1. A collection of questions about column spaces and projection matrices

   (a) What is the column rank of the following $\boldsymbol{X}$ matrix?

$$\begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \end{bmatrix}$$

   (b) The column space for the matrix in part 1a is what geometrical object (e.g., line, plane, cube, ..)?

   (c) Calculate the projection matrix, $\boldsymbol{P_X}$, for the matrix in part 1a.
   R notes: remember solve() computes $\boldsymbol{X}^{-1}$ when $\boldsymbol{X}$ is full rank and t($\boldsymbol{X}$) returns $\boldsymbol{X}'$.

   (d) What is the column rank of this $\boldsymbol{X}$ matrix?

$$\begin{bmatrix} 1 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \end{bmatrix}$$

   (e) The column space for the matrix in part 1d is what geometrical object (e.g., line, plane, cube, ..)?

   (f) Calculate the projection matrix, $\boldsymbol{P_X}$, for the matrix in part 1d.
   R notes: the MASS library has a ginv() function that computes a generalized inverse of matrix, so ginv($\boldsymbol{X}$) returns $\boldsymbol{X}^{-}$. It may help to use round($\boldsymbol{Z}$, 5) which rounds $\boldsymbol{Z}$ to 5 digits and hence cleans up the 'not quite zero' values.

   (g) The coefficients of the projection matrix calculated in part 1f "make sense". Explain why by relating the model using the $\boldsymbol{X}$ matrix in part 1a, the expected values for the 6 observations in that data set, and the coefficients in the projection matrix.

   (h) Your projection matrices for the two $\boldsymbol{X}$ matrices should be the same. Explain why this will always happen for a pair of $\boldsymbol{X}$ matrices, where one represents a cell means model and the other represents a factor effects model for the same data. Your pair of $\boldsymbol{X}$ matrices may have any number of rows and any number of groups in the design (and hence any number of columns).

2. The proof that a projection matrix is idempotent is trivial when the associated $\boldsymbol{X}$ matrix is full rank. (See me if you don't see that). The proof of idempotency on Friday's handout (a copy is in the graphs section of the class web site if you didn't get it) relied on $\boldsymbol{G\,A\,G} = \boldsymbol{G}$, where $\boldsymbol{G} = \boldsymbol{A}^{-}$. That is an unjustified restriction. (Although I did mumble something about needing the right type of generalized inverse when I presented in class).

   The usual definition of a generalized inverse of $\boldsymbol{A}$ is any matrix $\boldsymbol{G}$ that satisfies $\boldsymbol{A\,G\,A} = \boldsymbol{A}$. No other conditions! In particular, many generalized inverses satisfy $\boldsymbol{A\,G\,A} = \boldsymbol{A}$ but not $\boldsymbol{GAG} = \boldsymbol{G}$. The Moore-Penrose inverse is a very special type of generalized inverse that does have both properties: $\boldsymbol{AG\,A} = \boldsymbol{A}$ and $\boldsymbol{G\,A\,G} = \boldsymbol{G}$ (and two more that we). For any matrix $\boldsymbol{A}$, there are many possible generalized inverses $\boldsymbol{G}$, but only one Moore-Penrose inverse. FYI: the ginv() function in the MASS library of R calculates the Moore-Penrose inverse.

   (a) Consider the matrix
$$\boldsymbol{A} = \begin{bmatrix} 4 & 2 & 2 \\ 2 & 2 & 0 \\ 2 & 0 & 2 \end{bmatrix}.$$
   What is the Moore-Penrose inverse of $\boldsymbol{A}$ (use the ginv() function in R)?

   (b) Define $\boldsymbol{G}$ as the Moore-Penrose inverse of $\boldsymbol{A}$. What is $\boldsymbol{A\,G\,A}$?
   What is $\boldsymbol{G\,A\,G}$?
   Does this $\boldsymbol{G}$ matrix satisfy the conditions necessary to use the proof of idempotency on my Friday handout? Briefly explain why or why not.

   (c) Another matrix is
$$\boldsymbol{G}_2 = \begin{bmatrix} -6 & 2.5 & 1.5 \\ -5/6 & 29/6 & 16/3 \\ -7/6 & 14/3 & 37/6 \end{bmatrix}.$$
   Is $\boldsymbol{G}_2$ a generalized inverse of $\boldsymbol{A}$? Briefly explain why or why not.
   Is $\boldsymbol{G}_2$ a Moore-Penrose inverse of $\boldsymbol{A}$? Briefly explain why or why not.

   (d) The $\boldsymbol{A}$ matrix in part 2a is the $\boldsymbol{X'X}$ matrix for
$$\boldsymbol{X} = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \end{bmatrix}$$
   Is the projection matrix calculated using the generalized inverse in part 2a the same as that calculated using $\boldsymbol{G}_2$ from part 2c?
   What is that projection matrix (if the same) or what are the two projection matrices (if not the same)?

   (e) Prove that a projection matrix calculated using **any** type of generalized inverse is idempotent.
   Hint: You can not (or at least I can not) figure out how to use the line of proof used on the handout. Instead, consider $\boldsymbol{P_X}\,\boldsymbol{P_X}$ and expand one of the $\boldsymbol{P_X}$ matrices. Use one

or more properties of projection matrices (see Friday's handout) and the proof will be a few lines long.

3. Consider a sequence of models M_1, M_2, M_3, and M_4 with $\boldsymbol{X}$ matrices $\boldsymbol{X}_1$, $\boldsymbol{X}_2$, $\boldsymbol{X}_3$, and $\boldsymbol{X}_4$. These $\boldsymbol{X}$ matrices are:

$$\boldsymbol{X}_1 = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} \quad \boldsymbol{X}_2 = \begin{bmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 3 \\ 1 & 1 \\ 1 & 2 \\ 1 & 3 \end{bmatrix} \quad \boldsymbol{X}_3 = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 2 & 0 \\ 1 & 3 & 0 \\ 1 & 1 & 1 \\ 1 & 2 & 1 \\ 1 & 3 & 1 \end{bmatrix} \quad \boldsymbol{X}_4 = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & 2 & 0 & 0 \\ 1 & 3 & 0 & 0 \\ 1 & 1 & 1 & 1 \\ 1 & 2 & 1 & 2 \\ 1 & 3 & 1 & 3 \end{bmatrix}$$

Your thoughts on the following questions may be aided by being told that the first three observations are on men; the second three are on women (i.e. two groups), that data were collected at 3 levels of a continuous variable (1, 2, 3) and one goal of the study is to estimate the regression slope of Y on that continuous variable, perhaps with a different slope for men and women.

Note: To avoid potential confusion, I have numbered the coefficients β_1, β_2, β_3, and β_4, to correspond to their columns of the $\boldsymbol{X}$ matrix and to the model to which they are added.

(a) Describe in words what model M_4 (and hence $\boldsymbol{X}_4\boldsymbol{\beta}$) represents. For example, a cell means model for two groups could be described as "men have one mean; women have a second mean".
Hint: It may help to work out what M_4 represents for men (observations 1, 2, and 3) and what it represents for women (observations 4, 5, and 6).

(b) What hypothesis is tested by the comparison of M_3 and M_4?
Provide a verbal interpretation of this hypothesis.
For example, for a comparison of an equal means model to a model where two groups have two different means, you might answer the first part as $\mu_A = \mu_B$ and the second part as "the two groups have the same mean".

(c) What hypothesis is tested by the comparison of M_2 and M_4?
Provide a verbal interpretation of this hypothesis.

The Y values for these 6 observations are $Y = [0.91, 2.00, 3.01, 3.94, 7.94, 11.94]'$.
For all remaining parts of this question, the $\boldsymbol{\beta}$ vector corresponds to model M_4, that is $\boldsymbol{\beta} = [\beta_1\ \beta_2\ \beta_3\ \beta_4]'$.

(d) Report the estimated parameters ($\hat{\boldsymbol{\beta}}$) for model M_4.
R notes: You can estimate parameters for an $\boldsymbol{X}$ matrix by separating each column of the $\boldsymbol{X}$ matrix as a separate "variable" or by estimating β by lm(y~X) where both y and X are variables in your main workspace.
If X includes an intercept term (column of 1's), R will return NA (missing value) for the first column of X, you can avoid this by telling R to omit the default intercept by lm(y

-1 + X).
SAS notes: If you want to supress the default intercept in SAS, add /noint; at the end of the model statement. Each column of the $\boldsymbol{X}$ matrix needs to be stored in a separate variable. All variables are then included in the model statement.

(e) Use a $C\beta$ test to test $\beta_4 = 0$. Report your T statistic and p-value.

R notes: The quick way to do this is to use summary() on the object produced by lm(). That gives lots of information, including tests (T statistics and p-values) of each coefficient = 0. Or, you can extract the coefficient and the se of that coefficient, as the sqrt(appropriate element of vcov()), and construct the T statistic.
SAS notes: If you have no class statement in your proc glm, you get this test by default. If the four columns of $\boldsymbol{X}$ are X1, X2, X3, and X4, then model Y = X1 X2 X3 X4/noint; will fit model $M4$ and if there is no class statement, SAS will report the 4 $\hat{\beta}$'s and T statistics and p-values for the test of each β= 0. If there is a class statement (for some other reason), adding /solution to the model statement or an additional estimate statement will provide the estimate, its se, and a test.

(f) Use a model comparison approach to test $\beta_4 = 0$. Report what you are using as the full and reduced models (referring back to the list of models M_1, etc. at the beginning of the problem is fine), the SSE for the full model, the SSE for the reduced model, the F statistic, and the p-value.

(g) Are the p-values in parts 3e and 3f the same?

(h) Use a $C\beta$ test to test $\beta_3 = 0$ when fitting model M_4 (with the columns in the order given above). β_3 is the coefficient for the 3rd column. Report your T statistic and p-value.

R/SAS Note: Since this is a test of a single coefficient = 0, you can use the "quick" methods described a few parts previously.

(i) Use a model comparison approach (i.e. comparing model M_2 to model M_3) to test $\beta_3 = 0$. The hypothesis SS should be SS($M_3 \mid M_2$), i.e. the sequential SS going from model 2 to model 3. Estimate $\hat{\sigma}^2$ from the MSE from M_4. Report the SSE for the full model, the SSE for the reduced model, the F statistic, and the p-value.

R notes: You can get the change in SS between models M_2 and M_3 by fitting M_4 with a separate column for each variable, e.g. lm(y x1 + x2 + x3 + x4). Using anova() on the object produced by lm() gives you the chain of sequential SS. You just need to identify the appropriate line of output. To get the SSE's for the full and reduced models, I suggest you fit M_2 and M_3 separately and use anova() on those to get the SSE's.
SAS notes: Similar approach to R: you can get the change in SSE and the F test based on that change in SSE from the TYPE I SS and F test. You can get each SSE by fitting each model separately.

(j) Are the p-values in parts 3h and 3i the same?

(k) Test whether a single regression line (one intercept, one slope) is appropriate for these data or whether men and women need different lines (perhaps different intercepts, perhaps different slopes or perhaps both). Your choice of model comparison or $C\beta$. You don't have to do both. Report your test statistic and p-value, and state the distribution of the test statistic when Ho is true.

We'll talk about the differences between these approaches next week. Please note that I don't ask you to explain / understand the similarities or differences. Discovering them yourself will reinforce what I say in lecture on Wednesday and/or Friday (or perhaps Monday the week after).