THE UNIVERSITY OF CHICAGO
Booth School of Business
Business 41912, Spring Quarter 2020, Mr. Ruey S. Tsay

Midterm

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

Signature: **Name**: **ID**:

Notes:

1. The exam has two parts. The first part consists of some simple questions and the second part focuses on data analysis.

2. The total time of exam is 180 minutes, including submission of your answers, once you start your exam.

3. All tests use 5% type-I error.

4. The transpose of the matrix $\boldsymbol{A}$ is $\boldsymbol{A}'$. Also, the (i, j)th element of the matrix $\boldsymbol{A}$ is a_{ij}.

Part One. Basic concepts.

Problem A. (10 points). Are the following statement *True or False*?:

1. If X_1 and X_2 are normally distributed, then $\boldsymbol{X} = (X_1, X_2)'$ is bivariate normal.

2. If X_1 and X_2 are normally distributed, then $X = 2X_1 - 3X_2$ is normal.

3. If X_1, X_2, and X_3 are jointly normal, X_1 is independent of X_2, and X_1 is also independent of X_3, then X_2 and X_3 are independent.

4. If X_1 is normal and $X_2|X_1 = x$ is normal for all x, then $\boldsymbol{X} = (X_1, X_2)'$ is bivariate normal.

5. If $\boldsymbol{X}$ is a p-dimensional normal random vector, then $\boldsymbol{a}'\boldsymbol{X}$ is normal, where $\boldsymbol{a}$ is a p-dimensional constant vector in R^p.

Problem B. (10 points). Suppose that $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ are random samples from a p-dimensional multivariate normal distribution with mean $\boldsymbol{\mu}$ and positive-definite covariance matrix $\boldsymbol{\Sigma}$. Let $\boldsymbol{S}_x = \frac{1}{n-1}\sum_{i=1}^n (\boldsymbol{X}_i - \bar{\boldsymbol{X}})(\boldsymbol{X}_i - \bar{\boldsymbol{X}})'$ be the sample covariance matrix and $\bar{\boldsymbol{X}}$ be the sample mean. Let $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n$ be another random sample from the same distribution. Denote the sample mean and sample covariance matrix of this 2nd sample by $\bar{\boldsymbol{Y}}$ and $\boldsymbol{S}_y$, respectively. The two random samples are independent.

1. Obtain the mean and covariance matrix of $\bar{\boldsymbol{X}} - \bar{\boldsymbol{Y}}$.

2. What is the distribution of $\boldsymbol{X}_1 - \boldsymbol{Y}_1$?

3. What is the pooled covariance matrix estimate?

4. Consider the null hypothesis $H_0 : \boldsymbol{\mu} = \mathbf{0}$ vs $H_a : \boldsymbol{\mu} \neq \mathbf{0}$. Write down the test statistic, which is the most powerful for the hypotheses.

5. What is the distribution of $d_i^2 = n(\boldsymbol{X}_i - \boldsymbol{Y}_i)'\boldsymbol{S}_p^{-1}(\boldsymbol{X}_i - \boldsymbol{Y}_i)$, where $\boldsymbol{S}_p$ is the pooled covariance matrix estimate?

Problem C: (16 points). Consider the multiple linear regression (mlr) model,

$$y_i = \boldsymbol{Z}_i'\boldsymbol{\beta} + \epsilon_i, \quad i = 1, \ldots, n,$$

where $\boldsymbol{Z}_i = (Z_{1i}, \ldots, Z_{pi})'$ is a p dimensional vector and ϵ_i is a random noise with mean 0 and variance σ^2. In matrix form, the model can be written as

$$\boldsymbol{Y}_{n\times 1} = \boldsymbol{Z}_{n\times p}\boldsymbol{\beta}_{p\times 1} + \boldsymbol{\epsilon}_{n\times 1},$$

where $\boldsymbol{Z}$ is a full rank design matrix and we assume that ϵ_i are uncorrelated. Assume that $n > p$. Let $\widehat{\boldsymbol{Y}} = \boldsymbol{Z}\widehat{\boldsymbol{\beta}}$ be the fitted value of the mlr model and $\widehat{\boldsymbol{\epsilon}} = \boldsymbol{Y} - \widehat{\boldsymbol{Y}}$ be the residual vector, where $\widehat{\boldsymbol{\beta}}$ is the ordinary least squares estimate of $\boldsymbol{\beta}$. Also, let $\boldsymbol{H} = \boldsymbol{Z}(\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}'$ be the hat matrix of the regression.

1. (True or False) Let $\boldsymbol{u}$ be a n-dimensional vector of 1. Then, $\boldsymbol{u}'\widehat{\boldsymbol{\epsilon}} = 0$.

2. (True of False) $\sum_{i=1}^n h_{ii} = p$.

3. (True or False) $h_{ii} > 0$ for all i.

4. (True or False) $\text{Cov}(\widehat{\boldsymbol{\beta}}) = \sigma^2(\boldsymbol{Z}'\boldsymbol{Z})^{-1}$.

5. (True or False) Residuals $\hat{\epsilon}_i$ and $\hat{\epsilon}_j$ are uncorrelated if $i \neq j$.

6. (True or False) $E(\frac{1}{n}\widehat{\boldsymbol{\epsilon}}'\widehat{\boldsymbol{\epsilon}}) = \sigma^2$.

7. (True of False) The eigenvalues of $\boldsymbol{H}$ are either 0 or 1.

8. (True or False) Influential points are always outliers.

Problem D. (8 points). Let $\boldsymbol{X} = (X_1, X_2, X_3)'$ be a random vector with mean zero and covariance matrix

$$\boldsymbol{\Sigma} = \begin{bmatrix} 2.0 & 0.3 & 0.4 \\ 0.3 & 3.0 & -0.5 \\ 0.4 & -0.5 & 2.0 \end{bmatrix}.$$

1. Find the linear combination of $\boldsymbol{X}$ that has the maximum variance.
2. Find the linear combination of $\boldsymbol{X}$ that has the smallest variance.
3. Let $\boldsymbol{x}$ be a 3-dimensional vector. Find the value of
$$\max_{\boldsymbol{x} \neq \boldsymbol{0}} \frac{\boldsymbol{x}'\boldsymbol{\Sigma}\boldsymbol{x}}{\boldsymbol{x}'\boldsymbol{x}}.$$
4. Find the value of
$$\min_{\boldsymbol{x} \neq \boldsymbol{0}} \frac{\boldsymbol{x}'\boldsymbol{\Sigma}\boldsymbol{x}}{\boldsymbol{x}'\boldsymbol{x}}.$$

Problem E. (10 points). Let $\{x_1, \ldots, x_n\}$ be a sample of n data points. Let μ and σ^2 be the mean and variance of x_i. Answer briefly the following questions.

1. If the sample is random, what is the variance of the sample mean $\bar{x} = \sum_{i=1}^{n} x_i/n$.
2. If x_t follows the model $x_t = \phi_0 + \phi_1 x_{t-1} + a_t$, where $0 < |\phi_1| < 1$, ϕ_0 is a constant, and $\{a_t\}$ is an iid sequence of normal random variates with mean zero and variance 1, what is the variance of the sample mean $\sqrt{n}\bar{x}$ as $n \to \infty$?
3. Let s^2 be the sample variance. If the sample is random, what is the limiting distribution of $\sqrt{n}(\bar{x} - \mu)/s$?
4. If x_t follows the AR(1) model of question 2, what is the limiting distribution of $\sqrt{n}(\bar{x} - \mu)/s$ as $n \to \infty$?
5. Discuss the impact of serial correlations on the one-sample t-test.

Part Two: Data analysis

Problem F. (10 points). In an experiment sometime ago, George Box studied the amount of fabric wear y_1, y_2, and y_3 in three successive periods: (1) the first 1000 revolutions, (2) the second 1000 revolutions, and (3) the third 1000 revolutions of the abrasive wheel. There are three factors: proportion of filter (P), type of abrasive surface (S), and type of filter (I). There are also two replications, the variation of which is ignored for this problem. The data are in the file `WEAR.DAT` of the course web. Answer the following questions:

1. Perform a multivariate analysis of variance. Show the table.
2. Provide a summary using the Wilk's test.
3. Is there a three-way interaction P*S*I?
4. Are there any two-way interaction (P*S, S*I, P*I)?
5. Are the three main effects significant?

Problem G. (18 points). Consider some temperature measurements. The data are in the file `TEMPERATURE.DAT`. The measurements are

1. y_1: maximum daily air temperature
2. y_2: minimum daily air temperature
3. y_3: integrated area under daily air temperature curve (i.e., average air temperature)
4. y_4: maximum daily soil temperature
5. y_5: minimum daily soil temperature
6. y_6: integrated area under daily soil temperature curve
7. y_7: maximum daily relative humidity
8. y_8: minimum daily relative humidity
9. y_9: integrated area under daily relatively humidity
10. y_{10}: total wind, measured in miles per day
11. y_{11}: evaporation

Let $\boldsymbol{Y}_1 = (y_1, y_2, y_3)'$, $\boldsymbol{Y}_2 = (y_4, y_5, y_6)'$, $\boldsymbol{Y}_3 = (y_7, y_8, y_9)'$. Perform the necessary analysis to answer the following questions:

1. Are the covariance matrices of $\boldsymbol{Y}_1$ and $\boldsymbol{Y}_2$ the same? Why?

2. Are the mean vectors of $\boldsymbol{Y}_1$ and $\boldsymbol{Y}_2$ the same? Why?

3. Construct the 95% Hotelling T^2, Bonferroni, marginal Student-t, and asymptotic confidence intervals for the difference in means between $\boldsymbol{Y}_1$ and $\boldsymbol{Y}_2$.

4. Regress $\boldsymbol{Y}_1$ on $\boldsymbol{Y}_2$, i.e., $\boldsymbol{Y}_1$ is the dependent variable. Write down the estimated coefficient matrix? Is the regression significant?

5. What are the fitted values of $\boldsymbol{Y}_1$ if $\boldsymbol{Y}_2 = (90.7, 70.1, 190.5)'$?

6. Construct simultaneously 95% confidence intervals for the elements of $\boldsymbol{Y}_1$ given $\boldsymbol{Y}_2 = (90.7, 70.1, 190.5)'$.

7. Construct simultaneous 95% prediction intervals for the elements of $\boldsymbol{Y}_1$ given $\boldsymbol{Y}_2 = (90.7, 70.1, 190.5)'$.

8. Regress the three humidity measures on $\boldsymbol{Y}_1$. Write down the estimated coefficient matrix? Is the regression significant?

9. Regress the three humidity measures on $\boldsymbol{Y}_1$ and $\boldsymbol{Y}_2$. Is the contribution of $\boldsymbol{Y}_2$ significant in this particular instance? Why?

Problem H: (10 points). Consider, again, the Temperature data of Problem G. Now, let y_{11} (evaporation) be the dependent variable and all the other variables as predictors. Answer the following questions:

1. What is the model selected by the stepwise method? What is the minimum AIC value?

2. Based on the selected model in part 1, how does evaporation relate to temperature and relative humidity?

3. What is the model selected by the Mallow's C_p?

4. What is the best model selected if two predictors are used?

5. What is the best model selected if three predictors are used?

Problem I: (8 points). Consider the data set **ProblemI.csv**, where the first column is the dependent variable and the other columns are predictors. There are 100 observations with 500 predictors. Use the package **glmnet** to answer the following questions.

1. Obtain a coefficient profile plot of the Lasso regression.

2. Use cross validation to select the penalty parameter. What is the selected penalty parameter?

3. Use 0.2 as the threshold to select the predictors. That is, find the Lasso coefficients greater than 0.2 (in absolute value). Identify those predictors and their coefficient estimates.

4. Run a linear regression of the dependent variable on the selected predictors of question 3. Compare the estimates with those of the Lasso regression. Comment on the biases of the Lasso regression.