Notes on Statistical and Machine Learning

Linear Methods for Regression

Chapter: *5* **Prepared by:** *Chenxi Zhou*

This note is prepared based on

- *Chapter 3, Linear Methods for Regression* in Hastie, Tibshirani, and Friedman (2009),
- *Chapter 5, Model Assessment and Selection in Multiple Regression* in Izenman (2009),
- *Chapter 2, The Lasso for Linear Models* in Hastie, Tibshirani, and Wainwright (2015), and
- *Chapter 4, Generalizations of the Lasso Penalty* in Hastie, Tibshirani, and Wainwright (2015).

I. Introduction

1. **Basic Assumption:** Suppose we have an input vector $X := (X_1, \cdots, X_p)^\top \in \mathcal{X} \subseteq \mathbb{R}^p$ and wish to predict a real-valued output Y. The *linear models* for regression assumes that

   (a) the regression function $\mathbb{E}[Y \mid X]$ is linear in the input vector $X := (X_1, \cdots, X_p)^\top$, or

   (b) the linear model is a *reasonable* approximation.

   Under this assumption, the linear regression model has the form

   $$Y = f(X) + \varepsilon, \tag{1}$$

   where

   $$f(X) := \beta_0 + \sum_{j=1}^{p} \beta_j X_j, \tag{2}$$

   ε is an unobservable random variable (the *error component*) with mean 0 and variance σ^2 (typically unknown), and $(\beta_0, \boldsymbol{\beta}^\top)^\top$ is the unknown coefficient vector to be estimated with $\boldsymbol{\beta} := (\beta_1, \cdots, \beta_p)^\top \in \mathbb{R}^p$.

2. **Advantages of Linear Regression Models:**

   - They are simple and provide an adequate and interpretable description of how inputs affect the output;

- Linear methods can be applied to transformations of inputs and allow more flexibility.

3. **Goals:** The goals are the following

   (a) estimation of the (true) value of the coefficient vector $(\beta_0, \boldsymbol{\beta}^\top)^\top$ and σ^2;

   (b) selection of a subset of useful independent variables to model Y in the case where too many independent variables are present;

   (c) prediction of future values of Y, given unseen values of the input variables;

   (d) quantification of the accuracy of the predictions.

II. Least Squares Regression Function and Linear Regression Model

1. **Setup:** Suppose that the input vector, $X := (X_1, \cdots, X_p)^\top \in \mathcal{X} \subseteq \mathbb{R}^p$, and the random output variable, $Y \in \mathcal{Y} \subseteq \mathbb{R}$, are both random and are jointly distributed according to the probability distribution $\mathbb{P}_{X,Y} : \mathcal{X} \times \mathcal{Y} \to [0, 1]$. Assume

   (a) $\mathbb{E}[X] = \boldsymbol{\mu}_X \in \mathbb{R}^p$,

   (b) $\mathbb{E}[Y] = \mu_Y \in \mathbb{R}$,

   (c) the covariance matrix of X is $\boldsymbol{\Sigma}_{XX} \in \mathbb{R}^{p\times p}$,

   (d) the variance of Y is $\Sigma_Y = \sigma_Y^2 > 0$, and

   (e) the covariance between X and Y is $\boldsymbol{\Sigma}_{XY} \in \mathbb{R}^p$.

2. **Loss Function, Risk, Bayes Rule, and Bayes Risk:**

   (a) *Loss Function:* Let $f : \mathcal{X} \to \mathcal{Y}$ be a function. Let $L(y, f(x))$ be a measurement of the accuracy which gives the loss incurred if y is predicted by $f(x)$. The function $L : \mathcal{Y} \times \mathcal{Y} \to [0, \infty)$ is called the *loss function.*

   (b) *Risk:* The expected loss is the *risk function* given by

   $$R(f) := \mathbb{E}\big[L(Y, f(X))\big], \tag{3}$$

   which measures the quality of f as a predictor. In (3), the expectation is taken under $\mathbb{P}_{X,Y}$, the joint distribution of X and Y.

   (c) *Bayes Rule and Bayes Risk:* The *Bayes rule* is the function f^* that minimizes $R(f)$, that is,

   $$f^* := \arg\min_f R(f),$$

   and the corresponding *Bayes risk* is $R(f^*)$.

3. **Squared Error Loss:** If we let

$$L(y, f(\mathbf{x})) = \frac{1}{2}(y - f(\mathbf{x}))^2,$$

this particular loss function is called the *squared error loss*.

4. **Derivation of Regression Function:** Suppose that the loss function is the *squared error loss*. The corresponding risk function is

$$R(f) = \mathbb{E}\left[\frac{1}{2}(Y - f(X))^2\right] = \mathbb{E}_X\left[\mathbb{E}_{Y|X}\left[\frac{1}{2}(Y - f(\mathbf{x}))^2 \,|\, X = \mathbf{x}\right]\right].$$

We can minimize $R(f)$ pointwise at each $X = \mathbf{x}$. Note that

$$Y - f(\mathbf{x}) = \big(Y - \mu(\mathbf{x})\big) + \big(\mu(\mathbf{x}) - f(\mathbf{x})\big),$$

where $\mu(\mathbf{x}) := \mathbb{E}_{Y|X}[Y \,|\, X = \mathbf{x}]$ is the mean of the conditional distribution of Y given $X = \mathbf{x}$ and is called the *regression function* of Y on X.

Then, note the following

$$\begin{aligned}\mathbb{E}_{Y|X}\big[(Y - f(\mathbf{x}))^2 \,|\, X = \mathbf{x}\big] &= \mathbb{E}_{Y|X}\big[(Y - \mu(\mathbf{x}) + \mu(\mathbf{x}) - f(\mathbf{x}))^2 \,|\, X = \mathbf{x}\big] \\ &= \mathbb{E}_{Y|X}[(Y - \mu(\mathbf{x}))^2 \,|\, X = \mathbf{x}] + (\mu(\mathbf{x}) - f(\mathbf{x}))^2,\end{aligned}$$

which is minimized with respect to f by

$$f^*(\mathbf{x}) = \mu(\mathbf{x}) = \mathbb{E}_{Y|X}[Y \,|\, X = \mathbf{x}].$$

Therefore, the pointwise minimum of $\mathbb{E}_{Y|X}[(Y - f(\mathbf{x}))^2 \,|\, X = \mathbf{x}]$ is

$$\mathbb{E}_{Y|X}[(Y - f^*(\mathbf{x}))^2 \,|\, X = \mathbf{x}] = \mathbb{E}_{Y|X}[(Y - \mu(\mathbf{x}))^2 \,|\, X = \mathbf{x}].$$

Finally, the Bayes risk is given by

$$R(f^*) = \frac{1}{2}\,\mathbb{E}_X\big[\mathbb{E}_{Y|X}[(Y - \mu(\mathbf{x}))^2 \,|\, X = \mathbf{x}]\big] = \frac{1}{2}\,\mathbb{E}\big[(Y - \mu(X))^2\big].$$

5. **Linear Regression Model:** Suppose that we are in the linear model framework so that the regression function μ is a linear combination of components of X, i.e.,

$$\mu(X) = \beta_0 + \sum_{j=1}^{p} \beta_j X_j = \beta_0 + \boldsymbol{\beta}^\top X, \tag{4}$$

where β_0 is the intercept, $\boldsymbol{\beta} := (\beta_1, \beta_2, \cdots, \beta_p)^\top \in \mathbb{R}^p$, and both β_0 and $\boldsymbol{\beta}$ are unknown. We choose β_0 and $\boldsymbol{\beta}$ to minimize the following risk function

$$\widetilde{R}(\beta_0, \boldsymbol{\beta}) := R(\mu) = \mathbb{E}\left[\frac{1}{2}(Y - \underbrace{(\beta_0 + \boldsymbol{\beta}^\top X)}_{=\mu(X)})^2\right].$$

Let

$$(\beta_0^*, \boldsymbol{\beta}^{*\top})^\top := \arg\min_{\beta_0, \boldsymbol{\beta}} \widetilde{R}(\beta_0, \boldsymbol{\beta}).$$

Differentiating $\widetilde{R}$ with respect to β_0 and $\boldsymbol{\beta}$ yield

$$\begin{aligned}
\frac{\partial \widetilde{R}}{\partial \beta_0} &= -\mathbb{E}[Y - \beta_0 - \boldsymbol{\beta}^\top X] \\
&= -\Big(\mu_Y - \beta_0 - \boldsymbol{\beta}^\top \boldsymbol{\mu}_X\Big), \\
\frac{\partial \widetilde{R}}{\partial \boldsymbol{\beta}} &= -\mathbb{E}[X^\top (Y - \beta_0 - \boldsymbol{\beta}^\top X)] \\
&= -\Big(\mathbb{E}[X^\top Y] - \beta_0 \mathbb{E}[X^\top] - \mathbb{E}[XX^\top]\boldsymbol{\beta}\Big) \\
&= -\Big(\big(\boldsymbol{\Sigma}_{XY} + \boldsymbol{\mu}_X^\top \mu_Y\big) - \beta_0 \boldsymbol{\mu}_X^\top - \big(\boldsymbol{\Sigma}_{XX} + \boldsymbol{\mu}_X \boldsymbol{\mu}_X^\top\big)\boldsymbol{\beta}\Big).
\end{aligned}$$

Setting $\frac{\partial \widetilde{R}}{\partial \beta_0} = 0$ and $\frac{\partial \widetilde{R}}{\partial \boldsymbol{\beta}} = \mathbf{0}_p$, we have

$$\beta_0^* = \mu_Y - \boldsymbol{\mu}_X^\top \boldsymbol{\beta}^*, \qquad \text{and} \qquad \boldsymbol{\beta}^* = \boldsymbol{\Sigma}_{XX}^{-1} \boldsymbol{\Sigma}_{XY}.$$

Finally, notice that $\widetilde{R}$ is a convex function of $(\beta_0, \boldsymbol{\beta}^\top)^\top$, any stationary point must be a global minimum. We conclude that, at $(\beta_0^*, \boldsymbol{\beta}^{*\top})^\top$, $\widetilde{R}$ achieves the global minimum.

III. Least Squares Regression with Data

1. **Data and Assumption:** We assume that we have a set of training data $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$, where $(\mathbf{x}_i, y_i) \overset{\text{i.i.d}}{\sim} \mathbb{P}_{X,Y}$ and $\mathbb{P}_{X,Y} : \mathcal{X} \times \mathcal{Y} \to [0,1]$ denotes the joint distribution of $(X^\top, Y)^\top$. Here, each $\mathbf{x}_i = (1, x_{i,1}, \cdots, x_{i,p})^\top \in \mathbb{R}^{p+1}$ is a vector of feature measurements for the i-th case.

   We assume that Y and X are related by the linear regression model (1). To reiterate, we assume

$$Y = \beta_0 + \sum_{j=1}^p \beta_j X_j + \varepsilon,$$

   where ε is an unobservable random variable (the *error component*) with mean 0 and variance σ^2.

2. **Parameter Estimation:** To estimate $\widetilde{\boldsymbol{\beta}} := (\beta_0, \boldsymbol{\beta}^\top)^\top \in \mathbb{R}^{p+1}$, we use the *least squares* method and pick the coefficients $\widetilde{\boldsymbol{\beta}}$ to minimize the *residual sum of squares*

$$\text{RSS}\,(\widetilde{\boldsymbol{\beta}}) := \frac{1}{2}\sum_{i=1}^n \big(y_i - f(\mathbf{x}_i)\big)^2 = \frac{1}{2}\sum_{i=1}^n \left(y_i - \beta_0 - \sum_{j=1}^p \beta_j x_{i,j}\right)^2. \tag{5}$$

3. **Matrix Form of** RSS**:** We write (5) in the following vector-matrix form

$$\mathrm{RSS}\,(\widetilde{\boldsymbol{\beta}}) = (\mathbf{Y} - \mathbf{X}\widetilde{\boldsymbol{\beta}})^\top (\mathbf{Y} - \mathbf{X}\widetilde{\boldsymbol{\beta}}),$$

where

- $\mathbf{X} \in \mathbb{R}^{n\times(p+1)}$ is the *design matrix* with each row being an input vector for each case $\mathbf{x}_i$ and each column corresponding to a variate (the first column is a vector of 1), and
- $\mathbf{Y} := (y_1, \cdots, y_n)^\top \in \mathbb{R}^n$ is the n-dimensional vector of outputs in the training set.

4. **Least Squares Estimator of $\widetilde{\boldsymbol{\beta}}$:** We minimize RSS over $\mathbb{R}^{p+1}$

$$\underset{\widetilde{\boldsymbol{\beta}}}{\text{minimize}}\ \mathrm{RSS}(\widetilde{\boldsymbol{\beta}}),$$

which is an unconstrained convex optimization problem.

Differentiating RSS with respect to $\widetilde{\boldsymbol{\beta}}$ and setting the derivative to $\mathbf{0}_{p+1}$ yield

$$\frac{\partial \mathrm{RSS}\,(\widetilde{\boldsymbol{\beta}})}{\partial \widetilde{\boldsymbol{\beta}}} = -\mathbf{X}^\top (\mathbf{Y} - \mathbf{X}\widetilde{\boldsymbol{\beta}}) \overset{\text{set}}{=} 0. \tag{6}$$

Assuming that the design matrix $\mathbf{X}$ is of full column rank, the least squares estimator for $\widetilde{\boldsymbol{\beta}}$ is

$$\widehat{\boldsymbol{\beta}} := (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{Y},$$

and this estimator is unique.

Taking the second derivative of RSS, we obtain

$$\frac{\partial^2 \mathrm{RSS}\,(\widetilde{\boldsymbol{\beta}})}{\partial \widetilde{\boldsymbol{\beta}} \partial \widetilde{\boldsymbol{\beta}}^\top} = \mathbf{X}^\top \mathbf{X} \succeq 0,$$

implying that RSS achieves the minimum at $\boldsymbol{\beta} = \widehat{\boldsymbol{\beta}}$.

5. **Fitted Value and Projection Matrix:** The *fitted values* at the training data are

$$\widehat{\mathbf{Y}} := (\hat{y}_1, \hat{y}_2, \cdots, \hat{y}_n)^\top = \mathbf{X}\widehat{\boldsymbol{\beta}} = \mathbf{X}(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{Y} = \mathbf{H}\mathbf{Y},$$

where the matrix $\mathbf{H} := \mathbf{X}(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top$ is called the *hat* or *projection* matrix.

6. **Residual Vector:** The residual vector, defined as

$$\hat{\boldsymbol{\varepsilon}} := \mathbf{Y} - \widehat{\mathbf{Y}} = (\mathbf{I}_n - \mathbf{H})\mathbf{Y},$$

is the least squares estimates of the unobserved errors $\boldsymbol{\varepsilon} := (\varepsilon_1, \cdots, \varepsilon_n)^\top$.

7. **Geometric Interpretation of Least Squares:** Let

$$\mathbf{X} = \begin{bmatrix} \mathbf{c}_0 & \mathbf{c}_1 & \cdots & \mathbf{c}_p \end{bmatrix} \in \mathbb{R}^{n\times(p+1)},$$

where $\mathbf{c}_j \in \mathbb{R}^n$ denotes the j-th column of $\mathbf{X}$, for all $j = 0, \cdots, p$, and $\mathbf{c}_0 := (1, \cdots, 1)^\top \in \mathbb{R}^n$. Then, these $(p+1)$ n-dimensional vectors span a subspace of $\mathbb{R}^n$, called the *column space* of $\mathbf{X}$. In addition, note that $\mathbf{X}\widetilde{\boldsymbol{\beta}}$ represents a linear combination of columns of $\mathbf{X}$.

In the least squares, we minimize $\text{RSS}(\widetilde{\boldsymbol{\beta}}) = \frac{1}{2}\|\mathbf{Y} - \mathbf{X}\widetilde{\boldsymbol{\beta}}\|_2^2$ by choosing $\widehat{\boldsymbol{\beta}}$ so that the *residual vector* $\mathbf{Y} - \widehat{\mathbf{Y}} = \mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}}$ is orthogonal to the column space of $\mathbf{X}$ (see (6)). Therefore, the resulting estimate $\widehat{\mathbf{Y}}$ is the *orthogonal projection* of $\mathbf{Y}$ onto this column subspace.

Remark. Since $\mathbf{X}\widehat{\boldsymbol{\beta}} = \mathbf{H}\mathbf{Y}$, the hat matrix $\mathbf{H}$ computes the orthogonal projection. Hence, it is called the *projection matrix*.

8. **Property of H:**

   (a) Both $\mathbf{H}$ and $\mathbf{I}_n - \mathbf{H}$ are symmetric and idempotent;

   (b) $\mathbf{H}(\mathbf{I}_n - \mathbf{H}) = \mathbf{0}_{n\times n}$; and

   (c) $\mathbf{H}\mathbf{X} = \mathbf{X}$ and $(\mathbf{I}_n - \mathbf{H})\mathbf{X} = \mathbf{0}_{n\times n}$.

9. **Variance of $\widehat{\boldsymbol{\beta}}$:**

   (a) *Assumptions:* Assume that

      i. the observations y_i's are *uncorrelated* and have *constant variance* σ^2, and

      ii. the $\mathbf{x}_i$'s are fixed (not random).

   (b) *Variance of $\widehat{\boldsymbol{\beta}}$:* Since $\widehat{\boldsymbol{\beta}} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{Y}$, then

$$\begin{aligned}
\text{Var}\big[\widehat{\boldsymbol{\beta}}\big] &= \text{Var}\big[(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{Y}\big] \\
&= (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top \,\text{Var}[\mathbf{Y}]\mathbf{X}^\top(\mathbf{X}^\top\mathbf{X})^{-1} \\
&= (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top(\sigma^2\mathbf{I}_n)\mathbf{X}^\top(\mathbf{X}^\top\mathbf{X})^{-1} \\
&= \sigma^2(\mathbf{X}^\top\mathbf{X})^{-1}(\mathbf{X}^\top\mathbf{X}^\top)(\mathbf{X}^\top\mathbf{X})^{-1} \\
&= \sigma^2(\mathbf{X}^\top\mathbf{X})^{-1}.
\end{aligned}$$

10. **Estimator of σ^2:** Typically, the true value of σ^2 is unknown. We can estimate it by

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

which is an *unbiased* estimator of σ^2; that is, $\mathbb{E}[\hat{\sigma}^2] = \sigma^2$.

11. **Additional Assumptions:** We require the following assumptions to make inferences about the linear regression model (1):

(a) The linear regression model (1) is the *correct* model for the mean; that is, the conditional expectation of Y is linear in $X_1, \cdots, X_p$;

(b) The deviation of Y around its conditional expectation is *additive* and follows a *normal distribution* with mean 0 and variance σ^2.

Mathematically, we obtain the following model

$$Y = \mathbb{E}[Y \mid X_1, \cdots, X_p] + \varepsilon = \beta_0 + \sum_{i=1}^{p} \beta_j X_j + \varepsilon, \tag{7}$$

where $\varepsilon \sim \text{Normal}(0, \sigma^2)$.

12. Statistical Properties of $\widehat{\boldsymbol{\beta}}$ and $\hat{\sigma}^2$:

(a) $\widehat{\boldsymbol{\beta}}$ follows a multivariate normal distribution with mean $\widetilde{\boldsymbol{\beta}}$ and variance $\sigma^2(\mathbf{X}^\top\mathbf{X})^{-1}$, that is,

$$\widehat{\boldsymbol{\beta}} \sim \text{Normal}\big(\widetilde{\boldsymbol{\beta}}, (\mathbf{X}^\top\mathbf{X})^{-1}\sigma^2\big).$$

(b) The unbiased estimator $\hat{\sigma}^2$ follows a scaled chi-square distribution with the degrees of freedom $n-p-1$, that is,

$$(n-p-1)\hat{\sigma}^2 \sim \sigma^2\chi^2_{n-p-1}.$$

(c) $\widehat{\boldsymbol{\beta}}$ and $\hat{\sigma}^2$ are statistically independent.

13. Hypothesis Testing of a Single Coefficient: To test

$$H_0 : \beta_j = 0 \qquad \text{vs.} \qquad H_1 : \beta_j \neq 0,$$

for $j = 1, \cdots, p$, we use the test statistic

$$t := \frac{\hat{\beta}_j}{\hat{\sigma}\sqrt{\big[(\mathbf{X}^\top\mathbf{X})^{-1}\big]_{j,j}}},$$

where $\big[(\mathbf{X}^\top\mathbf{X})^{-1}\big]_{j,j}$ denotes the (j,j)-th element of the matrix $(\mathbf{X}^\top\mathbf{X})^{-1}$.

Under the null hypothesis that $\beta_j = 0$, t follows a t-distribution with the degrees of freedom $n-p-1$.

14. Confidence Interval of a Single Coefficient: The pointwise $1-2\alpha$ confidence interval for β_j is

$$\left(\hat{\beta}_j - t_{n-p-1,1-\frac{\alpha}{2}} \times \hat{\sigma}\sqrt{\big[(\mathbf{X}^\top\mathbf{X})^{-1}\big]_{j,j}},\ \hat{\beta}_j + t_{n-p-1,1-\frac{\alpha}{2}} \times \hat{\sigma}\sqrt{\big[(\mathbf{X}^\top\mathbf{X})^{-1}\big]_{j,j}}\right),$$

where $t_{n-p-1,1-\frac{\alpha}{2}}$ is the $(1-\frac{\alpha}{2})$-th percentile of the t-distribution with the degrees of freedom $n-p-1$.

15. **Hypothesis Testing of a Group of Coefficients or Nested Linear Models:** Supposing we want to test whether a group of coefficients is the zero vector or not or compare two *nested* linear models, we use the following test statistic

$$F = \frac{(\mathrm{RSS}_0 - \mathrm{RSS}_1)/(p_1 - p_0)}{\mathrm{RSS}_1/(n - p - 1)}, \tag{8}$$

where RSS_1 is the residual sum of squares for the least squares fit of the *bigger* model with $p_1 + 1$ parameters, and RSS_0 is the residual sum of squares for the nested *smaller* model with $p_0 + 1$ parameters, and there are $p_1 - p_0$ parameters constrained to be zero. Here, $p_1 > p_0$.

The F-statistic measures the change in the residual sum of squares per additional parameter in the bigger model and is normalized by an estimate of σ^2 under the bigger model. Note that the denominator is the unbiased estimator of σ^2 under the bigger model.

Under the Gaussian assumption and the null hypothesis that the smaller model is correct, the F statistic (8) follows an $F_{p_1-p_0,n-p_1-1}$ distribution.

16. **Confidence Set for $\widetilde{\boldsymbol{\beta}}$:** An approximate confidence set for the entire parameter vector $\widetilde{\boldsymbol{\beta}}$ is

$$C_{\widetilde{\boldsymbol{\beta}}} := \left\{ \widetilde{\boldsymbol{\beta}} \;\middle|\; (\widehat{\boldsymbol{\beta}} - \widetilde{\boldsymbol{\beta}})^\top \mathbf{X}^\top \mathbf{X} (\widehat{\boldsymbol{\beta}} - \widetilde{\boldsymbol{\beta}}) \le \hat{\sigma}^2 (p+1) F_{p+1,n-p-1,1-\alpha} \right\}, \tag{9}$$

where $F_{p+1,n-p-1,1-\alpha}$ is the $(1-\alpha)$-percentile of the F distribution with $(p+1, n-p-1)$ degrees of freedom.

Geometrically, the confidence set (9) is an $(p+1)$-dimensional ellipsoid with center $\widehat{\boldsymbol{\beta}}$ and orientation controlled by the matrix $\mathbf{X}^\top \mathbf{X}$.

17. **Statistical Properties of Fitted Value Vector $\widehat{\mathbf{Y}}$:** Under the model assumption (7), we have

$$\mathbb{E}[\widehat{\mathbf{Y}}] = \mathbf{Y}, \qquad \text{and} \qquad \mathrm{Cov}[\widehat{\mathbf{Y}}] = \sigma^2 \mathbf{H}.$$

18. **More Properties of H:**

    (a) The (i, j)-th component $h_{i,j}$ of $\mathbf{H}$ is the amount of leverage (or impact) that the observed value of y_j exerts on the fitted value $\hat{y}_i$. The hat matrix $\mathbf{H}$ is, therefore, used to identify *high-leverage points*.

    (b) The diagonal components $h_{i,i}$ satisfy $0 \le h_{i,i} \le 1$;

    (c) Under the assumption that $\mathbf{X}$ is of full column rank, the sum of the diagonal elements of $\mathbf{H}$ is $p + 1$; the average leverage magnitude is $(p + 1)/n$;

    (d) One way to define high-leverage points is those points having

$$h_{i,i} > \frac{2(p+1)}{n}.$$

19. Statistical Properties of Residual Vector $\hat{\boldsymbol{\varepsilon}}$: Under the model assumption (7), we have

$$\mathbb{E}[\hat{\boldsymbol{\varepsilon}}] = \mathbf{0}_n, \qquad \text{and} \qquad \mathrm{Cov}[\hat{\boldsymbol{\varepsilon}}] = (\mathbf{I}_n - \mathbf{H})\sigma^2.$$

Hence, $\mathrm{Var}[\hat{\varepsilon}_i] = (1 - h_{i,i})\sigma^2$, where $\hat{\varepsilon}_i := y_i - \hat{y}_i$ is the i-th residual and $h_{i,i}$ is the i-th diagonal element of $\mathbf{H}$.

20. Internally Studentized Residuals: The i-th *internally Studentized residual*, denoted by $\tilde{\varepsilon}_i$, is obtained by dividing the residual by its variance, i.e.,

$$\tilde{\varepsilon}_i = \frac{\hat{\varepsilon}_i}{\hat{\sigma}\sqrt{1 - h_{i,i}}},$$

where $h_{i,i}$ is the i-th diagonal element of $\mathbf{H}$.

21. Diagnostics: Because the fitted value vector $\widehat{\mathbf{Y}} = \mathbf{HY}$ and the (raw) residual vector $\hat{\boldsymbol{\varepsilon}} = (\mathbf{I}_n - \mathbf{H})\mathbf{Y}$ have zero covariance and, hence, are uncorrelated, it follows that the regression of $\mathbf{Y}$ on $\hat{\boldsymbol{\varepsilon}}$ has zero slope.

If the multiple regression model is correct, then a scatterplot of residuals (or internally Studentized residuals) against fitted values should show no discernible pattern (i.e., a slope of approximately zero).

22. Analysis of Variance (ANOVA): Let $\bar{y} := \frac{1}{n}\sum_{i=1}^n y_i$. Note the following identity

$$y_i - \bar{y} = (y_i - \hat{y}_i) + (\hat{y}_i - \bar{y}).$$

Then, we have

$$\begin{aligned}
\sum_{i=1}^n (y_i - \bar{y})^2 &= \sum_{i=1}^n \big[(y_i - \hat{y}_i) + (\hat{y}_i - \bar{y})\big]^2 \\
&= \sum_{i=1}^n \big[(y_i - \hat{y}_i)^2 + 2(y_i - \hat{y}_i)(\hat{y}_i - \bar{y}) + (\hat{y}_i - \bar{y})^2\big] \\
&= \sum_{i=1}^n (y_i - \hat{y}_i)^2 + 2\sum_{i=1}^n (y_i - \hat{y}_i)(\hat{y}_i - \bar{y}) + \sum_{i=1}^n (\hat{y}_i - \bar{y})^2.
\end{aligned}$$

We show $\sum_{i=1}^n (y_i - \hat{y}_i)(\hat{y}_i - \bar{y}) = 0$. Note the following

$$\begin{aligned}
\sum_{i=1}^n (y_i - \hat{y}_i)(\hat{y}_i - \bar{y}) &= \sum_{i=1}^n (y_i - \hat{y}_i)\hat{y}_i - \sum_{i=1}^n (y_i - \hat{y}_i)\bar{y} \\
&= (\mathbf{X}\widehat{\boldsymbol{\beta}})^\top (\mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}}) + (\mathbf{1}^\top \mathbf{Y})\mathbf{1}^\top (\mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}}) \\
&= \widehat{\boldsymbol{\beta}}^\top [\mathbf{X}^\top (\mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}})] + (\mathbf{1}_n^\top \mathbf{Y})[\mathbf{1}_n^\top (\mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}})] \\
&= 0,
\end{aligned}$$

where $\mathbf{X}^\top(\mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}}) = \mathbf{0}_{p+1}$ is due to the normal equation (6), and $\mathbf{1}_n^\top(\mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}}) = 0$ since the first column of $\mathbf{X}$ is $\mathbf{1}_n$ and the residual vector $\mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}}$ must be orthogonal to it, again by the normal equation (6).

As a consequence, we obtain the decomposition

$$\text{ToSS} = \text{RegSS} + \text{RSS},$$

where

$$\text{ToSS} := \sum_{i=1}^{n}(y_i - \bar{y})^2,$$

is the *total sum of squares*,

$$\text{RegSS} := \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2,$$

is the *regression sum of squares*, and

$$\text{RSS} := \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

is the *residual sum of squares* as before.

We have the following ANOVA table

Source of Variation	df	Sum of Squares
Regression on $X_1, \cdots, X_p$	p	$\text{RegSS} = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$
Residuals	$n-p-1$	$\text{RSS} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$
Total	$n-1$	$\text{ToSS} = \sum_{i=1}^{n}(y_i - \bar{y})^2$

23. R^2 and Adjusted-R^2: The *squared multiple correlation coefficient* is defined to be

$$R^2 := 1 - \frac{\text{RSS}}{\text{ToSS}},$$

and is always between 0 and 1, and is used to measured the proportion of the total variation in Y that can be explained by a linear regression model on $X_1, \cdots, X_p$.

Adding more variables to the linear model will *always* increase the R^2 value, even if the covariates are useless predictors of Y individually. To guard against this, the adjusted R^2 is given by

$$\text{adj}{-}R^2 = 1 - \frac{\text{RSS}/(n-p-1)}{\text{ToSS}/(n-1)}.$$

The adjusted R^2 can become smaller if the variable added does *not* add "greatly" to the model.

Remark. A small R^2 does *not* always imply a bad fit to the model.

24. Gauss-Markov Theorem: Consider to estimate a linear combination of the coefficient vector $\theta = \mathbf{a}^\top \widetilde{\boldsymbol{\beta}}$, where the vector $\mathbf{a} \in \mathbb{R}^{p+1}$ is fixed. Under the assumption that $\mathbf{X}$ is fixed, the least squares estimator of θ is

$$\hat{\theta} := \mathbf{a}^\top \widehat{\boldsymbol{\beta}} = \mathbf{a}^\top (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{Y},$$

which is a linear function of the response vector $\mathbf{Y}$. Then, we have the following

(a) $\mathbf{a}^\top \widehat{\boldsymbol{\beta}}$ is unbiased, since

$$\mathbb{E}\big[\mathbf{a}^\top \widehat{\boldsymbol{\beta}}\big] = \mathbf{a}^\top (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \, \mathbb{E}[\mathbf{Y}] = \mathbf{a}^\top (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{X} \widetilde{\boldsymbol{\beta}} = \mathbf{a}^\top \widetilde{\boldsymbol{\beta}};$$

(b) if $\tilde{\theta} = \tilde{\mathbf{c}}^\top \mathbf{Y}$ is any other linear estimator that is unbiased for $\mathbf{a}^\top \widetilde{\boldsymbol{\beta}}$, then

$$\mathrm{Var}\big[\mathbf{a}^\top \widehat{\boldsymbol{\beta}}\big] \le \mathrm{Var}\big[\tilde{\mathbf{c}}^\top \mathbf{Y}\big].$$

To see this, first note, since $\mathbf{Y} = \mathbf{X}\widetilde{\boldsymbol{\beta}} + \boldsymbol{\varepsilon}$, where $\boldsymbol{\varepsilon} := (\varepsilon_1, \varepsilon_2, \cdots, \varepsilon_n)^\top$, we have

$$\mathrm{Var}\big[\tilde{\mathbf{c}}^\top \mathbf{Y}\big] = \mathbb{E}\big[\big(\tilde{\mathbf{c}}^\top \mathbf{Y} - \mathbb{E}[\tilde{\mathbf{c}}^\top \mathbf{Y}]\big)^2\big] = \mathbb{E}\big[(\tilde{\mathbf{c}}^\top \boldsymbol{\varepsilon})^2\big] = \sigma^2 \mathbf{c}^\top \mathbf{c}.$$

Then, we solve the following minimization problem

$$\underset{\mathbf{c}}{\text{minimize}} \ \frac{1}{2}\sigma^2 \mathbf{c}^\top \mathbf{c} \qquad \text{subject to } \mathbf{c}^\top \mathbf{X} \widetilde{\boldsymbol{\beta}} = \mathbf{a}^\top \widetilde{\boldsymbol{\beta}},$$

which is equivalent to the following problem

$$\underset{\mathbf{c}}{\text{minimize}} \ \frac{1}{2}\mathbf{c}^\top \mathbf{c} \qquad \text{subject to } \mathbf{X}^\top \mathbf{c} = \mathbf{a}.$$

Note that this is a convex minimization problem under the equality constraint, and any stationary point must be the global minimizer.
The Lagrangian function is

$$\mathcal{L}(\mathbf{c}, \boldsymbol{\lambda}) := \frac{1}{2}\mathbf{c}^\top \mathbf{c} + \boldsymbol{\lambda}^\top (\mathbf{X}^\top \mathbf{c} - \mathbf{a}).$$

Differentiating with respect to $\mathbf{c}$ and setting the derivative to $\mathbf{0}$ yield

$$\frac{\partial \mathcal{L}}{\partial \mathbf{c}}(\mathbf{c}, \boldsymbol{\lambda}) = \mathbf{c} - \mathbf{X}\boldsymbol{\lambda} \overset{\text{set}}{=} \mathbf{0},$$

i.e., $\mathbf{c} = \mathbf{X}\boldsymbol{\lambda}$. Plugging into the equality constraint yields

$$\mathbf{X}^\top \mathbf{X} \boldsymbol{\lambda} - \mathbf{a} = \mathbf{0},$$

yielding $\boldsymbol{\lambda} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{a}$. It follows the optimal $\mathbf{c}$ is $\mathbf{c}^* := \mathbf{X}(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{a}$. Then, $\mathbf{a}^\top \widehat{\boldsymbol{\beta}} = \mathbf{c}^{*\top} \mathbf{Y}$. The desired result follows.

25. Mean Squared Error: The *mean squared error* of an estimator $\tilde{\theta}$ of θ is

$$\mathrm{MSE}(\tilde{\theta}) := \mathbb{E}\big[(\tilde{\theta}-\theta)^2\big] = \mathrm{Var}[\tilde{\theta}] + \big(\mathbb{E}[\tilde{\theta}]-\theta\big)^2.$$

Remark 1. If $\tilde{\theta}$ is an unbiased estimator of θ, i.e., $\mathbb{E}[\tilde{\theta}] = \theta$, we then have

$$\mathrm{MSE}(\tilde{\theta}) = \mathrm{Var}[\tilde{\theta}].$$

Remark 2. Due to the Gauss-Markov Theorem, the least squares estimator has the smallest mean squared error of all linear estimators with no bias. However, there may exist a *biased estimator* with smaller mean squared error, as they trade a little bias for a larger reduction in variance.

26. Expected Prediction Error: Consider the prediction of the new response at $\mathbf{x}_0$,

$$Y_0 = f(\mathbf{x}_0) + \varepsilon_0.$$

The expected prediction error of an estimate $\tilde{f}(\mathbf{x}_0) = \mathbf{x}_0^\top \widetilde{\boldsymbol{\beta}}$ is

$$\mathbb{E}\big[(Y_0 - \tilde{f}(\mathbf{x}_0))^2\big] = \sigma^2 + \mathbb{E}\big[(\mathbf{x}_0^\top\widetilde{\boldsymbol{\beta}} - f(\mathbf{x}_0))^2\big] = \sigma^2 + \mathrm{MSE}(\tilde{f}(\mathbf{x}_0)).$$

Thus, the expected prediction error and the mean squared error differ only by the constant σ^2, representing the variance of the new observation Y_0.

27. Rank Deficiency Case:

(a) *Definition:* The columns of $\mathbf{X}$ are *not* linearly independent so that $\mathbf{X}$ is *not* of full rank. Then, $\mathbf{X}^\top\mathbf{X}$ is singular.

(b) *When This Can Happen?* This may happen when

i. $\mathbf{X}$ is ill-conditioned, or

ii. the columns of $\mathbf{X}$ are collinear, or

iii. there are more variables than observations ($p > n$).

(c) *Impacts:* In this case,

i. the least squares coefficient vector $\widehat{\boldsymbol{\beta}}$ is *not* uniquely determined, but

ii. the fitted value vector $\widehat{\mathbf{Y}}$ is still the projection of $\mathbf{Y}$ onto the column space of $\mathbf{X}$ and is uniquely determined.

(d) *Condition Number:* We can measure the ill-condition of $\mathbf{X}$ by looking at the *condition number*,

$$\kappa(\mathbf{X}) := \frac{\sigma_1}{\sigma_p},$$

where σ_1 is the largest singular value of $\mathbf{X}$ and σ_p is the smallest singular value of $\mathbf{X}$. If $\kappa(\mathbf{X})$ is large, $\mathbf{X}$ is said to be *ill-conditioned*.

Remark. When exact collinearity occurs, $\kappa(\mathbf{X}) = \infty$, since $\sigma_p = 0$.

(e) *Collinearity Indices and Variance Inflation Factor (VIF):* We can use the *collinearity indices* to measure the degree of collinearity among variables. The *j-th collinearity index* is defined to be

$$\kappa_j := \sqrt{\text{VIF}_j}, \qquad \text{for all } j = 1, \cdots, p,$$

where

$$\text{VIF}_j := \frac{1}{1 - R_j^2} \tag{10}$$

is the j-th variance inflation factor and R_j^2 is the squared multiple correlation coefficient of the j-th column of $\mathbf{X}$ on the other $p-1$ columns of $\mathbf{X}$.

Remark 1. Large values of VIF_j (typically, $\text{VIF}_j > 10$) imply that R_j^2 is close to unity, which in turn suggests near collinearity may be present.

Remark 2. The collinearity indices have value at least one and are invariant under scale changes of the columns of $\mathbf{X}$.

(f) *How to deal with this rank-deficiency issue?*

i. Rank deficiency may occur when some columns are linearly dependent. We can drop redundant columns in $\mathbf{X}$;

ii. We can fit *regularized* linear regression model.

28. Univariate Linear Regression Without Intercept: Consider a univariate linear model ($p = 1$) with no intercept,

$$Y = X\beta + \varepsilon.$$

Let $\mathbf{X} = (x_1, \cdots, x_n)^\top \in \mathbb{R}^n$ be the design matrix and $\mathbf{Y} = (y_1, \cdots, y_n)^\top \in \mathbb{R}^n$ be the response vector. The least squares estimator of β and the residuals are

$$\hat{\beta} = \frac{\langle \mathbf{X}, \mathbf{Y} \rangle}{\|\mathbf{X}\|_2^2} = \frac{\sum_{i=1}^n x_i y_i}{\sum_{i=1}^n x_i^2}, \qquad \text{and} \qquad r_i = y_i - x_i\hat{\beta},$$

where

$$\langle \mathbf{X}, \mathbf{Y} \rangle = \sum_{i=1}^n x_i y_i, \qquad \text{and} \qquad \langle \mathbf{X}, \mathbf{X} \rangle = \|\mathbf{X}\|_2^2 = \sum_{i=1}^n x_i^2.$$

29. From Univariate Linear Regression to Multiple Linear Regression With Orthogonal Columns of X and Without Intercept: Suppose that we have a multiple linear regression model with *no intercept* and that the columns of the design matrix $\mathbf{X}$, denoted by $\mathbf{c}_1, \cdots, \mathbf{c}_p$, are orthogonal, i.e., $\langle \mathbf{c}_i, \mathbf{c}_j \rangle = 0$ for all $i \neq j$. Then, the least squares estimator of $\boldsymbol{\beta}$ is

$$\begin{aligned}\widehat{\boldsymbol{\beta}} &= (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}\mathbf{Y} \\ &= \text{diag}\Big(\langle \mathbf{c}_1, \mathbf{c}_1 \rangle, \langle \mathbf{c}_2, \mathbf{c}_2 \rangle, \cdots, \langle \mathbf{c}_p, \mathbf{c}_p \rangle \Big)^{-1} \mathbf{X}\mathbf{Y},\end{aligned}$$

and hence, for all $j = 1, \cdots, p$

$$\hat{\beta}_j = \frac{\langle \mathbf{c}_j, \mathbf{Y} \rangle}{\langle \mathbf{c}_j, \mathbf{c}_j \rangle},$$

which is the univariate estimators. In other words, when the columns in the design matrix are orthogonal, they have *no effect* on each other's parameter estimators in the model.

30. Univariate Linear Regression With an Intercept: Consider a univariate linear model ($p = 1$) with an intercept,

$$Y = \beta_0 + X\beta_1 + \varepsilon.$$

The least squares estimator of β_1 is

$$\hat{\beta}_1 = \frac{\langle \mathbf{X} - \bar{x}\mathbf{1}_n, \mathbf{Y} \rangle}{\langle \mathbf{X} - \bar{x}\mathbf{1}_n, \mathbf{X} - \bar{x}\mathbf{1}_n \rangle},$$

where $\mathbf{X} = (x_1, \cdots, x_n)^\top \in \mathbb{R}^n$, $\mathbf{Y} = (y_1, \cdots, y_n)^\top \in \mathbb{R}^n$, $\bar{x} = \sum_{i=1}^n x_i$ and $\mathbf{1}_n = (1, \cdots, 1)^\top \in \mathbb{R}^n$. One can view this derivation of $\hat{\beta}_1$ as of two steps:

Step 1: regress $\mathbf{X}$ on $\mathbf{1}_n$ and produce the residual $\mathbf{Z} := \mathbf{X} - \bar{x}\mathbf{1}_n$;

Step 2: regress $\mathbf{Y}$ on the residual $\mathbf{Z}$ to produce the estimate $\hat{\beta}_1$.

Remark 1. In Step 1, we orthogonalized $\mathbf{X}$ with respect to $\mathbf{c}_0 = \mathbf{1}_n$. The orthogonalization does *not* change the subspace spanned by columns but just produces an orthogonal basis for representing this subspace.

31. From Univariate Linear Regression to Multiple Linear Regression With an Intercept: Suppose we have an intercept and p covariates. By mimicking the procedures described above, we have the following algorithm, called the *regression by successive orthogonalization* or the *Gram-Schmidt procedure*.

Algorithm 1 Regression by Successive Orthogonalization

1: Initialize $\mathbf{z}_0 = \mathbf{c}_0 = \mathbf{1}_n$.

2: For $j = 1, \cdots, p$, regress $\mathbf{c}_j$ on $\mathbf{z}_0, \mathbf{z}_1, \cdots, \mathbf{z}_{j-1}$ to produce coefficients

$$\hat{\gamma}_{\ell,j} = \frac{\langle \mathbf{z}_\ell, \mathbf{c}_j \rangle}{\langle \mathbf{z}_\ell, \mathbf{z}_\ell \rangle}, \qquad \text{for all } \ell = 0, \cdots, j-1,$$

and the residual vector

$$\mathbf{z}_j = \mathbf{c}_j - \sum_{k=0}^{j-1} \hat{\gamma}_{k,j} \mathbf{z}_k.$$

3: Regress $\mathbf{Y}$ on the residual $\mathbf{z}_p$ to obtain the estimate

$$\hat{\beta}_p = \frac{\langle \mathbf{z}_p, \mathbf{Y} \rangle}{\langle \mathbf{z}_p, \mathbf{z}_p \rangle}. \tag{11}$$

Remark 1. Each $\mathbf{c}_j$ is a linear combination of $\mathbf{z}_0, \mathbf{z}_1, \cdots, \mathbf{z}_j$. Since the $\mathbf{z}_j$'s are all orthogonal, they form an *orthogonal basis* for the column space of $\mathbf{X}$. Hence, the least squares projection onto this column subspace is $\widehat{\mathbf{Y}}$.

Remark 2. The j-th multiple regression coefficient $\hat{\beta}_j$ represents the additional contribution of $\mathbf{c}_j$ to $\mathbf{Y}$, *after* $\mathbf{c}_j$ has been adjusted for $\mathbf{c}_0, \mathbf{c}_1, \cdots, \mathbf{c}_{j-1}, \mathbf{c}_{j+1}, \cdots, \mathbf{c}_p$.

Remark 3. If $\mathbf{c}_p$ is highly correlated with some other $\mathbf{c}_k$'s, the residual vector $\mathbf{z}_p$ will be very small and the estimator $\hat{\beta}_p$ will be very *unstable*. From (11), the variance of $\hat{\beta}_p$ can be obtained as

$$\mathrm{Var}\big[\hat{\beta}_p\big] = \frac{\sigma^2}{\|\mathbf{z}_p\|_2^2};$$

in particular, the precision with which we can estimate $\hat{\beta}_p$ depends on $\|\mathbf{z}_p\|_2$, which represents how much of $\mathbf{c}_p$ is *unexplained* by the other $\mathbf{c}_k$'s.

32. Obtaining $\widehat{\boldsymbol{\beta}}$ from Algorithm 1: Note that Step 2 in Algorithm 1 can be written as

$$\mathbf{X} = \mathbf{Z}\boldsymbol{\Gamma},$$

where the columns of $\mathbf{Z} \in \mathbb{R}^{n\times(p+1)}$ are $\mathbf{z}_1, \cdots, \mathbf{z}_p$, and $\boldsymbol{\Gamma} \in \mathbb{R}^{(p+1)\times(p+1)}$ is the upper triangular matrix with entries $\hat{\gamma}_{k,j}$. Let $\mathbf{D} = \mathrm{diag}\big(\|\mathbf{z}_0\|_2, \|\mathbf{z}_1\|_2, \cdots, \|\mathbf{z}_p\|_2\big) \in \mathbb{R}^{(p+1)\times(p+1)}$, and

$$\mathbf{X} = \mathbf{Z}\mathbf{D}^{-1}\mathbf{D}\boldsymbol{\Gamma} = \mathbf{Q}\mathbf{R},$$

where $\mathbf{Q}$ is an $n \times (p+1)$ orthogonal matrix satisfying $\mathbf{Q}^\top\mathbf{Q} = \mathbf{I}_{p+1}$, and $\mathbf{R}$ is a $(p+1)\times(p+1)$ upper triangular matrix. Under this $\mathbf{QR}$ decomposition of $\mathbf{X}$, the least squares solution is

$$\widehat{\boldsymbol{\beta}} = \mathbf{R}^{-1}\mathbf{Q}^\top\mathbf{Y}, \qquad \widehat{\mathbf{Y}} = \mathbf{Q}\mathbf{Q}^\top\mathbf{Y}.$$

In particular, it is easy to compute $\widehat{\boldsymbol{\beta}}$ using this procedure, since $\mathbf{R}$ is upper triangular.

IV. Subset Selection

1. Issues with the Least Squares Estimates:

- *Prediction Accuracy:*
  (a) The least squares estimates with too many variates included often have *low bias* but *large variance*, and *overfitting* will take place; and
  (b) The least squares estimates with too few variates included often have *low variance* but *high bias*, and *underfitting* will take place.

By shrinking or setting some (but *not* too many) coefficients to zero, we sacrifice a little bit of bias to reduce the variance of the predicted values, in order to improve the overall prediction accuracy.

- *Interpretation:* We often would like to determine a *smaller* subset that exhibit the strongest effects and sacrifice some small details.

We describe some techniques to select variables within linear regression.

2. Best-Subset Selection:

(a) *Main Idea:* Best subset regression finds, for each $k \in \{0, 1, 2, \cdots, p\}$, the subset of size k variables that gives the largest or the smallest value of the criterion under consideration.

(b) *Number of Models to Consider:* There are $\binom{p}{k}$ different subsets of variables that have k variables. In total, we will need to 2^p models to fit and consider. It is obvious that when p is large, this method becomes infeasible.

(c) *Criterion to Choose Model:* If we let P denote the sub-model that has certain k variables, one criterion (or a family of criteria) to assess its performance is of the form

$$\frac{\text{RSS}_P}{n} + \lambda(k+1)\frac{\hat{\sigma}^2_{\text{full}}}{n}, \tag{12}$$

where

- $\lambda > 0$ is a penalty coefficient,
- $\hat{\sigma}^2_{\text{full}}$ is the estimate of the variance from the *full* model,
- RSS_P is the residual sum of squares for the sub-model P, and
- the term $\lambda(k+1)\frac{\hat{\sigma}^2_{\text{full}}}{n}$ is called the *model complexity term*.

Special Cases of Criteria (12)*:*

i. If we let $\lambda = 2$, we obtain the *Akaike information criterion* (AIC);

ii. If we let $\lambda = \log n$, we obtain the *Bayesian information criterion* (BIC).

Remark 1. We cannot choose $\lambda = 0$ in (12), which reduces to the residual sum of squares of the model P. If we do so, the best-subset curve, by connecting the lowest residual sum of squares for each k, is necessarily decreasing, leading us to choose the model with all variables included.

Remark 2. The choice of k is a tradeoff between bias and variance, along with the more subjective desire for parsimony.

3. Forward-Stepwise Selection:

(a) *Procedures:* Forward-stepwise selection starts with the intercept, and then *sequentially* adds the predictor that most improves the fit.

In order to assess the amount of improvement, we can use the F-ratio

$$F = \frac{(\mathrm{RSS}_0 - \mathrm{RSS}_1)/1}{\mathrm{RSS}_1/\mathrm{df}_1}$$

and add the variable with the largest F-ratio, where $\mathrm{df}_1 = n - k - 1$ and k is the number of variables in the larger model, RSS_0 and RSS_1 are the residual sums of squares of the smaller and larger models, respectively. After adding the variable, we refit the model.

We stop selecting variables for the model when the F-ratio for each variable not currently in the model is smaller than some predetermined value F_0.

(b) *Comments:* Forward-stepwise selection

i. is a greedy algorithm, producing a nested sequence of models, which might produce a sub-optimal solution compared to the best-subset selection;

ii. has lower variance but maybe more bias; and

iii. is relatively easy to compute comparing to the best-subset selection.

4. Backward-Stepwise Selection:

(a) *Procedures:* Backward-stepwise selection starts with the full model (the model with *all* variables included), and sequentially deletes the predictor that has the least impact on the fit. The candidate variable for dropping is the one with the smallest Z-score or the lowest F-ratio

$$F = \frac{(\mathrm{RSS}_0 - \mathrm{RSS}_1)/1}{\mathrm{RSS}_1/\mathrm{df}_1},$$

where RSS_0 is the residual sum of squares for the smaller model (with df_0 degrees of freedom), and RSS_1 is the residual sum of squares for the larger model (with df_1 degrees of freedom), the "smaller" model is a sub-model of the "larger" model. Then, we refit the reduced model and iterate again. Here, $\mathrm{df}_0 - \mathrm{df}_1 = 1$ and $\mathrm{df}_1 = n - k - 1$, where k is the number of variables in the larger model.

Remark. Since $t_\nu^2 = F_{1,\nu}$, dropping the variable with the lowest Z score and dropping the one with lowest F-ratio are equivalent.

(b) *Comment:* Backward-stepwise selection can only be used when $n > p$, while forward-stepwise selection can always be used.

5. Criticisms of Forward- and Backward-Stepwise Selections:

(a) Stepwise selection methods ignore multiple testing problems;

(b) The maximum or minimum of a set of correlated F statistics is *not* an F statistic. Hence, the decision rules used in stepwise regression to add or drop an input variable can be misleading;

(c) There is *no* guarantee that the subsets obtained from forward- and backward-stepwise selection procedures will contain the same variables or even be the "best" subset;

(d) When there are more variables than observations $p > n$, backward-stepwise selection is typically *not* a feasible procedure;

(e) A stepwise procedure produces a *single* answer (a very specific subset) to the variable selection problem, although several different subsets may be equally good for regression purposes.

6. Forward Stagewise Regression:

(a) *Assumptions and Notation:*

i. The response vector $\mathbf{Y}$ has mean zero;

ii. The design matrix $\mathbf{X}$ has been standardized so that each column has zero mean and unit variance; let $\mathbf{c}_1, \cdots, \mathbf{c}_p$ be the columns of $\mathbf{X}$ (note that we do not have the constant unit vector here);

iii. Let $\widehat{\boldsymbol{\beta}}$ be the "current" estimate of the coefficient vector, $\widehat{\boldsymbol{\mu}} := \mathbf{X}\widehat{\boldsymbol{\beta}}$ be the "current" estimate of the response vector, and $\mathbf{r} := \mathbf{Y} - \widehat{\boldsymbol{\mu}}$ be the "current" residual vector.

(b) *Procedures:*

i. Initialize $\widehat{\boldsymbol{\beta}} = \mathbf{0}_{p+1}$, so that $\widehat{\boldsymbol{\mu}} = \mathbf{0}_n$ and $\mathbf{r} = \mathbf{Y} - \widehat{\boldsymbol{\mu}} = \mathbf{Y}$;

ii. Find the column vector $\mathbf{c}_{j'}$ that is most highly correlated correlation with $\mathbf{r}$, i.e.,

$$j' := \underset{j \in \{1,2,\cdots,p\}}{\arg\max} |\langle \mathbf{c}_j, \mathbf{r} \rangle|;$$

iii. Update $\hat{\beta}_{j'} \leftarrow \hat{\beta}_{j'} + \delta_{j'}$ and keep other coordinates unchanged, where $\delta := \varepsilon \cdot \text{sign}(\langle \mathbf{c}_{j'}, \mathbf{r} \rangle)$ and $\varepsilon > 0$ is a small constant that controls the step-length;

iv. Update $\widehat{\boldsymbol{\mu}} \leftarrow \widehat{\boldsymbol{\mu}} + \delta_{j'}\mathbf{c}_{j'}$ and $\mathbf{r} \leftarrow \mathbf{r} - \delta_{j'}\mathbf{c}_{j'}$;

v. Repeat Steps ii. - iv. many times until $\mathbf{X}^\top \mathbf{r} = \mathbf{0}_p$. This is the OLS solution.

(c) *Comments*:

i. In the forward stagewise regression, *no* adjustment of the other variables in the model is made when a new variable is added in — this is the key difference between the *forward-stepwise regression* and the *forward stagewise regression*;

ii. Forward stagewise regression may need *more than* p steps to reach the least squares fit, which is somewhat computationally inefficient.

V. Shrinkage Methods

1. Drawback of Subset Selection Methods: Subset selection is a *discrete* process — variables are either retained or discarded — it often exhibits high variance, and doesn't reduce the prediction error of the full model.

Shrinkage methods are a *continuous* process and do *not* suffer high variability.

2. A Slight Change of Notation: In this section *only*, we let

(a) $\mathbf{X} \in \mathbb{R}^{n\times p}$ whose first column is *not* the constant column vector, and

(b) $\boldsymbol{\beta} := (\beta_1, \cdots, \beta_p)^\top \in \mathbb{R}^p$ that does *not* contain the intercept.

3. Ridge Regression:

(a) *Formulation:* The ridge regression minimizes a penalized residual sum of squares

$$(\hat{\beta}_0^{\text{ridge}}, \hat{\beta}_1^{\text{ridge}}, \cdots, \hat{\beta}_p^{\text{ridge}})^\top := \underset{\beta_0,\beta_1,\cdots,\beta_p}{\arg\min} \left\{ \sum_{i=1}^n \left(y_i - \beta_0 - \sum_{j=1}^p \beta_j x_{i,j} \right)^2 + \lambda \sum_{j=1}^p \beta_j^2 \right\}, \tag{13}$$

where $\lambda \geq 0$ is a complexity parameter that controls the amount of shrinkage. The larger the value of λ, the greater the amount of shrinkage. The coefficients are shrunk toward zero.

(b) *Equivalent Formulation:* An equivalent formulation of ridge regression problem (13) is

$$(\hat{\beta}_0^{\text{ridge}}, \hat{\beta}_1^{\text{ridge}}, \cdots, \hat{\beta}_p^{\text{ridge}})^\top = \underset{\beta_0,\beta_1,\cdots,\beta_p}{\arg\min} \left\{ \sum_{i=1}^n \left(y_i - \beta_0 - \sum_{j=1}^p \beta_j x_{i,j} \right)^2, \text{ subject to } \sum_{j=1}^p \beta_j^2 \leq t \right\},$$

where $t \geq 0$.

Remark. There is a one-to-one correspondence between λ in (13) and t in (14).

(c) *Matrix Formulation:* In the matrix notation, the objective function in (13) can be written as

$$\text{RSS}_\lambda^{\text{ridge}}(\beta_0, \boldsymbol{\beta}) := (\mathbf{Y} - \beta_0 \mathbf{1}_n - \mathbf{X}\boldsymbol{\beta})^\top (\mathbf{Y} - \beta_0 \mathbf{1}_n - \mathbf{X}\boldsymbol{\beta}) + \lambda \|\boldsymbol{\beta}\|_2^2, \tag{14}$$

where $\mathbf{1}_n := (1, \cdots, 1)^\top \in \mathbb{R}^n$.

(d) *Solution to* $(\hat{\beta}_0^{\text{ridge}}, \hat{\beta}_1^{\text{ridge}}, \cdots, \hat{\beta}_p^{\text{ridge}})^\top$*:* Taking the first derivative of $\text{RSS}_\lambda^{\text{ridge}}$ with respect to β_0 and $\boldsymbol{\beta}$, respectively, and setting the derivatives to 0, we have $\hat{\beta}_0^{\text{ridge}}$ and $\widehat{\boldsymbol{\beta}}^{\text{ridge}} := (\hat{\beta}_1^{\text{ridge}}, \cdots, \hat{\beta}_p^{\text{ridge}})^\top$ must satisfy

$$(\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I}_n) \widehat{\boldsymbol{\beta}}^{\text{ridge}} = \mathbf{X}^\top (\mathbf{Y} - \hat{\beta}_0^{\text{ridge}} \mathbf{1}_n), \tag{15}$$

$$n \hat{\beta}_0^{\text{ridge}} = \mathbf{1}_n^\top (\mathbf{Y} - \mathbf{X} \widehat{\boldsymbol{\beta}}^{\text{ridge}}). \tag{16}$$

(e) *Suggestion in Implementation:* The ridge solutions are *not* equi-variant under scaling of the inputs. It is suggested to standardize the covariates before solving (13).

(f) *Another Equivalent Formulation of Ridge Regression:* We show that the ridge regression formulation (13) is equivalent to the following one

$$\left(\hat{\beta}_0^c, \widehat{\boldsymbol{\beta}}^{\text{ridge},c} \right) := \underset{\beta_0^c, \boldsymbol{\beta}^c}{\arg\min} \left\{ \sum_{i=1}^n \left(y_i - \beta_0^c - \sum_{j=1}^p \beta_j^c (x_{i,j} - \bar{x}_j) \right)^2 + \lambda \sum_{j=1}^p \beta_j^{c2} \right\},$$

where $\bar{x}_j$ is the sample mean of the j-th column in the design matrix. Start from (13), we have

$$\begin{aligned}\left(\hat{\beta}_0^{\text{ridge}}, \widehat{\boldsymbol{\beta}}^{\text{ridge}}\right) &= \arg\min_{\beta_0,\boldsymbol{\beta}} \left\{ \sum_{i=1}^n \left(y_i - \beta_0 - \sum_{j=1}^p \beta_j x_{i,j} \right)^2 + \lambda \sum_{j=1}^p \beta_j^2 \right\} \\ &= \arg\min_{\beta_0,\boldsymbol{\beta}} \left\{ \sum_{i=1}^n \left(y_i - \beta_0 - \sum_{j=1}^p \beta_j \bar{x}_j - \sum_{j=1}^p \beta_j (x_{i,j} - \bar{x}_j) \right)^2 + \lambda \sum_{j=1}^p \beta_j^2 \right\}.\end{aligned}$$

Hence, we have

$$\hat{\beta}_0^c = \hat{\beta}_0^{\text{ridge}} + \sum_{j=1}^p \hat{\beta}_j^{\text{ridge}} \bar{x}_j, \qquad \text{and} \qquad \hat{\beta}_j^{\text{ridge},c} = \hat{\beta}_j^{\text{ridge}},$$

for all $j = 1, \cdots, p$. But, from (16), we see that

$$\hat{\beta}_0^{\text{ridge}} = \frac{1}{n} \mathbf{1}_n^\top (\mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}}^{\text{ridge}}) = \frac{1}{n} \sum_{i=1}^n y_i - \sum_{j=1}^p \hat{\beta}_j^{\text{ridge}} \bar{x}_j.$$

Hence, the solution can be obtained by the following procedure:

Step 1. Center all covariates, i.e., each $x_{i,j}$ is replaced by $x_{i,j} - \bar{x}_j$;
Step 2. Estimate β_0 by $\bar{y} = \frac{1}{n}\sum_{i=1}^n y_i$;
Step 3. Remaining coefficients are estimated by a ridge regression without intercept by using the centered covariates.

(g) *Solution to Ridge Regression without an Intercept:* Assume all covariates have been centered so that there is no intercept. The ridge problem amounts to minimizing

$$\widetilde{\text{RSS}}_\lambda^{\text{ridge}}(\boldsymbol{\beta}) := (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})^\top (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) + \lambda \|\boldsymbol{\beta}\|_2^2.$$

By taking the derivative with respect to $\boldsymbol{\beta}$ and setting the derivative to 0, we solve for $\boldsymbol{\beta}$ and obtain

$$\widehat{\boldsymbol{\beta}}^{\text{ridge}}(\lambda) := (\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I}_p)^{-1} \mathbf{X}^\top \mathbf{Y}, \tag{17}$$

which again is a linear function in $\mathbf{Y}$.

Remark. By adding $\lambda \mathbf{I}_p$ to $\mathbf{X}^\top \mathbf{X}$ with $\lambda > 0$, $\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I}_p$ is always invertible, even if $\mathbf{X}^\top \mathbf{X}$ is singular.

(h) *Bayesian Interpretation:* Suppose that $y_i \sim \text{Normal}(\beta_0 + \boldsymbol{\beta}^\top \mathbf{x}_i, \sigma^2)$ and the parameters $\boldsymbol{\beta} \sim \text{Normal}_p(\mathbf{0}_p, \tau^2 \mathbf{I}_p)$, where $\text{Normal}_p(\mu, \Sigma)$ denotes the multivariate normal distribution with mean μ and covariance matrix Σ. Then, under the assumption that σ^2 and τ^2 are known, the negative log-posterior density of $\boldsymbol{\beta}$ is

$$\begin{aligned}-\log f(\boldsymbol{\beta}; \sigma^2, \tau^2) &\propto \frac{1}{\sigma^2} \sum_{i=1}^n (y_i - \beta_0 - \boldsymbol{\beta}^\top \mathbf{x}_i)^2 + \frac{1}{\tau^2} \|\boldsymbol{\beta}\|_2^2 \\ &\propto \sum_{i=1}^n (y_i - \beta_0 - \boldsymbol{\beta}^\top \mathbf{x}_i)^2 + \frac{\sigma^2}{\tau^2} \|\boldsymbol{\beta}\|_2^2,\end{aligned}$$

where we see $\lambda = \sigma^2/\tau^2$. Thus, the ridge regression estimator of $\boldsymbol{\beta}$ is the *mode* of the posterior distribution. Since the posterior distribution is normal, which is symmetric, we conclude that this estimator is also the posterior mean.

(i) *Singular Value Decomposition:* The singular value decomposition (SVD) of the matrix $\mathbf{X} \in \mathbb{R}^{n\times p}$ has the form

$$\mathbf{X} = \mathbf{U}\mathbf{D}\mathbf{V}^\top \in \mathbb{R}^{n\times p},$$

where $\mathbf{U}$ and $\mathbf{V}$ are $n \times p$ and $p \times p$ orthogonal matrices, respectively, and $\mathbf{D}$ is a $p \times p$ diagonal matrix with entries $d_1 \geq \cdots \geq d_p$ called the *singular values* of $\mathbf{X}$. Here $\mathbf{U}$ spans the column space of $\mathbf{X}$ and $\mathbf{V}$ spans the row space of $\mathbf{X}$.

(j) *SVD Analysis of Least Squares Solution:* Consider the least squares solution

$$\begin{aligned}\widehat{\boldsymbol{\beta}}^{\text{ls}} &:= \arg\min_{\boldsymbol{\beta}} (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})^\top (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})^\top \\ &= (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{Y},\end{aligned}$$

where we have centered both $\mathbf{Y}$ and each column of $\mathbf{X}$.

The corresponding least squares fitted values are

$$\widehat{\mathbf{Y}} = \mathbf{X}\widehat{\boldsymbol{\beta}}^{\text{ls}} = \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{Y} = \mathbf{U}\mathbf{U}^\top\mathbf{Y} = \sum_{j=1}^{p} \mathbf{u}_j\mathbf{u}_j^\top\mathbf{Y},$$

where $\mathbf{U}^\top\mathbf{Y}$ are the coordinates of $\mathbf{Y}$ with respect to the orthonormal basis $\mathbf{U}$ and $\mathbf{u}_j$ denotes the j-th column of $\mathbf{U}$ for all $j = 1, 2, \cdots, p$.

(k) *SVD Analysis of Ridge Regression:* The fitted values of the ridge regression is

$$\begin{aligned}\mathbf{X}\widehat{\boldsymbol{\beta}}^{\text{ridge}}(\lambda) &= \mathbf{X}(\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^\top\mathbf{Y} \\ &= \mathbf{U}\mathbf{D}(\mathbf{D}^2 + \lambda\mathbf{I})^{-1}\mathbf{D}\mathbf{U}^\top\mathbf{Y} \\ &= \sum_{j=1}^{p} \frac{d_j^2}{d_j^2 + \lambda}\mathbf{u}_j\mathbf{u}_j^\top\mathbf{Y}.\end{aligned}$$

Hence, ridge regression also computes the coordinates of $\mathbf{Y}$ with respect to the orthonormal basis $\mathbf{U}$, but shrinks by the factors $d_j^2/(d_j^2 + \lambda)$.

Remark. A greater amount of shrinkage is applied to the coordinates of basis vectors with smaller d_j^2, which corresponds to the directions in the column space of $\mathbf{X}$ having smaller variances.

(l) *Effective Degrees of Freedom:* Define the *effective degrees of freedom* of the ridge regression to be

$$\text{df}(\lambda) = \text{trace}\big[\mathbf{X}(\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}\big] = \sum_{j=1}^{p} \frac{d_j^2}{d_j^2 + \lambda}.$$

This is a monotone decreasing function of λ. Note that the following two special cases:

i. $\mathrm{df}(\lambda) = p$ if $\lambda = 0$, corresponding to the case with no regularization and the least squares case, and

ii. $\mathrm{df}(\lambda) \to 0$ as $\lambda \to \infty$.

(m) *Bias-Variance Trade-off:* We consider the mean squared error of the ridge regression estimator

$$\begin{aligned}\mathrm{MSE}(\lambda) &= \mathbb{E}\big[(\widehat{\boldsymbol{\beta}}^{\mathrm{ridge}}(\lambda) - \boldsymbol{\beta})^\top (\widehat{\boldsymbol{\beta}}^{\mathrm{ridge}}(\lambda) - \boldsymbol{\beta})\big] \\ &= \big(\mathrm{Bias}(\lambda)\big)^2 + \mathrm{Var}[\widehat{\boldsymbol{\beta}}^{\mathrm{ridge}}(\lambda)].\end{aligned}$$

For the variance term, we have

$$\begin{aligned}\mathrm{Var}[\widehat{\boldsymbol{\beta}}^{\mathrm{ridge}}(\lambda)] &= \sigma^2 \,\mathrm{trace}\Big((\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I}_p)^{-1}\mathbf{X}^\top\mathbf{X}(\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I}_p)^{-1}\Big) \\ &= \sigma^2 \,\mathrm{trace}\big((\mathbf{D}^2 + \lambda\mathbf{I}_p)^{-1}\mathbf{D}^2(\mathbf{D}^2 + \lambda\mathbf{I}_p)^{-1}\big) \\ &= \sigma^2 \sum_{j=1}^{p} \frac{d_j^2}{(d_j^2 + \lambda)^2}.\end{aligned}$$

For the squared bias term, we first note

$$\begin{aligned}\mathbb{E}[(\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I}_p)^{-1}\mathbf{X}^\top\mathbf{Y}] - \boldsymbol{\beta} &= \big((\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I}_p)^{-1}\mathbf{X}^\top\mathbf{X} - \mathbf{I}_p\big)\boldsymbol{\beta} \\ &= \mathbf{V}[(\mathbf{D}^2 + \lambda\mathbf{I}_p)^{-1}\mathbf{D}^2 - \mathbf{I}_p]\mathbf{V}^\top\boldsymbol{\beta}.\end{aligned}$$

Therefore, if we let $\boldsymbol{\alpha} := (\alpha_1, \alpha_2, \cdots, \alpha_p)^\top = \mathbf{V}^\top\boldsymbol{\beta}$, we have

$$\begin{aligned}\big(\mathrm{Bias}(\lambda)\big)^2 &= \big[\mathbb{E}[(\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I}_p)^{-1}\mathbf{X}^\top\mathbf{Y}] - \boldsymbol{\beta}\big]^\top \big[\mathbb{E}[(\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I}_p)^{-1}\mathbf{X}^\top\mathbf{Y}] - \boldsymbol{\beta}\big] \\ &= \boldsymbol{\alpha}^\top[\mathbf{D}^2(\mathbf{D}^2 + \lambda\mathbf{I}_p)^{-1} - \mathbf{I}_p][(\mathbf{D}^2 + \lambda\mathbf{I}_p)^{-1}\mathbf{D}^2 - \mathbf{I}_p]\boldsymbol{\alpha} \\ &= \sum_{j=1}^{p} \frac{\lambda^2\alpha_j^2}{(d_j^2 + \lambda)^2}.\end{aligned}$$

Therefore, we have

$$\mathrm{MSE}(\lambda) = \sum_{j=1}^{p} \frac{\lambda^2\alpha_j^2 + \sigma^2 d_j^2}{(d_j^2 + \lambda)^2}.$$

Note the following:

i. when $\lambda = 0$, the squared-bias term is zero;

ii. the variance term decreases monotonically as λ increases from zero, whereas the squared-bias term increases;

iii. for large values of λ, the squared-bias term dominates the mean squared error.

4. Least absolute shrinkage and selection operator (Lasso):

(a) *Formulation:* The Lagrangian formulation of the lasso regression is

$$\left(\hat{\beta}_0^{\text{lasso}}, \widehat{\boldsymbol{\beta}}^{\text{lasso}}\right) := \arg\min_{\beta_0, \boldsymbol{\beta}} \left\{ \frac{1}{2} \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{i,j} \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right\}, \tag{18}$$

where $\widehat{\boldsymbol{\beta}}^{\text{lasso}} := (\hat{\beta}_1^{\text{lasso}}, \cdots, \hat{\beta}_p^{\text{lasso}})^\top \in \mathbb{R}^p$, and $\lambda \geq 0$ is a complexity parameter that controls the amount of shrinkage. The *constrained* formulation of the lasso regression is

$$\begin{aligned} \left(\hat{\beta}_0^{\text{lasso}}, \widehat{\boldsymbol{\beta}}^{\text{lasso}}\right) = \arg\min_{\beta_0, \boldsymbol{\beta}} \frac{1}{2} \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{i,j} \right)^2, \\ \text{subject to } \sum_{j=1}^{p} |\beta_j| \leq t. \end{aligned} \tag{19}$$

(b) *Solution to β_0:* Suppose that the columns of $\mathbf{X}$ has been standardized so that each column has zero mean and unit norm. The solution to β_0 is

$$\hat{\beta}_0^{\text{lasso}} = \frac{1}{n} \sum_{i=1}^{n} y_i.$$

Denote the objective function in (18) by $\text{RSS}_\lambda^{\text{lasso}}$. Taking the partial derivative of $\text{RSS}_\lambda^{\text{lasso}}$ with respect to β_0 and setting the result to zero yield

$$\begin{aligned} \hat{\beta}_0^{\text{lasso}} &= \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{p} \hat{\beta}_j^{\text{lasso}} x_{i,j} \right) \\ &= \sum_{i=1}^{n} y_i - \sum_{j=1}^{p} \hat{\beta}_j^{\text{lasso}} \left(\sum_{i=1}^{n} x_{i,j} \right) \\ &= \sum_{i=1}^{n} y_i, \end{aligned}$$

where we obtain the last equality since columns of $\mathbf{X}$ has zero mean.

Hence, in the sequel, we fit a model without an intercept.

(c) *Continuous Variable Selection:* Due to the nature of the constraint, making t in (19) sufficiently small or λ in (18) sufficiently large will cause some of the coefficients to be exactly zero. Thus, the lasso does a kind of *continuous subset selection*.

(d) *Special Choices of t:* If t is chosen larger than $t_0 := \sum_{j=1}^{p} |\hat{\beta}_j|$, where $\hat{\beta}_j$'s are the least squares estimates, then the lasso estimates are equal to the least squares estimates.

(e) *Optimality Conditions:* By the theory of convex analysis, the necessary and sufficient conditions for $\widehat{\boldsymbol{\beta}}^{\text{lasso}}$ is

$$-\frac{1}{n}\left\langle \mathbf{c}_j, \mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}}^{\text{lasso}} \right\rangle + \lambda s_j = 0, \qquad \text{for all } j = 1, 2, \cdots, p,$$

where $\mathbf{c}_j$ denotes the j-th column of $\mathbf{X}$ and

$$s_j = \begin{cases} \text{sign}(\hat{\beta}_j^{\text{lasso}}), & \text{if } \hat{\beta}_j^{\text{lasso}} \neq 0, \\ \text{any value between } -1 \text{ and } 1, & \text{if } \hat{\beta}_j^{\text{lasso}} = 0. \end{cases}$$

(f) *Explicit Solution in Single Predictor Case:* Assume we only have one predictor, and both $\{y_i\}_{i=1}^n$ and $\{x_i\}_{i=1}^n$ have been standardized so that

$$\sum_{i=1}^n y_i = 0, \qquad \sum_{i=1}^n x_i = 0, \qquad \text{and} \qquad \sum_{i=1}^n x_i^2 = 1.$$

Consider the lasso problem with only one predictor and with no intercept

$$\underset{\beta}{\text{minimize}} \left\{ \frac{1}{2}\sum_{i=1}^n (y_i - \beta x_i)^2 + \lambda|\beta| \right\}. \tag{20}$$

Denote the objective function in (20) by $f(\beta)$, and note

$$\begin{aligned} f(\beta) &= \frac{1}{2}\sum_{i=1}^n (y_i^2 - 2x_i y_i \beta + \beta^2 x_i^2) + \lambda|\beta| \\ &= \frac{1}{2}\left(\sum_{i=1}^n x_i^2\right)\beta^2 - \left(\sum_{i=1}^n x_i y_i\right)\beta + \lambda|\beta| + \frac{1}{2}\left(\sum_{i=1}^n y_i^2\right) \\ &= \frac{1}{2}\beta^2 - \left(\sum_{i=1}^n x_i y_i\right)\beta + \lambda|\beta| + \frac{1}{2}\left(\sum_{i=1}^n y_i^2\right) \\ &= \frac{1}{2}\beta^2 - \hat{\beta}^{\text{ls}}\beta + \lambda|\beta| + \frac{1}{2}\left(\sum_{i=1}^n y_i^2\right), \end{aligned}$$

by noting that

$$\hat{\beta}^{\text{ls}} = \arg\min_{\beta} \left\{ \frac{1}{2}\sum_{i=1}^n (y_i - x_i\beta)^2 \right\} = \sum_{i=1}^n x_i y_i.$$

We now only need to consider to minimize

$$h(\beta) := \frac{1}{2}\beta^2 - \hat{\beta}^{\text{ls}}\beta + \lambda|\beta|,$$

since $\frac{1}{2}\sum_{i=1}^n y_i^2$ term does *not* affect β.
Consider the following two cases:

- if $\hat{\beta}^{\text{ls}} \geq 0$, then it is necessary that the minimizer of h must be nonnegative. Hence, we only need to minimize

$$\begin{aligned}\tilde{h}(\beta) :=& \frac{1}{2}\beta^2 - \hat{\beta}^{\text{ls}}\beta + \lambda\beta \\ =& \frac{1}{2}\beta^2 - (\hat{\beta}^{\text{ls}} - \lambda)\beta, \qquad \text{for } \beta \geq 0.\end{aligned}$$

  Then, if $\hat{\beta}^{\text{ls}} - \lambda \geq 0$, the minimizer is at $\hat{\beta} := \hat{\beta}^{\text{ls}} - \lambda$; if $\hat{\beta}^{\text{ls}} - \lambda < 0$, the minimizer is at 0.
- if $\hat{\beta}^{\text{ls}} < 0$, then it is necessary that the minimizer of h must be negative. Hence, we only need to minimize

$$\begin{aligned}\tilde{h}(\beta) :=& \frac{1}{2}\beta^2 - \hat{\beta}^{\text{ls}}\beta - \lambda\beta \\ =& \frac{1}{2}\beta^2 - (\hat{\beta}^{\text{ls}} + \lambda)\beta, \qquad \text{for } \beta < 0.\end{aligned}$$

  Then, if $\hat{\beta}^{\text{ls}} + \lambda < 0$, the minimizer is at

$$\hat{\beta} := \hat{\beta}^{\text{ls}} + \lambda = -|\hat{\beta}^{\text{ls}}| + \lambda = \text{sign}(\hat{\beta}^{\text{ls}})(|\hat{\beta}^{\text{ls}}| - \lambda);$$

  if $\hat{\beta}^{\text{ls}} - \lambda \geq 0$, the minimizer is at 0.

Summarizing two cases, we conclude that the minimizer to (**??**), denoted by $\hat{\beta}^{\text{lasso}}$, is

$$\hat{\beta}^{\text{lasso}} = \text{sign}(\hat{\beta}^{\text{ls}})\big(|\hat{\beta}^{\text{ls}}| - \lambda\big)_+,$$

which is called the *soft-thresholding operator* and $(x)_+ = \max\{x, 0\}$.

(g) *Soft-thresholding Operator:* We denote the *soft-thresholding operator* as

$$S(x, \lambda) = \text{sign}(x)(|x| - \lambda)_+, \qquad \text{for all } x \in \mathbb{R} \text{ and } \lambda \geq 0. \tag{21}$$

Note that $S(x, \lambda)$ translates its argument x toward zero by the amount λ and sets it to zero if $|x| \leq \lambda$.

(h) *Explicit Solution under Orthonormal Design:* Suppose the design matrix $\mathbf{X}$ has orthonormal columns, meaning that

$$\langle \mathbf{c}_i, \mathbf{c}_j \rangle = \begin{cases} 0, & \text{if } i \neq j, \\ 1, & \text{if } i = j. \end{cases}$$

Then, we have, for all $j = 1, 2, \cdots, p$,

$$\hat{\beta}_j^{\text{lasso}} = S(\hat{\beta}_j^{\text{ls}}, \lambda) = \text{sign}(\hat{\beta}_j^{\text{ls}})\big(|\hat{\beta}_j^{\text{ls}}| - \lambda\big)_+, \tag{22}$$

where $\hat{\beta}_j^{\text{ls}}$ denotes the least squares estimate for the j-th coefficient.

Recall that the lasso minimizes

$$\begin{aligned}\text{RSS}_\lambda^{\text{lasso}} &= \frac{1}{2}(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta})^\top(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta})+\lambda\|\boldsymbol{\beta}\|_1 \\ &= \frac{1}{2}\mathbf{Y}^\top\mathbf{Y}-\mathbf{Y}^\top\mathbf{X}\boldsymbol{\beta}+\frac{1}{2}\boldsymbol{\beta}^\top\mathbf{X}^\top\mathbf{X}\boldsymbol{\beta}+\lambda\sum_{j=1}^{p}|\beta_j| \\ &\overset{\text{(i)}}{=} \frac{1}{2}\mathbf{Y}^\top\mathbf{Y}+\sum_{j=1}^{p}\left(\frac{1}{2}\beta_j^2-(\mathbf{Y}^\top\mathbf{c}_j)\beta_j+\lambda|\beta_j|\right) \\ &\overset{\text{(ii)}}{=} \frac{1}{2}\mathbf{Y}^\top\mathbf{Y}+\sum_{j=1}^{p}\left(\frac{1}{2}\beta_j^2-\hat{\beta}_j^{\text{ls}}\beta_j+\lambda|\beta_j|\right), \end{aligned} \tag{23}$$

where we use the orthonormality of $\mathbf{X}$ in (i) and $\hat{\beta}_j^{\text{ls}} = \mathbf{Y}^\top\mathbf{c}_j$ in (ii). To see $\hat{\beta}_j^{\text{ls}} = \mathbf{Y}^\top\mathbf{c}_j$, note that

$$\widehat{\boldsymbol{\beta}}^{\text{ls}} = \arg\min_{\boldsymbol{\beta}}(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta})^\top(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta}) = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{Y} = \mathbf{X}^\top\mathbf{Y}.$$

What remains to show is that $S(\hat{\beta}^{\text{ls}},\lambda)$ minimizes

$$h(\beta) = \frac{1}{2}\beta^2-\hat{\beta}^{\text{ls}}\beta+\lambda|\beta|,$$

by noting that (23) is separable in each coordinate. The desired result follows from the solution to the single predictor case.

(i) *Computation – Cyclic Coordinate Descent Algorithm:* Explicit solution of the lasso problem under the general case does *not* exist. We develop a *cyclic coordinate descent algorithm* to compute the solution.

i. Main Idea: Fix the penalty parameter λ in the Lagrangian form of the Lasso problem (18). We repeatedly cycle through the predictors in order, where at the j-th step, we update the coefficient β_j by minimizing the objective function in this coordinate while holding all other coefficients fixed at the current value.

ii. *Derivation of Algorithm:* Let $\hat{\beta}_k(\lambda)$ be the current estimate of β_k at the penalty parameter λ, for $k = 1,\cdots,p$ and $k \neq j$. We can write the objective function for the lasso problem as

$$\frac{1}{2}\sum_{i=1}^{n}\left(y_i-\sum_{k\neq j}\hat{\beta}_k(\lambda)x_{i,k}-\beta_j x_{i,j}\right)^2+\lambda\sum_{k\neq j}|\hat{\beta}_k(\lambda)|+\lambda|\beta_j|.$$

This can be viewed as a univariate Lasso problem with the response variable being the partial residual

$$r_i^{(j)} := y_i-\sum_{k\neq j}\hat{\beta}_k(\lambda)x_{i,k},$$

where $\hat{\beta}_k(\lambda)$'s are viewed as fixed. In terms of the partial residuals, β_j is updated as

$$\hat{\beta}_j = S\big(\langle \mathbf{c}_j, \mathbf{r}^{(j)} \rangle, \lambda\big),$$

where $\mathbf{r}^{(j)} := (r_1^{(j)}, r_2^{(j)}, \cdots, r_n^{(j)})^\top \in \mathbb{R}^n$ and $S(\,\cdot\,, \lambda)$ is the soft-thresholding operator.
In addition, if we let $\hat{\beta}_j(\lambda)$ denote the current value of β_j, we have

$$\begin{aligned}
\langle \mathbf{c}_j, \mathbf{r}^{(j)} \rangle &= \langle \mathbf{c}_j, \mathbf{r}^{(j)} - \hat{\beta}_j(\lambda)\mathbf{c}_j + \hat{\beta}_j(\lambda)\mathbf{c}_j \rangle \\
&= \langle \mathbf{c}_j, \mathbf{r} + \hat{\beta}_j(\lambda)\mathbf{c}_j \rangle \\
&= \langle \mathbf{c}_j, \mathbf{r} \rangle + \hat{\beta}_j(\lambda)\langle \mathbf{c}_j, \mathbf{c}_j \rangle \\
&= \langle \mathbf{c}_j, \mathbf{r} \rangle + \hat{\beta}_j,
\end{aligned}$$

where $\mathbf{r} = (r_1, r_2, \cdots, r_n)^\top \in \mathbb{R}^n$ is the full residual vector with $r_i = y_i - \sum_{j=1}^p \hat{\beta}_j(\lambda) x_{i,j}$, for all $i = 1, 2, \cdots, n$, and we obtain the last equality since each column $\mathbf{c}_j$, for all $j = 1, 2, \cdots, p$, has unit norm.
Therefore, we can update the coefficient β_j as

$$\hat{\beta}_j(\lambda) \leftarrow S\Big(\hat{\beta}_j(\lambda) + \langle \mathbf{c}_j, \mathbf{r} \rangle, \lambda\Big), \tag{24}$$

We repeat iteration in (24) by cycling each variable in turn until convergence.

iii. Correctness of Algorithm: Note that the objective function (18) is a convex function of $\boldsymbol{\beta}$ and has no local minima. The cyclic coordinate descent algorithm derived above minimizes this convex objective function along each coordinate at a time, and (under certain mild conditions) converges to a global minimum.

5. **Comparisons of Subset Selection, Ridge Regression and Lasso:** In the case of the orthonormal design matrix $\mathbf{X}$, the solutions to three methods are

   - *Best Subset Selection of Size M:* $\hat{\beta}_j^{\text{ls}} \times \mathbb{1}(|\hat{\beta}_j^{\text{ls}}| \geq |\hat{\beta}_{(M)}^{\text{ls}}|)$;
   - *Ridge Regression:* $\hat{\beta}_j^{\text{ls}}/(1+\lambda)$;
   - *Lasso:* $\text{sign}(\hat{\beta}_j^{\text{ls}})\big(|\hat{\beta}_j^{\text{ls}}| - \lambda\big)_+$,

   where $\hat{\beta}_j^{\text{ls}}$'s are the least squares estimates. We see that

   - Ridge regression does a *proportional* shrinkage;
   - Lasso translates each coefficient by a constant factor $\lambda > 0$, truncating at zero. This is called "soft thresholding";
   - Best-subset selection drops all variables with coefficients smaller than the M-th largest; this is a form of "hard-thresholding".

6. **Generalizations of Ridge and Lasso:**

(a) *Problem Formulation:* Consider the criterion

$$\left(\tilde{\beta}_0, \widetilde{\boldsymbol{\beta}}\right) := \underset{\beta_0, \boldsymbol{\beta}}{\arg\min} \left\{ \frac{1}{2} \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{i,j} \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j|^q \right\}, \tag{25}$$

for $q \geq 0$.

(b) *Bayes Estimate Interpretation:* We can think of $|\beta_j|^q$ as the log-prior density for β_j, where

- $q = 0$ corresponds to the subset selection,
- $q = 1$ corresponds to the lasso, and
- $q = 2$ corresponds to the ridge regression.

In this view, the lasso, ridge regression and best subset selection are *Bayes estimates* with different priors, and they all can be derived as *posterior modes*.

(c) *Picking $q \in (1, 2)$:*

- Choosing $q \in (1, 2)$ results in a compromise between the lasso and the ridge;
- $|\beta_j|^q$ is differentiable at 0;
- These choices of q *cannot* set coefficients exactly to 0.

7. Elastic Net: The *elastic net penalty* is

$$\lambda \sum_{j=1}^{p} \left(\frac{1}{2}(1-\alpha)\beta_j^2 + \alpha|\beta_j| \right),$$

where $\lambda \geq 0$ and $\alpha \in [0, 1]$.

(a) *Comments:* The elastic net penalty

- selects variables like the lasso, and shrinks together the coefficients of correlated predictors like ridge;
- has considerable computational advantages over the L_q penalties; and
- for any $\alpha < 1$ and $\lambda > 0$, the elastic net problem is *strictly convex*, implying a unique solution exists irrespective of the correlations or duplications of predictors.

(b) *Coordinate Descent Algorithm for Solving Elastic Net Problem:* We first center each column of the design matrix $\mathbf{X}$ so that the sum of all entries in each column is 0, and compute the optimal intercept

$$\hat{\beta}_0 = \frac{1}{n} \sum_{i=1}^{n} y_i.$$

The update of each coordinate coefficient is

$$\hat{\beta}_j \leftarrow \frac{S(\sum_{i=1}^{n} r_{i,j} x_{i,j}, \lambda\alpha)}{\sum_{i=1}^{n} x_{i,j}^2 + \lambda(1-\alpha)}, \tag{26}$$

where $S(x,u) = \text{sign}(x)(|x|-u)_+$ is the soft-thresholding operator, and

$$r_{i,j} = y_i - \hat{\beta}_0 - \sum_{k \neq j} x_{i,k}\hat{\beta}_k$$

is the partial residual. We cycle over the updates (26) until convergence.

8. Least Angle Regression (LAR):

(a) *Main Idea:* LAR builds a model sequentially, adds one variable at a time "as much" of a variable as it deserves – LAR moves the coefficient of this variable continuously toward its least squares value.

(b) *Algorithm Details:*

Algorithm 2 Least Angle Regression

1: Standardize the design matrix so that each column has zero mean and unit norm.
2: Start with the residual

$$\mathbf{r} := \mathbf{Y} - \frac{1}{n}\mathbf{1}_n^\top \mathbf{Y}$$

and $\beta_1 = \beta_2 = \cdots = \beta_p = 0$.
3: Find the predictor X_j most correlated with $\mathbf{r}$.
4: Move β_j from 0 towards its least-squares coefficient $\langle X_j, \mathbf{r}\rangle$, until some other competitor X_k has as much correlation with the current residual as does X_j.
5: Move β_j and β_k in the direction defined by their joint least squares coefficient of the current residual on (X_j, X_k), until some other competitor X_l has as much correlation with the current residual.
6: Continue in this way until all p predictors have been entered. After $\min(n-1, p)$ steps, we arrive at the full least squares solution.

(c) *Update Direction:* Suppose $\mathcal{A}_k$ is the active set of variables at the beginning of the kth step, and let $\boldsymbol{\beta}_{\mathcal{A}_k}$ be the coefficient vector for these variables at this step; there will be $k-1$ nonzero values, and the one just entered will be zero.
If $\mathbf{r}_k := \mathbf{Y} - \mathbf{X}_{\mathcal{A}_k}\boldsymbol{\beta}_{\mathcal{A}_k}$ is the current residual, then the direction for this step is

$$\boldsymbol{\delta}_k := (\mathbf{X}_{\mathcal{A}_k}^\top \mathbf{X}_{\mathcal{A}_k})^{-1}\mathbf{X}_{\mathcal{A}_k}\mathbf{r}_k,$$

The coefficient profile evolves as $\boldsymbol{\beta}_{\mathcal{A}_k}(\alpha) := \boldsymbol{\beta}_{\mathcal{A}_k} + \alpha\boldsymbol{\delta}_k$.
Along the direction $\boldsymbol{\delta}_k$, the correlations tie and keep decreasing.

(d) *Update of Fitted Value Vector:* If the fitted value vector at the beginning of the k-th step is $\widehat{\mathbf{f}}_k$, then it evolves as

$$\widehat{\mathbf{f}}_k(\alpha) = \widehat{\mathbf{f}}_k + \alpha \cdot \mathbf{u}_k,$$

where $\mathbf{u}_k := \mathbf{X}_{\mathcal{A}_k}\boldsymbol{\delta}_k$ is the new fit direction.

Remark. The name "least angle" arises from a geometric interpretation of this process: $\mathbf{u}_k$ makes the smallest (and equal) angle with each of the predictors in the active set $\mathcal{A}_k$.

(e) *Modification of LAR to Compute Lasso Solution:* Modify Step 5 in Algorithm 2 to the following.

Algorithm 3 Modified Least Angle Regression to Compute Lasso Solution

If a non-zero coefficient hits zero, drop its variable from the active set of variables and recompute the current joint least squares direction.

(f) *Similarity between LAR and Lasso:* Suppose $\mathcal{A}$ is the active set of variables at some stage in the algorithm, tied in their absolute inner-product with the current residuals $\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}_{\mathcal{A}}$. We can express this as

$$\langle \mathbf{c}_j, \mathbf{Y} - \mathbf{X}\boldsymbol{\beta}_{\mathcal{A}} \rangle = \gamma s_j, \qquad \text{for all } j \in \mathcal{A}, \tag{27}$$

where $s_j \in \{-1, 1\}$ indicates the sign of the inner-product, and γ is the common correlation value. In addition, we have

$$\left|\langle \mathbf{x}_k, \mathbf{Y} - \mathbf{X}\boldsymbol{\beta}_{\mathcal{A}} \rangle\right| \le \gamma, \qquad \text{for all } k \notin \mathcal{A}. \tag{28}$$

Now consider the lasso criterion (18), which we write the objective function in vector form

$$\mathrm{RSS}_\lambda^{\text{lasso}}(\boldsymbol{\beta}) := \frac{1}{2}\|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_1.$$

Let $\mathcal{B}$ be the active set of variables in the solution for a given value of λ. For these variables, $\mathrm{RSS}_\lambda^{\text{lasso}}$ is differentiable, and the stationarity conditions give

$$\langle \mathbf{x}_j, \mathbf{Y} - \mathbf{X}\boldsymbol{\beta} \rangle = \lambda \cdot \mathrm{sign}(\beta_j), \qquad \text{for all } j \in \mathcal{B}. \tag{29}$$

For variables not in $\mathcal{B}$, using the subgradient, we have

$$\left|\langle \mathbf{x}_k, \mathbf{Y} - \mathbf{X}\boldsymbol{\beta} \rangle\right| \le \lambda, \qquad \text{for all } k \notin \mathcal{B}. \tag{30}$$

Comparing (27) with (29), they are identical only if the sign of β_j matches the sign of the inner product. Comparing (28) with (30), they are almost identical. These reveal the similarity between LAR and Lasso.

9. Degrees-of-Freedom Formula for LAR and Lasso:

(a) *Definition in Classical Statistics:* In classical statistics, the *degrees of freedom* means the number of linearly independent parameters.

Example:

i. If we fit a linear regression model using a pre-specified subset of k features without reference to the training data, the degrees of freedom used in the fitted model is defined to be k.

ii. If we carry out a best subset selection to determine the "optimal" set of k predictors, the resulting model has k parameters, but in some sense we have used up more than k degrees of freedom.

(b) *New Definition:* We define the *degrees of freedom* of the fitted value vector $\widehat{\mathbf{Y}} := (\hat{y}_1, \hat{y}_2, \cdots, \hat{y}_n)^\top$ to be

$$\mathrm{df}(\widehat{\mathbf{Y}}) := \frac{1}{\sigma^2} \sum_{i=1}^{n} \mathrm{Cov}(y_i, \hat{y}_i). \tag{31}$$

Here, $\mathrm{Cov}(y_i, \hat{y}_i)$ refers to the sampling covariance between the predicted value $\hat{y}_i$ and its corresponding outcome value y_i.

Intuitive Explanation: The harder that we fit to the data, the larger this covariance and hence $\mathrm{df}(\widehat{\mathbf{Y}})$.

Comment: One can be applied (31) to any model prediction $\widehat{\mathbf{Y}}$, including models that are adaptively fitted to the training data.

(c) *Examples:*

i. Linear regression model with fixed k variables: It is easy to obtain $\mathrm{df}(\widehat{\mathbf{Y}}) = k$, by noting

$$\begin{aligned}
\mathrm{df}(\widehat{\mathbf{Y}}) &= \frac{1}{\sigma^2} \mathrm{trace}(\mathrm{Cov}(\widehat{\mathbf{Y}}, \mathbf{Y})) \\
&= \frac{1}{\sigma^2} \mathrm{trace}\big(\mathrm{Cov}\big(\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{Y}, \mathbf{Y}\big)\big) \\
&= \frac{1}{\sigma^2} \mathrm{trace}\big(\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top \mathrm{Cov}(\mathbf{Y}, \mathbf{Y})\big) \\
&= \frac{1}{\sigma^2} \mathrm{trace}\big(\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top \sigma^2 \mathbf{I}_n\big) \\
&= \mathrm{trace}\big(\mathbf{X}^\top\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\big) \\
&= k.
\end{aligned}$$

ii. *Ridge regression:* The fitted value vector of the ridge regression for a fixed value of λ is

$$\widehat{\mathbf{Y}}^{\mathrm{ridge}} = \mathbf{X}(\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^\top\mathbf{Y}.$$

Then,

$$\begin{aligned}
\mathrm{df}(\widehat{\mathbf{Y}}^{\mathrm{ridge}}) &= \frac{1}{\sigma^2} \mathrm{trace}\big(\mathrm{Cov}(\widehat{\mathbf{Y}}^{\mathrm{ridge}}, \mathbf{Y})\big) \\
&= \frac{1}{\sigma^2} \mathrm{trace}\big(\mathrm{Cov}(\mathbf{X}(\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^\top\mathbf{Y}, \mathbf{Y})\big) \\
&= \frac{1}{\sigma^2} \mathrm{trace}\big(\mathbf{X}(\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^\top \mathrm{Cov}(\mathbf{Y}, \mathbf{Y})\big) \\
&= \mathrm{trace}\big(\mathbf{X}(\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^\top\big).
\end{aligned}$$

iii. *LAR:* After the k-th step of the LAR procedure, the effective degrees of freedom of the fit vector is exactly k.

iv. *Lasso:* Using Algorithm 3, at any stage $\mathrm{df}(\widehat{\mathbf{Y}})$ approximately equals the number of predictors in the model.

VI. Methods Using Derived Input Variables

1. **General Ideas of Methods Using Derived Input Variables:** Rather than using a large number of correlated covariates, use a small number of linear combinations, $\{Z_k\}_{k=1}^M$, of the original variables $\{X_j\}_{j=1}^p$, with $M < p$ or even $M \ll p$.

2. **Principal Components Regression:**

   (a) *Principal Components and Derived Variables:* Let $\mathbf{X}$ be the centered design matrix with the singular value decomposition $\mathbf{X} = \mathbf{U}\mathbf{D}\mathbf{V}^\top$. Then, the eigen-decomposition of $\mathbf{X}^\top\mathbf{X}$ is

   $$\mathbf{X}^\top\mathbf{X} = \mathbf{V}\mathbf{D}^2\mathbf{V}^\top.$$

   The eigenvectors $\mathbf{v}_1, \cdots, \mathbf{v}_p$, columns of $\mathbf{V}$, are called the *principal components directions* of $\mathbf{X}$. Form the derived variables $\mathbf{z}_1, \cdots, \mathbf{z}_p$ by

   $$\mathbf{z}_j = \mathbf{X}\mathbf{v}_j, \qquad \text{for all } j = 1, \cdots, p.$$

   (b) *Orthogonality of $\mathbf{z}_j$'s:* The vectors $\mathbf{z}_j$'s are orthogonal. To see this, let $j \neq k$ and notice

   $$\langle \mathbf{z}_j, \mathbf{z}_k \rangle = \mathbf{v}_j^\top\mathbf{X}^\top\mathbf{X}\mathbf{v}_k = \mathbf{v}_j^\top\mathbf{V}\mathbf{D}^2\mathbf{V}^\top\mathbf{v}_k = {\mathbf{I}_n^{(j)}}^\top\mathbf{D}^2\mathbf{I}_n^{(k)} = 0,$$

   and

   $$\langle \mathbf{z}_j, \mathbf{z}_j \rangle = \mathbf{v}_j^\top\mathbf{X}^\top\mathbf{X}\mathbf{v}_j = \mathbf{v}_j^\top\mathbf{V}\mathbf{D}^2\mathbf{V}^\top\mathbf{v}_j = {\mathbf{I}_n^{(j)}}^\top\mathbf{D}^2\mathbf{I}_n^{(j)} = d_j^2,$$

   where $\mathbf{I}_n^{(j)}$ denotes the j-th column of the $n \times n$ identity matrix $\mathbf{I}_n$.

   (c) *Principal Components Regression:* Principal components regression regress $\mathbf{Y}$ on $\mathbf{z}_1, \cdots, \mathbf{z}_M$ for some $M \leq p$ so that

   $$\widehat{\mathbf{Y}}^{\mathrm{pcr}} = \frac{1}{n}\mathbf{1}_n^\top\mathbf{Y}\mathbf{1}_n + \sum_{m=1}^M \hat{\theta}_m\mathbf{z}_m,$$

   where $\hat{\theta}_m = \langle \mathbf{z}_m, \mathbf{Y} \rangle / \langle \mathbf{z}_m, \mathbf{z}_m \rangle$.
   Since $\mathbf{z}_m$'s are linear combinations of columns of $\mathbf{X}$, then

   $$\sum_{m=1}^M \hat{\theta}_m\mathbf{z}_m = \sum_{m=1}^M \hat{\theta}_m\mathbf{X}\mathbf{v}_m = \mathbf{X}\sum_{m=1}^M \hat{\theta}_m\mathbf{v}_m.$$

   Hence, $\widehat{\boldsymbol{\beta}}^{\mathrm{pcr}}(M) = \sum_{m=1}^M \hat{\theta}_m\mathbf{v}_m$. Here, we use the notation "(M)" to explictly denote that the coefficient vector estimate depends on the number of principal components chosen.

(d) *Special Case When $M = p$:* If $M = p$,

$$\begin{aligned}\widehat{\boldsymbol{\beta}}^{\text{pcr}}(p) &= \sum_{m=1}^{p} \hat{\theta}_m \mathbf{v}_m = \sum_{m=1}^{p} \frac{\langle \mathbf{z}_m, \mathbf{Y} \rangle}{\langle \mathbf{z}_m, \mathbf{z}_m \rangle} \mathbf{v}_m \\ &= \sum_{m=1}^{p} \frac{\langle \mathbf{X}\mathbf{v}_m, \mathbf{Y} \rangle}{\langle \mathbf{X}\mathbf{v}_m, \mathbf{X}\mathbf{v}_m \rangle} \mathbf{v}_m \\ &= \sum_{m=1}^{p} \frac{\langle \mathbf{U}\mathbf{D}\mathbf{V}^\top \mathbf{v}_m, \mathbf{Y} \rangle}{\langle \mathbf{U}\mathbf{D}\mathbf{V}^\top \mathbf{v}_m, \mathbf{U}\mathbf{D}\mathbf{V}^\top \mathbf{v}_m \rangle} \mathbf{v}_m \\ &= \sum_{m=1}^{p} \frac{\mathbf{v}_m}{d_m} \mathbf{U}^\top \mathbf{Y} \\ &= \mathbf{V}\mathbf{D}^{-1}\mathbf{U}^\top \mathbf{Y}.\end{aligned}$$

But, on the other hand, we have the least squares estimate of $\boldsymbol{\beta}$, $\widehat{\boldsymbol{\beta}}^{\text{ls}}$, is

$$\widehat{\boldsymbol{\beta}}^{\text{ls}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{Y} = \mathbf{V}\mathbf{D}^{-1}\mathbf{U}^\top \mathbf{Y}.$$

That is, when $M = p$, the principal components estimator and the least squares estimator are identical. If $M < p$, we obtain a reduced regression.

(e) *Comparison of Principal Components Regression and Ridge Regression:*

- Both principal components regression and ridge regression operate via the principal components of the design matrix.
- Ridge regression shrinks the coefficients of the principal components depending on the size of the corresponding eigenvalue; principal components regression discards the $p - M$ smallest eigenvalue components.

3. Partial Least Squares:

(a) *Assumption:* Each column of the design matrix $\mathbf{X}$ is standardized to have mean 0 and norm 1.

(b) *Main Idea:* In partial least-squares regression, the derived variables are specifically constructed to retain most of the information in the predictors that helps predict $\mathbf{Y}$, while at the same time reducing the dimensionality of the regression. In particular, partial least-squares regression uses data on *both* the input and output variables.

(c) *Explanation:* PLS begins by computing $\widehat{\varphi}_{1,j} := \langle \mathbf{c}_j, \mathbf{Y} \rangle$ for each $j = 1, 2, \cdots, p$. Then, we construct the derived input

$$\mathbf{z}_1 := \sum_{j=1}^{p} \widehat{\varphi}_{1,j} \mathbf{c}_j,$$

which is the first partial least squares direction. Note that in the construction of each $\mathbf{z}_m$, the inputs are weighted by the *strength* (i.e., covariance) of their univariate effect on $\mathbf{Y}$. The outcome $\mathbf{Y}$ is regressed on $\mathbf{z}_1$, giving coefficient $\hat{\theta}_1$.

We orthogonalize $\mathbf{c}_1, \mathbf{c}_2, \cdots, \mathbf{c}_p$ with respect to $\mathbf{z}_1$. We continue this process, until $M \leq p$ directions have been obtained.

Remark. If we choose $M = p$, we get back to the least squares solution; if $M < p$, this procedure produces a reduced regression.

(d) *Algorithm:* The complete algorithm of the partial least squares is shown in Algorithm 4.

Algorithm 4 Partial Least Squares

1: Standardize the design matrix $\mathbf{X}$ so that each column has mean 0 and norm 1. Let $\mathbf{c}_1, \mathbf{c}_2, \cdots, \mathbf{c}_p$ be columns of $\mathbf{X}$;

2: Set $\widehat{\mathbf{Y}} = \frac{1}{n}\mathbf{1}_n^\top \mathbf{Y}\mathbf{1}_n$, and $\mathbf{c}_j^{(0)} = \mathbf{c}_j$ for all $j = 1, \cdots, p$;

3: For $m = 1, \cdots, M$ for some $M \leq p$:

i. $\mathbf{z}_m = \sum_{j=1}^p \widehat{\varphi}_{m,j}\mathbf{c}_j^{(m-1)}$, where $\widehat{\varphi}_{m,j} = \langle \mathbf{c}_j^{(m-1)}, \mathbf{Y}\rangle$;

ii. Regress $\mathbf{Y}$ on $\mathbf{z}_m$ and obtain the coefficient

$$\hat{\theta}_m = \frac{\langle \mathbf{z}_m, \mathbf{Y}\rangle}{\langle \mathbf{z}_m, \mathbf{z}_m\rangle};$$

iii. Set $\widehat{\mathbf{Y}}^{(m)} = \widehat{\mathbf{Y}}^{(m-1)} + \hat{\theta}_m \mathbf{z}_m$;

iv. Orthogonalize each $\mathbf{c}_j^{(m-1)}$ with respect to $\mathbf{z}_m$, i.e.,

$$\mathbf{c}_j^{(m)} = \mathbf{c}_j^{(m-1)} - \frac{\langle \mathbf{z}_m, \mathbf{c}_j^{(m-1)}\rangle}{\langle \mathbf{z}_m, \mathbf{z}_m\rangle}\mathbf{z}_m, \qquad \text{for all } j = 1, 2, \cdots, p.$$

4: Output the sequence of fitted vectors $\{\widehat{\mathbf{Y}}^{(m)}\}_{m=1}^M$. Since $\{\mathbf{z}^{(m)}\}_{m=1}^M$ are linear in the original $\mathbf{c}_j$, so is $\widehat{\mathbf{Y}}^{(m)} = \mathbf{X}\widehat{\boldsymbol{\beta}}^{\text{pls}}(m)$. The coefficients can be recovered from the sequence of the PLS transformation.

(e) *Some Remarks:*

i. Partial least squares seeks directions that have *high variance* and have *high correlation* with the response vector $\mathbf{Y}$.

ii. If the design matrix $\mathbf{X}$ is orthogonal, the partial least squares finds the least squares estimates after $m = 1$. Subsequent steps have no effect.

To see this, fix $m = 1$, let $\mathbf{c}_j^{(1)} = \mathbf{c}_j$, and note

$$\begin{aligned}
\mathbf{z}_1 &= \sum_{j=1}^{p} \langle \mathbf{c}_j, \mathbf{Y} \rangle \mathbf{c}_j, \\
\langle \mathbf{z}_1, \mathbf{Y} \rangle &= \left\langle \sum_{j=1}^{p} \langle \mathbf{c}_j, \mathbf{Y} \rangle \mathbf{c}_j, \mathbf{Y} \right\rangle = \sum_{j=1}^{p} \big(\langle \mathbf{c}_j, \mathbf{Y} \rangle \big)^2, \\
\langle \mathbf{z}_1, \mathbf{z}_1 \rangle &= \left\langle \sum_{j=1}^{p} \langle \mathbf{c}_j, \mathbf{Y} \rangle \mathbf{c}_j, \sum_{k=1}^{p} \langle \mathbf{c}_k, \mathbf{Y} \rangle \mathbf{c}_k \right\rangle \\
&= \sum_{j=1}^{p} \sum_{k=1}^{p} \langle \mathbf{c}_j, \mathbf{Y} \rangle \langle \mathbf{c}_k, \mathbf{Y} \rangle \langle \mathbf{c}_j, \mathbf{c}_k \rangle \\
&= \sum_{j=1}^{p} \big(\langle \mathbf{c}_j, \mathbf{Y} \rangle \big)^2,
\end{aligned}$$

where we use the assumption that columns of $\mathbf{X}$ are orthogonal and have unit norm to derive the last equality. It follows that $\hat{\theta}_1 = 1$. Then, for any $j = 1, 2, \cdots, m$, we have

$$\mathbf{c}_j^{(2)} = \mathbf{c}_j - \frac{\langle \mathbf{z}_1, \mathbf{c}_j \rangle}{\langle \mathbf{z}_1, \mathbf{z}_1 \rangle} \mathbf{z}_1,$$

where

$$\langle \mathbf{z}_1, \mathbf{c}_j \rangle = \left\langle \sum_{k=1}^{p} \langle \mathbf{c}_k, \mathbf{Y} \rangle \mathbf{c}_k, \mathbf{c}_j \right\rangle = \sum_{k=1}^{p} \langle \mathbf{c}_k, \mathbf{Y} \rangle \langle \mathbf{c}_k, \mathbf{c}_j \rangle = \langle \mathbf{c}_j, \mathbf{Y} \rangle.$$

Now, proceed to $m = 2$. Note that, for any $j = 1, \cdots, p$,

$$\begin{aligned}
\widehat{\varphi}_{2,j} &= \langle \mathbf{c}_j^{(2)}, \mathbf{Y} \rangle \\
&= \langle \mathbf{c}_j, \mathbf{Y} \rangle - \frac{\langle \mathbf{c}_j, \mathbf{Y} \rangle}{\sum_{j=1}^{p} (\langle \mathbf{c}_j, \mathbf{Y} \rangle)^2} \left\langle \sum_{k=1}^{p} \langle \mathbf{c}_k, \mathbf{Y} \rangle \mathbf{c}_k, \mathbf{Y} \right\rangle \\
&= \langle \mathbf{c}_j, \mathbf{Y} \rangle - \frac{\langle \mathbf{c}_j, \mathbf{Y} \rangle}{\sum_{j=1}^{p} (\langle \mathbf{c}_j, \mathbf{Y} \rangle)^2} \sum_{k=1}^{p} (\langle \mathbf{c}_k, \mathbf{Y} \rangle)^2 \\
&= \langle \mathbf{c}_j, \mathbf{Y} \rangle - \langle \mathbf{c}_j, \mathbf{Y} \rangle \\
&= 0.
\end{aligned}$$

4. A Summary of Selection and Shrinkage Methods:

- *Ridge regression* shrinks all directions, but shrinks the low-variance directions more;
- *Principal components regression* leaves M high-variance directions alone and discards the rest;

- *Partial least squares regression* tends to shrink the low-variance directions but inflate some of the higher variance directions, which makes it unstable and leads to higher prediction errors;
- *Lasso* falls between ridge regression and best subset selection.

VII. More on the Lasso and Related Path Algorithms

1. **Garotte:** Let $\widehat{\boldsymbol{\beta}}^{\text{ls}}$ be the least squares estimator and $\mathbf{W} := \text{diag}(\mathbf{w}) \in \mathbb{R}^{p\times p}$ be a diagonal matrix with nonnegative weights $\mathbf{w} := (w_1, w_2, \cdots, w_p)^\top \in \mathbb{R}^p$ along the diagonal. The *Garotte* finds $\mathbf{w}$ that minimizes

$$\frac{1}{2}\big(\mathbf{Y} - \mathbf{X}\mathbf{W}\widehat{\boldsymbol{\beta}}^{\text{ls}}\big)^\top\big(\mathbf{Y} - \mathbf{X}\mathbf{W}\widehat{\boldsymbol{\beta}}^{\text{ls}}\big), \tag{32}$$

subject to one of the following two constraints:

(a) nonnegative Garotte: $\mathbf{w} \geq \mathbf{0}_p$ and $\mathbf{1}_p^\top\mathbf{w} = \sum_{j=1}^p w_j \leq c$,

(b) Garotte: $\mathbf{w}^\top\mathbf{w} \leq c$.

Either version seeks to find some desirable scaling of the regression coefficients.

In particular, as c is decreased, more components of $\mathbf{w}$ become 0 (thus eliminating those particular variables from the regression function), while the nonzero $\widehat{\boldsymbol{\beta}}^{\text{ls}}$ shrink toward 0.

Special Case — Orthogonal Design: Suppose that columns of $\mathbf{X}$ are orthogonal and that c is in the range where the equality $\sum_{j=1}^p w_j = c$ can be satisfied, the solution to $\mathbf{w}$ in (32) is given by

$$\hat{w}_j = \left(1 - \frac{\lambda}{(\hat{\beta}_j^{\text{ls}})^2}\right)_+, \qquad \text{for all } j = 1, 2, \cdots, p,$$

where λ is chosen so that $\sum_{j=1}^p \hat{w}_j = c$. Hence, if $\hat{\beta}_j^{\text{ls}}$ is large, the shrinkage factor will be close to 1, leading to no shrinkage; if it is small, the coefficient will be shrunken to 0.

Remark. Both versions of the Garotte depend upon the existence of $\widehat{\boldsymbol{\beta}}^{\text{ls}}$ and fail in situations where $p > n$.

2. **Incremental Forward Stagewise Regression:**

(a) FS_ε *Algorithm:*

Algorithm 5 Incremental Forward Stagewise Regression – FS_ε

1: Start with the residual vector $\mathbf{r}$ equal to $\mathbf{Y}$ and $\beta_1, \beta_2, \cdots, \beta_p = 0$. All the predictors are standardized to have mean zero and unit norm.
2: Find the predictor X_j most correlated with $\mathbf{r}$;
3: Update

$$\beta_j \leftarrow \beta_j + \delta_j,$$

where $\delta_j = \varepsilon \cdot \text{sign}(\langle \mathbf{x}_j, \mathbf{r} \rangle)$ and $\varepsilon > 0$ is a small step size, and set

$$\mathbf{r} \leftarrow \mathbf{r} - \delta_j \mathbf{x}_j.$$

4: Repeat Steps 2 and 3 many times, until the residuals are uncorrelated with all the predictors.

(b) *Infinitesimal Forward Stagewise Regression,* FS_0*:*

Algorithm 6 Least Angle Regression – FS_0 Modification

1: Standardize the variables to have mean zero and unit norm.
2: Start with the residual

$$\mathbf{r} := \mathbf{Y} - \frac{1}{n}\mathbf{1}_n^\top \mathbf{Y}$$

and $\beta_1 = \beta_2 = \cdots = \beta_p = 0$.
3: Find the predictor X_j most correlated with $\mathbf{r}$.
4: Move β_j from 0 towards its least-squares coefficient $\langle X_j, \mathbf{r} \rangle$, until some other competitor X_k has as much correlation with the current residual as does X_j.
5: Find the new direction by solving the following constrained least squares problem

$$\begin{aligned} &\underset{b}{\text{minimize}} \, \|\mathbf{r} - \mathbf{X}_{\mathcal{A}} b\|_2^2, \\ &\text{subject to } b_j s_j \geq 0, j \in \mathcal{A}, \end{aligned}$$

where $s_j = \text{sign}(\langle \mathbf{c}_j, \mathbf{r} \rangle)$.

(Texts in red highlight the difference with Algorithm 2.)

6: Continue in this way until all p predictors have been entered. After $\min(n-1, p)$ steps, we arrive at the full least squares solution.

The modification occurs at Step 5, which amounts to solving a non-negative least squares regression problem.

3. **Piecewise-Linear Path Algorithms:** Consider the general problem

$$\widehat{\boldsymbol{\beta}}(\lambda) := \arg\min_{\boldsymbol{\beta}} \Big\{ R(\boldsymbol{\beta}) + \lambda J(\boldsymbol{\beta}) \Big\},$$

where

$$R(\boldsymbol{\beta}) := \sum_{i=1}^{n} L\bigg(y_i, \beta_0 + \sum_{j=1}^{p} \beta_j x_{i,j}\bigg),$$

and both the loss function L and the penalty function J are convex.

Then, if the following two conditions are satisfied, the solution path $\widehat{\boldsymbol{\beta}}(\lambda)$ is piecewise linear:

(a) R is quadratic or piecewise-quadratic as a function of $\boldsymbol{\beta}$, and

(b) J is piecewise linear in $\boldsymbol{\beta}$.

4. **Dantzig Selector:**

(a) *Problem Formulation:* The *Dantzig Selector* solves the following minimization problem

$$\begin{aligned} &\underset{\boldsymbol{\beta}}{\text{minimize}} \, \|\boldsymbol{\beta}\|_1 \\ &\text{subject to } \|\mathbf{X}^\top(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})\|_\infty \le s. \end{aligned} \tag{33}$$

Equivalently, (33) can be written as

$$\begin{aligned} &\underset{\boldsymbol{\beta}}{\text{minimize}} \, \|\mathbf{X}^\top(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})\|_\infty \\ &\text{subject to } \|\boldsymbol{\beta}\|_1 \le t. \end{aligned} \tag{34}$$

In the formulations above, $\|\cdot\|_\infty$ denotes the maximum absolute value of the components of the vector.

The problem (34) can be solved using the linear programming.

(b) *Similarity to Lasso:* Note that (34) is very similar to the Lasso problem (19). The difference is that (34) replaces the squared error loss by the maximum absolute value of its gradient.

(c) *Remarks:*

i. In (34), as t gets large, both Dantzig selector and Lasso yield the least squares solution if $n > p$;

ii. If $p \ge N$, both Dantzig selector and Lasso yield the least squares solution with minimum L_1 norm.

iii. Dantzig selector tries to minimize the maximum inner product of the current residual with all the predictors. Hence, it can achieve a *smaller* maximum than the lasso.

5. **Grouped Lasso:**

(a) *Motivation:* In some problems, the predictors belong to pre-defined groups. In this situation, it may be desirable to shrink and select the members of a group *together*. The grouped lasso is one way to achieve this.

(b) *Setup:* Suppose that the p predictors are divided into L groups, with p_ℓ predictors in Group ℓ. We use a matrix $\mathbf{X}_\ell \in \mathbb{R}^{n\times p_\ell}$ to represent the predictors corresponding to the ℓ-th group, with corresponding coefficient vector $\boldsymbol{\beta}_\ell \in \mathbb{R}^{p_\ell}$.

(c) *Problem Formulation:* The grouped Lasso solves the following optimization problem

$$\underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\text{minimize}} \left\{ \frac{1}{2}\left\| \mathbf{Y} - \sum_{\ell=1}^{L} \mathbf{X}_\ell \boldsymbol{\beta}_\ell \right\|_2^2 + \lambda \sum_{\ell=1}^{L} \sqrt{p_\ell}\|\boldsymbol{\beta}_\ell\|_2 \right\}, \tag{35}$$

where the p_ℓ term accounts for the varying group sizes and we center $\mathbf{Y}$ and each column of $\mathbf{X}$ so that there is no intercept term.

Remark 1. Since the Euclidean norm of a vector $\boldsymbol{\beta}_\ell$ is zero if and only if *all* of its components are zero, this procedure encourages sparsity at both the group and individual levels. That is, for some values of λ, an *entire* group of predictors may drop out of the model.

Remark 2. If $L = p$ and $p_\ell = 1$ for all $\ell = 1, 2, \cdots, L$ so that each individual predictor forms a single group, (35) becomes the ordinary lasso problem.

(d) *Computation:* Differentiating (35) with respect to $\boldsymbol{\beta}_\ell$ for all $\ell = 1, 2, \cdots, L$ and setting the derivatives to zero yield

$$-\mathbf{X}_\ell^\top \left(\mathbf{Y} - \sum_{\ell=1}^{L} \mathbf{X}_\ell \widehat{\boldsymbol{\beta}}_\ell \right) + \lambda\sqrt{p_\ell}\hat{\mathbf{s}}_\ell = \mathbf{0}_{p_\ell}, \qquad \text{for all } \ell = 1, \cdots, L, \tag{36}$$

where $\hat{\mathbf{s}}_\ell \in \mathbb{R}^{p_\ell}$ is the subgradient of the norm $\|\cdot\|_2$ evaluated at $\widehat{\boldsymbol{\beta}}_\ell$ and is given by

$$\hat{\mathbf{s}}_\ell = \begin{cases} \widehat{\boldsymbol{\beta}}_\ell / \|\widehat{\boldsymbol{\beta}}_\ell\|_2, & \text{if } \widehat{\boldsymbol{\beta}}_\ell \neq \mathbf{0}_{p_\ell}, \\ \text{any vector satisfying } \|\widehat{\boldsymbol{\beta}}_\ell\|_2 \leq 1, & \text{if } \widehat{\boldsymbol{\beta}}_\ell = \mathbf{0}_{p_\ell}, \end{cases} \tag{37}$$

and $(\widehat{\boldsymbol{\beta}}_1^\top, \cdots, \widehat{\boldsymbol{\beta}}_L^\top)^\top$ denote the minimizer of (35) with respect to $(\boldsymbol{\beta}_1^\top, \cdots, \boldsymbol{\beta}_L^\top)^\top$. One approach to compute $(\widehat{\boldsymbol{\beta}}_1^\top, \cdots, \widehat{\boldsymbol{\beta}}_L^\top)^\top$ is to use the *block coordinate descent algorithm* — hold all block vectors $\{\widehat{\boldsymbol{\beta}}_k\}_{k\neq\ell}$ fixed and solve for $\widehat{\boldsymbol{\beta}}_\ell$. With all $\{\widehat{\boldsymbol{\beta}}_k\}_{k\neq\ell}$ fixed, we can rewrite (36) as

$$-\mathbf{X}_\ell^\top \left(\mathbf{r}^{(\ell)} - \mathbf{X}_\ell \widehat{\boldsymbol{\beta}}_\ell \right) + \lambda\sqrt{p_\ell}\hat{\mathbf{s}}_\ell = \mathbf{0}_{p_\ell},$$

where $\mathbf{r}^{(\ell)} = \mathbf{Y} - \sum_{k\neq\ell} \mathbf{X}_k \widehat{\boldsymbol{\beta}}_k$ is the ℓ-th partial residual.

From (37), we must have $\widehat{\boldsymbol{\beta}}_\ell = \mathbf{0}_{p_\ell}$ if $\|\mathbf{X}_\ell^\top \mathbf{r}^{(\ell)}\|_2 < \lambda$, and, otherwise, $\widehat{\boldsymbol{\beta}}_\ell$ must satisfy

$$\widehat{\boldsymbol{\beta}}_\ell = \left(\mathbf{X}_\ell^\top \mathbf{X}_\ell + \frac{\lambda\sqrt{p_\ell}}{\|\widehat{\boldsymbol{\beta}}_\ell\|_2} \mathbf{I}_{p_\ell} \right)^{-1} \mathbf{X}_\ell^\top \mathbf{r}^{(\ell)}, \tag{38}$$

which is similar to the solution to the ridge regression.

Remark. Note that (38) depends on $\|\widehat{\boldsymbol{\beta}}_\ell\|_2$ and is not directly applicable to compute $\widehat{\boldsymbol{\beta}}_\ell$. But, when $\mathbf{X}_\ell$ is orthonormal, we have

$$\widehat{\boldsymbol{\beta}}_\ell = \left(1 - \frac{\lambda\sqrt{p_\ell}}{\|\mathbf{X}_\ell^\top \mathbf{r}^{(\ell)}\|_2}\right)_+^{-1} \mathbf{X}_\ell^\top \mathbf{r}^{(\ell)}.$$

6. Sparse Grouped Lasso:

(a) *Motivation:* When a group is included in a group lasso, *all* the coefficients in that group are nonzero, which is a consequence of the L_2 norm. Sometimes, we want sparsity with respect to both which groups are selected and which coefficients are nonzero within a group. The *sparse grouped lasso* is designed to achieve the *within*-group sparsity.

(b) *Problem Formulation:* In the sparse grouped lasso, we consider the following optimization problem

$$\underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\text{minimize}} \left\{ \frac{1}{2}\left\|\mathbf{Y} - \sum_{\ell=1}^{L} \mathbf{X}_\ell \boldsymbol{\beta}_\ell\right\|_2^2 + \lambda \sum_{\ell=1}^{L} \left[(1-\alpha)\|\boldsymbol{\beta}_\ell\|_2 + \alpha\|\boldsymbol{\beta}_\ell\|_1\right] \right\}, \tag{39}$$

where $\alpha \in [0,1]$ and we center $\mathbf{Y}$ and each column of $\mathbf{X}$ so that there is no intercept term. Note that when $\alpha = 0$, (39) becomes the grouped lasso problem, and when $\alpha = 1$, (39) becomes the ordinary lasso problem.

Remark. Note that (39) is a convex problem.

(c) *Optimality Condition:* Differentiating (39) with respect to $\boldsymbol{\beta}_\ell$ and setting the derivatives to zero yield

$$-\mathbf{X}_\ell^\top \left(\mathbf{Y} - \sum_{\ell=1}^{L} \mathbf{X}_\ell \widehat{\boldsymbol{\beta}}_\ell\right) + \lambda(1-\alpha)\hat{\mathbf{s}}_\ell + \lambda\alpha\hat{\mathbf{t}}_\ell = \mathbf{0}_{p_\ell}, \quad \text{for all } \ell = 1, 2, \cdots, L, \tag{40}$$

where $\hat{\mathbf{s}}_\ell \in \mathbb{R}^{p_\ell}$ belongs to the subdifferential of $\|\cdot\|_2$ evaluated at $\widehat{\boldsymbol{\beta}}_\ell$ and $\hat{\mathbf{t}}_\ell \in \mathbb{R}^{p_\ell}$ belongs to the subdifferential of $\|\cdot\|_1$ evaluated at $\widehat{\boldsymbol{\beta}}_\ell$.

(d) *Computation:* We can use the *block coordinate descent algorithm* to solve (39). Let

$$\mathbf{r}^{(\ell)} = \mathbf{Y} - \sum_{k\neq\ell} \mathbf{X}_k \widehat{\boldsymbol{\beta}}_k$$

be the ℓ-th partial residual. Then, $\widehat{\boldsymbol{\beta}}_\ell = \mathbf{0}_{p_\ell}$ if and only if the equation

$$\mathbf{X}_\ell^\top \mathbf{r}^{(\ell)} = \lambda(1-\alpha)\hat{\mathbf{s}}_\ell + \alpha\hat{\mathbf{t}}_\ell$$

has a solution with $\|\hat{\mathbf{s}}_\ell\|_2 \leq 1$ and each component of $\hat{\mathbf{t}}_\ell$ is between -1 and 1, inclusively, and if and only if

$$\|S(\mathbf{X}_\ell^\top \mathbf{r}^{(\ell)}, \lambda\alpha)\|_2 \leq \lambda(1-\alpha),$$

where $S(\cdot, \lambda\alpha)$ is the soft-thresholding operator applied to each component of $\mathbf{X}_\ell^\top \mathbf{r}^{(\ell)}$.

After we have checked $\widehat{\boldsymbol{\beta}}_\ell \neq \mathbf{0}_{p_\ell}$, finding $\widehat{\boldsymbol{\beta}}_\ell$ amounts to solving the following subproblem

$$\underset{\boldsymbol{\beta}_\ell \in \mathbb{R}^{p_\ell}}{\text{minimize}} \left\{ \frac{1}{2}\|\mathbf{r}^{(\ell)} - \mathbf{X}_\ell \boldsymbol{\beta}_\ell\|_2^2 + \lambda(1-\alpha)\|\boldsymbol{\beta}_\ell\|_2 + \lambda\alpha\|\boldsymbol{\beta}_\ell\|_1 \right\}. \tag{41}$$

The problem (41) can be solved by iterating

$$\begin{aligned} \boldsymbol{\omega} &\leftarrow \boldsymbol{\beta}_\ell + \nu \mathbf{X}_\ell^\top (\mathbf{r}^{(\ell)} - \mathbf{X}_\ell \boldsymbol{\beta}_\ell), \\ \boldsymbol{\beta}_\ell &\leftarrow \left(1 - \frac{\nu\lambda(1-\alpha)}{\|S(\boldsymbol{\omega}, \lambda\alpha)\|_2}\right)_+ S(\boldsymbol{\omega}, \lambda\alpha), \end{aligned}$$

until convergence, where $\nu > 0$ is the step size.

7. Relaxed Lasso:

(a) *Main Idea:* Use cross-validation to estimate the initial penalty parameter for the lasso, and then again for a second penalty parameter applied to the selected set of predictors.

(b) *Motivation:* Since the variables in the second step have less "competition" from noise variables, cross-validation will tend to pick a smaller value for λ, and hence their coefficients will be shrunken *less* than those in the initial estimate.

(c) *Procedure:*

i. Use the Lasso to select the set of non-zero predictors; and

ii. Apply the lasso again, but using only the selected predictors from the first step.

8. Adaptive Lasso: The *adaptive lasso* uses a weighted penalty of the form

$$\sum_{j=1}^{p} w_j |\beta_j|, \qquad \text{where } w_j = \frac{1}{|\hat{\beta}_j|^\nu}, \tag{42}$$

and $\hat{\beta}_j$ is the ordinary least squares estimate and $\nu > 0$.

Remarks.

(a) The adaptive Lasso (42) is a practical approximation to the $|\boldsymbol{\beta}|^q$ penalties ($q = 1 - \nu$ here);

(b) The adaptive lasso yields consistent estimates of the parameters while retaining the attractive convexity property of the lasso.

9. Fused Lasso:

(a) *Motivation:* In real applications, we assume that the regression function is piecewise-constant over contiguous regions.

(b) *Problem Formulation:* The fused lasso solves the following optimization problem

$$\underset{\boldsymbol{\theta}\in\mathbb{R}^n}{\text{minimize}}\left\{\frac{1}{2}\sum_{i=1}^{n}(y_i-\theta_i)^2+\lambda_1\sum_{i=1}^{n}|\theta_i|+\lambda_2\sum_{i=2}^{n}|\theta_i-\theta_{i-1}|\right\}, \tag{43}$$

where both λ_1 and λ_2 are nonnegative. Note there are two penalization terms:

i. The first penalty term, $\lambda_1\sum_{i=1}^{n}|\theta_i|$ is the usual L_1 penalty and attempts to shrinks each individual θ_i toward 0, and

ii. the second penalty term, $\lambda_2\sum_{i=2}^{n}|\theta_i-\theta_{i-1}|$, encourages neighboring coefficient θ_i to be similar and will cause some to be identical.

This second penalty term is known as the *total-variation denoising*.

(c) *Extension:* A possible extension of the fused lasso problem in (43) is

$$\underset{(\beta_0,\boldsymbol{\beta}^\top)^\top}{\text{minimize}}\left\{\frac{1}{2}\sum_{i=1}^{n}\left(y_i-\beta_0-\sum_{j=1}^{p}\beta_j x_{i,j}\right)^2+\lambda_1\sum_{j=1}^{p}|\beta_j|+\lambda_2\sum_{j=2}^{p}|\beta_j-\beta_{j-1}|\right\}.$$

(d) *Fitting the Fused Lasso:* Note that (43) is *not* a separable function of $\theta_1,\cdots,\theta_n$. Hence, the coordinate descent algorithm is *not* applicable.

Let $\widehat{\boldsymbol{\theta}}(\lambda_1,\lambda_2)$ be the optimal solution of (43). We have the following lemma:

Lemma 1. *For any $\lambda_1'>\lambda_1$, we have*

$$\widehat{\theta}_i(\lambda_1',\lambda_2)=S\big(\widehat{\theta}_i(\lambda_1,\lambda_2),\lambda_1'-\lambda_1\big),\qquad \textit{for all } i=1,2,\cdots,n,$$

where S is the soft-thresholding operator $S(x,\lambda):=\text{sign}(x)(|x|-\lambda)_+$.

One special case is the following

$$\widehat{\theta}_i(\lambda_1,\lambda_2)=S\big(\widehat{\theta}_i(0,\lambda_2),\lambda_1\big).$$

Consequently, if we solve the fused lasso problem with $\lambda_1=0$, all other solutions can be obtained immediately by soft thresholding.

Hence, we consider the problem in the sequel

$$\underset{\boldsymbol{\theta}\in\mathbb{R}^n}{\text{minimize}}\left\{\frac{1}{2}\sum_{i=1}^{n}(y_i-\theta_i)^2+\lambda_2\sum_{i=2}^{n}|\theta_i-\theta_{i-1}|\right\}. \tag{44}$$

(e) *Reparametrization:* Let $\boldsymbol{\gamma} = \mathbf{M}\boldsymbol{\theta}$ for an invertible matrix $\mathbf{M} \in \mathbb{R}^{n\times n}$ such that

$$\gamma_1 = \theta_1, \qquad \text{and} \qquad \gamma_i = \theta_i - \theta_{i-1} \text{ for all } i = 2, \cdots, n.$$

Then, we can rewrite (44) as

$$\underset{\boldsymbol{\gamma}}{\text{minimize}} \left\{ \frac{1}{2}\|\mathbf{Y} - \mathbf{M}^{-1}\boldsymbol{\gamma}\|_2^2 + \lambda\|\boldsymbol{\gamma}\|_1 \right\}. \tag{45}$$

The *benefit* of introducing this reparametrization is that the penalty term is now additive.

We can solve (45) by coordinate descent algorithm or projected gradient descent algorithm.

(f) *Dual Path Algorithm:* We can rewrite (44) as

$$\begin{aligned} \underset{(\boldsymbol{\gamma},\mathbf{z})\in\mathbb{R}^{n\times(n-1)}}{\text{minimize}} & \left\{ \frac{1}{2}\|\mathbf{Y} - \boldsymbol{\theta}\|_2^2 + \lambda\|\mathbf{z}\|_1 \right\}, \\ & \text{subject to } \mathbf{D}\boldsymbol{\theta} = \mathbf{z}, \end{aligned} \tag{46}$$

where $\mathbf{D} \in \mathbb{R}^{(n-1)\times n}$ is the matrix of the first differences given by

$$\mathbf{D} = \begin{pmatrix} -1 & 1 & 0 & 0 & \cdots & 0 & 0 \\ 0 & -1 & 1 & 0 & \cdots & 0 & 0 \\ 0 & 0 & -1 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & -1 & 1 \end{pmatrix}$$

and $\mathbf{z} \in \mathbb{R}^{n-1}$ is an auxlliary variable.

The Lagrangian function associated with (46) is

$$L(\boldsymbol{\theta}, \mathbf{z}; \mathbf{u}) := \frac{1}{2}\|\mathbf{Y} - \boldsymbol{\theta}\|_2^2 + \lambda\|\mathbf{z}\|_1 + \mathbf{u}^\top(\mathbf{D}\boldsymbol{\theta} - \mathbf{z}),$$

where $\mathbf{u} \in \mathbb{R}^{n-1}$ is a vector of Lagrangian multipliers. Then, the Lagrangian dual function is

$$Q(\mathbf{u}) := \inf_{(\boldsymbol{\gamma},\mathbf{z})\in\mathbb{R}^{n\times(n-1)}} L(\boldsymbol{\theta}, \mathbf{z}; \mathbf{u}) = \begin{cases} -\frac{1}{2}\|\mathbf{Y} - \mathbf{D}^\top\mathbf{u}\|_2^2, & \text{if } \|\mathbf{u}\|_\infty \le \lambda, \\ -\infty, & \text{otherwise.} \end{cases}$$

The Lagrangian dual problem is to maximize Q with respect to $\mathbf{u}$; given an optimal $\widehat{\mathbf{u}} := \arg\max_{\mathbf{u}} Q(\mathbf{u})$, we can recover an optimal solution to $\boldsymbol{\theta}$ as $\widehat{\boldsymbol{\theta}} = \mathbf{Y} - \mathbf{D}^\top\widehat{\mathbf{u}}$.

10. Nearly Isotonic Regression:

(a) *Review of the Classic Isotonic Regression:* The *classic isotonic regression* solves the following optimization problem

$$\begin{aligned} \underset{\boldsymbol{\theta}\in\mathbb{R}^n}{\text{minimize}} & \left\{ \sum_{i=1}^n (y_i - \theta_i)^2 \right\} \\ & \text{subject to } \theta_1 \le \theta_2 \le \cdots \le \theta_n. \end{aligned} \tag{47}$$

The resulting solution to (47) gives the best non-increasing fit to the data. The solution to (47) is unique and can be obtained by using the *pool adjacent violators algorithm*.

(b) *Problem Formulation for Nearly Isotonic Regression:* The *nearly isotonic regression* is a relaxation of the classic isotonic regression problem and solves the following problem

$$\underset{\boldsymbol{\theta}\in\mathbb{R}^n}{\text{minimize}} \left\{ \sum_{i=1}^{n} (y_i - \theta_i)^2 + \lambda \sum_{i=1}^{n} (\theta_i - \theta_{i+1})_+ \right\}. \tag{48}$$

(c) *How the Penalty Term in (48) Works:* Note that the penalty term penalizes adjacent pairs that violate the monotonicity property, i.e., $\theta_i > \theta_{i+1}$. Note that

i. as $\lambda = 0$, the solution interpolates the data, and

ii. as $\lambda \to \infty$, we recover the the solution to the classic isotonic regression problem (47).

Intermediate values of λ yield non-monotone solutions that trade off monotonicity with goodness of fit.

References

Hastie, Trevor, Robert Tibshirani, and Jerome Friedman (2009). *The Elements of Statistical Learning*. Vol. 1. Springer Series in Statistics. New York, NY, USA: Springer New York Inc.

Hastie, Trevor, Robert Tibshirani, and Martin Wainwright (2015). *Statistical Learning with Sparsity: The Lasso and Generalizations*. Chapman & Hall/CRC. ISBN: 1498712169.

Izenman, Alan J (Mar. 2009). *Modern Multivariate Statistical Techniques: Regression, Classification, and Manifold Learning*. en. Springer Science & Business Media.