Chapter 3. Linear regression for the full-rank model

The linear regression model is probably the most fundamental and widely used statistical model. Consider the following general linear model in matrix form

$$\boldsymbol{Y}_{n\times 1} = \boldsymbol{X}_{n\times p}\boldsymbol{\beta}_{0,p\times 1} + \boldsymbol{\varepsilon}_{n\times 1}, \tag{1}$$

where $(Y, X^\top)$ is a pair of response and p-dimensional vector of covariates and ε is unobservable error term, $\boldsymbol{\beta}_0$ is the true value of $\boldsymbol{\beta}$. The least squares (LS) and the least absolute deviation (LAD) are among the most widely-used criterions in statistical estimation for linear regression model. As a standard case, we consider $\boldsymbol{X}$ is of full column rank, that is $r(\boldsymbol{X}) = p$.

3.1 Ordinary least squares estimation

For ordinary least squares estimation, it is commonly assumed that $E(\boldsymbol{\varepsilon}) = \boldsymbol{0}$ and $Cov(\varepsilon) = \sigma^2\boldsymbol{I}$, among which the mean-zero condition is an identifiability condition for the intercept component of β. The celebrated least squares estimate is to minimize

$$(\boldsymbol{Y} - \boldsymbol{X\beta})^\top(\boldsymbol{Y} - \boldsymbol{X\beta})$$

over $\boldsymbol{\beta}$. Simple calculations yields that

$$\begin{aligned} L(\boldsymbol{\beta}) &\equiv (\boldsymbol{Y} - \boldsymbol{X\beta})^\top(\boldsymbol{Y} - \boldsymbol{X\beta}) \\ &= \boldsymbol{Y}^\top\boldsymbol{Y} - \boldsymbol{2Y}^\top\boldsymbol{X\beta} + \boldsymbol{\beta}^\top\boldsymbol{X}^\top\boldsymbol{X\beta}. \end{aligned}$$

Then,

$$\frac{\partial L(\boldsymbol{\beta})}{\partial\boldsymbol{\beta}} = -2\boldsymbol{X}^\top\boldsymbol{Y} + \boldsymbol{2X}^\top\boldsymbol{X\beta} = \boldsymbol{0}$$

leads to the so-called *normal equation*

$$\begin{aligned} &\boldsymbol{X}^\top\boldsymbol{Y} = \boldsymbol{X}^\top\boldsymbol{X}\hat{\boldsymbol{\beta}} \\ \Rightarrow\ &\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^\top\boldsymbol{X})^{-\boldsymbol{1}}\boldsymbol{X}^\top\boldsymbol{Y} \end{aligned}$$

Compared with other existing methods, the LS is easy to implement and most popular, as its objective function $L(\boldsymbol{\beta})$ is convex and the solution $\hat{\boldsymbol{\beta}}$ is of a closed form.

Remark 1. Note that

$$\begin{aligned}
&(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta})^\top(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta})\\
=&[\boldsymbol{Y}-\boldsymbol{X}\hat{\boldsymbol{\beta}}+\boldsymbol{X}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})]^\top[\boldsymbol{Y}-\boldsymbol{X}\hat{\boldsymbol{\beta}}+\boldsymbol{X}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})]\\
=&(\boldsymbol{Y}-\boldsymbol{X}\hat{\boldsymbol{\beta}})^\top(\boldsymbol{Y}-\boldsymbol{X}\hat{\boldsymbol{\beta}})+(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})^\top\boldsymbol{X}^\top\boldsymbol{X}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})+\boldsymbol{2}(\boldsymbol{Y}-\boldsymbol{X}\hat{\boldsymbol{\beta}})^\top\boldsymbol{X}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}).
\end{aligned}$$

But

$$\begin{aligned}
(\boldsymbol{Y}-\boldsymbol{X}\hat{\boldsymbol{\beta}})^\top\boldsymbol{X}&=(\boldsymbol{Y}-\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-\boldsymbol{1}}\boldsymbol{X}^\top\boldsymbol{Y})^\top\boldsymbol{X}\\
&=\boldsymbol{0}.
\end{aligned}$$

Then,

$$\begin{aligned}
&(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta})^\top(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta})\\
=&(\boldsymbol{Y}-\boldsymbol{X}\hat{\boldsymbol{\beta}})^\top(\boldsymbol{Y}-\boldsymbol{X}\hat{\boldsymbol{\beta}})+(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})^\top\boldsymbol{X}^\top\boldsymbol{X}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})\\
\geq&0,
\end{aligned}$$

which achieves its minimum when $\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}$.
Therefore, $\hat{\boldsymbol{\beta}}=(\boldsymbol{X}^\top\boldsymbol{X})^{-\boldsymbol{1}}\boldsymbol{X}^\top\boldsymbol{Y}$ is the least squares estimate of $\boldsymbol{\beta_0}$.

3.1.1 Properties of the least squares estimate.

Given the least squares estimate, we define the vector of residuals as $\hat{\boldsymbol{\varepsilon}}=\boldsymbol{Y}-\boldsymbol{X}\hat{\boldsymbol{\beta}}$. Hence

$$\begin{aligned}
\hat{\varepsilon}&=\boldsymbol{Y}-\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-\boldsymbol{1}}\boldsymbol{X}^\top\boldsymbol{Y}\\
&=[\boldsymbol{I}-\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-\boldsymbol{1}}\boldsymbol{X}^\top]\boldsymbol{Y}\\
&=[\boldsymbol{I}-\boldsymbol{H}]\boldsymbol{Y}
\end{aligned}$$

where $\boldsymbol{H}=\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-\boldsymbol{1}}\boldsymbol{X}^\top$ is the so-called hat matrix of order $n\times n$. To obtain a fitted value of $\boldsymbol{Y}$, we plug in $\hat{\boldsymbol{\beta}}$ and get

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

There are a number of properties here:

Properties:

1. The hat matrix $\boldsymbol{H}$ is symmetric idempotent;

   Proof: Note that $(\boldsymbol{X}^\top \boldsymbol{X})$ is symmetric and $(\boldsymbol{X}^\top \boldsymbol{X})^{-1}$ is also symmetric. Then,

   $$\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top = \boldsymbol{H}^\top, \qquad \boldsymbol{H}\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top \boldsymbol{X}(\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top = \boldsymbol{H}.$$

2. $\boldsymbol{X}^\top \hat{\boldsymbol{\varepsilon}} = \boldsymbol{0}$; (This holds because of $\boldsymbol{X}^\top \boldsymbol{H} = \boldsymbol{X}^\top$, $\boldsymbol{H}\boldsymbol{X} = \boldsymbol{X}$ and $\boldsymbol{X}^\top(\boldsymbol{I}-\boldsymbol{H}) = \boldsymbol{0}$, $(\boldsymbol{I}-\boldsymbol{H})\boldsymbol{X} = \boldsymbol{0}$.)

   Proof: Since

   $$\begin{aligned}\boldsymbol{X}^\top \boldsymbol{H} &= \boldsymbol{X}^\top \boldsymbol{X}(\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top = \boldsymbol{X}^\top,\\ \boldsymbol{H}\boldsymbol{X} &= \boldsymbol{X}(\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top \boldsymbol{X} = \boldsymbol{X},\end{aligned}$$

   then,

   $$\begin{aligned}\boldsymbol{X}^\top(\boldsymbol{I}-\boldsymbol{H}) &= \boldsymbol{X}^\top - \boldsymbol{X}^\top \boldsymbol{H} = \boldsymbol{X}^\top - \boldsymbol{X}^\top = \boldsymbol{0},\\ (\boldsymbol{I}-\boldsymbol{H})\boldsymbol{X} &= \boldsymbol{X} - \boldsymbol{H}\boldsymbol{X} = \boldsymbol{X} - \boldsymbol{X} = \boldsymbol{0}.\end{aligned}$$

   Clearly,

   $$\boldsymbol{X}^\top \hat{\boldsymbol{\varepsilon}} = \boldsymbol{X}^\top(\boldsymbol{I}-\boldsymbol{H})\boldsymbol{Y} = \boldsymbol{0}.$$

3. $\hat{\boldsymbol{Y}}^\top \hat{\boldsymbol{\varepsilon}} = \boldsymbol{0}$;

   Proof: Write

   $$\begin{aligned}(\boldsymbol{H}\boldsymbol{Y})^\top(\boldsymbol{I}-\boldsymbol{H})\boldsymbol{Y} &= \boldsymbol{Y}^\top \boldsymbol{H}^\top(\boldsymbol{I}-\boldsymbol{H})\boldsymbol{Y} = \boldsymbol{Y}^\top \boldsymbol{H}(\boldsymbol{I}-\boldsymbol{H})\boldsymbol{Y}\\ &= \boldsymbol{Y}^\top \boldsymbol{0}\boldsymbol{Y} = \boldsymbol{0}.\end{aligned}$$

4. $\boldsymbol{I}-\boldsymbol{H}$ is symmetric idempotent;

5. $E(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta}_0$ (unbiased estimate);

   Proof:

   $$E(\hat{\boldsymbol{\beta}}) = E((\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top \boldsymbol{Y}) = (\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top E(\boldsymbol{Y}) = (\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top \boldsymbol{X}\boldsymbol{\beta}_0 = \boldsymbol{\beta}_0.$$

6. $Cov(\hat{\boldsymbol{\beta}}) = (\boldsymbol{X}^\top \boldsymbol{X})^{-1}\sigma^2$;

   Proof:

   $$\begin{aligned}Cov(\hat{\boldsymbol{\beta}}) &= Cov((\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top \boldsymbol{Y}) = (\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top Cov(\boldsymbol{Y})\boldsymbol{X}(\boldsymbol{X}^\top \boldsymbol{X})^{-1}\\ &= \sigma^2(\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top \boldsymbol{I}\boldsymbol{X}(\boldsymbol{X}^\top \boldsymbol{X})^{-1} = \sigma^2(\boldsymbol{X}^\top \boldsymbol{X})^{-1}.\end{aligned}$$

7. $tr(\boldsymbol{I}_n - \boldsymbol{H}) = n - p$;

   Proof: Note that

$$tr(\boldsymbol{H}) = tr(\boldsymbol{X(X^\top X)^{-1}X^\top}) = tr(\boldsymbol{X^\top X(X^\top X)^{-1}}) = tr(\boldsymbol{I}_p) = p.$$

   Then,

$$tr(\boldsymbol{I}_n - \boldsymbol{H}) = tr(\boldsymbol{I}_n) - tr(\boldsymbol{H}) = n - p.$$

8. $\hat{\boldsymbol{\varepsilon}}^\top \hat{\boldsymbol{\varepsilon}} = tr(\boldsymbol{YY^\top(I - H)})$;

   Proof: We can easily show that

$$\begin{aligned}\hat{\boldsymbol{\varepsilon}}^\top \hat{\boldsymbol{\varepsilon}} &= \boldsymbol{Y^\top(I-H)^\top(I-H)Y} = \boldsymbol{Y^\top(I-H)Y} \\ &= tr(\boldsymbol{Y^\top(I-H)Y}) = tr(\boldsymbol{YY^\top(I-H)}).\end{aligned}$$

9. $E(\boldsymbol{YY^\top}) = \sigma^2\boldsymbol{I} + \boldsymbol{X\beta\beta^\top X^\top}$;

10. $\hat{\boldsymbol{\varepsilon}}^\top \hat{\boldsymbol{\varepsilon}}/(n-p)$ is an unbiased estimate of σ^2, that is

$$E(\frac{\hat{\boldsymbol{\varepsilon}}^\top \hat{\boldsymbol{\varepsilon}}}{n-p}) = \sigma^2.$$

   Proof: Write

$$\begin{aligned}E(\hat{\boldsymbol{\varepsilon}}^\top \hat{\boldsymbol{\varepsilon}}) &= E(\boldsymbol{Y^\top(I-H)(I-H)Y}) = E(\boldsymbol{Y^\top(I-H)Y}) \\ &= tr((\boldsymbol{I}-\boldsymbol{H})\boldsymbol{\Sigma}) + \boldsymbol{\beta^\top X^\top (I-H)X\beta} \\ &= \sigma^2 tr(\boldsymbol{I}-\boldsymbol{H}) = \sigma^2(n-p).\end{aligned}$$

   Thus, $E(\frac{\hat{\boldsymbol{\varepsilon}}^\top \hat{\boldsymbol{\varepsilon}}}{n-p}) = \sigma^2$.

   Remark 2. Note that in this course, we mostly consider fixed design, that is the covariate X is fixed and determistic. For random design, the least square estimation is still valid and its theoretical properties can be established without further difficulties.

3.2 The weighted least square estimation.

For a general case that $Cov(\boldsymbol{\varepsilon}) = \boldsymbol{\Sigma}$ and $\boldsymbol{\Sigma}$ is known, the weighted least squares will be used to estimate $\boldsymbol{\beta}$ in model (1). Note that $\boldsymbol{\Sigma} \neq \boldsymbol{I}$ in general but is positive definite, Recall that the ordinary least squares is to minimize $(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta})^\top(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta})$ and $\hat{\boldsymbol{\beta}} = \boldsymbol{(X^\top X)^{-1}X^\top Y}$. The weighted least squars (WLS) or generalized least squares (GLS) estimator is defined as the minimizer of

$$(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta})^\top\boldsymbol{\Sigma}^{-1}(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta})$$

over $\boldsymbol{\beta}$.

Similar to section 2.1, we let

$$\begin{aligned} S(\boldsymbol{\beta}) &= \boldsymbol{(Y-X\beta)^\top\Sigma^{-1}(Y-X\beta)} \\ &= \boldsymbol{Y^\top\Sigma^{-1}Y - 2Y^\top\Sigma^{-1}X\beta + \beta^\top X^\top\Sigma^{-1}X\beta}. \end{aligned}$$

Then,

$$\begin{aligned} \frac{\partial S(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} &= -2\boldsymbol{X^\top\Sigma^{-1}Y + 2X^\top\Sigma^{-1}X\beta} \\ &= \boldsymbol{0} \\ \Rightarrow \boldsymbol{X^\top\Sigma^{-1}Y} &= \boldsymbol{X^\top\Sigma^{-1}X\beta} \\ \Rightarrow \tilde{\boldsymbol{\beta}} =& \boldsymbol{(X^\top\Sigma^{-1}X)^{-1}X^\top\Sigma^{-1}Y}. \end{aligned}$$

Note that $E(\tilde{\boldsymbol{\beta}}) = \boldsymbol{\beta}_0$ and $Cov(\tilde{\boldsymbol{\beta}}) = (\boldsymbol{X^\top\Sigma^{-1}X})^{-1}$.

Remark 3. When $\boldsymbol{\Sigma} = \sigma^2\boldsymbol{I}$, the WLS or GLS reduces to the OLS.

Remark 4. We provide another aspect to motivate the WLS. Since $\boldsymbol{\Sigma}$ is positive definite, $\boldsymbol{\Sigma}^{-1/2}$ exists such that $\boldsymbol{\Sigma}^{-1/2}\boldsymbol{\Sigma}^{-1/2} = \boldsymbol{\Sigma}^{-1}$. Thus,

$$\boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{Y} = \boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{X\beta} + \boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{\varepsilon}.$$

Now $E(\boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{\varepsilon}) = \boldsymbol{0}$ and $Cov(\boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{\varepsilon}) = \boldsymbol{I}_n$ satisfy the conditions of the ordinary least squares. Thereby,

$$\begin{aligned} \tilde{\boldsymbol{\beta}} &= \left\{(\boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{X})^\top(\boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{X})\right\}^{-1}(\boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{X})^\top\boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{Y} \\ &= \boldsymbol{(X^\top\Sigma^{-1}X)^{-1}X^\top\Sigma^{-1}Y}. \end{aligned}$$

3.3 The Best linear unbiased estimator (b.l.u.e. or BLUE) (Gauss-Markov Theorem)

Let $\boldsymbol{t} \in \mathbb{R}^p$ be a vector. We consider the problem of finding the b.l.u.e. of $\boldsymbol{t}^\top\boldsymbol{\beta}$. Let $\boldsymbol{\lambda}^\top\boldsymbol{Y}$ be a linear function of the observations and an estimator of $\boldsymbol{t}^\top\boldsymbol{\beta}$. To find the BLUE of $\boldsymbol{t}^\top\boldsymbol{\beta}$ is to determine $\boldsymbol{\lambda}$ such that $\boldsymbol{\lambda}^\top\boldsymbol{Y}$ is unbiased for $\boldsymbol{t}^\top\boldsymbol{\beta}$ and has minimum variance among all the linear unbiased estimates. To this end,

1. First, if $\boldsymbol{\lambda}^\top\boldsymbol{Y}$ is an unbiased estimator of $\boldsymbol{t}^\top\boldsymbol{\beta}$, $E(\boldsymbol{\lambda}^\top\boldsymbol{Y}) = \boldsymbol{t}^\top\boldsymbol{\beta}$. But $E(\boldsymbol{\lambda}^\top\boldsymbol{Y}) = \boldsymbol{\lambda}^\top E(\boldsymbol{Y}) = \boldsymbol{\lambda}^\top\boldsymbol{X}\boldsymbol{\beta}$ according to model (1). Hence,

$$\boldsymbol{\lambda}^\top\boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{t}^\top\boldsymbol{\beta}$$

   which is true for all $\boldsymbol{\beta}$. Thus, $\boldsymbol{\lambda}^\top\boldsymbol{X} = \boldsymbol{t}^\top$.

2. Second, we need to find the linear unbiased estimator of $\boldsymbol{t}^\top\boldsymbol{\beta}$ which has minimum variance. Note that

$$Var(\boldsymbol{\lambda}^\top\boldsymbol{Y}) = \boldsymbol{\lambda}^\top\boldsymbol{\Sigma}\boldsymbol{\lambda}.$$

   Using $\boldsymbol{2\theta}$ as a vector of Lagrange multipliers, we need to minimize

$$W(\boldsymbol{\lambda}, \boldsymbol{\theta}) = \boldsymbol{\lambda}^\top\boldsymbol{\Sigma}\boldsymbol{\lambda} - 2\boldsymbol{\theta}^\top(\boldsymbol{X}^\top\boldsymbol{\lambda} - \boldsymbol{t}),$$

   where $\boldsymbol{X}^\top\boldsymbol{\lambda} = \boldsymbol{t}$ is the unbiasedness condition. Thus,

$$\frac{\partial W(\boldsymbol{\lambda}, \boldsymbol{\theta})}{\partial \boldsymbol{\lambda}} = 2\boldsymbol{\Sigma}\boldsymbol{\lambda} - 2\boldsymbol{X}\boldsymbol{\theta} = 0,$$
$$\frac{\partial W(\boldsymbol{\lambda}, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = 2\boldsymbol{X}^\top\boldsymbol{\lambda} - 2\boldsymbol{t} = \mathbf{0}.$$

   Solving the above two equations for $\boldsymbol{\lambda}$ and $\boldsymbol{\theta}$, we have

$$\boldsymbol{\lambda}^\top = \boldsymbol{t}^\top(\boldsymbol{X}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{\Sigma}^{-1}.$$

   Therefore, the BLUE of $\boldsymbol{t}^\top\boldsymbol{\beta}$ is

$$\boldsymbol{\lambda}^\top\boldsymbol{Y} = \boldsymbol{t}^\top(\boldsymbol{X}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{Y},$$

   with variance

$$\begin{aligned} Var(\boldsymbol{\lambda}^\top\boldsymbol{Y}) &= \boldsymbol{t}^\top(\boldsymbol{X}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{\Sigma}^{-1}(\Sigma)(\boldsymbol{\Sigma}^{-1})\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{X})^{-1}\boldsymbol{t} \\ &= \boldsymbol{t}^\top(\boldsymbol{X}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{X})^{-1}\boldsymbol{t}. \end{aligned}$$

Remark 5. In a special case that $\boldsymbol{\Sigma} = \sigma^2 \boldsymbol{I}$, the BLUE of $\boldsymbol{t}^\top \boldsymbol{\beta}$ is

$$\boldsymbol{t}^\top (\boldsymbol{X}^\top (\boldsymbol{I}\boldsymbol{\sigma}^{\mathbf{2}})^{-\mathbf{1}} \boldsymbol{X})^{-\mathbf{1}} \boldsymbol{X}^\top (\boldsymbol{I}\boldsymbol{\sigma}^{\mathbf{2}})^{-\mathbf{1}} \boldsymbol{Y} = \boldsymbol{t}^\top (\boldsymbol{X}^\top \boldsymbol{X})^{-\mathbf{1}} \boldsymbol{X}^\top \boldsymbol{Y},$$

with variance

$$\boldsymbol{t}^\top (\boldsymbol{X}^\top (\boldsymbol{I}\boldsymbol{\sigma}^{\mathbf{2}})^{-\mathbf{1}} \boldsymbol{X})^{-\mathbf{1}} \boldsymbol{t} = \sigma^2 \boldsymbol{t}^\top (\boldsymbol{X}^\top \boldsymbol{X})^{-\mathbf{1}} \boldsymbol{t}.$$

Remark 6. By letting $\boldsymbol{t}^\top$ be, in turn, each row of $\boldsymbol{I}_k$, we can easily obtain the BLUE of $\boldsymbol{\beta} = \tilde{\boldsymbol{\beta}} = (\boldsymbol{X}^\top \boldsymbol{\Sigma}^{-\mathbf{1}} \boldsymbol{X})^{-\mathbf{1}} \boldsymbol{X}^\top \boldsymbol{\Sigma}^{-\mathbf{1}} \boldsymbol{Y}$, which is precisely the weighted least square estimate or generalized least square estimate.

Remark 7. When $\boldsymbol{\Sigma} = \sigma^2 \boldsymbol{I}$, the BLUE of $\boldsymbol{\beta}$ is $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^\top \boldsymbol{X})^{-\mathbf{1}} \boldsymbol{X}^\top \boldsymbol{Y}$.

In summary, the least square estimate of β_0 in (1) is the best linear unbiased estimate.

THEOREM 1. *$W = \boldsymbol{\lambda}^\top \boldsymbol{\Sigma} \boldsymbol{\lambda}$ is minimized if*

$$\boldsymbol{\lambda}^\top = \boldsymbol{t}^\top (\boldsymbol{X}^\top \boldsymbol{\Sigma}^{-\mathbf{1}} \boldsymbol{X})^{-\mathbf{1}} \boldsymbol{X}^\top \boldsymbol{\Sigma}^{-\mathbf{1}}$$

subject to the constraint that

$$\boldsymbol{X}^\top \boldsymbol{\lambda} = \boldsymbol{t}.$$

Proof. Let $\boldsymbol{\lambda}_1^\top = \boldsymbol{t}^\top (\boldsymbol{X}^\top \boldsymbol{\Sigma}^{-\mathbf{1}} \boldsymbol{X})^{-\mathbf{1}} \boldsymbol{X}^\top \boldsymbol{\Sigma}^{-\mathbf{1}}$. Let $\boldsymbol{\lambda}_2$ be another vector that is different from $\boldsymbol{\lambda}$ but also satisfies $\boldsymbol{X}^\top \boldsymbol{\lambda}_{\mathbf{2}} = \boldsymbol{t}$. Then,

$$\begin{aligned} W^\top &= \boldsymbol{\lambda}_{\mathbf{2}}^\top \boldsymbol{\Sigma} \boldsymbol{\lambda}_{\mathbf{2}} \\ &= [(\boldsymbol{\lambda}_{\mathbf{2}} - \boldsymbol{\lambda}_{\mathbf{1}}) + \boldsymbol{\lambda}_{\mathbf{1}}]^\top \boldsymbol{\Sigma} [(\boldsymbol{\lambda}_{\mathbf{2}} - \boldsymbol{\lambda}_{\mathbf{1}}) + \boldsymbol{\lambda}_{\mathbf{1}}] \\ &= (\boldsymbol{\lambda}_{\mathbf{2}} - \boldsymbol{\lambda}_{\mathbf{1}})^\top \boldsymbol{\Sigma} (\boldsymbol{\lambda}_{\mathbf{2}} - \boldsymbol{\lambda}_{\mathbf{1}}) + \boldsymbol{\lambda}_{\mathbf{1}}^\top \boldsymbol{\Sigma} \boldsymbol{\lambda}_{\mathbf{1}} + (\boldsymbol{\lambda}_{\mathbf{2}} - \boldsymbol{\lambda}_{\mathbf{1}})^\top \boldsymbol{\Sigma} \boldsymbol{\lambda}_{\mathbf{1}} + \boldsymbol{\lambda}_{\mathbf{1}}^\top \boldsymbol{\Sigma} (\boldsymbol{\lambda}_{\mathbf{2}} - \boldsymbol{\lambda}_{\mathbf{1}}). \end{aligned}$$

Nevertheless,

$$\begin{aligned} (\boldsymbol{\lambda}_{\mathbf{2}} - \boldsymbol{\lambda}_{\mathbf{1}})^\top \boldsymbol{\Sigma} \boldsymbol{\lambda}_{\mathbf{1}} &= (\boldsymbol{\lambda}_{\mathbf{2}} - \boldsymbol{\lambda}_{\mathbf{1}})^\top \boldsymbol{\Sigma} [\boldsymbol{\Sigma}^{-\mathbf{1}} \boldsymbol{X} (\boldsymbol{X}^\top \boldsymbol{\Sigma}^{-\mathbf{1}} \boldsymbol{X})^{-\mathbf{1}} \boldsymbol{t}] \\ &= (\boldsymbol{\lambda}_{\mathbf{2}} - \boldsymbol{\lambda}_{\mathbf{1}})^\top \boldsymbol{X} (\boldsymbol{X}^\top \boldsymbol{\Sigma}^{-\mathbf{1}} \boldsymbol{X})^{-\mathbf{1}} \boldsymbol{t} \\ &= \mathbf{0} (\text{this is because } \boldsymbol{\lambda}_{\mathbf{1}}^\top \boldsymbol{X} = \boldsymbol{t}^\top \text{and } \boldsymbol{\lambda}_{\mathbf{2}}^\top \boldsymbol{X} = \boldsymbol{t}^\top). \end{aligned}$$

Also,

$$\boldsymbol{\lambda}_{\mathbf{1}}^\top \boldsymbol{\Sigma} (\boldsymbol{\lambda}_{\mathbf{2}} - \boldsymbol{\lambda}_{\mathbf{1}}) = (\boldsymbol{\lambda}_{\mathbf{2}} - \boldsymbol{\lambda}_{\mathbf{1}})^\top \boldsymbol{\Sigma} \boldsymbol{\lambda}_{\mathbf{1}} = 0.$$

As a result,

$$W^\top = (\boldsymbol{\lambda_2} - \boldsymbol{\lambda_1})^\top \boldsymbol{\Sigma} (\boldsymbol{\lambda_2} - \boldsymbol{\lambda_1}) + \boldsymbol{\lambda_1}^\top \Sigma \boldsymbol{\lambda_1}.$$

which is minimized if $\boldsymbol{\lambda_2} = \boldsymbol{\lambda_1}$. The proof is complete. $\square$

3.4 Least squares theory when the parameters are random variables (random-effect model)

In this section, we assume that the parameters of the regression models are random variables with a known mean and variance. Consider the linear model

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{b} + \boldsymbol{e}, \tag{2}$$

where $(Y_i, b_i, e_i), i = 1, \ldots, n$ are independent and identically distributed (i.i.d) copies of (Y, b, e), and $E(\boldsymbol{b}) = \boldsymbol{\theta}$ and $Cov(\boldsymbol{b}) = \boldsymbol{F}$, $\boldsymbol{\theta}$ is a k-dimensional vector and $\boldsymbol{F}$ is a $k \times k$ positive definite matrix. Also assume that

$$E(\boldsymbol{e}|\boldsymbol{b}) = 0, \quad Cov(\boldsymbol{e}|\boldsymbol{b}) = \boldsymbol{V}.$$

We then show how to find the best linear estimator (predictor) of a random variable $\boldsymbol{p}^\top \boldsymbol{b}$, where $\boldsymbol{p} \in \mathbb{R}^k$ is a given vector. The following formulae connect the conditional and unconditional means and variances.

$$\begin{aligned} E(\boldsymbol{Y}) &= E(E(\boldsymbol{Y}|\boldsymbol{e})), \\ Var(\boldsymbol{Y}) &= E\{Var(\boldsymbol{Y}|\boldsymbol{b})\} + Var\{E(\boldsymbol{Y}|\boldsymbol{b})\} = \boldsymbol{V} + \boldsymbol{X}\boldsymbol{F}\boldsymbol{X}^\top, \\ Cov(\boldsymbol{Y}, \boldsymbol{p}^\top \boldsymbol{b}) &= E\{Cov(\boldsymbol{Y}, \boldsymbol{p}^\top \boldsymbol{b}|\boldsymbol{b})\} + Cov[E(\boldsymbol{Y}|\boldsymbol{b}), \boldsymbol{p}^\top \boldsymbol{b}] = \boldsymbol{X}\boldsymbol{F}\boldsymbol{p}. \end{aligned} \tag{3}$$

Students need to show the above formula by themselves as basic exercises on conditional expectation. The third equation above is by the **law of total covariance**, that is,

$$Cov(X, Y) = E[Cov(X, Y|Z)] + Cov(E(X|Z), E(Y|Z)).$$

The objective is to determine a linear function $a + \boldsymbol{L}^\top \boldsymbol{Y}$ such that

$$E(\boldsymbol{p}^\top \boldsymbol{b} - a - \boldsymbol{L}^\top \boldsymbol{Y}) = 0, \tag{4}$$

and

$$v \equiv Var(\boldsymbol{p}^\top \boldsymbol{b} - a - \boldsymbol{L}^\top \boldsymbol{Y}) \qquad \text{achieves its minimum.} \tag{5}$$

THEOREM 2. *The optimum estimator/predictor that satisfies (4) and (5) takes the form*

$$\boldsymbol{p}^\top \hat{\boldsymbol{b}} = \boldsymbol{p}^\top \boldsymbol{\theta} + \boldsymbol{p}^\top \boldsymbol{F}\boldsymbol{X}^\top (\boldsymbol{V} + \boldsymbol{X}\boldsymbol{F}\boldsymbol{X}^\top)^{-1}(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\theta}) \tag{6}$$

$$= \boldsymbol{p}^\top \boldsymbol{\theta} + \boldsymbol{p}^\top (\boldsymbol{F}^{-1} + \boldsymbol{X}^\top \boldsymbol{V}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^\top \boldsymbol{V}^{-1}(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\theta}). \tag{7}$$

Proof: The expectation in (4) yields

$$a = (\boldsymbol{p}^\top - \boldsymbol{L}^\top \boldsymbol{X})\boldsymbol{\theta}. \tag{8}$$

Employing the three formula in (3), the quantity to be minimized in (5) is

$$v = \boldsymbol{p}^\top \boldsymbol{F}\boldsymbol{p} + \boldsymbol{L}^\top (\boldsymbol{X}\boldsymbol{F}\boldsymbol{X}^\top + \boldsymbol{V})\boldsymbol{L} - 2\boldsymbol{L}^\top \boldsymbol{X}\boldsymbol{F}\boldsymbol{p}.$$

Then, differentiating v with respect to $\boldsymbol{L}$ and setting the results equal to zero, we obtain

$$(\boldsymbol{X}\boldsymbol{F}\boldsymbol{X}^\top + \boldsymbol{V})\boldsymbol{L} = \boldsymbol{X}\boldsymbol{F}\boldsymbol{p}$$

and the optimizing $\boldsymbol{L}$ is

$$\boldsymbol{L} = (\boldsymbol{X}\boldsymbol{F}\boldsymbol{X}^\top + \boldsymbol{V})^{-1}\boldsymbol{X}\boldsymbol{F}\boldsymbol{p}. \tag{9}$$

Substitution of (8) and (9) into $a + \boldsymbol{L}^\top \boldsymbol{Y}$ yields (6). The equivalence of the two expressions in (7) is established by using the following Woodbury (1950) matrix identity

$$(\boldsymbol{A} + \boldsymbol{B}\boldsymbol{C}\boldsymbol{D})^{-1} = \boldsymbol{A}^{-1} - \boldsymbol{A}^{-1}\boldsymbol{B}(\boldsymbol{C}^{-1} + \boldsymbol{D}\boldsymbol{A}^{-1}\boldsymbol{B})^{-1}\boldsymbol{D}\boldsymbol{A}^{-1},$$

where $\boldsymbol{A} = \boldsymbol{V}$, $\boldsymbol{B} = \boldsymbol{X}$, $\boldsymbol{C} = \boldsymbol{F}$ and $\boldsymbol{D} = \boldsymbol{X}^\top$. The proof is complete.

Substitution into (9) gives the minimum variance

$$\begin{aligned} v_{min} &= \boldsymbol{p}^\top \boldsymbol{F}\boldsymbol{p} - \boldsymbol{p}^\top \boldsymbol{F}\boldsymbol{X}^\top (\boldsymbol{X}\boldsymbol{F}\boldsymbol{X}^\top + \boldsymbol{V})^{-1}\boldsymbol{X}\boldsymbol{F}\boldsymbol{p} \\ &= \boldsymbol{p}^\top (\boldsymbol{X}^\top \boldsymbol{V}^{-1}\boldsymbol{X})^{-1}\boldsymbol{p} - (\boldsymbol{X}^\top \boldsymbol{V}^{-1}\boldsymbol{X})^{-1}(\boldsymbol{F} + (\boldsymbol{X}^\top \boldsymbol{V}^{-1}\boldsymbol{X})^{-1})^{-1}(\boldsymbol{X}^\top \boldsymbol{V}^{-1}\boldsymbol{X})^{-1}\boldsymbol{p}. \end{aligned}$$

Notice that v_{min} is less than the variance of the least-square estimator.

Remark 8. When $\boldsymbol{F} = \sigma^2 \boldsymbol{G}^{-1}$, $\boldsymbol{V} = \sigma \boldsymbol{I}$ and $\boldsymbol{\theta} = \mathbf{0}$, the estimator in (6) reduces to

$$\boldsymbol{p}^\top \hat{\boldsymbol{b}} = \boldsymbol{p}^\top (\boldsymbol{X}^\top \boldsymbol{X} + \boldsymbol{G})^{\mathbf{-1}} \boldsymbol{X}^\top \boldsymbol{Y},$$

the *generalized ridge regression* estimator of C.R. Rao (1975). When $\boldsymbol{G} = k\boldsymbol{I}$, it reduces to the ridge regression estimator of Hoerl and Kennard (1970). We will introduce the ridge regression in details in later sections.