Homework 6 Statistics 8320 (Wikle) April 17, 2016

DUE: April 28 , 2016

1. Consider an experiment where there were two treatment groups with repeated measurements at 4 times on each subject. Group 1 and 2 had 16 and 11 experimental units, respectively. Let $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$ represent the means for these two groups.

   (a) One question of interest is whether the mean profile for boys and girls are the same. Hotelling's T^2 statistic is calculated to be 16.5. Use an F-test to test if the mean profiles are the same (test at $\alpha = 0.05$).

   (b) Consider the test for time by group interaction given by $H_0 : \mathbf{C}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) = \mathbf{0}$. The T^2 statistic for this test is calculated to be 8.8. Write out the form of the matrix $\mathbf{C}$ for this test and conduct an F-test for this interaction.

   (c) Based on the results in part (b), what test or tests would you consider next?

2. A famous repeated measures dataset is from Pothoff and Roy (1964), which consists of dental measurements from the center of the pituitary to the pteryomaxillary fissure in 11 girls and 16 boys at ages 8, 10, 12, and 14. The subjects are individual children and there are four repeated measurements on each. The data are given in `PothoffRoy1964.dat` on the Blackboard class site. The first column is the person identifier, the next column is gender (F- female, M-male), and then the four repeated measurements (at 8, 10, 12, and 14). Use SAS PROC GLM to answer the following questions.

   (a) Consider the Mauchly test for sphericity applied to orthogonal components. What is the purpose of this test and what is being tested? What is the result and what does this suggest we should do with respect to the analysis of repeated measures here?

   (b) Regardless of your results in part (a), perform a profile analysis of these data. Specifically, report the Wilks' Lambda value, F value, degrees of freedom, p-value and result for a test of (i) parallel profiles (age*gender interaction), (ii) horizontal profiles (age effect), and (iii) coincident profiles (between subject effects).

   (c) Report test results for the univariate tests for age*gender interaction and age (if appropriate).

   (d) What are the Greenhouse-Geisser and Huynh-Feldt adjustments used in the univariate tests?

   (e) Use PROC MIXED to analyze these data with a repeated measurement on the subjects. Assume a compound symmetry covariance matrix for the within subject covariance matrix. Report your results regarding the interaction and main effects. [Note: PROC MIXED and GLM differ in several notable ways. Among them: PROC MIXED uses REML to estimate random effects parameters and GLM uses method of moments. In addition, MIXED allows you to specify the covariance structure of the random effects, and GLM assumes implicitly that the effects are unstructured. In addition, GLM is not as flexible when it comes to handling missing data - which is not a problem here.]

3. Consider the following data on a study of learning for 15 rats that were randomly assigned to three different reinforcement schedules and then given a maze to go through under four different experimental conditions. The sequence in which the four conditions were presented in the experiment was randomized independent for each animal. The response variable in the study was the number of seconds taken to run the maze. The data are given below:

Reinforcement Schedule	Rat	Cond. 1	Cond. 2	Cond. 3	Cond. 4
1	1	29	20	21	18
1	2	24	15	10	8
1	3	31	19	10	31
1	4	41	11	15	42
1	5	30	20	27	53
2	1	25	17	19	17
2	2	20	12	8	8
2	3	35	16	9	28
2	4	35	8	14	40
2	5	26	18	18	51
3	1	10	18	16	14
3	2	9	10	18	11
3	3	7	18	19	12
3	4	8	19	20	5
3	5	11	20	17	6

   (a) Plot the mean profiles for each reinforcement schedule. Describe the plots and what they might indicate with respect to the reinforcement schedules across conditions.

   (b) Test the hypothesis of an overall reinforcement schedule effect.

   (c) The scientists are interested in the hypothesis that schedule 1 and 2 behave the same, and they both behave differently than schedule 3. Describe the tests and results that you conduct to evaluate their hypotheses.

4. Consider the data below that represent 10 measures on each of 3 traits:

$$\mathbf{X} = \begin{bmatrix} 7 & 4 & 3 \\ 4 & 1 & 8 \\ 6 & 3 & 5 \\ 8 & 6 & 1 \\ 8 & 5 & 7 \\ 7 & 2 & 9 \\ 5 & 3 & 3 \\ 9 & 5 & 8 \\ 7 & 4 & 5 \\ 8 & 2 & 2 \end{bmatrix}$$

(a) Use R software to calculate the covariance matrix. Report the estimated covariance matrix.

(b) What is the total variance?

(c) Report the values of $\mathbf{A}$ and $\mathbf{\Lambda}$ in $\mathbf{C} = \mathbf{A\Lambda A'}$, where $\mathbf{C}$ is the covariance matrix you calculated above. Make sure your decomposition is such that eigenvalues are ordered so that $\lambda_1 \geq \lambda_2 \geq \lambda_3$, where $\mathbf{\Lambda} = diag(\lambda_1, \lambda_2, \lambda_3)$.

(d) How much variation is accounted for by the first principal component?

(e) What is the cumulative variation accounted for by the first two principal components?

(f) Report the first two principal components (e.g., y_1, and y_2).

(g) Report the loading coefficients for the first two principal components (i.e., report the correlation between the principal component and the attributes of $\mathbf{X}$. Interpret the loadings.

(h) Plot the first two principle components as a scatter plot. What to you notice?

(i) Report the estimated correlation matrix of the original data.

(j) As before, report the eigenvectors and eigenvalues of the symmetric decomposition of this correlation matrix. How much variation is accounted for by the first principal component and the second principal component from this decomposition?

5. The file

`decathlon.dat`

contains results for the men's decathlon in the 1988 Olympics. Each column of the data set contains the following information for a given individual athlete:

- *col 1:* 100 meter race
- *col 2:* long jump
- *col 3:* shot put
- *col 4:* high jump
- *col 5:* 400 meter race
- *col 6:* hurdles
- *col 7:* discus
- *col 8:* pole vault
- *col 9:* javelin
- *col 10:* 1500 meter run
- *col 11:* Total Score

Note that the "Total Score" is based on traditional conversion tables for each event (i.e., non-statistical). We will investigate how alternative "statistically based" scores might compare to these. Note, before we do the analysis it will facilitate interpretation if all the events were "scored" in the same direction (i.e., biggest distance in long jump is good, whereas smallest time in running event is good) so you should make the running events (100 meters, 400 meters, hurdles, and 1500 meters) negative so that a big value is "good" for all events.

Consider a Principal Components (PC) analysis on these data using, using the correlation matrix in SAS.

(a) Why does it make sense to consider the correlation matrix in this case?

(b) How much variability is accounted for by the first 2 PCs?

(c) Why is it reasonable to consider just the first 2 PCs?

(d) Interpret the first two PCs.

(e) Plot the first PC scores versus the 2nd PC scores. Can you identify any pattern in this plot relative to an athlete's overall ranking in the detcathlon point total? Explain.

(f) Correlate the first PC score to the overall decathlon point score. Do the same, but for the second PC scores and the overall decathlon score. Interpret these results to suggest what is being measured by the overall decathlon score.