topics in modern linear models.

Linear models were largely developed in the pre-computer age of statistics, but even in today's computer era there are still good reasons to study and use them. They are simple and often provide an adequate and interpretable description of how the inputs (predictor) affect the output. For prediction purposes they can sometimes outperform fancier nonlinear models, especially in situations with small numbers of training cases, low signal-to-noise ratio or sparse data. In this chapter, we describe the linear methods for regression in penalized regularization framework, such as ridge regression, the least absolute shrinkage and selection operator (LASSO), in details. It is my firm belief that an understanding of linear methods is essential for understanding nonlinear ones. Please refer to the two books entitled "*The elements of statistical learning*" by Hastie, Tibshirani and Friedman (2011) and "*Statistical learning with sparsity: the Lasso and generalizations*" by Hastie, Tibshirani and Wainwright (2015) for more detailed discussions on the topics covered in this chapter.

## 5.1 The Bias-Variance Trade-off of estimating $\boldsymbol{\beta}$

Generally, one may assume that the output/response $Y$ and the input/predictor $Z$ satisfy

$$y = f(\mathbf{Z}) + \varepsilon,$$

where $E(\varepsilon) = 0$ and $Var(\varepsilon) = \sigma^2$. When $f(\mathbf{z})$ is assumed to be $\mathbf{Z}\boldsymbol{\beta}$, so far, we've been dealing with, $\hat{\boldsymbol{\beta}}^{ls}$, the least square estimate, which has well-known properties (e.g. the Gauss-Markov, MLE in normal case). That is, the least squares estimate of the parameter $\beta$ have the smallest variance among all linear unbiased estimates. But, can we do better?

Consider the mean squared error (MSE) of an estimator $\hat{\theta}$ in estimating $\theta$:

$$MSE(\hat{\theta}) = E(\hat{\theta} - \theta)^2 = Var(\hat{\theta}) + \{E(\hat{\theta}) - \theta\}^2.$$

The first term is the variance, while the second term is the squared bias. The Gauss-Markov theorem implies that the least squares estimator has the smallest mean squared error of all linear estimators with no bias. However, the restriction to unbiased estimates is not necessarily a wise one. There may well exist a biased estimator with smaller mean squared error. Such an estimator would trade a little bias for a larger reduction in variance. Biased estimators are commonly used. Any method that shrinks or sets to zero some of the least squares coefficients may result in a biased estimate.

some estimated value.

For illustration, suppose we have an estimator $\hat{f}(\mathbf{z}) = \mathbf{Z}\hat{\boldsymbol{\beta}}$. To see whether it is a good candidate, there are usually two questions we need to consider: (1) Is $\hat{\boldsymbol{\beta}}$ close to the true $\boldsymbol{\beta}$? (bias). (2) Will $\hat{f}(\mathbf{Z})$ fit future observations well? To answer the questions, we might consider the **mean squared error** of our estimate $\hat{\boldsymbol{\beta}}$,

$$MSE(\hat{\boldsymbol{\beta}}) = E(\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|) = E[(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^\top(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})].$$

$MSE(\hat{\beta}) = \mathbb{E}(\hat{\beta}-\beta_0)^T(\hat{\beta}-\beta_0)$
$= \mathbb{E}(\hat{\beta}^T\hat{\beta}) - \beta_0^T\beta_0$

$\mathbb{E}(\hat{\beta}^T\hat{\beta}) = \mathbb{E}(Y^TX(X^TX)^{-1}(X^TX)^{-1}X^TY)$
$= tr(X(X^TX)^{-1}(X^TX)^{-1}X^TYY^T)$
$= tr(X(X^TX)^{-1}(X^TX)^{-1}X^T\sigma I^2)$
$+ tr(X(X^TX)^{-1}(X^TX)^{-1}X^T\ldots$
$= \sigma^2 tr(X^TX)^{-1} + \beta_0^T\beta_0$

For example, in the least-square estimation, we have

$$MSE(\hat{\boldsymbol{\beta}}^{ls}) = E(\|\hat{\boldsymbol{\beta}}^{ls} - \boldsymbol{\beta}\|) = tr(Var(\hat{\boldsymbol{\beta}}^{ls})) = \sigma^2 tr[(\mathbf{X}^\top\mathbf{X})^{-1}].$$

$\hat{f}(\mathbf{z})$ fits our data well doesn't mean that it will be a good fit to new data. In fact, suppose that we take new measurements $y_i'$ at the same $\mathbf{z}_i$'s:

$$(\mathbf{z}_1, y_1'), (\mathbf{z}_2, y_2'), \ldots, (\mathbf{z}_n, y_n').$$

If $\hat{f}(\cdot)$ is a good model, then $\hat{f}(\mathbf{z})$ should also be close to the new target $y_i'$. This is the notion of **prediction error** (PE). So good estimation should, on average, have small prediction error.

Consider the expected prediction error of an estimate $\hat{f}(\mathbf{z})$ at a particular point $\mathbf{z}_0$,

in paper, $\mathcal{Z}$ is usually a random observation.

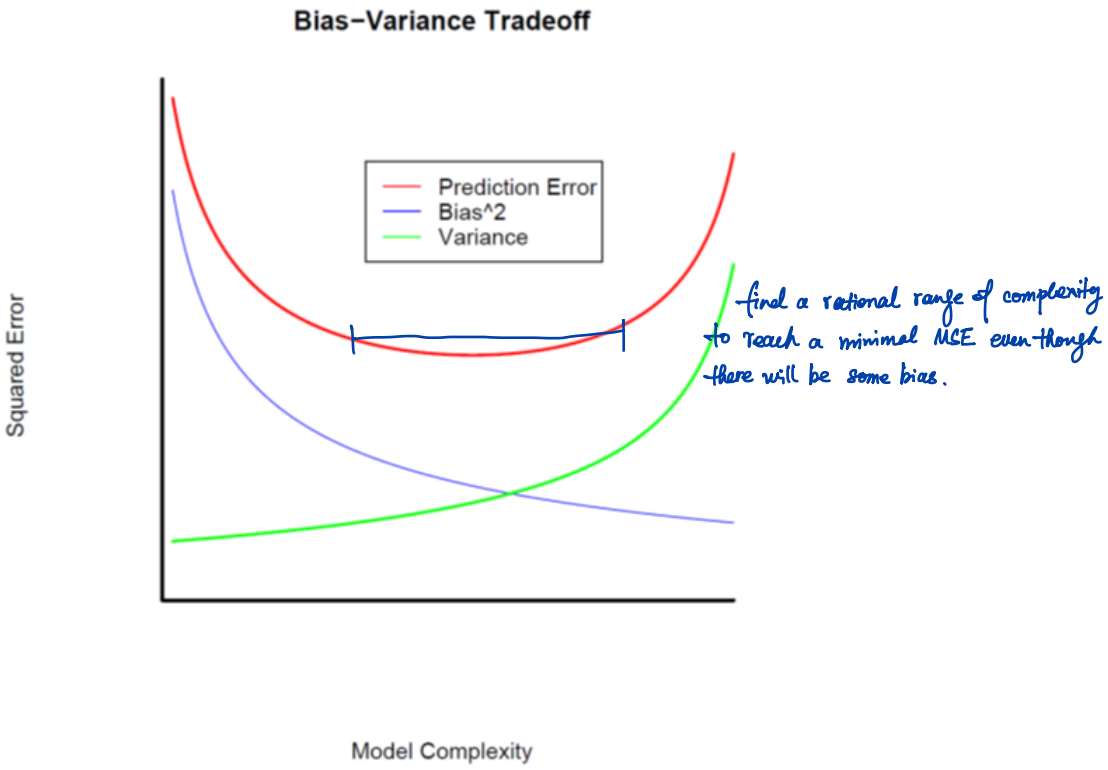
$$\begin{aligned}\mathrm{PE}(\mathbf{z}_0) &= E_{Y|\mathbf{Z}=\mathbf{z_0}}\left[(Y - \hat{f}(\mathbf{Z}))^2|\mathbf{Z} = \mathbf{z_0}\right] \\ &= \sigma_\varepsilon^2 + \mathrm{Bias}^2(\hat{f}(\mathbf{z}_0)) + \mathrm{Var}(\hat{f}(\mathbf{z}_0)) \\ &= \sigma_\varepsilon^2 + MSE(\hat{f}(\mathbf{z_0})). \qquad (1)\end{aligned}$$

$PE = \mathbb{E}_{Y,\mathcal{Z}}[(Y - \hat{f}(\mathcal{Z}))^2]$.

The relationship in (1) tells that the mean squared error is intimately related to prediction accuracy. In other words, the expected prediction error and mean squared error differ only by the constant $\sigma_\varepsilon^2$, representing the variance of the new observation $y_0$.

Such a decomposition in (1) is known as the **bias-variance trade-off**. As model becomes more complex (more terms included), local structure/curvature can be picked up. But coefficient estimates suffer from high variance as more terms are included in the model. So introducing a little bias in our estimate for $\beta$ might lead to a substantial decrease in variance, and hence to a substantial decrease in PE.

$PE(\mathcal{Z}_0) = \mathbb{E}[(Y - \hat{f}(\mathcal{Z}))^2 | \mathcal{Z} = \mathcal{Z}_0]$
$= \mathbb{E}[\varepsilon + f(\mathcal{Z}_0) - \hat{f}(\mathcal{Z}_0)]^2$
$= \mathbb{E}[\varepsilon^2 + \underbrace{(f(\mathcal{Z}_0) - \hat{f}(\mathcal{Z}_0))^2}_{MSE} + \underbrace{2\varepsilon(f(\mathcal{Z}_0) - \hat{f}(\mathcal{Z}_0))}_{0}]$ ($\mathbb{E}\varepsilon = 0$)
$= \sigma_\varepsilon^2 + \underbrace{[f(\mathcal{Z}_0) - \mathbb{E}\hat{f}(\mathcal{Z}_0)]^2}_{\text{bias}} + \underbrace{\mathbb{E}[\hat{f}(\mathcal{Z}_0) - \mathbb{E}\hat{f}(\mathcal{Z}_0)]^2}_{\text{Variance}}$ over the training process.


Bias–Variance Tradeoff
Prediction Error
Bias^2
Variance
Squared Error
Model Complexity
find a rational range of complexity to reach a minimal MSE even though there will be some bias.

## 5.2 Shrinkage methods

As introduced before, we have an input vector $Z = (Z_1, Z_2, \ldots, Z_p)$ and an output or response variable $Y$. The linear regression model has the form

$$Y = \beta_0 + \sum_{j=1}^{p} Z_j \beta_j + \varepsilon = \boldsymbol{Z\beta} + \boldsymbol{\varepsilon}.$$

As explained in section 5.1, there are two reasons why we are often not satisfied with the least squares estimates.

(1) The first is *prediction accuracy*: the least square estimates often have low bias but large variance. Prediction accuracy can sometimes be improved by shrinking or setting some coefficients to zero. By doing so, we sacrifice a little bit bias to reduce the variance of the predicted values, and hence may improve the overall prediction accuracy.

(2) The second reason is *interpretation*. With a large number of predictors, we often would like to determine a smaller subset that exhibit the strongest effects. In order to get the "big picture", we are willing to sacrifice some of the small details.

There are a number of approaches to variable subset selection with linear regression, e.g. best-subset selection, forward and backward-Stepwise selection, etc. Due to limitation of time, I will skip the introduction of the subset selection methods and focus on the shrinkage methods.

By retaining a subset of the predictors and discarding the rest, subset selection produces a model that is interpretable and has possibly lower prediction error than the full model. However, because it is a discrete process–variables are either retained or discarded–it often exhibits high variance, and so doesn't reduce the prediction error of the full model. Shrinkage methods are more continuous and stable, and don't suffer as much from high variability.

### 5.2.1 Ridge regression

Ridge regression shrinks the regression coefficients by imposing a penalty on their size. The ridge regression is to minimize


$$\sum_{i=1}^{n} (y_i - \mathbf{Z}_i \boldsymbol{\beta})^2 \quad \text{s.t.} \quad \sum_{j=1}^{p} \beta_j^2 \leqslant t$$

$$\text{i.e.} \quad (\boldsymbol{Y} - \boldsymbol{Z\beta})^\top (\boldsymbol{Y} - \boldsymbol{Z\beta}) \quad \text{s.t.} \quad \sum_{j=1}^{p} \beta_j^2 \leqslant t$$

By convention,

(1) $\mathbf{Z}$ is assumed to be standardized (mean=0, variance=1)

(2) $\boldsymbol{Y}$ is assumed to be centered

Equivalently, we can write the ridge coefficients as the minimizer of the following penalized residual sum of square (PRSS):

$$\hat{\beta}^{\text{ridge}} = \arg\min_{\beta} PRSS(\beta)_{\ell_2} = \arg\min_{\beta} \left\{ \sum_{i=1}^{n} (y_i - \mathbf{Z}_i\boldsymbol{\beta})^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \right\}$$

$$= \arg\min_{\beta} \|\boldsymbol{Y} - \boldsymbol{Z}\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_2^2 \tag{2}$$

Its solution may have smaller average PE than $\hat{\boldsymbol{\beta}}^{ls}$. Note that $PRSS(\boldsymbol{\beta})_{\ell_2}$ is convex, and hence has a unique solution. Taking derivatives, we obtain:

$$\frac{\partial PRSS(\boldsymbol{\beta})_{\ell_2}}{\partial \boldsymbol{\beta}} = -2\boldsymbol{Z}^\top(\boldsymbol{Y} - \boldsymbol{Z}\boldsymbol{\beta}) + 2\lambda\boldsymbol{\beta}$$

$$\hat{\boldsymbol{\beta}}_\lambda^{ridge} = (\boldsymbol{Z}^\top\boldsymbol{Z} + \lambda\mathbf{I}_p)^{-1}\boldsymbol{Z}^\top\boldsymbol{Y}$$

where $\boldsymbol{Z}$ is standardized and $\boldsymbol{Y}$ is centered. The solution is indexed by a tuning parameter $\lambda$, which makes the problem non-singular even if $\boldsymbol{Z}^\top\boldsymbol{Z}$ is not invertible. This was the original motivation for ridge regression (Hoerl and Kennard, 1970).

*Remark 1.* It is worthwhile to point out that, the ridge solutions are not equivariant under scaling of the inputs, and so this is why one normally standardizes the inputs before solving (2). In addition, notice that the intercept $\beta_0$ has been left out of the penalty term. Penalization of the intercept would make the procedure depend on the origin chosen for $Y$; that is, adding a constant $c$ to each of the targets $y_i$ would not simply result in a shift of the predictions by the same amount $c$. Henceforth we assume that this centering has been done, so that the input/design matrix has $p$ (rather than $p+1$) columns.

Note that the solution is indexed by $\lambda$, a complexity parameter that controls the amount of shrinkage: the larger the value of $\lambda$, the greater the amount of shrinkage:

(1) $\lambda$ controls the size of the coefficients

(2) $\lambda$ controls the amount of regulation

(3) As $\lambda \downarrow 0$, we obatin the least square solutions

(4) As $\lambda \uparrow \infty$, we have $\hat{\boldsymbol{\beta}}_\lambda^{ridge}$=0 (intercept-only model)

*Remark 2.* We next give a simple proof that $\hat{\boldsymbol{\beta}}_\lambda^{ridge}$ is biased. To this end, let $\boldsymbol{R} = \boldsymbol{Z}^\top \boldsymbol{Z}$ and assume $\boldsymbol{Z}$ has full column rank. Write

$$\begin{aligned}
\hat{\boldsymbol{\beta}}_\lambda^{ridge} &= (\boldsymbol{Z}^\top \boldsymbol{Z} + \lambda \mathbf{I}_p)^{-1} \boldsymbol{Z}^\top \boldsymbol{Y} \\
&= (\boldsymbol{R} + \lambda \mathbf{I}_p)^{-1} \boldsymbol{R} (\boldsymbol{R}^{-1} \boldsymbol{Z}^\top \boldsymbol{Y}) \\
&= [\boldsymbol{R}(\mathbf{I}_p + \lambda \boldsymbol{R}^{-1})]^{-1} \boldsymbol{R} \hat{\boldsymbol{\beta}}^{ls} \\
&= (\mathbf{I}_p + \lambda \boldsymbol{R}^{-1})^{-1} \hat{\boldsymbol{\beta}}^{ls}.
\end{aligned}$$

Thus,

$$\begin{aligned}
E(\hat{\boldsymbol{\beta}}_\lambda^{ridge}) &= E[(\mathbf{I}_p + \lambda \boldsymbol{R}^{-1})^{-1} \hat{\boldsymbol{\beta}}^{ls}] \\
&= (\mathbf{I}_p + \lambda \boldsymbol{R}^{-1})^{-1} \boldsymbol{\beta} \\
&\neq \boldsymbol{\beta} \qquad (\forall \boldsymbol{\beta} \neq 0, \lambda \neq 0)
\end{aligned}$$

Note that 1 is not an eigenvalue of the matrix $(\mathbf{I}_p + \lambda \boldsymbol{R}^{-1})^{-1}$, as $\lambda \neq 0$ and $\boldsymbol{R}$ is positive definite.

On the other hand, ridge regression can be translated into the least square problem alike with **the data augmentation approach.** The $\ell_2$-PRSS can be written as:

$$\begin{aligned}
PRSS(\boldsymbol{\beta})_{\ell_2} &= \sum_{i=1}^{n} (y_i - \mathbf{Z}_i \boldsymbol{\beta})^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \\
&= \sum_{i=1}^{n} (y_i - \mathbf{Z}_i \boldsymbol{\beta})^2 + \sum_{j=1}^{p} (0 - \sqrt{\lambda} \beta_j)^2.
\end{aligned}$$

Hence, the $\ell_2$ criterion can be recast as another least squares problem for another data set.

The $\ell_2$ criterion is simply the RSS for the augmented data set:

$$\boldsymbol{Z}_\lambda = \begin{pmatrix} z_{1,1} & z_{1,2} & z_{1,3} & \cdots & z_{1,p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ z_{n,1} & z_{n,2} & z_{n,3} & \cdots & z_{n,p} \\ \sqrt{\lambda} & 0 & 0 & \cdots & 0 \\ 0 & \sqrt{\lambda} & 0 & \cdots & 0 \\ 0 & 0 & \sqrt{\lambda} & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & \sqrt{\lambda} \end{pmatrix}, \quad \boldsymbol{Y}_\lambda = \begin{pmatrix} y_1 \\ \vdots \\ y_n \\ 0 \\ 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}$$

augmented data (added p data points).

So,

$$\boldsymbol{Z}_\lambda = \begin{pmatrix} \boldsymbol{Z} \\ \sqrt{\lambda}\mathbf{I}_p \end{pmatrix}, \quad Y_\lambda = \begin{pmatrix} Y \\ 0 \end{pmatrix}$$

The "least squares" solution for the augmented data set is:

$$\begin{aligned} (\boldsymbol{Z}_\lambda^\top \boldsymbol{Z}_\lambda)^{-1} \boldsymbol{Z}_\lambda^\top \boldsymbol{Y}_\lambda &= \left( (\boldsymbol{Z}^\top, \sqrt{\lambda}\mathbf{I}_p) \begin{pmatrix} \boldsymbol{Z} \\ \sqrt{\lambda}\mathbf{I}_p \end{pmatrix} \right)^{-1} (\boldsymbol{Z}^\top, \sqrt{\lambda}\mathbf{I}_p) \begin{pmatrix} \boldsymbol{Y} \\ 0 \end{pmatrix} \\ &= (\boldsymbol{Z}^\top \boldsymbol{Z} + \lambda \mathbf{I}_p)^{-1} \boldsymbol{Z}^\top \boldsymbol{Y} \end{aligned}$$

which is simply the ridge solution.

**Computing the ridge solutions via the SVD.**

When computing $\hat{\boldsymbol{\beta}}_\lambda^{ridge}$ numerically, matrix inversion should be avoided, since inverting $\boldsymbol{Z}^\top \boldsymbol{Z}$ can be computationally expensive: $O(p^3)$. Instead, the **singular value decomposition** is utilized, that is,

$$\boldsymbol{Z} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^\top,$$

$n \times p$ $p \times p$ $p \times p$.

Where

$U = (\mathbf{u_1}, \mathbf{u_2}, \ldots, \mathbf{u_p})$ is an $n \times p$ orthogonal matrix.

$D = \text{diag}(d_1, d_2, \ldots, d_p)$ is a $p \times p$ diagonal matrix consisting of the singular values $d_1 \geq d_2 \geq \cdots \geq d_p \geq 0$.

$$\hat{\beta}_{ls} = (Z^T Z)^{-1} Z^T Y$$

$$\begin{aligned} \hat{\beta}_r &= (Z^T Z + \lambda I_p)^{-1} Z^T Y \\ &= (I_p + \lambda (Z^T Z)^{-1})^{-1} (Z^T Z)^{-1} Z^T Y \\ &= (I_p + \lambda (Z^T Z)^{-1})^{-1} \hat{\beta}_{ls} \end{aligned}$$

$= (VDU^TUDV^T + \lambda I_p)^{-1} VDU^TY$

$= (VD^TV^T + V\lambda I_p V^T)^{-1} VDU^TY$

$= V\,diag(d_i^2+\lambda)^{-1}\, V^TV\, diag(d_i)\, U^TY$

$= V\,diag\frac{d_i}{d_i^2+\lambda}\, U^TY$

$V^\top = \begin{pmatrix} \mathbf{v_1}^\top \\ \mathbf{v_2}^\top \\ \vdots \\ \mathbf{v_p}^\top \end{pmatrix}$ is a $p \times p$ orthogonal matrix.

It can be shown that

$V\,diag(d_j^2)V^T + V\,diag(\lambda)V^T$

$$\begin{aligned}\hat{\boldsymbol{\beta}}_\lambda^{ridge} &= (\boldsymbol{Z}^\top\boldsymbol{Z} + \lambda\mathbf{I}_p)^{-1}\boldsymbol{Z}^\top\boldsymbol{Y} \\ &= \boldsymbol{V}\text{diag}\left(\frac{d_j}{d_j^2+\lambda}\right)\boldsymbol{U}^\top\boldsymbol{Y}\end{aligned}$$

by using the following eigen (or spectral) decomposition of $\boldsymbol{Z}^\top\boldsymbol{Z}$:

$$\begin{aligned}\boldsymbol{Z}^\top\boldsymbol{Z} &= (\boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^\top)^\top(\boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^\top) \\ &= \boldsymbol{V}\boldsymbol{D}^\top\boldsymbol{U}^\top\boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^\top \\ &= \boldsymbol{V}\boldsymbol{D}^\mathbf{2}\boldsymbol{V}^\top\end{aligned}$$

**Proof:**

A consequence is that:

$$\begin{aligned}\hat{\boldsymbol{Y}}^{ridge} &= \boldsymbol{Z}\hat{\boldsymbol{\beta}}_{\lambda}^{ridge} \\ &= \sum_{j=1}^{p}\left(\mathbf{u_j}\frac{d_j^2}{d_j^2+\lambda}\mathbf{u_j}^{\top}\right)\boldsymbol{Y}\end{aligned}$$

*Remark 3.* (Optional) Ridge regression also has a relationship with principal components analysis (PCA). The fact is that, the derived variable $\gamma_j = \boldsymbol{Z}\mathbf{v}_j = d_j\mathbf{u}_j$ is the $j$th principal component (PC) of $\boldsymbol{Z}$, hence, ridge regression projects $\boldsymbol{Y}$ onto these components with large $d_j$. Ridge regression shrinks the coefficients of low-variance (small singular value) components.

**Orthonormal $Z$ in ridge regression**

In the special case of an orthonormal design matrix, that is $\boldsymbol{Z}$ is orthonormal, then $\boldsymbol{Z}^{\top}\boldsymbol{Z} = I_p$. In this case, a couple of closed form properties exist. Let $\hat{\boldsymbol{\beta}}^{ls}$ denote the LS solution for our orthonormal $\boldsymbol{Z}$, then

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

and

$$\hat{\beta}_j^{ridge} = \frac{\hat{\beta}_j^{ols}}{1+\lambda}.$$

The optimal choice of $\lambda$ minimizing the expected prediction error is:

$$\lambda^* = \frac{p\sigma^2}{\sum_{j=1}^{p}\beta_j^2}$$

where $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_p)$ is the true coefficient vector.

This illustrates the essential "shrinkage" feature of ridge regression. Applying the ridge regression penalty has the effect of shrinking the estimates towards zero. The penalty introduces biase but reducing the variance of the estimate. Ridge estimator does not threshold, since the shrinkage is smooth (proportional to the original coefficient).

Furthermore, it can be checked that, the variance of the ridge regression estimate is

$$\mathrm{Var}(\hat{\boldsymbol{\beta}}^{ridge}) = \sigma^2\boldsymbol{W_\lambda}(\boldsymbol{Z}^{\top}\boldsymbol{Z})^{-1}\boldsymbol{W_\lambda} = \sigma^2(\boldsymbol{Z}^{\top}\boldsymbol{Z}+\lambda I_p)^{-1}\boldsymbol{Z}^{\top}\boldsymbol{Z}(\boldsymbol{Z}^{\top}\boldsymbol{Z}+\lambda I_p)^{-1},$$

where $\boldsymbol{W}_\lambda = (\boldsymbol{Z}^\top \boldsymbol{Z} + \lambda I_p)^{-1} \boldsymbol{Z}^\top \boldsymbol{Z}$. Recall that $\hat{\boldsymbol{\beta}}^{ridge} = W_\lambda \hat{\boldsymbol{\beta}}^{ls}$ The bias of the ridge regression estimate is

$$\text{bias}(\hat{\boldsymbol{\beta}}^{ridge}) = -\lambda \boldsymbol{W}_{\boldsymbol{\lambda}} (\boldsymbol{Z}^\top \boldsymbol{Z})^{-1} \boldsymbol{\beta} = -\lambda (\boldsymbol{Z}^\top \boldsymbol{Z} + \lambda I_p)^{-1} \beta$$

It can be shown that

(1) the total variance $\sum_{j=1}^{p} \text{Var}(\hat{\beta}_j^{ridge})$ is a monotone decreasing sequence with respect to $\lambda$.

(2) the total squared bias $\sum_{j=1}^{p} \text{Bias}^2(\hat{\beta}_j^{ridge})$ is a monotone increasing sequence with respect to $\lambda$.

**Existence Theorem**

There always exists a $\lambda$ such that the MSE of $\hat{\boldsymbol{\beta}}^{ridge}$ is less than the MSE of $\hat{\boldsymbol{\beta}}^{OLS}$.

*The proof of this existence theorem is left for students' take-home exercise.* A rather surprising result with somewhat radical implications: even if the model we fit is exactly correct and follows the exact distribution we specify, we can always obtain a better (in terms of MSE) estimator by shrinking towards zero.

**Smoother matrix and effective degrees of freedom**

A **Smoother matrix** $S$ is a linear operator satisfying:

$$\hat{\boldsymbol{Y}} = \boldsymbol{S}\boldsymbol{Y}$$

Example: In ordinary least squares, recall the hat matrix

$$\boldsymbol{H} = \boldsymbol{Z}(\boldsymbol{Z}^\top \boldsymbol{Z})^{-1} \boldsymbol{Z}^\top$$

Recall that $\text{rank}(\boldsymbol{Z}) = \text{tr}(H) = p$, which is how many degrees of freedom are used in the model. Analogously, we can define the **effective degrees of freedom** (or effective number of parameters) for a smoother to be:

$$\text{df}(\boldsymbol{S}) \triangleq \text{tr}(\boldsymbol{S})$$

In ridge regression, the fits are given by:

$$\hat{\boldsymbol{Y}} = \boldsymbol{Z}(\boldsymbol{Z}^\top \boldsymbol{Z} + \lambda \mathbf{I}_p)^{-1} \boldsymbol{Z}^\top \boldsymbol{Y}$$

So the smoother or "hat" matrix in ridge takes the form:

$$\boldsymbol{S}_\lambda = \boldsymbol{Z}(\boldsymbol{Z}^\top \boldsymbol{Z} + \lambda \mathbf{I}_p)^{-1} \boldsymbol{Z}^\top$$

So the effective degrees of freedom in ridge regression are given by:

$$\mathrm{df}(\lambda) = \mathrm{tr}(\boldsymbol{S}_\lambda) = \mathrm{tr}\left(\boldsymbol{Z}(\boldsymbol{Z}^\top \boldsymbol{Z} + \lambda \mathbf{I}_p)^{-1} \boldsymbol{Z}^\top\right) = \sum_{j=1}^{p} \frac{d_j^2}{d_j^2 + \lambda}$$

Note that $\mathrm{df}(\lambda)$ is monotone decreasing in $\lambda$.