Design of Experiment: Prelim Problems
May 2025

Please Do All Problems. Each of the 14 parts carries an equal weight of 10 points.

1. Four types of fuel additives (factor A with fixed effects) for cars are compared to determine which is best to reduce nitrogen oxides. Four types of cars (factor B with fixed effects) and several randomly selected drivers (factor C with random effects) can be used in the experiment. Consider the following 4 scenarios and write down the statistical linear model and related assumptions. Assume there are no interaction effects.

   a) In a complete randomized design, for each combination of the four types of fuel additives and the four types of cars, the car is driven for five minutes twice and the nitrogen oxides level are measured for each time. There are 32 observations in total. The order of the runs of expreiments are all random. The gas tanks are completely cleaned between each driving.

   b) In a randomized complete block design,the experiment in 1(a) is repeated in random order for four randomly selected drivers (factor C with random effects).

   c) Suppose 32 randomly selected drivers are available to participate this experiment. They are equally randomly assigned to 16 combinations of the four types of fuel additives and the four types of cars. Each driver drives the car for five minutes twice and the nitrogen oxides level are measured for each time. There are 64 observations in total. The order of the runs of expreiments are all random.

   d) Suppose 4 drivers can participate the study. In a 4×4 Latin square design, the 4 drivers and 4 types of cars are the row and column factors and the fuel additive type is the treatment factor.

2. Cheese is made by bacterial fermentation of Pasteurized milk. Most of the bacteria are purposefully added to do the fermentation; these are the starter cultures. Some "wild" bacteria are also present in cheese; these are the nonstarter bacteria. One hypothesis is that nonstarter bacteria may affect the quality of a cheese, so that otherwise identical cheese making facilities produce different cheeses due to their different indigenous nonstarter bacteria.

   Two strains of nonstarter bacteria were isolated at a premium cheese facility: R50#10 and R21#2. We will add these nonstarter bacteria to cheese to see if they affect quality. Our four treatments (factor A with fixed effects) will be control (level 1), addition of R50 (level 2), addition of R21 (level 3), and addition of a blend of R50 and R21 (level 4). Twelve cheeses (factor B with random effects) are made, three for each of the four treatments, with the treatments being randomized to the cheeses. Each cheese is then divided into four portions, and the four portions for each cheese are randomly assigned to one of four aging times(factor C with fixed effects): 1 day, 28 days, 56 days, and 84 days. Each portion is measured for total free amino acids (a measure of bacterial activity) after it has aged for its specified number of days (data from Peggy Swearingen). There are 48 observations in total. Denote y_{ijk} as the total free amino acids for $j^{th}cheese$ of i^{th} treatment with k^{th} aging time. The linear model is $y_{ijk} = \mu + \alpha_i + B_{j(i)} + \gamma_k + \alpha\gamma_{ik} + \epsilon_{ijk}$, $i = 1, 2, 3, 4$, $j = 1, 2, 3$, $k = 1, 2, 3, 4$.

Treatment	Average total free amino acids
Control	1.57
R50	1.66
R21	1.67
R50+R21	1.97

   a) Write down the assumptions for the unrestricted model and explain each term in the linear model.

   b) Complete the following ANOVA table.

Source	d.f.	MS	E(MS)
A	_____	_____	_____
B(A)	_____	0.13	_____
C	_____	15.05	_____
A*C	_____	0.11	_____
Error	_____	0.07	_____
Total	_____		

   c) Write down a contrast for interaction effect between addition of R50 and R21. Construct a 95% confidence interval for the contrast.

   d) Conduct hypothesis testing for teatment effects at 0.05 significance level.

   e) Estimate the variance for random effect of cheese. Construct a 95% confidence interval.

   f) Derive the formula for E(Mean Squares for Treatment)

3. The effect of five different ingredients (A, B, C, D, E) on the reaction time in minutes of a chemical process is being studied. Each batch of new material is only large enough to permit five runs to be made. Furthermore, each run requires approximately 1.5 hours, so only five runs can be made in one day. The experimenter decides to run the experiment as repeated Latin squares so that day and batch effects may be systematically controlled. The design is as follows. Denote y_{lijk} as the reaction time for the j^{th}day using k^{th} ingredient and i^{th} batch of material in the l^{th} replicate. $y_{lijk} = \mu + \theta_l + A_{i(l)} + B_{j(l)} + \gamma_k + \epsilon_{lijk}$. $l = 1, 2$, $i = 1, 2, 3, 4, 5$, $j = 1, 2, 3, 4, 5$, $k = 1, 2, 3, 4, 5$. μ is the baseline for assembly time. θ_l is the random effect of l^{th} replicate. $A_{i(l)}$ is the random effect of i^{th} batch of material in the l^{th} replicate. $B_{j(l)}$ is the random effect of j^{th} day in the l^{th} replicate. γ_k is the fixed effect of k^{th} ingredient.

Replicated 1

	Day				
Batch	1	2	3	4	5
1	A	B	D	C	E
2	C	E	A	D	B
3	B	A	C	E	D
4	D	C	E	B	A
5	E	D	B	A	C

Replicated 2

	Day				
Batch	6	7	8	9	10
6	C	D	E	A	B
7	B	C	D	E	A
8	E	A	B	C	D
9	A	B	C	D	E
10	D	E	A	B	C

a) Please complete the assumptions of the unrestricted model.

b) Complete the following ANOVA table.

Source	d.f.	MS	E(MS)
Replicate	_____	2.5	_____
Batch(Replicate)	_____	3.9	_____
Day(Replicate)	_____	3.1	_____
Ingredient	_____	35.4	_____
Error	_____	3.1	_____
Total	_____		

c) Construct hypothesis testing for ingredient effect on reaction time at 0.05 significance level.

d) Derive the formula for E(mean square for Batch(Replicate))

Applied Statistics Preliminary Examination
Theory of Linear Models
May 2025

Instructions:

- Do all 3 Problems. Neither calculators nor electronic devices of any kind are allowed. Show all your work, clearly stating any theorem or fact that you use. Each of the 12 parts carries an equal weight of 10 points.
- Abbreviations/Acronyms.
  - IID (independent and identically distributed).
  - LSE (least squares estimator); BLUE (best linear unbiased estimator). Sometimes the LSE may be designated OLS (ordinary least squares) estimator, in order to differentiate it from the GLS (generalized least squares) estimator.
- Notation.
  - $\boldsymbol{x}^T$ or $\boldsymbol{A}^T$: indicates transpose of vector $\boldsymbol{x}$ or matrix $\boldsymbol{A}$.
  - $\text{tr}(\boldsymbol{A})$ and $|\boldsymbol{A}|$: denotes the trace and determinant, respectively, of matrix $\boldsymbol{A}$.
  - $\boldsymbol{I}_n$: the $n \times n$ identity matrix.
  - $\boldsymbol{j}_n = (1, \ldots, 1)^T$ is an n-vector of ones, and $\boldsymbol{J}_{m,n}$ is an $m \times n$ matrix of ones.
  - $\mathbb{E}(\boldsymbol{x})$ and $\mathbb{V}(\boldsymbol{x})$: expectation and variance of random vector $\boldsymbol{x}$.
  - $\boldsymbol{x} \sim N_m(\boldsymbol{\mu}, \boldsymbol{\Sigma})$: the m-dimensional random vector $\boldsymbol{x}$ has a normal distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$.
  - $X \sim t(n, \lambda)$: a t distribution with n degrees of freedom and noncentrality parameter λ. If $\lambda = 0$ we write simply: $X \sim t(n)$.
  - $X \sim F(n_1, n_2, \lambda)$: an F distribution with n_1 and n_2 numerator and denominator degrees of freedom respectively, and noncentrality parameter λ. If $\lambda = 0$ we write simply: $X \sim F(n_1, n_2)$.
- Possibly useful results.
  - For a generic linear model $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, where the design matrix $\boldsymbol{X}$ is of dimensions $n \times k$, and $\widehat{\boldsymbol{\beta}}$ is the LSE of $\boldsymbol{\beta}$, the R^2 is defined as:

$$R^2 = \left(\widehat{\boldsymbol{\beta}}^T \boldsymbol{X}^T \boldsymbol{y} - n\overline{y}^2\right) / S_{yy}, \qquad \overline{y} = \sum_{i=1}^{n} y_i/n, \qquad S_{yy} = \sum_{i=1}^{n} (y_i - \overline{y})^2.$$

  We also establish the following notation: if x_{ij} is the (i, j)-th element of the design matrix $\boldsymbol{X}$, let $\overline{x}_j = \sum_{i=1}^{n} x_{ij}/n$ be the mean of the j-th column.

1. Let $\boldsymbol{X}_k$ be a full-rank $n \times k$ matrix whose column space $C(\boldsymbol{X}_k)$ consists of the linear span of the vectors $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_k\}$, i.e., $C(\boldsymbol{X}_k) = \mathcal{L}(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_k)$, and let $\boldsymbol{X}_{k+1}$ be a full-rank $n \times (k+1)$ matrix whose column space consists of the linear span of the vectors that constitute $C(\boldsymbol{X}_k)$ and the extra vector $\boldsymbol{x}_{k+1}$, so that $C(\boldsymbol{X}_{k+1}) = \mathcal{L}(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_k, \boldsymbol{x}_{k+1})$. In addition, let $\boldsymbol{G}_k = (\boldsymbol{X}_k^T \boldsymbol{X}_k)^{-1}$, $\boldsymbol{G}_{k+1} = (\boldsymbol{X}_{k+1}^T \boldsymbol{X}_{k+1})^{-1}$, $\boldsymbol{P}_k = \boldsymbol{X}_k \boldsymbol{G}_k \boldsymbol{X}_k^T$, and $\boldsymbol{P}_{k+1} = \boldsymbol{X}_{k+1} \boldsymbol{G}_{k+1} \boldsymbol{X}_{k+1}^T$.

   (a) Show that $\boldsymbol{P}_k$ is the projection matrix onto $C(\boldsymbol{P}_k)$.

   (b) Show that $C(\boldsymbol{P}_k) = C(\boldsymbol{X}_k)$.

   (c) Show that $\boldsymbol{P}_{k+1} - \boldsymbol{P}_k$ is a projection matrix.

   (d) Is $C(\boldsymbol{P}_k)$ orthogonal to $C(\boldsymbol{P}_{k+1} - \boldsymbol{P}_k)$? Justify.

2. Consider the two full-rank linear regression models for the same response vector $\boldsymbol{y} = (y_1, \ldots, y_n)^T$ with sample mean $\overline{y}$ and sums of squares about the mean S_{yy} (see these definitions in the "Possibly useful results" of page 1):

$$\boldsymbol{y} = \boldsymbol{X}_k \boldsymbol{\beta} + \boldsymbol{\epsilon}, \tag{1}$$

   and

$$\boldsymbol{y} = \boldsymbol{X}_{k+1} \boldsymbol{b} + \boldsymbol{e}, \tag{2}$$

   where the design matrices are as defined in Problem 1, so that $C(\boldsymbol{X}_k) \subset C(\boldsymbol{X}_{k+1})$. As the notation suggests, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_k)^T$ and $\boldsymbol{b} = (b_1, \ldots, b_{k+1})^T$ are the corresponding coefficient vectors of appropriate dimension, with associated LSEs $\widehat{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{b}}$. The ultimate goal of this problem is to prove that, if R_k^2 and R_{k+1}^2 are the R^2 values for models (1) and (2) respectively, then R_{k+1}^2 cannot be smaller than R_k^2. (To simplify the problem we also assume that the first column of each design matrix consists of a vector of 1's, i.e., $\boldsymbol{x}_1 = \boldsymbol{j}_n$.) In what follows, x_{ij} is the (i,j)-th element of the design matrix $\boldsymbol{X}_k$, and $\overline{x}_j = \sum_{i=1}^n x_{ij}/n$ is the mean of the j-th column (see "Possibly useful results" of page 1).

   (a) Show that $\overline{y} = \sum_{j=1}^k \hat{\beta}_j \overline{x}_j$.

   (b) If x_{ij} is the (i,j)-th element of $\boldsymbol{X}_k$, and $\overline{x}_j = \sum_{i=1}^n x_{ij}/n$ is the mean of the j-th column, show that $S_{yy} R_k^2 = \sum_{j=1}^k \sum_{i=1}^n \hat{\beta}_j (x_{ij} - \overline{x}_j) y_i$.

   (c) By considering the difference $S_{yy}(R_{k+1}^2 - R_k^2)$ and noting the results of Problem 1, or otherwise, show that $R_k^2 \leq R_{k+1}^2$.

3. Consider the observations $\{y_{ij}\}$ from the following linear model on 6 distinct subjects:

$$y_{ij} = \alpha_i + (\mu + \gamma_i) x_{ij} + \epsilon_{ij}, \qquad i = 1, 2, \quad j = 1, 2, 3,$$

   where x_{ij} is the age of subject j whose gender is indicated by the index i. Let $\boldsymbol{\beta}^T = (\alpha_1, \alpha_2, \mu, \gamma_1, \gamma_2)$, assume that all subjects have different ages, and that $\{\epsilon_{ij}\} \sim$ IID $\mathrm{N}(0, \sigma^2)$.

   (a) Express the model in the usual vector-matrix form, $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, write down the design matrix $\boldsymbol{X}$, and determine its rank.

   (b) Is the hypothesis $H_0 : \alpha_1 = \alpha_2 = \mu = 0$ testable? Carefully justify your answer.

   (c) It is desired to test if the model can be reduced to common intercepts ($\alpha_1 = \alpha_2$) and slopes ($\gamma_1 = \gamma_2$) for the two genders. Carefully express this as an appropriate null hypothesis, and show that it is testable.

   (d) Propose a test statistic for the test in (c) and specify its distribution under H_0. Give detailed explanations and provide numerical values for all the quantities that can actually be calculated.