

THE UNIVERSITY OF CHICAGO
Graduate School of Business
Business 41912, Spring Quarter 2008, Mr. Ruey S. Tsay

Midterm

GSB Honor Code:

I pledge my honor that I have not violated the Honor Code during this examination.

Signature:

Name:

ID:

Notes:

1. Open book and notes. The exam time is 120 minutes.
 2. Write your answers in a bluebook. Mark the solution clearly.
 3. All tests are based on the 5% significance level.
 4. Some R output is attached for the problems.
1. (20 points) Suppose that $\mathbf{X} = (X_1, X_2, X_3)'$ follows a 3-dimensional normal distribution with mean $\boldsymbol{\mu} = (1, 2, 3)'$ and covariance matrix

$$\boldsymbol{\Sigma} = \begin{bmatrix} 3 & 1 & 1 \\ 1 & 2 & 0 \\ 1 & 0 & 3 \end{bmatrix}.$$

Answer the following equations:

- (a) What is the distribution of $\mathbf{Z} = (X_1 + X_2, X_1 - X_2)'$?
- (b) What is the distribution of X_1 given $X_2 = 1$ and $X_3 = 2$?
- (c) Find a linear combination of \mathbf{X} of length 1 that has the minimum variance among all possible linear combinations.
- (d) Find the coefficient $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2)'$ of the least squares regression

$$X_3 = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon.$$

2. (25 points) Consider the Moody's bond rating data of Problem 6.21. The four variables used are

- X_1 = current ratio (a measure of short-term liquidity)
- X_2 = long-term interest rate (a measure of interest coverage)
- X_3 = debt-to-equity ratio (a measure of financial risk or leverage)
- X_4 = rate of return on equity (a measure of profitability).

A sample of 20 bonds each from the Aa and Baa categories is collected. The summary statistics are as follows:

(a) Aa bond companies: $n_1 = 20$, $\hat{\mathbf{x}}'_1 = [2.287, 12.600, 0.347, 14.830]$, and

$$\mathbf{S}_1 = \begin{bmatrix} 0.459 & 0.254 & -0.026 & -0.244 \\ 0.254 & 27.465 & -0.589 & -0.267 \\ -0.026 & -0.589 & 0.030 & 0.102 \\ -0.244 & -0.267 & 0.102 & 6.854 \end{bmatrix}.$$

(b) Baa bond companies: $n_2 = 20$, $\hat{\mathbf{x}}'_2 = [2.404, 7.155, 0.524, 12.840]$, and

$$\mathbf{S}_2 = \begin{bmatrix} 0.944 & -0.089 & 0.002 & -0.719 \\ -0.089 & 16.432 & -0.400 & 19.044 \\ 0.002 & -0.400 & 0.024 & -0.094 \\ -0.719 & 19.044 & -0.094 & 61.854 \end{bmatrix}.$$

The pooled sample covariance matrix is

$$\mathbf{S}_{pool} = \begin{bmatrix} 0.701 & 0.083 & -0.012 & -0.481 \\ 0.083 & 21.949 & -0.494 & 9.388 \\ -0.012 & -0.494 & 0.027 & 0.004 \\ -0.481 & 9.388 & 0.004 & 34.354 \end{bmatrix}.$$

Answer the following questions:

- Using the pooled covariance matrix, test for the equality of the mean vectors between the two categories. Write down the hypotheses and draw your conclusion.
- Construct the 95% simultaneous confidence intervals for the components of the differences in mean vectors.
- Construct the Bonferroni 95% simultaneous confidence intervals for the components of the differences in mean vectors.
- Test the equality between the two covariance matrices. Draw the conclusion.
- Assuming the populations are normal with unequal covariance matrices. Test for the equality in the mean vectors. Draw your conclusion.

3. (20 points) Consider the psychological profile data on Table 4.6 of the textbook, p. 207. These data are collected from a psychological test administered to Peruvian teenagers (ages 15, 16 and 17). The sample size is 130. The five response variables are scores for *independence*, *support*, *benevolence*, *conformity*, and *leader*. The gender (male = 1, female = 2) and the socioeconomic status (low = 1 and medium = 2) are also recorded. In our analysis, we treat gender and socioeconomic status as two factors, and our goal is to investigate whether the two factors have significant effects on the test result. Answer the following questions.
- Are the data normally distributed at the 5% level? (See the qqchi2 result)
 - Is there any interaction effect between gender and socioeconomic status? Why?
 - Did the socioeconomic status affect the test result? Why?
 - Did the gender affect the test result? Why?
 - Focus on the score of *leader*. Construct a univariate two-way analysis of variance table from the output.
4. (25 points) Consider the multiple linear regression

$$\mathbf{Y}_{n \times 1} = \mathbf{Z}_{n \times (r+1)} \boldsymbol{\beta}_{(r+1) \times 1} + \boldsymbol{\epsilon}_{n \times 1},$$

where n is the sample size, r is the number of explanatory, and the first column of \mathbf{Z} is the n -dimensional vector of 1. Answer the following questions:

- Write down the assumptions for which the ordinary least squares estimates are also the maximum likelihood estimates of the vector $\boldsymbol{\beta}$.
- Let $\mathbf{H} = \mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'$ be the hat-matrix. Then, the least squares estimate of $\boldsymbol{\beta}$ is $\hat{\boldsymbol{\beta}} = (\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{Y}$. Let $\hat{\boldsymbol{\epsilon}} = \mathbf{Y} - \mathbf{Z}\hat{\boldsymbol{\beta}}$ be the residual vector. Show that $\sum_{j=1}^n \hat{\epsilon}_j = 0$, the $\hat{\epsilon}_j$ is the j th element of $\hat{\boldsymbol{\epsilon}}$.
- Derive the covariance matrix of $\hat{\boldsymbol{\epsilon}}$.
- Let $tr(\mathbf{H})$ be the trace of the hat matrix \mathbf{H} . Prove that $tr(\mathbf{H}) = r + 1$ and $h_{jj} > 0$ for all j , where h_{jj} is the (j, j) th element of \mathbf{H} .
- A high h_{jj} indicates high leverage. Why? Please explain.

5. (10 points) Briefly answer the following questions:

- (a) If \mathbf{X}_1 and \mathbf{X}_2 are multivariate normal with mean zero and covariance matrix Σ_1 and Σ_2 , respectively, then $\mathbf{X} = (\mathbf{X}'_1, \mathbf{X}'_2)'$ follows a multivariate normal distribution. True or false? Why?
- (b) Let $\bar{\mathbf{X}}$ be the sample mean of the data $\mathbf{X}_1, \dots, \mathbf{X}_n$, where $E(\mathbf{X}_i) = \boldsymbol{\mu}$ and $\text{Cov}(\mathbf{X}_i) = \Sigma$ for all i . Then, $\sqrt{n}(\bar{\mathbf{X}} - \boldsymbol{\mu})$ is approximately $N_p(\mathbf{0}, \Sigma)$ for large n , where $p = \dim(\mathbf{X}_i)$. True or false? Why?
- (c) Give two methods that can be used to select a linear regression model when there are r explanatory variables.
- (d) Define the Cook distance of a linear regression model with r explanatory variables and n data points.
- (e) If \mathbf{X} is distributed as $N_p(\mathbf{0}, \Sigma)$, where Σ is positive definite, then $\mathbf{X}'\Sigma^{-1}\mathbf{X}$ is distributed as χ_p^2 . True or false? Why?

```
*** Data analysis for the midterm. May 2008.
```

```
> setwd("C:/teaching/ama")
```

```
** Problem 1 **
```

```
> S=matrix(c(3,1,1,1,2,0,1,0,3),3,3)
```

```
> S
```

```
      [,1] [,2] [,3]
[1,]    3    1    1
[2,]    1    2    0
[3,]    1    0    3
```

```
> mm=eigen(S)
```

```
> mm$values
```

```
[1] 4.246980 2.554958 1.198062
```

```
> mm$vectors
```

```
      [,1]      [,2]      [,3]
[1,] 0.7369762 -0.3279853  0.5910090
[2,] 0.3279853 -0.5910090 -0.7369762
[3,] 0.5910090  0.7369762 -0.3279853
```

```
** Problem 3 **
```

```
> x=read.table("t4-6.dat")
```

```
> dim(x)
```

```
[1] 130  7
```

```
> fac1=factor(x[,7])
```

```
> fac2=factor(x[,6])
```

```
> da=x[,1:5]
```

```
> apply(da,2,mean)
```

```
      V1      V2      V3      V4      V5
15.66923 17.07692 18.78462 15.50000 11.73077
```

```
> S=cov(da)
```

```
> print(S,digits=3)
```

```
      V1      V2      V3      V4      V5
V1  34.75 -4.28 -18.07 -15.97  5.72
V2  -4.28 17.51  0.42  -7.87  -8.72
V3 -18.07  0.42 29.84  9.35 -13.94
V4 -15.97 -7.87  9.35 33.04  -9.94
V5  5.72  -8.72 -13.94  -9.94 26.96
```

```
> source("r-qqchi2.txt")
```

```

> qqchi2(da)
[1] "correlation coefficient:"
[1] 0.9962366

> da=as.matrix(da)
> m1=manova(da~fac1+fac2+fac1*fac2)
> m1
Call:
  manova(da ~ fac1 + fac2 + fac1 * fac2)

Terms:
              fac1      fac2 fac1:fac2 Residuals
resp 1          215.412   53.727    0.077  4213.561
resp 2           0.641   66.821    3.660  2188.108
resp 3          172.816   24.027    4.863  3648.263
resp 4          654.514   13.673    0.068  3594.245
resp 5           73.575  308.726    4.147  3091.129
Deg. of Freedom      1        1        1        126

Residual standard error: 5.782816 4.167246 5.380935 5.340951 4.953057
Estimated effects may be unbalanced
> summary(m1,'Wilks')
              Df  Wilks approx F num Df den Df   Pr(>F)
fac1           1 0.7873   6.5930     5   122 1.825e-05 ***
fac2           1 0.8717   3.5906     5   122 0.004593 **
fac1:fac2      1 0.9955   0.1109     5   122 0.989764
Residuals 126
---

> v1=aov(da[,1]~fac1+fac2+fac1*fac2)
> summary(v1)
              Df Sum Sq Mean Sq F value Pr(>F)
fac1           1  215.4   215.4   6.4415 0.01237 *
fac2           1   53.7    53.7   1.6066 0.20730
fac1:fac2      1    0.1     0.1   0.0023 0.96168
Residuals    126 4213.6    33.4

```