

Note: There are four problems on this Exam worth a total of 100 points.

1. [50 points] Consider a two-way cross-classified ANOVA model

$$y_{ij} = \mu + \alpha_i + \tau_j + \varepsilon_{ij}, i = 1, 2; j = 1, 2, 3.$$

The parameters  $\mu, \alpha_i$  and  $\tau_j$  are unknown.

- a. [5 points] Suppose that the errors  $\{\varepsilon_{ij}\}$  have mean zero, variance  $\sigma^2$ , and are uncorrelated. Express these observations into a general linear model form  $Y = X\beta + \varepsilon$ , where  $\beta' = (\mu, \alpha_1, \alpha_2, \tau_1, \tau_2, \tau_3)$ , i.e., define the vector  $Y$ , the matrix  $X$ , and the covariance matrix of the error vector  $\varepsilon$ .

Let us stack the data points in the vector  $Y$  given by

$$Y' = (Y_{11}, Y_{12}, Y_{13}, Y_{21}, Y_{22}, Y_{23}), \text{ and let } \varepsilon' = (\varepsilon_{11}, \varepsilon_{12}, \varepsilon_{13}, \varepsilon_{21}, \varepsilon_{22}, \varepsilon_{23}), \text{ and } \beta = (\mu, \alpha_1, \alpha_2, \tau_1, \tau_2, \tau_3).$$

$$\text{Then the design matrix } X = \begin{bmatrix} 1_3 & 1_3 & 0_3 & I_3 \\ 1_3 & 0_3 & 1_3 & I_3 \end{bmatrix}. \text{ Furthermore, } \Sigma_\varepsilon = \sigma^2 \mathbf{I}_6.$$

- b. [10 points] What is the rank,  $r$  of the matrix  $X$ ? Give a set of basis vectors for the null space  $N(X)$ .

Since, the sum of columns 2 and 3 equals column 1, as well as the sum of columns 4-6 equals column 1, the rank of  $X$  can be at most 4. Furthermore, the rank of the sub-matrix formed by rows 1-4 and columns 4-6 is 4, therefore  $\text{rank}(X) = 4$ . Therefore, dimension of  $N(X) = 6 - \text{rank}(X) = 2$ .

A set of basis vectors for  $N(X)$  consists of two linearly independent vectors satisfying  $X\gamma = \mathbf{0}$ . Since the sum of columns 2-3 equals column 1, it follows that

$\gamma^{(1)'} = [1, -1, -1, 0, 0, 0]$  satisfies this condition. Similarly, since the sum of columns 4-6 equals column 1, it follows that  $\gamma^{(2)'} = [1, 0, 0, -1, -1, -1]$  also satisfies this condition. Clearly, these vectors are linearly independent.

- c. [10 points] Give a set of  $r$  linearly independent estimable functions  $\lambda'\beta$ . Consider the parametric functions  $\alpha_1 - \alpha_2, \tau_1 - \tau_2$  and  $\tau_1 + \tau_2 - 2\tau_3$ . Do the coefficient vectors of these parametric functions form an orthogonal basis for the row space of  $X$ ?

Note that the three functions given in the problem are contrasts in  $\alpha$ 's or  $\tau$ 's. Hence, they are estimable functions for the balanced two-way ANOVA. Furthermore, since their coefficient vectors are mutually orthogonal, they are also linearly independent. Now, add another linearly independent estimable function (say)

$\mu + \alpha_1 + \tau_1$  to this set, we get four linearly independent estimable functions, as desired. Clearly, the three given functions can't form a basis for a space with dimension 4.

- d. [10 points] Write down the normal equations for this model, and find the OLS estimators of three parametric functions in part (c) above and their variance covariance matrix.

The normal equations  $[\mathbf{X}'\mathbf{X}]\boldsymbol{\beta} = \mathbf{X}'\mathbf{Y}$ , using the usual dot notations, are given by

$$\mathbf{X}'\mathbf{X} = \begin{bmatrix} 6 & 3*\mathbf{1}'_2 & 2*\mathbf{1}'_3 \\ 3*\mathbf{1}_2 & 3*\mathbf{I}_{2 \times 2} & \mathbf{J}_{2 \times 3} \\ 2*\mathbf{1}_3 & \mathbf{J}_{3 \times 2} & 2*\mathbf{I}_{3 \times 3} \end{bmatrix}, (\mathbf{X}'\mathbf{Y}) = [Y_{..} \quad Y_{1.} \quad Y_{2.} \quad Y_{.1} \quad Y_{.2} \quad Y_{.3}]'$$

It is easy to check that one solution of these equations is

$$\boldsymbol{\beta} = [\bar{Y}_{..} \quad \bar{Y}_{1.} - \bar{Y}_{..} \quad \bar{Y}_{2.} - \bar{Y}_{..} \quad \bar{Y}_{.1} - \bar{Y}_{..} \quad \bar{Y}_{.2} - \bar{Y}_{..} \quad \bar{Y}_{.3} - \bar{Y}_{..}]'$$

Thus,

$$\lambda'_1 \hat{\boldsymbol{\beta}} = \hat{\alpha}_1 - \hat{\alpha}_2 = \bar{Y}_{1.} - \bar{Y}_{2.}; \lambda'_2 \hat{\boldsymbol{\beta}} = \hat{\tau}_1 - \hat{\tau}_2 = \bar{Y}_{.1} - \bar{Y}_{.2}; \text{ and } \lambda'_3 \hat{\boldsymbol{\beta}} = \hat{\tau}_1 + \hat{\tau}_2 - 2\hat{\tau}_3 = \bar{Y}_{.1} + \bar{Y}_{.2} - 2\bar{Y}_{.3}.$$

$$\text{Now, } \begin{pmatrix} \lambda'_1 \hat{\boldsymbol{\beta}} \\ \lambda'_2 \hat{\boldsymbol{\beta}} \\ \lambda'_3 \hat{\boldsymbol{\beta}} \end{pmatrix} = \mathbf{L}\mathbf{Y}, \text{ where } \mathbf{L} = \begin{pmatrix} 1/3 & 1/3 & 1/3 & -1/3 & -1/3 & -1/3 \\ 1/2 & -1/2 & 0 & 1/2 & -1/2 & 0 \\ 1/2 & 1/2 & -1 & 1/2 & 1/2 & -1 \end{pmatrix}.$$

$$\text{Therefore, } \text{Cov} \begin{pmatrix} \lambda'_1 \hat{\boldsymbol{\beta}} \\ \lambda'_2 \hat{\boldsymbol{\beta}} \\ \lambda'_3 \hat{\boldsymbol{\beta}} \end{pmatrix} = \text{Cov}(\mathbf{L}\mathbf{Y}) = \sigma^2 \mathbf{L}\mathbf{L}', \text{ where, } \mathbf{L}\mathbf{L}' = \begin{pmatrix} 6/9 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 3 \end{pmatrix}.$$

Note: OLS estimators of these mutually orthogonal contrasts are uncorrelated.

- e. [5 points] Is  $Y_{11}$  the best linear unbiased estimator of  $\alpha_1 + \tau_1$ ? Explain

Since,  $E[Y_{11}] = \mu + \alpha_1 + \tau_1 \neq \alpha_1 + \tau_1$  for all  $\beta$ 's,  $Y_{11}$  is not even an unbiased estimator for  $\alpha_1 + \tau_1$ . Trivially, it can't be the best linear unbiased estimator.

- f. [10 points] Consider a reduced model for this problem under the restriction  $\alpha_1 - \alpha_2 = 0$ . Find the difference of the ERROR Sum of Squares for the reduced model and the full model.

Note that  $\alpha_1 - \alpha_2$  is an estimable function (a contrast), and from part (d) above,

$\hat{\alpha}_1 - \hat{\alpha}_2 = \bar{Y}_{1.} - \bar{Y}_{2.}$ ,  $\text{Var}(\hat{\alpha}_1 - \hat{\alpha}_2) = \frac{2}{3}\sigma^2$ . Therefore, from the handout on estimation under linear restrictions, the difference of the ERROR Sum of Squares for the reduced model and the full model is given by  $(\hat{\alpha}_1 - \hat{\alpha}_2)^2 / (2/3) = 1.5(\bar{Y}_{1.} - \bar{Y}_{2.})^2$ .

2. [20 points] Suppose that the experiment in Problem 1 above, was not performed as planned. Two of the experimental units were not run properly – cells (2,2) and (1,3)-those two observations became replicates for the cell (2,3). These mistakes led to the following relabeling of observations, a slight change in the model and the design matrix, but the same parametric vector:

$$y_{ijk} = \mu + \alpha_i + \tau_j + \varepsilon_{ijk}, i = 1, 2; j = 1, 2, 3; k = 1, \dots, n_{ij},$$

$$\text{where, } n_{11} = n_{12} = n_{21} = 1, n_{23} = 3, \text{ but } n_{13} = n_{22} = 0.$$

- a. [5 points] Express these observations into a general linear model form  $Y = X\beta + \varepsilon$ , where  $\beta' = (\mu, \alpha_1, \alpha_2, \tau_1, \tau_2, \tau_3)$ , i.e., define the vector  $Y$ , the matrix  $X$ , and the covariance matrix of the error vector  $\varepsilon$ .

Let us stack the data points in the vector  $Y$  given by

$$Y' = (Y_{111}, Y_{121}, Y_{211}, Y_{231}, Y_{232}, Y_{233}), \text{ and let } \varepsilon' = (\varepsilon_{111}, \varepsilon_{121}, \varepsilon_{211}, \varepsilon_{231}, \varepsilon_{232}, \varepsilon_{233}), \text{ and } \beta = (\mu, \alpha_1, \alpha_2, \tau_1, \tau_2, \tau_3). \text{ Then the design matrix } X \text{ is given by}$$

$$X = \begin{bmatrix} 1 & 1 & 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 & 0 & 1 \end{bmatrix}. \text{ Furthermore, } \Sigma_{\varepsilon} = \sigma^2 \mathbf{I}_6.$$

- b. [5 points] What is the rank of the matrix  $X$ ? Give a set of basis vectors for the null space  $N(X)$ .

Since the rows 4-6 of  $X$  are identical,  $\text{rank}(X)$  is at most 4. Furthermore, the sub-matrix defined by rows 1-4, columns 3-6 is of rank 4. Thus  $\text{rank}(X) = 4$ . The discussion in Problem 1(b) above shows that the null space  $N(X)$  for this unbalanced two way design matrix is same as the null space of the balanced two way design matrix, and two vectors given in the solution of Problem 1(b) above form a basis of the null space.

- c. [10 points] For all the parametric functions listed in Problem 1(c), determine whether they are estimable or not.

Since the null spaces of the two design matrices are same, it follows that the row spaces of the two matrices are same. Thus these three parametric functions are estimable.

3. [15 points] For the general linear model  $Y = X\beta + \varepsilon$ , with uncorrelated errors having mean zero and variance  $\sigma^2$ , suppose that the design matrix  $X$  is not necessarily of full rank. Let  $\hat{Y} = X\hat{\beta}$  denote the vector of fitted values, where  $\hat{\beta}$  is any OLS solution to the normal equations  $X'X\beta = X'Y$ . Let  $e = Y - \hat{Y}$  denote the vector of estimated residuals. Find the variance-covariance matrix  $\Sigma_e$  of  $e$ . Furthermore show that  $\sum_{i=1}^n \text{Var}(e_i) = \sigma^2\{n - \text{rank}(X)\}$ .

For the GLM,  $\hat{Y} = X\hat{\beta} = X(X'X)^{-1}X'Y = PY$ , where  $P = X(X'X)^{-1}X'$  is the orthogonal projection matrix onto the column space  $C(X)$ , is uniquely defined, even if the column rank( $X$ ) is not full. Now,  $e = Y - \hat{Y} = (I - P)Y$

$$\Sigma_e = \text{Cov}[(I - P)Y] = (I - P)\text{Cov}(Y)(I - P) = \sigma^2(I - P)I(I - P).$$

However, since  $P$  is a symmetric idempotent matrix,  $(I - P)$  is also a symmetric idempotent matrix. Thus,

$$\Sigma_e = \sigma^2(I - P)(I - P) = \sigma^2(I - P).$$

Furthermore, the eigen-values  $\lambda_i, i = 1, 2, \dots, p$ , of  $P$  are either 0 or 1 [Eigen-values of  $P^2$  are equal to the square of the eigen-values of  $P$ . However  $P^2 = P$  implies that  $\lambda_i, i = 1, 2, \dots, p$ , must be 0 or 1]. Therefore,  $\text{trace}[P] = \sum_{i=1}^p \lambda_i = \text{rank}(P) = \text{rank}(X)$ .

Now,

$$\sum_{i=1}^n \text{Var}(e_i) = \text{trace}[\Sigma_e] = \sigma^2 \text{trace}[(I - P)] = \sigma^2[n - \text{trace}(P)] = \sigma^2(n - \text{rank}(X)).$$

4. [15 points] Consider two general linear models for the same training data:

- i.  $Y = X\beta + \varepsilon$ , where  $X$  is a  $n \times p$  matrix, and
- ii.  $Y = W\gamma + \varepsilon$ , where  $W$  is a  $n \times t$  matrix with full column rank.

Suppose that  $C(X) = C(W)$ , i.e., the two models are *reparametrizations* of each other. Thus there exist matrices  $S$  and  $T$  such that  $W = XT$  and  $X = WS$ . Suppose that we have  $T$ , give an easy way to find  $S$ .

First of all, since the column rank of  $W$  is full,  $W'W$  is a non-singular matrix, and its inverse exists. Now,  $X = WS \Rightarrow W'X = W'WS \Leftrightarrow S = [W'W]^{-1}W'X$ .

Thus, even if  $T$  is not given,  $S$  can be found directly. [Note that  $W$  itself is not invertible.] In addition, one can also argue,

$$\begin{aligned} W = XT &\Rightarrow W'W = W'XT = W'WST \\ &\Rightarrow ST = I, \text{ since } W'W \text{ is invertible.} \end{aligned}$$

Now, given  $T$ , one can solve the systems of equations  $ST = I$  for the variables in  $S$ .