Stat 5100 **Assignment 10**

Due: For practice only.

The following questions are courtesy of Dr. Nettleton. Consider Questions 1 – 4 for practice for the Final Exam. Consider the remaining questions practice for the PhD Qualifying exam.

1. Let $(\boldsymbol{X}^\top\boldsymbol{X})^-$ be any generalized inverse of $\boldsymbol{X}^\top\boldsymbol{X}$. A generalized inverse of a symmetric matrix is not necessarily symmetric. Thus, we cannot assume that

$$\left[(\boldsymbol{X}^\top\boldsymbol{X})^-\right]^\top = \left[(\boldsymbol{X}^\top\boldsymbol{X})^\top\right]^-$$

always holds. Find a matrix $\boldsymbol{X}$ such that $\left[(\boldsymbol{X}^\top\boldsymbol{X})^-\right]^\top \neq \left[(\boldsymbol{X}^\top\boldsymbol{X})^\top\right]^-$.

However, it is also true that a symmetric generalized inverse can always be found for a symmetric matrix.

2. For each of the following special cases, derive the REML estimator of σ^2.

   (a) Suppose $y_1, y_2, y_3 \overset{iid}{\sim} \mathcal{N}(\mu, \sigma^2)$.

   (b) Suppose

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \mu_1 \\ \mu_1 \\ \mu_2 \\ \mu_2 \end{bmatrix}, \begin{bmatrix} \sigma^2 & \sigma^2/2 & 0 & 0 \\ \sigma^2/2 & \sigma^2 & 0 & 0 \\ 0 & 0 & \sigma^2 & \sigma^2/2 \\ 0 & 0 & \sigma^2/2 & \sigma^2 \end{bmatrix} \right).$$

3. Suppose 100 maize genotypes were assigned to 304 plots in a field using an unbalanced completely randomized design in which some genotypes were assigned to only one plot while others were assigned to as many as six plots. Plots were planted with seed from their assigned genotypes, and yield in bushels per acre was recorded for each plot at the end of the growing season. The dataset is available in Canvas.

   Consider the model

$$y_{ij} = \mu + g_i + e_{ij},$$

   where $\mu + g_i$ is the mean yield for the ith genotype, and $e_{ij} \sim \mathcal{N}(0, \sigma_e^2)$ for all i and j, with independence among all e_{ij} terms.

   (a) Find the BLUE of $\mu + g_i$ for each $i = 1, \ldots, 100$.

   (b) For this and all subsequent parts of this problem, assume $g_1, \ldots, g_{100} \overset{iid}{\sim} \mathcal{N}(0, \sigma_g^2)$ and independent of all the e_{ij} terms. Find the REML estimates of σ_g^2 and σ_e^2.

   (c) Find the BLUP of $\mu + g_i$ for each $i = 1, \ldots, 100$.

   (d) Make a plot of the BLUPs (vertical axis) vs. the BLUEs from part (a) (horizontal axis) with one point for each genotype. Add the $y = x$ line to your plot. Explain why the plot looks the way it does.

(e) According to the BLUEs from part (a), list the top five highest yielding genotypes.

(f) According to the BLUPs, list the top five highest yielding genotypes.

(g) Why is the top-yielding genotype according to the BLUEs from part (a) not so highly rated according to the BLUPs?

4. *This is a repeated measures analysis.* An experiment was designed to compare the effect of three drugs (A, B, and C) on the heart rate of women. Fifteen women were randomly assigned to the drugs using a completely randomized design with five women for each drug. The heart rate (in beats per minute) of each woman was measured at 0, 5, 10, and 15 minutes after the drug was administered. The data are provided in the file HeartRate.txt.

   Let y_{ijk} denote the heart rate at the kth time point for the jth woman treated with the ith drug.

   Suppose

   $$y_{ijk} = \mu_{ik} + \epsilon_{ijk},$$

   where μ_{ik} is an unknown constant for each combination of $i = 1, 2, 3$ and $k = 1, 2, 3, 4, 5$ and ϵ_{ijk} is a normally distributed error term with mean 0 for all $i = 1, 2, 3$, $j = 1, 2, 3, 4, 5$, and $k = 1, 2, 3, 4$. For all $i = 1, 2, 3$ and $j = 1, 2, 3, 4, 5$, let

   $$\boldsymbol{\epsilon}_{ij} = (\epsilon_{ij1}, \epsilon_{ij2}, \epsilon_{ij3}, \epsilon_{ij4})^\top.$$

   Suppose all the $\boldsymbol{\epsilon}_{ij}$ vectors are mutually independent, and let $\boldsymbol{W}$ be the variance-covariance matrix of $\boldsymbol{\epsilon}_{ij}$, which is assumed to be the same for all $i = 1, 2, 3$ and $j = 1, 2, 3, 4, 5$.

   (a) Find the REML estimate of $\boldsymbol{W}$ under the assumption that $\boldsymbol{W}$ is a positive definite, compound symmetric matrix.

   (b) Find AIC and BIC for the case where $\boldsymbol{W}$ is a positive definite, compound symmetric matrix.

   (c) Find the REML estimate of $\boldsymbol{W}$ under the assumption that $\boldsymbol{W}$ is a positive definite matrix with constant variance and an AR(1) correlation structure.

   (d) Find AIC and BIC for the case where $\boldsymbol{W}$ is a a positive definite matrix with constant variance and an AR(1) correlation structure.

   (e) Find the REML estimate of $\boldsymbol{W}$ under the assumption that $\boldsymbol{W}$ is a positive definite, symmetric matrix.

   (f) Find AIC and BIC for the case where $\boldsymbol{W}$ is a positive definite, symmetric matrix.

   (g) Which of the three structures for $\boldsymbol{W}$ is preferred for this dataset?

   (h) Using the preferred structure for $\boldsymbol{W}$, compute a 95% confidence interval for the mean heart rate 10 minutes after treatment with drug A minus the mean heart rate 10 minutes after treatment with drug B.

   (i) Using the preferred structure for $\boldsymbol{W}$, compute a 95% confidence interval for the mean heart rate 10 minutes after treatment with drug A minus the mean heart rate 5 minutes after treatment with drug A.

5. Consider a class of 50 students. Suppose each student is required to take two midterm exams and one final exam. The midterms are coded as 1 and 2, and the final exam is coded as 3 in the dataset available in a data set called ExamScores.txt.

In the dataset, student 1 has scores for only exams 1 and 2. Suppose student 1 was not able to take the final exam due to a medical emergency. For $i = 1, \ldots, 50$ and $j = 1, 2, 3$, let y_{ij} be the score for student i on exam j. For $i = 1, \ldots, 50$ and $j = 1, 2, 3$, suppose

$$y_{ij} = s_i + \mu_j + e_{ij}, \tag{1}$$

where μ_j is an unknown parameter, $s_i \sim \mathcal{N}(0, \sigma_s^2)$, $e_{ij} \sim \mathcal{N}(0, \sigma_j^2)$ for some unknown variance parameter $\sigma_j^2 > 0$, and all s_i and e_{ij} terms are independent. For this model, the variance of the error terms is not constant and instead depends on the exam. This model may be fit to the data using the R code

```
library(nlme)
lme(score ~ 0 + exam, random = ~ 1 | student,
                      weights = varIdent(form = ~ 1 | exam),
data = d)
```

(This code assumes the data are in a data.frame `d`, where `exam` and `student` are factors. Be careful with cutting and pasting from this pdf to R as some characters (e.g., $\sim$) may not translate properly.)

(a) Use R (or SAS if you prefer) to find REML estimates of σ_s^2, σ_1^2, σ_2^2, and σ_3^2.

(b) Use R (or SAS if you prefer), to find the EBLUP of student 1's exam 3 score.

(c) Find an expression for $\mathrm{E}(\mathrm{y}_{13}|\mathrm{y}_{11}, \mathrm{y}_{12})$ in terms of model (1) parameters.

(d) Compute an estimate of $\mathrm{E}(\mathrm{y}_{13}|\mathrm{y}_{11}, \mathrm{y}_{12})$ by replacing parameters with their estimates and replacing y_{11} and y_{12} with their observed values.

(e) For $i = 1, \ldots, 50$, let $\boldsymbol{y}_i = [y_{i1}, y_{i2}, y_{i3}]^\top$, and suppose $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_{50} \overset{iid}{\sim} \mathcal{N}([\mu_1, \mu_2, \mu_3]^\top, \boldsymbol{W})$, where

$$\boldsymbol{W} = \begin{bmatrix} w_{11} & w_{12} & w_{13} \\ w_{12} & w_{22} & w_{23} \\ w_{13} & w_{23} & w_{33} \end{bmatrix}$$

is an unknown, positive definite variance-covariance matrix. Repeat parts (c) and (d) for this new model.

(f) Multiple linear regression could also be used to predict the missing exam 3 score for student 1. Using the data from students 2 through 50, fit a multiple linear regression of exam 3 score on exam 1 score and exam 2 score. Use an intercept and assume additive effects for exam 1 score and exam 2 score. Provide the estimated regression equation and provide a prediction for the exam 3 score of student 1 based on the estimated regression equation.

[*Note: This problem has asked for three predictions of the exam 3 score for student 1. Two of the three predictions should be the same.*]

6. Let y_{i1} denote the weight (in kg) gained by a calf i from birth to 1 week of age. Let y_{i2} denote the weight (in kg) gained by calf i from 1 week of age to 12 weeks of age. Suppose it is known that

$$\mathrm{Var}(\mathrm{y}_{\mathrm{i}1}) = \mathrm{Var}(\mathrm{y}_{\mathrm{i}2}) = 4 \text{ for all i}$$

and that the correlation between y_{i1} and y_{i2} is 0.5 for all i. Suppose weight gained by any one calf is independent of the weight gained by any other calf. Suppose the following information is available for three randomly selected calves. (Note that the periods in the table denote data that are missing completely at random.)

Calf Number (i)	y_{i1}	y_{i2}
1	51	54
2	48	.
3	52	.

(a) Determine the best linear unbiased estimator of the expected total weight gained by a calf from birth to 12 weeks of age.

(b) Predict the weight gains that are missing from the table above; i.e., predict y_{22} and y_{32}.

7. An experiment was conducted at 15 research stations around the country to determine how dose of a chemical mixture affects the leaf area of a certain type of plant. Instructions for creating the chemical mixture and for carrying out the experiment were sent to the managers at each of the 15 research stations. At each station, a completely randomized design was used to assign 5 doses of the chemical (0, 25, 50, 75, and 100 mL/day) to 20 plants with 4 plants per dose. Each plant grew in its own pot and received its assigned chemical dose each day of the experiment. At three weeks of age, the leaf area of each plant was recorded. The data for this experiment are available in the file

http://dnett.github.io/S510/LeafArea.txt

For $i = 1, \ldots, 15$, $j = 1, \ldots, 5$, and $k = 1, \ldots, 4$, let y_{ijk} be the leaf area for the kth plant that received dose j in research station i, and suppose

$$y_{ijk} = (\beta_1 + b_{1i}) + (\beta_2 + b_{2i})x_j + e_{ijk} \tag{2}$$

In this model (2), β_1 and β_2 are unknown parameters, $x_1 = 0$, $x_2 = 25$, $x_3 = 50$, $x_4 = 75$, $x_5 = 100$, the e_{ijk} terms are $iid\ N(0, \sigma_e^2)$, and the b_{1i} and b_{2i} terms are normal random effects independent of the e_{ijk} terms. More specifically, let

$$\boldsymbol{b}_i = \begin{bmatrix} b_{1i} \\ b_{2i} \end{bmatrix} \text{ for all } i = 1, \ldots, 15.$$

We assume $\boldsymbol{b}_1, \ldots, \boldsymbol{b}_{15} \overset{iid}{\sim} N(\boldsymbol{0}, \boldsymbol{\Sigma_b})$ for some positive definite 2×2 variance matrix $\boldsymbol{\Sigma_b}$.

Model (2) is a special case of what is sometimes referred to as a *random coefficient model* because the regression coefficients are assumed to be random variables rather than fixed parameters. It is straightforward to fit such a model in `R` using code like the following.

```
d = read.delim("http://dnett.github.io/S510/LeafArea.txt")
library(lme4)
o = lmer(LeafArea ~ Dose + (1 + Dose | ResearchStation), data = d)
```

The approximate BLUEs of β_1 and β_2 can be obtained with code like

```
fixef(o)
```

As usual, the estimated variance of the estimator of the fixed effects parameters is given by `vcov(o)`.

The empirical BLUPs of b_{1i} and b_{2i} can be obtained with code like

`ranef(o)`

Typing `summary(o)` provides you with enough information to determine the REML estimates of σ_e^2 and $\boldsymbol{\Sigma_b}$. The estimate of the matrix $\boldsymbol{\Sigma_b}$ is not provided directly, but you can compute it from the given estimates of the variances and the provided estimate of the correlation between b_{1i} and b_{2i} labeled `Corr` in the `Random effects` portion of the output.

(a) Provide the REML estimate of σ_e^2.

(b) Provide the REML estimate of $\boldsymbol{\Sigma_b}$.

(c) Make a scatterplot of leaf area vs. dose for the data from the 7th research station. Add a black line to the plot that shows the estimate of the regression function $\beta_1 + \beta_2 x$ for $x \in (0, 100)$.

(d) Find the prediction of the regression function for the 7th research station; i.e., predict $(\beta_1 + b_{17}) + (\beta_2 + b_{27})x$ for $x \in (0, 100)$.

(e) Using only the data from the 7th research station, find the ordinary least squares estimate of the regression function for the simple linear regression of leaf area on dose of the chemical.

(f) To the plot in part (c), add a red line that shows the regression function predicted in part (d) and a blue line that shows the regression function estimated in part (e).

(g) Compute the likelihood ratio statistic for testing $H_0 : \beta_2 = 0$.

(h) Find AIC for the fit of model (2) to the data from all 15 research stations.

(i) Find AIC for a simplified version of model (2) that assumes there is one slope coefficient common to all research stations.

(j) Find AIC for a simplified version of model (2) that assumes there is one intercept coefficient common to all research stations and one slope coefficient common to all research stations.

(k) According to AIC, which model is preferred among model (2), the model considered in part (i), and the model considered in part (j).

8. Model (2) from problem 1 can be written in linear mixed-effects model form as $\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{Zu} + \boldsymbol{e}$. Define $\boldsymbol{X}$, $\boldsymbol{\beta}$, $\boldsymbol{Z}$, $\boldsymbol{u}$, $\boldsymbol{G} = \mathrm{Var}(\boldsymbol{u})$, and $\boldsymbol{R} = \mathrm{Var}(\boldsymbol{e})$ using terms from model (2). Assume the response vector $\boldsymbol{y}$ is ordered as in the dataset LeafArea.txt.

9. In 1846, a group of pioneers traveling west became stranded in the eastern Sierra Nevada mountains. By the time the last survivor was rescued in the spring of 1847, 40 of 87 members in the original group had died from starvation and exposure to extreme cold. The group became known as the Donner Party. The dataset

http://dnett.github.io/S510/Donner.txt

contains the age, sex, and status (survived or died) of the members of the group that were 15 years of age or older. Conduct an analysis of this data set to determine how age and sex are associated with the probability of survival. Support your answer with appropriate tests and/or confidence intervals. State your conclusions in ways that will be easily interpretable by nonstatisticians.

10. Consider an experiment with three treatments (A, B, and C). Suppose there are 10 experimental units for each of treatments A and B. Suppose there are 50 experiment units for treatment C. Imagine that the response for each experimental unit has a binomial distribution with $m = 20$ trials (same for all experiment units) and a success probability that depends on treatment. Suppose that (unknown to the researcher) the success probabilities for treatments A, B, and C are 0.5, 0.5, and 0.95, respectively. Rather than using logistic regression for analysis, a researcher decides to use a standard three-treatment ANOVA assuming a normal response [`lm(y~trt)`]. The researcher is primarily interested in a comparison of treatments A and B, so he examines the `R` output for the coefficients to see if the "trtB" (the name R would use) coefficient is significant because he knows that provides a test for the difference in treatment A and B means due to the set first to zero constraints. Explain why this might not be a safe analysis strategy.