Exam 1

Advanced Linear Models (PHST 781)
Department of Biostatistics and Bioinformatics

12^{th} October, 2018

Name:

- There are are a total of 115 points in this Question Paper. Answer as much as you can. If your acquired score is greater than equal to 100, it will be counted as 100%.

- The Exam is scheduled for 3 hours. "Time Left" reminders will be posted in 1.5 hrs, 2:30 hrs, 2:45 hrs from the beginning of the Exam time.

- There are three $\star$ marked problems that are more involved than the rest. In case you are stuck in one of those, it might be a good idea to consider solving other problems first and then continue with the * marked problems.

- You may take help from the "Exam Assistance Note" containing a few required definitions, lemma and theorem statements.

1. (a)

Let Y be Response variable, X_1, X_2 denote the Explanatory variables, ε be unknown random errors and $\beta_1, \ldots, \beta_3$ are unknown parameters of interest. Determine whether the following relationship equation is a linear model. Relationship Equation: $Y = \beta_0 + \beta_1 e^{X_1} + \beta_2 X_2 + \beta_3 X_1 X_2 + \varepsilon$.

Score:
Total Score: 5

Ans:

☐ Not Linear Model ☐ Linear Model

(b)

Identify if the following matrix is a Orthogonal Projection matrix.

$$M = \frac{1}{9}\begin{bmatrix} 3 & 2 & 4 \\ 2 & 1 & 2 \\ 4 & 2 & 4 \end{bmatrix}$$

Score:
Total Score: 5

Ans:

☐ Orthogonal Projection ☐ Not Orthogonal Projection

(c)

Consider the matrices $A = \begin{bmatrix} 2 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix}$. Does the matrix A have a Generalized Inverse ? If yes, Construct a generalized inverse of A.

Score:
Total Score: 1+5

Ans: The matrix A have a generalized inverse.

The generalized inverse of A is $A^{-} = \begin{bmatrix} \frac{1}{2} & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix}$.

(d) Consider the matrix

$$D = \frac{1}{2}\begin{bmatrix} 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 \end{bmatrix}.$$

i. Represent D as a Kronecker product of two lower-dimensional matrices.
ii. Prove that D is a **Orthogonal Projection** matrix.

Score: ______
Total Score: 3+7

Ans:

(e) Consider the model $\underset{\sim}{\mathbf{y}} = \mathbf{X}\underset{\sim}{\beta} + \underset{\sim}{\varepsilon}$, where $\underset{\sim}{\beta} = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix}$, $\mathbf{X} = \begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 1 \\ 0 & 1 & 0 \\ -1 & 0 & -1 \end{bmatrix}$ and $\underset{\sim}{\varepsilon}$ is a mean zero error vector with variance co-variance matrix $\sigma^2 I_{4\times 4}$. Prove that the parameter β_1 is not estimable

Score: ______

Total Score: 7

Ans:

2. Let $\underset{\sim}{\mathbf{x}}_1 = \begin{bmatrix}1 & 1 & \dots & 1\end{bmatrix}^T_{2n\times 1}$ and $\underset{\sim}{\mathbf{x}}_2 = \begin{bmatrix}1 & -1 & 1 & \dots & -1\end{bmatrix}^T_{2n\times 1}$, i.e. the i^{th} entry of the vector $\underset{\sim}{\mathbf{x}}_2$ is $(-1)^{i-1}$ for $i = 1, \dots, 2n$. Consider the linear model

$$\underset{\sim}{\mathbf{Y}} = \mathbf{X}_{2n\times 2}\underset{\sim}{\beta} + \underset{\sim}{\varepsilon}_{2n\times 1},$$

where the two columns of the design matrix $\mathbf{X}$ are $\underset{\sim}{\mathbf{x}}_1$ and $\underset{\sim}{\mathbf{x}}_2$, $\underset{\sim}{\beta} = \begin{bmatrix}\beta_1 & \beta_2\end{bmatrix}^T$. Furthermore, assume that $\underset{\sim}{\varepsilon}_{2n\times 1} \sim N(\underset{\sim}{\mathbf{0}}, \sigma^2 I_{2n\times 2n})$. $\underset{\sim}{\mathbf{Y}}$ denotes the random vector corresponding to the **response variable**. Let a statistician is interested in inference for the parametric function $\theta = \beta_1 - \beta_2$

(a) Is θ a linear parametric function of $\underset{\sim}{\beta}$? If yes, find a vector $\underset{\sim}{\lambda}_1$ such that $\theta = \underset{\sim}{\lambda}_1^T \underset{\sim}{\beta}$

Score: ______
Total Score: 1+ 4

Ans:

(b) Is θ estimable parametric function? provide appropriate justification to your answer.

Score: ______
Total Score: 1+ 6

Ans:

(c) Find the **Orthogonal Projection Matrix** for $\mathscr{C}(\mathbf{X})$, the column space of $\mathbf{X}$. (Justify your steps with the reference to the results/theorem/lemma you are using.)

Score: ____
Total Score: 8

Ans:

Find the **Best Linear Unbiased Estimator** for θ? (Show your steps and justify your steps with the reference to the results/theorem/lemma you are using.)

(d) Score: ______
Total Score: 10

Ans:

3. Consider the data set

Response:	15.8	20.1	19.1	16.7	14.3	19.1
Explanatory Variable:	7	6	4	6.5	3	2

.

A statistician believes that the data is being generated from a piece-wise linear model specified as,

$$Y = \begin{cases} \alpha_0 + \alpha_1 X + \varepsilon & \text{if } X \leq 5 \\ \gamma_0 + \gamma_1 X + \varepsilon & \text{if } X > 5. \end{cases} \quad (1)$$

(a) Represent the above model in terms of the matrix form of the standard linear model

$$\underset{\sim}{\mathbf{y}} = \mathbf{X}\underset{\sim}{\beta} + \underset{\sim}{\varepsilon},$$

by explicitly specifying (In terms of numbers when known and in terms of appropriate symbols when unknown) the following quantities a) $\mathbf{X}$, the design Matrix, b) $\underset{\sim}{\mathbf{y}}$, the response vector and Vector of regression coefficient, $\underset{\sim}{\beta}$.

Score: ______
Total Score: 6+1+1

Ans:

(b) The model defined in Equation (1) is not guaranteed to be continuous. Show that the fitted model will be continuous if the constraint $\alpha_0 - \gamma_0 = 5\gamma_1 - 5\alpha_1$ is imposed on the regression coefficients. Is it a linear constraint on the parameters?

Score: ______
Total Score: 2+1

Ans:

(c) $\star$ Construct a model

$$\underset{\sim}{\mathbf{y}}_\star = \mathbf{X}_\star \underset{\sim}{\beta}_\star + \underset{\sim}{\varepsilon},$$

incorporating the constraint in part(b). Specify (In terms of numbers when known and in terms of appropriate symbols when unknown) the following quantities a) $\mathbf{X}_\star$, the design Matrix, b) $\underset{\sim}{\mathbf{y}}_\star$, the response vector and c) Vector of regression coefficient, $\underset{\sim}{\beta}_\star$.

Score: ______
Total Score: 7+1+3

Ans: (You have more space in page 10 to write answer to this question.)

4. Consider a linear model given by

$$\mathbf{Y} = \mathbf{X}\underset{\sim}{\beta} + \underset{\sim}{\varepsilon}$$

where $\mathbf{Y} \in \mathbb{R}^n, \underset{\sim}{\beta} \in \mathbb{R}^p$ and $\underset{\sim}{\varepsilon}$ is a Normally distributed random vector with mean $\underset{\sim}{\mathbf{0}}$ and variance $\sigma^2 I_{n\times n}$ and $\mathbf{X}$ is an $n \times p$ matrix with rank $r < p < n$ (i.e. The design matrix $\mathbf{X}$ does not have full column rank). The parameter σ^2 is an unknown positive number. Assume that $\underset{\sim}{\varepsilon} \sim N\left(\underset{\sim}{\mathbf{0}}, \sigma^2 I_{n\times n}\right)$. Let $\mathscr{C}(\mathbf{X})$ denotes the column space of $\mathbf{X}$, the vector space containing all possible linear combination of the column vectors of the matrix $\mathbf{X}$. It can be shown that $P = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-}\mathbf{X}^T$ is the **Orthogonal Projection** matrix for the $\mathscr{C}(\mathbf{X})$, where $(\mathbf{X}^T\mathbf{X})^{-}$ is any Generalized inverse of $(\mathbf{X}^T\mathbf{X})$.

(a) Show that $E(\mathbf{Y}^T(I_{n\times n} - P)\mathbf{Y}) = (n-r)\sigma^2$.

Score: ______
Total Score: 5

Ans:

(b) Prove that the statistics $\mathbf{Y}^T\mathbf{P}\mathbf{Y}$ and $\mathbf{Y}^T(\mathbf{I}_{n\times n}-\mathbf{P})\mathbf{Y}$ are independent (Mention if you are using any result).

Score: ____
Total Score:5

Ans:

Derive the distribution of $\frac{(\mathbf{Y}^T P\mathbf{Y})/r}{(\mathbf{Y}^T(I_{n\times n}-P)\mathbf{Y})/(n-r)}$? (Show your steps)

(c)

Score:
Total Score: 10

Ans:

(d) $\star$ Let $\widehat{\underset{\sim}{\beta}}_{LSE}$ be the **Least Square Estimator** for the parameter $\underset{\sim}{\beta}$. Show that $(\underset{\sim}{\mathbf{Y}} - \mathbf{X}\underset{\sim}{\beta})^T P(\underset{\sim}{\mathbf{Y}} - \mathbf{X}\underset{\sim}{\beta}) = (\widehat{\underset{\sim}{\beta}}_{LSE} - \underset{\sim}{\beta})^T \mathbf{X}^T \mathbf{X}(\widehat{\underset{\sim}{\beta}}_{LSE} - \underset{\sim}{\beta})$.

Score:

Total Score: 5

Ans:

(e) $\star$ Derive the distribution of $\frac{\|X(\widehat{\underset{\sim}{\beta}}_{LSE} - \underset{\sim}{\beta})\|^2}{\sigma^2}$. (You may use the relation in part(d) to get your answer.)

Score:
Total Score: 5

Ans:

5. Consider the linear model when the observed response variable

$$Y_{i,j} = \mu + \tau_i + \varepsilon_{i,j} \text{ for } i = 1,2,\dots 5; j = 1,2,\dots 3,$$

where $\varepsilon_{i,j} \overset{iid}{\sim} N(0,\sigma^2)$ and $\sigma^2, \mu, \tau_1, \dots, \tau_5$ are unknown parameters of the model. μ is called the 'baseline effect' or 'mean effect' while $\tau_1, \dots, \tau_5$ are called 'treatment effects'. Consider the notations,

$$\underset{\sim}{\mathbf{Y}} = \left[\overbrace{Y_{1,1}, \dots, Y_{1,3}}^{1^{st}\text{ Treatment}} \,\vdots\, \dots \,\vdots\, \overbrace{Y_{i,1}, \dots Y_{i,3}}^{i^{th}\text{ Treatment}} \,\vdots\, \dots \,\vdots\, \overbrace{Y_{5,1} \dots Y_{5,3}}^{5^{th}\text{ Treatment}} \right]^T,$$

$$\underset{\sim}{\beta} = \left[\mu, \tau_1, \dots, \tau_5\right]^T = \left[\mu \quad \underset{\sim}{\tau}^T\right]^T, \text{ with } \underset{\sim}{\tau}^T = \left[\tau_1, \dots, \tau_5\right],$$

$$\underset{\sim}{\varepsilon} = \left[\overbrace{\varepsilon_{1,1}, \dots, \varepsilon_{1,3}}^{1^{st}\text{ Treatment}} \,\vdots\, \dots \,\vdots\, \overbrace{\varepsilon_{i,1}, \dots \varepsilon_{i,3}}^{i^{th}\text{ Treatment}} \,\vdots\, \dots \,\vdots\, \overbrace{\varepsilon_{5,1} \dots \varepsilon_{5,3}}^{5^{th}\text{ Treatment}} \right]^T,$$

$$\underset{\sim}{\mathbf{1}}_5 = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}_{5\times 1}, \underset{\sim}{\mathbf{1}}_3 = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}_{3\times 1} \underset{\sim}{\mathbf{0}} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}_{3\times 1} \text{ and}$$

$$\mathbf{X} = \begin{bmatrix} \underset{\sim}{\mathbf{1}}_3 & \underset{\sim}{\mathbf{1}}_3 & \underset{\sim}{\mathbf{0}} & \dots & \underset{\sim}{\mathbf{0}} \\ \underset{\sim}{\mathbf{1}}_3 & \underset{\sim}{\mathbf{0}} & \underset{\sim}{\mathbf{1}}_3 & \dots & \underset{\sim}{\mathbf{0}} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \underset{\sim}{\mathbf{1}}_3 & \underset{\sim}{\mathbf{0}} & \dots & \dots & \underset{\sim}{\mathbf{1}}_3 \end{bmatrix}_{15\times 6} = \left[\underbrace{\underset{\sim}{\mathbf{1}}_5 \otimes \underset{\sim}{\mathbf{1}}_3} \,\vdots\, \underbrace{I_5 \otimes \underset{\sim}{\mathbf{1}}_3} \right] = \left[X_\mu \,\vdots\, X_\tau \right],$$

where $X_\mu = \underset{\sim}{\mathbf{1}}_5 \otimes \underset{\sim}{\mathbf{1}}_3$, $X_\tau = I_5 \otimes \underset{\sim}{\mathbf{1}}_3$, then we can represent the model as

$$\underset{\sim}{\mathbf{Y}} = \mathbf{X}\underset{\sim}{\beta} + \underset{\sim}{\varepsilon}.$$

(a) Consider a linear parametric function $\underset{\sim}{\alpha}^T \underset{\sim}{\tau}$ where $\underset{\sim}{\alpha} \in \mathbb{R}^5$. **Prove that, if $\underset{\sim}{\alpha}^T \underset{\sim}{\tau}$ is estimable then it must be a contrast between the treatments.**

(b) Let a linear parametric function $\underset{\sim}{\alpha}^T \underset{\sim}{\tau}$ is a contrast between the treatment effects for some $\underset{\sim}{\alpha}^T \in \mathbb{R}^5$. **Prove that $\underset{\sim}{\alpha}^T \underset{\sim}{\tau}$ is estimable.**

(c) Note that $\tau_1 - \tau_2$ estimable as $\tau_1 - \tau_2$ is a contrast between the treatments.

Derive $\widehat{\tau_1 - \tau_2}$ the Best Linear Unbiased Estimator for $\tau_1 - \tau_2$ and express your answer in terms of $\overline{Y}_{i,\bullet}$ for $i = 1, \dots 5$.

$$\text{It is given that } P_X \underset{\sim}{\mathbf{Y}} = \begin{bmatrix} \overline{Y}_{1,\bullet} \\ \vdots \\ \overline{Y}_{5,\bullet} \end{bmatrix} \otimes \underset{\sim}{\mathbf{1}}_3 \text{ where } \overline{Y}_{i,\bullet} = \frac{1}{3}\sum_{j=1}^{3} Y_{i,j} \text{ for } i = 1, \dots 5.$$

Here P_X denotes the Orthogonal Projection matrix for the column space of $\mathbf{X}$

HINT: You may have already constructed a vector $\underset{\sim}{\mathbf{a}} \in \mathbb{R}^{15}$ such that $\begin{bmatrix} 0 & \underset{\sim}{\alpha}^T \end{bmatrix} = \underset{\sim}{\mathbf{a}}^T \mathbf{X}$.

(d) What is P_X (mention if you are using any nontrivial result to derive P_X)?
Derive the variance of $\widehat{\tau_1 - \tau_2}$?

(e) What is the value of the constant ν so that

$$\hat{\sigma}^2 = \frac{\underset{\sim}{\mathbf{Y}}^T (I - P_X) \underset{\sim}{\mathbf{Y}}}{\nu} = \frac{\sum_{i=1}^{5} \sum_{j=1}^{3} Y_{i,j}^2 - 3 \sum_{i=1}^{5} \overline{Y}_{i,\bullet}^2}{\nu}$$

is unbiased estimator of σ^2. (You do not need to prove)

(f) How many unique primary contrasts of the treatments $\tau_1, \ldots, \tau_5$ are there ?

(Note: primary contrasts are the contrasts of the form $\tau_i - \tau_j$ for $1 \le i \neq j \le 5$. Note that , $\tau_i - \tau_j$ and $\tau_j - \tau_i$ are assumed to be 'same' primary contrast for a pair of distinct indices $i \neq j$.)

(g) Write down the definition of the **simultaneous confidence intervals** for all the unique primary contrasts.

(h) Construct a 95% Bonferroni's simultaneous confidence intervals for all the unique primary contrasts above.