STAT 575 – Assignment 2 – due date is on course outline

1. The LR criterion (6-48), for testing the equality of the covariance matrices in a one-way MANOVA, is actually a modification of the LR test in which sample sizes are replaced by degrees of freedom. Show that the likelihood ratio test rejects for small values of
$$\Lambda^* = k \cdot \prod_{l=1}^{g} \left(\frac{|\mathbf{S}_l|}{|\mathbf{S}_{pooled}|\,|} \right)^{n_l/2}, \text{ where } k = \frac{\prod_{l=1}^{g} \left(\frac{n_l-1}{n_l}\right)^{pn_l/2}}{\left(\frac{n-g}{n}\right)^{pn/2}}.$$

2. (a) Consider the 'husbands and wives' example discussed in Lecture 8. In our discussion the two samples were independent. Suppose instead that the data (on the course web site) were a sample of couples, with husbands and wives each rating their own spouses. Formulate and test the same first two hypotheses as were discussed in class. Suggest a way to formulate and test the third hypothesis.

   (b) For the situation discussed in class, show that an equivalent formulation of the hypothesis of parallel profiles is '$H_{01} : \boldsymbol{\mu}_2 = \boldsymbol{\mu}_1 + \theta \mathbf{1}_p$ for some θ'. What then is the additional hypothesis of coincident profiles?

3. Show that $vec\,(\mathbf{ABC}) = \left(\mathbf{A} \otimes \mathbf{C}^T\right) vec\mathbf{B}$. [Hint: First establish this in the case that $\mathbf{B}$ is of the form $\mathbf{x}\mathbf{y}^T$. Then show that *any* matrix can be expressed as a sum of such 'rank 1' matrices.]

4. Recall that, in the multivariate regression model
$$\mathbf{Y}_{n\times m} = \mathbf{Z}_{n\times(r+1)}\mathbf{B}_{(r+1)\times m} + \mathbf{E}_{n\times m},$$
a $100\,(1-\alpha)\,\%$ confidence region for one linear combination $\boldsymbol{\gamma} = \mathbf{B}^T\mathbf{z}$ is
$$\left\{ \boldsymbol{\gamma} | \frac{(\hat{\boldsymbol{\gamma}}-\boldsymbol{\gamma})^T \left[\frac{\mathbf{W}}{n-r-1}\right]^{-1} (\hat{\boldsymbol{\gamma}}-\boldsymbol{\gamma})}{\mathbf{z}^T \left(\mathbf{Z}^T\mathbf{Z}\right)^{-1} \mathbf{z}} \leq \frac{m\,(n-r-1)}{n-r-m} F^m_{n-r-m}\,(\alpha) = c^2 \right\},$$
where $\hat{\boldsymbol{\gamma}} = \hat{\mathbf{B}}^T\mathbf{z}$. Show that a simultaneous region for *all* such linear combinations is obtained by replacing c^2 by $(n-r-1)$ times the $1-\alpha$ quantile of the distribution of
$$\Lambda = ch_{\max}\mathbf{W}^{-1}\mathbf{R},$$
where $\mathbf{W} \sim W_m\,(n-r-1, \boldsymbol{\Sigma})$ independently of $\mathbf{R} \sim W_m\,(r+1, \boldsymbol{\Sigma})$ and we can take $\boldsymbol{\Sigma} = \mathbf{I}_m$ without loss of generality. [You might repeatedly need the result stated in Question 5a) of Assignment 1.]

... over

5. See Example 6.13. Carry out an analysis on R. Write a script that results in the following output: (a) a printout of all $\hat{\tau}$ (call them tau1 and tau2, i.e. indices '1' and '2' will denote respectively the low and high levels of a factor), all $\hat{\beta}$ (call them beta1 and beta2) and all $\hat{\gamma}$ (call them gam11, etc.), together with a 2×2 table expressing the mean responses at each of the 4 combinations of levels of the factors; (b) the matrix $\mathbf{SSP}_{Res}$; (c) the p-values for the tests of H_{01} : no interactions, H_{02} : no factor 1 effects, H_{03} : no factor 2 effects. Do this both for Pillai's test and Wilks's test. I will be looking to see that you do much of this by retrieving the relevant information from R, without doing a lot of detailed computing yourself. This might require some time spent alternating between Google and the R help files. *E-mail your script to me – I should be able to run it exactly as I receive it and obtain the required output.*
(d) Assess the normality of the residuals, with appropriate plots and tests. Comment.

6. MANOVA can be phrased as a regression problem, and if the data are unbalanced then this might be the only way to do it, at least on some packages. To formulate two-way manova as a regression problem, first let $\mathbf{Y}$ be the $n \times m$ data matrix (here n is the *total* number of responses vectors – what was previously written gbn in manova notation). With g levels of factor 1 and b levels of factor 2, define vectors, which will become columns of the $\mathbf{Z}$-matrix, as follows. Let $\mathbf{u}_j (j = 1, ..., g-1)$ be an n-vector with a $+1$ in the i^{th} location if this response is obtained using the $(j+1)^{th}$ level of factor 1, and a -1 otherwise. Define vectors $\mathbf{v}_j$ $(j = 1, ..., b-1)$ similarly, but indicating the levels of factor 2. Let $\mathbf{w}_j$ $(j = 1, ..., (g-1)(b-1))$ consist of all the (elementwise) products of the $\mathbf{u}_\alpha$ with the $\mathbf{v}_\beta$. Then form the $\mathbf{Z}$-matrix from these columns – first a column of ones, then all of the $\mathbf{u}_j$, then all of the $\mathbf{v}_j$, finally all of the $\mathbf{w}_j$. Now fit the multivariate regression model $\mathbf{Y} = \mathbf{ZB} + \mathbf{E}$.

   Here you are asked to verify, in the context of the data from the preceding question, that H_{01} is now the hypothesis that the final $(g-1)(b-1)$ rows of $\mathbf{B}$ vanish; if so then H_{02} is the hypothesis that rows $2, ..., g$ vanish, and H_{03} is the hypothesis that rows $g+1, ..., g+b-1$ vanish. Thus:

   (a) In the context of Example 6.13, formulate the regression model. Prepare a 2×2 table expressing the mean responses at each of the 4 combinations of levels of the factors, in both manova and regression notation. Use the notation $\boldsymbol{\mu} + \boldsymbol{\tau} + \boldsymbol{\beta} + \boldsymbol{\gamma}$ to denote the mean response when both factors are at the high levels, and then represent the other three means in these symbols. Solve for the regression parameters (the rows of $\mathbf{B}$) in terms of $\boldsymbol{\mu}, \boldsymbol{\tau}, \boldsymbol{\beta}, \boldsymbol{\gamma}$.

   (b) Fit the regression model, and use $\hat{\mathbf{B}}$ to evaluate the terms in your 2×2 table. Verify that you obtain the same numbers as in part (a) of the previous question.

... over

(c) Carry out (c) of the previous question. Clearly indicate the R commands which yield the results, when applied to your regression fit. Verify that the p-values are the same as previously obtained.

(d) Suppose now that the design was unbalanced, and that there was an additional observation $(3, 6, 8)$ with each factor at its high level. Prepare a 3×2 table giving the p-values of these three hypotheses, as obtained from (i) the manova output and (ii) the regression output.