L07 Model with normal distributions

1. Maximum likelihood Estimators

   (1) Model
   Model $Y = X\beta + \epsilon$ with $\epsilon \sim N(0, \sigma^2\Sigma)$ allows us to approach the estimating β and σ^2 via maximum likelihood methods.
   The likelihood funaction is the joint pdf of Y treated as a function of β and σ^2.

$$\begin{aligned} L(\beta, \sigma^2) &= \tfrac{1}{(2\pi)^{n/2}|\sigma^2\Sigma|^{1/2}} \exp\left[\tfrac{-1}{2}(Y - X\beta)'(\sigma^2\Sigma)^{-1}(Y - X\beta)\right] \\ &= \tfrac{1}{(2\pi\sigma^2)^{n/2}|\Sigma|^{1/2}} \exp\left[\tfrac{-1}{2\sigma^2}(Y - X\beta)'\Sigma^{-1}(Y - X\beta)\right] \\ &= \tfrac{1}{(2\pi\sigma^2)^{n/2}|\Sigma|^{1/2}} \exp\left(\tfrac{1}{-2\sigma^2}\|Y - X\beta\|^2_{\Sigma^{-1}}\right). \end{aligned}$$

   (2) MLE for β
   Since Y was observed, it is reasonable to believe that β and σ^2 have the values to make the pdf high at Y. Thus we call $\widehat{\beta}$ and $\widehat{\sigma}^2$ maximum likelihood estimators (MLEs) if $L(\beta, \sigma^2) \le L(\widehat{\beta}, \widehat{\sigma}^2)$ for all β and σ^2.
   Let MLE(β) be the collection of all MLEs for β. Then

$$\text{MLE}(\beta) = \text{GLSE}_{V^{-1}}(\beta) = \left(\Sigma^{-1/2}X\right)^+ \Sigma^{-1/2}Y + \mathcal{N}(X).$$

   Proof. By the form of $L(\beta, \sigma^2)$ in (1),

$$\begin{aligned} L(\beta, \sigma^2) \le L(\widehat{\beta}, \sigma^2) \text{ for all } \beta \text{ and } \sigma^2 &\iff \|Y - X\beta\|^2_{V^{-1}} \ge \|Y - X\widehat{\beta}\|^2_{V^{-1}} \text{ for all } \beta \\ &\iff \widehat{\beta} \in \text{GLSE}_{V^{-1}}(\beta). \end{aligned}$$

   (3) MLE for σ^2
   With $\widehat{\beta} \in \text{MLE}(\beta)$, let $\text{SSE}_{\Sigma^{-1}} = \|Y - X\widehat{\beta}\|^2_{\Sigma^{-1}}$. Then $\dfrac{\text{SSE}_{\Sigma^{-1}}}{n}$ is MLE for σ^2.
   Proof. $L(\widehat{\beta}, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{n/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2\sigma^2}\text{SSE}_{\Sigma^{-1}}\right)$ is a function of σ^2.
   By conventional first derivative test or second derivative test, one can see that this function is maximized at $\sigma^2 = \dfrac{\text{SSE}_{\Sigma^{-1}}}{n}$.
   Comment: $L(\widehat{\beta}, \widehat{\sigma}^2) = \left(\frac{n}{2\pi e}\right)^{n/2} |\Sigma|^{-1/2} \left(\text{SSE}_{\Sigma^{-1}}\right)^{-n/2}$.

2. MVUE

   (1) Cramer-Rao lower bound
   Suppose $Y \in R^n$ has pdf $f(y, \theta)$, $\theta \in R^k$. With respect to $\theta \in R^k$, $\nabla \ln f(Y, \theta) \in R^k$ is a random vector with variance-covariance matrix $I(\theta) \in R^{k\times k}$ called the information matrix for the pdf $f(y, \theta)$.
   Suppose statistic vector $T(Y) \in R^q$ has mean $E[T(Y)] = g(\theta) \in R^q$. It can be shown (in Stat771-772 or Stat870-871) that

$$\text{Cov}(T(Y)) \ge \left[\frac{\partial\, g(\theta)}{\partial\, \theta^T}\right] [I(\theta)]^{-1} \left[\frac{\partial\, g(\theta)}{\partial\, \theta^T}\right]'.$$

   This lower bound for Cov($T(Y)$) is called the Cramer-Rao lower bound which is the lowest risk for all UEs for $g(\theta)$.

(2) MVUE

If $\text{Cov}(T(Y))$ reaches the Cramer-Rao lower bound, then it is the best estimator among all UEs for $g(\theta)$. This best estimator is called the minimum variance-covariance unbiased estimator (MVUE).

(3) Theorem

Suppose in model $Y = X\beta + \epsilon$, $\epsilon \sim N(0, \sigma^2\Sigma)$, X has full column rank. Then all $H\beta$ are estimable, and $H\left(\Sigma^{-1/2}X\right)^{+}\Sigma^{-1/2}Y$ is MVUE for $H\beta$.

Proof. For $H \in R^{q\times p}$, $H = [H(X'X)^{-1}X']X$. So $H\beta$ is estimable with $\frac{\partial H\beta}{\partial \beta'} = H$.
For $Y \sim N(X\beta, \sigma^2\Sigma)$, $\ln f(y, \beta) = -\frac{n}{2}\ln 2\pi\sigma^2 - \frac{1}{2}\ln|\Sigma| - \frac{1}{2\sigma^2}(Y - X\beta)'\Sigma^{-1}(Y - X\beta)$.
$\frac{\partial \ln f(y,\beta)}{\partial \beta'} = -\frac{1}{2\sigma^2}(Y - X\beta)'2\Sigma^{-1}(-X)$. So $\nabla \ln f(Y, \beta) = \frac{1}{\sigma^2}X'\Sigma^{-1}(Y - X\beta)$.
Thus $I(\beta) = \frac{1}{\sigma^2}X'\Sigma^{-1}X$. Hence $\text{CRLB}(H\beta) = \sigma^2 H(X'\Sigma^{-1}X)^{-1}H'$.
But $\text{Cov}(H\left(\Sigma^{-1/2}X\right)^{+}\Sigma^{-1/2}Y) = \sigma^2 H(X'\Sigma^{-1}X)^{+}H' = \sigma^2 H(X'\Sigma^{-1}X)^{-1}H'$ which is $\text{CRLB}(H\beta)$. Hence $H\left(\Sigma^{-1/2}X\right)^{+}\Sigma^{-1/2}Y$ is the MVUE for $H\beta$.

3. Sampling distributions

(1) The MVUE for $H\beta$, $H\left(\Sigma^{-1/2}X\right)^{+}\Sigma^{-1/2}Y \sim N\left(H\beta, \sigma^2 H(X'\Sigma^{-1}X)^{+}H'\right)$.

Proof. With $A = H\left(\Sigma^{-1/2}X\right)^{+}\Sigma^{-1/2}$ and $Y \sim N(X\beta, \sigma^2\Sigma)$,
$AY \sim N\left(AX\beta, \sigma^2 A\Sigma A'\right) = N\left(H\beta, \sigma^2 H(X'\Sigma^{-1}X)^{+}H'\right)$.

Comment: Because X has full column rank, $(X'\Sigma^{-1}X)^{+} = (X'\Sigma X)^{-1}$.

Ex1: The distribution of $\widehat{Y} = X\left(\Sigma^{-1/2}X\right)^{+}\Sigma^{-1/2}Y$, the MVUE for $X\beta = E(Y)$, is $\widehat{Y} \sim N\left(X\beta, \sigma^2 X(X'\Sigma^{-1}X)^{-1}X'\right)$.

(2) $\frac{\text{SSE}}{\sigma^2} \sim \chi^2(n - r)$ where $r = \text{rank}(X)$.

Proof. Note that

$$\begin{aligned}
\text{SSE} &= \|Y - \widehat{Y}\|^2_{\Sigma^{-1}} = \|\Sigma^{-1/2}Y - \Sigma^{-1/2}\widehat{Y}\|^2 \\
&= \|\Sigma^{-1/2}Y - \Sigma^{-1/2}X\left(\Sigma^{-1/2}X\right)^{+}\Sigma^{-1/2}Y\|^2 \\
&= \|\left[I - \left(\Sigma^{-1/2}X\right)\left(\Sigma^{-1/2}X\right)^{+}\right]\left(\Sigma^{-1/2}Y\right)\|^2 \\
&= \left(\Sigma^{-1/2}Y\right)'\left[I - \left(\Sigma^{-1/2}X\right)\left(\Sigma^{-1/2}X\right)^{+}\right]\left(\Sigma^{-1/2}Y\right).
\end{aligned}$$

So $\frac{\text{SSE}}{\sigma^2} = Z'BZ$ where $Z = \Sigma^{-1/2}Y \sim N\left(\Sigma^{-1/2}X\beta, \sigma^2 I\right)$ and

$$B = \frac{1}{\sigma^2}\left[I - \left(\Sigma^{-1/2}X\right)\left(\Sigma^{-1/2}X\right)^{+}\right].$$

But $B\sigma^2 IB = B$, $\left(\Sigma^{-1/2}X\beta\right)'B\left(\Sigma^{-1/2}X\beta\right) = 0$ and $\text{tr}(B\sigma^2 I) = n - r$.
The above imply that $\frac{\text{SSE}}{\sigma^2}\chi^2(n - r)$.

(3) $H\left(\Sigma^{-1/2}X\right)^{+}\Sigma^{-1/2}Y$ and SSE are independent.

Proof. With $Y \sim N(X\beta, \sigma^2\Sigma)$, $A = H\left(\Sigma^{-1/2}X\right)^{+}\Sigma^{-1/2}$, and $\text{SSE} = \sigma^2 Y'\left(\Sigma^{-1/2}B\Sigma^{-1/2}\right)Y$, $A\left(\sigma^2\Sigma\right)\left(\Sigma^{-1/2}B\Sigma^{-1/2}\right) = 0$ from which the conclusion of the independence of $AY = H\left(\Sigma^{-1/2}X\right)^{+}\Sigma^{-1/2}Y$ and SSE follows.

L08: A biased estimator: Ridge estimator

1. The problem of multicollinearity

   (1) Biased estimators
   When ξ is estimated by $\widehat{\xi}$, the risks $r(\widehat{\xi}, \xi) = \text{Cov}(\widehat{\xi}\,) + [E(\widehat{\xi}\,) - \xi][E(\widehat{\xi}\,) - \xi]'$ and $\text{MSE}(\widehat{\xi}, \xi) = \text{tr}[r(\widehat{\xi}, \xi)] = \text{tr}[\text{Cov}(\widehat{\xi}\,)] + \|E(\widehat{\xi}\,) - \xi\|^2$.
   Reducing the large $\text{tr}[\text{Cov}(\widehat{\xi}\,)]$ may cause the increase in the bias and result in a biased estimator. However, if the reduction in $\text{tr}[\text{Cov}(\widehat{\xi}\,)]$ is greater than the increment in the bias, then it is worthwhile to do so.

   (2) BLUE of β
   In Model $Y = X\beta + \epsilon$, $\epsilon \sim (0, \sigma^2 I_n)$, if the columns of X are linearly independent, then β is estimable since $\beta = I_p\beta$ and $I_p = X^+X$. The BLUE for β

   $$\widehat{\beta} = X^+Y = (X'X)^{-1}X'Y \sim \left(\beta, \sigma^2(X'X)^{-1}\right).$$

   Let $X'X = P\Lambda P'$ be the EVD. Then $r(\widehat{\beta}, \beta) = \text{Cov}(\widehat{\beta}) = \sigma^2(X'X)^{-1} = \sigma^2 P\Lambda^{-1}P'$ and $\text{MSE}(\widehat{\xi}, \xi) = \text{tr}\left(\sigma^2 P\Lambda^{-1}P'\right) = \frac{\sigma^2}{\lambda_1} + \cdots + \frac{\sigma^2}{\lambda_p}$.

   (3) The problem of multicollinearity in X
   Note that the columns of X are linearly independent if and only if $|X'X| = \prod_i \lambda_i > 0$.
   We say that there is a multicollinearity in X if the columns of X are almost linearly dependent interpreted as $|X'X| = \prod_i \lambda_i$ is almost 0.
   So the multicollinearity will make $\text{MSE}(\widehat{\beta}, \beta) = \text{tr}[\text{Cov}(\widehat{\beta}\,)] = \frac{\sigma^2}{\lambda_1} + \cdots + \frac{\sigma^2}{\lambda_p}$ large. Thus while $\widehat{\beta}$ is still a BLUE, but it is not stable due to large total variances, also its risk $\text{MSE}(\widehat{\beta}, \beta)$ is high.

2. Ridge estimator

   (1) Ridge estimator
   One naive idea on the remedy for the problem caused by the smaller λ_i, $i = 1, .., p$, in

   $$\widehat{\beta} = (X'X)^{-1}X'Y = (P\Lambda P')^{-1}X'Y$$

   is to replace λ_i by $\lambda_i + k_i$ where $k_i > 0$, i.e., to replace Λ by $\Lambda + K$ where $K = \text{diag}(k_1, .., k_p)$ to have

   $$\widehat{\beta}(K) = [P(\Lambda + K)P']^{-1}X'Y = P(\Lambda + K)^{-1}P'X'Y$$

   called a ridge estimator for β. The ridge estimator is still a linear estimator for β.

   (2) $\text{tr}\left[\text{Cov}\left(\widehat{\beta}(K)\right)\right] = \sigma^2 \sum_i \frac{\lambda_i}{(\lambda_i + k_i)^2}$

   Proof. With $\widehat{\beta}(K) = P(\Lambda + K)^{-1}P'X'Y$ and $Y \sim (X\beta, \sigma^2 I_n)$,

   $$\begin{aligned}\text{Cov}\left[\widehat{\beta}(K)\right] &= [P(\Lambda + K)^{-1}P'X']\sigma^2 I[P(\Lambda + K)^{-1}P'X']' \\ &= \sigma^2 P(\Lambda + K)^{-1}P'X'XP(\Lambda + K)^{-1}P' \\ &= \sigma^2 P(\Lambda + K)^{-1}\Lambda(\Lambda + K)^{-1}P'\end{aligned}$$

   So $\text{tr}\left[\text{Cov}\left(\widehat{\beta}(K)\right)\right] = \sigma^2 \text{tr}\left[(\Lambda + K)^{-1}\Lambda(\Lambda + K)^{-1}\right] = \sigma^2 \sum_i \frac{\lambda_i}{(\lambda_i + k_i)^2}$

Ex1: $\text{tr}\left[\text{Cov}\left(\widehat{\beta}(K)\right)\right] = \sigma^2 \sum_i \frac{\lambda_i}{(\lambda_i+k_i)^2} \le \sigma^2 \sum_i \frac{1}{\lambda_i} = \text{tr}\left(\text{Cov}(\widehat{\beta}\ )\right).$

(3) $\|\widehat{\beta}(K) - \beta\|^2 = \sum_i \frac{k_i^2}{(\lambda_i+k_i^2)^2}\,[(P'\beta)_i]^2$

Proof. First, $\beta - E[\widehat{\beta}(K)] = \beta - P(\Lambda+K)^{-1}P'X'X\beta = \beta - P(\Lambda+K)^{-1}\Lambda P'\beta$
$= P[I - (\Lambda+K)^{-1}\Lambda]P'\beta.$

But $(\Lambda+k)^{-1} = \Lambda^{-1} - \Lambda^{-1}(\Lambda^{-1}+K^{-1})^{-1}\Lambda^{-1}$ since

$$\begin{aligned}
& (\Lambda+K)[\Lambda^{-1} - \Lambda^{-1}(\Lambda^{-1}+K^{-1})^{-1}\Lambda^{-1}] \\
= \;& I - (\Lambda^{-1}+K^{-1})^{-1}\Lambda^{-1} + K\Lambda^{-1} - K\Lambda^{-1}(\Lambda^{-1}+K^{-1})^{-1}\Lambda^{-1} \\
= \;& I + K[-K^{-1}(\Lambda^{-1}+K^{-1})^{-1} + I - \Lambda^{-1}(\Lambda^{-1}+K^{-1})^{-1}]\Lambda^{-1} \\
= \;& I + K[I - (K^{-1}+\Lambda^{-1})(\Lambda^{-1}+K^{-1})^{-1}]\Lambda^{-1} = I.
\end{aligned}$$

$$\begin{aligned}
\text{So } \|\beta - E[\widehat{\beta}(K)]\|^2 &= \|P[I - (\Lambda+K)^{-1}\Lambda]P'\beta\|^2 \\
&= \|P\{I - [\Lambda^{-1} - \Lambda^{-1}(\Lambda^{-1}+K^{-1})^{-1}\Lambda^{-1}]\Lambda\}P'\beta\|^2 \\
&= \|[\Lambda^{-1}(\Lambda^{-1}+K^{-1})^{-1}]P'\beta\|^2 \\
&= \sum_i \left[\frac{1/\lambda_i}{(1/\lambda_i)+(1/k_i)}(P'\beta)_i\right]^2 = \sum_i \frac{k_i^2}{(\lambda_i+k_i)^2}\,[(P'\beta)_i]^2\,.
\end{aligned}$$

Ex2: $\|E(\widehat{\beta}(K)) - \beta\|^2 \ge 0 = \|E(\widehat{\beta}\ ) - \beta\|^2.$

3. Making ridge estimator better than BLUE

(1) MSE($\widehat{\beta}(K)$, β)

$$\begin{aligned}
\text{MSE}(\widehat{\beta}(K),\, \beta) &= \text{tr}\left[\text{Cov}\left(\widehat{\beta}(K)\right)\right] + \|E(\widehat{\beta}(K)) - \beta\|^2 \\
&= \sum_i \frac{\sigma^2\lambda_i}{(\lambda_i+k_i)^2} + \sum_i \frac{k_i^2}{(\lambda_i+k_i)^2}\,[(P'\beta)_i]^2 = \sum_i \frac{k_i^2[(P'\beta)_i]^2 + \lambda_i\sigma^2}{(\lambda_i+k_i)^2} \\
&= \sum_i f_i(k_i) \qquad\qquad \text{where } f_i(k_i) = \frac{k_i^2[(P'\beta)_i]^2 + \lambda_i\sigma^2}{(\lambda_i+k_i)^2}.
\end{aligned}$$

(2) Minimizing MSE($\widehat{\beta}(K)$, β)

$f'(k_i) = \cdots = \frac{2k_i\lambda_i[(P'\beta)_i]^2 - 2\lambda\sigma^2}{(\lambda_i+k_i)^3} = \frac{2\lambda_i[(P'\beta)_i]^2}{(\lambda_i+k_i)^3}\left[k_i - \frac{\sigma^2}{[(P'\beta)_i]^2}\right]$. By the first derivative test, $f(k_i)$ is minimized at $k_i = \frac{\sigma^2}{[(P'\beta)_i]^2}$, $i = 1, ..., p$, So is MSE($\widehat{\beta}(K)$, β).

(3) Ridge estimator could be better than the BLUE

$$\begin{aligned}
\text{MSE}(\widehat{\beta}(K),\, \beta)_{k_i = \frac{\sigma^2}{[(P'\beta)_i]^2}} &= \sum_{i=1}^p \frac{\frac{\sigma^4}{[(P'\beta)_i]^2} + \lambda_i\sigma^2}{\left[\lambda_i + \frac{\sigma^2}{[(P'\beta)_i]^2}\right]^2} = \sum_{i=1}^p \frac{\sigma^2\left[\lambda_i + \frac{\sigma^2}{[(P'\beta)_i]^2}\right]}{\left[\lambda_i + \frac{\sigma^2}{[(P'\beta)_i]^2}\right]^2} \\
&= \sum_{i=1}^p \frac{\sigma^2}{\lambda_i + \frac{\sigma^2}{[(P'\beta)_i]^2}} \le \sum_{i=1}^p \frac{\sigma^2}{\lambda_i} = \text{MSE}(\widehat{\beta})
\end{aligned}$$

Comment: $k_i = \frac{\sigma^2}{[(P'\beta)_i]^2}$ is a theoretical value since it depends on unknown parameter σ^2 and β. In practice one can estimate σ^2 and β, and use the estimated value of k_i.