# STAT 510 Homework 3

**Due Date:** 11:00 A.M., Wednesday, February 1

1. Case Study 5.1.1 from *The Statistical Sleuth* describes a dietary restriction study. Female mice were assigned to one of the following six treatment groups:

   **(1)** NP: unlimited, nonpurified, standard feed

   **(2)** N/N85: normal diet before weaning and normal diet (85 kcal/week) after weaning

   **(3)** N/R50: normal diet before weaning and reduced calorie (50 kcal/week) after weaning

   **(4)** R/R50: reduced calorie diet before and after weaning (50 kcal/week)

   **(5)** N/R50 lopro: normal diet before weaning, reduced calorie (50 kcal/week) after weaning, and reduced protein

   **(6)** N/R40: normal diet before weaning and severely reduced calorie (40 kcal/week) after weaning

   The response of interest was mouse lifetime in months. Download the corresponding data file at http://www.statisticalsleuth.com/ or access it by installing and loading the R package `Sleuth3` and examining `case0501`. Complete the following parts under the assumption that a Gauss-Markov model with normal errors and an unrestricted mean for each of the six treatment groups is appropriate for these data.

   (a) Create side-by-side boxplots of the response for this dataset, with one boxplot for each treatment group. Be sure to clearly label the axes of your plot.

   (b) Find the SSE (sum of squared errors) for the full model with one unrestricted mean for each of the six treatment groups.

   (c) Compute $\hat{\sigma}^2$ for the full model.

   (d) Find the SSE for a reduced model that has one common mean for the *N/R50* and *N/R50 lopro* treatment groups and an unrestricted mean for each of the other four treatment groups.

   (e) Use the answers from parts (b) through (d) to compute an $F$ statistic for testing the null hypothesis that mean of the response vector is in the column space associated with the reduced model vs. the alternative that the mean of the response vector is in the column space of the full model but not in the column space of the reduced model.

   (f) Explain to the scientists conducting this study what the $F$ statistic in part (e) can be used to test. Consider the context of the study (i.e., pay attention to the description of the experiment and the descriptions of the treatments) and use terms non-statistician scientists will understand.

   (g) Consider an $F$ statistic of the form given on slide 18 of the notes entitled *A Review of Some Key Linear Model Results*. Provide the $\boldsymbol{C}$ matrix and $\boldsymbol{d}$ vector and compute the $F$ statistic corresponding to the test of the hypotheses in part (e).

2. Provide an example that shows that a generalized inverse of a symmetric matrix need not be symmetric. (*Comment: For this reason, we cannot assume that* $(\boldsymbol{X}'\boldsymbol{X})^- = [(\boldsymbol{X}'\boldsymbol{X})^-]'$.)

3. A useful result from linear algebra (that you may use in STAT510 without proof) is as follows:

$$\text{rank}(\boldsymbol{UV}) \leq \min\{\text{rank}(\boldsymbol{U}), \text{rank}(\boldsymbol{V})\}$$

   for any two matrices $\boldsymbol{U}$ and $\boldsymbol{V}$ with dimensions that allow multiplication (number of columns a $\boldsymbol{U}$ equals the number of rows of $\boldsymbol{V}$). In words, this result says that the rank of a product of matrices is no greater than the rank of any matrix in the product. Show the following:

(a) For any matrix $\boldsymbol{X}$, $\text{rank}(\boldsymbol{X}) = \text{rank}(\boldsymbol{X}'\boldsymbol{X})$.

(b) For any matrix $\boldsymbol{X}$, $\text{rank}(\boldsymbol{X}) = \text{rank}(\boldsymbol{P_X})$. (*Comment: We have already shown in a previous homework assignment that the column spaces of $\boldsymbol{X}$ and $\boldsymbol{P_X}$ are the same. This implies* $\text{rank}(\boldsymbol{X}) = \text{rank}(\boldsymbol{P_X})$*, but try to prove this result about ranks using the result about the rank of a matrix product.*)

(c) Let $\boldsymbol{X}$ be an $n \times p$ matrix. Suppose $\boldsymbol{C}$ is a $q \times p$ matrix of rank $q$. Suppose there exists a matrix $\boldsymbol{A}$ such that $\boldsymbol{C} = \boldsymbol{AX}$. Then $\boldsymbol{C}(\boldsymbol{X}'\boldsymbol{X})^{-}\boldsymbol{C}'$ is a $q \times q$ matrix of rank $q$. (*Comment: We need this result to guarantee the existence of* $[\boldsymbol{C}(\boldsymbol{X}'\boldsymbol{X})^{-}\boldsymbol{C}']^{-1}$ *on slide 18 of slide set 2.*)

(d) If $\boldsymbol{X}$ is an $n \times p$ matrix and $\boldsymbol{A}$ is a matrix with $n$ columns satisfying $\boldsymbol{AP_X} = \boldsymbol{A}$, then $\text{rank}(\boldsymbol{AX}) = \text{rank}(\boldsymbol{A})$.

4. An experiment was conducted to study the durability of coated fabric subjected to abrasive tests. Three factors were considered. One factor was filler type with two levels (F1 and F2). Another was surface treatment with two levels (S1 and S2). The third factor was proportion of filler with three levels (25%, 50%, and 75%). Using a completely randomized design with two fabric samples per treatment, the amount of fabric lost in milligrams for each fabric sample was recorded following testing. Data are available in a tab delimited text file at http://dnett.github.io/S510/FabricLoss.txt.

(a) Consider a cell means model for these data. Estimate the mean and standard error for the treatment corresponding to F2, S1, and 50% filler.

(b) The concept of LSMEANS has been explained carefully in lecture and course notes for the special case of a two-factor study. The concept generalizes easily to multi-factor studies. For example, in a three-factor study, the LSMEAN for level $i$ of the first factor is the OLS estimator of $\bar{\mu}_{i\cdot\cdot}$, the average of the cell means for all treatments that involve level $i$ of the first factor. Find LSMEANS for the levels of the factor filler type.

(c) We can also compute LSMEANS for estimable marginal means like $\bar{\mu}_{\cdot jk}$, the average of the cell means for all treatments involving level $j$ of the second factor and level $k$ of the third factor. Find the LSMEAN for surface treatment S2 and 25% filler.

(d) Provide a standard error for the estimate computed in part (c).

(e) In a three-factor study we would say there are no main effects for the first factor if $\bar{\mu}_{i\cdot\cdot} = \bar{\mu}_{i^*\cdot\cdot}$ for all levels $i \neq i^*$. Conduct a test for filler type main effects. Provide an $F$-statistic, a $p$-value, and a conclusion.

(f) In a three-factor study in which the third factor has $K$ levels, we would say there are no three-way interactions if, for all $i \neq i^*$ and $j \neq j^*$,

$$\mu_{ij1}-\mu_{ij^*1}-\mu_{i^*j1}+\mu_{i^*j^*1} = \mu_{ij2}-\mu_{ij^*2}-\mu_{i^*j2}+\mu_{i^*j^*2} = \cdots = \mu_{ijK}-\mu_{ij^*K}-\mu_{i^*jK}+\mu_{i^*j^*K}.$$

Note that each linear combination above can be viewed as a two-way interaction effect involving the first two factors while holding the level of the third factor fixed. If these interaction effects are all the same regardless of which level of the third factor is selected, we say there are no three way interactions. Put another equivalent way, there are no three-factor interactions if

$$\mu_{ijk} - \mu_{ij^*k} - \mu_{i^*jk} + \mu_{i^*j^*k} - \mu_{ijk^*} + \mu_{ij^*k^*} + \mu_{i^*jk^*} - \mu_{i^*j^*k^*} = 0$$

for all $i \neq i^*$, $j \neq j^*$, and $k \neq k^*$. Conduct a test for three-way interactions among the factors filler type, surface treatment, and filler proportion. Provide an $F$-statistic, a $p$-value, and a conclusion.

(g) In a three-factor study, we would say there are no two-way interactions between the first and third factors if

$$\bar{\mu}_{i\cdot k} - \bar{\mu}_{i\cdot k^*} - \bar{\mu}_{i^*\cdot k} + \bar{\mu}_{i^*\cdot k^*} = 0$$

for all $i \neq i^*$ and $k \neq k^*$. Conduct a test for two-way interactions between the factors filler type and filler proportion. Provide an $F$-statistic, a $p$-value, and a conclusion.