# Applied Statistics Preliminary Examination
# Theory of Linear Models
# May 2021

**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 14 parts carries an equal weight of 10 points, except that 3(b), 3(c), and 3(d) are worth 12 points each.
- 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 $\boldsymbol{A}$.
  - $\mathrm{tr}(\boldsymbol{A})$ 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.
  - $\mathbb{E}(X)$ and $\mathbb{V}(X)$: expectation and variance of random variable $X$.
  - $\boldsymbol{x} \sim N_m(\boldsymbol{\mu}, \boldsymbol{\Sigma})$: 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 $\lambda$. 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 $\lambda$. If $\lambda = 0$ we write simply: $X \sim F(n_1, n_2)$.
- Possibly useful results.
  - If $\boldsymbol{x} \sim N_m(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ is given in partitioned form as

$$\boldsymbol{x} = \begin{pmatrix} \boldsymbol{x}_1 \\ \boldsymbol{x}_2 \end{pmatrix}, \qquad \boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{pmatrix}, \qquad \boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{pmatrix},$$

    with $m_1 = \dim(\boldsymbol{x}_1)$, then the conditional distribution of $\boldsymbol{x}_1$ given $\boldsymbol{x}_2$ is

$$\boldsymbol{x}_1 | \boldsymbol{x}_2 \sim N_{m_1}\left(\boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\boldsymbol{x}_2 - \boldsymbol{\mu}_2),\ \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}\right).$$

1. Let the scalar $y$ and the $p$-dimensional vector $\boldsymbol{x}$ be jointly multivariate normally distributed as follows:

$$\begin{pmatrix} y \\ \boldsymbol{x} \end{pmatrix} \sim \mathrm{N}\left( \begin{pmatrix} \mu_y \\ \boldsymbol{\mu}_x \end{pmatrix}, \begin{pmatrix} \sigma_y^2 & \boldsymbol{\sigma}_{xy}^T \\ \boldsymbol{\sigma}_{xy} & \boldsymbol{\Sigma}_{xx} \end{pmatrix} \right).$$

   (a) Show that the conditional mean of $y$ given $\boldsymbol{x}$ takes on the form of the linear regression:

$$\mathbb{E}(y|\boldsymbol{x}) = a + \boldsymbol{\beta}^T(\boldsymbol{x} - \boldsymbol{b}),$$

   and find the value of the scalar $a$, the vector $\boldsymbol{b}$, and the coefficient vector $\boldsymbol{\beta}$.

   (b) Define the *mean squared error* (MSE) due to regression as $\sigma_\epsilon^2 := \mathbb{E}(y - \mathbb{E}(y|\boldsymbol{x}))^2$, and find an expression for it in terms of the elements of the covariance matrix of $(y, \boldsymbol{x})^T$.

   (c) If the *theoretical* $R^2$ is defined as the proportion of the variance of $y$ that is accounted for by $\mathbb{E}(y|\boldsymbol{x})$, i.e., $\rho_{y|\boldsymbol{x}}^2 := (\sigma_y^2 - \sigma_\epsilon^2)/\sigma_y^2$, show that

$$\rho_{y|\boldsymbol{x}}^2 = \frac{1}{\sigma_y^2} \boldsymbol{\sigma}_{xy}^T \boldsymbol{\Sigma}_{xx}^{-1} \boldsymbol{\sigma}_{xy}.$$

   (d) Define the transformed regression vector $\boldsymbol{z} := \boldsymbol{R}\boldsymbol{x}$, where $\boldsymbol{R}$ is a $p \times p$ full-rank orthonormal matrix, i.e., $\boldsymbol{R}\boldsymbol{R}^T = \boldsymbol{R}^T\boldsymbol{R} = \boldsymbol{I}_p$. Find the (joint) distribution of $(y, \boldsymbol{z})^T$.

   (e) Show that $\mathbb{E}(y - \mathbb{E}(y|\boldsymbol{x}))^2 = \mathbb{E}(y - \mathbb{E}(y|\boldsymbol{z}))^2$, and hence deduce that the theoretical $R^2$ is unchanged by the transformation in (d). That is, show that $\rho_{y|\boldsymbol{z}}^2 = \rho_{y|\boldsymbol{x}}^2$.

2. Consider the vector of $n = 5$ observations $\boldsymbol{y} = (y_1, \ldots, y_5)^T$ from the linear model $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, where the full-rank matrix $\boldsymbol{X} = (\boldsymbol{x}_1, \boldsymbol{x}_2)$ consists of the columns vectors $\boldsymbol{x}_1 = (x_{1,1}, \ldots, x_{5,1})^T$ and $\boldsymbol{x}_2 = (x_{1,2}, \ldots, x_{5,2})^T$, $\boldsymbol{\beta} = (\beta_1, \beta_2)^T$, and $\boldsymbol{\epsilon} \sim \mathrm{N}(\boldsymbol{0}, \sigma^2 \boldsymbol{I}_5)$. Define the following summary statistics:

$$s_{11} = \sum_{i=1}^5 x_{i,1}^2, \quad s_{12} = \sum_{i=1}^5 x_{i,1}x_{i,2}, \quad s_{22} = \sum_{i=1}^5 x_{i,2}^2, \quad t_1 = \sum_{i=1}^5 x_{i,1}y_i, \quad t_2 = \sum_{i=1}^5 x_{i,2}y_i.$$

   (a) Derive an expression for $\hat{\boldsymbol{\beta}} = (\hat{\beta}_1, \hat{\beta}_2)^T$, the LSE of $\boldsymbol{\beta}$, in terms of the above summary statistics.

   (b) Find an expression for the covariance matrix $\mathrm{var}(\hat{\boldsymbol{\beta}})$ of the LSE, and hence calculate the correlation coefficient between $\hat{\beta}_1$ and $\hat{\beta}_2$.

   (c) Give an expression for the unbiased estimate of $\sigma^2$, and compute the coefficient of determination ($R^2$) for the fitted model.

   (d) Construct a level $\alpha$ test to determine if $\boldsymbol{x}_2$ is needed in the model that already has $\boldsymbol{x}_1$. That is, is it sufficient to have the model matrix consist of the single column vector $\boldsymbol{X} = \boldsymbol{x}_1$?

3. Consider the two-factor linear model $y_{ij} = \mu + \alpha_i + \beta_j + \epsilon_{ij}$, where $i = 1, 2, 3$ and $j = 1, 2, 3$, the $\epsilon_{ij}$ are IID $N(0, \sigma^2)$, but not all combinations $(i, j)$ of factors A and B are observed. There are 3 cases to consider, each defined by the matrices below, where an asterisk $(*)$ in the $(i, j)$ entry indicates that $y_{ij}$ was observed. Thus, for example, in Case 1 the following vector of responses is observed: $\boldsymbol{y} = (y_{11}, y_{12}, y_{21}, y_{22}, y_{23}, y_{32}, y_{33}, )^T$.

Case 1

| | | |
|---|---|---|
| * | * | |
| * | * | * |
| | * | * |

Case 2

| | | |
|---|---|---|
| * | * | |
| | * | * |
| | | * |

Case 3

| | | |
|---|---|---|
| * | * | |
| * | * | |
| | | * |

Our aim in this Problem is to make inference on all *estimable* contrasts of factor A and factor B main effects, i.e., all differences of parameters of the form $\alpha_i - \alpha_{i'}$ for $i \neq i'$, and $\beta_j - \beta_{j'}$ for $j \neq j'$. Recall that in the generic linear model $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, the linear combination $\boldsymbol{\lambda}^T\boldsymbol{\beta}$ is defined to be estimable if there exists a constant vector $\boldsymbol{a}$ such that $\mathbb{E}(\boldsymbol{a}^T\boldsymbol{y}) = \boldsymbol{\lambda}^T\boldsymbol{\beta}$. Also, denote by $k$ the rank of the model matrix, i.e., $k = \mathrm{rk}(\boldsymbol{X})$.

(a) In Case 1, provide an easy argument to show that $3 \leq k \leq 6$. Explicitly find the value of $k$.

(b) In Case 1, and by finding an appropriate vector $\boldsymbol{a}$ in each case, show that all $\alpha_i - \alpha_{i'}$ and $\beta_j - \beta_{j'}$ contrasts are estimable.

(c) In Case 2, determine which $\alpha_i - \alpha_{i'}$ and $\beta_j - \beta_{j'}$ contrasts are estimable. For each contrast that is estimable, find an appropriate vector $\boldsymbol{a}$.

(d) In Case 3, determine which $\alpha_i - \alpha_{i'}$ and $\beta_j - \beta_{j'}$ contrasts are estimable. For each contrast that is estimable, find an appropriate vector $\boldsymbol{a}$.

(e) In Case 1, determine if the hypothesis stated below is *testable*, carefully justifying your answer. If it is testable, show how to construct a test statistic for it, and state the distribution of the test statistic under $H_0$:

$$H_0 : \alpha_1 - \alpha_2 = 1 \quad \text{and} \quad \alpha_1 - \alpha_3 = 2 \quad \text{and} \quad \alpha_2 - \alpha_3 = 3.$$

# Design of Experiment: Prelim Problems

May 2021

Please Do All Problems. Each of the 16 parts carries an equal weight of 10 points.

1. Consider a paper manufacturer who is interested in three different pulp preparation methods (factor A) and four different cooking temperatures (factor B) for the pulp and who wishes to study the effect of these two factors on the tensile strength of the paper. The experimenter has decided to run 3 replicates (factor C if needed). Please write down the statistical models for each of the following designs. Assume factor A has fixed effect and factor B has random effect. Replicates have random effects. Interaction effects between factors A and B are the only interaction effects.

   a) Suppose the order of the experimentation is completely randomized within each of the three replicates.

   b) Suppose in each replicate, a batch of pulp is produced by one of the three methods under study. Then this batch is divided into four samples, and each sample is cooked at one of the four temperatures. Then a second batch of pulp is made up using another of the three methods. This second batch is also divided into four samples that are tested at the four temperatures. Then a third batch of pulp is made up using the remaining method. This third batch is also divided into four samples that are tested at the four temperatures. The process is then repeated, until all three replicates of the experiment are obtained.

   c) Suppose the experimenter decides to conduct the experiment as follows. A batch of pulp is produced by one of the three methods under study. Then this batch is divided into 12 samples, and each sample is cooked at one of the four temperatures so that each temperature is randomly applied to 3 samples. Then a second batch of pulp is made up using another of the three methods. This second batch is also divided into 12 samples that are tested at the four temperatures. Then a third batch of pulp is made up using the remaining method. This third batch is also divided into 12 samples that are tested at the four temperatures.

2. A fishing line experiment was in order to compare the strengths of two brands of fishing line exposed to two different levels of stress. Two different reels of fishing line were purchased for each of the two brands, and sections of line were cut from each reel. Thus the reels were automatically assigned to the levels of factor A (Brand), and constituted the four whole plots. There were no blocks in this experiment. The split plots constituted sixteen sections of line, four cut from each of the four reels (that is, 16 split plots in total, 4 per whole plot). The split plots were randomly assigned to two different stress levels so that each stress level was assigned two split plots per whole plot. The stress level 1 was induced by hanging half of a brick from the assigned section of line for 14 hours. The stress level 2 was induced by hanging one brick from the assigned section of line for 14 hours. Although this did not precisely mimic the stress induced during fishing, it was still expected to give some information about the strength of the lines under stress. The strength test was accomplished by hanging a bucket on the end of the line, which was suspended from a beam. The bucket was gradually filled with water through a small hole in the lid until the line broke. The data are the resulting weights of water to the nearest 0.01 lb. Assume Brand has fixed effects, stress has random effects, and there are no interaction effects. Denote $y_{iujt}$ as the weight of water when the line broke for $u^{th}$ whole plot of $i^{th}$ brand and $t^{th}$ split plot of $j^{th}$ stress level. $\overline{y}_{1...} - \overline{y}_{2...} = 0.20lb$, $\overline{y}_{..1.} - \overline{y}_{..2.} = 0.10lb$.

   a) Please write down a model for this experiment. Clearly define all the terms and state all relevant assumptions.

   b) Fill in the ANOVA table below.

| Source | d.f. | MS | E(MS) |
|---|---|---|---|
| | | | |
| Error(W) | | 0.1 | |
| Total(W) | | | |
| | | | |
| Error(S) | | 0.05 | |
| Total(S) | | | |

   c) State and conduct statistical tests for Brand effect at 0.05 significance level, and interpret the test results.

   d) State and conduct statistical tests for Stress effect at 0.05 significance level, and interpret the test results.

   e) Calculate E(Mean Square for Brand) and E(MSE(W)).

   f) Construct 95% confidence interval for the effect difference of the two brands.

   g) If we want to estimate the Brand effect more accurately compared to Stress effect, what adjustment do you want to make about the experimental design.

3. An experiment was planned to examine whether different brands of golf balls travel on average the same distances when hit by amateur golfers. The experiment was planned with a specific selection of $v = 3$ golf brands of balls and some number $b$ of golfers to be determined. The experiment was to be run as a general complete block design with fixed treatment effects and random golfer effects. Since the golfer is aware of which brand of ball he or she is hitting, there may well be a golfer×brand interaction. However, the differences between brands averaged over the interaction is important here. A small pilot experiment was conducted. There were only two golfers, and each hit s = 6 balls of each brand in a random order. Mis-hits were ignored. The distances that the balls traveled were recorded in yards. $Y_{\theta it}$ denotes the distance for the $t^{th}$ ball of $i^{th}$ brand hit by the $\theta^{th}$ golfer. The average distances for the balls of the three brands are $\overline{y}_{.1.} = 211.9167$, $\overline{y}_{.2.} = 217.1667$, $\overline{y}_{.3.} = 205.75$, respectively.

   a) Write a linear model equation appropriate for analyzing the responses $Y_{\theta it}$ from this experiment, clearly state the assumptions and conditions for this model.

   b) Fill in the ANOVA table below.

| source | d.f. | MS | E(MS) |
|---|---|---|---|
| golfer | _____ | 770.22 | _____ |
| brand | _____ | _____ | _____ |
| _____ | _____ | 198.028 | _____ |
| error | _____ | 96.2222 | _____ |
| total | _____ | | |

   c) Use the pilot experiment data to calculate a 95% upper bound for the error variance $\sigma^2$.

   d) Use the pilot experiment data to calculate a 95% upper bound for the variance component for golfer.

   e) Test whether the variance component for brand× golfer interaction is higher than the variance of random error at 0.05 significance level.

   f) The experimenter wanted the main experiment to be able to calculate a set of simultaneous 95% confidence intervals for the pairwise differences in the brands, and he wanted the widths of these intervals to be at most 20 yards. Assuming that each golfer would hit about 18 balls in total, as in the pilot experiment, how many randomly selected golfers would be needed?