ST 557: Applied Multivariate Analysis
Fall 2019

Homework 4

Due: Friday November 8
90 Points

1. **(20 points)** A group of $n = 66$ students were given two different reading tests, Test 1 and Test 2, both before and after participating in a reading instruction program. Each student produced four test scores: Pre1, Pre2, Post1, and Post2. The pre- and post-instruction scores for both tests are given in the ReadingTest.csv file available on Canvas. Let μ_{11} denote the population average score on Test 1 before instruction; μ_{12} denote the population average score on Test 2 before instruction; μ_{21} denote the population average score on Test 1 after instruction; and μ_{22} denote the population average score on Test 2 after instruction. Then $\boldsymbol{\mu}_1 = [\mu_{11}, \mu_{12}]^T$ is the before instruction population mean vector, and $\boldsymbol{\mu}_2 = [\mu_{21}, \mu_{22}]^T$ is the after instruction population mean vector.

   (a) Is a paired test appropriate to test the hypothesis $H_0 : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2$ vs. $H_A : \boldsymbol{\mu}_1 \neq \boldsymbol{\mu}_2$? Why or why not?

   (b) Perform a level $\alpha = 0.05$ test of $H_0 : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2$ vs. $H_A : \boldsymbol{\mu}_1 \neq \boldsymbol{\mu}_2$. Based on the result of this hypothesis test, would you conclude that the reading instruction produces a difference in performance on the two tests?

   (c) Construct 95% simultaneous Bonferroni confidence intervals for the differences in means (Post - Pre) for the two tests.

   (d) Construct the 95% Hotelling's T^2 confidence region for the difference vector $\boldsymbol{\delta} = \boldsymbol{\mu}_2 - \boldsymbol{\mu}_1$.

2. **(10 points)** Monthly temperature data for 20 different weather stations within 100 miles of Corvallis was 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?

3. **(30 points)** It is often claimed that professional athletes perform better in years when they are facing a contract renewal/are eligible to be free agents. Performance statistics for 337 non-pitcher MLB players from the 1992 season are given in the file BaseballData.csv. The first column of the data file is an indicator variable for whether that player was free agent eligible in the 1992 season (0 = not eligible, 1 = eligible). A player may declare himself a free agent, and is thus free agent eligible (able to negotiate with any team for a new contract) if:

- He has at least 6 years of Major League service; and
- He is not already under contract for the next season.

The remaining columns are various statistics used to measure how well the player is playing (note that high numbers are considered good for all but Walks, Strike-outs, and Errors–Walks are arguably neutral, though they do contribute to OBP: on-base percentage). These columns are:

- Batting average
- On-base percentage (OBP)
- Number of runs
- Number of hits
- Number of doubles
- Number of triples
- Number of home runs
- Number of runs batted in (RBI)
- Number of walks
- Number of strike-outs
- Number of stolen bases
- Number of errors

(a) To use this data to test the hypothesis that average performance is the same whether or not a player is eligible for free agent status, would a paired test be appropriate? Why or why not?

(b) Compute and compare the covariance matrices for the non-eligible and the eligible players. Do the covariance matrices seem roughly similar?

(c) Test the hypothesis that performance is the same whether or not a player is free agent eligible at level $\alpha = 0.05$ using Hotelling's T^2 test for equal covariance matrices.

(d) Test the hypothesis that performance is the same whether or not a player is free agent eligible at level $\alpha = 0.05$ using Hotelling's T^2 test for unequal covariance matrices (use the asymptotic chi-square critical value). Does the result/decision differ from part c?

(e) Would it be reasonable to always perform both the tests for equal and unequal covariance matrices, and then reject the null if either one of the tests rejected? Why or why not?

(f) Recall that the critical value for moderate sample sizes is

$$\frac{\nu p}{\nu - p + 1} F_{(p,\nu-p+1)}(\alpha)$$

where the denominator degrees of freedom parameter ν is estimated as

$$\nu = \frac{p + p^2}{\sum_{\ell=1}^{2} \frac{1}{n_\ell} \left\{ \operatorname{tr}\left[\left(\frac{1}{n_\ell} \mathbf{S}_\ell \left(\frac{1}{n_1}\mathbf{S}_1 + \frac{1}{n_2}\mathbf{S}_2 \right)^{-1} \right)^2 \right] + \left(\operatorname{tr}\left[\frac{1}{n_\ell} \mathbf{S}_\ell \left(\frac{1}{n_1}\mathbf{S}_1 + \frac{1}{n_2}\mathbf{S}_2 \right)^{-1} \right] \right)^2 \right\}}$$

where $\min(n_1, n_2) \leq \nu \leq n_1 + n_2$. Compare the chi-square critical value $\chi^2_{(p)}$ to the minimum and maximum possible F critical values for this problem. Does it matter much whether the F critical value or the χ^2 critical value is used?

4. **(15 points)** 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.

5. **(15 points)** 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 $Y_1 = NO_2$ and $Y_2 = O_3$ 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)?