Lecture 9: Some Applications in Statistics

The linear statistical model

Given a random vector $\mathbf{x} = (\mathbf{x}_i)$ with **expected value** $\mathrm{E}\,\mathbf{x} = \boldsymbol{\mu} = (\boldsymbol{\mu}_i)$, its **covariance matrix** is

$$\mathrm{Cov}\,\mathbf{x} = \mathrm{E}\,\{(\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^T\} = \left[\mathrm{E}\,(\mathbf{x}_i-\boldsymbol{\mu}_i)(\mathbf{x}_j-\boldsymbol{\mu}_j)\right] \; .$$

A **linear statistical model** is

$$\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon} \qquad (1)$$

- $\mathbf{y} \in \mathbb{R}^n$ is **observed**, or measured in some experimental set-up,
- the **parameters** $\boldsymbol{\beta} \in \mathbb{R}^p$ are unknown,
- the matrix $X \in \mathbb{R}^{n\times p}$ (the **design matrix**) is given, and
- $\boldsymbol{\varepsilon} \in \mathbb{R}^n$ is a random vector representing the **errors** of observing $\mathbf{y}$, which are not systematic, i.e.,

$$\mathrm{E}\,\boldsymbol{\varepsilon} = \mathbf{0}\;,\;\; \mathrm{Cov}\,\boldsymbol{\varepsilon} = V^2\;,\;\; \text{assumed known.} \qquad (2)$$

The linear statistical model (cont'd)

The story so far:

$$\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{1}$$

$$\mathrm{E}\,\boldsymbol{\varepsilon} = \mathbf{0}\ ,\ \ \mathrm{Cov}\,\boldsymbol{\varepsilon} = V^2\ , \tag{2}$$

From (1)–(2) it follows that

$$\mathrm{E}\,\mathbf{y} = X\boldsymbol{\beta}\ ,\ \ \mathrm{Cov}\,\mathbf{y} = V^2\ . \tag{3}$$

This model has several names, including: **linear statistical model** (or just **linear model**), **linear regression** and the **Gauss–Markov model**. We denote this model by $(\mathbf{y}, X\boldsymbol{\beta}, V^2)$.

The **problem**: estimate a **linear function** of the **parameters**, say

$$B\boldsymbol{\beta}\ ,\ \text{for a given matrix } B \in \mathbb{R}^{m\times p}\ , \tag{4}$$

from the observed $\mathbf{y}$ (the problem of estimating the variance V^2, if unknown, is not treated here.)

The linear statistical model (cont'd)

A **linear estimator** (abbreviated **LE**) of $B\boldsymbol{\beta}$ is

$$A\mathbf{y} \ , \quad \text{for some } A \in \mathbb{R}^{m\times n} \ . \tag{5}$$

It is **unbiased** (abbreviated **LUE**) if

$$\mathrm{E}\{A\mathbf{y}\} = B\boldsymbol{\beta} \ , \quad \text{for all } \boldsymbol{\beta} \in \mathbb{R}^p \ , \tag{6}$$

and it is the **best linear unbiased estimator** (**BLUE**) if its variance is minimal, in some sense, among all LUE's. In general, not all linear functions have LUE's.

The function $B\boldsymbol{\beta}$ is called **estimable** if it has an **LUE**, i.e., if there is a matrix $A \in \mathbb{R}^{m\times n}$ such that (6) holds.

The **unbiasedness condition** (6) reduces to an identity

$$AX\boldsymbol{\beta} = B\boldsymbol{\beta} \ , \ \text{for all } \boldsymbol{\beta} \ , \ \text{or equivalently, } AX = B \ , \tag{7}$$

4 main cases of the model $(\mathbf{y}, X\boldsymbol{\beta}, V^2)$

There are 2 cases for the **design matrix** $X \in \mathbb{R}_r^{n\times p}$:

(A) X is of **full column rank** $(r = p)$, or

(B) X is of **rank** $r < p$,

and 2 cases for the **covariance matrix** V^2 (which is PSD):

(1) V is **nonsingular**, i.e. V^2 is **positive definite** (PD), or

(2) V is **singular**.

giving 4 cases for the model, (A1), (B1), (A2) and (B2).

The simplest case is studied next.

X full column rank, V nonsingular

Consider the model $(\mathbf{y}, X\boldsymbol{\beta}, V^2)$ with V **nonsingular**, and the $n \times p$ matrix X is of **full column rank**, i.e., $R(X^T) = \mathbb{R}^p$.

Then any linear function $B\boldsymbol{\beta}$ is estimable. In particular, for $B = I$ the linear equation (7) reduces to $AX = I$, and we conclude that $A\mathbf{y}$ is an LUE of $\boldsymbol{\beta}$ whenever A is a left–inverse of X. The set of LUE's of $\boldsymbol{\beta}$ is therefore

$$\mathrm{LUE}(\boldsymbol{\beta}) = \{X^{(1)}\mathbf{y} : \; X^{(1)} \in X\{1\}\} \, .$$

and the **minimum–norm LUE** of $\boldsymbol{\beta}$ is

$$\widehat{\boldsymbol{\beta}} = \left(X^T X\right)^{-1} X^T \mathbf{y} = X^\dagger \mathbf{y} \, . \tag{8}$$

Without loss of generality we can assume

$$V^2 = \sigma^2 I$$

i.e., the errors have **equal variances** and are **uncorrelated**.

The Gauss–Markov Theorem

Theorem. Consider the linear model $(\mathbf{y}, X\boldsymbol{\beta}, \sigma^2 I)$ with X of full column–rank. Then for any $B \in \mathbb{R}^{m\times p}$:

(a) The linear function $B\boldsymbol{\beta}$ is estimable.

(b) The estimator $B\widehat{\boldsymbol{\beta}} = BX^{\dagger}\mathbf{y}$ is BLUE in the sense that

$$\operatorname{Cov} A\mathbf{y} \ \succcurlyeq \ \operatorname{Cov} B\widehat{\boldsymbol{\beta}} \tag{9}$$

for any other LUE $A\mathbf{y}$ of $B\boldsymbol{\beta}$.

(c) The BLUE $B\widehat{\boldsymbol{\beta}} = BX^{\dagger}\mathbf{y}$ belongs to the class of estimators

$$\mathcal{E}(X) := \{A\mathbf{y} : \ A = KX^T \ , \text{ for some matrix } K\} \ . \tag{10}$$

If $A\mathbf{y}$ is any LUE in $\mathcal{E}(X)$ (i.e. the rows of A are in $R(X)$) then

$$A\mathbf{y} = B\widehat{\boldsymbol{\beta}} \quad \text{with probability 1.} \tag{11}$$

Proof

(a) was shown above.

(b) Let $A\mathbf{y}$ be any LUE of $B\boldsymbol{\beta}$. Then:

(b1) The covariance of $A\mathbf{y}$ is $\operatorname{Cov} A\mathbf{y} = \sigma^2 A A^T$.

(b2) The covariance of $B\widehat{\boldsymbol{\beta}}$ is

$$\begin{aligned}\operatorname{Cov} B\left(X^T X\right)^{-1} X^T\mathbf{y} &= \sigma^2 B\left(X^T X\right)^{-1} B^T \\ &= \sigma^2 A X\left(X^T X\right)^{-1} X^T A^T \ , \ (\because \ B = AX) \ .\end{aligned}$$

$$\therefore \ \operatorname{Cov} A\mathbf{y} - \operatorname{Cov} B\widehat{\boldsymbol{\beta}} = \sigma^2 A\left(I - X\left(X^T X\right)^{-1} X^T\right) A^T \ . \qquad (12)$$

(c) The estimate $BX^\dagger\mathbf{y}$ is in $\mathcal{E}(X)$ since $X^\dagger = (X^T X)^\dagger X^T$. Then (11) follows from

$$\text{RHS}(12) = \sigma^2 A P_{N(X^T)} A^T = O \ ,$$

if $A = KX^T$ for some K. $\square$

The Gauss-Markov Theorem for functionals

Consider the problem of estimating **linear functionals** $\langle \mathbf{b}, \boldsymbol{\beta} \rangle$. A linear estimate $\langle \mathbf{a}, \mathbf{y} \rangle$ is in the class $\mathcal{E}(X)$ if and only if $\mathbf{a} \in R(X^T)$. The G–M Theorem then reduces to:

Corollary. Let $(\mathbf{y}, X\boldsymbol{\beta}, \sigma^2 I)$ and X be of full column rank. Then for any $\mathbf{b} \in \mathbb{R}^p$:

(a) The linear functional $\langle \mathbf{b}, \boldsymbol{\beta} \rangle$ is estimable.

(b) The estimator $\langle \mathbf{b}, \widehat{\boldsymbol{\beta}} \rangle = \langle \mathbf{b}, BX^\dagger \mathbf{y} \rangle$ is BLUE in the sense that

$$\mathrm{Var}\, \langle \mathbf{a}, \mathbf{y} \rangle \;\geq\; \mathrm{Var}\, \langle \mathbf{b}, \widehat{\boldsymbol{\beta}} \rangle$$

for any other LUE $\langle \mathbf{a}, \mathbf{y} \rangle$ of $\langle \mathbf{b}, \widehat{\boldsymbol{\beta}} \rangle$.

(c) If $\langle \mathbf{a}, \mathbf{y} \rangle$ is any LUE of $\langle \mathbf{b}, \widehat{\boldsymbol{\beta}} \rangle$ with $\mathbf{a} \in R(X^T)$ then $\langle \mathbf{a}, \mathbf{y} \rangle = \langle \mathbf{b}, \widehat{\boldsymbol{\beta}} \rangle$ with probability 1. □

The general $(\mathbf{y}, X\boldsymbol{\beta}, V^2)$

Theorem (Generalized Gauss–Markov Theorem). Let $(\mathbf{y}, X\boldsymbol{\beta}, V^2)$ be a linear model, and let $\langle \mathbf{b}, \boldsymbol{\beta} \rangle$ be any estimable functional. Then:

(a) $\langle \mathbf{b}, \boldsymbol{\beta} \rangle$ has a unique BLUE $\langle \mathbf{b}, \widetilde{\boldsymbol{\beta}} \rangle$ where

$$\widetilde{\boldsymbol{\beta}} = X^{\dagger} \left(I - (V P_{N(X^T)})^{\dagger} V \right)^T \mathbf{y} \ . \tag{1}$$

(b) $\widetilde{\boldsymbol{\beta}} \in R(X^T)$, and if $\boldsymbol{\beta}^*$ is any other LUE in $R(X^T)$,

$$\operatorname{Cov} \boldsymbol{\beta}^* \ \succcurlyeq \ \operatorname{Cov} \widetilde{\boldsymbol{\beta}} \ .$$

Regularization

Let $A \in \mathbb{C}_r^{m\times n}$ and let $\{\mathbf{u}_1, \ldots, \mathbf{u}_r\}$ and $\{\mathbf{v}_1, \ldots, \mathbf{v}_r\}$ be o.n. bases of $R(A^*)$ and $R(A)$, respectively, related by,

$$A\,\mathbf{v}_i = \sigma_i\,\mathbf{u}_i\ ,\ \text{ and } A^*\mathbf{u}_i = \sigma_i\,\mathbf{v}_i\ ,\ \ i \in \overline{1,r}\ .$$

Consider the equation

$$A\mathbf{x} = \mathbf{b} \tag{1}$$

where $\mathbf{b} \in R(A)$ is

$$\mathbf{b} = \sum_{i=1}^{r} \beta_i \mathbf{v}_i\ .$$

The **least–norm solution** is

$$\mathbf{x} = A^\dagger \mathbf{b} = \sum_{i=1}^{r} \frac{\beta_i}{\sigma_i} \mathbf{u}_i \tag{2}$$

and is **sensitive** to **errors** ε in the **smaller singular values**, as seen from

$$\frac{1}{\sigma + \varepsilon} \approx \frac{1}{\sigma} - \frac{1}{\sigma^2}\varepsilon + \frac{1}{\sigma^3}\varepsilon^2 + \cdots \tag{3}$$

Regularization (cont'd)

Instead of (2), consider the **approximate solution**

$$\mathbf{x}(\lambda) = (A^*A + \lambda I)^{-1}A^*\mathbf{b} = \sum_{i=1}^{r} \frac{\sigma_i\,\beta_i}{\sigma_i^2 + \lambda}\,\mathbf{u}_i \tag{4}$$

where λ is positive. It is **less sensitive** to errors in the singular values, as shown by

$$\frac{(\sigma+\varepsilon)}{(\sigma+\varepsilon)^2+\lambda} \approx \frac{\sigma}{\sigma^2+\lambda} - \frac{\sigma^2-\lambda}{(\sigma^2+\lambda)^2}\varepsilon + \frac{\sigma(\sigma^2-3\lambda)}{(\sigma^2+\lambda)^3}\varepsilon^2 + \cdots \tag{5}$$

where the choice $\lambda = \sigma^2$ gives

$$\frac{(\sigma+\varepsilon)}{(\sigma+\varepsilon)^2+\lambda} \approx \frac{1}{2\sigma} - \frac{1}{4\sigma^3}\varepsilon^2 + \cdots$$

Ridge regression

Consider the **linear model**

$$\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{1}$$

with $X \in \mathbb{R}_p^{n\times p}$ (full column rank), and the **error** $\varepsilon \sim N(\mathbf{0}, \sigma^2 I)$.

If $X^T X$ is **ill–conditioned**, then the **BLUE** of $\boldsymbol{\beta}$

$$\widehat{\boldsymbol{\beta}} = \left(X^T X\right)^{-1} X^T \mathbf{y} \tag{2}$$

is unsatisfactory. To see this, consider the **SVD** of X,

$$U^T X V = \Lambda = \begin{bmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_p \\ 0 & \cdots & 0 \\ \vdots & & \vdots \\ 0 & \cdots & 0 \end{bmatrix}, \quad \text{where the } \textbf{singular values} \text{ are denoted by } \lambda_i \tag{3}$$

Ridge regression (cont'd)

The transformation

$$\mathbf{z} := U^T \mathbf{y} \ , \ \boldsymbol{\gamma} = V^T \boldsymbol{\beta} \ , \ \boldsymbol{\nu} = U^T \boldsymbol{\varepsilon} \ . \tag{4}$$

takes the model (1) into

$$\mathbf{z} = \Lambda \boldsymbol{\gamma} + \boldsymbol{\nu} \tag{5}$$

where $\boldsymbol{\nu} \sim N(\mathbf{0}, \sigma^2 I)$ ($\because$ V is **orthogonal**), and the **parameters** to be **estimated** are $\boldsymbol{\gamma} = (\gamma_i)$. The components z_i of $\mathbf{z}$ are also normal

$$z_i \sim N(\lambda_i \gamma_i, \sigma^2) \ , \quad i \in \overline{1, p} \ , \tag{6a}$$

$$z_i \sim N(0, \sigma^2) \ , \quad i \in \overline{p+1, n} \ . \tag{6b}$$

For $i \in \overline{1, p}$, the BLUE of γ_i is

$$\widehat{\gamma}_i = \frac{z_i}{\lambda_i} \ , \ \text{with variance } \operatorname{Var} \widehat{\gamma}_i = \operatorname{E} \left(\frac{z_i}{\lambda_i} - \gamma_i \right)^2 = \frac{\sigma^2}{\lambda_i^2} \tag{7}$$

Dropping the U out of the BLUE

The **ridge regression estimator** (abbreviated **RRE**) of $\boldsymbol{\beta}$ is

$$\widehat{\boldsymbol{\beta}}(k) = \left(X^T X + kI\right)^{-1} X^T \mathbf{y} \ , \tag{8}$$

where k is a positive parameter. The RRE is a family of estimators $\{\widehat{\boldsymbol{\beta}}(k) : \ k > 0\}$, parameterized by k. with the BLUE for $k = 0$.

For the transformed model (4), the RRE of $\boldsymbol{\gamma}$ is

$$\widehat{\boldsymbol{\gamma}}(k) = \left(\Lambda^T \Lambda + kI\right)^{-1} \Lambda^T \mathbf{z} \ ,$$

and for $i \in \overline{1,p}$,

$$\widehat{\gamma}_i(k) = \frac{\lambda_i z_i}{\lambda_i^2 + k} \ . \tag{9}$$

The RRE **shrinks** every component of the observation vector **z**., by a **factor**

$$c(\lambda_i, k) = \frac{\lambda_i}{\lambda_i^2 + k} \ , \tag{10}$$

The MSE of the RRE

If $\boldsymbol{\beta}^*$ is an estimator of a parameter $\boldsymbol{\beta}$, its

(a) **bias** is $\text{bias}(\boldsymbol{\beta}^*) = \text{E}\,\boldsymbol{\beta}^* - \boldsymbol{\beta}$, and its

(b) **mean square error**(**MSE**) is $\text{MSE}(\boldsymbol{\beta}^*) = \text{E}\,(\boldsymbol{\beta}^* - \boldsymbol{\beta})^2$

which is equal to variance of $\boldsymbol{\beta}^*$ if $\boldsymbol{\beta}^*$ is unbiased.

The RRE (8) is biased, $\text{bias}(\widehat{\boldsymbol{\gamma}}(k)) = -k\left(\Lambda^T\Lambda + kI\right)^{-1}\boldsymbol{\gamma}$,

with

$$\text{bias}(\widehat{\gamma}_i(k)) = -k\frac{\gamma_i}{\lambda_i^2 + k}\,, \quad i \in \overline{1,p}\,.$$

$$\text{Var}(\widehat{\gamma}_i(k)) = \frac{\lambda_i^2\sigma^2}{(\lambda_i^2 + k)^2}\,,$$

$$\begin{aligned}\text{MSE}(\widehat{\boldsymbol{\gamma}}(k)) &= \sum_{i=1}^{p}\frac{\lambda_i^2\sigma^2}{(\lambda_i^2 + k)^2} + \sum_{i=1}^{p}\frac{k^2\gamma_i^2}{(\lambda_i^2 + k)^2} \\ &= \sum_{i=1}^{p}\frac{\lambda_i^2\sigma^2 + k^2\gamma_i^2}{(\lambda_i^2 + k)^2}\,. \qquad (11)\end{aligned}$$

An RRE with smaller MSE than the BLUE

$$\text{MSE}(\widehat{\boldsymbol{\gamma}}(k)) = \sum_{i=1}^{p} \frac{\lambda_i^2 \sigma^2 + k^2 \gamma_i^2}{(\lambda_i^2 + k)^2} \ . \qquad (11)$$

Theorem. There is a $k > 0$ for which the MSE of the RRE is smaller than that of the BLUE,

$$\text{MSE}(\widehat{\boldsymbol{\beta}}(k)) < \text{MSE}(\widehat{\boldsymbol{\beta}}(0)) \ .$$

Proof. Let $f(k) = \text{RHS}(11)$. We have to show that f is decreasing at zero, i.e. $f'(0) < 0$. This follows since

$$f'(k) = 2 \sum_{i=1}^{p} \frac{\lambda_i^2 (k \gamma_i^2 - \sigma^2)}{(\lambda_i^2 + k)^3} \ . \qquad \square$$

An **optimal RRE** $\widehat{\boldsymbol{\beta}}(k^*)$ may be defined as corresponding to a value k^* where $f(k)$ is minimum.