

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

Solutions to Midterm

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 $\mathbf{Z} = (X_1 + X_2, X_1 - X_2)'$?

Answer: Bivariate normal with mean $\boldsymbol{\mu} = [3, -1]'$ and covariance matrix $\begin{bmatrix} 7 & 1 \\ 1 & 3 \end{bmatrix}$.

- (b) What is the distribution of X_1 given $X_2 = 1$ and $X_3 = 2$?

Answer: Using Result 4.6, the distribution is $N(1/6, 13/6)$.

- (c) Find a linear combination of \mathbf{X} of length 1 that has the minimum variance among all possible linear combinations.

Answer: The linear combination is $(0.591, -0.737, -0.328)'$, i.e. the eigenvector of the smallest eigenvalue.

- (d) Find the coefficient $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2)'$ of the least squares regression

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

Answer: From the least squares formula

$$\begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix} = \begin{bmatrix} 3 & 1 \\ 1 & 2 \end{bmatrix}^{-1} \begin{bmatrix} 1 \\ 0 \end{bmatrix} = \begin{bmatrix} 0.4 \\ -0.2 \end{bmatrix}.$$

Also, $\beta_0 = \mu_3 - \beta_1 \mu_1 - \beta_2 \mu_2 = 3 - 0.4 \times 1 + 0.2 \times 2 = 3$.

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:

- (a) 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.

Answer: The null hypothesis is $H_o : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2$ and the alternative hypothesis is $H_a : \boldsymbol{\mu}_1 \neq \boldsymbol{\mu}_2$, where $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$ are the mean vector for the Aa and Baa bonds. The Hotelling T^2 statistic is 15.84 with p-value 0.014. Thus, we reject the null hypothesis, i.e. the mean vectors of the two categories of bonds are significantly different.

- (b) Construct the 95% simultaneous confidence intervals for the components of the differences in mean vectors.

Answer: The 95% simultaneous confidence intervals for the components of the differences in mean vectors are

$$\begin{aligned} [1] & -1.014 \quad 0.78 \\ [2] & 0.43 \quad 10.46 \\ [3] & -0.355 \quad -0.001 \\ [4] & -4.29 \quad 8.27 \end{aligned}$$

- (c) Construct the Bonferroni 95% simultaneous confidence intervals for the components of the differences in mean vectors.

Answer: The Bonferroni confidence intervals are

```
> tt=abs(qt(0.05/8,38))
> for (i in 1:4){
+ lo=di[i]-tt*sqrt(Sp[i,i]*(1/10))
+ up=di[i]+tt*sqrt(Sp[i,i]*(1/10))
+ print(c(lo,up))
+ }
[1] -0.811  0.577
[1]  1.560  9.330
[1] -0.313 -0.041
[1] -2.870  6.850
```

- (d) Test the equality between the two covariance matrices. Draw the conclusion.

Answer: Use the Box-M test. The test statistic is 27.39 with p-value 0.002.

```
> det1=det(S1)
> det2=det(S2)

> u=(2/19-1/38)*(2*16+12-1)/(6*5*1)
> u
[1] 0.1131579

> M=38*log(det(Sp))-(19*log(det1)+19*log(det2))
> M
[1] 30.87986
> C=(1-u)*M
> C
[1] 27.38556
> v=0.5*4*5*1
> pp=1-pchisq(C,v)
> pp
[1] 0.002262336
```

- (e) Assuming the populations are normal with unequal covariance matrices. Test for the equality in the mean vectors. Draw your conclusion.

Answer: The answer the same as that of Part (a), because the two groups have the same sample size.

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.

- (a) Are the data normally distributed at the 5% level? (See the qqchi2 result)

Answer: The Chi-square QQ plot has a correlation 0.996, which compared with critical values of Table 4.2 cannot reject the normality assumption.

- (b) Is there any interaction effect between gender and socioeconomic status? Why?

Answer: No, there is no interactions. The Wilks' Lambda test statistic has a p-value of 0.9898.

- (c) Did the socioeconomic status affect the test result? Why?

Answer: Yes, the socioeconomic status affects the test result. The p-value of the Wilks' lambda statistic is 0.0046, which is small.

- (d) Did the gender affect the test result? Why?

Answer: Yes, the gender has a significant impact on the test scores at the 5% level.

- (e) Focus on the score of *leader*. Construct a univariate two-way analysis of variance table from the output.

Answer: From the output the analysis of variance table is

Source	Df	Sum Sq.	Mean Sq.	F-value	Pr(> F)
fac1	1	73.575	73.573	2.999	0.086
fac2	1	308.726	308.726	12.584	0.0005
fac1:fac2	1	4.147	4.147	0.169	0.68
Residuals	126	3091.129	24.53		

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:

- (a) Write down the assumptions for which the ordinary least squares estimates are also the maximum likelihood estimates of the vector $\boldsymbol{\beta}$.

Answer: The assumptions are

- i. \mathbf{Z} is of full rank $r + 1$.
- ii. $E(\boldsymbol{\epsilon}) = \mathbf{0}$.
- iii. $\text{Cov}(\boldsymbol{\epsilon}) = \sigma^2 \mathbf{I}$.

iv. $\boldsymbol{\epsilon}$ is multivariate normal.

- (b) 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 $\hat{\boldsymbol{\epsilon}}$.

Answer: $\mathbf{Z}'\hat{\boldsymbol{\epsilon}} = \mathbf{Z}'(\mathbf{Y} - \mathbf{Z}\hat{\boldsymbol{\beta}}) = \mathbf{Z}'(\mathbf{I} - \mathbf{H})\mathbf{Y} = (\mathbf{Z}' - \mathbf{Z}')\mathbf{Y} = \mathbf{0}$. Since the first column of \mathbf{Z} is the n -dimensional vector of 1, we have $\sum_{i=1}^n \hat{\epsilon}_i = 0$.

- (c) Derive the covariance matrix of $\hat{\boldsymbol{\epsilon}}$.

Answer: Using $\hat{\boldsymbol{\epsilon}} = (\mathbf{I} - \mathbf{H})\mathbf{Y}$, we have $\text{Cov}(\hat{\boldsymbol{\epsilon}}) = (\mathbf{I} - \mathbf{H})\text{Cov}(\mathbf{Y})(\mathbf{I} - \mathbf{H})' = \sigma^2(\mathbf{I} - \mathbf{H})^2 = \sigma^2(\mathbf{I} - \mathbf{H})$, where we use the fact that $\mathbf{I} - \mathbf{H}$ is idempotent.

- (d) Let $\text{tr}(\mathbf{H})$ be the trace of the hat matrix \mathbf{H} . Prove that $\text{tr}(\mathbf{H}) = r + 1$ and $h_{jj} > 0$ for all j , where h_{jj} is the (j, j) th element of \mathbf{H} .

Answer: From the definition, $h_{jj} = \mathbf{Z}'_j(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}_j$, where \mathbf{Z}'_j is the j th row of the design matrix \mathbf{Z} . Since $\mathbf{Z}'\mathbf{Z}$ is positive definite, $h_{jj} > 0$.

Next, $\text{tr}(\mathbf{H}) = \text{tr}(\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}') = \text{tr}[(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{Z}] = \text{tr}(\mathbf{I}_{r+1}) = r + 1$.

- (e) A high h_{jj} indicates high leverage. Why? Please explain.

Answer: The fitted values are $\hat{\mathbf{Y}} = \mathbf{Z}\hat{\boldsymbol{\beta}} = \mathbf{H}\mathbf{Y}$ so that $\text{hat}Y_j = \sum_{k=1}^n h_{jk}Y_k = h_{jj}Y_j + \sum_{k \neq j} h_{jk}Y_k$. Thus, the fitted value \hat{Y}_j is closer to Y_j if h_{jj} is larger. In this way, h_{jj} plays an important role in determining the linear regression line and it is called the *leverage*. A high h_{jj} means the j th data point has high leverage.

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 $\boldsymbol{\Sigma}_1$ and $\boldsymbol{\Sigma}_2$, respectively, then $\mathbf{X} = (\mathbf{X}'_1, \mathbf{X}'_2)'$ follows a multivariate normal distribution. True or false? Why?

Answer: False, the marginal distributions cannot determine the joint distribution. The only exception is when \mathbf{X}_1 and \mathbf{X}_2 are independent.

- (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 $\text{Cov}(\mathbf{X}_i) = \boldsymbol{\Sigma}$ for all i . Then, $\sqrt{n}(\bar{\mathbf{X}} - \boldsymbol{\mu})$ is approximately $N_p(\mathbf{0}, \boldsymbol{\Sigma})$ for large n , where $p = \dim(\mathbf{X}_i)$. True or false? Why?

Answer: False. When \mathbf{X}_i are serially correlated, then the variance matrix of the sample mean is not the covariance matrix of \mathbf{X}_i . See the vector AR(1) example discussed in the class.

- (c) Give two methods that can be used to select a linear regression model when there are r explanatory variables.

Answer: Any two of (a) Stepwise regression, (b) Mallows's C_p , (c) AIC or BIC criterion, or (d) the Boosting method.

- (d) Define the Cook distance of a linear regression model with r explanatory variables and n data points.

Answer: The Cook distance is defined as

$$D_i = \frac{(\hat{\boldsymbol{\beta}}_{(i)} - \hat{\boldsymbol{\beta}})'(\mathbf{Z}'\mathbf{Z})(\hat{\boldsymbol{\beta}}_{(i)} - \hat{\boldsymbol{\beta}})}{(r+1)s^2},$$

where $\hat{\boldsymbol{\beta}}_{(i)}$ is the least squares estimate of $\boldsymbol{\beta}$ when the i th data point is removed from the data and s^2 is the residual mean squared errors of the full regression, i.e. $s^2 = \hat{\boldsymbol{\epsilon}}'\hat{\boldsymbol{\epsilon}}/(n-r-1)$, where $\hat{\boldsymbol{\epsilon}} = (\mathbf{I} - \mathbf{H})\mathbf{Y}$.

- (e) If \mathbf{X} is distributed as $N_p(\mathbf{0}, \boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma}$ is positive definite, then $\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{X}$ is distributed as χ_p^2 . True or false? Why?

Answer: Yes, because $\mathbf{Z} = \boldsymbol{\Sigma}^{-1/2}\mathbf{X}$ is $N_p(\mathbf{0}, \mathbf{I})$ so that $\mathbf{Z}'\mathbf{Z} = \mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{X} \sim \chi_p^2$.