Ch 3: Linear Methods for Regression

Advanced Statistical Data Mining

Jaejik Kim

Department of Statistics
Sungkyunkwan University

Fall 2020

Linear Regression Models

- The linear regression model: $f(X) = \beta_0 + \sum_{j=1}^{p} X_j \beta_j$.
- f: Regression function $\mathrm{E}[Y|X]$.
- Sources of inputs X_j:
  - qualitative inputs,
  - transformations of quantitative inputs,
  - basis expansions leading to polynomial representation,
  - numeric or "dummy" coding of qualitative inputs,
  - interactions between variables such as $X_3 = X_1 \cdot X_2$.
- The model is linear in the parameters $\beta_j,\ j = 1, \ldots, p$.

Least Squares Estimation

- The least squares method minimizes the residual sum of squares (RSS)
  $RSS(\beta) = \sum_{i=1}^{N}(y_i - f(x_i))^2 = (\boldsymbol{y} - \boldsymbol{X}\beta)^\top(\boldsymbol{y} - \boldsymbol{X}\beta)$
- $\hat{\beta} = (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{y}$
- Fitted values of training inputs: $\hat{\boldsymbol{y}} = \boldsymbol{X}\hat{\beta} = \boldsymbol{H}\boldsymbol{y}$
- $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top$ (*hat matrix*).
- $\hat{\boldsymbol{y}}$: The orthogonal projection of $\boldsymbol{y}$ onto the column space of $\boldsymbol{X}$.

Orthogonal Projection

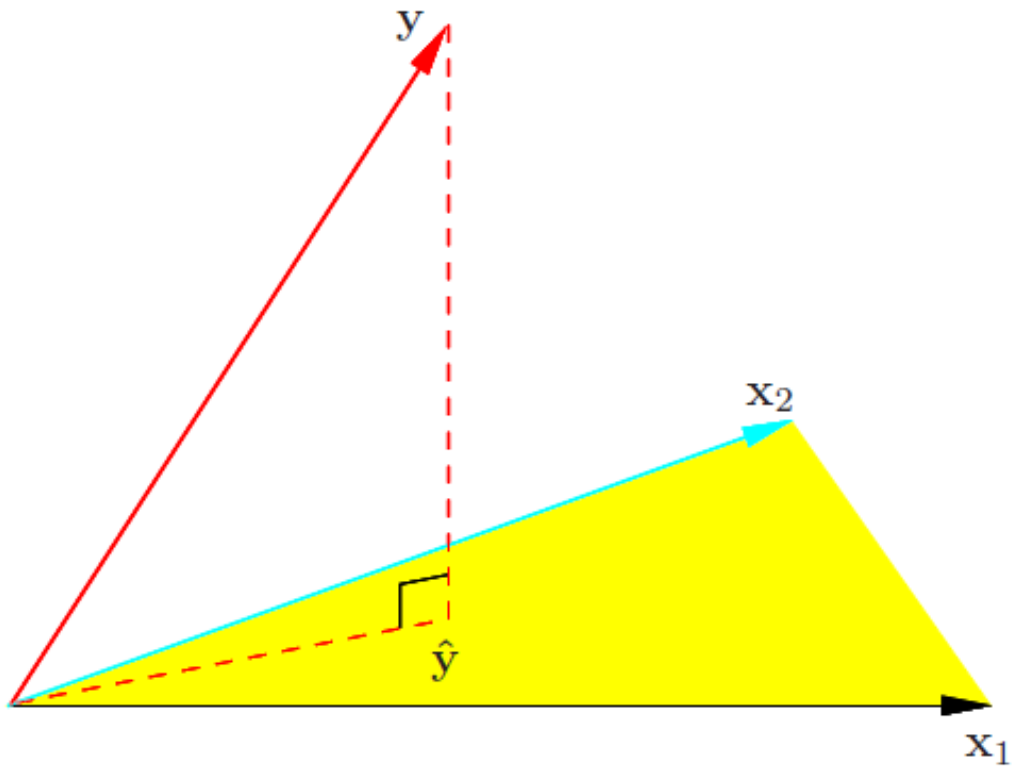


FIGURE 3.2. *The N-dimensional geometry of least squares regression with two predictors. The outcome vector* $\mathbf{y}$ *is orthogonally projected onto the hyperplane spanned by the input vectors* $\mathbf{x}_1$ *and* $\mathbf{x}_2$*. The projection* $\hat{\mathbf{y}}$ *represents the vector of the least squares predictions*

Basic Least Squares Assumptions

- So far, no distributional assumptions have been made.
- Now we make some assumptions:
  - x_i's are fixed.
  - $\mathrm{Cov}(Y_i, Y_j) = \begin{cases} \sigma^2 & \text{if } i = j \\ 0 & \text{if } i \neq j \end{cases}$
- $\hat{\beta}$: Unbiased estimator of β; i.e., $\mathrm{E}[\hat{\beta}] = \beta$.
- $\mathrm{Var}(\hat{\beta}) = (\boldsymbol{X}^\top \boldsymbol{X})^{-1}\sigma^2$
- $\hat{\sigma}^2 = \dfrac{RSS(\hat{\beta})}{N - p - 1}$; $\mathrm{E}[\hat{\sigma}^2] = \sigma^2$

Multivariate Expectations and Variances

- $\mathrm{E}\left[\sum_{i=1}^{N} a_i Y_i\right] = \sum_{i=1}^{N} a_i \mathrm{E}\left[Y_i\right] \Longleftrightarrow \mathrm{E}[\boldsymbol{a}^\top \boldsymbol{Y}] = \boldsymbol{a}^\top \mathrm{E}[\boldsymbol{Y}]$
- $\mathrm{Var}\left[\sum_{i=1}^{N} a_i Y_i\right] = \sum_{i=1}^{N} \sum_{j=1}^{N} a_i a_j \mathrm{Cov}[Y_i, Y_j]$
  $\Longleftrightarrow \mathrm{Var}[\boldsymbol{a}^\top \boldsymbol{Y}] = \boldsymbol{a}^\top \mathrm{Var}[\boldsymbol{Y}]\boldsymbol{a}$
- $\mathrm{Cov}\left[\sum_{i=1}^{N} a_i Y_i, \sum_{j=1}^{N} b_i Y_i\right] = \sum_{i=1}^{N} \sum_{j=1}^{N} a_i b_j \mathrm{Cov}[Y_i, Y_j]$
  $\Longleftrightarrow \mathrm{Cov}[\boldsymbol{a}^\top \boldsymbol{Y}, \boldsymbol{b}^\top \boldsymbol{Y}] = \boldsymbol{a}^\top \mathrm{Var}[\boldsymbol{Y}]\boldsymbol{b}$
- $\mathrm{E}[\boldsymbol{A}\boldsymbol{Y}] = \boldsymbol{A}\ \mathrm{E}[\boldsymbol{Y}]$
- $\mathrm{Var}[\boldsymbol{A}\boldsymbol{Y}] = \boldsymbol{A}\ \mathrm{Var}[\boldsymbol{Y}]\ \boldsymbol{A}^\top$
- $\mathrm{Cov}[\boldsymbol{A}\boldsymbol{Y}, \boldsymbol{B}\boldsymbol{Y}] = \boldsymbol{A}\ \mathrm{Var}[\boldsymbol{Y}]\ \boldsymbol{B}^\top$

Gauss-Markov Theorem

- Consider a linear combination of the form $a^\top \hat{\beta}$ for an arbitrary $a \in \mathbb{R}^{p+1}$.
- LSE of $a^\top \beta = a^\top \hat{\beta} = a^\top (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{y}$.
- $\mathrm{E}[a^\top \hat{\beta}] = a^\top \beta$
- *Gauss-Markov Theorem*: The LSE $a^\top \hat{\beta}$ has variance no bigger than that of any other linear unbiased estimator $\tilde{\theta} = \boldsymbol{c}^\top \boldsymbol{Y}$ that is unbiased for $a^\top \beta$.
- $\mathrm{MSE}(\tilde{\theta}) = \mathrm{E}(\tilde{\theta} - \theta)^2 = \mathrm{Var}(\tilde{\theta}) + [\mathrm{E}(\tilde{\theta}) - \theta]^2 \Rightarrow$ Gauss-Markov Theorem implies that the LSE has the smallest MSE (variance) of all unbiased linear estimators.
- Conversely, there may exist biased estimators which have smaller MSE than the LSE.

Inferential Assumptions

- Inference $\Rightarrow$ Stronger assumptions:
  - $Y = E(Y|X_1, \ldots, X_p) + \epsilon = \beta_0 + \sum_{j=1}^{p} X_j \beta_j + \epsilon$
  - $\epsilon \sim \text{Normal}(0, \sigma^2)$
- $\hat{\beta} \sim \text{Normal}(\beta, (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \sigma^2)$
- $\dfrac{(N - p - 1)\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{N-p-1}$
- $\hat{\beta}$ & $\hat{\sigma}^2$ are independent

Hypothesis Testing

- Since the marginal distribution of a multivariate Normal is a univariate Normal, $\hat{\beta}_j \sim \text{Normal}(\beta_j, \sigma^2 v_j)$ where v_j is the jth diagonal element of $(\boldsymbol{X}^\top \boldsymbol{X})^{-1}$.
- $\dfrac{\hat{\beta}_j - \beta_j}{\sigma\sqrt{v_j}} \sim \text{Normal}(0, 1)$
- $z_j = \dfrac{\hat{\beta}_j - \beta_j}{\hat{\sigma}\sqrt{v_j}} \sim t_{N-p-1}$
- Reject H_0 from $H_0 : \beta_j = 0$ vs. $H_a : \beta_j \neq 0$ if $|z_j| > t_{N-p-1,\ (1-\alpha/2)}$.

Hypothesis Testing

- Test for a group of coefficients:
$$F = \frac{(RSS_0 - RSS_1)/(p_1 - p_0)}{RSS_1/(N - p_1 - 1)} \sim F_{p_1 - p_0, N - p_1 - 1}.$$

- RSS_1: The RSS for the full model ($p_1 + 1$ parameters): $Y = \beta_0 + \sum_{j=1}^{p_1} X_j \beta_j + \epsilon$

- RSS_0: RSS for the reduced model ($p_0 + 1$ parameters, $p_0 < p_1$): $Y = \beta_0 + \sum_{j=1}^{p_0} X_j \beta_j + \epsilon$

- Reject H_0 from $H_0 : \beta_{p_0+1} = \ldots = \beta_{p_1} = 0$ vs. H_a : "H_0 is not true" if $F > F_{p_1 - p_0, N - p_1 - 1,\ (1-\alpha)}$.

Simple Regression and Orthogonality

- Univariate regression model with no intercept where $X \in \mathbb{R}$:
$$Y = X\beta + \epsilon.$$

- Given data $(x_1, y_1), \ldots, (x_N, y_N)$,
$$\hat{\beta} = \frac{\sum_{i=1}^{N} x_i y_i}{\sum_{i=1}^{N} x_i^2} = \frac{\langle \boldsymbol{x}, \boldsymbol{y} \rangle}{\langle \boldsymbol{x}, \boldsymbol{x} \rangle}.$$
  - $\boldsymbol{r} = \boldsymbol{y} - \hat{\boldsymbol{y}} = \boldsymbol{y} - \boldsymbol{x}\hat{\beta}$.
  - $\hat{\boldsymbol{y}}$: Projection of $\boldsymbol{y}$ onto $\boldsymbol{x}$.
  - $\boldsymbol{r}$: Orthogonal complement of that projection.

- Multivariate regression: If the inputs are orthogonal ($\langle \boldsymbol{x}_j, \boldsymbol{x}_k \rangle = 0$ if $j \neq k$), then $\hat{\beta}_j = \frac{\langle \boldsymbol{x}_j, \boldsymbol{y} \rangle}{\langle \boldsymbol{x}_j, \boldsymbol{x}_j \rangle}$ for all j.
  $\Rightarrow$ No effect on each other's parameter estimates in the model.

Gram-Schmidt Orthogonalization

Theorem Gram-Schmidt Orthogonalization

Let V be an inner product space and $S = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_p\}$ be a linearly independent subset of V. Define $S' = \{\boldsymbol{z}_1, \boldsymbol{z}_2, \ldots, \boldsymbol{z}_p\}$, where $\boldsymbol{z}_1 = \boldsymbol{x}_1$ and

$$\boldsymbol{z}_j = \boldsymbol{x}_j - \sum_{k=1}^{j-1} \frac{\langle \boldsymbol{z}_k, \boldsymbol{x}_j \rangle}{\langle \boldsymbol{z}_k, \boldsymbol{z}_k \rangle} \boldsymbol{z}_k, \ 2 \leq j \leq p.$$

Then, $\boldsymbol{z}_1, \boldsymbol{z}_2, \ldots, \boldsymbol{z}_p$ are orthogonal (orthogonal basis of V).

$$span(S) = span(S').$$

$\Rightarrow$ Orthogonalization does NOT change the subspace spanned by $\boldsymbol{x}_j$'s.

Logic of Orthogonalization

- Simple linear regression: $Y = \beta_0 + \beta_1 X + \epsilon$.
- LSE $\hat{\beta}_1 = \dfrac{\langle \boldsymbol{x} - \bar{x}\mathbf{1}, \boldsymbol{y} \rangle}{\langle \boldsymbol{x} - \bar{x}\mathbf{1}, \boldsymbol{x} - \bar{x}\mathbf{1} \rangle}$.

Step 1 $\boldsymbol{x} = \gamma_0 \mathbf{1} \;\Rightarrow\; \hat{\gamma}_0 = \dfrac{\langle \mathbf{1}, \boldsymbol{x} \rangle}{\langle \mathbf{1}, \mathbf{1} \rangle} = \bar{x} \;\Rightarrow\; \boldsymbol{z} = \boldsymbol{x} - \bar{x}\mathbf{1}.$

Step 2 $\boldsymbol{y} = \gamma_1 \boldsymbol{z} \;\Rightarrow\; \hat{\gamma}_1 = \dfrac{\langle \boldsymbol{z}, \boldsymbol{y} \rangle}{\langle \boldsymbol{z}, \boldsymbol{z} \rangle} = \dfrac{\langle \boldsymbol{x} - \bar{x}\mathbf{1}, \boldsymbol{y} \rangle}{\langle \boldsymbol{x} - \bar{x}\mathbf{1}, \boldsymbol{x} - \bar{x}\mathbf{1} \rangle} = \hat{\beta}_1.$

- Multiple regression: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$.

Step 1 $\boldsymbol{x}_1 = \gamma_{01} \mathbf{1} \;\Rightarrow\; \hat{\gamma}_{01} = \dfrac{\langle \mathbf{1}, \boldsymbol{x}_1 \rangle}{\langle \mathbf{1}, \mathbf{1} \rangle} \;\Rightarrow\; \boldsymbol{z}_1 = \boldsymbol{x}_1 - \hat{\gamma}_{01}\mathbf{1}.$

Step 2 $\boldsymbol{x}_2 = \gamma_{12}\boldsymbol{z}_1 + \gamma_{02}\mathbf{1} \;\Rightarrow\; \hat{\gamma}_{12} = \dfrac{\langle \boldsymbol{z}_1, \boldsymbol{x}_2 \rangle}{\langle \boldsymbol{z}_1, \boldsymbol{z}_1 \rangle},\; \hat{\gamma}_{02} = \dfrac{\langle \mathbf{1}, \boldsymbol{x}_2 \rangle}{\langle \mathbf{1}, \mathbf{1} \rangle}$

$\Rightarrow\; \boldsymbol{z}_2 = \boldsymbol{x}_2 - \hat{\gamma}_{12}\boldsymbol{z}_1 - \hat{\gamma}_{02}\mathbf{1}.$

Step 3 $\boldsymbol{y} = \gamma_2 \boldsymbol{z}_2 \;\Rightarrow\; \hat{\gamma}_2 = \dfrac{\langle \boldsymbol{z}_2, \boldsymbol{y} \rangle}{\langle \boldsymbol{z}_2, \boldsymbol{z}_2 \rangle} = \hat{\beta}_2.$

Gram-Schmidt Orthogonalization

- If the inputs are not orthogonal, we can orthogonalize the inputs and obtain the coefficient estimate for the last input.

- Gram-Schmidt Procedure:

  0. $\boldsymbol{z}_0 = \boldsymbol{x}_0 = \mathbf{1}$
  1. $\boldsymbol{z}_1 = \boldsymbol{x}_1 - \frac{\langle \boldsymbol{z}_0, \boldsymbol{x}_1 \rangle}{\langle \boldsymbol{z}_0, \boldsymbol{z}_0 \rangle} \boldsymbol{z}_0$
  2. $\boldsymbol{z}_2 = \boldsymbol{x}_2 - \frac{\langle \boldsymbol{z}_0, \boldsymbol{x}_2 \rangle}{\langle \boldsymbol{z}_0, \boldsymbol{z}_0 \rangle} \boldsymbol{z}_0 - \frac{\langle \boldsymbol{z}_1, \boldsymbol{x}_2 \rangle}{\langle \boldsymbol{z}_1, \boldsymbol{z}_1 \rangle} \boldsymbol{z}_1$
  $\vdots$
  p. $\boldsymbol{z}_p = \boldsymbol{x}_p - \sum_{j=0}^{p-1} \frac{\langle \boldsymbol{z}_j, \boldsymbol{x}_p \rangle}{\langle \boldsymbol{z}_j, \boldsymbol{z}_j \rangle} \boldsymbol{z}_j$

- Then $\hat{\beta}_p = \frac{\langle \boldsymbol{z}_p, \boldsymbol{y} \rangle}{\langle \boldsymbol{z}_p, \boldsymbol{z}_p \rangle}$.

Meaning of G-S Orthogonalization

- Any one of $\boldsymbol{x}_j,\ j = 1, \ldots, p$ can be in the last position. $\Rightarrow$ All β_j can be estimated by the G-S orthogonalization.

- $\hat{\beta}_j$: Simple regression coefficient of $\boldsymbol{y}$ on $\boldsymbol{z}^*$, where $\boldsymbol{z}^*$ is the residual after regressing

$$\boldsymbol{x}_j = \gamma_0 \boldsymbol{x}_0 + \gamma_1 \boldsymbol{x}_1 + \cdots + \gamma_{j-1} \boldsymbol{x}_{j-1} + \gamma_{j+1} \boldsymbol{x}_{j+1} + \cdots + \gamma_p \boldsymbol{x}_p.$$

$\Rightarrow \hat{\beta}_j$: Additional contribution of $\boldsymbol{x}_j$ on $\boldsymbol{y}$, after x_j has been adjusted for $\boldsymbol{x}_0, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_{j-1}, \boldsymbol{x}_{j+1}, \ldots, \boldsymbol{x}_p$.

Problem of Least Squares

- When $\boldsymbol{x}_p$ (the observed values for the pth variable) is highly correlated with some of the other $\boldsymbol{x}_k$'s, the residual vector $\boldsymbol{z}_p$ will be close to 0.
- $\text{Var}(\hat{\beta}_p) = \text{Var}\left(\frac{\boldsymbol{z}_p^\top \boldsymbol{y}}{\boldsymbol{z}_p^\top \boldsymbol{z}_p}\right) = \frac{\boldsymbol{z}_p^\top (\sigma^2 \boldsymbol{I}) \boldsymbol{z}_p}{(\boldsymbol{z}_p^\top \boldsymbol{z}_p)^2} = \frac{\sigma^2 \boldsymbol{z}_p^\top \boldsymbol{z}_p}{(\boldsymbol{z}_p^\top \boldsymbol{z}_p)^2} = \frac{\sigma^2}{\|\boldsymbol{z}_p\|^2}$
- If $\boldsymbol{x}_p$ and other inputs are highly correlated, $\|\boldsymbol{z}_p\| \downarrow \Rightarrow$ $Var(\hat{\beta}_p) \uparrow$.

Alternatives to Least Squares

- There are two reasons we may not be satisfied with least squares estimates:
  - *Prediction Accuracy*: LS estimates often have large variance (when predictors are correlated). A little bit of bias can be added to reduce the variance of the predicted values.
  - *Interpretation*: Often we would like to determine a smaller subset of variables that exhibit the strongest effects. To get the "big picture", some small details can be sacrificed.

- Three methods to fix these problems:
  - Subset Selection
  - Coefficient Shrinkage
  - Methods Using Derived Input Directions

Subset Selection

- *Best Subset Selection*: For each $k \in \{0, 1, \ldots, p\}$, RSS values of all possible models are computed. $\Rightarrow$ Model with the smallest RSS at each k. $\Rightarrow$ Final model determined by AIC, BIC, or cross-validation. $\Rightarrow$ The optimal model. If p is large, then this may be infeasible.
- *Forward Selection*: Start with the intercept, then sequentially add inputs that most improves the fit until the last input is not significant. $\Rightarrow$ greedy algorithm $\Rightarrow$ suboptimal (No problem for large p even when $p \geq N$).
- *Backward Selection*: Start with the full model, then sequentially delete inputs that contribute least to the RSS until all inputs are significant (Only available when $N > p$).
- *Stepwise Regression*: Consider forward and backward moves at each stage. $\Rightarrow$ For large p, computationally expensive.

Best subset

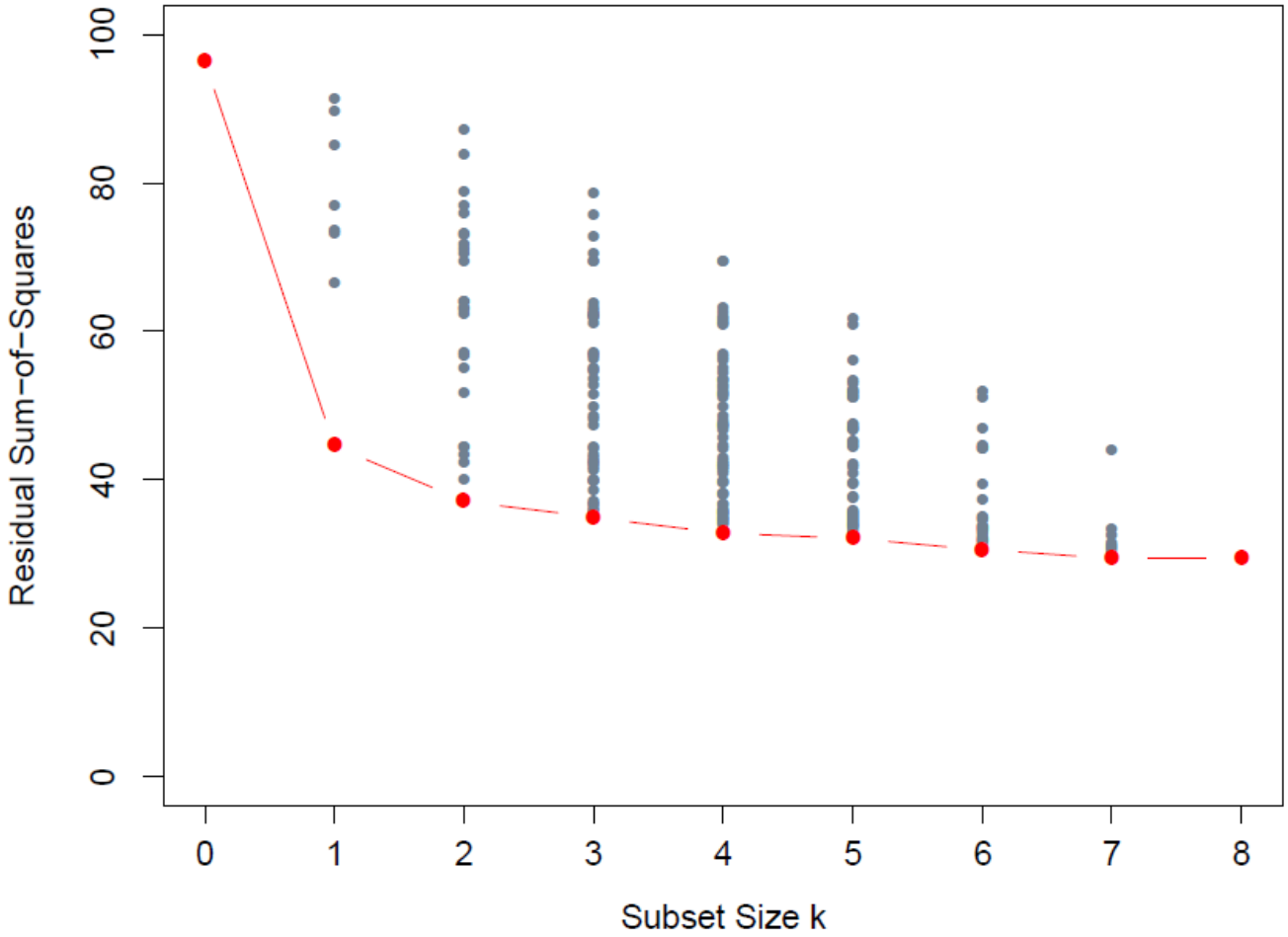


FIGURE 3.5. *All possible subset models for the prostate cancer example. At each subset size is shown the residual sum-of-squares for each model of that size.*

Shrinkage Methods

- Subset selection $\Rightarrow$ Discrete process (retain or discard) $\Rightarrow$ Sometimes high model variance.
- Shrinkage methods $\Rightarrow$ More continuous process.
- Shrinkage methods:
  - Ridge regression.
  - Lasso, Elastic net, SCAD.
  - Least Angle Regression (LAR).

Ridge Regression

- Another way to control the variance is to impose a size constraint on the coefficients.
- Each input variable and the output should be standardized; that is $\bar{y} = 0$ and $\sum_{i=1}^{N} x_{ij} = 0$ & $\sum_{i=1}^{N} x_{ij}^2/(N-1) = 1$.
- Ridge regression minimizes $RSS(\beta)$ subject to the constraint $\sum_{j=1}^{p} \beta_j^2 \leq s$.
- This is equivalent to minimizing the penalized RSS:

$$PRSS(\beta; \lambda) = \sum_{i=1}^{N} (y_i - f(\boldsymbol{x}_i))^2 + \lambda \sum_{j=1}^{p} \beta_j^2.$$

with respect to β, considering λ to be fixed.

Ridge Solution

$$PRSS(\beta;\lambda) = (\boldsymbol{y} - \boldsymbol{X}\beta)^\top(\boldsymbol{y} - \boldsymbol{X}\beta) + \lambda\beta^\top\beta$$

$$\frac{\partial PRSS}{\partial \beta} = -2\boldsymbol{X}^\top(\boldsymbol{y} - \boldsymbol{X}\hat{\beta}^{\text{ridge}}) + 2\lambda\hat{\beta}^{\text{ridge}} = 0$$

$$0 = -\boldsymbol{X}^\top\boldsymbol{y} + \boldsymbol{X}^\top\boldsymbol{X}\hat{\beta}^{\text{ridge}} + \lambda\boldsymbol{I}\hat{\beta}^{\text{ridge}}$$

$$(\boldsymbol{X}^\top\boldsymbol{X} + \lambda\boldsymbol{I})\hat{\beta}^{\text{ridge}} = \boldsymbol{X}^\top\boldsymbol{y}$$

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

Singular Value Decomposition

- The singular value decomposition (SVD) of $\boldsymbol{X} \in \mathbb{R}^{N \times p}$ has the form
$$\boldsymbol{X} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^{\top}.$$
- $\boldsymbol{U}$ is a $N \times p$ orthogonal matrix with columns which span the column space of $\boldsymbol{X}$.
- $\boldsymbol{V}$ is a $p \times p$ orthogonal matrix with columns which span the row space of $\boldsymbol{X}$.
- $\boldsymbol{D}$ is a $p \times p$ diagonal matrix with diagonal entries $d_1 \geq d_2 \geq \cdots \geq d_r > d_{r+1} = \ldots = d_p = 0$ where $r = rank(\boldsymbol{X})$. The entries $d_1, \ldots, d_p$ are called the *singluar values* of $\boldsymbol{X}$.

SVD and LSE

- Applying the singular value decomposition of $\boldsymbol{X}$ to the least square regression:

$$\begin{aligned}
\hat{\boldsymbol{y}} &= \boldsymbol{X}\hat{\beta}^{ls} \\
&= \boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{y} \\
&= \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^\top[(\boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^\top)^\top\boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^\top]^{-1}(\boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^\top)^\top\boldsymbol{y} \\
&= \boldsymbol{U}\boldsymbol{U}^\top\boldsymbol{y}.
\end{aligned}$$

- $\boldsymbol{U}^\top\boldsymbol{y}$ is the coordinates of $\boldsymbol{y}$ w.r.t. the orthogonal basis $\boldsymbol{U}$.

SVD and Ridge Regression

- Applying the singular value decomposition of $\boldsymbol{X}$ to the ridge regression formulas:
  - $$\begin{aligned}\hat{\beta}^{\text{ridge}} &= (\boldsymbol{X}^\top \boldsymbol{X} + \lambda \boldsymbol{I})^{-1} \boldsymbol{X}^\top \boldsymbol{y} \\ &= \boldsymbol{V}(\boldsymbol{D}^2 + \lambda \boldsymbol{I})^{-1} \boldsymbol{D} \boldsymbol{U}^\top \boldsymbol{y} \\ &= \sum_{j=1}^{p} v_j \frac{d_j}{d_j^2 + \lambda} \boldsymbol{u}_j^\top \boldsymbol{y}\end{aligned}$$
  - $$\begin{aligned}\hat{\boldsymbol{y}}^{\text{ridge}} &= \boldsymbol{X} \hat{\beta}^{\text{ridge}} \\ &= \boldsymbol{U}\boldsymbol{D}(\boldsymbol{D}^2 + \lambda \boldsymbol{I})^{-1} \boldsymbol{D} \boldsymbol{U}^\top \boldsymbol{y} \\ &= \sum_{j=1}^{p} \boldsymbol{u}_j \frac{d_j^2}{d_j^2 + \lambda} \boldsymbol{u}_j^\top \boldsymbol{y}\end{aligned}$$
- Since $\lambda \geq 0$, $\frac{d_j^2}{d_j^2+\lambda} \leq 1$.
- Like the LSE, ridge regression computes the coordinates of $\boldsymbol{y}$ w.r.t. the orthogonal basis $\boldsymbol{U}$. But, these coordinates are shrunken by $\frac{d_j^2}{d_j^2+\lambda}$.

SVD and Ridge Regression

- For fixed λ, small $d_j^2 \Rightarrow$ great amount of shrinkage.
- The meaning of small d_j^2:
  - Eigen decomposition of $\boldsymbol{X}^\top \boldsymbol{X}$:
    $\boldsymbol{X}^\top \boldsymbol{X} = \boldsymbol{V}\boldsymbol{D}^2\boldsymbol{V}^\top$, where $\boldsymbol{V} = (\boldsymbol{v}_1, \ldots, \boldsymbol{v}_p)$, where $\boldsymbol{v}_j$ is eigenvectors of $\boldsymbol{X}^\top \boldsymbol{X}$.
  - $\boldsymbol{z}_j = \boldsymbol{X}\boldsymbol{v}_j$: j^{th} principal component.
  - $\boldsymbol{z}_1 = \boldsymbol{X}\boldsymbol{v}_1 = \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^\top \boldsymbol{v}_1 = \boldsymbol{u}_1 d_1$.
  - Sample variance of $\boldsymbol{z}_1$: $Var(\boldsymbol{z}_1) = \frac{\boldsymbol{z}_1^\top \boldsymbol{z}_1}{N} = \frac{d_1^2 \boldsymbol{u}_1^\top \boldsymbol{u}_1}{N}$.
  - Since $d_1 \geq d_2 \geq \cdots \geq d_p$,
    $Var(\boldsymbol{z}_1) \geq Var(\boldsymbol{z}_2) \geq \cdots \geq Var(\boldsymbol{z}_p)$.
  - Small d_j^2 means the directions with small variance in the column space of $\boldsymbol{X} \Rightarrow$ ridge regression shrinks these directions.

Ridge Regression: df

- The *effective degrees of freedom* of the ridge regression fit is

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

- Effective d.f. $=$ the $\#$ of free parameters.
- Effective d.f. of linear regression $= tr(\boldsymbol{H}) = p$, when the intercept term is removed by standardization.
- Note that $\mathrm{df}(0) = p$ and $\mathrm{df}(\infty) = 0$.
- Note that the coefficients tend to get closer to 0 as λ increases.

d.f. vs. Coefficient

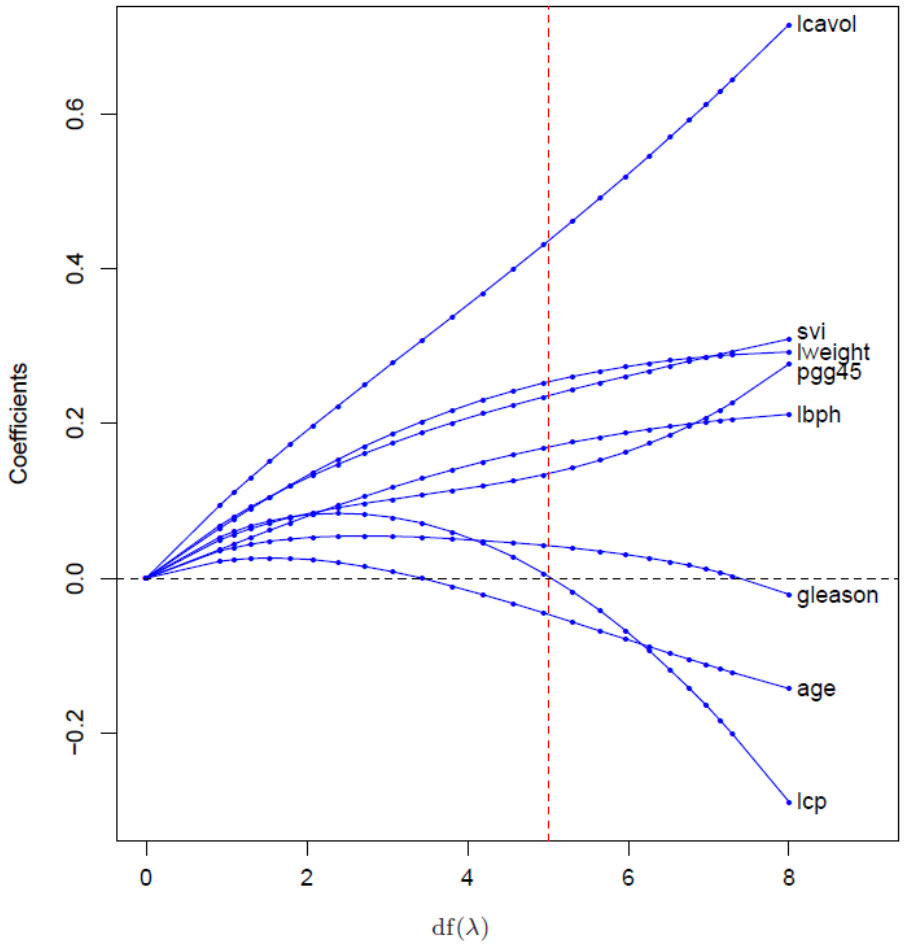


FIGURE 3.8. *Profiles of ridge coefficients for the prostate cancer example, as the tuning parameter λ is varied. Coefficients are plotted versus $\mathrm{df}(\lambda)$, the effective degrees of freedom. A vertical line is drawn at $\mathrm{df} = 5.0$, the value chosen by cross-validation.*

Lasso

- Standardization of input variables.
- Lasso estimator:

$$\hat{\beta}^L = \arg\min_{\beta} \left[\sum_{i=1}^{N} (y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right].$$

$$\hat{\beta}^L = \arg\min_{\beta} \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2 \text{ subject to } \sum_{j=1}^{p} |\beta_j| \leq t.$$

- Unlike ridge regression, as $\lambda \uparrow$ (or $t \downarrow$), some of $\hat{\beta}_j^L \to 0$.
- If $t > \sum_{j=1}^{p} |\beta_j^{ls}|$, $\hat{\beta}_j^L = \hat{\beta}_j^{ls}$.

Lasso fit vs. Restriction

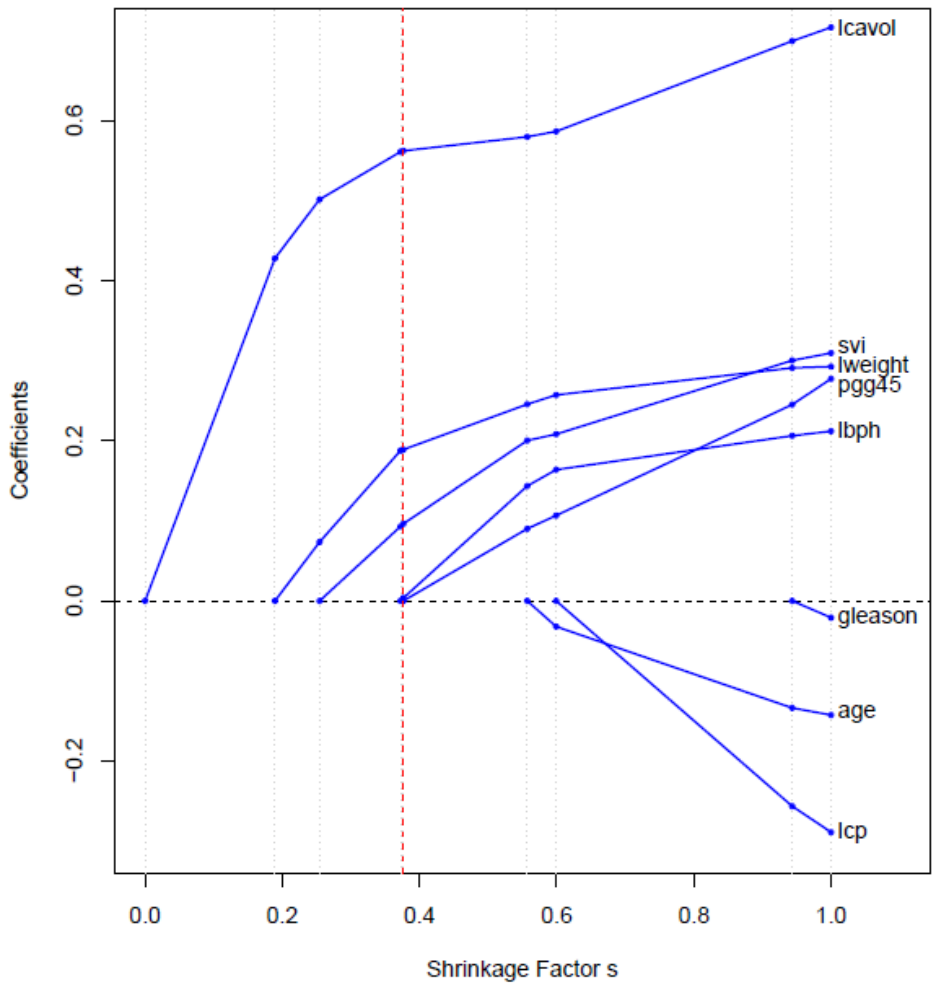
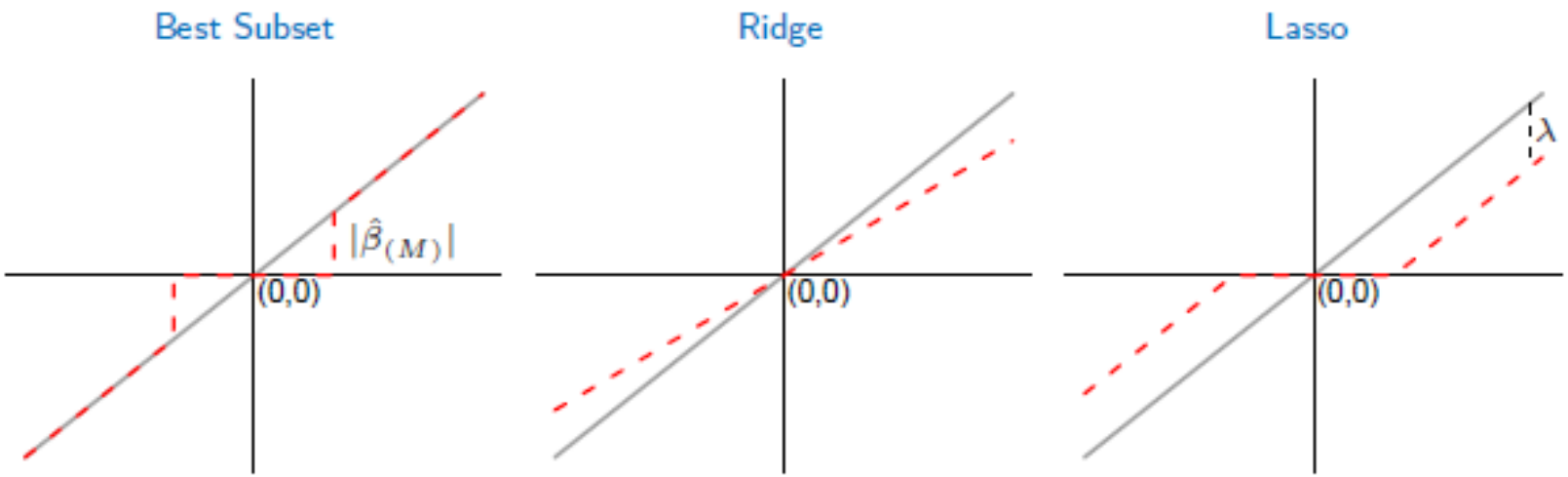


FIGURE 3.10. *Profiles of lasso coefficients, as the tuning parameter t is varied. Coefficients are plotted versus $s = t/\sum_1^p |\hat{\beta}_j|$. A vertical line is drawn at $s = 0.36$, the value chosen by cross-validation. Compare Figure 3.8 on page 65; the lasso profiles hit zero, while those for ridge do not. The profiles are piece-wise linear, and so are computed only at the points displayed; see Section 3.4.4 for details.*

Lasso vs. Ridge

When the columns of $\boldsymbol{X}$ are orthonormal.



Lasso vs. Ridge

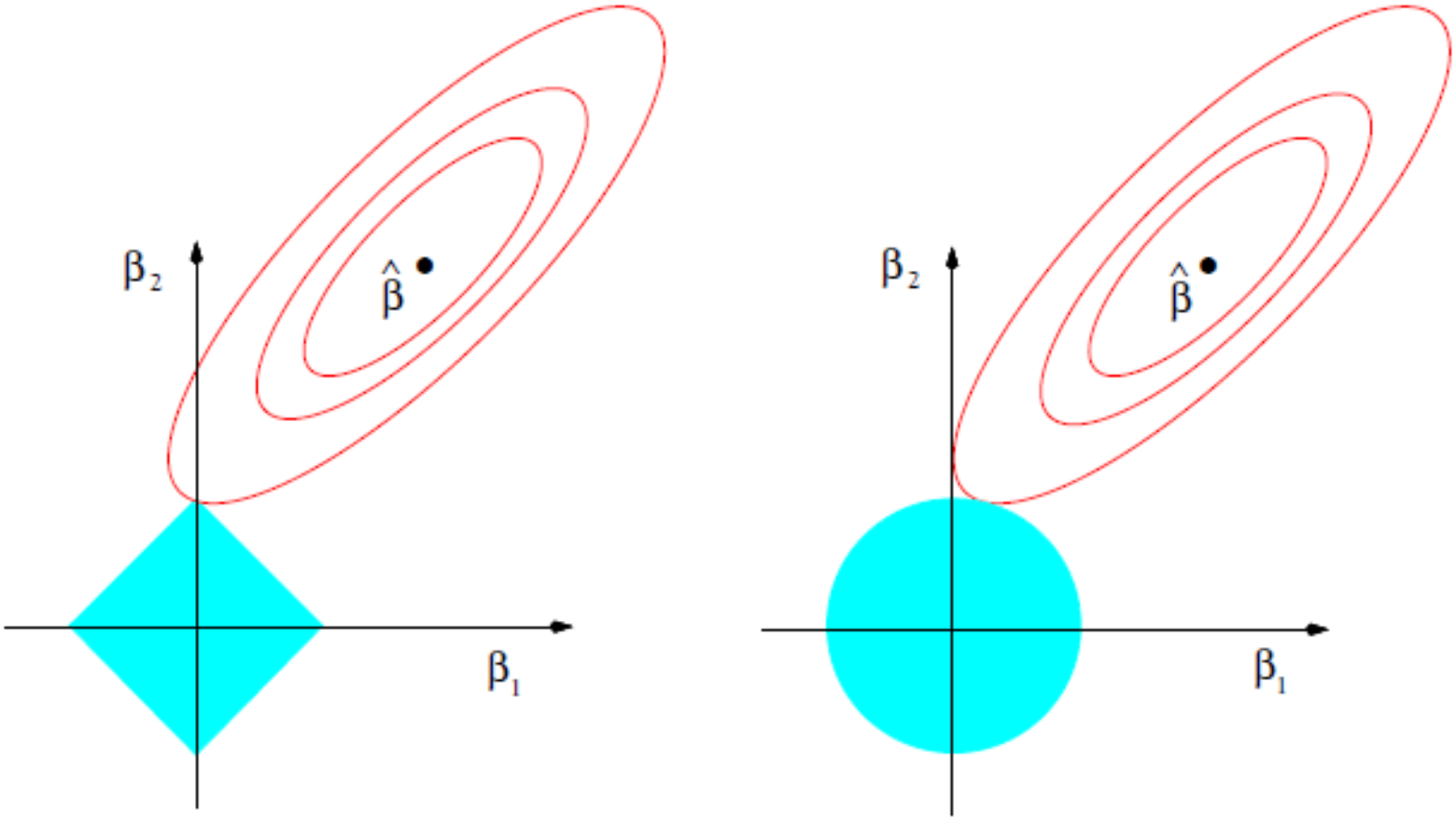


FIGURE 3.11. *Estimation picture for the lasso (left) and ridge regression (right). Shown are contours of the error and constraint functions. The solid blue areas are the constraint regions* $|\beta_1| + |\beta_2| \leq t$ *and* $\beta_1^2 + \beta_2^2 \leq t^2$*, respectively, while the red ellipses are the contours of the least squares error function.*

Generalization

- Generalization

$$\hat{\beta} = \arg\min_{\beta} \left[\sum_{i=1}^{N} (y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} |\beta_j|^q \right].$$

  - $q = 0$: Subset selection.
  - $q = 1$: Lasso regression.
  - $q = 2$: Ridge regression.

- Problems of lasso:
  - No analytical solution.
  - Not consistent estimator (Non-zero coefficients to be biased towards zero).

Elastic Net & Grouped Lasso

- Elastic net penalty:

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

  - $0 \leq \alpha \leq 1$. Compromise between ridge & lasso.
  - It selects variables like the lasso and shrinks together the coefficients of correlated inputs.

- Grouped lasso: Inputs with pre-defined groups (e.g., genes with the same pathway, a set of dummy variables, etc)

$$\min_{\beta} \left[||\boldsymbol{y} - \beta_0 \mathbf{1} - \sum_{l=1}^{L} \boldsymbol{X}_l \beta_l||_2^2 + \lambda \sum_{l=1}^{L} \sqrt{p_l} ||\beta_l||_2 \right].$$

  - L groups of p predictors.
  - $\boldsymbol{X}_l$: Predictors of the lth group.
  - $\sqrt{p_l}$ accounts for the group size.

SCAD

- SCAD (Smoothly Clipped Absolute Deviation) penalty

$$\begin{cases} \lambda|\beta| & \text{if } |\beta| \leq \lambda \\ -\frac{|\beta|^2 - 2a\lambda|\beta| + \lambda^2}{2(a-1)} & \text{if } \lambda < |\beta| \leq a\lambda \\ \frac{(a+1)\lambda^2}{2} & \text{if } |\beta| > a\lambda \end{cases}$$

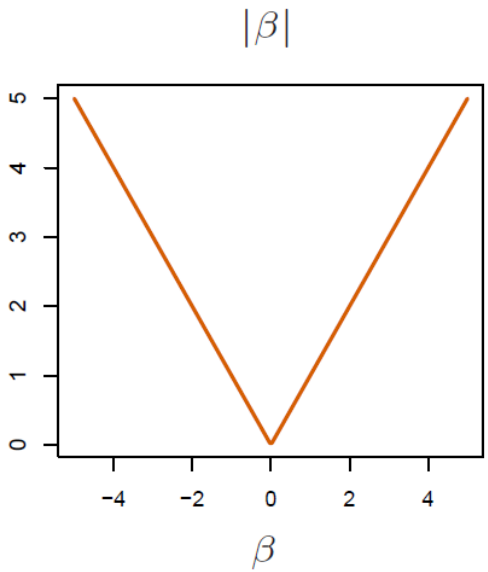
  - $a > 2,\ \lambda > 0.$

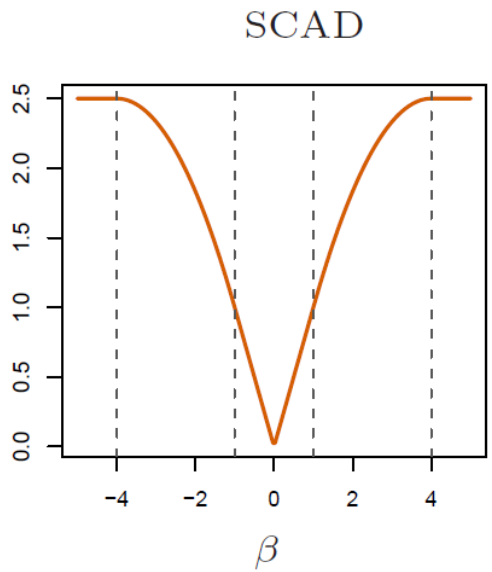
- Solution of the SCAD penalty:

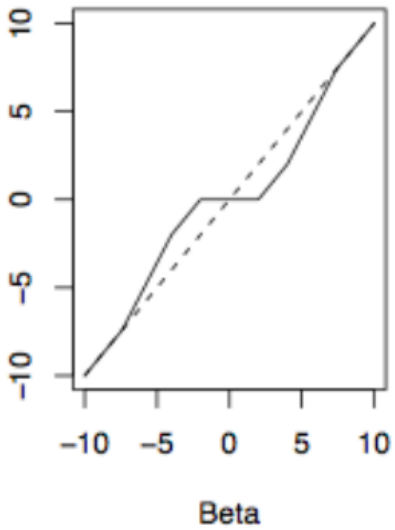
$$\begin{cases} (|\beta_j| - \lambda)_+ sign(\hat{\beta}_j) & \text{if } |\hat{\beta}_j| \leq 2\lambda \\ \{(a-1)\hat{\beta}_j - sign(\hat{\beta}_j)a\lambda\}/(a-2) & \text{if } 2\lambda < |\hat{\beta}_j| \leq a\lambda \\ \hat{\beta}_j & \text{if } |\hat{\beta}_j| > a\lambda \end{cases}$$

  - Small coefficients being set to zero, a few other coefficients being shrunk towards zero while retaining the large coefficients. Thus, SCAD can produce sparse set of solution and approximately unbiased coefficients for large coefficients.

Lasso vs. SCAD







Adaptive Lasso

- Adaptive lasso penalty:

$$\sum_{j=1}^{p} w_j |\beta_j|,$$

  - $w_j = 1/|\hat{\beta}_j|^{\nu}$, where $\hat{\beta}_j$ is LSE and $\nu > 0$.
  - Consistent estimator.

Regularization Path Algorithm

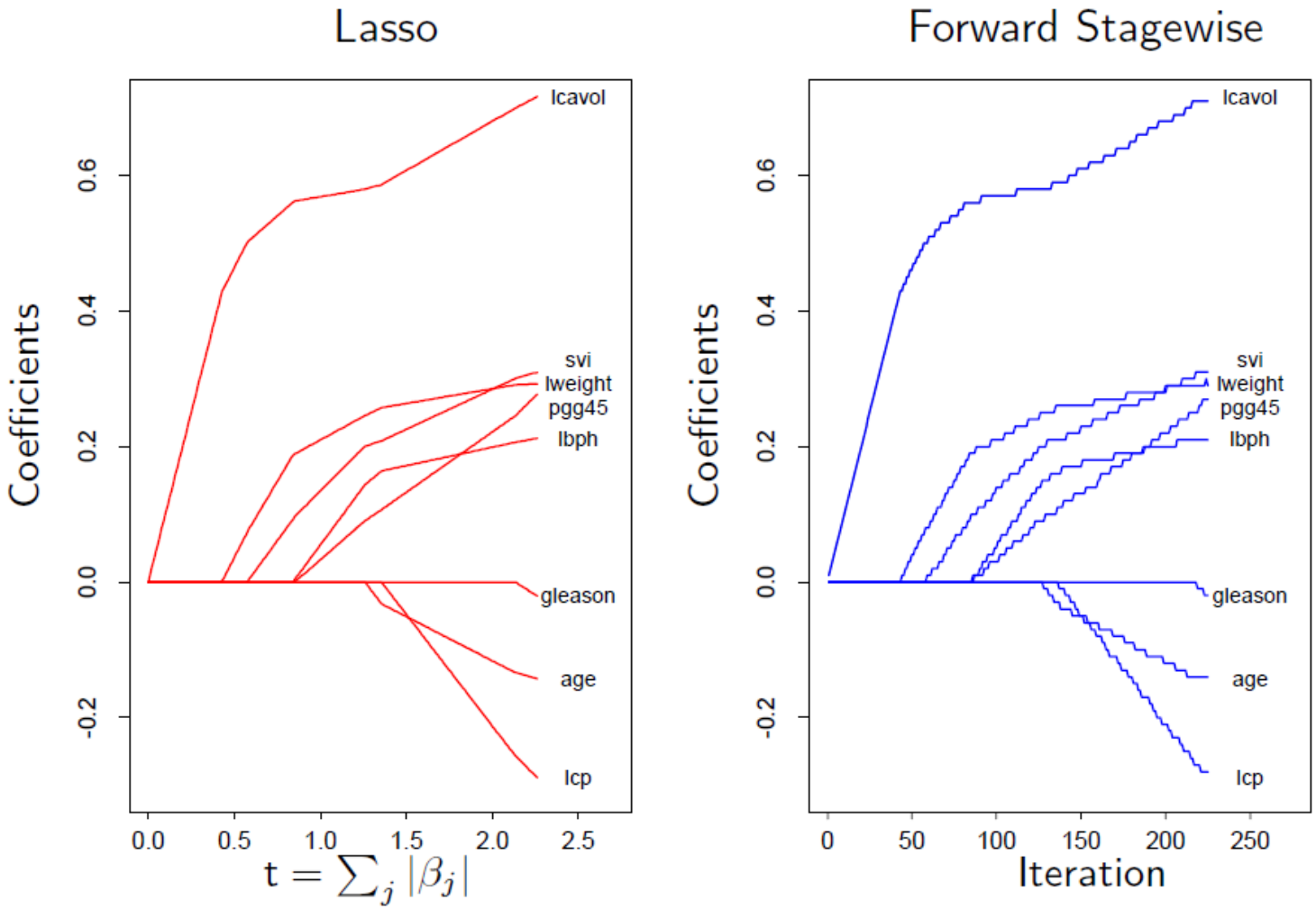
- Lasso solution:
  - If columns of input matrix $\boldsymbol{X}$ are not orthogonal, there does not exist analytical lasso solution.
  - Efficient algorithms for lasso solution.
  - Algorithms developed from boosting (ensemble learning) $\Rightarrow$ Incremental forward stagewise regression.
- Path algorithm:
  - Incremental forward stagewise regression.
  - Least angle regression (LAR).

Incremental Forward Stagewise Regression

- Version of least squares boosting for linear regression.
- Forward selection: Discrete process.
- Incremental Forward Stagewise Regression: Almost continuous process.
- Standardized inputs are used.

- Algorithm:
  1. Start with the residual $\boldsymbol{r} = \boldsymbol{y}$ and $\beta_1, \ldots, \beta_p = 0$. All $\boldsymbol{x}_j$'s are standardized.
  2. Find the predictor $\boldsymbol{x}_j$ most correlated with $\boldsymbol{r}$.
  3. Update $\beta_j \leftarrow \beta_j + \delta_j$, where $\delta_j = \epsilon \cdot \text{sign}(\langle \boldsymbol{x}_j, \boldsymbol{r} \rangle)$ and $\epsilon > 0$ is a small step size, and set $\boldsymbol{r} \leftarrow \boldsymbol{r} - \delta_j \boldsymbol{x}_j$.
  4. Repeat steps 2 and 3 many times, until the residuals are uncorrelated with all the predictors.

Incremental Forward Stagewise Regression vs. Lasso



Least Angle Regression (LAR)

- LAR: Democratic version of the forward selection method.
  - LAR coefficients: As much of a predictor as it deserves.
- Modified version of LAR $\Rightarrow$ Lasso solution.
- Algorithm:
  1. Standardize inputs to have mean zero and unit norm.
  2. Start with $\boldsymbol{r} = \boldsymbol{y} - \bar{y}\mathbf{1}$ & $\beta_1 = \cdots = \beta_p = 0$.
  3. Find the input $\boldsymbol{x}_j$ most correlated with $\boldsymbol{r}$.
  4. Move β_j from 0 towards its LSE $\langle \boldsymbol{x}_j, \boldsymbol{r} \rangle$, until some other input $\boldsymbol{x}_k$ has as much correlation with the current residual as does $\boldsymbol{x}_j$.
  5. Move β_j & β_k in the direction defined by LSE of $\boldsymbol{r} = \beta_j \boldsymbol{x}_j + \beta_k \boldsymbol{x}_k$, until some other input $\boldsymbol{x}_l$ has as much correlation with the current residual.
     - 5-1. If a non-zero coefficient hits zero, drop its variable from the active set of variables & recompute the current joint least square direction. (optional for lasso solution)
  6. Continue in this way until all p inputs have been entered.

Derived Input Directions

- Derived input direction methods:
  - Principal component regression (PCR): Directions with high variance in $\boldsymbol{X}$.
  - Partial Least Squares (PLS): Directions with high variance in $\boldsymbol{X}$ and high correlation with $\boldsymbol{y}$.

Principal Components

- Applying the SVD of $\boldsymbol{X}$ to obtain an expression for $\boldsymbol{X}^\top \boldsymbol{X}$, we obtain
$$\boldsymbol{X}^\top \boldsymbol{X} = \boldsymbol{V}\boldsymbol{D}\boldsymbol{U}^\top \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^\top = \boldsymbol{V}\boldsymbol{D}^2\boldsymbol{V}^\top.$$
- The columns of $\boldsymbol{V}$ are the eigenvectors of $\boldsymbol{X}^\top \boldsymbol{X}$.
- The jth principal component of $\boldsymbol{X}$ is defined as
$$\boldsymbol{z}_j = \boldsymbol{X} v_j = \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^\top v_j = d_j \boldsymbol{u}_j.$$
- The first principal component has the largest sample variance among all normalized linear combinations of the columns of $\boldsymbol{X}$. The last principal component has the minimum variance.

Principal Components Regression

- Regress the output $\boldsymbol{y}$ on the first $M(\leq p)$ principal components.
- $\hat{\boldsymbol{y}}^{\text{pcr}} = \bar{y}\mathbf{1} + \sum_{m=1}^{M} \hat{\theta}_m \boldsymbol{z}_m$, $\hat{\theta}_m = \dfrac{\langle \boldsymbol{z}_m, \boldsymbol{y} \rangle}{\langle \boldsymbol{z}_m, \boldsymbol{z}_m \rangle}$
- Since each $\boldsymbol{z}_m$ is a linear combination of the original $\boldsymbol{x}$'s, we can give an estimate of the original slope coefficients:

$$\hat{\beta}^{\text{pcr}}(M) = \sum_{m=1}^{M} \hat{\theta}_m v_m.$$

- If $M = p$, then $\hat{\beta}^{\text{pcr}}(p) = \hat{\beta}$.
- The proportion variation in the inputs explained by first M principal components is $\dfrac{\sum_{j=1}^{M} d_j^2}{\sum_{j=1}^{p} d_j^2}$.

Partial Least Squares (PLS)

- Unlike PCR, PLS uses $\boldsymbol{y}$ to derive input directions.

- Algorithm:
  1. Standardize inputs to have mean zero and unit variance.
  2. Set $\hat{\boldsymbol{y}}^{(0)} = \bar{y}\mathbf{1}$ & $\boldsymbol{x}_j^{(0)} = \boldsymbol{x}_j,\ j = 1, \ldots, p$.
  3. For $m = 1, \ldots, p$,
     - 3-1. $\boldsymbol{z}_m = \sum_{j=1}^{p} \hat{\phi}_{mj} \boldsymbol{x}_j^{(m-1)}$, where $\hat{\phi}_{mj} = \langle \boldsymbol{x}_j^{(m-1)}, \boldsymbol{y} \rangle$.
     - 3-2. $\hat{\theta}_m = \frac{\langle \boldsymbol{z}_m, \boldsymbol{y} \rangle}{\langle \boldsymbol{z}_m, \boldsymbol{z}_m \rangle}$.
     - 3-3. $\hat{\boldsymbol{y}}^{(m)} = \hat{\boldsymbol{y}}^{(m-1)} + \hat{\theta}_m \boldsymbol{z}_m$.
     - 3-4. Orthogonalize each $\boldsymbol{x}_j^{(m-1)}$ w.r.t. $\boldsymbol{z}_m$ (i.e., $\boldsymbol{x}_j^{(m)} = \boldsymbol{x}_j^{(m-1)} - \frac{\langle \boldsymbol{z}_m, \boldsymbol{x}_j^{(m-1)} \rangle}{\langle \boldsymbol{z}_m, \boldsymbol{z}_m \rangle} \boldsymbol{z}_m,\ j = 1, \ldots, p$).
  4. Output: The fitted vector $\hat{\boldsymbol{y}}^{(m)}$.