Review: Mean and Covariance

- The mean of a random vector $\mathbf{Z}$ is a m-by-1 vector with the i-th element equal to $\mathbb{E}(Z_i)$.

$$\boldsymbol{\mu}_{m\times 1} = \mathbb{E}[\mathbf{Z}] = \begin{pmatrix} \mathbb{E}Z_1 \\ \cdots \\ \mathbb{E}Z_m \end{pmatrix}.$$

- The covariance of $\mathbf{Z}$ is a symmetric m-by-m matrix with the (i, j)-th element equal to $\mathsf{Cov}(Z_i, Z_j)$.

$$\begin{aligned}\Sigma_{m\times m} = \mathsf{Cov}(\mathbf{Z}) &= \mathbb{E}\Big[(\mathbf{Z}-\boldsymbol{\mu})(\mathbf{Z}-\boldsymbol{\mu})^t\Big] \\ &= \begin{pmatrix} \mathsf{Var}(Z_1) & \cdots & \mathsf{Cov}(Z_1, Z_m) \\ \cdots & \cdots & \cdots \\ \mathsf{Cov}(Z_m, Z_1) & \cdots & \mathsf{Var}(Z_m) \end{pmatrix}.\end{aligned}$$

  The covariance matrix Σ is positive semi-definite (psd), that is, for any vector $\mathbf{a} \in \mathbb{R}^m$, $\mathbf{a}^t\Sigma\mathbf{a} \geq 0$.

- Affine transformations: $\mathbf{W} = \mathbf{a}_{n\times 1} + \mathbf{B}_{n\times m}\mathbf{Z}$,

$$\mathbb{E}[\mathbf{W}] = \mathbf{a} + \mathbf{B}\boldsymbol{\mu}, \quad \mathsf{Cov}(\mathbf{W}) = \mathbf{B}\boldsymbol{\Sigma}\mathbf{B}^t.$$

Especially, for $W = v_1 Z_1 + \cdots v_m Z_m = \mathbf{v}^t\mathbf{Z}$,

$$\begin{aligned}
\mathbb{E}[W] &= \mathbf{v}^t\boldsymbol{\mu} = \sum_{i=1}^{m} v_i\mu_i, \\
\mathsf{Var}(W) &= \mathbf{v}^t\Sigma\mathbf{v} = \sum_{i=1}^{m} v_i^2\mathsf{Var}(Z_i) + 2\sum_{i<j} v_i v_j \mathsf{Cov}(Z_i, Z_j).
\end{aligned}$$

Means and Covariances of LS Estimates

Recall our assumption: $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$ with

$$\mathbb{E}(\mathbf{e}) = \mathbf{0}, \quad \text{Cov}(\mathbf{e}) = \sigma^2 \mathbf{I}_n,$$

that is, $\mathbb{E}(\mathbf{y}) = \mathbf{X}\boldsymbol{\beta}, \quad \text{Cov}(\mathbf{y}) = \sigma^2 \mathbf{I}_n$.

Only $\mathbf{e}$ is random, the design matrix $\mathbf{X}$ are treated as given, and $\boldsymbol{\beta}$ are constants (although unknown). Suppose $\mathbf{X}$ is of full rank, so the LS estimate $\hat{\boldsymbol{\beta}}$ is unique. Under this assumption,

$$\begin{aligned} \mathbb{E}(\hat{\boldsymbol{\beta}}) &= \mathbb{E}(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\mathbf{y} \\ &= (\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\mathbb{E}\mathbf{y} \\ &= (\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\mathbf{X}\boldsymbol{\beta} = \boldsymbol{\beta} \end{aligned}$$

$$
\begin{aligned}
\mathsf{Cov}(\hat{\boldsymbol{\beta}}) &= (\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\mathsf{Cov}(\mathbf{y})\left[(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\right]^t \\
&= (\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\sigma^2\mathbf{X}(\mathbf{X}^t\mathbf{X})^{-1} \\
&= \sigma^2(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\mathbf{X}(\mathbf{X}^t\mathbf{X})^{-1} = \sigma^2(\mathbf{X}^t\mathbf{X})^{-1}; \\
\mathbb{E}(\hat{\mathbf{y}}) &= \mathbf{X}\boldsymbol{\beta}, \quad \mathsf{Cov}(\hat{\mathbf{y}}) = \sigma^2\mathbf{H}; \\
\mathbb{E}(\mathbf{r}) &= \mathbf{0}, \quad \mathsf{Cov}(\mathbf{r}) = \sigma^2(\mathbf{I}_n - \mathbf{H}) \\
\mathbb{E}\mathbf{r}^t\mathbf{r} &= \mathbb{E}\ \mathsf{tr}\left[\mathbf{r}^t\mathbf{r}\right] = \mathbb{E}\ \mathsf{tr}\left[\mathbf{r}\mathbf{r}^t\right] = \mathsf{tr}\left[\mathbb{E}\mathbf{r}\mathbf{r}^t\right] \\
&= \mathsf{tr}\left[\mathsf{Cov}(\mathbf{r})\right] = \sigma^2\mathsf{tr}[\mathbf{I}_n - \mathbf{H}] = (n-p)\sigma^2
\end{aligned}
$$

So

$$
\frac{1}{n-p}\mathbb{E}\mathbf{r}^t\mathbf{r} = \mathsf{RSS}/(n-p)
$$

is an unbiased estimator of σ^2.

- So the LS estimate $\hat{\boldsymbol{\beta}}$ is unbiased.

- We can plug-in the estimated error variance $\hat{\sigma}^2$ to obtain the variance estimate of $\hat{\boldsymbol{\beta}}$, i.e.,

$$\mathsf{Cov}(\hat{\boldsymbol{\beta}}) = \hat{\sigma}^2(\mathbf{X}^t\mathbf{X})^{-1}.$$

- We often use the standard error of $\hat{\boldsymbol{\beta}}$ in our later inference. For example

$$\mathsf{se}(\hat{\beta}_1) = \sqrt{\mathsf{Var}(\hat{\beta}_1)} = \hat{\sigma}\sqrt{[(\mathbf{X}^t\mathbf{X})^{-1}]_{11}}$$

where $[(\mathbf{X}^t\mathbf{X})^{-1}]_{11}$ denotes the $(1,1)$-th entry of the matrix $(\mathbf{X}^t\mathbf{X})^{-1}$.

Mean Squared Error

We often compare various estimators for a parameter θ by their Mean Squared Error (MSE):

$$\begin{aligned} \mathsf{MSE}(\hat{\theta}) &= \mathbb{E}(\theta - \hat{\theta})^2 \\ &= (\theta - \mathbb{E}(\theta))^2 + \mathbb{E}(\theta - \mathbb{E}(\theta))^2 \\ &= \mathsf{Bias}^2 + \mathsf{Var} \end{aligned}$$

If $\hat{\theta}$ is unbiased then its MSE is just equal to its variance.

Estimable Linear Combinations

Recall that if $\lambda^t\boldsymbol{\beta}$ is estimable, then there exists an $n \times 1$ vector "$\mathbf{a}$" such that $\mathbf{X}^t\mathbf{a} = \lambda$ and $\mathbf{a}^t\mathbf{y}$ is a linear and unbiased estimator of $\lambda^t\boldsymbol{\beta}$.

- There are many such "$\mathbf{a}$" vectors, i.e., $\mathbf{a}$ is not unique.
- Suppose "$\mathbf{a}$" satisfies $\mathbf{X}^t\mathbf{a} = \lambda$, so does $\hat{\mathbf{a}}$, the projection of $\mathbf{a}$ onto $C(\mathbf{X})$. This is because

$$\mathbf{X}^t\mathbf{a} = (\mathbf{HX})^t\mathbf{a} = \mathbf{X}^t\mathbf{H}^t\mathbf{a} = \mathbf{X}^t\mathbf{H}\mathbf{a} = \mathbf{X}^t\hat{\mathbf{a}}.$$

- $\mathbf{a}$ is not unique but $\hat{\mathbf{a}}$ is. Suppose $\mathbf{a}_1$ and $\mathbf{a}_2$ are two $n \times 1$ vectors satisfying $\lambda = \mathbf{X}\mathbf{a}_1 = \mathbf{X}\mathbf{a}_2$. So

$$\lambda = \mathbf{X}\hat{\mathbf{a}}_1 = \mathbf{X}\hat{\mathbf{a}}_2 \Longrightarrow \mathbf{X}(\hat{\mathbf{a}}_1 - \hat{\mathbf{a}}_2) = \mathbf{0},$$

  which implies that $(\hat{\mathbf{a}}_1 - \hat{\mathbf{a}}_2)$ is orthogonal to $C(\mathbf{X})$ but apparently it's also in $C(\mathbf{X})$, so $(\hat{\mathbf{a}}_1 - \hat{\mathbf{a}}_2) = \mathbf{0}$.

The LS Estimate of $\lambda^t\boldsymbol{\beta}$

- Suppose $\lambda^t\boldsymbol{\beta}$ is estimable, then its LS estimate is defined to be

$$\widehat{\lambda^t\boldsymbol{\beta}}_{\mathsf{LS}} = \lambda^t\hat{\boldsymbol{\beta}}_{\mathsf{LS}},$$

  where $\hat{\boldsymbol{\beta}}_{\mathsf{LS}}$ is any element from the set $\{\mathbf{v} : \mathbf{X}\mathbf{v} = \hat{\mathbf{y}}\}$.

- The LS estimate $\hat{\lambda^t\boldsymbol{\beta}}_{\mathsf{LS}}$ is unique even if $\hat{\boldsymbol{\beta}}_{\mathsf{LS}}$ is not: recall that

$$\lambda = \mathbf{X}^t\mathbf{a}, \tag{1}$$

  so

$$\lambda^t\hat{\boldsymbol{\beta}}_{\mathsf{LS}} = \mathbf{a}^t\mathbf{X}\hat{\boldsymbol{\beta}}_{\mathsf{LS}} = \mathbf{a}^t\hat{\mathbf{y}} = \mathbf{a}^t\mathbf{H}\mathbf{y} = (\mathbf{H}\mathbf{a})^t\mathbf{y} = \hat{\mathbf{a}}^t\mathbf{y},$$

  and $\hat{\mathbf{a}}$ is unique among all $\mathbf{a}$'s satisfying (1).

The Gauss-Markov Theorem

Gauss-Markov Theorem: For an estimable linear combination $\lambda^t\boldsymbol{\beta}$, its LS estimate $\lambda^t\hat{\boldsymbol{\beta}}$ is the BLUE (best linear unbiased estimator)

Proof: Suppose $\mathbf{a}^t\mathbf{y} + b$ is a linear unbiased estimator of $\lambda^t\boldsymbol{\beta}$. It is easy to compute its variance that is equal to $\sigma^2\|\mathbf{a}\|^2$.

Since it's unbiased, we have

$$\lambda^t\boldsymbol{\beta} = \mathbb{E}\mathbf{a}^t\mathbf{y} + b = \mathbf{a}^t\mathbf{X}\boldsymbol{\beta} + b,$$

which holds true for any value of $\boldsymbol{\beta}$. Therefore $b = 0$ and $\mathbf{a}^t\mathbf{X} = \lambda^t$.

Instead of directly computing the variance of the LS estimate $\lambda^t\hat{\boldsymbol{\beta}}$, we use an

alternative expression for $\lambda^t\hat{\boldsymbol{\beta}}$ which involves $\mathbf{a}$ (see the discussion on P9):

$$\lambda^t\hat{\boldsymbol{\beta}} = \mathbf{a}^t\mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{a}^t\hat{\mathbf{y}} = \mathbf{a}^t\mathbf{H}\mathbf{y} = (\mathbf{H}\mathbf{a})^t\mathbf{y} = \hat{\mathbf{a}}^t\mathbf{y}.$$

So the variance of the LS estimate $\lambda^t\hat{\boldsymbol{\beta}}$ is equal to $\sigma^2\|\hat{\mathbf{a}}\|^2$, which is smaller than $\sigma^2\|\mathbf{a}\|^2$.

That is, we can improve (still unbiased, but with smaller variance) any linear estimator $\mathbf{a}^t\mathbf{y}$ by using $\hat{\mathbf{a}}$ as the new weights on the n data points $\mathbf{y}$.

How to Compute $\widehat{\lambda^t \boldsymbol{\beta}}_{\mathbf{LS}}$?

Three methods:

$$\widehat{\lambda^t \boldsymbol{\beta}}_{\mathsf{LS}} = \mathbf{a}^t \hat{\mathbf{y}} = \hat{\mathbf{a}}^t \mathbf{y} = \lambda^t \hat{\boldsymbol{\beta}}_{\mathsf{LS}}.$$

- Since $\lambda^t \boldsymbol{\beta}$ is estimable, find one of the $\mathbf{a}$ vectors. Also compute the projection matrix $\mathbf{H}$. Then you can either use $\mathbf{a}^t \hat{\mathbf{y}}$ or $\hat{\mathbf{a}}^t \mathbf{y}$.
- Compute one $\hat{\boldsymbol{\beta}}_{\mathsf{LS}}$. For example, for the one-way ANOVA example, you can solve $(\mu, \alpha_1, \alpha_2)$ by setting $\mu = 0$. And then plug in $\hat{\boldsymbol{\beta}}_{\mathsf{LS}}$ to $\lambda^t \hat{\boldsymbol{\beta}}_{\mathsf{LS}}$.