ST 557: Applied Multivariate Analysis

Homework 3

1. Monthly temperature data for 20 different weather stations within 100 miles of Corvallis were obtained for the period 1950 - 2009. From this data, decade averages were computed for each station and are given in the **TempData.csv** file available on Canvas. Let $\mu_1, \mu_2, \mu_3, \mu_4, \mu_5, \mu_6$ denote the average temperature for decades 1950s, 1960s, 1970s, 1980s, 1990s, 2000s respectively.

   (a) Test the null hypothesis $H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5 = \mu_6$ vs. H_A : not all μ_j are equal at level $\alpha = 0.05$ using Hotelling's T^2 test. Explain how you performed this test. Based on the result of this hypothesis test, would you conclude that the average temperature around Corvallis has stayed constant over the past 60 years?

   (b) Construct simultaneous 95% Bonferroni confidence intervals for $\mu_2 - \mu_1$, $\mu_3 - \mu_1$, $\mu_4 - \mu_1$, $\mu_5 - \mu_1$, and $\mu_6 - \mu_1$. Do any of these confidence intervals include 0? What would you conclude based on these confidence intervals?

2. Researchers have suggested that a change in skull size over time is evidence of the interbreeding of a resident population with immigrant populations. Samples of 30 male Egyptian skulls were obtained for five different time periods: 4000 B.C., 3300 B.C., 1850 B.C., 200 B.C., and 180 A.D. For each skull, measurements of four dimensions were taken, and are given in the file **SkullData.csv** available on Canvas. The measured variables are

   - MB: Maximal Breadth of Skull
   - BH: Basibregmatic Height of Skull
   - BL: Basialveolar Length of Skull
   - NH: Nasal Height of Skull

   The first column in the data file indicates the time period of the sample, with negative numbers indicating B.C. and positive numbers indicating A.D.

   (a) Compute and compare the covariance matrices for each time period. Do they seem approximately similar?

   (b) Perform a level $\alpha = 0.05$ test of the hypothesis that population mean vectors for all of these time periods are the same (assume equal covariance matrices). Based on this hypothesis test, does there seem to be evidence of interbreeding (if the researchers' theory that skull size change indicates interbreeding is correct)?

   (c) Perform separate univariate ANOVAs for each variable at level $\alpha^* = \frac{\alpha}{p}$. Are any of these univariate ANOVAs significant? If we reject $H_0 : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = \boldsymbol{\mu}_3 = \boldsymbol{\mu}_4 = \boldsymbol{\mu}_5$ if any of the univariate ANOVAs are significant at level $\alpha^* = \frac{\alpha}{p}$, will the overall probability of a Type I error (the overall significance level) be controlled at level α (that is, will it be $\leq \alpha$)? Explain.

3. Air-pollution measurements were taken at 12:00 noon in Los Angeles on 42 different days. For each day, the amount of wind and solar radiation were measured, along with quantities of 2 different pollutants (NO_2 and O_3). The data are available on Canvas in the file **PollutionData.csv**.

(a) Perform a multivariate multiple regression analysis using both responses $Y_1 = \mathrm{NO}_2$ and $Y_2 = \mathrm{O}_3$ and predictors X_1 = Wind and X_2 = Solar Radiation. Test the null hypothesis that $\boldsymbol{\beta}_2 = \mathbf{0}$. What would you conclude based on this test?

(b) Test the null hypothesis that $\boldsymbol{\beta}_1 = \mathbf{0}$. What would you conclude based on this test?

(c) Test the null hypothesis that $\boldsymbol{\beta}_1 = \boldsymbol{\beta}_2 = \mathbf{0}$. What would you conclude based on this test? How does this result fit with the results of parts (a) and (b)?

4. Data on national track records for men for 55 different countries are given in the file **TrackData.csv**, available on Canvas. For each country, the national record for eight different distances is recorded. The first column of the data set contains the country name, and the second column contains a three-letter abbreviation. The next eight columns contain the national records for each of the eight distances: the first three distances (100m, 200m, 400m) are recorded in seconds, and the remaining five distances (800m, 1500m, 5000m, 10000m, Marathon) are recorded in minutes.

   (a) Obtain the sample covariance matrix **S** and the sample correlation matrix **R** for the distances based on this data. Which of these matrices would you find more interesting/appropriate to use for a principal component analysis of this data, and why?

   (b) Determine the eigenvalues and eigenvectors of **S**.

   (c) Determine the eigenvalues and eigenvectors of **R**.

   (d) Construct four plots of the loadings for the first four principal components of **S**. Include the loadings for the corresponding principal component of **R** on the same plot, in a different color or plotting character.

   (e) How would you interpret the loadings for the first principal component found using **S**?

   (f) How would you interpret the loadings for the first principal component found using **R**?

   (g) What does the second principal component of **R** seem to represent?

   (h) Plot the scree plot and cumulative variance explained plot for the principal components of both **S** and **R**.

   (i) How many principal components of **R** would you want to keep? Explain.

5. The weekly rates of return for five stocks listed on the New York Stock Exchange are given in the file **NYSEData.csv**, available on Canvas. For each of the 103 weeks (two years), the return rates are reported for all five stocks. The first three stocks are for financial companies, and the last two are energy/petroleum stocks.

   (a) Construct the sample covariance matrix **S** for the five stocks, and find the sample principal components.

   (b) What proportion of the total sample variance is explained by the first three principal components?

   (c) How would you interpret the first three principal components (that is, interpret the loadings for each of these components on the original variables–what do each of these directions represent)?