Chapter 4

Multicollinearity

4.1 Multicollinearity

-A set of predictors $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_p$ is said to have "multicollinearity" if there exist linear or near-linear dependencies among predictors.

-In case there exists a linear dependency among the predictors, the columns of $\boldsymbol{X} = (\mathbf{1}, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_p)$ are linearly dependent, or equivalently, the centered columns $\boldsymbol{x}_1 - \bar{x}_1\mathbf{1}, \ldots, \boldsymbol{x}_p - \bar{x}_p\mathbf{1}$ are linearly dependent, so that the matrix $\boldsymbol{X}$ and $\boldsymbol{X}^\top\boldsymbol{X}$ are not of full rank.

Multicollinearity not only makes the computation of the parametric estimates erratic, but also increase the variance of the estimates

$$\sum_{j=0}^{p} Var(\hat{\beta}_j) = \mathrm{tr}(Var(\hat{\boldsymbol{\beta}})) = \sigma^2 \mathrm{tr}((\boldsymbol{X}^\top\boldsymbol{X})^{-1}) = \sigma^2 \sum_{j=0}^{p} \frac{1}{\kappa_j},$$

where κ_j's are eigenvalues of $\boldsymbol{X}^\top\boldsymbol{X}$.

Let $S_{jj} = \sum_{i=1}^{n}(x_{ij} - \bar{x}_j)^2$ and R_j^2 denote the coefficient of determinant in regressing the jth predictor x_j on the remaining $(x_k : k \neq j)$.Then,

$$Var(\hat{\beta}_j) = \frac{1}{1 - R_j^2}\frac{\sigma^2}{S_{jj}}, \quad 1 \leq j \leq p$$

Proof. Assume j=1 without loss of generality. Recalling that

$$\hat{\beta}_A = (\boldsymbol{X}_A^T\boldsymbol{X}_A)^{-1}\boldsymbol{X}_A^T(Y - \boldsymbol{X}_B\hat{\beta}_B), \quad \hat{\beta}_B = (\boldsymbol{X}_{B,\perp}^\top\boldsymbol{X}_{B,\perp})^{-1}\boldsymbol{X}_{B,\perp}^\top\boldsymbol{Y}$$

in the regression $\boldsymbol{Y} = \boldsymbol{X}\beta + \varepsilon$, where

$$\beta = (\beta_A^T, \beta_B^T)^T, \quad \boldsymbol{X} = (\boldsymbol{X}_A, \boldsymbol{X}_B) \quad \text{with } \boldsymbol{X}\beta = \boldsymbol{X}_A\beta_A + \boldsymbol{X}_B\beta_B, \quad \hat{\beta} = (\hat{\beta}_A^T, \hat{\beta}_B^T)^T.$$

This with choice $\boldsymbol{X}_A = (\mathbf{1}, \boldsymbol{X}_2, \cdots, \boldsymbol{X}_p), \boldsymbol{X}_B = (\boldsymbol{X}_1)$ gives

$$\hat{\boldsymbol{\beta}}_1 = (\boldsymbol{X}_{\mathbf{1},\perp}^T \boldsymbol{X}_{\mathbf{1},\perp})^{-1} \boldsymbol{X}_{\mathbf{1},\perp}^T \boldsymbol{Y} \quad with\ \boldsymbol{X}_{\mathbf{1},\perp} = \boldsymbol{X}_1 - \Pi(\boldsymbol{X}_1 | C_{\mathbf{1},\boldsymbol{X}_2,\cdots,\boldsymbol{X}_p})$$

so that

$$Var(\hat{\boldsymbol{\beta}}_1) = (\boldsymbol{X}_{\mathbf{1},\perp}^T \boldsymbol{X}_{\mathbf{1},\perp})^{-1} \boldsymbol{X}_{\mathbf{1},\perp}^T \underbrace{Var(Y)}_{\sigma^2 \cdot I} \boldsymbol{X}_{\mathbf{1},\perp} (\boldsymbol{X}_{\mathbf{1},\perp}^T \boldsymbol{X}_{\mathbf{1},\perp}) = (\boldsymbol{X}_{\mathbf{1},\perp}^T \boldsymbol{X}_{\mathbf{1},\perp})^{-1} \cdot \sigma^2$$

Think of fitting the regression model

$$x_{i1} = \alpha_0 + \alpha_1 x_{i2} + \cdots + \alpha_{p-1} x_{ip} + \varepsilon_i, \quad i = 1, \cdots, n$$

Note that $\boldsymbol{X}_{\mathbf{1},\perp}$ is the residual vector in the above regression. so that $\boldsymbol{X}_{\mathbf{1},\perp}^T \boldsymbol{X}_{\mathbf{1},\perp}$ is the residual sum squares. Since the total sum of squares in this case is S_{11},

$$R_1^2 = \frac{S_{11} - \boldsymbol{X}_{\mathbf{1},\perp}^T \boldsymbol{X}_{\mathbf{1},\perp}}{S_{11}} \Leftrightarrow \boldsymbol{X}_{\mathbf{1},\perp}^T \boldsymbol{X}_{\mathbf{1},\perp} = S_{11}(1 - R_1^2)$$

□

4.2 Diagnostics

- Variance inflation factor
$$VIF_j := (1 - R_j^2)^{-1}$$
VIF is simply the inflation rate of $Var(\hat{\beta}_j)$ in comparison with the case where x_j is not correlated with other predictor, i.e.
$$S_{jk} = \sum_{i=1}^{n} (x_{ij} - \bar{x_j})(x_{ik} - \bar{x_k}) = 0, \quad \forall k \neq j$$
Large VIF_j for one or multiple j's indicates multicollinearity. The inspection of all pairwise correlation between two predictors is not sufficient for detecting multicollinearity in general.

- Condition number of $\boldsymbol{X}^T \boldsymbol{X}$:
$$\kappa = \frac{(\text{Maximal eigenvalue of } \boldsymbol{X}^T \boldsymbol{X})}{(\text{Minimal eigenvalue of } \boldsymbol{X}^T \boldsymbol{X})}$$
Typically, the existence of $VIF_j > 10$ or $\kappa > 100$ is considered as an indication of severe multicollinearity.

4.3 Preliminaries

- Spectral decomposition(of real symmetric matrix)

  Let $A : m \times m$ real symmetric matrix.

  $(\lambda_j, \nu_j),\ i = 1, \cdots, m$: eigenvalue and corresponding orthonormal eignvectors of A, where "orthonormal" means that $\nu_i^T \nu_j = \begin{cases} 1, & i = j \\ 0, & \text{o.w.} \end{cases}$

  $\Rightarrow$ A is decomposed as

$$A = P\Lambda P^T = \lambda_1 \nu_1 \nu_1^T + \cdots + \lambda_m \nu_m \nu_m^T,$$

  where $P = \begin{pmatrix} \nu_1 & \nu_2 & \cdots & \nu_m \end{pmatrix} = \begin{pmatrix} \nu_{11} & \cdots & \nu_{1m} \\ \vdots & & \vdots \\ \nu_{m1} & \cdots & \nu_{mm} \end{pmatrix}$ *and* $\Lambda = \begin{pmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & & \ddots & 0 \\ 0 & & \cdots & \lambda_m \end{pmatrix}$.

- Properties of spectral decomposition $A = P\Lambda P^T$
  - $PP^T = P^T P = I$ so that $P^{-1} = P^T$
  - $A^{-1} = P\Lambda^{-1}P^T, \quad \Lambda^{-1} = \begin{pmatrix} \frac{1}{\lambda_1} & 0 & \cdots & 0 \\ 0 & & \ddots & 0 \\ 0 & & \cdots & \frac{1}{\lambda_m} \end{pmatrix}$
  - $A^K = P\Lambda^K P^T, \quad \Lambda^K = \begin{pmatrix} \lambda_1^K & 0 & \cdots & 0 \\ 0 & & \ddots & 0 \\ 0 & & \cdots & \lambda_m^K \end{pmatrix}$

- Principal Component Analysis(PCA)

  Recall that $\boldsymbol{X}_{A,\perp} = \boldsymbol{X}_A - H_{\mathbf{1}}\boldsymbol{X}_A$, where $H_{\mathbf{1}} = \mathbf{1}(\mathbf{1}^T\mathbf{1})^{-1}\mathbf{1}^T = \frac{1}{n}\mathbf{1}\mathbf{1}^T$ and $\boldsymbol{X} = \begin{pmatrix} \mathbf{1} & \boldsymbol{X}_A \end{pmatrix}$. Here, $\boldsymbol{X}_{A,\perp}$ is the design matrix consisting of the centered $\boldsymbol{X}_j's$. The entries of $\boldsymbol{X}_{A,\perp}^\top \boldsymbol{X}_{A,\perp}$ are $(n-1)s_{jk} = \sum_i (x_{ij} - \bar{x}_j)(x_{ik} - \bar{x}_k)$.

  PCA of the predictor:

  Conduct the spectral decomposition of $\boldsymbol{X}_{A,\perp}^T \boldsymbol{X}_{A,\perp}$. That is,

$$\boldsymbol{X}_{A,\perp}^T \boldsymbol{X}_{A,\perp} = P\Lambda P^T, \quad \Lambda = \begin{pmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ 0 & 0 & \ddots & 0 \\ 0 & 0 & \cdots & \lambda_p \end{pmatrix} P = \begin{pmatrix} \nu_1 & \nu_2 & \cdots & \nu_p \end{pmatrix},$$

where $\nu_j's$ are the orthonormal eigenvectors of $\boldsymbol{X}_{A,\perp}^T\boldsymbol{X}_{A,\perp}$ ordered in terms of the respective eigenvalues $\lambda_1 \geq \lambda_2 \cdots \geq \lambda_p$.

Write $\nu_j = (\nu_{j1}, \cdots, \nu_{jp})^T$. Then the jth Principal component scores $z_j := \boldsymbol{X}_{A,\perp}\nu_j$ are observed vector of new variables $z_j = \nu_{j1}(x_1 - \bar{x}_1) + \cdots + \nu_{jp}(x_p - \bar{x}_p)$. Therefore,

$$\|z_j\|^2 = \lambda_j,\ 1 \leq j \leq p \quad \& \quad z_j^T z_k = 0 \quad j \neq k$$

so that $z_j's$ are uncorrelated. ($\because \boldsymbol{X}_{A,\perp}P = (z_1, \cdots, z_p)$ by def)

Identification of sources of multicollinearity:

$$\lambda_j = 0 \iff z_j = 0$$

In other words, the observed values of the predictors satisfy the equation

$$\nu_{j1}(x_1 - \bar{x}_1) + \cdots + \nu_{jp}(x_p - \bar{x}_p) = 0$$

4.4 Working with Collinear Data

4.4.1 Principal Component Regression

Recall the definition of z_j in the PCA of the predictors:

$$z_j := \boldsymbol{X}_{A,\perp}\nu_j,$$

where $\nu_j's$ are the orthonomal eigenvectors of $\boldsymbol{X}_{A,\perp}^T\boldsymbol{X}_{A,\perp}$ ordered in terms of the respective eigenvalues $\lambda_1 \geq \lambda_2 \cdots \geq \lambda_p$.

One may rewrite the regression equation using the centered predictors

$$\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p = \beta_0^* + \beta_1(x_1 - \bar{x}_1) + \cdots + \beta_p(x_p - \bar{x}_p).$$

That is,

$$\begin{aligned}\boldsymbol{X}\beta &= \mathbf{1}\beta_0^* + \boldsymbol{X}_{A,\perp}\boldsymbol{\beta}_A = \mathbf{1}\beta_0^* + \boldsymbol{X}_{A,\perp}P\underbrace{P^T\boldsymbol{\beta}_A}_{\text{let } \boldsymbol{\alpha}}, \quad \text{where } \mathbf{1} = (1, \cdots, 1)^\top \\ &= \mathbf{1}\beta_0^* + \boldsymbol{Z}\alpha, \quad \boldsymbol{Z} = (\boldsymbol{z}_1, \cdots, \boldsymbol{z}_p)\end{aligned}$$

so that the regression model becomes

$$\boldsymbol{Y} = \mathbf{1}\beta_0^* + \boldsymbol{Z}\boldsymbol{\alpha} + \boldsymbol{\epsilon}.$$

Since each $\boldsymbol{z}_j$ is orthogonal to $\mathbf{1}$, the LSE for $\beta_0^*,\ \boldsymbol{\alpha}$ is

$$\hat{\beta}_0{}^* = \bar{Y} \quad \& \quad \hat{\boldsymbol{\alpha}} = (Z^T Z)^{-1} Z^T Y = \Lambda^{-1} Z^T Y.$$

Correspondingly, we have

$$\hat{\boldsymbol{\beta}}_A = P\hat{\boldsymbol{\alpha}} = \sum_{j=1}^{P} \hat{\alpha}_j \nu_j$$

$$\hat{\boldsymbol{\beta}}_0 = \hat{\beta}_0^* - \hat{\boldsymbol{\beta}}_1 \bar{x}_1 - \cdots - \hat{\boldsymbol{\beta}}_p \bar{x}_p = \bar{Y} - \hat{\boldsymbol{\beta}}_1 \bar{x}_1 - \cdots - \hat{\boldsymbol{\beta}}_p \bar{x}_p.$$

Practically, in particular in the presence of multicollinearity, i.e., $\lambda_j \approx 0$ for some $j's$ use a subset of the PCs $\boldsymbol{z}_j$ instead of the full set. Choose the first q PCs $\boldsymbol{z}_1, \cdots, \boldsymbol{z}_q$ with $q < p$ and fit the resulting model

$$Y = \mathbf{1}\beta_0^* + Z_q \alpha_q^* + \varepsilon, \quad \text{where } Z_q = (z_1, \cdots, z_q) \ \& \ \alpha_q^* = (\alpha_1, \cdots, \alpha_q)^T$$

4.4.2 Ridge Regression

Note that $\boldsymbol{X}\beta = \mathbf{1}\beta_0^* + \boldsymbol{X}_{A,\perp}\beta_A$ with $\beta^T = (\beta_0, \beta_A^T)$ and that β_0^* is estimated by $\bar{Y}$.

Ridge estimator

$$\hat{\beta}_{A,R}(k) = (\boldsymbol{X}_{A,\perp}^T \boldsymbol{X}_{A,\perp} + kI)^{-1} \boldsymbol{X}_{A,\perp}^T \boldsymbol{Y}$$

Add a positive constant $k > 0$ to the diagonal entries of $\boldsymbol{X}_{A,\perp}^T \boldsymbol{X}_{A,\perp}$ such that the resulting (added) matrix $\boldsymbol{X}_{A,\perp}^T \boldsymbol{X}_{A,\perp} + kI$ is invertible even if $\boldsymbol{X}_{A,\perp}^T \boldsymbol{X}_{A,\perp},\ (\boldsymbol{X}^T \boldsymbol{X})$ is not.

- The ridge estimator $\hat{\beta}_{A,R}$ is biased, i.e.

$$E(\hat{\beta}_{A,R}) \underset{\substack{\uparrow \\ \text{(HW!!)}}}{=} (\boldsymbol{X}_{A,\perp}^T \boldsymbol{X}_{A,\perp} + kI)^{-1} \boldsymbol{X}_{A,\perp}^T \boldsymbol{X}_{A,\perp} \beta_A \neq \beta_A,\ k > 0$$

  but it has smaller (total) variance.

- It does not give best fit (as the LSE does), but may do better job in out-of-sample prediction.

- The ridge estimator is defined to be the minimizer of this penalized least square estimation problem:

$$\|\boldsymbol{Y} - \bar{Y} \cdot \mathbf{1} - \boldsymbol{X}_{A,\perp}\beta_A\|^2 + k\|\beta_A\|^2 \quad \text{(HW)}$$

- Constrained optimization problem:

$$\min_{\boldsymbol{\beta}_A!} \|\boldsymbol{Y} - \bar{Y} \cdot \mathbf{1} - \boldsymbol{X}_{A,\perp}\boldsymbol{\beta}_A\|^2 \quad \text{subject to } \|\boldsymbol{\beta}_A\|^2 \leq d \cdots (*)$$
$$\iff \min_{\beta_A!} \|\boldsymbol{Y} - \bar{Y} \cdot \mathbf{1} - \boldsymbol{X}_{A,\perp}\boldsymbol{\beta}_A\|^2 + k(\|\boldsymbol{\beta}_A\|^2 - d) \cdots (**)$$

where $k > 0$, $d = \hat{\boldsymbol{\beta}}_A^T \left(I + k(\boldsymbol{X}_{A,\perp}^T \boldsymbol{X}_{A,\perp})^{-1} \right)^{-2} \hat{\boldsymbol{\beta}}_A$ and $\hat{\boldsymbol{\beta}}_A = (\boldsymbol{X}_{A,\perp}^T \boldsymbol{X}_{A,\perp})^{-1} \boldsymbol{X}_{A,\perp}^T \boldsymbol{Y}$: LSE

Why?

d is chosen such that the minimizer of the unconstrained optimization problem (**), $\hat{\boldsymbol{\beta}}_{A,R}(k)$ satisfies $\|\boldsymbol{\beta}_A\|^2 - d = 0$, i.e.

$$\begin{aligned}
\|\boldsymbol{\beta}_{A,R}\|^2 &= \boldsymbol{Y}^T \boldsymbol{X}_{A,\perp} (kI + \boldsymbol{X}_{A,\perp}^T \boldsymbol{X}_{A,\perp})^{-2} \boldsymbol{X}_{A,\perp}^T \boldsymbol{Y} \\
&= \hat{\boldsymbol{\beta}}_A^T (\boldsymbol{X}_{A,\perp}^T \boldsymbol{X}_{A,\perp})(kI + \boldsymbol{X}_{A,\perp}^T \boldsymbol{X}_{A,\perp})^{-2} (\boldsymbol{X}_{A,\perp}^T \boldsymbol{X}_{A,\perp}) \hat{\boldsymbol{\beta}}_A \\
&= \hat{\boldsymbol{\beta}}_A^T \left(I + k(\boldsymbol{X}_{A,\perp}^T \boldsymbol{X}_{A,\perp})^{-1} \right)^{-2} \hat{\boldsymbol{\beta}}_A = d,
\end{aligned}$$

and the constrained optimization problem (*) has a unique solution. Therefore, for any $\boldsymbol{\beta}_A$ with $\|\boldsymbol{\beta}_A\|^2 \leq d$

$$\begin{aligned}
\|\boldsymbol{Y} - \bar{Y} \cdot \mathbf{1} - \boldsymbol{X}_{A,\perp}\boldsymbol{\beta}_A\|^2 &\geq \|\boldsymbol{Y} - \bar{Y} \cdot \mathbf{1} - \boldsymbol{X}_{A,\perp}\boldsymbol{\beta}_A\|^2 + \underbrace{k(|\boldsymbol{\beta}_A\|^2 - d)}_{<0} \\
&\geq \|\boldsymbol{Y} - \bar{Y} \cdot \mathbf{1} - \boldsymbol{X}_{A,\perp}\hat{\boldsymbol{\beta}}_{A,R}\|^2 + k(|\hat{\boldsymbol{\beta}}_{A,R}\|^2 - d) \\
&= \|Y - \bar{Y} \cdot \mathbf{1} - \boldsymbol{X}_{A,\perp}\hat{\boldsymbol{\beta}}_{A,R}\|^2
\end{aligned}$$

i.e. $\hat{\boldsymbol{\beta}}_{A,R}$ minimizes the constrained problem (*).

- Note that $\|\hat{\boldsymbol{\beta}}_A\|^2 > d$ because

$$\begin{aligned}
d &= \hat{\boldsymbol{\beta}}_A^\top \underbrace{\left(I + k(\boldsymbol{X}_{A,\perp}^\top \boldsymbol{X}_{A,\perp})^{-1} \right)^{-2}}_{=P(I+k\Lambda^{-1})^{-2}P^\top} \hat{\boldsymbol{\beta}}_A \\
&= \sum_j \left(\frac{1}{1 + k\lambda_j^{-1}} \right)^2 (P^\top \hat{\boldsymbol{\beta}}_A)_j^2 \\
&= \sum_j \left(\frac{\lambda_j}{\lambda_j + k} \right)^2 (P^\top \hat{\boldsymbol{\beta}}_A)_j^2 < \sum_{j=1}^P \mathbf{1} \cdot (P^\top \hat{\boldsymbol{\beta}}_A)_j^2 = \|P^\top \hat{\boldsymbol{\beta}}_A\|^2 = \|\hat{\boldsymbol{\beta}}_A\|^2
\end{aligned}$$

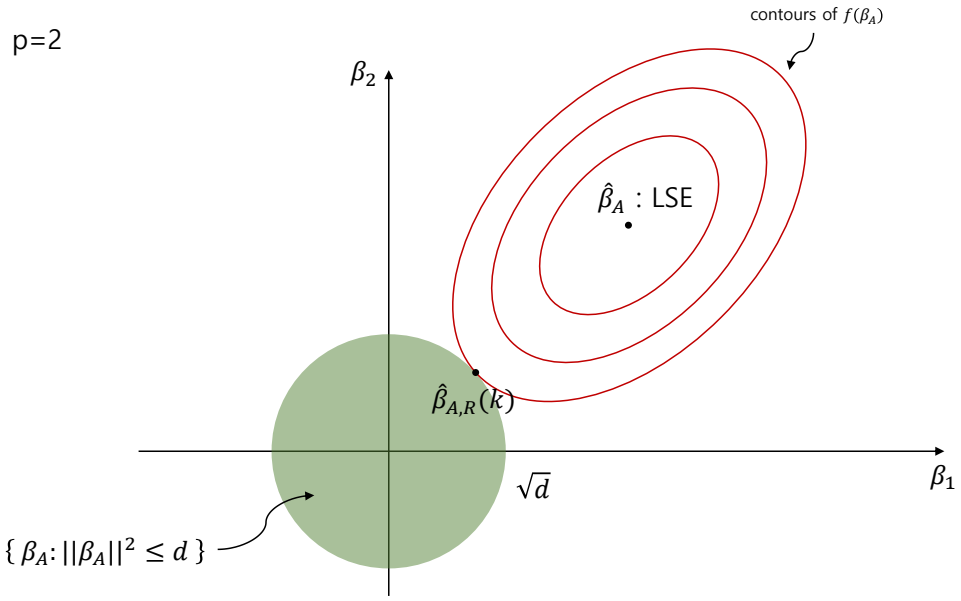
- Note that

$$\|\boldsymbol{Y}-\bar{Y}\cdot\mathbf{1}-\boldsymbol{X}_{A,\perp}\boldsymbol{\beta}_A\|^2=\|\boldsymbol{Y}-\bar{Y}\cdot\mathbf{1}-\boldsymbol{X}_{A,\perp}\hat{\boldsymbol{\beta}}_A\|^2+\underbrace{(\boldsymbol{\beta}_A-\hat{\boldsymbol{\beta}}_A)^\top\boldsymbol{X}_{A,\perp}^\top\boldsymbol{X}_{A,\perp}(\boldsymbol{\beta}_A-\hat{\boldsymbol{\beta}}_A)}_{=f(\boldsymbol{\beta}_A)}$$

- When $p=2$,

  - Geometry of Ridge

p=2



The penalty term $k\|\boldsymbol{\beta}_A\|^2$ in the penalized least squares criterion shrinks $\boldsymbol{\beta}$ toward 0 so that the ridge estimator $\hat{\boldsymbol{\beta}}_{A,R}(k)$ can be seen as a shrinkage estimator

Chapter 5

Variable Selection (chapter 9)

5.1 Motivation of Variable selection

Assume (true) regression model (of no intercept) with two covariates x_{i1} & x_{i2} for the data $\{(x_{i1}, x_{i2}, Y_i),\ i = 1, \cdots, n\}$:

$$Y_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i.$$

Here, $\sum_i x_{i1} = \sum_i x_{i2} = 0 \quad \& \quad \sum_i x_{i1}^2 = \sum_i x_{i2}^2 = 1$ for simplicity. We regard the model (1) as full model. In the full model (1), the LSE $\hat{\boldsymbol{\beta}}_F$ of $\beta = (\beta_1, \beta_2)^T$ is

$$\hat{\boldsymbol{\beta}}_F = (\hat{\beta}_{F,1}, \hat{\beta}_{F,2})^\top = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^\top Y \quad \text{with} \quad \boldsymbol{X} = \begin{pmatrix} x_{11} & x_{12} \\ \vdots & \\ x_{n1} & x_{n2} \end{pmatrix} \quad \text{and} \quad Y = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}$$

so the variance of $\hat{\beta}_{F,1}$ is

$$Var(\hat{\beta}_{F,1}) = \Big(\text{ the } (1,1) \text{ element of } (\boldsymbol{X}^\top\boldsymbol{X})^{-1} \Big)\sigma^2 = \frac{\sigma^2}{1 - r_{12}^2}$$

with (sample) correlation coefficient r_{12} between x_{i1} and x_{i2}. And $\hat{\beta}_{F,1}$ is unbiased, that is, $E(\hat{\beta}_{F,1}) = \beta_1$.

Consider the subset model (using only x_{i1})

$$Y_i = \beta_1 x_{i1} + \epsilon_i, \quad i = 1, \cdots, n$$

The LSE $\hat{\beta}_{S,1}$ *of* β_1 under the subset model is

$$\hat{\beta}_{S,1} = \sum_{i=1}^{n} x_{i1} Y_i,$$

thus,

$$E(\hat{\beta}_{S,1}) = \sum_{i=1}^{n} x_{i1} \underbrace{E(Y_i)}_{\beta_1 x_{i1} + \beta_2 x_{i2}} = \beta_1 + r_{12}\beta_2 \neq \beta_1,$$

i.e. $\hat{\beta}_{S,1}$ is biased and $Var(\hat{\beta}_{S,1}) = \sigma^2$.

In order to compare estimations of β_1 from the full and subset models, we compute the mean squared errors (MSE) of $\hat{\beta}_{F,1}$ and $\hat{\beta}_{S,1}$

$$\text{MSE}(\hat{\beta}_{F,1}) = E(\hat{\beta}_{F,1} - \beta_1)^2 = Var(\hat{\beta}_{F,1}) = \sigma^2 \frac{1}{1 - r_{12}^2}$$
$$\text{MSE}(\hat{\beta}_{S,1}) = E(\hat{\beta}_{S,1} - \beta_1)^2 = Var(\hat{\beta}_{S,1}) + (\text{bias}\,(\hat{\beta}_{S,1}))^2.$$

The subset model will estimate more precisely whenever

$$\text{MSE}(\hat{\beta}_{S,1}) < \text{MSE}(\hat{\beta}_{F,1}), \quad \text{i.e., } \frac{|\beta_2|}{\sigma} < \frac{1}{\sqrt{1 - r_{12}^2}}$$

$\Rightarrow$

- If $|\beta_2| = 0$, then the subset model is understood as the true model so that the subset model is preferred.
- When $r_{12}^2 \approx 1$, i.e., "multicollinearity" is severe (even if $\beta_2 \neq 0$), the subset model is almost always better than the full model.
- For any value of r_{12}^2, the subset model will be preferred if $|\beta_2|/\sigma < 1/\sqrt{1 - r_{12}^2}$
- If the β's and σ^2 were known, deletion of variables with small $|\beta|/\sigma$ would be desirable.

5.2 Criteria for selecting subsets (model selection)

To decide if one subset is better than another, we need some criteria for subset selection, which is often called "model selection"

Adjusted R^2

How to select useful covariaes?

One may consider the coefficient of determination R^2, which turn out to be SSR/SST. But, this is not a good criterion because it is nondecreasing as a new predictor enters the model.

Suppose that the totality of all predictors at hands are $x_1, \cdots, x_p$. We want to select a subset S of the index set $1, \cdots, p$. Let $|S|$ denote the cardinality of the set S and $q = |S| + 1$. Let $R^2(S)$ and $\text{SSE}(S)$ denote the coefficient of determination and the residual sum of squares, respectively, when Y is regressed on $x_j,\ j \in S$ with an intercept term: $R^2(S) = \text{SSR}(x_j,\ j \in S)/\text{SST}$

We define adjusted R^2 (for the subset S) as

$$R_a^2(S) = 1 - \frac{(n-1)}{(n-q)}(1 - R^2(S))$$

and mean squared residual as

$$\text{MSE}(S) = \frac{\text{SSE}(S)}{n-q}$$

In order to select as optimal subset (model), we choose the model S that maximizes $R_a^2(S)$ (w.r.t $S \subset \{1, \cdots, p\}$)). Notice that maximizing $R_a^2(S)$ is equivalent to minimizing $\text{MSE}(S)$

Mallow's C_p, AIC and BIC

There are some other criteria for model selection

- **Mallow's C_p:**
$$C_p(S) = \frac{\text{SSE}(S)}{\hat{\sigma}^2} - n + 2(|S| + 1),$$
where $\hat{\sigma}^2 = \text{SSE}(1, \ldots, p)/(n - p - 1)$ is obtained from the full model.

- **Akaike Information Criterion(AIC):**
$$\text{AIC}(S) = \log \text{SSE}(S) + \frac{2}{n}((|S| + 1)$$
This can be obtained from the general definition of AIC using the likelihood. (HW!!)

- **Bayesian Information Criterion (BIC):**
$$\text{BIC}(S) = \log \text{SSE}(S) + \frac{\log n}{n}(|S| + 1)$$

Some remarks

- The model selection criteria, $C_p(S)$, AIC(S) and BIC(S) take the form

$$\text{(Goodness-of-fit) + (Model complexity)}$$

- BIC penalizes larger (more complex) models more heavily than AIC when $\log n > 2$, so that it prefers smaller models in comparison with AIC.

5.3 Computational techniques

All possible regressions

Suppose that we have p predictors, $x_1, \cdots, x_p$ at hands. In order to select a subset of predictor $x_1, \cdots, x_p$, one may fit all submodels

$$Y_i = \beta_0 + \sum_{j \in S} \beta_j x_{ij} + \varepsilon_i, \quad S \subset \{1, \cdots, p\}$$

and compute a selection criterion such as C_p, AIC or BIC. But, this requires fitting

$$\binom{p}{0} + \binom{p}{1} + \cdots + \binom{p}{p} = 2^p$$

regression (sub)models so that the number 2^p of the regression models to be conducted increases rapidly as p increases.

Are there simpler methods than all possible regression?

5.3.1 Forward Selection

(i) Step 0: Consider that there are no predictors in the model other than the intercept, i.e. set

$$S_0 = \phi \quad \text{and} \quad M_0 = \mathcal{C}_{\mathbf{1}} = \{\beta_0 \mathbf{1} : \beta_0 \in R\}.$$

(ii) Step $k (k = 1, 2, \ldots)$: Add to the model the predictor that has the highest (sample) partial correlation (in absolute value). With response, adjusting for the predictors in the model M_{k-1} in the previous step. That is, find an index $j = j_k$ that maximize $r_{k,j}^2$, where

$$r_{k,j} = \frac{(\boldsymbol{x}_j - \Pi(\boldsymbol{x}_j | M_{k-1}))^\top (\boldsymbol{Y} - \Pi(\boldsymbol{Y} | M_{k-1}))}{\|\boldsymbol{x}_j - \Pi(\boldsymbol{x}_j | M_{k-1})\| \cdot \|\boldsymbol{Y} - \Pi(\boldsymbol{Y} | M_{k-1})\|}$$

and update the subset by $S_k = S_{k-1} \cup \{j_k\}$ and the column space by $M_k = \mathcal{C}_{\mathbf{1}, (\boldsymbol{x}_\ell : \ell \in S_k)}$.

The procedure stops at the kth step and choose $S^* = S_{k-1}$ as final model if a **stopping rule** is met.

Stopping rules: e.g.

1. the cardinality $|S_k|$ exceeds a predetermined size p^*
2. the partial F statistic $F_{k,j}$ for testing $H : \beta_{jk} = 0$ is less than predetermined number, say, F_{IN}, where
$$F_{k,j} = \frac{SSR(X_j|X_l;\ l \in S_{k-1})}{SSE(X_l, l \in S_k)/(n-k-1)}$$

Remark 5.3.1 *1. When $k = 1$, $r_{1,j}$ is the (sample) correlation between x_j and Y.*

2. The $r_{k,j}$ is called the (sample) partial correlation between x_j and Y given $(x_\ell : \ell \in S_{k-1})$.

3. It can be shown that
$$F_{k,j} = (n-k-1) \cdot \frac{r_{k,j}^2}{1-r_{k,j}^2} \quad \textit{(HW!!).}$$
Thus, choosing j that maximizing $r_{k,j}^2$ is equivalent to choosing j that maximizes $F_{k,j}$. It is also equivalent to choosing j that maximizes the coefficient of determination R^2 when the variable x_j is added to the existing group $(x_l : l \in S_{k-1})$. (HW!!)

5.3.2 Backward elimination

Similar idea, but in an opposite direction can be applied.

(i) Step 0: Start with the full model, i.e. $S_0 = \{1, \cdots, p\}$.

(ii) Step $k (k = 1, 2, \ldots)$: Remove the variable that meets the following equivalent criteria:

- it has the smallest partial correlation (in absolute value) with response, adjusting for all the other variables left in the model.
- it has the smallest partial F statistic ($F_{k,j}$) value.
- removing the variable gives the smallest changes in R^2.

The procedure continue removing one variable at a time until a stopping rule is met.

Stopping rule: e.g.,

1. stop with a subset of predetermined size p^*
2. stop if the partial F statistic is larger than or equal to predetermined number, say F_{OUR}.

5.3.3 Stepwise Regression

It is a combination of forward selection and backward elimination.

(i) Step 0,1 : Do as in of the forward selection

(ii) Step k ($k \geq 2$):

Find j_k as in the kth step of the forward selection. If $F_{k,j_k} < F_{IN}$, then choose $S^* = S_{k-1}$ as a final model. If $F_{k,j_k} < F_{IN}$ then identify a set of $i \in S_{k-1}$, denoted by D_{k-1}, such that

$$\frac{\mathrm{SSR}\Big(x_i \;\Big|\; \{x_\ell : \ell \neq i, \ell \in S_{k-1}\}, x_{jk}\Big)}{\mathrm{SSE}\Big(\{x_\ell : \ell \in S_{k-1}\}, x_{jk}\Big)/(n-k-1)} < F_{OUT},$$

then set $S_k = S_{k-1} \cup j_k - D_{k-1}$.

5.4 Least Absolute Shrinkage Selection Operator

Recall that $\boldsymbol{X}\boldsymbol{\beta} = \mathbf{1}\beta_0^* + \boldsymbol{X}_{A,\perp}\boldsymbol{\beta}_A$ with $\boldsymbol{\beta} = (\beta_0, \beta_A^T)^T$ and that β_0^* is estimated by $\bar{Y}$.

LASSO (Tibshirani, JRSSB 1991)

$$\hat{\boldsymbol{\beta}}_{A,L} \equiv \hat{\boldsymbol{\beta}}_{A,L}^{(\lambda)} = \arg\min_{\boldsymbol{\beta}_A} \|\boldsymbol{Y} - \bar{Y}\cdot\mathbf{1} - \boldsymbol{X}_{A,\perp}\boldsymbol{\beta}_A\|^2 + \lambda\|\boldsymbol{\beta}_A\|_1 \cdots (*) \tag{0.1}$$

for some λ_0, where $\|a\|_1 = \sum_{j=1}^p |a_j|$ is the l_1 norm for a vector $a = (a_1, \cdots, a_p)$.

Remark 5.4.1 *1. Instead of $\|\boldsymbol{\beta}_A\|^2$ in ridge estimation, LASSO uses $\|\boldsymbol{\beta}_A\|_1$ as a penalty for large β_j's.*

2. *Take the penalty constant $\lambda = 0$ corresponds to the ordinary least squares estimatior.*

3. *And choosing $\lambda = \infty$ would give $\hat{\boldsymbol{\beta}}_{A,L} = 0$.*

4. *Equivalently, the LASSO estimator $\hat{\beta}_{A,L}$ minimizes the constrained least square criterion:*

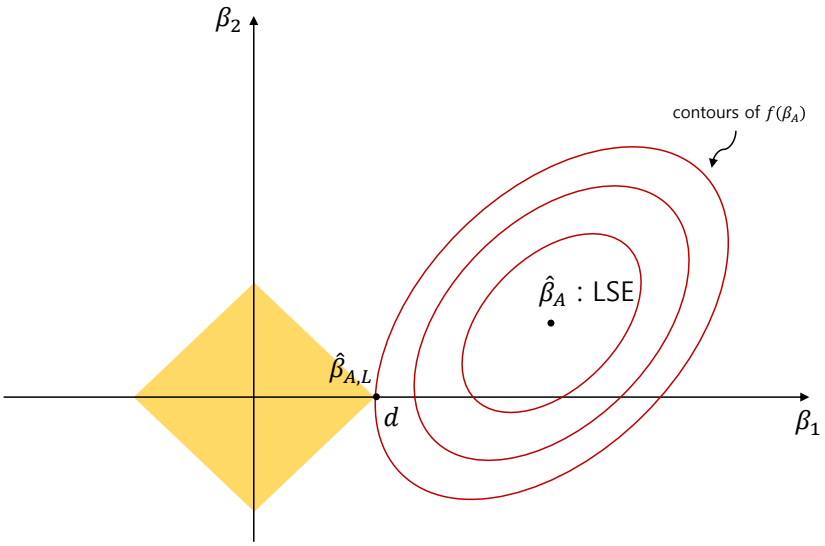
$$\min_{\boldsymbol{\beta}_A}! \|Y - \bar{Y} \cdot \mathbf{1} - \boldsymbol{X}_{A,\perp}\beta_A\|^2 \quad \textit{subject to} \quad \|\beta_A\| \leq d,$$

subject to $\|\beta_A\|_1 \leq d$, where $d = \|\hat{\beta}^{(\lambda)}_{A,L}\|_1$ and $\hat{\boldsymbol{\beta}}^{(\lambda)}_{A,L}$ is the minimizer of the LASSO problem(0.1)

5. $\|\hat{\boldsymbol{\beta}}_A\|_1 > \|\hat{\beta}^{(\lambda)}_{A,L}\|_1 = d$ *where* $\hat{\boldsymbol{\beta}}_A$*:LSE.*

- Geometry of LASSO

p=2



constrained region: $\{\beta : \|\beta\|_1 = |\beta_1| + |\beta_2| \leq d\}$

Shrinkage and selection by LASSO

1. LASSO as a shrinkage estimator: the penalty term $\lambda\|\beta_A\|_1$ in the penalized least squares criterion shrinks $\hat{\boldsymbol{\beta}}_A$ toward 0.

2. LASSO as a variable selector: when some components of $\hat{\beta_A}$ are small, the l_1 penalty makes the components by exactly zero, so that LASSO exactly performs both the selection of predictors and the estimation of the regression parameters in one step.

LASSO for the orthogonal design

When all columns of the design matrix $\boldsymbol{X}_{A,\perp} = (x_{i1} - \bar{x}_1, x_{i2} - \bar{x}_2, \cdots, x_{ip} - \bar{x}_p)$ are orthogonal, i.e. $\sum_{i=1} n(x_{ij} - \bar{x}_j)(x_{ik} - \bar{x}_k) = 0$ for all $j \neq k$, the LASSO estimator can be given explicitly.

Define $K_j := \sum_{i=1} n(x_{ij} - \bar{x}_j)^2$ and let $\hat{\boldsymbol{\beta}}_A \equiv (\hat{\boldsymbol{\beta}}_1, \cdots, \hat{\boldsymbol{\beta}}_p)^T$ denote the LSE. Then,

$$\hat{\beta}_{j,L} = sgn(\hat{\boldsymbol{\beta}}_j) \cdot \left(\|\hat{\boldsymbol{\beta}}_j\| - \frac{\lambda}{2K_j} \right)_+,$$

where $x_+ = \max\{x, 0\}$ and $\hat{\boldsymbol{\beta}}_{A,L} = (\hat{\beta}_{1,L}, \cdots, \hat{\beta}_{p,L})$.

$\because$

Recall that

$$f(\boldsymbol{\beta}_A) = (\boldsymbol{\beta}_A - \hat{\boldsymbol{\beta}_A})^\top \boldsymbol{X}_{A,\perp}^\top \boldsymbol{X}_{A,\perp} (\boldsymbol{\beta}_A - \hat{\boldsymbol{\beta}_A}) = \sum_{j=1}^{p} K_j (\beta_j - \hat{\beta}_j)^2$$

Therefore, the LASSO objective function is

$$(\text{constant term}) + f(\beta_A) + \lambda \sum_{i=1}^{p} |\beta_j| = (\text{constant term}) + \sum_{j=1}^{p} \{K_j (\beta_j - \hat{\boldsymbol{\beta}}_j)^2 + \lambda |\beta_j|\}.$$

Thus, the minimization of the LASSO objective function can be done by componentwise minimization of $K_j(\beta_j - \hat{\boldsymbol{\beta}}_j)^2 + \lambda|\beta_j|$. Note that

$$\begin{aligned} K_j(\beta_j - \hat{\boldsymbol{\beta}}_j)^2 + \lambda|\beta_j| &= K_j \left((\beta_j - \hat{\boldsymbol{\beta}}_j)^2 + \frac{\lambda}{K_j} |\beta_j| \right) \\ &= \left[\left(\beta_j - \left(\hat{\boldsymbol{\beta}}_j - sgn(\beta_j) \frac{\lambda}{2K_j} \right) \right)^2 + (\text{constant term}) \right] \end{aligned}$$

so that it is minimized by $\hat{\beta}_{j,L} = \hat{\beta}_j - sgn(\beta_j)\lambda/(2K_j)$ when $|\hat{\beta}_j| > \lambda/2K_j$ and 0 when $|\hat{\beta}_j| \leq \lambda/2K_j$ (HW!!)

LARS algorithm

Contrary to the case of ridge estimation, the LASSO estimator does not have a closed form except in the orthonomal design case.

The LARS algorithm (Efron, Hastie, Johnstone & Tibshirani, AoS 2004) give the complete set of the LASSO estimates for all $0 < \lambda < \infty$. The R package calle "LARS" and its manual are available from the webpage

http:// www-stat.stanford.edu/ hastie/Papers/#LARS

LASSO for high-dimensional data

1. "high dimensionality"
   - What is "high-dimensionality"?
     In our linear regression problem, the number of covariates, p is large, even $p > n$
   - "high-dimensionality" in modern applications: e.g. gene expression data, econometrics, etc.
   - Notice that the LSE method breaks down when $p > n$
2. The LASSO is very useful and effective in high-dimensional situations as it does estimation and (variable) selection at one step.
3. Figure