

1. [10 points] For the general linear model $Y = X\mathbf{b} + \mathbf{e}$, with uncorrelated errors having mean zero and variance \mathbf{s}^2 , suppose that the design matrix \mathbf{X} is not necessarily of full rank. Let $\hat{Y} = X\hat{\mathbf{b}}$ denote the projection of Y onto the space spanned by the columns of \mathbf{X} , where $\hat{\mathbf{b}}$ is any OLS solution to the normal equations $X'X\mathbf{b} = X'Y$. Find the variance-covariance matrix $\Sigma_{\hat{Y}}$ of \hat{Y} . Furthermore show that $\text{Trace}(\Sigma_{\hat{Y}}) = \mathbf{s}^2 \text{rank}(X)$.

For the GLM $\mathbf{Y} = \mathbf{X}\mathbf{b} + \mathbf{e}$, $E[\mathbf{e}] = 0$, $\text{Cov}[\mathbf{e}] = ((\text{cov}(\mathbf{e}_i, \mathbf{e}_j))) = \mathbf{s}^2 \mathbf{I}$, the estimator $\hat{\mathbf{Y}} = \mathbf{X}\hat{\mathbf{b}} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} = \mathbf{P}\mathbf{Y}$, where $\mathbf{P} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ is the orthogonal projection matrix onto the space spanned by the columns of \mathbf{X} , is uniquely defined, even if the column rank(\mathbf{X}) is not full. Now,

$$\Sigma_{\hat{Y}} = \text{Cov}(\mathbf{P}\mathbf{Y}) = \mathbf{P}\text{Cov}(\mathbf{Y})\mathbf{P}' = \mathbf{s}^2 \mathbf{P}\mathbf{P}'.$$

However, since \mathbf{P} is a symmetric idempotent matrix,

$$\Sigma_{\hat{Y}} = \mathbf{s}^2 \mathbf{P}\mathbf{P}' = \mathbf{s}^2 \mathbf{P}^2 = \mathbf{s}^2 \mathbf{P}.$$

Furthermore, the eigen-values $I_i, i=1,2,\dots,p$, of \mathbf{P} are either 0 or 1 [Eigen-values of \mathbf{P}^2 are equal to the square of the eigen-values of \mathbf{P} , and $\mathbf{P}^2 = \mathbf{P}$ implies that these values must be either 0 or 1]. Now,

$$\sum_{i=1}^n \text{Var}(\hat{Y}_i) = \text{trace}[\Sigma_{\hat{Y}}] = \text{trace}[\mathbf{s}^2 \mathbf{P}] = \mathbf{s}^2 \text{trace}[\mathbf{P}] = \mathbf{s}^2 \sum_{i=1}^p I_i = \mathbf{s}^2 \text{rank}(\mathbf{X}).$$

2. [50 points] Consider a two-way ANOVA model

$$y_{ij} = \mathbf{a}_i + \mathbf{t}_j + \mathbf{e}_{ij}, i = 1, 2; j = 1, 2.$$

The parameters \mathbf{a}_i and \mathbf{t}_j are unknown.

- a. [5 points] Suppose that the errors \mathbf{e}_{ij} have mean zero, variance \mathbf{s}^2 , and are uncorrelated. Express these observations into a general linear model form $Y = X\mathbf{b} + \mathbf{e}$, where $\mathbf{b}' = (\mathbf{a}_1, \mathbf{a}_2, \mathbf{t}_1, \mathbf{t}_2)$, i.e., define the vector \mathbf{Y} and the matrix \mathbf{X} , and the covariance matrix of the error vector \mathbf{e} .

Let us stack the data points in the vector \mathbf{Y} given by

$$\mathbf{Y}' = (Y_{11}, Y_{12}, Y_{21}, Y_{22}), \text{ and let } \mathbf{e}' = (\mathbf{e}_{11}, \mathbf{e}_{12}, \mathbf{e}_{21}, \mathbf{e}_{22}), \text{ and } \mathbf{b} = (\mathbf{a}_1, \mathbf{a}_2, \mathbf{t}_1, \mathbf{t}_2).$$

$$\text{Then the design matrix } \mathbf{X} = \begin{bmatrix} 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 1 \end{bmatrix}. \text{ Furthermore, } \Sigma_{\mathbf{e}} = \mathbf{s}^2 \mathbf{I}_4.$$

- b. [10 points] Show that for this design, a parametric function $c_1\mathbf{a}_1 + c_2\mathbf{a}_2 + d_1\mathbf{t}_1 + d_2\mathbf{t}_2$ is estimable if and only if $c_1 + c_2 = d_1 + d_2$.

The parametric function $\mathbf{l}'\boldsymbol{\beta} = c_1\mathbf{a}_1 + c_2\mathbf{a}_2 + d_1\mathbf{t}_1 + d_2\mathbf{t}_2$ is estimable, iff

$$\exists \text{ a vector } \mathbf{t} \text{ such that } \mathbf{l}' = \mathbf{t}'\mathbf{X} \Leftrightarrow (c_1 \quad c_2 \quad d_1 \quad d_2) = (t_1 + t_2 \quad t_3 + t_4 \quad t_1 + t_3 \quad t_2 + t_4).$$

Therefore, $c_1 + c_2 = t_1 + t_2 + t_3 + t_4 = d_1 + d_2$.

Now suppose that $c_1 + c_2 = d_1 + d_2$ holds, then $c_1 = d_1 + d_2 - c_2$. Now

$$(c_1 \quad c_2 \quad d_1 \quad d_2) = (d_1 + d_2 - c_2 \quad c_2 \quad d_1 \quad d_2) = (d_1 - c_2 + \Delta \quad d_2 - \Delta \quad c_2 - \Delta \quad \Delta)\mathbf{X}.$$

- c. [5 points] Consider the parametric functions $\mathbf{a}_1 - \mathbf{a}_2$, $\mathbf{t}_1 - \mathbf{t}_2$ and $\mathbf{a}_1 + \mathbf{a}_2 + \mathbf{t}_1 + \mathbf{t}_2$. Show that the coefficient vectors in these parametric functions form an orthogonal basis of the row space of \mathbf{X} .

Note that these functions can be expressed as $\mathbf{K}\boldsymbol{\beta}$, where $\mathbf{K} = \begin{bmatrix} 1 & -1 & 0 & 0 \\ 0 & 0 & 1 & -1 \\ 1 & 1 & 1 & 1 \end{bmatrix}$.

Easy to check that $\mathbf{K}\mathbf{K}'$ is a diagonal matrix. Hence the rows of \mathbf{K} are orthogonal.

Furthermore, $\text{rank}(\mathbf{K})=3$, and $\mathbf{X} = \frac{1}{2} \begin{bmatrix} 1 & 1 & 1 \\ 1 & -1 & 1 \\ -1 & 1 & 1 \\ -1 & -1 & 1 \end{bmatrix} \mathbf{K}$.

Thus, each row of \mathbf{X} can be written as linear combinations of rows of \mathbf{K} . Therefore, rows of \mathbf{K} form an orthogonal basis of the row space $\mathfrak{R}[\mathbf{X}']$.

- d. [15 points] A g-inverse of the matrix $\mathbf{X}'\mathbf{X}$ for the above model is given below:

$$(\mathbf{X}'\mathbf{X})^- = \frac{1}{4} \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 4 & -2 & -2 \\ 0 & -2 & 3 & 1 \\ 0 & -2 & 1 & 3 \end{bmatrix}$$

Find the best linear unbiased estimators of three parametric functions in part (c) above and their variance covariance matrix.

Given a generalized inverse of $\mathbf{X}'\mathbf{X}$, the BLUEs of $\mathbf{K}\boldsymbol{\beta}$ are given by

$$\mathbf{K}\hat{\boldsymbol{\beta}} = \mathbf{K}[\mathbf{X}'\mathbf{X}]^- \mathbf{X}'\mathbf{Y} = \mathbf{L}\mathbf{Y}, \text{ where } \mathbf{L} = \frac{1}{2} \begin{bmatrix} 1 & 1 & -1 & -1 \\ 1 & -1 & 1 & -1 \\ 1 & 1 & 1 & 1 \end{bmatrix}.$$

Note that the three rows of \mathbf{L} are orthogonal, and the length of each row vector equals one. Hence the var-cov matrix of $\mathbf{K}\hat{\boldsymbol{\beta}} = \boldsymbol{\Sigma}_{\mathbf{K}\hat{\boldsymbol{\beta}}} = \mathbf{s}^2\mathbf{L}\mathbf{L}' = \mathbf{s}^2\mathbf{I}_3$.

e. [5 points] Is Y_{11} the best linear unbiased estimator of $\mathbf{a}_1 + \mathbf{t}_1$? Explain

No, it is not. Since $\mathbf{a}_1 + \mathbf{t}_1$ is estimable, therefore its BLUE is unique.

It is easy to check that

$$(1 \ 0 \ 1 \ 0)\hat{\mathbf{b}} = \frac{1}{4}(3Y_{11} + Y_{12} + Y_{21} - Y_{22}) \neq Y_{11}.$$

Note that the variance of the BLUE is $\frac{9}{16}\mathbf{s}^2 < \text{Var}(Y_{11})$.

f. [10 points] Consider a reduced model for this problem under the restriction $\mathbf{a}_1 - \mathbf{a}_2 = 0$. Find the difference of the ERROR Sum of Squares for the reduced model and the full model.

From Part (c) above, $(\hat{\mathbf{a}}_1 - \hat{\mathbf{a}}_2) = \frac{1}{2}(1 \ 1 \ -1 \ 1)Y$, with $\text{Var}(\hat{\mathbf{a}}_1 - \hat{\mathbf{a}}_2) = \mathbf{s}^2$. The handout on "Optimization of Error SS under Linear Restrictions on parameter vector", it is known for estimable linear restrictions,

$$\begin{aligned} \text{Error SS(Reduced model under the restriction } \mathbf{a}_1 - \mathbf{a}_2 = 0 - \text{Error SS(Full model)} \\ = \mathbf{s}^2(\hat{\mathbf{a}}_1 - \hat{\mathbf{a}}_2)^2 / \text{Var}(\hat{\mathbf{a}}_1 - \hat{\mathbf{a}}_2) = (\hat{\mathbf{a}}_1 - \hat{\mathbf{a}}_2)^2. \end{aligned}$$

3. [20 points] Assume that the 4-dimensional random vector Y follows the model $Y_i = \mathbf{m} + \mathbf{e}_i$, $i=1, 2, 3, 4$, where the errors have mean zero, and given the scalar c , the variance covariance matrix of \mathbf{e} is given by

$$\mathbf{s}^2 \begin{bmatrix} 1 & c & c & 0 \\ c & 1 & 0 & c \\ c & 0 & 1 & c \\ 0 & c & c & 1 \end{bmatrix}.$$

a) [5 points] Find all values of c for which the above matrix is a covariance matrix.

For the matrix V above to be a covariance matrix, it must be n.n.d. Thus all its eigenvalues must be non-negative. Considering the appropriate 2x2 partitioned matrices in V ,

$$\begin{aligned} |V - II| &= \begin{vmatrix} 1-I & c \\ c & 1-I \end{vmatrix} \cdot \begin{vmatrix} 1-I & c \\ c & 1-I \end{vmatrix} - \begin{bmatrix} c & 0 \\ 0 & c \end{bmatrix} \begin{bmatrix} 1-I & c \\ c & 1-I \end{bmatrix}^{-1} \begin{bmatrix} c & 0 \\ 0 & c \end{bmatrix} \\ &= \begin{vmatrix} 1-I & c \\ c & 1-I \end{vmatrix}^2 - c^2 I = \begin{vmatrix} (1-I)^2 + c^2 & 2c(1-I) \\ 2c(1-I) & (1-I)^2 + c^2 \end{vmatrix} - c^2 I = \begin{vmatrix} (1-I)^2 & 2c(1-I) \\ 2c(1-I) & (1-I)^2 \end{vmatrix} \\ &= (1-I)^4 - 4c^2(1-I)^2 = (1-I)^2((1-I)^2 - 4c^2). \end{aligned}$$

Thus the roots of the characteristic polynomial $|V - \mathbf{I}I| = 0$ are $I = 1$ (with multiplicity 2) and $1 - I = 2|c|$.

Now $I \geq 0 \Rightarrow 1 - 2|c| \geq 0 \Leftrightarrow |c| \leq \frac{1}{2}$.

b) [10 points] Find the Gauss Markov estimator $\hat{\mathbf{m}}$ of \mathbf{m} based on the vector Y .

Note that for this model the design matrix is $\mathbf{1}$, a column of 1's. Furthermore, $VX = (1 + 2c)\mathbf{1}$. Therefore, the column space of VX is same as the column space of X .

Hence the Gauss Markov estimator of $\mathbf{m} = \text{OLS of } \mathbf{m} = X\bar{Y} / X'X = \frac{1}{4} \sum_1^4 Y_i = \bar{Y}$.

c) [5 points] Find the ratio of the variances of $\hat{\mathbf{m}}$ and the OLS of \mathbf{m} .

Now $\text{Var}(\bar{Y}) = \sigma^2 \frac{(1 + 2c)}{4}$. Of course, since the two estimators are same, the ratio of their variances equals 1. Note that for $c = -\frac{1}{2}$, $\text{Var}(\bar{Y}) = 0$, and $\bar{Y} = \mathbf{m}$ with Prob 1.

4. [5 points each] Explain why each of the following statement is True or False. If you make correct choice, but provide incorrect explanation, you will not receive any credit.

a. [True/False] In a general linear model, $Y = X\mathbf{b} + \mathbf{e}$, let x_i denote the i^{th} column of X , $i = 1, 2, \dots, p$. The parametric function $c_1\mathbf{b}_1 + c_2\mathbf{b}_2$ is estimable if the vectors $\{x_i, i = 1, 2\}$ do not belong to the space spanned by the vectors $\{c_2x_1 - c_1x_2, x_3, \dots, x_p\}$.

False, but something close to this holds. Since the vectors $(c_1 \ c_2)$ and $(c_2 \ -c_1)$ are orthogonal, one can reparametrize $\mathbf{b}' = (\mathbf{b}_1, \mathbf{b}_2, \mathbf{b}_3, \dots, \mathbf{b}_p)$ by $\mathbf{b}^{*'} = (\mathbf{b}_1^*, \mathbf{b}_2^*, \mathbf{b}_3, \dots, \mathbf{b}_p)$ using $\mathbf{b}_1^* = c_1\mathbf{b}_1 + c_2\mathbf{b}_2$ and $\mathbf{b}_2^* = c_2\mathbf{b}_1 - c_1\mathbf{b}_2$. Now solve for $\mathbf{b}_1, \mathbf{b}_2$ in terms of $\mathbf{b}_1^*, \mathbf{b}_2^*$ i.e., $\begin{pmatrix} \mathbf{b}_1 \\ \mathbf{b}_2 \end{pmatrix} = \frac{1}{c_1^2 + c_2^2} \begin{bmatrix} c_1 & c_2 \\ c_2 & -c_1 \end{bmatrix} \begin{pmatrix} \mathbf{b}_1^* \\ \mathbf{b}_2^* \end{pmatrix}$. Assume, without loss of generality, that $c_1^2 + c_2^2 = 1$, and substitute for $\mathbf{b}_1, \mathbf{b}_2$ in the original model in terms of the $\mathbf{b}_1^*, \mathbf{b}_2^*$. Then first two columns in the design matrix of the reparametrized model are $c_1X_1 + c_2X_2, c_2X_1 - c_1X_2$. Now \mathbf{b}_1^* is estimable if the first column of the new matrix, i.e., $c_1X_1 + c_2X_2$ does not belong to the space spanned by the last $(p-1)$ columns of the new matrix, i.e. $c_1X_1 + c_2X_2 \notin \mathfrak{R}[c_2X_1 - c_1X_2, X_3, \dots, X_p]$. [This was a home work problem, and discussed in the class.] The key is to think of reparametrization. The problem statement said

b. [True/False] If the marginal distribution of X_1 and X_2 are normal with means zero and variance 1, then their joint distribution must be a bivariate normal distribution.

False, since the marginals do not determine the joint distribution.

- c. [True/False] In the sample model $Y_i = \mathbf{b}'x_i + \mathbf{e}_i, i = 1, 2, \dots, 4$, with errors $\{\mathbf{e}_i\}$ having mean zero, variance \mathbf{S}^2 , and pair-wise correlations \mathbf{r} , the B.L.U.E. of the parameter \mathbf{b} is given by the ratio estimator $\hat{\mathbf{b}} = \frac{\sum_{i=1}^4 Y_i}{\sum_{i=1}^4 x_i}$.

In this case, since the X does not contain the column of 1's, the GLS and OLS may be different, unless $\mathbf{r} = 0$. However, since

$$[(1 - \mathbf{r})I + \mathbf{r}J]^{-1} = c_1I + c_2J, \text{ where } c\text{'s are non-zero,}$$

$$X'V^{-1}X = c_1x'x + c_2(x'x)^2 \text{ and } X'V^{-1}Y = c_1x'y + c_2(x'1)(1'y).$$

Thus, the G-M estimator is given by their ratio. This is not equal to the ratio estimator $(1'y)/(x'1)$.

- d. [True/False] Let $c'Y$ and $d'Y$ be both BLUE of some parametric function $l'b$. Then c and d must be equal.

This is false if the function $l'b$ is not estimable, since $l'b^0 = l'(X'X)^-X'Y$ depends on the choice of g-inverse. However, if it is estimable, the BLUE is unique. Thus $c'Y = d'Y$ for all Y and hence $c = d$.