Math tools

Note: Statistical Inference

Oct 2024

Lecturer:

Typed by: Zhuohua Shen

Contents

Chapter 1

Preliminary

1.1 Distributions

1.1.1 χ^2 distribution

Let $X \sim \chi_k^2$, then if $m > -k/2$, the moment $\mathbb{E}(X^m)$ exists and is equal to:

$$\mathbb{E}(X^m) = 2^m \frac{\Gamma\left(\frac{k}{2} + m\right)}{\Gamma\left(\frac{k}{2}\right)}. \tag{1.1}$$

See [link] for the proof.

Chapter 2

Statistical inference fundamentals

References: most of the contents are from the undergraduate course STA3020 (by Prof. Jianfeng Mao in 2022-2023 T1, and Prof. Jiasheng Shi in 2023-2024 T2) and postgraduate course STAT5010 (by Kin Wai Keith Chan in 2024-2025 T1), with main textbook Casella and Berger [1]

2.1 Statistical Models

See Chapter 3 of [1]. Suppose $X_i \sim_{\text{iid}} \mathbb{P}_*$, where $\mathbb{P}_*$ refers to the unknown data generating process (DGPg), we find $\widehat{\mathbb{P}} \approx \mathbb{P}_*$. A statistical model is a set of distributions $\mathscr{F} = \{\mathbb{P}_\theta : \theta \in \Theta\}$, where Θ is the parameter space. A parametric model is the model with $\dim(\Theta) < \infty$, while a nonparametric model satisfies $\dim(\Theta) = \infty$.

Exponential family

Definition 2.1.1. A k-dimensional exponential family (EF) $\mathscr{F} = \{f_\theta : \theta \in \Theta\}$ is a model consisting of pdfs of the form

$$f_\theta(x) = c(\theta)h(x)\exp\left\{\sum_{j=1}^{k} \eta_j(\theta)T_j(x)\right\} \tag{2.1}$$

where $c(\theta), h(x) \geq 0$, $\Theta = \{\theta : c(\theta) \geq 0, \eta_j(\theta)$ being well defined for $1 \leq j \leq k\}$. Let $\eta_j = \eta_j(\theta)$, the canonical form is

$$f_\eta(x) = b(\eta)h(x)\exp\left\{\sum_{j=1}^{k} \eta_j T_j(x)\right\}, \tag{2.2}$$

- k-dim natural exponential family (NEF): $\mathscr{F}' = \{f_\eta : \eta \in \Xi\}$;
- natural parameter $\eta = (\eta_1, \ldots, \eta_k)^\top$;
- natural parameter space: $\Xi = \{\eta \in \mathbb{R}^k : 0 < b(\eta) < \infty\}$;
- the NEF $\mathscr{F}'$ is of full rank if Ξ contains an open set in $\mathbb{R}^k$;
- the EF is a curved exponential family if $p = \dim(\Theta) < k$.

Properties of EF:

- Let $X \sim f_\eta$, where $\eta \in \Xi$ such that (i) f_η is of the form (2.2) with $B(\eta) = -\log b(\eta)$, and (ii) Ξ contains an open set in $\mathbb{R}^k$. Then, for $j, j' = 1, \ldots, k$, $\mathbb{E}\{T_j(X)\} = \partial B(\eta)/\partial \eta_j$ and $\mathbb{C}\text{ov}\{T_j(X), T_{j'}(X)\} = \partial^2 B(\eta)/(\partial \eta_j \partial \eta_{j'})$.
- Stein's identity:

Location-scale family

Definition 2.1.2. Let f be a density.

- A location-scale family is given by $\mathscr{F} = \{f_{\mu,\sigma} : \mu \in \mathbb{R}, \sigma \in \mathbb{R}^{++}\}$, where $f_{\mu,\sigma}(x) = f\left((x-\mu)/\sigma\right)/\sigma$.
- location parameter: μ; scale parameter: σ; standard density: f;
- A location family is $\mathscr{F} = \{f_{\mu,1} : \mu \in \mathbb{R}\}$.
- A scale family is $\mathscr{F} = \{f_{0,\sigma} : \sigma \in \mathbb{R}^{++}\}$

Representation: $X = \mu + \sigma Z$, $Z \sim f_{0,1}(\cdot)$.

- See some examples in Example 3.9, Keith's note 3, and Table 1 in Shi's note L1.
- Transform between location parameter and scale parameter by taking log.

Identifiable family

Definition 2.1.3. If $\forall \theta_1, \theta_2 \in \Theta$ that

$$\theta_1 \neq \theta_2 \quad \Rightarrow \quad f_{\theta_1}(\cdot) \neq f_{\theta_2}(\cdot),$$

then $\mathscr{F}$ is said to be an identifiable family, or equivalently $\theta \in \Theta$ is identifiable.

A typical feature of non-identifiable EF is that $p = \dim(\Theta) > k$. Typically,

- $p < k$, curved (must).
- $p = k$, of full rank.
- $p > k$, non-identifiable.

2.2 Principles of Data Reduction

Statistics: $T = T(X_{1:n})$, a function of $X_{1:n}$ and free of any unknown parameter.

2.2.1 Sufficiency Principle

Sufficiency principle: If $T = T(X_{1:n})$ is a "sufficient statistics" for θ, then any inference on θ will depend on $X_{1:n}$ only through T.

Sufficient, minimal sufficient, ancillary, and complete statistics

Definition 2.2.1. Suppose $X_{1:n} \sim_{\text{iid}} \mathbb{P}_\theta$, where $\theta \in \Theta$. Let $T = T(X_{1:n})$ be a statistic. Then T is sufficient (SS) for θ

- $\Leftrightarrow$ (def) $[X_{1:n} \mid T = t]$ is free of θ for each t.
- $\Leftrightarrow$ (technical lemma) $T(x_{1:n}) = T(x'_{1:n})$ implies that $f_\theta(x_{1:n})/f_\theta(x'_{1:n})$ is free of θ.
- $\Leftrightarrow$ (Neyman-Fisher factorization theorem) $\forall \theta \in \Theta$, $x_{1:n} \in \mathscr{X}^n$, $f_\theta(x_{1:n}) = A(t, \theta) B(x_{1:n})$.
- $\Leftrightarrow$ Define $\Lambda(\theta', \theta'' \mid x_{1:n}) := f_{\theta'}(x_{1:n})/f_{\theta''}(x_{1:n})$. $\forall \theta', \theta'' \in \Theta$, $\exists$ function $C_{\theta', \theta''}$ such that $\Lambda(\theta', \theta'' \mid x_{1:n}) = C_{\theta', \theta''}(t)$, for all $x_{1:n} \in \mathscr{X}^n$ where $t = T(x_{1:n})$.

T is minimal sufficient (MSS) for θ

- $\Leftrightarrow$ (def) (1) T is a SS for θ; (2) $T = g(S)$ for any other SS S.
- $\Leftrightarrow$ (1) T is a SS for θ; (2) $S(x_{1:n}) = S(x'_{1:n})$ implies $T(x_{1:n}) = T(x'_{1:n})$ for any SS S.
- $\Leftrightarrow$ (Lehmann-Scheffé theorem) $\forall x_{1:n}, x'_{1:n} \in \mathscr{X}^n$, $f_\theta(x_{1:n})/f_\theta(x'_{1:n})$ is free of $\theta \Leftrightarrow T(x_{1:n}) = T(x'_{1:n})$.

$A = A(X_{1:n})$ is ancillary (ANS) if the distribution of A does not depend on θ.
T is complete (CS) if $\forall \theta \in \Theta$, $\mathbb{E}_\theta g(T) = 0$ implies $\forall \theta \in \Theta$, $\mathbb{P}_\theta\{g(T) = 0\} = 1$.

Properties

- (Transformation) If $T = r(T')$, then (i) T is SS $\Rightarrow$ T' is SS; (ii) T' is CS $\Rightarrow$ T is CS; (iii) r is one-to-one, then if one is SS/MSS/CS, then the another is.
- (Basu's Lemma) $X_i \sim_{\text{iid}} \mathbb{P}_\theta$, A is ANS and T s CSS, then $A \perp\!\!\!\perp T$.
- (Bahadur's theorem) $X_i \sim_{\text{iid}} \mathbb{P}_\theta$, if an MSS exists, then any CSS is also an MSS.
  - Then if a CSS exists, then any MSS is also a CSS $\Rightarrow$ CSS=MSS.
  - All or nothing: start with MSS T, check whether T is CS. (i) Yes, it is both CSS and MSS, then the set of MSS=CSS; (ii) No, there is no CSS at all.
- (Exp-family) If $X_i \sim_{\text{iid}} f_\eta$ in (2.2), then $T = (\sum_{i=1}^n T_1(X_i), \ldots, \sum_{i=1}^n T_k(X_i))$ is a SS, called natural sufficient statistic. If Ξ contains an open set in $\mathbb{R}^k$ (i.e., $\mathscr{F}'$ is of full rank), then T is MSS and CSS.

Proof techniques

- Prove T is not sufficient for θ: show if $\exists x_{1_n}, x'_{1:n} \in \mathcal{X}^n$ and $\theta', \theta'' \in \Theta$, such that $T(x_{1:n}) = T(x'_{1:n})$ and $\Lambda(\theta', \theta'' \mid x_{1:n}) \neq \Lambda(\theta', \theta'' \mid x'_{1:n})$.
- Prove A is an ANS: consider location-scale representation.
- Prove T is a CS: use definition or take $\mathrm{d}\mathbb{E}_\theta g(T)/\mathrm{d}\theta = 0$.
- Disprove T is CS:
  - Construct an ANS $S(T)$ based on T, then $\mathbb{E}S(T)$ is free of θ, then $g(T) = S(T) - \mathbb{E}S(T)$ is free of θ but $g(T) \neq 0$ w.p.1.
  - (Cancel the 1st moment) Find two unbiased estiamtors for θ as a function of T. E.g., $X_1, X_2 \sim_{\text{iid}} \mathrm{N}(\theta, \theta^2)$, $T = (X_1, X_2)$, $g(T) = X_1 - X_2 \sim \mathrm{N}(0, 2\theta^2)$.

Remark 2.2.2. • ANS A is useless on its own, but useful together with other information.

- $\mathbb{P}(A(\boldsymbol{X}) \mid \theta)$ is free of θ, but for non-SS T, $\mathbb{P}(A(\boldsymbol{X}) \mid T(\boldsymbol{X}))$ is not necessarily free of θ.

2.2.2 Likelihood principle

2.3 M-estimation

Estimating equation

Definition 2.3.1. Let the unknown $\boldsymbol{b} \in \mathbb{R}^p$. Define the *estimating equation*

$$\bar{\boldsymbol{m}}(W, \boldsymbol{b}) = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{m}(w_i, \boldsymbol{b}) = \mathbf{0}_p, \tag{2.3}$$

where $\boldsymbol{m}(\cdot,\cdot) \in \mathbb{R}^p$, and $W = \{w_i\}_{i=1}^n$ are the observed data. Denote $\hat{\boldsymbol{\beta}}$ and $\boldsymbol{\beta}$ the solution of $\bar{\boldsymbol{m}}(W, \boldsymbol{b}) = \mathbf{0}_p$ and $\mathbb{E}\{\bar{\boldsymbol{m}}(W, \boldsymbol{b})\} = \mathbf{0}_p$, respectively:

$$\bar{\boldsymbol{m}}(W, \hat{\boldsymbol{\beta}}) = \mathbf{0}_p,$$
$$\mathbb{E}\{\bar{\boldsymbol{m}}(W, \boldsymbol{\beta})\} = \mathbf{0}_p.$$

Under mild regularity conditions, $\hat{\boldsymbol{\beta}} \xrightarrow{\mathrm{p}} \boldsymbol{\beta}$ and $\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{\mathrm{d}} \mathrm{N}_p(\mathbf{0}_p, \Sigma)$.

2.3.1 Asymptotic properties

Asymptotic properties of M-estimator under with iid data

Theorem 2.3.2. *Assume that $W = \{w_i\}_{i=1}^n$ are iid with the same distribution as w. The true parameter $\boldsymbol{\beta} \in \mathbb{R}^p$ is the unique solution of*

$$\mathbb{E}\{\boldsymbol{m}(w, \boldsymbol{b})\} = \mathbf{0}_p,$$

and the estimator $\hat{\boldsymbol{\beta}} \in \mathbb{R}^p$ is the solution of

$$\bar{\boldsymbol{m}}(W, \boldsymbol{b}) = \mathbf{0}_p.$$

Under some regularity conditions,

$$\sqrt{n}\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right) \xrightarrow{\mathrm{d}} \mathrm{N}_p(\mathbf{0}_p, B^{-1}MB^{-\mathrm{T}}),$$

where

$$B = -\frac{\partial \mathbb{E}\{\boldsymbol{m}(w, \boldsymbol{\beta})\}}{\partial \boldsymbol{b}^\top}, \quad M = \mathbb{E}\{\boldsymbol{m}(w, \boldsymbol{\beta})\boldsymbol{m}(w, \boldsymbol{\beta})^\top\}.$$

Asymptotic properties of M-estimator under with independent data

Theorem 2.3.3. *Assume that $W = \{w_i\}_{i=1}^n$ are independent observations. The true parameter $\boldsymbol{\beta} \in \mathbb{R}^p$ is the unique solution of*

$$\mathbb{E}\{\bar{\boldsymbol{m}}(W, \boldsymbol{b})\} = \mathbf{0}_p,$$

and the estimator $\hat{\boldsymbol{\beta}} \in \mathbb{R}^p$ is the solution of

$$\bar{\boldsymbol{m}}(W, \boldsymbol{b}) = \mathbf{0}_p.$$

Under some regularity conditions,

$$\sqrt{n}\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right) \xrightarrow{\mathrm{d}} \mathrm{N}_p(\mathbf{0}_p, B^{-1}MB^{-\mathrm{T}}),$$

where

$$B = -\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} \frac{\partial \mathbb{E}\{\boldsymbol{m}(w_i, \boldsymbol{\beta})\}}{\partial \boldsymbol{b}^\top}, \quad M = \lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} \mathbb{C}\mathrm{ov}\{\boldsymbol{m}(w_i, \boldsymbol{\beta})\}.$$

MLE

Example 2.3.4. *Suppose $y_1, \ldots, y_n \sim_{\mathrm{iid}} f(y \mid \boldsymbol{\theta})$, $\boldsymbol{\theta} \in \mathbb{R}^p$. The MLE satisfies the following estimating equation*

$$\mathbb{E}\left\{\frac{\partial \log f(y \mid \boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right\} = \mathbf{0}_p,$$

which is Bartlett's first identity. Under regularity conditions, $\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \xrightarrow{\mathrm{d}} \mathrm{N}_p(\mathbf{0}_p, B^{-1}MB^{-1})$, where

$$B = I(\boldsymbol{\theta}) = -\frac{\partial}{\partial \boldsymbol{\theta}^\top}\mathbb{E}\left(\frac{\partial \log f(y \mid \boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right) = \mathbb{E}\left\{-\frac{\partial^2 \log f(y \mid \boldsymbol{\theta})}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}^\top}\right\},$$
$$M = J(\boldsymbol{\theta}) = \mathbb{E}\left\{\frac{\partial \log f(y \mid \boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\frac{\partial \log f(y \mid \boldsymbol{\theta})}{\partial \boldsymbol{\theta}^\top}\right\}.$$

If the model is correct, Bartlett's second identity ensures that $I(\boldsymbol{\theta}) = J(\boldsymbol{\theta})$, and therefore $\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \xrightarrow{\mathrm{d}}$ $\mathrm{N}_p(\mathbf{0}_p, I(\boldsymbol{\theta})^{-1})$

Chapter 3

Multivariate Inference Fundamentals

Reference:

- Robb J. Muirhead - Aspects of multivariate statistical theory [5].
- CUHK STAT4002 - Applied Multivariate Analysis (2023 Spring), by Zhixiang Lin.
- CUHK STAT5030 - Linear Models (2025 Spring), by Yuanyuan Lin.
- Peng Ding - Linear Model and Extensions.
- Ronald Christensen - Plane answers to complex questions: the theory of linear models [2].

3.1 Random vectors and distributions

Definition 3.1.1. Let $\boldsymbol{x} = (x_1, \ldots, x_p)^\top \in \mathbb{R}^p$ be a random vector,

- Mean $\mathbb{E}\boldsymbol{x} = \boldsymbol{\mu} = (\mathbb{E}x_1, \ldots, \mathbb{E}x_p)^\top = (\mu_j)$.
- Covariance matrix $\mathbb{V}\mathrm{ar}(\boldsymbol{x}) = \mathbb{C}\mathrm{ov}(\boldsymbol{x}) = \Sigma = \mathbb{E}[(\boldsymbol{x} - \mathbb{E}\boldsymbol{x})(\boldsymbol{x} - \mathbb{E}\boldsymbol{x})^\top] = \mathbb{E}\boldsymbol{x}\boldsymbol{x}^\top - \mathbb{E}\boldsymbol{x}\mathbb{E}\boldsymbol{x}^\top = (\sigma_{ij})$, $\Sigma \succeq \mathbf{0}$.
- Correlation matrix $R = D^{-1/2}\Sigma D^{-1/2}$, where $D = \mathrm{diag}(\sigma_{11}, \ldots, \sigma_{pp})$. We have $R_{ij} = \rho_{ij} = \sigma_{ij}/(\sqrt{\sigma_{ii}}\sqrt{\sigma_{jj}})$.
- If $\boldsymbol{y} \in \mathbb{R}^q$ random vector, then $\mathbb{C}\mathrm{ov}(\boldsymbol{x}, \boldsymbol{y}) = \mathbb{E}[(\boldsymbol{x} - \mathbb{E}\boldsymbol{x})(\boldsymbol{y} - \mathbb{E}\boldsymbol{y})^\top] = \mathbb{E}\boldsymbol{x}\boldsymbol{y}^\top - \mathbb{E}\boldsymbol{x}\mathbb{E}\boldsymbol{y}^\top \in \mathbb{R}^{p\times q}$.
- MSE $\mathrm{MSE}(\hat{\boldsymbol{\beta}}; \boldsymbol{\beta}) = \mathbb{E}\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|^2 = \|\mathbb{E}\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|^2 + \mathrm{tr}[\mathbb{V}\mathrm{ar}(\hat{\boldsymbol{\beta}})]$.

If $\boldsymbol{Z} = (z_{ij}) \in \mathbb{R}^{p\times q}$ is a random matrix,

- $\mathbb{E}\boldsymbol{Z} = (\mathbb{E}z_{ij})$.

Proposition 3.1.2. *Let $\boldsymbol{x} \in \mathbb{R}^p$ be a random vector, $\boldsymbol{a}, \boldsymbol{b} \in \mathbb{R}^p$ be vectors, $A \in \mathbb{R}^{r_1\times p}, B \in \mathbb{R}^{r_2\times p}$ be matrices,*

- $\mathbb{E}\boldsymbol{a}^\top\boldsymbol{x} = \boldsymbol{a}^\top\mathbb{E}\boldsymbol{x}$, $\mathbb{V}\mathrm{ar}(\boldsymbol{a}^\top\boldsymbol{x}) = \boldsymbol{a}^\top\Sigma\boldsymbol{a}$, *and* $\mathbb{C}\mathrm{ov}(\boldsymbol{a}^\top\boldsymbol{x}, \boldsymbol{b}^\top\boldsymbol{x}) = \boldsymbol{a}^\top\Sigma\boldsymbol{b}$.
- $\mathbb{E}A\boldsymbol{x} = A\mathbb{E}\boldsymbol{x}$, $\mathbb{V}\mathrm{ar}(A\boldsymbol{x}) = A\Sigma A^\top$, *and* $\mathbb{C}\mathrm{ov}(A\boldsymbol{x}, B\boldsymbol{x}) = A\Sigma B^\top$.
- *If* $\boldsymbol{y} = A\boldsymbol{x} + \boldsymbol{b}$, *where* $A \in \mathbb{R}^{q\times p}$, $\boldsymbol{b} \in \mathbb{R}^q$, *then* $\boldsymbol{\mu}_{\boldsymbol{y}} = A\boldsymbol{\mu}_{\boldsymbol{x}} + \boldsymbol{b}$ *and* $\Sigma_{\boldsymbol{y}} = A\Sigma_{\boldsymbol{x}}A^\top$.
- $\mathbb{E}(\boldsymbol{x}^\top A\boldsymbol{x}) = \mathrm{tr}(A\Sigma) + \boldsymbol{\mu}^\top A\boldsymbol{\mu}$.

Let $\boldsymbol{Z} \in \mathbb{R}^{p\times q}$ be a random matrix, $B \in \mathbb{R}^{m\times p}$, $C \in \mathbb{R}^{q\times n}$, and $D \in \mathbb{R}^{m\times n}$ constants, then

- $\mathbb{E}(B\boldsymbol{Z}C + D) = B\mathbb{E}(\boldsymbol{Z})C + D$.

- The $\Sigma \in \mathbb{R}^{p\times p}$ is a covariance matrix (i.e., $\Sigma = \mathbb{C}\mathrm{ov}(\boldsymbol{x})$ for some random vector $\boldsymbol{x} \in \mathbb{R}^p$) iff $\Sigma \succeq \mathbf{0}$.
  - ($\Leftarrow$): suppose $\mathrm{r}(\Sigma) = r \le p$, write full rank decomposition $\Sigma = CC^\top$, $C \in \mathbb{R}^{p\times r}$. Let $\boldsymbol{y} \sim [\mathbf{0}_r, I_r]$, then $\mathbb{C}\mathrm{ov}(C\boldsymbol{y}) = \Sigma$.
- If Σ is not PD, then $\exists \boldsymbol{a} \ne \mathbf{0}_p$ s.t. $\mathbb{V}\mathrm{ar}(\boldsymbol{a}^\top\boldsymbol{x}) = 0$ so w.p.1., $\boldsymbol{a}^\top\boldsymbol{x} = k$, i.e., $\boldsymbol{x}$ lies in a hyperplane.

Theorem 3.1.3. *If $\boldsymbol{x} \in \mathbb{R}^p$ random, then its distribution is uniquely determined by the distributions of $\boldsymbol{a}^\top\boldsymbol{x}$, $\forall\boldsymbol{a} \in \mathbb{R}^p$.*

The proof uses the fact that a distribution in $\mathbb{R}^p$ is uniquely determined by its ch.f., see Theorem 1.2.2. [5].

Definition 3.1.4. Dataset contains p variables and n observations are represented by $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)^\top$, where the ith row $\boldsymbol{x}_i^\top = (x_{i1}, \ldots, x_{ip})$ is the ith observation vector, $i = 1, \ldots, n$.

- (Sample mean vector) $\bar{\boldsymbol{x}} = n^{-1}\sum_{i=1}^n \boldsymbol{x}_i = (\bar{x}_1, \ldots, \bar{x}_p)^\top$, where $\bar{x}_j = n^{-1}\sum_{i=1}^n x_{ij}$.
- (Sum of squares and cross product (SSCP) matrix) $\boldsymbol{A} = \sum_{i=1}^n (\boldsymbol{x}_i - \bar{\boldsymbol{x}})(\boldsymbol{x}_i - \bar{\boldsymbol{x}})^\top$.
- (Sample covariance matrix) $\boldsymbol{S} = (n-1)^{-1}\boldsymbol{A}$.
- (Sample correlation matrix) $\boldsymbol{R} = D^{-1/2}\boldsymbol{S}D^{-1/2}$, where $D^{-1/2} = \mathrm{diag}(1/\sqrt{s_{11}}, \ldots, 1/\sqrt{s_{pp}})$.

- $\bar{\boldsymbol{x}} = n^{-1}\boldsymbol{X}^\top \mathbf{1}_n$, and

$$\begin{aligned} \boldsymbol{A} &= \sum_{i=1}^{n} (\boldsymbol{x}_i - \boldsymbol{\mu})(\boldsymbol{x}_i - \boldsymbol{\mu})^\top - n(\bar{\boldsymbol{x}} - \boldsymbol{\mu})(\bar{\boldsymbol{x}} - \boldsymbol{\mu})^\top \\ &= (\boldsymbol{X} - \mathbf{1}_n \bar{\boldsymbol{x}}^\top)^\top (\boldsymbol{X} - \mathbf{1}_n \bar{\boldsymbol{x}}^\top) \succeq \mathbf{0}. \end{aligned}$$

- $\mathbb{E}\bar{\boldsymbol{x}} = \boldsymbol{\mu}$, $\mathbb{V}\mathrm{ar}(\bar{\boldsymbol{x}}) = n^{-1}\Sigma$, $\mathbb{E}\boldsymbol{A} = (n-1)\Sigma$, and $\mathbb{E}\boldsymbol{S} = \Sigma$.

3.1.1 Multivariate normal distribution

Original definition of multivariate normal

Definition 3.1.5. The random vector $\boldsymbol{x} \in \mathbb{R}^p$ is said to have an p-variate normal distribution ($\boldsymbol{x} \sim \mathrm{N}_p$) if $\forall \boldsymbol{a} \in \mathbb{R}^p$, the distribution of $\boldsymbol{a}^\top \boldsymbol{x}$ is univariate normal.

Fundamental properties

Theorem 3.1.6. *Let $\boldsymbol{x} \sim \mathrm{N}_p$, we have*

1. *Both $\boldsymbol{\mu} = \mathbb{E}\boldsymbol{x}$ and $\Sigma = \mathbb{C}\mathrm{ov}(\boldsymbol{x})$ exist and the distribution of $\boldsymbol{x}$ is determined by $\boldsymbol{\mu}$ and Σ. Write $\boldsymbol{x} \sim \mathrm{N}_p(\boldsymbol{\mu}, \Sigma)$.*
2. *(Representation) Let $\Sigma \succeq \mathbf{0}_{p\times p}$, $\mathrm{r}(\Sigma) = r \le p$, and $u_{1:r} \sim_{\mathrm{iid}} \mathrm{N}(0,1)$, i.e., $\boldsymbol{u} \sim \mathrm{N}_r(\mathbf{0}_r, I_r)$, then if C is the full rank decomposition of Σ and $\boldsymbol{\mu} \in \mathbb{R}^p$, then $\boldsymbol{x} = C\boldsymbol{u} + \boldsymbol{\mu} \sim \mathrm{N}_p(\boldsymbol{\mu}, \Sigma)$.*
   - *Let $\Sigma = HDH^\top$ be the spectral decomposition, then $\boldsymbol{x} = HD^{1/2}\boldsymbol{z} + \boldsymbol{\mu}$, where $\boldsymbol{z} \sim \mathrm{N}_p(\mathbf{0}_p, I_p)$.*
3. *If $\boldsymbol{x} \sim \mathrm{N}_p(\boldsymbol{\mu}, \Sigma)$, then its ch.f. $\phi_{\boldsymbol{x}}(\boldsymbol{t}) = \exp(i\boldsymbol{\mu}^\top \boldsymbol{t} - \boldsymbol{t}^\top \Sigma \boldsymbol{t}/2)$.*
4. *(Density) $\boldsymbol{x} \sim \mathrm{N}_p(\boldsymbol{\mu}, \Sigma)$ with $\Sigma \succ \mathbf{0}$, then $\boldsymbol{x}$ has pdf*

$$f(\boldsymbol{x}) = \frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}} \exp\left\{-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu})^T \Sigma^{-1} (\boldsymbol{x} - \boldsymbol{\mu})\right\}. \tag{3.1}$$

Note that we guarantee the existence of $\mathrm{N}_p(\boldsymbol{\mu}, \Sigma)$ by means of the representation in point 2. By its density, we have MVN kernel: If

$$f(\boldsymbol{x}) \propto \exp\left\{-\frac{1}{2}(\boldsymbol{x}^\top A\boldsymbol{x} - 2\boldsymbol{x}^\top B)\right\} = \exp\left\{-\frac{1}{2}(\boldsymbol{x} - A^{-1}B)^\top A(\boldsymbol{x} - A^{-1}B) - B^\top A^{-1}B\right\},$$

then $\boldsymbol{x} \sim \mathrm{N}_p(A^{-1}B, A^{-1})$.

Properties of multivariate normal

Theorem 3.1.7. *If $\boldsymbol{x} \sim \mathrm{N}_p(\boldsymbol{\mu}, \Sigma)$, then we have*

1. *(Linearity) Let $B \in \mathbb{R}^{q\times p}, \boldsymbol{b} \in \mathbb{R}^q$ nonrandom, and $B\Sigma B^\top \succ \mathbf{0}$, then $B\boldsymbol{x} + \boldsymbol{b} \sim \mathrm{N}_q(B\boldsymbol{\mu} + \boldsymbol{b}, B\Sigma B^\top)$.*
2. *(Linear combinations) If $\boldsymbol{x}_k \sim \mathrm{N}_p(\boldsymbol{\mu}_k, \Sigma_k) \perp\!\!\!\perp$ for $k = 1, \ldots, N$, then for any fixed constants $\alpha_1, \ldots, \alpha_N$, $\sum_{k=1}^N \alpha_k \boldsymbol{x}_k \sim \mathrm{N}_p(\sum_{k=1}^N \alpha_k \boldsymbol{\mu}_k, \sum_{k=1}^N \alpha_k^2 \Sigma_k)$.*
   - *The sample mean $\bar{\boldsymbol{x}} \sim \mathrm{N}_p(\boldsymbol{\mu}, \Sigma/N)$.*
3. *(Subset) The marginal distribution of any subset of $k(< p)$ components of $\boldsymbol{x}$ is k-variate normal.*
4. *(Marginal distribution) Partition*

$$\boldsymbol{x} = \begin{bmatrix} \boldsymbol{x}_1 \\ \boldsymbol{x}_2 \end{bmatrix}, \quad \boldsymbol{\mu} = \begin{bmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{bmatrix}, \quad \boldsymbol{\Sigma} = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}, \quad \boldsymbol{x}_1 \in \mathbb{R}^q, \boldsymbol{x}_2 \in \mathbb{R}^{p-q}, \Sigma_{12} \in \mathbb{R}^{q\times(p-q)}.$$

   Then $\boldsymbol{x}_1 \sim \mathrm{N}_q(\boldsymbol{\mu}_1, \Sigma_{11})$, $\boldsymbol{x}_1 \perp\!\!\!\perp \boldsymbol{x}_2$ iff $\Sigma_{12} = \mathbf{0}$.
5. *(Conditional distribution) Let Σ_{22}^- be a generalized inverse of Σ_{22} (i.e., $\Sigma_{22}\Sigma_{22}^-\Sigma_{22} = \Sigma_{22}$), then*
   (a) $\boldsymbol{x}_1 - \Sigma_{12}\Sigma_{22}^-\boldsymbol{x}_2 \sim \mathrm{N}_q(\boldsymbol{\mu}_1 - \Sigma_{12}\Sigma_{22}^-\boldsymbol{\mu}_2, \Sigma_{11} - \Sigma_{12}\Sigma_{22}^-\Sigma_{21})$, and $\perp\!\!\!\perp \boldsymbol{x}_2$.
   (b) $[\boldsymbol{x}_1 \mid \boldsymbol{x}_2] \sim \mathrm{N}_q(\boldsymbol{\mu}_1 + \Sigma_{12}\Sigma_{22}^-(\boldsymbol{x}_2 - \boldsymbol{\mu}_2), \Sigma_{11} - \Sigma_{12}\Sigma_{22}^-\Sigma_{21})$.
6. *(Cramér) If $p \times 1$ random vectors $\boldsymbol{x} \perp\!\!\!\perp \boldsymbol{y}$ and $\boldsymbol{x} + \boldsymbol{y} \sim \mathrm{N}_p$, then both $\boldsymbol{x}, \boldsymbol{y} \sim \mathrm{N}_p$.*
7. *(MLE) of (μ, Σ) is $(\bar{\boldsymbol{x}}, A/n)$.*
8. *(Inverse of Σ and conditional independence) Denote $\Sigma^{-1} = (v^{jk})_{1\le j,k\le p}$. Then $\forall j \neq k$, $v^{jk} = 0 \Leftrightarrow x_j \perp\!\!\!\perp x_k \mid \boldsymbol{x} \setminus \{x_j, x_k\}$.*

For point 3, each component of a random vector is (marginally) normal does not imply that the vector has a multivariate normal distribution. Counterexample: let $U_1, U_2, U_3 \sim_{\mathrm{iid}} \mathrm{N}(0,1)$, $Z \perp\!\!\!\perp U_{1:3}$. Define

$$X_1 = \frac{U_1 + ZU_3}{\sqrt{1+Z^2}}, \quad X_2 = \frac{U_2 + ZU_3}{\sqrt{1+Z^2}}.$$

Then $[X_1|Z] \sim \mathrm{N}(0,1)$, free of Z, so $X_1 \sim \mathrm{N}(0,1)$, and $X_2 \sim \mathrm{N}(0,1)$. But (X_1, X_2) not normal. The converse is true if the components of $\boldsymbol{x}$ are all independent and normal, or if $\boldsymbol{x}$ consists of independent subvectors, each of which is normally distributed.

For the proof of point 5, we use the lemma: if $\Sigma \succeq \mathbf{0}$, then $\ker(\Sigma_{22}) \subset \ker(\Sigma_{12})$, and $\mathrm{range}(\Sigma_{21}) \subset \mathrm{range}(\Sigma_{22})$. So $\exists B \in \mathbb{R}^{q\times(p-q)}$ satisfying $\Sigma_{12} = B\Sigma_{22}$.

Quadratic form of MVN

MGF

Proposition 3.1.8. *If $\boldsymbol{x} \sim \mathrm{N}(\boldsymbol{\mu}, \Sigma)$, A is symmetric and σ IS non-singular, then*

$$M_{\boldsymbol{x}^\top \boldsymbol{A}\boldsymbol{x}}(t) = |I - 2tA\Sigma|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}\boldsymbol{\mu}^\top [I - (I - 2t\boldsymbol{A}\Sigma)^{-1}]\Sigma^{-1}\boldsymbol{\mu}\right\}.$$

Variance

Proposition 3.1.9. *If $\boldsymbol{x} \sim \mathrm{N}(\boldsymbol{\mu}, \Sigma)$, then*

$$\mathbb{V}\mathrm{ar}(\boldsymbol{x}^\top A\boldsymbol{x}) = 2\mathrm{tr}(A\Sigma A\Sigma) + 4\boldsymbol{\mu}^\top A\Sigma A\boldsymbol{\mu},$$
$$\mathbb{C}\mathrm{ov}(\boldsymbol{x}^\top A_1\boldsymbol{x}, \boldsymbol{x}^\top A_2\boldsymbol{x}) = 2\mathrm{tr}(A_1\Sigma A_2\Sigma) + 4\boldsymbol{\mu}^\top A_1\Sigma A_2\boldsymbol{\mu}.$$

See Problem B 2.14 in Peng DING.

Independence of quadratic form

Theorem 3.1.10. *Assume $\boldsymbol{x} \sim N_p(\mu, \Sigma)$.*

1. *For two symmetric matrix $A, B \in \mathbb{R}^{p\times p}$, $\boldsymbol{x}^\top A\boldsymbol{x} \perp\!\!\!\perp \boldsymbol{x}^\top B\boldsymbol{x}$ iff*

$$\Sigma A\Sigma B\Sigma = \mathbf{0}, \quad \Sigma A\Sigma B\boldsymbol{\mu} = \Sigma B\Sigma A\boldsymbol{\mu} = \mathbf{0}, \quad \boldsymbol{\mu}^\top A\Sigma B\boldsymbol{\mu} = 0.$$

   If $\Sigma \succ \mathbf{0}$, then iff $A\Sigma B = \mathbf{0}$.

2. *If $\Sigma \succ \mathbf{0}$, $A \in \mathbb{R}^{p\times p}$ symmetric, and $B \in \mathbb{R}^{r\times p}$, then $\boldsymbol{x}^\top A\boldsymbol{x} \perp\!\!\!\perp B\boldsymbol{x}$ iff $B\Sigma A = \mathbf{0}$.*

See Theorem B.11 in Peng DING, and Problem 1.22–1.23 in [5].

Quadratic form of Σ

Theorem 3.1.11. *If $\boldsymbol{x}, \boldsymbol{x}_{1:N} \overset{}{\sim}_{\mathrm{iid}} \mathrm{N}_p(\boldsymbol{\mu}, \Sigma)$, where Σ is nonsingular, then*

- $(\boldsymbol{x} - \boldsymbol{\mu})^\top \Sigma^{-1}(\boldsymbol{x} - \boldsymbol{\mu}) \sim \chi^2_p$,
- $\boldsymbol{x}^\top \Sigma^{-1}\boldsymbol{x} \sim \chi^2_p(\boldsymbol{\mu}^\top\Sigma^{-1}\boldsymbol{\mu})$,
- *partition*

$$\boldsymbol{x} = \begin{bmatrix}\boldsymbol{x}_1\\ \boldsymbol{x}_2\end{bmatrix}, \ \boldsymbol{\mu} = \begin{bmatrix}\boldsymbol{\mu}_1\\ \boldsymbol{\mu}_2\end{bmatrix}, \ \Sigma = \begin{bmatrix}\Sigma_{11} & \Sigma_{12}\\ \Sigma_{21} & \Sigma_{22}\end{bmatrix}, \quad \boldsymbol{x}_1, \boldsymbol{\mu}_1 \in \mathbb{R}^k, \ \Sigma_{11} \in \mathbb{R}^{k\times k}, \ \textit{then}$$

$$Q = (\boldsymbol{x} - \boldsymbol{\mu})^\top\Sigma^{-1}(\boldsymbol{x} - \boldsymbol{\mu}) - (\boldsymbol{x}_1 - \boldsymbol{\mu}_1)^\top\Sigma_{11}^{-1}(\boldsymbol{x}_1 - \boldsymbol{\mu}_1) \sim \chi^2_{p-k}.$$

- $N(\bar{\boldsymbol{x}}_N - \boldsymbol{\mu})^\top\Sigma^{-1}(\bar{\boldsymbol{x}}_N - \boldsymbol{\mu}) \sim \chi^2_p$,
- *the Mahalanobis distance* $d_i^2 = (\boldsymbol{x}_i - \bar{\boldsymbol{x}}_N)^\top \boldsymbol{S}^{-1}(\boldsymbol{x}_i - \bar{\boldsymbol{x}}_N) \xrightarrow{\mathrm{d}} \chi^2_p$.

If $\mathrm{r}(\Sigma) = k \le p$, then

1.

$$(\boldsymbol{x} - \boldsymbol{\mu})^\top\Sigma^{-}(\boldsymbol{x} - \boldsymbol{\mu}) \sim \chi^2_k,$$
$$(\boldsymbol{x} - \boldsymbol{\mu})^\top\Sigma^{+}(\boldsymbol{x} - \boldsymbol{\mu}) \sim \chi^2_k.$$

2. *If Σ is idempotent with $\mathrm{r}(\Sigma) = k$, then $(\boldsymbol{x} - \boldsymbol{\mu})^\top(\boldsymbol{x} - \boldsymbol{\mu}) \sim \chi^2_k$.*
   - *If $\boldsymbol{\mu} \in \mathrm{Col}(\Sigma)$, then $\boldsymbol{x}^\top\boldsymbol{x} \sim \chi^2_k(\boldsymbol{\mu}^\top\boldsymbol{\mu})$.*

Quadratic form of any matrices

Theorem 3.1.12. *If $\boldsymbol{x} \sim \mathrm{N}_p(\boldsymbol{\mu}, \Sigma)$,*

1. *if Σ is nonsingular, $B \in \mathbb{R}^{p\times p}$ is symmetric, then $\boldsymbol{x}^\top B\boldsymbol{x} \sim \chi^2_k(\delta)$ iff $B\Sigma$ is idempotent (equiv., $B\Sigma B = B$), in which case $k = \mathrm{r}(B)$ and $\delta = \boldsymbol{\mu}^\top B\boldsymbol{\mu}$;*
2. *for $A \in \mathbb{R}^{p\times p}$, $\boldsymbol{x}^\top A\boldsymbol{x} \sim \chi^2_{\mathrm{r}(A\Sigma)}(\boldsymbol{\mu}^\top A\boldsymbol{\mu})$ if*

$$(1)\Sigma A\Sigma A\Sigma = \Sigma A\Sigma, \quad (2)\boldsymbol{\mu}^\top A\Sigma A\boldsymbol{\mu} = \boldsymbol{\mu}^\top A\boldsymbol{\mu}, \quad (3)\Sigma A\Sigma A\boldsymbol{\mu} = \Sigma A\boldsymbol{\mu};$$

3. *let $A_i \in \mathbb{R}^{p\times p}$ be symmetric with rank k_i for $i = 1, \ldots, m$. Denote $A = \sum_{i=1}^m A_i$, which is symmetric with rank k. Then $\boldsymbol{x}^\top A_i\boldsymbol{x} \sim \chi^2_{k_i}(\boldsymbol{\mu}^\top A_i\boldsymbol{\mu})$, $\boldsymbol{x}^\top A_i\boldsymbol{x}$ are pairwise independent and $\boldsymbol{x}^\top A\boldsymbol{x} \sim \chi^2_k(\boldsymbol{\mu}^\top A\boldsymbol{\mu})$, iff*
   - *(I) any two of the following are true: (a) $A_i\Sigma$ idempotent, $\forall i$; (b) $A_i\Sigma A_j = 0$, $\forall i < j$; and (c) $A\Sigma$ idempotent;* ***OR***
   - *(II) (c) is true and (d) $k = \sum_{i=1}^m k_i$;* ***OR***
   - *(III) (c) is true and (e) $A_1\Sigma, \ldots, A_{m-1}\Sigma$ are idempotent and $A_m\Sigma \succeq \mathbf{0}$.*
4. *(Cochran's theorem) $\boldsymbol{x} \sim \mathrm{N}_p(\mathbf{0}_p, I_p)$ and A_i is symmetric of rank k_i, for $i = 1, \ldots, m$ with $\sum_{i=1}^m A_i = I_p$,*

then $\boldsymbol{x}^\top A_i \boldsymbol{x} \sim \chi^2_{k_i}$ *independently iff* $\sum_{i=1}^m k_i = p$.

3.1.2 The noncentral χ^2 and F distribution

3.2 Asymptotic properties

3.2.1 Asymptotic distributions of sample means and covariance matrices

Refer to section 1.2.2, [5].

CLT for sample means

Theorem 3.2.1. *Let* $\boldsymbol{x}_{1:n} \sim_{\text{iid}} [\boldsymbol{\mu}, \Sigma]$, *then*

$$\sqrt{n}(\bar{\boldsymbol{x}}_n - \boldsymbol{\mu}) = \frac{1}{\sqrt{n}} \sum_{i=1}^n (\boldsymbol{x}_i - \boldsymbol{\mu}) \xrightarrow{\text{d}} \text{N}_p(\mathbf{0}_p, \Sigma).$$

CLT for sample covariance matrices

Theorem 3.2.2. *Let* $\boldsymbol{x}_{1:n} \sim_{\text{iid}} [\boldsymbol{\mu}, \Sigma]$ *with finite fourth moments, SSCP matrix* $\boldsymbol{A} = \sum_{i=1}^n (\boldsymbol{x}_i - \bar{\boldsymbol{x}})(\boldsymbol{x}_i - \bar{\boldsymbol{x}})^\top$, *and* $\boldsymbol{S} = (n-1)^{-1}\boldsymbol{A}$. *Let* $V = \mathbb{C}\text{ov}[\text{vec}((\boldsymbol{x}_1 - \boldsymbol{\mu})(\boldsymbol{x}_1 - \boldsymbol{\mu})^\top)]$, *then*

$$\frac{1}{\sqrt{n}}(\text{vec}(\boldsymbol{A}) - n \cdot \text{vec}(\Sigma)) \xrightarrow{\text{d}} \text{N}_{p^2}(\mathbf{0}, V),$$

$$\sqrt{n-1}(\text{vec}(\boldsymbol{S}) - \text{vec}(\Sigma)) \xrightarrow{\text{d}} \text{N}_{p^2}(\mathbf{0}, V).$$

Note that $V \in \mathbb{R}^{p^2 \times p^2}$ is singular as the LHS vectors above have repeated elements.

Chapter 4

Linear Models

Reference:

1. CUHK STAT5030 - Linear Models (2025 Spring), by Yuanyuan Lin.
2. CUHKSZ STA3001 - Linear Models (2022 Fall), by Zhou Zhou.
3. Peng Ding - Linear Model and Extensions.

4.1 Linear regression for the full-rank model

Full-rank linear model (fixed design $\boldsymbol{X}$)

Definition 4.1.1.

$$\boldsymbol{Y}_n = \boldsymbol{X}_{n\times p}\boldsymbol{\beta}_p + \boldsymbol{\varepsilon}_n, \quad \mathrm{r}(\boldsymbol{X}) = p, \tag{4.1}$$

where $(\boldsymbol{Y}, X)$ is a pair of response, $\boldsymbol{X}$ is fixed and determistic, $\boldsymbol{\beta}$ is a vector of covariates, and $\boldsymbol{\varepsilon}$ is unobservable error term.

$$\boldsymbol{Y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \ \boldsymbol{X} = \begin{pmatrix} \boldsymbol{x}_1^\top \\ \boldsymbol{x}_2^\top \\ \vdots \\ \boldsymbol{x}_n^\top \end{pmatrix} = \begin{pmatrix} \boldsymbol{X}_1 & \dots & \boldsymbol{X}_p \end{pmatrix}.$$

4.1.1 Ordinary least squares (OLS) estimation

The *least squares estimate (LSE)* minimizes

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\arg\min} \ L(\boldsymbol{\beta}) \equiv \sum_{i=1}^n (y_i - \boldsymbol{x}_i^\top\boldsymbol{\beta})^2 = (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})^\top(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}),$$

which satisfies

$$\text{(FONC)} \qquad \partial L(\boldsymbol{\beta})/\partial\boldsymbol{\beta} = -2\boldsymbol{X}^\top\boldsymbol{Y} + 2X^\top X\boldsymbol{\beta} = \mathbf{0}_p,$$

$$\textit{(Normal equation)} \qquad \boldsymbol{X}^\top\boldsymbol{Y} = \boldsymbol{X}^\top\boldsymbol{X}\hat{\boldsymbol{\beta}}.$$

As $\boldsymbol{X}^\top\boldsymbol{X}$ has full rank,

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{Y} = \left(\sum_{i=1}^n \boldsymbol{x}_i\boldsymbol{x}_i^\top\right)^{-1}\left(\sum_{i=1}^n \boldsymbol{x}_i y_i\right). \tag{4.2}$$

We decompose

$$\boldsymbol{Y} = \hat{\boldsymbol{Y}} + \hat{\varepsilon}, \quad \text{(fitted vector) } \hat{\boldsymbol{Y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}}, \quad \text{(residual) } \hat{\varepsilon} = \boldsymbol{Y} - \hat{\boldsymbol{Y}} = \boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}$$

The geometry of LSE

Proposition 4.1.2. *The OLS problem is to find the vector in* $\mathrm{Col}(\boldsymbol{X})$ *that is the closest to* $\boldsymbol{Y}$. *So* $\hat{\boldsymbol{Y}}$ *is the projection of* $\boldsymbol{Y} \in \mathrm{Col}(\boldsymbol{X})$ *onto* $\mathrm{Col}(X)$, *and the residual* $\hat{\varepsilon} \perp \mathrm{Col}(\boldsymbol{X})$ *algebraically.*

1. $\boldsymbol{X}_j^\top\hat{\varepsilon} = 0 \Longleftrightarrow \boldsymbol{X}^\top\hat{\varepsilon} = \mathbf{0}_p \Longleftrightarrow \boldsymbol{X}^\top(\boldsymbol{Y} - \hat{\boldsymbol{Y}}) = \mathbf{0}_p$.
2. $\hat{\boldsymbol{Y}}^\top\hat{\varepsilon} = 0$, *the Pythagorean Theorem implies that* $\|\boldsymbol{Y}\|^2 = \|\hat{\boldsymbol{Y}}\|^2 + \|\hat{\varepsilon}\|^2$.
3. $\forall \boldsymbol{b} \in \mathbb{R}^p$, $(\boldsymbol{X}\hat{\boldsymbol{\beta}} - \boldsymbol{X}\boldsymbol{b})^\top\hat{\varepsilon} = 0$, *we have the following decomposition*
$$\|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{b}\|^2 = \|\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}\|^2 + \|\boldsymbol{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{b})\|^2,$$
which implies that $\|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{b}\|^2 \geq \|\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}\|^2$ *with equality holding iff* $\boldsymbol{b} = \hat{\boldsymbol{\beta}}$.

Let $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top = (h_{ij})_{1\le i,j\le n}$ be the so-called *hat matrix*.

$$\hat{\boldsymbol{\varepsilon}} = (I_n - \boldsymbol{H})\boldsymbol{Y}, \qquad \hat{\boldsymbol{Y}} = \boldsymbol{H}\boldsymbol{Y}, \qquad \hat{y}_i = h_{ii}y_i + \sum_{j\neq i} h_{ij}y_j.$$

h_{ii} is called the *leverage score*.

Properties of hat matrix $\boldsymbol{H}$

Proposition 4.1.3.

1. *$\boldsymbol{H}$ is symmetric idempotent (projection matrix), so $\boldsymbol{H}\boldsymbol{X} = \boldsymbol{X}$, $(I_n - \boldsymbol{H})\boldsymbol{X} = \boldsymbol{0}_n$, and $\mathrm{r}(\boldsymbol{H}) = \mathrm{r}(\boldsymbol{X})$.*
2. *$I_n - \boldsymbol{H}$ is symmetric idempotent, so $\mathrm{tr}(I_n - \boldsymbol{H}) = n - \mathrm{r}(\boldsymbol{X})$.*
3. *$h_{ii} = h_{ii}^2 + \sum_{j\neq i} h_{ij}^2$, so $0 \le h_{ii} \le 1$.*
4. *$h_{ij}^2 \le 1/4$, $\forall j \neq i$.*
5. *$\hat{\boldsymbol{\varepsilon}}^\top\hat{\boldsymbol{\varepsilon}} = \boldsymbol{Y}^\top(I_n - \boldsymbol{H})\boldsymbol{Y} = \boldsymbol{Y}^\top\boldsymbol{Y} - \hat{\boldsymbol{\beta}}^\top\boldsymbol{X}^\top\boldsymbol{Y} = \mathrm{tr}(\boldsymbol{Y}\boldsymbol{Y}^\top(I_n - \boldsymbol{H}))$.*

Gauss-Markov model

Gauss-Markov model assumption

Assumption 4.1.4. For the linear model in Definition 4.1.1, we assume

$$\mathbb{E}(\boldsymbol{\varepsilon}) = \boldsymbol{0}_n, \qquad \mathbb{C}\mathrm{ov}(\boldsymbol{\varepsilon}) = \sigma^2 I_n. \tag{4.3}$$

The unknown parameters are $(\boldsymbol{\beta}, \sigma^2)$.

The mean-zero condition is an identifiability condition for the intercept component of $\boldsymbol{\beta}$.

Statistical properties of LSE

Proposition 4.1.5. *Under Assumption 4.1.4, for the LSE $\hat{\boldsymbol{\beta}}$ in (4.2), we have*

1. *(Unbiased estimate) $\mathbb{E}(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta}$.*
2. *$\mathbb{C}\mathrm{ov}(\hat{\boldsymbol{\beta}}) = \sigma^2(\boldsymbol{X}^\top\boldsymbol{X})^{-1}$.*
3. *$\hat{\boldsymbol{Y}}$ and $\hat{\boldsymbol{\varepsilon}}$ are uncorrealted:*

$$\mathbb{E}\begin{pmatrix}\hat{\boldsymbol{Y}}\\ \hat{\boldsymbol{\varepsilon}}\end{pmatrix} = \begin{pmatrix}\boldsymbol{X}\boldsymbol{\beta}\\ \boldsymbol{0}_n\end{pmatrix}, \qquad \mathbb{C}\mathrm{ov}\begin{pmatrix}\hat{\boldsymbol{Y}}\\ \hat{\boldsymbol{\varepsilon}}\end{pmatrix} = \sigma^2\begin{pmatrix}\boldsymbol{H} & \boldsymbol{0}_{n\times n}\\ \boldsymbol{0}_{n\times n} & I_n - \boldsymbol{H}\end{pmatrix}$$

   - *$\mathbb{C}\mathrm{ov}(\hat{y}_i, \hat{y}_j) = \sigma^2 h_{ij}$ and $\mathbb{C}\mathrm{ov}(\hat{\varepsilon}_i, \hat{\varepsilon}_j) = -\sigma^2 h_{ij}$ for $i \neq j$.*

4. *Let the residual error sum of squares(RSS)/sum of squares due to error(SSE) be*

$$SSE = RSS = \hat{\boldsymbol{\varepsilon}}^\top\hat{\boldsymbol{\varepsilon}} = \sum_{i=1}^n \hat{\varepsilon}_i^2, \tag{4.4}$$

then

$$\mathbb{E}(\hat{\varepsilon}_i) = \sigma^2(1 - h_{ii}), \quad \mathbb{E}(\hat{\sigma}^2) := \mathbb{E}\left\{\frac{RSS}{n - \mathrm{r}(\boldsymbol{X})}\right\} = \sigma^2. \tag{4.5}$$

Gauss-Markov theorem

Theorem 4.1.6. *Under Assumption 4.1.4, the LSE $\hat{\boldsymbol{\beta}}$ for $\boldsymbol{\beta}$ is the best linear unbiased estimator (BLUE) in the sense that*

$$\mathbb{C}\mathrm{ov}(\tilde{\boldsymbol{\beta}}) \succeq \mathbb{C}\mathrm{ov}(\hat{\boldsymbol{\beta}})$$

for $\forall\tilde{\boldsymbol{\beta}}$ satisfying: (C1) $\tilde{\boldsymbol{\beta}} = A\boldsymbol{Y}$ for some $A \in \mathbb{R}^{p\times n}$ not depending on $\boldsymbol{Y}$; and (C2) $\mathbb{E}\tilde{\boldsymbol{\beta}} = \boldsymbol{\beta}$ for any $\boldsymbol{\beta}$. In the process of the proof, we have that for all such $\tilde{\boldsymbol{\beta}}$,

$$\mathbb{C}\mathrm{ov}(\tilde{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\beta}}) = \boldsymbol{0}_p, \qquad \mathbb{C}\mathrm{ov}(\tilde{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}) = \mathbb{C}\mathrm{ov}(\tilde{\boldsymbol{\beta}}) - \mathbb{C}\mathrm{ov}(\hat{\boldsymbol{\beta}}).$$

It implies that

$$\begin{aligned}\boldsymbol{c}^\top\mathbb{C}\mathrm{ov}(\tilde{\boldsymbol{\beta}})\boldsymbol{c} \ge \boldsymbol{c}^\top\mathbb{C}\mathrm{ov}(\hat{\boldsymbol{\beta}})\boldsymbol{c} \quad &\Leftrightarrow \quad \mathbb{V}\mathrm{ar}(\boldsymbol{c}^\top\tilde{\boldsymbol{\beta}}) \ge \mathbb{V}\mathrm{ar}(\boldsymbol{c}^\top\hat{\boldsymbol{\beta}}),\\ &\quad\quad \mathbb{V}\mathrm{ar}(\tilde{\beta}_j) \ge \mathbb{V}\mathrm{ar}(\hat{\beta}_j),\ j = 1, \dots, p.\end{aligned}$$

4.1.2 Model with intercept

Model with intercept

Proposition 4.1.7. *Consider the linear model in Definition 4.1.1 with intercept $\boldsymbol{Y}_n = \boldsymbol{X}\boldsymbol{\beta}_{k+1} + \boldsymbol{\varepsilon}_n$, where $\boldsymbol{X} = (\boldsymbol{1}_n, \boldsymbol{X}_1)$, $\boldsymbol{\beta}^\top = (\beta_0, \boldsymbol{b}^\top)$, and $\boldsymbol{b}^\top = (\beta_1, \dots, \beta_k)$. Let*

$$\bar{\boldsymbol{X}} = (\bar{x}_{\cdot 1}, \dots, \bar{x}_{\cdot k})^\top \in \mathbb{R}^k, \quad \boldsymbol{X}_c = \boldsymbol{X}_1 - \boldsymbol{1}_n\bar{\boldsymbol{X}}^\top \in \mathbb{R}^{n\times k}, \quad \hat{\boldsymbol{\beta}} = (\hat{\beta}_0, \hat{\boldsymbol{b}}^\top)^\top \in \mathbb{R}^{k+1}.$$

Then we have the following facts:

1. $\boldsymbol{1}_n^\top \boldsymbol{Y} = n\bar{y}$, $\boldsymbol{1}_n^\top \boldsymbol{X}_1 = n\bar{\boldsymbol{X}}^\top$, *and* $\boldsymbol{X}_1^\top \boldsymbol{X}_1 - n\bar{\boldsymbol{X}}\bar{\boldsymbol{X}}^\top = \boldsymbol{X}_c^\top \boldsymbol{X}_c$.
2. *(The geometry of LSE)* $\boldsymbol{1}_n^\top \hat{\boldsymbol{\varepsilon}} = 0 \Rightarrow n^{-1}\sum_{i=1}^n \hat{\varepsilon}_i = 0$.
3. *(Hat matrix)*
   - $\boldsymbol{H}\boldsymbol{1}_n = \boldsymbol{1}_n \Rightarrow \sum_{j=1}^n h_{ij} = 1$, $\forall i = 1, \ldots, n$.
   - $\boldsymbol{H} = n^{-1}\boldsymbol{1}_n\boldsymbol{1}_n^\top + \boldsymbol{X}_c(\boldsymbol{X}_c^\top \boldsymbol{X}_c)^{-1}\boldsymbol{X}_c$.
   - $h_{ii} \geq n^{-1}$. *Hence* $n^{-1} \leq h_{ii} \leq 1$.
4. *(LSE decomposition)*
$$\begin{pmatrix} \hat{\beta}_0 \\ \hat{\boldsymbol{b}} \end{pmatrix} = \begin{bmatrix} \bar{y} - \bar{\boldsymbol{X}}^\top \hat{\boldsymbol{b}} \\ (\boldsymbol{X}_c^\top \boldsymbol{X}_c)^{-1}\boldsymbol{X}_c^\top \boldsymbol{Y} \end{bmatrix} = \begin{bmatrix} n^{-1}\boldsymbol{1}_n^\top \boldsymbol{Y} - \bar{\boldsymbol{X}}^\top \hat{\boldsymbol{b}} \\ (\boldsymbol{X}_c^\top \boldsymbol{X}_c)^{-1}\boldsymbol{X}_c^\top \boldsymbol{Y} \end{bmatrix}.$$
5. *(Variance) Suppose Assumption 4.1.4 holds, then*
$$\mathbb{V}\mathrm{ar}\begin{pmatrix} \hat{\beta}_0 \\ \hat{\boldsymbol{b}} \end{pmatrix} = \sigma^2 \begin{bmatrix} \frac{1}{n} + \bar{\boldsymbol{X}}^\top (\boldsymbol{X}_c^\top \boldsymbol{X}_c)^{-1}\bar{\boldsymbol{X}} & -\bar{\boldsymbol{X}}^\top (\boldsymbol{X}_c^\top \boldsymbol{X}_c)^{-1} \\ -(\boldsymbol{X}_c^\top \boldsymbol{X}_c)^{-1}\bar{\boldsymbol{X}} & (\boldsymbol{X}_c^\top \boldsymbol{X}_c)^{-1} \end{bmatrix},$$
$$\mathbb{V}\mathrm{ar}(\hat{\boldsymbol{b}}) = \sigma^2(\boldsymbol{X}_c^\top \boldsymbol{X}_c)^{-1}, \quad \mathbb{V}\mathrm{ar}(\hat{\beta}_0) = \frac{\sigma^2}{n} + \bar{\boldsymbol{X}}^\top \mathbb{V}\mathrm{ar}(\hat{\boldsymbol{b}})\bar{\boldsymbol{X}}, \quad \mathbb{C}\mathrm{ov}(\hat{\beta}_0, \hat{\boldsymbol{b}}^\top) = -\bar{\boldsymbol{X}}^\top \mathbb{V}\mathrm{ar}(\hat{\boldsymbol{b}}).$$

4.1.3 Weighted least square (WLS) estimation

WLS assumption

Assumption 4.1.8. For the linear model in Definition 4.1.1, we assume
$$\mathbb{E}(\boldsymbol{\varepsilon}) = \boldsymbol{0}_n, \qquad \mathbb{C}\mathrm{ov}(\boldsymbol{\varepsilon}) = \Sigma, \tag{4.6}$$
where $\Sigma \succ \boldsymbol{0}$ is known.

The *weighted least squares estimator* (WLSE) or *generalized least squares estimator* (GLSE) is defined as the minimizer
$$\hat{\boldsymbol{\beta}}^{\text{WLS}} = \operatorname*{arg\,min}_{\boldsymbol{\beta}\in\mathbb{R}^p} S(\boldsymbol{\beta}) \equiv (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})^\top \Sigma^{-1}(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}),$$
which satisfies
$$\text{(FONC)} \qquad \partial S(\boldsymbol{\beta})/\partial\boldsymbol{\beta} = -2\boldsymbol{X}^\top \Sigma^{-1}\boldsymbol{Y} + 2\boldsymbol{X}^\top \Sigma^{-1}\boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{0}_p,$$
$$\textit{(Normal equation)} \qquad \boldsymbol{X}^\top \Sigma^{-1}\boldsymbol{Y} = \boldsymbol{X}^\top \Sigma^{-1}\boldsymbol{X}\hat{\boldsymbol{\beta}}^{\text{WLS}}.$$
As $\boldsymbol{X}^\top \Sigma^{-1}\boldsymbol{X}$ has full rank since $\boldsymbol{X}$ and Σ^{-1} have full rank,
$$\hat{\boldsymbol{\beta}}^{\text{WLS}} = (\boldsymbol{X}^\top \Sigma^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^\top \Sigma^{-1}\boldsymbol{Y}, \qquad \mathbb{E}\hat{\boldsymbol{\beta}}^{\text{WLS}} = \boldsymbol{\beta}. \tag{4.7}$$

Remark 4.1.9.

1. Another aspect to motivate the WLS: Since $\Sigma \succ \boldsymbol{0}$, $\Sigma^{-1/2}$ exists and thus
$$\underbrace{\Sigma^{-\frac{1}{2}}\boldsymbol{Y}}_{\tilde{\boldsymbol{Y}}} = \underbrace{\Sigma^{-\frac{1}{2}}\boldsymbol{X}}_{\tilde{\boldsymbol{X}}}\boldsymbol{\beta} + \underbrace{\Sigma^{-\frac{1}{2}}\boldsymbol{\varepsilon}}_{\tilde{\boldsymbol{\varepsilon}}}., \qquad \hat{\boldsymbol{\beta}}^{\text{WLS}} = (\tilde{\boldsymbol{X}}^\top \tilde{\boldsymbol{X}})^{-1}\tilde{\boldsymbol{X}}\tilde{\boldsymbol{Y}}. \tag{4.8}$$
Now $\mathbb{E}(\tilde{\boldsymbol{\varepsilon}}) = \boldsymbol{0}_n$ and $\mathbb{C}\mathrm{ov}(\tilde{\boldsymbol{\varepsilon}}) = I_n$ satisfy Assumption 4.1.4 of the OLS.
2. Theorem 4.1.6 states that the LSE with $\Sigma = I_n$ has the smallest covariance matrix under the Gauss-Markov model.

WLS version of Gauss-Markov theorem

Corollary 4.1.10. *Under Assumption 4.1.8, the WLSE* $\hat{\boldsymbol{\beta}}^{\text{WLS}}$ *for* $\boldsymbol{\beta}$ *is the BLUE:*
$$\mathbb{C}\mathrm{ov}(\tilde{\boldsymbol{\beta}}) \succeq \mathbb{C}\mathrm{ov}(\hat{\boldsymbol{\beta}}^{\text{WLS}})$$
for $\forall\tilde{\boldsymbol{\beta}}$ *satisfying: (C1)* $\tilde{\boldsymbol{\beta}} = A\boldsymbol{Y}$ *for some* $A \in \mathbb{R}^{p\times n}$ *not depending on* $\boldsymbol{Y}$*; and (C2)* $\mathbb{E}\tilde{\boldsymbol{\beta}} = \boldsymbol{\beta}$ *for any* $\boldsymbol{\beta}$*. For all such* $\tilde{\boldsymbol{\beta}}$*,*
$$\mathbb{C}\mathrm{ov}(\tilde{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}^{\text{WLS}}, \hat{\boldsymbol{\beta}}^{\text{WLS}}) = \boldsymbol{0}_p, \qquad \mathbb{C}\mathrm{ov}(\tilde{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}^{\text{WLS}}) = \mathbb{C}\mathrm{ov}(\tilde{\boldsymbol{\beta}}) - \mathbb{C}\mathrm{ov}(\hat{\boldsymbol{\beta}}^{\text{WLS}}).$$

Remark 4.1.11. Let $\boldsymbol{t} \in \mathbb{R}^p$, by similar technique, the BLUE of $\boldsymbol{t}^\top\boldsymbol{\beta}$ is $\boldsymbol{t}^\top\hat{\boldsymbol{\beta}}^{\text{WLS}}$, i.e., for any $\boldsymbol{\lambda} \in \mathbb{R}^p$ satisfies $\mathbb{E}(\boldsymbol{\lambda}^\top\boldsymbol{Y}) = \boldsymbol{t}^\top\boldsymbol{\beta}$, we have
$$\mathbb{V}\mathrm{ar}(\boldsymbol{\lambda}^\top\boldsymbol{Y}) \geq \mathbb{V}\mathrm{ar}(\boldsymbol{t}^\top\hat{\boldsymbol{\beta}}^{\text{WLS}}) = \mathbb{V}\mathrm{ar}(\boldsymbol{t}^\top(\boldsymbol{X}^\top\Sigma^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^\top\Sigma^{-1}\boldsymbol{Y}).$$

4.1.4 LS theory of random-effect model

Full-rank linear model (random effect)

Definition 4.1.12.

$$\underset{n}{\boldsymbol{Y}} = \underset{n\times k}{\boldsymbol{X}}\underset{k}{\boldsymbol{b}} + \underset{n}{\boldsymbol{e}}, \tag{4.9}$$

where $\{Y_i, b_i, e_i\}_{i=1}^n$ are iid copies of (Y, b, e), and $\mathbb{E}\boldsymbol{b} = \boldsymbol{\theta}$ and $\mathbb{C}\mathrm{ov}(\boldsymbol{b}) = \boldsymbol{F}$, $baq \in \mathbb{R}^k$ and $\boldsymbol{F} \in \mathbb{R}^{k\times k}$ PD. Also assume

$$\mathbb{E}(\boldsymbol{e} \mid \boldsymbol{b}) = \boldsymbol{0}, \quad \mathbb{C}\mathrm{ov}(\boldsymbol{e} \mid \boldsymbol{b}) = \boldsymbol{V}.$$

We want to find the BLUE of $\boldsymbol{p}^\top\boldsymbol{b}$ where $\boldsymbol{p} \in \mathbb{R}^k$ is given.

$$\mathbb{E}\boldsymbol{Y} = \mathbb{E}\{\mathbb{E}(\boldsymbol{Y} \mid \boldsymbol{b})\} = \boldsymbol{X}\boldsymbol{\theta},$$

$$\mathbb{V}\mathrm{ar}(\boldsymbol{Y}) = \mathbb{E}\{\mathbb{V}\mathrm{ar}(\boldsymbol{Y} \mid \boldsymbol{b})\} + \mathbb{V}\mathrm{ar}\{\mathbb{E}(\boldsymbol{Y}|\boldsymbol{b})\} = \boldsymbol{V} + \boldsymbol{X}\boldsymbol{F}\boldsymbol{X}^\top,$$

$$\mathbb{C}\mathrm{ov}(\boldsymbol{Y}, \boldsymbol{p}^\top\boldsymbol{b}) = \mathbb{E}\{\mathbb{C}\mathrm{ov}(\boldsymbol{Y}, \boldsymbol{p}^\top\boldsymbol{b}|\boldsymbol{b})\} + \mathbb{C}\mathrm{ov}[\mathbb{E}(\boldsymbol{Y}|\boldsymbol{b}), \boldsymbol{p}^\top\boldsymbol{b}] = \boldsymbol{X}\boldsymbol{F}\boldsymbol{p}.$$

4.1.5 Normal linear model

Normal linear model

Assumption 4.1.13. For the linear model in Definition 4.1.1, we assume $\boldsymbol{\varepsilon} \sim \mathrm{N}_n(\boldsymbol{0}_n, \sigma^2 I_n)$. The unknown parameters are $(\boldsymbol{\beta}, \sigma^2)$.

By Assumption 4.1.13, we have

$$\boldsymbol{Y} \sim \mathrm{N}_n(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2 I_n) \qquad \Longleftrightarrow \qquad y_i \sim_{\text{iid}} \mathrm{N}(\boldsymbol{x}_i^\top\boldsymbol{\beta}, \sigma^2),\ i = 1, \ldots, n.$$

The likelihood function is

$$L(\boldsymbol{\beta}, \sigma^2) = \left(\frac{1}{2\pi\sigma^2}\right)^{\frac{n}{2}} \exp\left\{-\frac{1}{2\sigma^2}(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})^\top(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})\right\}.$$

By setting

$$\frac{\partial \log L}{\partial \boldsymbol{\beta}} = \frac{1}{\sigma^2}(\boldsymbol{X}^\top\boldsymbol{Y} - \boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{\beta}) = \boldsymbol{0}_p, \qquad \frac{\partial \log L}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})^\top(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}) = 0,$$

the MLE of $\boldsymbol{\beta}$ and σ^2 are

$$\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}^{\text{MLE}} = (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{Y}, \qquad \hat{\sigma}^{2\text{MLE}} = \frac{1}{n}(\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}^{\text{MLE}})^\top(\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}^{\text{MLE}}) = \frac{1}{n}\boldsymbol{Y}^\top(I_n - \boldsymbol{H})\boldsymbol{Y}, \tag{4.10}$$

that is, LSE is the MLE. Note that $\hat{\sigma}^{2\text{MLE}}$ is biased for finite n.

Properties under normal assumption

Proposition 4.1.14. *Under Assumption 4.1.13, we have:*

1. $\hat{\boldsymbol{\beta}}^{\text{MLE}} \sim \mathrm{N}_p(\boldsymbol{\beta}, (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\sigma^2)$.
2. *The MLE/LSE $\hat{\boldsymbol{\beta}}$ and SSE (4.4) are independent.*
3. *Recall the variance estimator $\hat{\sigma} = SSE/(n - \mathrm{r}(\boldsymbol{X}))$ in (4.5), we have*

$$\frac{SSE}{\sigma^2} \sim \chi^2_{n-\mathrm{r}(\boldsymbol{X})}, \qquad \text{or} \qquad \frac{(n - \mathrm{r}(\boldsymbol{X}))\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{n-\mathrm{r}(\boldsymbol{X})}.$$

- $\mathbb{E}\hat{\sigma} = \sqrt{\frac{2\sigma^2}{n-p}\frac{\Gamma((n-p+1)/2)}{\Gamma((n-p)/2)}}$, *and* $\mathbb{E}(\hat{\sigma}^{-2}) = \frac{n-p}{2\sigma^2}\frac{\Gamma((n-p-2)/2)}{\Gamma((n-p)/2)}$.

ANOVA

The *total sum of squares (SST)*, *sum of squares due to regression (SSR)/reduction in sum of squares*, and SSE are

$$\mathrm{SST} = \boldsymbol{Y}^\top\boldsymbol{Y},$$

$$\mathrm{SSE} = \boldsymbol{Y}^\top\boldsymbol{Y} - \hat{\boldsymbol{\beta}}^\top\boldsymbol{X}^\top\boldsymbol{Y} = \boldsymbol{Y}^\top(I_n - \boldsymbol{H})\boldsymbol{Y},$$

$$\mathrm{SSR} = \mathrm{SST} - \mathrm{SSE} = \hat{\boldsymbol{\beta}}^\top\boldsymbol{X}^\top\boldsymbol{Y} = \boldsymbol{Y}^\top\boldsymbol{H}\boldsymbol{Y},$$

which have the following distributions

$$\frac{\mathrm{SST}}{\sigma^2} \sim \chi^2_n\left(\frac{\boldsymbol{\beta}^\top\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{\beta}}{\sigma^2}\right), \qquad \frac{\mathrm{SSE}}{\sigma^2} \sim \chi^2_{n-\mathrm{r}(\boldsymbol{X})}, \qquad \frac{\mathrm{SSR}}{\sigma^2} \sim \chi^2_{\mathrm{r}(\boldsymbol{X})}\left(\frac{\boldsymbol{\beta}^\top\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{\beta}}{\sigma^2}\right),$$

we can show SSR $\perp\!\!\!\perp$ SSE, thus

$$F(R) = \frac{\mathrm{MSR}}{\mathrm{MSE}} = \frac{\mathrm{SSR}/\mathrm{r}(\boldsymbol{X})}{\mathrm{SSE}/(n - \mathrm{r}(\boldsymbol{X}))} = \frac{\mathrm{SSR}/\mathrm{r}(\boldsymbol{X})}{\hat{\sigma}^2} \sim \mathrm{F}_{[\mathrm{r}(\boldsymbol{X}), \boldsymbol{n}-\mathrm{r}(\boldsymbol{X})]}\left(\frac{\boldsymbol{\beta}^\top\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{\beta}}{\sigma^2}\right).$$

SSM, SST$_m$, SSR$_m$

Definition 4.1.15. For the model without predictors $y_i = c_0 + \varepsilon_i$, the LSE of c_0 is $\hat{c}_0 = n^{-1}\mathbf{1}_n^\top \boldsymbol{Y} = \bar{y}$. Define the SSM to be the SSR without x:

$$\mathrm{SSM} = \hat{c}_0 \mathbf{1}_n^\top \boldsymbol{Y} = \boldsymbol{Y}^\top \left(\frac{\mathbf{1}_n \mathbf{1}_n^\top}{n} \right) \boldsymbol{Y} = n\bar{y}^2.$$

The *total sum of squares corrected for the mean* and the *sum of squares of regression corrected for the mean* are

$$\mathrm{SST}_m = \mathrm{SST} - \mathrm{SSM} = \boldsymbol{Y}^\top (\boldsymbol{I}_n - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^\top)\boldsymbol{Y},$$

$$\mathrm{SSR}_m = \mathrm{SSR} - \mathrm{SSM} = \hat{\boldsymbol{\beta}}^\top \boldsymbol{X}^\top \boldsymbol{Y} - \boldsymbol{Y}^\top n^{-1}\mathbf{1}_n\mathbf{1}_n^\top \boldsymbol{Y} = \hat{\boldsymbol{b}}^\top \boldsymbol{X}_c^\top \boldsymbol{Y} = \hat{\boldsymbol{b}}^\top (\boldsymbol{X}_c^\top \boldsymbol{X}_c)\hat{\boldsymbol{b}}.$$

Distributions of SSM, SST$_m$, and SSR$_m$

Proposition 4.1.16. *Under model (4.1) and normal assumption 4.1.13,*

$$\frac{SSM}{\sigma^2} \sim \chi_1^2 \left(\frac{(\mathbf{1}_n^\top \boldsymbol{X}\boldsymbol{\beta})^2}{n\sigma^2} \right), \quad \frac{SST_m}{\sigma^2} \sim \chi_{n-1}^2 \left(\frac{\boldsymbol{\beta}^\top \boldsymbol{X}^\top \boldsymbol{X}\boldsymbol{\beta} - n^{-1}(\mathbf{1}_n^\top \boldsymbol{X}\boldsymbol{\beta})^2}{\sigma^2} \right),$$

and if the model has intercept,

$$\frac{SSR_m}{\sigma^2} \sim \chi_{\mathrm{r}(\boldsymbol{X})-1}^2 \left(\frac{\boldsymbol{b}^\top (\boldsymbol{Z}^\top \boldsymbol{Z})\boldsymbol{b}}{\sigma^2} \right).$$

Note that $SST = (SSM + SSR_m) + SSE$, each terms are mutually independent.

Construct F-statistics:

$$F(M) = \frac{\mathrm{MSM}}{\mathrm{MSE}} = \frac{\mathrm{SSM}/1}{\mathrm{SSE}/(n - \mathrm{r}(\boldsymbol{X}))} \sim \mathrm{F}_{[1,n-\mathrm{r}(\boldsymbol{X})]} \left(\frac{(\mathbf{1}_n^\top \boldsymbol{X}\boldsymbol{\beta})^2}{n\sigma^2} \right),$$

$$F(R_m) = \frac{\mathrm{MSR}_m}{\mathrm{MSE}} = \frac{\mathrm{SSR}_m/(\mathrm{r}(\boldsymbol{X}) - 1)}{\mathrm{SSE}/(n - \mathrm{r}(\boldsymbol{X}))} \sim \mathrm{F}_{[\mathrm{r}(\boldsymbol{X})-1,n-\mathrm{r}(\boldsymbol{X})]} \left(\frac{\boldsymbol{b}^\top (\boldsymbol{Z}^\top \boldsymbol{Z})\boldsymbol{b}}{\sigma^2} \right)$$

General linear hypothesis and constrained LSE

The general hypothesis we consider is

$$H_0: \ \boldsymbol{K}^\top \boldsymbol{\beta} = \boldsymbol{m}, \quad \boldsymbol{K}^\top \in \mathbb{R}^{s\times(k+1)}, \ \boldsymbol{\beta} \in \mathbb{R}^{k+1}, \ \boldsymbol{m} \in \mathbb{R}^s,$$

where $\boldsymbol{K}$ and $\boldsymbol{m}$ are specified constants, and $\boldsymbol{K}^\top$ is assumed to be of full row rank to guarantee that the linear function of $\boldsymbol{\beta}$ must be linearly independent.

Distribution of test statistics

Proposition 4.1.17. *Under Assumption 4.1.13,*

$$\boldsymbol{K}^\top \hat{\boldsymbol{\beta}} - \boldsymbol{m} \sim \mathrm{N}_s[\boldsymbol{K}^\top \boldsymbol{\beta} - \boldsymbol{m}, \sigma^2 \boldsymbol{K}^\top (\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{K}].$$

Define

$$Q = (\boldsymbol{K}^\top \hat{\boldsymbol{\beta}} - \boldsymbol{m})^\top [\boldsymbol{K}^\top (\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{K}]^{-1}(\boldsymbol{K}^\top \hat{\boldsymbol{\beta}} - \boldsymbol{m}).$$

Then $Q \perp\!\!\!\perp SSE$, and

$$\frac{Q}{\sigma^2} \sim \chi_s^2 \left\{ \sigma^{-2}(\boldsymbol{K}^\top \boldsymbol{\beta} - \boldsymbol{m})^\top [\boldsymbol{K}^\top (\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{K}]^{-1}(\boldsymbol{K}^\top \boldsymbol{\beta} - \boldsymbol{m}) \right\}.$$

Thus we have the test statistics

$$F(H) = \frac{Q/s}{SSE/[n - \mathrm{r}(\boldsymbol{X})]} = \frac{Q}{s\hat{\sigma}^2} \sim \mathrm{F}_{[s,n-\mathrm{r}(\boldsymbol{X})]} \left\{ \sigma^{-2}(\boldsymbol{K}^\top \boldsymbol{\beta} - \boldsymbol{m})^\top [\boldsymbol{K}^\top (\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{K}]^{-1}(\boldsymbol{K}^\top \boldsymbol{\beta} - \boldsymbol{m}) \right\}.$$

When H_0 is true, the solution of the constrained LSE $\text{minimize}_{\boldsymbol{\beta}} \|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}\|^2$ subject to $\boldsymbol{K}^\top \boldsymbol{\beta} = \boldsymbol{m}$ is

$$\begin{aligned} \tilde{\boldsymbol{\beta}} &= (\boldsymbol{X}^\top \boldsymbol{X})^{-1}[\boldsymbol{X}^\top \boldsymbol{Y} - \boldsymbol{K}(\boldsymbol{K}^\top (\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{K})^{-1}(\boldsymbol{K}^\top (\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top \boldsymbol{Y} - \boldsymbol{m})] \\ &= \hat{\boldsymbol{\beta}} - (\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{K}(\boldsymbol{K}^\top (\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{K})^{-1}(\boldsymbol{K}^\top \hat{\boldsymbol{\beta}} - \boldsymbol{m}). \end{aligned}$$

Under the reduced model (H_0), the SSE is

$$\mathrm{SSE}_{H_0} = (\boldsymbol{Y} - \boldsymbol{X}\tilde{\boldsymbol{\beta}})^\top (\boldsymbol{Y} - \boldsymbol{X}\tilde{\boldsymbol{\beta}}) = \mathrm{SSE} + Q$$

4.1.6 Proofs

Proof of Proposition 4.1.5

Proof. 1.

$$\mathbb{E}(\hat{\boldsymbol{\beta}}) = \mathbb{E}((\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top \boldsymbol{Y}) = (\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top \mathbb{E}(\boldsymbol{Y}) = (\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top \boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{\beta}.$$

2.

$$\begin{aligned}\mathbb{C}\mathrm{ov}(\hat{\boldsymbol\beta}) &= \mathbb{C}\mathrm{ov}((\boldsymbol X^\top \boldsymbol X)^{-1}\boldsymbol X^\top \boldsymbol Y) = (\boldsymbol X^\top \boldsymbol X)^{-1}\boldsymbol X^\top \mathbb{C}\mathrm{ov}(\boldsymbol Y)\boldsymbol X(\boldsymbol X^\top \boldsymbol X)^{-1}\\ &= \sigma^2(\boldsymbol X^\top \boldsymbol X)^{-1}\boldsymbol X^\top I_n \boldsymbol X(\boldsymbol X^\top \boldsymbol X)^{-1} = \sigma^2(\boldsymbol X^\top \boldsymbol X)^{-1}.\end{aligned}$$

3.

$$\mathbb{E}\begin{pmatrix}\hat{\boldsymbol Y}\\ \hat{\varepsilon}\end{pmatrix} = \begin{pmatrix}\boldsymbol H\\ I_n-\boldsymbol H\end{pmatrix}\mathbb{E}(\boldsymbol Y) = \begin{pmatrix}\boldsymbol H\\ I_n-\boldsymbol H\end{pmatrix}\boldsymbol X\boldsymbol\beta = \begin{pmatrix}\boldsymbol H\boldsymbol X\boldsymbol\beta\\ (I_n-\boldsymbol H)\boldsymbol X\boldsymbol\beta\end{pmatrix} = \begin{pmatrix}\boldsymbol X\boldsymbol\beta\\ \boldsymbol 0_n\end{pmatrix}.$$

$$\begin{aligned}\mathbb{C}\mathrm{ov}\begin{pmatrix}\hat{\boldsymbol Y}\\ \hat{\varepsilon}\end{pmatrix} &= \begin{pmatrix}\boldsymbol H\\ I_n-\boldsymbol H\end{pmatrix}\mathbb{C}\mathrm{ov}(\boldsymbol Y)\begin{pmatrix}\boldsymbol H^\top & (I_n-\boldsymbol H)^\top\end{pmatrix} = \sigma^2\begin{pmatrix}\boldsymbol H\\ I_n-\boldsymbol H\end{pmatrix}\begin{pmatrix}\boldsymbol H & I_n-\boldsymbol H\end{pmatrix}\\ &= \sigma^2\begin{pmatrix}\boldsymbol H^2 & \boldsymbol H(I_n-\boldsymbol H)\\ (I_n-\boldsymbol H)\boldsymbol H & (I_n-\boldsymbol H)^2\end{pmatrix} = \sigma^2\begin{pmatrix}\boldsymbol H & \boldsymbol 0\\ \boldsymbol 0 & I_n-\boldsymbol H\end{pmatrix},\end{aligned}$$

4. By item 3, $\mathbb{E}(\hat\varepsilon_i) = \sigma^2(1-h_{ii})$, hence $\mathbb{E}(\mathrm{RSS}) = \sum_{i=1}^n \sigma^2(1-h_{ii}) = \sigma^2\{n-\mathrm{tr}(H)\} = \sigma^2(n-p)$. Alternatively, we can directly show

$$\mathbb{E}(\hat{\boldsymbol\varepsilon}^\top\hat{\boldsymbol\varepsilon}) = \mathbb{E}(\boldsymbol Y^\top(I_n-\boldsymbol H)\boldsymbol Y) = \mathrm{tr}((\boldsymbol I_n-\boldsymbol H)\Sigma) + \boldsymbol\beta^\top\boldsymbol X^\top(I_n-\boldsymbol H)\boldsymbol X\boldsymbol\beta = \sigma^2\mathrm{tr}(I_n-\boldsymbol H) = \sigma^2(n-p).$$

□

Proof of Theorem 4.1.6

Proof. We must verify that the OLS estimator itself satisfies (C1) and (C2). We have $\hat{\boldsymbol\beta} = \hat A\boldsymbol Y$ with $\hat A = (\boldsymbol X^\top\boldsymbol X)^{-1}\boldsymbol X^\top$, and it is unbiased by Proposition 4.1.5.

First, the unbiasedness requirement implies that

$$\mathbb{E}(\tilde{\boldsymbol\beta}) = \boldsymbol\beta \implies \mathbb{E}(A\boldsymbol Y) = A\mathbb{E}(\boldsymbol Y) = A\boldsymbol X\boldsymbol\beta = \boldsymbol\beta \implies A\boldsymbol X\boldsymbol\beta = \boldsymbol\beta, \quad \forall\boldsymbol\beta.$$

So $A\boldsymbol X = I_p$ must hold. In particular, the OLS estimator satisfies $\hat A\boldsymbol X = (\boldsymbol X^\top\boldsymbol X)^{-1}\boldsymbol X^\top\boldsymbol X = I_p$.

Second, we can decompose the covariance of $\tilde{\boldsymbol\beta}$ as

$$\begin{aligned}\mathbb{C}\mathrm{ov}(\tilde{\boldsymbol\beta}) &= \mathbb{C}\mathrm{ov}(\hat{\boldsymbol\beta} + \tilde{\boldsymbol\beta} - \hat{\boldsymbol\beta})\\ &= \mathbb{C}\mathrm{ov}(\hat{\boldsymbol\beta}) + \mathbb{C}\mathrm{ov}(\tilde{\boldsymbol\beta} - \hat{\boldsymbol\beta}) + \mathbb{C}\mathrm{ov}(\hat{\boldsymbol\beta}, \tilde{\boldsymbol\beta} - \hat{\boldsymbol\beta}) + \mathbb{C}\mathrm{ov}(\tilde{\boldsymbol\beta} - \hat{\boldsymbol\beta}, \hat{\boldsymbol\beta}).\end{aligned}$$

The last two terms are in fact zero. By symmetry, we only need to show that the third term is zero:

$$\begin{aligned}\mathbb{C}\mathrm{ov}(\hat{\boldsymbol\beta}, \tilde{\boldsymbol\beta} - \hat{\boldsymbol\beta}) &= \mathbb{C}\mathrm{ov}\{\hat A\boldsymbol Y, (A-\hat A)\boldsymbol Y\} = \sigma^2\hat A(A-\hat A)^\top\\ &= \sigma^2(\hat A A^\top - \hat A\hat A^\top)\\ &= \sigma^2\left\{(\boldsymbol X^\top\boldsymbol X)^{-1}\boldsymbol X^\top A^\top - (\boldsymbol X^\top\boldsymbol X)^{-1}\boldsymbol X^\top\boldsymbol X(\boldsymbol X^\top\boldsymbol X)^{-1}\right\}\\ &= \sigma^2\{(\boldsymbol X^\top\boldsymbol X)^{-1}I_p - (\boldsymbol X^\top\boldsymbol X)^{-1}\} = \boldsymbol 0_p.\end{aligned}$$

The above covariance decomposition simplifies to

$$\mathbb{C}\mathrm{ov}(\tilde{\boldsymbol\beta}) = \mathbb{C}\mathrm{ov}(\hat{\boldsymbol\beta}) + \mathbb{C}\mathrm{ov}(\tilde{\boldsymbol\beta} - \hat{\boldsymbol\beta}),$$

which implies

$$\mathbb{C}\mathrm{ov}(\tilde{\boldsymbol\beta}) - \mathbb{C}\mathrm{ov}(\hat{\boldsymbol\beta}) = \mathbb{C}\mathrm{ov}(\tilde{\boldsymbol\beta} - \hat{\boldsymbol\beta}) \succeq \boldsymbol 0.$$

□

Proof of Corollary 4.1.10

Proof. The WLSE satisfies conditions (C1)–(C2). We write WLS in the form of OLS (4.8) such that $\hat{\boldsymbol\beta}^{\mathrm{WLS}} = (\tilde{\boldsymbol X}^\top\tilde{\boldsymbol X})^{-1}\tilde{\boldsymbol X}\tilde{\boldsymbol Y}$. Then by Theorem 4.1.6,

$$\mathbb{C}\mathrm{ov}(\tilde{\boldsymbol\beta}') \succeq \mathbb{C}\mathrm{ov}(\hat{\boldsymbol\beta}^{\mathrm{WLS}})$$

for all $\tilde{\boldsymbol\beta}'$ satisfying: (1) $\tilde{\boldsymbol\beta}' = \tilde A\tilde{\boldsymbol Y}$ for some $\tilde A \in \mathbb{R}^{p\times n}$ not depending on $\tilde{\boldsymbol Y}$; and (2) $\mathbb{E}\tilde{\boldsymbol\beta}' = \boldsymbol\beta$. Suppose $\tilde{\boldsymbol\beta}$ satisfies: (C1) $\tilde{\boldsymbol\beta} = A\boldsymbol Y$ for some $A \in \mathbb{R}^{p\times n}$ not depending on $\boldsymbol Y$; and (C2) $\mathbb{E}\tilde{\boldsymbol\beta} = \boldsymbol\beta$. Then $\tilde{\boldsymbol\beta} = (A\Sigma^{1/2})\Sigma^{-1/2}\boldsymbol Y = (A\Sigma^{1/2})\tilde{\boldsymbol Y}$, i.e. satisfying (1); and (2) is satisfied by definition. So

$$\mathbb{C}\mathrm{ov}(\tilde{\boldsymbol\beta}) \succeq \mathbb{C}\mathrm{ov}(\hat{\boldsymbol\beta}^{\mathrm{WLS}}).$$

□

Proof of Proposition 4.1.14

Proof. Item 1 is obvious. Item 2 applies Theorem 3.1.10. Item 3 uses Theorem 3.1.12 to show the χ^2 distribution and (1.1) to show the moments. □

4.2 Linear regression for the non full-rank models

4.3 Quantile Regression

4.3.1 Sample quantile

For a random variable y, we can define its distribution function as $F(c) = \mathbb{P}(y \leq c)$ and its τth quantile as

$$F^{-1}(\tau) = \inf\{q : F(q) \geq \tau\}.$$

Check function and quantile

Proposition 4.3.1. *With a monotone distribution function and positive density at the τth quantile, we have*

$$F^{-1}(\tau) = \arg\min_{q\in\mathbb{R}} \mathbb{E}\left\{\rho_\tau(y-q)\right\}, \tag{4.11}$$

where

$$\rho_\tau(u) = u\left\{\tau - \mathbf{1}_{(u<0)}\right\} = \begin{cases} u\tau, & if\ u \geq 0, \\ -u(1-\tau), & if\ u < 0, \end{cases}$$

is the check function. In particular, the median of y is

$$\text{median}(y) = F^{-1}(0.5) = \arg\min_{q\in\mathbb{R}} \mathbb{E}\left\{|y-q|\right\}.$$

The empirical distribution function is $\hat{F}(c) = n^{-1}\sum_{i=1}^n \mathbf{1}_{(y_i \leq c)}$, which is a step function, increasing but not strictly monotone.

Sample quantile

Definition 4.3.2. We define the *sample quantile* as

$$\hat{F}^{-1}(\tau) = \arg\min_{q\in\mathbb{R}} n^{-1}\sum_{i=1}^n \rho_\tau(y_i - q). \tag{4.12}$$

$\hat{F}^{-1}(\tau)$ may not be unique even though the population quantile is. We can view $\hat{F}^{-1}(\tau)$ as a set containing all minimizers, and with large samples the values in the set do not differ much.

CLT for sample quantile

Theorem 4.3.3. *Assume that $(y_i)_{i=1}^n \sim_{\text{iid}} y$ with distribution function $F(\cdot)$ that is strictly increasing and density function $f(\cdot)$ that is positive at the τth quantile. The sample quantile is consistent for the true quantile and is asymptotically Normal:*

$$\sqrt{n}\left\{\hat{F}^{-1}(\tau) - F^{-1}(\tau)\right\} \xrightarrow{\text{d}} \text{N}\left(0, \frac{\tau(1-\tau)}{\left[f\left\{F^{-1}(\tau)\right\}\right]^2}\right).$$

In particular, the sample median satisfies

$$\sqrt{n}\left\{\hat{F}^{-1}(0.5) - \text{median}(y)\right\} \xrightarrow{\text{d}} \text{N}\left(0, \frac{1}{4\left[f\left\{\text{median}(y)\right\}\right]^2}\right).$$

4.3.2 Linear regression quantile

Conditional quantile function

Definition 4.3.4. Let $\boldsymbol{x} \in \mathbb{R}^p$. The *conditional quantile function* is defined as

$$F^{-1}(\tau \mid \boldsymbol{x}) = \arg\min_{q(\cdot)} \mathbb{E}[\rho_\tau\{y - q(\boldsymbol{x})\}].$$

We can use a linear function $\boldsymbol{x}^\top\boldsymbol{\beta}(\tau)$ to approximate the conditional quantile function with

$$\boldsymbol{\beta}(\tau) = \arg\min_{\boldsymbol{b}\in\mathbb{R}^p} \mathbb{E}\{\rho_\tau(y - \boldsymbol{x}^\top\boldsymbol{b})\}$$

called the τth *population regression quantile*, and

$$\hat{\boldsymbol{\beta}}(\tau) = \arg\min_{\boldsymbol{b}\in\mathbb{R}^p} \frac{1}{n}\sum_{i=1}^n \{\rho_\tau(y_i - \boldsymbol{x}_i^\top\boldsymbol{b})\}$$

called the τth *sample regression quantile*. As a special case, when $\tau = 0.5$, we have the *regression median/least absolute deviations (LAD)*:

$$\hat{\boldsymbol{\beta}}(0.5) = \arg\min_{\boldsymbol{b}\in\mathbb{R}^p} \frac{1}{n}\sum_{i=1}^n \left|y_i - \boldsymbol{x}_i^\top\boldsymbol{b}\right|.$$

Conditional quantile model: $F^{-1}(\tau \mid \boldsymbol{x}) = \boldsymbol{x}^\top \boldsymbol{\beta}(\tau)$, where $\beta_j(\tau)$ is the partial influence of x_{ij} on the τth conditional quantile of y_i given $\boldsymbol{x}_i$.

CLT for regression quantiles

Theorem 4.3.5. *Assume* $(y_i, \boldsymbol{x}_i)_{i=1}^n \sim_{\text{iid}} (y, \boldsymbol{x})$, *under some regularity conditions, we have*

$$\sqrt{n}\{\hat{\boldsymbol{\beta}}(\tau) - \boldsymbol{\beta}(\tau)\} \xrightarrow{\text{d}} \mathrm{N}_p(\mathbf{0}_p, B^{-1}MB^{-1}),$$

where

$$B = \mathbb{E}\left[f_{y|\boldsymbol{x}}\{\boldsymbol{x}^\top \boldsymbol{\beta}(\tau)\}\boldsymbol{x}\boldsymbol{x}^\top\right], \qquad M = \mathbb{E}\left[\{\tau - \mathbf{1}_{(y - \boldsymbol{x}^\top \boldsymbol{\beta}(\tau) \le 0)}\}^2 \boldsymbol{x}\boldsymbol{x}^\top\right].$$

4.3.3 Proofs

Proof of Proposition 4.3.1

Proof. We use Leibniz's integral rule. Write

$$\mathbb{E}\{\rho_\tau(y-q)\} = \mathbb{E}\{(y-q)(\tau - \mathbf{1}_{(y<q)})\} = \int_{-\infty}^{q} (\tau - 1)(c-q)\mathrm{d}F(c) + \int_{q}^{\infty} \tau(c-q)\mathrm{d}F(c).$$

To minimize it over q, we can solve the FONC by assuming that the derivative and integral can be exchanged:

$$\frac{\mathrm{d}\mathbb{E}\{\rho_\tau(y-q)\}}{\mathrm{d}q} = (1-\tau)\int_{-\infty}^{q} \mathrm{d}F(c) - \tau \int_{q}^{\infty} \mathrm{d}F(c) = (1-\tau)F(q) - \tau\{1 - F(q)\} = 0,$$

$$\Rightarrow \tau = F(q).$$

So the τth quantile satisfies the FONC. If y has density $f(\cdot)$, the second-order condition ensures it is the minimizer:

$$\left.\frac{\mathrm{d}^2\mathbb{E}\{\rho_\tau(y-q)\}}{\mathrm{d}q^2}\right|_{q=F^{-1}(\tau)} = f\{F^{-1}(\tau)\} > 0,$$

by Leibniz's integral rule again. □

Proof of Theorem 4.3.3

Proof. Recall the definition (4.12) of sample quantile

$$\hat{F}^{-1}(\tau) = \arg\min_{q \in \mathbb{R}} n^{-1} \sum_{i=1}^{n} \rho_\tau(y_i - q).$$

Since

$$\frac{\mathrm{d}\rho_\tau(y_i - q)}{\mathrm{d}q} = -\tau + \mathbf{1}_{(y_i < q)} = \begin{cases} -\tau & \text{if } y > q, \\ 1 - \tau & \text{if } y < q, \end{cases}$$

FONC gives the estimating equation (2.3)

$$\bar{m}(y_{1:n}, q) = \frac{1}{n}\sum_{i=1}^{n} \frac{\mathrm{d}\rho_\tau(y_i - q)}{\mathrm{d}q} = \frac{1}{n}\sum_{i=1}^{n} \underbrace{\{-\tau + \mathbf{1}_{(y_i<q)}\}}_{m(y_i,q)} = 0.$$

As $F(\cdot)$ is strictly increasing, and $\mathbb{E}\{m(y,q)\} = -\tau + F(q)$, the unique solution q_0 of $\mathbb{E}\{m(y,q)\} = 0$ is $q_0 = F^{-1}(\tau)$, the unique true τth quantile. Besides, the solution $\hat{q}$ of $\bar{m}(y_{1:n}, q) = 0$ is the sample quantile $\hat{q} = \hat{F}^{-1}(\tau)$ by definition.

Now we find B and M in Theorem 2.3.2:

$$\frac{\mathrm{d}\mathbb{E}m(y,q)}{\mathrm{d}q} = F'(q) = f(q) \;\Rightarrow\; B = -\frac{\mathrm{d}\mathbb{E}m(y,q_0)}{\mathrm{d}q} = -f(F^{-1}(\tau)),$$

$$\mathbb{E}m(y,q)^2 = \tau^2 - 2\tau F(q) + F(q) \;\Rightarrow\; M = \mathbb{E}m(y,q_0)^2 = \tau(1-\tau).$$

By Theorem 2.3.2,

$$\sqrt{n}(\hat{q} - q_0) = \sqrt{n}(\hat{F}^{-1}(\tau) - F^{-1}(\tau)) \xrightarrow{\text{d}} \mathrm{N}\left(0, \frac{\tau(1-\tau)}{\left[f\left\{F^{-1}(\tau)\right\}\right]^2}\right).$$

□

Proof of Theorem 4.3.5

Proof. □

4.4 Regularization

4.4.1 Motivation

4.4.2 Ridge regression

Consider the linear regression model

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \qquad \boldsymbol{X} = (\boldsymbol{x}_1, \dots, \boldsymbol{x}_n)^\top \in \mathbb{R}^{n \times p}.$$

Ridge regression shrinks the regression coeffcients by imposing a penalty on their size

Ridge regression

Definition 4.4.1. The *ridge regression* is

$$\underset{\boldsymbol{\beta} \in \mathbb{R}^p}{\text{minimize}} \; \|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}\|^2 \qquad \text{s.t.} \qquad \|\boldsymbol{\beta}\|^2 \le t, \tag{4.13}$$

where by convention, $\boldsymbol{Z}$ is assumed to be standardized and $\boldsymbol{Y}$ is assumed to be centered. Equivalently, we can write the ridge coeffcients as the minimizer of the following *penalized residual sum of square (PRSS)*:

$$\begin{aligned} \hat{\boldsymbol{\beta}}_\lambda^{\text{RIDGE}} &= \underset{\boldsymbol{\beta} \in \mathbb{R}^p}{\arg\min} \; \text{PRSS}(\boldsymbol{\beta})_{\ell_2} = \underset{\boldsymbol{\beta} \in \mathbb{R}^p}{\arg\min} \; \|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}\|^2 + \lambda \|\boldsymbol{\beta}\|^2 \\ &= (\boldsymbol{X}^\top \boldsymbol{X} + \lambda I_p)^{-1} \boldsymbol{X}^\top \boldsymbol{Y}, \end{aligned} \tag{4.14}$$

where $\lambda \ge 0$ is the tunning parameter.

For a given λ, we can always find a t such that the solutions from (4.13) and (4.14) are identical. The corresponding t and λ satisfy $t = \|\hat{\boldsymbol{\beta}}_\lambda^{\text{RIDGE}}\|^2$.

Remark 4.4.2. Ridge regression can be translated into the LS problem alike with the data augmentation approach. The ℓ_2-PRSS can be written as

$$\text{PRSS}(\boldsymbol{\beta})_{\ell_2} = \sum_{i=1}^n (y_i - \boldsymbol{x}_i^\top \boldsymbol{\beta})^2 + \sum_{j=1}^p (0 - \sqrt{\lambda}\beta_j)^2 = \|\boldsymbol{Y}_\lambda - \boldsymbol{X}_\lambda \boldsymbol{\beta}\|^2,$$

where

$$\boldsymbol{Y}_\lambda = \begin{pmatrix} \boldsymbol{Y} \\ \boldsymbol{0}_p \end{pmatrix}, \qquad \boldsymbol{X}_\lambda = \begin{pmatrix} \boldsymbol{X} \\ \sqrt{\lambda} I_p. \end{pmatrix}$$

The LSE for the augmented data set is $(\boldsymbol{X}_\lambda^\top \boldsymbol{X}_\lambda)^{-1} \boldsymbol{X}_\lambda^\top \boldsymbol{Y}_\lambda = \hat{\boldsymbol{\beta}}_\lambda^{\text{RIDGE}}$.

We can compute the ridge solutions via the SVD of $\boldsymbol{X} = UDV^\top$, where $U = (\boldsymbol{u}_1, \dots, \boldsymbol{u}_p) \in \mathbb{R}^{n \times p}$ is orthogonal, $D = \text{diag}(d_1, \dots, d_p)$ consisting of the singular values $d_1 \ge d_2 \ge \cdots \ge d_p \ge 0$, and $V = (\boldsymbol{v}_1, \dots, \boldsymbol{v}_p) \in \mathbb{R}^{p \times p}$ is orthogonal. We can show

$$\hat{\boldsymbol{\beta}}_\lambda^{\text{RIDGE}} = V \text{diag}\left(\frac{d_j}{d_j^2 + \lambda}\right) U^\top \boldsymbol{Y} = \left(\sum_{j=1}^p \frac{d_j}{d_j^2 + \lambda} \boldsymbol{v}_j \boldsymbol{u}_j^\top\right) \boldsymbol{Y},$$

$$\hat{\boldsymbol{Y}}^{\text{RIDGE}} = \boldsymbol{X}\hat{\boldsymbol{\beta}}_\lambda^{\text{RIDGE}} = \underbrace{U \text{diag}\left(\frac{d_j^2}{d_j^2 + \lambda}\right) U^\top}_{\boldsymbol{H}_\lambda = \boldsymbol{X}(\boldsymbol{X}^\top \boldsymbol{X} + \lambda I_p)^{-1} \boldsymbol{X}^\top} \boldsymbol{Y} = \left(\sum_{j=1}^p \frac{d_j^2}{d_j^2 + \lambda} \boldsymbol{u}_j \boldsymbol{u}_j^\top\right) \boldsymbol{Y}.$$

The *effective degrees of freedom* for $\boldsymbol{H}_\lambda$ is

$$\text{df}(\boldsymbol{H}_\lambda) = \text{tr}(\boldsymbol{H}_\lambda) = \sum_{j=1}^p \frac{d_j^2}{d_j^2 + \lambda}.$$

Remark 4.4.3. Ridge regression also has a relationship with PCA. $d_j \boldsymbol{u}_j$ is the jth PC of $\boldsymbol{X}$, hence, ridge regression projects $\boldsymbol{X}$ onto these components with large d_j. Ridge regression shrinks the coeffcients of low-variance (small singular value) components.

MSE of $\hat{\boldsymbol{\beta}}_\lambda^{\text{RIDGE}}$

Proposition 4.4.4. *Suppose Assumption 4.1.4 holds. Then*

$$\begin{aligned}
\mathbb{E}\hat{\boldsymbol{\beta}}_\lambda^{\text{RIDGE}} &= V\mathrm{diag}\left(\frac{d_j^2}{d_j^2+\lambda}\right)V^\top\boldsymbol{\beta},\\
\mathbb{E}\hat{\boldsymbol{\beta}}_\lambda^{\text{RIDGE}} - \boldsymbol{\beta} &= -\lambda(\boldsymbol{X}^\top\boldsymbol{X}+\lambda I_p)^{-1}\boldsymbol{\beta} = -\lambda V(D^2+\lambda I_p)^{-1}V^\top\boldsymbol{\beta},\\
\mathbb{V}\mathrm{ar}(\hat{\boldsymbol{\beta}}_\lambda^{\text{RIDGE}}) &= \sigma^2(\boldsymbol{X}^\top\boldsymbol{X}+\lambda I_p)^{-1}\boldsymbol{X}^\top\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X}+\lambda I_p)^{-1} = \sigma^2 V\mathrm{diag}\left(\frac{d_j^2}{(d_j^2+\lambda)^2}\right)V^\top,\\
MSE(\hat{\boldsymbol{\beta}}_\lambda^{\text{RIDGE}}) &= \lambda^2\sum_{j=1}^p\frac{\gamma_j^2}{(d_j^2+\lambda)^2} + \sigma^2\sum_{j=1}^p\frac{d_j^2}{(d_j^2+\lambda)^2},\ \gamma_j = \boldsymbol{v}_j^\top\boldsymbol{\beta}.
\end{aligned}$$

We can show $\frac{\mathrm{d}MSE(\hat{\boldsymbol{\beta}}_\lambda^{\text{RIDGE}})}{\mathrm{d}\lambda}|_{\lambda=0} < 0$*, hence* $\exists\lambda > 0$ *such that* $MSE(\hat{\boldsymbol{\beta}}_\lambda^{\text{RIDGE}}) < MSE(\hat{\boldsymbol{\beta}}_0^{\text{RIDGE}}) = MSE(\hat{\boldsymbol{\beta}}^{\text{LSE}})$

4.4.3 LASSO

Consider the linear regression model

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \qquad \boldsymbol{X} = (\boldsymbol{X}_1,\ldots,\boldsymbol{X}_p) = (\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n)^\top \in \mathbb{R}^{n\times p}.$$

LASSO

Definition 4.4.5. The *least absolute shrinkage and selection operator (LASSO)* is defined as the ℓ_1 optimization problem:

$$\underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\text{minimize}}\ \frac{1}{2n}\|\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta}\|^2, \quad \text{s.t.} \quad \|\boldsymbol{\beta}\|_1 \le t. \tag{4.15}$$

The Lagrangian form is

$$\underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\text{minimize}}\ \frac{1}{2n}\|\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta}\|^2 + \lambda\|\boldsymbol{\beta}\|_1,\ \lambda \ge 0. \tag{4.16}$$

Denote $\hat{\boldsymbol{\beta}}_\lambda^{\text{LASSO}}$ the resulting LASSO estimate of (4.16).

There is an one-to-one correspondence between the constrained problem (4.15) and the Lagrangian form (4.16): for each value of t in the range where the constraint $\|\boldsymbol{\beta}\|_1 \le t$ is active, there is a corresponding value of λ that yields the same solution from (4.16). Conversely, the solution $\hat{\boldsymbol{\beta}}_\lambda^{\text{LASSO}}$ solves the bound problem with $t = \|\hat{\boldsymbol{\beta}}_\lambda^{\text{LASSO}}\|_1$.

LASSO solution

Proposition 4.4.6. *Necessary and suffcient conditions for a solution to problem (4.16) take the form*

$$-\frac{1}{n}\langle\boldsymbol{X}_j, \boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta}\rangle + \lambda s_j = 0, \quad j = 1,\ldots,p, \tag{4.17}$$

where $s_j = \mathrm{sign}(\beta_j)$ *if* $\beta_j \ne 0$*, and* $s_j \in [-1,1]$ *if* $\beta_j = 0$*.*

In particular, for a single predictor setting based on $\{Z_i, y_i\}_{i=1}^n$, the problem is to solve

$$\underset{\beta\in\mathbb{R}}{\text{minimize}}\ \frac{1}{2n}\sum_{i=1}^n(y_i - x_i\beta)^2 + \lambda|\beta|, \qquad \boldsymbol{y} = (y_1,\ldots,y_n)^\top,\ \boldsymbol{x} = (x_1,\ldots,x_n)^\top.$$

The solution is

$$\hat{\beta}_\lambda^{\text{LASSO}} = \mathcal{S}_\lambda\left(\frac{1}{n}\langle\boldsymbol{y},\boldsymbol{z}\rangle\right) = \begin{cases}\frac{1}{n}\langle\boldsymbol{y},\boldsymbol{z}\rangle - \lambda, & \text{if } \frac{1}{n}\langle\boldsymbol{y},\boldsymbol{z}\rangle > \lambda,\\ 0, & \text{if } \frac{1}{n}|\langle\boldsymbol{y},\boldsymbol{z}\rangle| \le \lambda,\\ \frac{1}{n}\langle\boldsymbol{y},\boldsymbol{z}\rangle + \lambda, & \text{if } \frac{1}{n}\langle\boldsymbol{y},\boldsymbol{z}\rangle < -\lambda,\end{cases}$$

where $\mathcal{S}_\lambda(x) = \mathrm{sign}(x)(|x|-\lambda)_+$ is the *soft-thresholding operator*.

Chapter 5

Generalized Linear Models

5.1 Logistic regression

5.1.1 Logistic regression for binary outcomes

Logistic model

Definition 5.1.1. Let the binary outcomes $y_i \in \{0, 1\}$ and p-dim covariates $\boldsymbol{x}_i \sim f_{\boldsymbol{x}}$ for $i = 1, \ldots, n$. Model

$$\mathbb{P}(y_i = 1 \mid \boldsymbol{x}_i) = g(\boldsymbol{x}_i^\top \boldsymbol{\beta}),$$

where $g(\cdot) : \mathbb{R} \to [0, 1]$ is the link function. We have:

1. (*Logit model*) $g(z) = e^z / (1 + e^z)$. It is the distribution function of the standard logistic distribution with density $g'(z) = e^z/(1 + e^z)^2 = g(z)\{1 - g(z)\}$.
2. (*Probit model*) $g(z) = \Phi(z)$.
3. (*Cauchit model*) $g(z) = \pi^{-1} \arctan(z) + 1/2$. It is the distribution function of the standard Cauchy distribution with density $g'(z) = \pi^{-1}(1 + z^2)^{-1}$.
4. (*Cloglog model*) $g(z) = 1 - \exp(-e^z)$. It is the distribution function of the standard log-Weilbull distribution with density $g'(z) = \exp(z - e^z)$.

We mainly focus on the logit model:

$$\mathbb{P}(y_i = 1 \mid \boldsymbol{x}_i) = g(\boldsymbol{x}_i^\top \boldsymbol{\beta}) = \frac{e^{\boldsymbol{x}_i^\top \boldsymbol{\beta}}}{1 + e^{\boldsymbol{x}_i^\top \boldsymbol{\beta}}},$$

with likelihood, log-likelihood, and MLE are

$$\begin{aligned}
L(\boldsymbol{\beta}, f_{\boldsymbol{x}} \mid y_{1:n}, \boldsymbol{x}_{1:n}) &= \prod_{i=1}^n \{g(\boldsymbol{x}_i^\top \boldsymbol{\beta})\}^{y_i} \{1 - g(\boldsymbol{x}_i^\top \boldsymbol{\beta})\}^{1 - y_i} f_{\boldsymbol{x}}(\boldsymbol{x}_i), \\
\ell(\boldsymbol{\beta}, f_{\boldsymbol{x}} \mid y_{1:n}, \boldsymbol{x}_{1:n}) &= \sum_{i=1}^n \left[y_i \log g(\boldsymbol{x}_i^\top \boldsymbol{\beta}) + (1 - y_i) \log\{1 - g(\boldsymbol{x}_i^\top \boldsymbol{\beta})\} \right] + \sum_{i=1}^n \log f_{\boldsymbol{x}}(\boldsymbol{x}_i), \\
\hat{\boldsymbol{\beta}} &= \operatorname*{arg\,max}_{\boldsymbol{\beta} \in \mathbb{R}^p} \ell(\boldsymbol{\beta}, f_{\boldsymbol{x}} \mid y_{1:n}, \boldsymbol{x}_{1:n}).
\end{aligned}$$

Properties of logit model by M-estimation

Proposition 5.1.2.

1. *(FONC and estimating equation for $\hat{\boldsymbol{\beta}}$)*

$$\bar{\boldsymbol{m}}((y_{1:n}, \boldsymbol{x}_{1:n}), \boldsymbol{\beta}) = \frac{\partial \ell(\boldsymbol{\beta}, f_{\boldsymbol{x}} \mid \cdot)}{\partial \boldsymbol{\beta}} = \sum_{i=1}^n \underbrace{\{y_i - g(\boldsymbol{x}_i^\top \boldsymbol{\beta})\} \boldsymbol{x}_i}_{\boldsymbol{m}((y_i, \boldsymbol{x}_i), \boldsymbol{\beta})}.$$

2. *(Second-order derivative)*

$$\frac{\partial \boldsymbol{m}((y, \boldsymbol{x}), \boldsymbol{\beta})}{\partial \boldsymbol{\beta}^\top} = -g(\boldsymbol{x}^\top \boldsymbol{\beta})\{1 - g(\boldsymbol{x}^\top \boldsymbol{\beta})\} \boldsymbol{x} \boldsymbol{x}^\top.$$

3. *(CLT for $\hat{\boldsymbol{\beta}}$) Suppose $\boldsymbol{\beta}_0$ is the true parameter, then*

$$\begin{gathered}
\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \xrightarrow{\mathrm{d}} \mathrm{N}_p(\mathbf{0}_p, A^{-1} B A^{-\top}) = \mathrm{N}_p(\mathbf{0}_p, B^{-1}), \\
A = \mathbb{E} g(\boldsymbol{x}^\top \boldsymbol{\beta}_0)\{1 - g(\boldsymbol{x}^\top \boldsymbol{\beta}_0)\} \boldsymbol{x} \boldsymbol{x}^\top, \quad B = -A = I(\boldsymbol{\beta}_0).
\end{gathered}$$

Chapter 6

Bayesian Inference

Chapter 7

Structural Equation Model (SEM)

Reference:

- CUHK STAT5020 - Topics in multivariate analysis (2025 Spring), by Xin Yuan SONG.
- Sik-Yum Lee and Xin-Yuan Song - Basic and advanced Bayesian structural equation modeling: With applications in the medical and behavioral sciences [4].

7.1 SEM models

Goal: to examine the relationships among the variables of interest.
Approach: group observed variables to form *latent variables* (a combination of several observed variables).

- Reduce the number of variables compared to direct regression.
- As highly correlated observed variables are grouped into latent variables, the problem induced by multicollinearity is alleviated.
- It gives better assessments on the interrelationships of latent constructs.

Linear SEMs

Definition 7.1.1. Assume that the observed variables $\boldsymbol{y}_p \sim_{\text{iid}} \mathrm{N}_p$ with mean $\boldsymbol{\mu}_p$. Let $\boldsymbol{\omega}_q$ be latent variables. $\boldsymbol{\omega}_q = (\eta_{q_1}^\top, \boldsymbol{\xi}_{q_2}^\top)^\top$, where η_{q_1} is the key outcome latent variables, and $\boldsymbol{\xi}_{q_2}$ is the explanatory latent variables. Define

$$\begin{aligned} (\textbf{Measurement equation } (\text{MeaEq})) \qquad & \boldsymbol{y}_p = \boldsymbol{\mu}_p + \boldsymbol{\Lambda}_{p\times q}\boldsymbol{\omega}_q + \boldsymbol{\epsilon}_p, \\ (\textbf{Structural equation } (\text{StEq})) \qquad & \eta_{q_1} = \boldsymbol{\Gamma}_{q_1\times q_2}\boldsymbol{\xi}_{q_2} + \boldsymbol{\delta}_{q_1}, \end{aligned} \tag{7.1}$$

where $\boldsymbol{\Lambda}$ is the unknown *factor loading matrix*, $\boldsymbol{\Gamma}$ is the unknown matrix of regression coeffcients, and $\boldsymbol{\epsilon}$ and $\boldsymbol{\delta}$ are measurement (residual) errors.

The basic SEM is formulated by two components:

1. (MeaEq) *Confirmatory factor analysis (CFA)* model groups the highly correlated observed variables into latent variables. We know the information/structure of $\boldsymbol{\Lambda}$ (e.g., know y_1, y_2 are only affected by η and y_3, y_4 are affected by ξ).
   - *Exploratory factor analysis (EFA)*: we have $\boldsymbol{y}$ and just know they have lower-dimensional structure, but don't know the structure of $\boldsymbol{\Lambda}$ and q.

   It regresses $\boldsymbol{y}$ on a smaller number of latent variables.
2. (StEq) A regression type model, in which $\boldsymbol{\eta}$ is regressed on $\boldsymbol{\xi}$.

Standard linear SEMs

Assumption 7.1.2. For $i = 1, \ldots, n$,
(A1) $\boldsymbol{\epsilon}_i \sim_{\text{iid}} \mathrm{N}[\mathbf{0}_p, \boldsymbol{\Psi}_{\boldsymbol{\epsilon}}]$, where $\boldsymbol{\Psi}_{\boldsymbol{\epsilon}} \in \mathbb{R}^{p\times p}$ is diagonal.
(A2) $\boldsymbol{\xi}_i \sim_{\text{iid}} \mathrm{N}[\mathbf{0}_{q_2}, \boldsymbol{\Phi}]$, where $\boldsymbol{\Phi}$ is a general (we can evaluate the correlations of the latent variables).
(A3) $\boldsymbol{\delta}_i \sim_{\text{iid}} \mathrm{N}[\mathbf{0}_{q_1}, \boldsymbol{\Psi}_{\boldsymbol{\delta}}]$, where $\boldsymbol{\Psi}_{\boldsymbol{\delta}}$ is diagonal.
(A4) $\boldsymbol{\delta}_i \perp\!\!\!\perp \boldsymbol{\xi}_i$, and $\boldsymbol{\epsilon}_i \perp\!\!\!\perp \boldsymbol{\omega}_i, \boldsymbol{\delta}_i$.

These assumptions imply that

$$\eta_i \sim_{\text{iid}} \mathrm{N}_{q_1}(\mathbf{0}_{q_1}, \boldsymbol{\Gamma\Phi\Gamma}^\top + \boldsymbol{\Psi}_{\boldsymbol{\delta}}), \quad \boldsymbol{\omega}_i \sim_{\text{iid}} \mathrm{N}_q\left(\mathbf{0}_q, \boldsymbol{\Sigma}_{\boldsymbol{\omega}} = \begin{bmatrix} \boldsymbol{\Gamma\Phi\Gamma}^\top + \boldsymbol{\Psi}_{\boldsymbol{\delta}} & \boldsymbol{\Gamma\Phi} \\ \boldsymbol{\Phi\Gamma}^\top & \boldsymbol{\Phi} \end{bmatrix}\right), \quad \boldsymbol{y}_i \sim_{\text{iid}} \mathrm{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}(\boldsymbol{\theta}) = \boldsymbol{\Lambda\Sigma}_{\boldsymbol{\omega}}\boldsymbol{\Lambda}^\top + \boldsymbol{\Psi}_{\boldsymbol{\epsilon}}).$$

Identifiability issue: The measurement equation is identfied if $\forall \boldsymbol{\theta}_1, \boldsymbol{\theta}_2$, $\text{MeaEq}(\boldsymbol{\theta}_1) = \text{MeaEq}(\boldsymbol{\theta}_2)$ implies $\boldsymbol{\theta}_1 = \boldsymbol{\theta}_2$. The structural equation is identfied if $\forall \boldsymbol{\theta}_1, \boldsymbol{\theta}_2$, $\text{StEq}(\boldsymbol{\theta}_1) = \text{StEq}(\boldsymbol{\theta}_2)$ implies $\boldsymbol{\theta}_1 = \boldsymbol{\theta}_2$. The SEM is identified if both of its MeaEq and StEq are identfied. For example, consider (7.1) MeaEq:

$$\boldsymbol{y} = \boldsymbol{\mu} + (\boldsymbol{\Lambda M})(\boldsymbol{M}^{-1}\boldsymbol{\omega}) + \boldsymbol{\epsilon} = \boldsymbol{\mu} + \boldsymbol{\Lambda}^*\boldsymbol{\omega}^* + \boldsymbol{\epsilon}, \quad \boldsymbol{\omega}^* \sim \mathrm{N}(\boldsymbol{0}_q, \boldsymbol{M}^{-1}\mathbb{C}\mathrm{ov}(\boldsymbol{\omega})\boldsymbol{M}^{-\top}).$$

We have to impose restrictions on $\boldsymbol{\Lambda}$ and/or $\mathbb{C}\mathrm{ov}(\boldsymbol{\omega})$ such that the only nonsingular $\boldsymbol{M} = I_q$.

- A simple and common method is using a $\boldsymbol{\Lambda}$ with the *non-overlapping structure*, e.g.,

$$\boldsymbol{\Lambda}^\top = \begin{bmatrix} 1 & \lambda_{21} & \lambda_{31} & \lambda_{41} & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & \lambda_{62} & \lambda_{72} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & \lambda_{93} & \lambda_{10,3} \end{bmatrix}$$

  where 1's are fixed to introduce a scale to latent variables.
- We can also allow λ_{11}, λ_{52}, and/or λ_{83} to be unknowns, and fix $\mathrm{diag}(\mathbb{C}\mathrm{ov}(\boldsymbol{\omega})) = I_q$, such that $\mathbb{C}\mathrm{ov}(\boldsymbol{\omega})$ is a correlation matrix. But it is not convenient for identifying an SEM with StEq, and induces complication in the Bayesian analysis.

Example 7.1.3. Study the kidney disease of type 2 diabetic patients. We observe: plasma creatine (PCr), urinary albumin creatinine ratio (ACR), systolic blood pressure (SBP), diastolic blood pressure (DBP), body mass index (BMI), waist hip ratio (WHR), glycated hemoglobin (HbAlc), fasting plasma glucose (FPG). Group

- {PCr, ACR}: 'kidney disease (KD)'
- {SBP, DBP}: 'blood pressure (BP)'
- {BMI, WHR}: 'obesity (OB)'
- {HbA1c, FPG}: 'glycemic control (GC)'

$\boldsymbol{y} = (\mathrm{PCr}, \mathrm{ACR}, \mathrm{SBP}, \mathrm{DBP}, \mathrm{BMI}, \mathrm{WHR})^\top$, $\boldsymbol{\omega} = (\mathrm{KD}, \mathrm{BP}, \mathrm{OB})^\top$, $\boldsymbol{\eta} = \mathrm{KD}$, $\boldsymbol{\xi} = (\mathrm{BP}, \mathrm{OB})^\top$, $p = 6$, $q = 3, q_1 = 1, q_2 = 2$. Then

$$\text{(MeaEq)} \qquad \begin{bmatrix} \mathrm{PCr} \\ \mathrm{ACR} \\ \mathrm{SBP} \\ \mathrm{DBP} \\ \mathrm{BMI} \\ \mathrm{WHR} \end{bmatrix} = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \\ \mu_4 \\ \mu_5 \\ \mu_6 \end{bmatrix} + \begin{bmatrix} \lambda_{11} & 0 & 0 \\ \lambda_{21} & 0 & 0 \\ 0 & \lambda_{32} & 0 \\ 0 & \lambda_{42} & 0 \\ 0 & 0 & \lambda_{53} \\ 0 & 0 & \lambda_{63} \end{bmatrix} \begin{bmatrix} \mathrm{KD} \\ \mathrm{BP} \\ \mathrm{OB} \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \\ \epsilon_4 \\ \epsilon_5 \\ \epsilon_6 \end{bmatrix}.$$

We know that KD is only linked with PCr and ACR.

$$\text{(StEq)} \qquad \mathrm{KD} = \gamma_1 \mathrm{BP} + \gamma_2 \mathrm{OB} + \delta.$$

Suppose we wish to study the effects of $\boldsymbol{\xi}$ on $\boldsymbol{\eta} = (\mathrm{KD}, \eta_A)^\top$, $q_1 = 2$,

$$\begin{pmatrix} \mathrm{KD} \\ \eta_A \end{pmatrix} = \begin{pmatrix} 0 & 0 \\ \pi & 0 \end{pmatrix} \begin{pmatrix} \mathrm{KD} \\ \eta_A \end{pmatrix} + \begin{pmatrix} \gamma_1 & \gamma_2 \\ \gamma_3 & \gamma_4 \end{pmatrix} \begin{pmatrix} \mathrm{BP} \\ \mathrm{OB} \end{pmatrix} + \begin{pmatrix} \delta \\ \delta_A \end{pmatrix}.$$

More SEM models

Definition 7.1.4. Extensions of StEq:

- Basic StEq: $\boldsymbol{\eta}_{q_1} = \boldsymbol{\Gamma}_{q_1 \times q_2} \boldsymbol{\xi}_{q_2} + \boldsymbol{\delta}_{q_1}$.
- An extension:

$$\boldsymbol{\eta}_{q_1} = \boldsymbol{\Pi}_{q_1 \times q_1} \boldsymbol{\eta}_{q_1} + \boldsymbol{\Gamma}_{q_1 \times q_2} \boldsymbol{\xi}_{q_2} + \boldsymbol{\delta}_{q_1},$$

  where $\boldsymbol{\Pi}$ is a matrix of unknown coeffcients such that $I_{q_1} - \boldsymbol{\Pi}$ is nonsingular and the diagonal elements of $\boldsymbol{\Pi}$ are zero.
- Add fixed known covariates $\boldsymbol{d}$:

$$\boldsymbol{\eta}_{q_1} = \boldsymbol{B}_{q_1 \times r_2} \boldsymbol{d}_{r_2} + \boldsymbol{\Pi}_{q_1 \times q_1} \boldsymbol{\eta}_{q_1} + \boldsymbol{\Gamma}_{q_1 \times q_2} \boldsymbol{\xi}_{q_2} + \boldsymbol{\delta}_{q_1}.$$

- Add nonlinear structure:

$$\boldsymbol{\eta}_{q_1} = \boldsymbol{B}_{q_1 \times r_2} \boldsymbol{d}_{r_2} + \boldsymbol{\Pi}_{q_1 \times q_1} \boldsymbol{\eta}_{q_1} + \boldsymbol{\Gamma}_{q_1 \times t} \boldsymbol{F}_t(\boldsymbol{\xi}_{q_2}) + \boldsymbol{\delta}_{q_1}, \tag{7.2}$$

  where $\boldsymbol{F}(\boldsymbol{\xi}) = (f_1(\boldsymbol{\xi}), \ldots, f_t(\boldsymbol{\xi}))^\top$ with nonzero, known, and linearly independent differentiable functions $f_1, \ldots, f_t$, $t \geq q_2$. To identity StEq, structures like $\boldsymbol{F}_1(\boldsymbol{\xi}) = (\xi_1, \xi_2, \xi_1^2, \xi_1^2)^\top$ and $\boldsymbol{F}_2(\boldsymbol{\xi}) = (\xi_1, \xi_2, \xi_1\xi_2, 0)^\top$ are not allowed.
- Combine $\boldsymbol{d}$ and $\boldsymbol{F}$:

$$\boldsymbol{\eta}_{q_1} = \boldsymbol{\Pi}_{q_1 \times q_1} \boldsymbol{\eta}_{q_1} + \boldsymbol{\Lambda}_{\boldsymbol{\omega}, q_1 \times t} \boldsymbol{G}_t(\boldsymbol{d}, \boldsymbol{\xi}) + \boldsymbol{\delta}_{q_1}, \tag{7.3}$$

  where $\boldsymbol{G}(\boldsymbol{d\xi}) = (g_1(\boldsymbol{d}, \boldsymbol{\xi}), \ldots, g_t(\boldsymbol{d}, \boldsymbol{\xi}))^\top$ is a vector-valued function with nonzero, known, and linearly independent differentiable functions. (7.2) can be obtained by letting $\boldsymbol{\Lambda}_{\boldsymbol{\omega}} = (\boldsymbol{B}, \boldsymbol{\Gamma})$ and $\boldsymbol{G}(\boldsymbol{d}, \boldsymbol{\xi}) = (\boldsymbol{d}^\top, \boldsymbol{F}(\boldsymbol{\xi})^\top)^\top$.

An Extension of MeaEq:

- Add fixed known covariates $\boldsymbol{c}$ (If $\boldsymbol{\mu}_p$ is included, then $\boldsymbol{c} = [1, \boldsymbol{c}_2]^\top$.)

$$\boldsymbol{y}_p = \boldsymbol{A}_{p \times r_1} \boldsymbol{c}_{r_1} + \boldsymbol{\Lambda}_{p \times q} \boldsymbol{\omega}_q + \boldsymbol{\epsilon}_p. \tag{7.4}$$

Let $\boldsymbol{\Lambda}_k^\top$ be the kth row of $\boldsymbol{\Lambda}$, and $\boldsymbol{\Lambda}_k^\top = (\boldsymbol{\Lambda}_{k\boldsymbol{\eta}}^\top, \boldsymbol{\Lambda}_{k\boldsymbol{\xi}}^\top)$ be a partition correspondings to the partition of $\boldsymbol{\omega} = (\boldsymbol{\eta}^\top, \boldsymbol{\xi}^\top)^\top$.

For StEq (7.2) with $r_2 = 0$ and MeaEq (7.1),

$$\mathbb{E}(\boldsymbol{\xi}) = \mathbf{0}_{q_2}, \quad \mathbb{E}(\boldsymbol{\eta}) = [(I_{q_1} - \boldsymbol{\Pi})^{-1}\boldsymbol{\Gamma}]\mathbb{E}(\boldsymbol{F}(\boldsymbol{\xi})),$$
$$\mathbb{E}(y_k) = \mu_k + \boldsymbol{\Lambda}_{k\boldsymbol{\eta}}^\top[(I_{q_1} - \boldsymbol{\Pi})^{-1}\boldsymbol{\Gamma}]\mathbb{E}(\boldsymbol{F}(\boldsymbol{\xi})).$$

For StEq (7.3) and MeaEq (7.4), let $\boldsymbol{A}_k^\top$ be the kth row of $\boldsymbol{A}$,

$$\mathbb{E}(y_k) = \boldsymbol{A}_k^\top \boldsymbol{c} + \boldsymbol{\Lambda}_{k\boldsymbol{\eta}}^\top \mathbb{E}(\boldsymbol{\eta}) = \boldsymbol{A}_k^\top \boldsymbol{c} + \boldsymbol{\Lambda}_{k\boldsymbol{\eta}}^\top[(I_{q_1} - \boldsymbol{\Pi})^{-1}\boldsymbol{\Lambda}_{\boldsymbol{\omega}}]\mathbb{E}(\boldsymbol{G}(\boldsymbol{d}, \boldsymbol{\xi})).$$

Issues of developing a comprehensive SEM:

1. Make sure the sample size of $\boldsymbol{y}$ is large enough to achieve accurate statistical results.
2. If the size of the proposed SEM and the number of parameters are large, we may encounter diffculties in achieving convergence of the related computing algorithm for obtaining statistical results.
3. So far, the most general SEM is MeaEq (7.4) and StEq (7.3):
$$\boldsymbol{y} = \boldsymbol{A}\boldsymbol{c} + \boldsymbol{\Lambda}\boldsymbol{\omega} + \boldsymbol{\epsilon}, \qquad \boldsymbol{\eta} = \boldsymbol{\Pi}\boldsymbol{\eta} + \boldsymbol{\Lambda}_{\boldsymbol{\omega}}\boldsymbol{G}(\boldsymbol{d}, \boldsymbol{\xi}) + \boldsymbol{\delta}.$$
It has limitations. As $\boldsymbol{G}(\boldsymbol{d}, \boldsymbol{\xi})$ does not involve any η_k in $\boldsymbol{\eta}$, nonlinear terms related to η_k cannot be used to predict the other $\eta_{k'}$'s, i.e., η_k cannot be accommodated in $\boldsymbol{G}(\boldsymbol{d}, \boldsymbol{\xi})$ and nonlinear effects of η_k on $\eta_{k'}$ cannot be assessed.

7.2 Bayesian methods for estimating SEM

Motivation: A traditional method is the *covariance structure approach*, which focuses on fitting the covariance structure under the proposed model to the sample covariance matrix $\boldsymbol{S}$.

- In many complex situations, deriving the explicit covariance structure $\Sigma = \mathbb{C}\mathrm{ov}(\boldsymbol{y})$ or obtaining an appropriate $\boldsymbol{S}$ is difficult (e.g., when $\exists$ missing data).

Advarntages of Bayesian estimates (BE): $\log \mathbb{P}(\boldsymbol{\theta} \mid \boldsymbol{Y}) = \log \mathbb{P}(\boldsymbol{Y} \mid \boldsymbol{\theta}) + \log \mathbb{P}(\boldsymbol{\theta}) + \text{const}$

1. Methods are based on the first moment properties of $\boldsymbol{y}$ which are simpler than the second moment properties of $\boldsymbol{S}$. Hence, it has potential to be applied to more complex situations.
2. It estimates $\boldsymbol{\omega}$ directly, which cannot be obtained with classical methods.
3. As $n \to \infty$, $\log \mathbb{P}(\boldsymbol{Y} \mid \boldsymbol{\theta})$ could dominate $\log \mathbb{P}(\boldsymbol{\theta})$, hence the BE have the same optimal properties as the MLE.
4. It allows the use of genuine prior information for producing better results. With small or moderate sample sizes, $\mathbb{P}(\boldsymbol{\theta})$ plays a more substantial role in BE than $\mathbb{P}(\boldsymbol{Y} \mid \boldsymbol{\theta})$, and is useful for achieving better results.
5. It provides more easily assessable statistics for goodness-of-fit and model comparison, and also other useful statistics such as the posterior mean and percentiles. (Don't need to derive the asymptotic distributions.)
6. It can give more reliable results for small samples.

7.2.1 Priors for SEM

Informative prior:

Conjugate prior for SEMs

Assumption 7.2.1. In developing the Bayesian methods for analyzing SEMs, we usually assign fixed known values to the hyperparameters in the conjugate prior distributions. Consider

$$\boldsymbol{y}_i = \boldsymbol{\mu} + \boldsymbol{\Lambda}\boldsymbol{\omega}_i + \boldsymbol{\epsilon}_i,$$
$$\boldsymbol{\eta}_i = \boldsymbol{B}\boldsymbol{d}_i + \boldsymbol{\Pi}\boldsymbol{\eta}_i + \boldsymbol{\Gamma}\boldsymbol{F}(\boldsymbol{\xi}_i) + \boldsymbol{\delta}_i = \boldsymbol{\Lambda}_{\boldsymbol{\omega}}\boldsymbol{G}(\boldsymbol{\omega}_i) + \boldsymbol{\delta}_i,$$

where $\boldsymbol{\Lambda}_{\boldsymbol{\omega}} = (\boldsymbol{B}, \boldsymbol{\Pi}, \boldsymbol{\Gamma}) \in \mathbb{R}^{q_1 \times (r_2+q_1+t)}$, and $\boldsymbol{G}(\boldsymbol{\omega}_i) = (\boldsymbol{d}_i^\top, \boldsymbol{\eta}_i^\top, \boldsymbol{F}(\boldsymbol{\xi}_i)^\top)^\top \in \mathbb{R}^{r_2+q_1+t}$. Assumption 7.1.2 is satisfied. $\boldsymbol{\xi}_i \sim_{\text{iid}} \mathrm{N}[\mathbf{0}_{q_2}, \boldsymbol{\Phi}]$, $\boldsymbol{\epsilon}_i \sim_{\text{iid}} \mathrm{N}[\mathbf{0}_p, \boldsymbol{\Psi}_{\boldsymbol{\epsilon}} = \mathrm{diag}(\psi_{\boldsymbol{\epsilon}k})]$,and $\boldsymbol{\delta}_i \sim_{\text{iid}} \mathrm{N}[\mathbf{0}_{q_1}, \boldsymbol{\Psi}_{\boldsymbol{\delta}} = \mathrm{diag}(\psi_{\boldsymbol{\delta}k})]$.

- Prior (conjugate) for $\boldsymbol{\theta}_{\boldsymbol{y}} = (\boldsymbol{\mu}, \boldsymbol{\Lambda}, \boldsymbol{\Psi}_{\boldsymbol{\epsilon}})$: let $\boldsymbol{\Lambda}_k^\top$ be the kth row of $\boldsymbol{\Lambda}$,
$$\psi_{\boldsymbol{\epsilon}k} \sim \mathrm{IG}(\alpha_{0\boldsymbol{\epsilon}k}, \beta_{0\boldsymbol{\epsilon}k}), \quad [\boldsymbol{\Lambda}_k \mid \psi_{\boldsymbol{\epsilon}k}] \sim \mathrm{N}_q(\boldsymbol{\Lambda}_{0k}, \psi_{\boldsymbol{\epsilon}k}\boldsymbol{H}_{0\boldsymbol{y}k}),\ k = 1, \ldots, p,$$
$$\boldsymbol{\mu} \sim \mathrm{N}_p(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0).$$
- Prior (conjugate) for $\boldsymbol{\theta}_{\boldsymbol{\omega}} = (\boldsymbol{\Lambda}_{\boldsymbol{\omega}}, \boldsymbol{\Psi}_{\boldsymbol{\delta}}, \boldsymbol{\Phi})$: let $\boldsymbol{\Lambda}_{\boldsymbol{\omega}k}^\top$ be the kth row of $\boldsymbol{\Lambda}_{\boldsymbol{\omega}}$, $\boldsymbol{\Psi}_{\boldsymbol{\delta}}$,
$$\boldsymbol{\Phi} \sim \mathrm{IW}_{q_2}(\boldsymbol{R}_0^{-1}, \rho_0),\ \text{or } \boldsymbol{\Phi}^{-1} \sim \mathrm{W}_{q_2}(\boldsymbol{R}_0, \rho_0)$$
$$\psi_{\boldsymbol{\delta}k} \sim \mathrm{IG}(\alpha_{0\boldsymbol{\delta}k}, \beta_{0\boldsymbol{\delta}k}), \quad [\boldsymbol{\Lambda}_{\boldsymbol{\omega}k} \mid \psi_{\boldsymbol{\delta}k}] \sim \mathrm{N}_{r_2+q_1+t}(\boldsymbol{\Lambda}_{0\boldsymbol{\omega}k}, \psi_{\boldsymbol{\delta}k}\boldsymbol{H}_{0\boldsymbol{\omega}k}),\ k = 1, \ldots, q_1.$$
- Assume the prior $\boldsymbol{\theta}_{\boldsymbol{y}} \perp\!\!\!\perp \boldsymbol{\theta}_{\boldsymbol{\omega}}$.

Hyperparameter selection: If we have good prior information about a parameter, select the prior distribution with a small variance. E.g.,

- if $\mathbf{\Lambda}_k \approx \mathbf{\Lambda}_{0k}$, then $\boldsymbol{H}_{0yk} = 0.5I_q$. If not, select the prior with a larger variance;
- since $\epsilon_{ik} \sim \mathrm{N}(0, \psi_{\boldsymbol{\epsilon}k})$, if the variation is small, $\psi_{\boldsymbol{\epsilon}k}$ is small, then choose small $\mathbb{E}(\psi_{\boldsymbol{\epsilon}k}) = \beta_{0\boldsymbol{\epsilon}k}/(\alpha_{0\boldsymbol{\epsilon}k} - 1)$ and $\mathbb{V}\mathrm{ar}(\psi_{\boldsymbol{\epsilon}k}) = \beta_{0\boldsymbol{\epsilon}k}^2/\{(\alpha_{0\boldsymbol{\epsilon}k} - 1)^2(\alpha_{0\boldsymbol{\epsilon}k} - 2)\}$;
- if $\mathbf{\Phi} \approx \mathbf{\Phi}_0$, since $\mathbb{E}(\mathbf{\Phi}) = \boldsymbol{R}_0^{-1}/(\rho_0 - q_2 - 1)$, choose $\boldsymbol{R}_0^{-1} = (\rho_0 - q_2 - 1)\mathbf{\Phi}_0$.

Noninformative prior (Jeffrey): If information is not available and the sample size is small,

$$\mathbb{P}(\mathbf{\Lambda}, \mathbf{\Psi}_{\boldsymbol{\epsilon}}) \propto \mathbb{P}(\psi_{\boldsymbol{\epsilon}1}, \cdots, \psi_{\boldsymbol{\epsilon}p}) \propto \prod_{k=1}^{p} \psi_{\boldsymbol{\epsilon}k}^{-1}, \quad \mathbb{P}(\mathbf{\Lambda}_{\boldsymbol{\omega}}, \mathbf{\Psi}_{\boldsymbol{\delta}}) \propto \mathbb{P}(\psi_{\boldsymbol{\delta}1}, \cdots, \psi_{\boldsymbol{\delta}q_1}) \propto \prod_{k=1}^{q_1} \psi_{\boldsymbol{\delta}k}^{-1},$$
$$\mathbb{P}(\mathbf{\Phi}) \propto |\mathbf{\Phi}|^{-(q_2+1)/2}.$$

If the sample size is large, can use a portion of the data to estimate $\mathbf{\Lambda}_{0k}$, $\mathbf{\Lambda}_{0\boldsymbol{\omega}k}$ and $\mathbf{\Phi}_0$ with noninformative priors. If the sample size is moderate, can use the same data twice.

7.2.2 Bayesian estimation using MCMC

Model: Linear SEM with fixed covariates without intercept:

$$\begin{aligned} \boldsymbol{y}_i &= \mathbf{\Lambda}\boldsymbol{\omega}_i + \boldsymbol{\epsilon}_i, \\ \boldsymbol{\eta}_i &= \boldsymbol{B}\boldsymbol{d}_i + \mathbf{\Pi}\boldsymbol{\eta}_i + \mathbf{\Gamma}\boldsymbol{\xi}_i + \boldsymbol{\delta}_i = \mathbf{\Lambda}_{\boldsymbol{\omega}}\boldsymbol{v}_i + \boldsymbol{\delta}_i, \end{aligned}$$

where $\mathbf{\Lambda}_{\boldsymbol{\omega}} = (\boldsymbol{B}, \mathbf{\Pi}, \mathbf{\Gamma}) \in \mathbb{R}^{q_1 \times (r_2+q_1+q_2)}$, and $\boldsymbol{v}_i = (\boldsymbol{d}_i^\top, \boldsymbol{\eta}_i^\top, \boldsymbol{\xi}_i^\top)^\top \in \mathbb{R}^{r_2+q_1+q_2}$. That is, assume $\boldsymbol{\mu} = \mathbf{0}_p$ and $\boldsymbol{F}(\boldsymbol{\xi}_i) - \boldsymbol{\xi}_i$.

Denote data $\boldsymbol{Y} = (\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n) = (\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_p)^\top \in \mathbb{R}^{p \times n}$, $\boldsymbol{V} = (\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n) = (\boldsymbol{V}_1, \ldots, \boldsymbol{V}_{r_2+q_1+q_2})^\top$, $\mathbf{\Xi}_k = (\eta_{1k}, \cdots, \eta_{nk})^\top$ for $k = 1, \ldots, q_1$, matrix of latent variables $\mathbf{\Omega} = (\boldsymbol{\omega}_1, \ldots, \boldsymbol{\omega}_n) \in \mathbb{R}^{q \times n}$, $\mathbf{\Omega}_1 = (\boldsymbol{\eta}_1, \ldots, \boldsymbol{\eta}_n)$, $\mathbf{\Omega}_2 = (\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_n)$, and

$$\boldsymbol{\theta} = (\mathbf{\Lambda}, \boldsymbol{B}, \mathbf{\Pi}, \mathbf{\Gamma}, \mathbf{\Phi}, \mathbf{\Psi}_{\boldsymbol{\epsilon}}, \mathbf{\Psi}_{\boldsymbol{\delta}}) = (\underbrace{\mathbf{\Lambda}, \mathbf{\Psi}_{\boldsymbol{\epsilon}}}_{\boldsymbol{\theta}_{\boldsymbol{y}}}, \underbrace{\mathbf{\Lambda}_{\boldsymbol{\omega}}, \mathbf{\Phi}, \mathbf{\Psi}_{\boldsymbol{\delta}}}_{\boldsymbol{\theta}_{\boldsymbol{\omega}}}).$$

Proposition 7.2.2. *The above model has the following posterior distributions:*

1. *Conditional distribution* $\mathbb{P}(\mathbf{\Omega} \mid \boldsymbol{Y}, \boldsymbol{\theta}) = \prod_{i=1}^n \mathbb{P}(\boldsymbol{\omega}_i \mid \boldsymbol{y}_i, \boldsymbol{\theta}) \propto \prod_{i=1}^n \mathbb{P}(\boldsymbol{\omega}_i \mid \boldsymbol{\theta})\mathbb{P}(\boldsymbol{y}_i \mid \boldsymbol{\omega}_i, \boldsymbol{\theta})$, *where*

$$[\boldsymbol{\omega}_i \mid \boldsymbol{\theta}] \sim \mathrm{N}_q(\boldsymbol{\mu}_{\boldsymbol{\omega}_i}, \mathbf{\Sigma}_{\boldsymbol{\omega}}), \quad [\boldsymbol{y}_i \mid \boldsymbol{\omega}_i, \boldsymbol{\theta}] \sim \mathrm{N}_p(\mathbf{\Lambda}\boldsymbol{\omega}_i, \mathbf{\Psi}_{\boldsymbol{\epsilon}}),$$
$$[\boldsymbol{\omega}_i \mid \boldsymbol{y}_i, \boldsymbol{\theta}] \sim \mathrm{N}_q(\mathbf{\Sigma}^{*-1}(\mathbf{\Sigma}_{\boldsymbol{\omega}}^{-1}\boldsymbol{\mu}_{\boldsymbol{\omega}_i} + \mathbf{\Lambda}^\top\mathbf{\Psi}_{\boldsymbol{\epsilon}}^{-1}\boldsymbol{y}_i), \mathbf{\Sigma}^{*-1})$$

   where

$$\mathbf{\Pi}_0 = I_{q_1} - \mathbf{\Pi}, \ \boldsymbol{\mu}_{\boldsymbol{\omega}_i} = \begin{pmatrix} \mathbf{\Pi}_0^{-1}\boldsymbol{B}\boldsymbol{d}_i \\ \mathbf{0}_{q_2} \end{pmatrix}, \ \mathbf{\Sigma}_{\boldsymbol{\omega}} = \begin{bmatrix} \mathbf{\Pi}_0^{-1}(\mathbf{\Gamma}\mathbf{\Phi}\mathbf{\Gamma}^\top + \mathbf{\Psi}_{\boldsymbol{\delta}})\mathbf{\Pi}_0^{-\top} & \mathbf{\Pi}_0^{-1}\mathbf{\Gamma}\mathbf{\Phi} \\ \mathbf{\Phi}\mathbf{\Gamma}^\top\mathbf{\Pi}_0^{-\top} & \mathbf{\Phi} \end{bmatrix},$$
$$\mathbf{\Sigma}^* = \mathbf{\Sigma}_{\boldsymbol{\omega}}^{-1} + \mathbf{\Lambda}^\top\mathbf{\Psi}_{\boldsymbol{\epsilon}}^{-1}\mathbf{\Lambda}.$$

2. *Assume all elements of* $\mathbf{\Lambda}_k$ *and* $\mathbf{\Lambda}_{\boldsymbol{\omega}}$ *are unknown, conditional distribution* $\mathbb{P}(\boldsymbol{\theta} \mid \boldsymbol{Y}, \mathbf{\Omega}) = \mathbb{P}(\boldsymbol{\theta}_{\boldsymbol{y}} \mid \boldsymbol{Y}, \mathbf{\Omega})\mathbb{P}(\boldsymbol{\theta}_{\boldsymbol{\omega}} \mid \boldsymbol{Y}, \mathbf{\Omega})$, *where we can show* $[\mathbf{\Lambda}_k, \psi_{\boldsymbol{\epsilon}k} \mid \boldsymbol{Y}, \mathbf{\Omega}] \perp\!\!\!\perp$ *and* $[\mathbf{\Lambda}_{\boldsymbol{\omega}k}, \psi_{\boldsymbol{\delta}k} \mid \mathbf{\Omega}] \perp\!\!\!\perp$, *and*

$$\mathbb{P}(\boldsymbol{\theta}_{\boldsymbol{y}} \mid \boldsymbol{Y}, \mathbf{\Omega}) \propto \prod_{k=1}^{p} \mathbb{P}(\mathbf{\Lambda}_k, \psi_{\boldsymbol{\epsilon}k} \mid \boldsymbol{Y}, \mathbf{\Omega}),$$
$$\mathbb{P}(\boldsymbol{\theta}_{\boldsymbol{\omega}} \mid \boldsymbol{Y}, \mathbf{\Omega}) \propto \left[\prod_{k=1}^{q_1} \mathbb{P}(\mathbf{\Lambda}_{\boldsymbol{\omega}k}, \psi_{\boldsymbol{\delta}k} \mid \mathbf{\Omega})\right] \mathbb{P}(\mathbf{\Phi} \mid \mathbf{\Omega}_2)$$

   where

$$\begin{aligned} \mathbb{P}(\mathbf{\Lambda}_k, \psi_{\boldsymbol{\epsilon}k}^{-1} \mid \boldsymbol{Y}, \mathbf{\Omega}) &\propto \mathrm{N}_q(\boldsymbol{a}_k, \psi_{\boldsymbol{\epsilon}k}\boldsymbol{A}_k) \cdot \mathrm{Ga}(n/2 + \alpha_{0\boldsymbol{\epsilon}k}, \beta_{\boldsymbol{\epsilon}k}), \\ \mathbb{P}(\mathbf{\Lambda}_{\boldsymbol{\omega}k}, \psi_{\boldsymbol{\delta}k}^{-1} \mid \mathbf{\Omega}) &\propto \mathrm{N}_{r_2+q_1+q_2}(\boldsymbol{a}_{\boldsymbol{\omega}k}, \psi_{\boldsymbol{\delta}k}\boldsymbol{A}_{\boldsymbol{\omega}k}) \cdot \mathrm{Ga}(n/2 + \alpha_{0\boldsymbol{\delta}k}, \beta_{\boldsymbol{\delta}k}), \\ [\mathbf{\Phi} \mid \mathbf{\Omega}_2] &\sim \mathrm{IW}_{q_2}[(\mathbf{\Omega}_2\mathbf{\Omega}_2^\top + \boldsymbol{R}_0^{-1}), n + \rho_0]. \end{aligned}$$

   where

$$\begin{aligned} \boldsymbol{A}_k &= (\boldsymbol{H}_{0yk}^{-1} + \mathbf{\Omega}\mathbf{\Omega}^\top)^{-1}, \ \boldsymbol{a}_k = \boldsymbol{A}_k(\boldsymbol{H}_{0yk}^{-1}\mathbf{\Lambda}_{0k} + \mathbf{\Omega}\boldsymbol{Y}_k), \\ \beta_{\boldsymbol{\epsilon}k} &= \beta_{0\boldsymbol{\epsilon}k} + \frac{1}{2}(\boldsymbol{Y}_k^\top\boldsymbol{Y}_k - \boldsymbol{a}_k^\top\boldsymbol{A}_k^{-1}\boldsymbol{a}_k + \mathbf{\Lambda}_{0k}^\top\boldsymbol{H}_{0yk}^{-1}\mathbf{\Lambda}_{0k}), \\ \boldsymbol{A}_{\boldsymbol{\omega}k} &= (\boldsymbol{H}_{0\boldsymbol{\omega}k}^{-1} + \boldsymbol{V}_k\boldsymbol{V}_k^\top)^{-1}, \ \boldsymbol{a}_{\boldsymbol{\omega}k} = \boldsymbol{A}_{\boldsymbol{\omega}k}(\boldsymbol{H}_{0\boldsymbol{\omega}k}^{-1}\mathbf{\Lambda}_{0\boldsymbol{\omega}k} + \boldsymbol{V}_k\mathbf{\Xi}_k), \\ \beta_{\boldsymbol{\delta}k} &= \beta_{0\boldsymbol{\delta}k} + \frac{1}{2}(\mathbf{\Xi}_k^\top\mathbf{\Xi}_k - \boldsymbol{a}_{\boldsymbol{\omega}k}^\top\boldsymbol{A}_{\boldsymbol{\omega}k}^{-1}\boldsymbol{a}_{\boldsymbol{\omega}k} + \mathbf{\Lambda}_{0\boldsymbol{\omega}k}^\top\boldsymbol{H}_{0\boldsymbol{\omega}k}^{-1}\mathbf{\Lambda}_{0\boldsymbol{\omega}k}). \end{aligned}$$

Remark 7.2.3. For general nonlinear SEMs with MeaEq (7.4) and StEq (7.3), we can defind $\boldsymbol{u} = [\boldsymbol{c}^\top, \boldsymbol{\omega}^\top]^\top \in \mathbb{R}^{r_1+q}$ and use similar procedure to derive the full conditional distributions. But by the nonlinear structure $\boldsymbol{G}(\boldsymbol{\omega})$, $\mathbb{P}(\mathbf{\Omega} \mid \boldsymbol{Y}, \boldsymbol{\theta})$ may not have closed form like normal, while $\mathbb{P}(\boldsymbol{\theta} \mid \boldsymbol{Y}, \mathbf{\Omega})$ is not affected and keeps normal-Gamma. To handle fixed parameters, see Appendix 3.3 in [4] and Sec 4.3.1 and Appendix 4.3 in [3]. Also see STAT5020 HW1 Q3.

7.3 Bayesian model comparison

7.4 Hierarchical and Multisample Data

Two-level nonlinear SEM with mixed type variables

Definition 7.4.1. Consider a collection of p-variate random vectors $\boldsymbol{u}_{gi}$, $i = 1, \ldots, N_g$, nested within groups $g = 1, \ldots, G$.

$$\begin{aligned} \text{(Within-groups) } \boldsymbol{u}_{gi} &= \boldsymbol{v}_g + \boldsymbol{\Lambda}_{1g}\boldsymbol{\omega}_{1gi} + \boldsymbol{\epsilon}_{1gi}, \quad g = 1, \cdots, G, \quad i = 1, \cdots, N_g, \\ \text{(Between-groups) } \boldsymbol{v}_g &= \boldsymbol{\mu} + \boldsymbol{\Lambda}_2\boldsymbol{\omega}_{2g} + \boldsymbol{\epsilon}_{2g}, \quad g = 1, \ldots, G \end{aligned}$$

where $\boldsymbol{\Lambda}_{1g} \in \mathbb{R}^{p \times q_1}$, $\boldsymbol{\epsilon}_{1gi} \in \mathbb{R}^{q_1}$, $\boldsymbol{\epsilon}_{1gi} \in \mathbb{R}^{p} \sim \mathrm{N}(\mathbf{0}, \boldsymbol{\Psi}_{1g})$ independent of $\boldsymbol{\omega}_{1gi}$, where $\boldsymbol{\Psi}_{1g}$ is diagonal. Then

$$\mathbf{u}_{\mathrm{gi}} = \mu + \boldsymbol{\Lambda}_2\omega_{\mathrm{2g}} + \epsilon_{\mathrm{2g}} + \boldsymbol{\Lambda}_{\mathrm{1g}}\omega_{\mathrm{1gi}} + \epsilon_{\mathrm{1gi}}.$$

The SEM is

$$\begin{aligned} \eta_{1gi} &= \Pi_{1g}\eta_{1gi} + \Gamma_{1g}\mathbf{F}_1(\xi_{1gi}) + \delta_{1gi}, \\ \eta_{2g} &= \Pi_2\eta_{2g} + \Gamma_2\mathbf{F}_2(\xi_{2g}) + \delta_{2g}, \end{aligned}$$

Bibliography

[1] G. Casella and R. L. Berger. *Statistical inference*, volume 2. Duxbury Pacific Grove, CA, 2002. 2, 2.1

[2] R. Christensen et al. *Plane answers to complex questions*, volume 35. Springer, 2002. 3

[3] S.-Y. Lee. *Structural equation modeling: A Bayesian approach*. John Wiley & Sons, 2007. 7.2.3

[4] S.-Y. Lee and X.-Y. Song. *Basic and advanced Bayesian structural equation modeling: With applications in the medical and behavioral sciences*. John Wiley & Sons, 2012. 7, 7.2.3

[5] R. J. Muirhead. *Aspects of multivariate statistical theory*. John Wiley & Sons, 1982. 3, 3.1, 3.1.1, 3.2.1