STAT 8003 Final Exam, Fall 2013

This is a take home final exam and is due Friday, Dec. 13th by 5 pm. Please upload your answer (pdf file) to blackboard with the file name with your Last name and first name, for example,

XieJichun.pdf

Each sub-problem (marked by alphabetic letter) is 10 points, and the total score is 100 points.

You can consult any references but cannot speak with anyone (except for Prof. Xie) about the exam. If you do not understand a question, send Prof. Xie an e-mail (jichun@temple.edu). Good luck!

Problem 1. Scientists measured the weather and wind data at 3 hour intervals for Gabo Island, off the eastern tip of Victoria, during 1989. Observations were made 7 times a day for 7 days. Thus, in total there are 49 observations.

Variable	Description
WetTemp	Wet bulb temperature
Pressure	Barometer pressure
Humidity	Relative humidity in percent
Speed	Wind speed in knots

Now suppose the outcome is $\mathbf{Y} = (Y_1, \ldots, Y_n)^{\mathrm{T}}$, where Y_i is the wet bulb temperature of the ith observation. Suppose $\mathbf{X}_1$, $\mathbf{X}_2$ and $\mathbf{X}_3$ are the observations for pressure, humidity and speed, respectively.

a) Plot a scatterplots for the data. Describe the pattern you observe. Fit a linear model for the data. Interprete your results.

Problem 2. Consider the same example in Problem 1. Since the observation are recorded subsequently, it is resonable to assume that the unmeasured the effects of the temperature are correlated. We consider the following model:

$$Y_k = \beta_0 + \beta_1 X_{k1} + \beta_2 X_{k2} + \beta_3 X_{k3} + \epsilon_k. \tag{1}$$

Here $\epsilon_k \sim \mathrm{N}(0, \sigma^2)$, but they are not independent from each other. We assume that $\boldsymbol{\epsilon} = (\epsilon_1, \dots, \epsilon_n) \sim \mathrm{N}(\mathbf{0}, \boldsymbol{\Sigma})$, with $\boldsymbol{\Sigma}$ known.

a) When the error terms are correlated, we cannot use ordinary least squares to find the solution. But instead we can use a modified version. Define the new RSS function:

$$RSS = (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}), \tag{2}$$

with $\mathbf{X}$ is the design matrix with the intercept. The general least square estimator $\hat{\boldsymbol{\beta}}$ is obtained by minimizing (2). Please show that

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} \mathbf{X})^{-1} (\mathbf{X}^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} \mathbf{Y}).$$

Also show $\mathtt{E}(\hat{\boldsymbol{\beta}})$ and $\mathtt{Var}(\hat{\boldsymbol{\beta}})$.

b) Suppose a statistician ignores the effect of $\boldsymbol{\Sigma}$, so that under model (1) he uses the ordinary least square estimators $\tilde{\boldsymbol{\beta}} = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}(\mathbf{X}^{\mathrm{T}}\mathbf{Y})$ to estimate $\boldsymbol{\beta}$. For any given vector $\mathbf{a}$, suppose we are interested in estimating $\mathbf{a}^{\mathrm{T}}\boldsymbol{\beta}$. Now the statistician uses $\mathbf{a}^{\mathrm{T}}\tilde{\boldsymbol{\beta}}$ to estimate $\mathbf{a}^{\mathrm{T}}\boldsymbol{\beta}$. Is the estimator unbiased? And what's the variance? If you are the statistician, will you use $\mathbf{a}^{\mathrm{T}}\tilde{\boldsymbol{\beta}}$ or $\mathbf{a}^{\mathrm{T}}\hat{\boldsymbol{\beta}}$ to estimate $\mathbf{a}^{\mathrm{T}}\boldsymbol{\beta}$? Why? (Hint: Let $A = \boldsymbol{\Sigma}^{-1/2}$. Consider a equivalent linear model $\tilde{\mathbf{Y}} = A\mathbf{Y}$ and $\tilde{\mathbf{X}} = A\mathbf{X}$, and obtain its LSE.)

Problem 3. In many studies, a large number of independent variables, $X_1, \dots, X_p$, are measured. However, it may be impractical to include all these variables in a linear regression model. One way to reduce the dimensionality of the model is via principal components; a linear combination of the variables. The ith principal component, Z_i, is given by

$$Z_i = \mathbf{a}_i^{\mathrm{T}} \mathbf{X} = a_{i1} X_1 + \cdots + a_{ip} X_p$$

such that

$$\begin{gathered} \mathbf{a}_i^{\mathrm{T}}\mathbf{X} \text{ maximizes } \mathtt{Var}(\mathbf{a}_i^{\mathrm{T}}\mathbf{X}) \\ \text{subject to } \mathbf{a}_i^{\mathrm{T}}\mathbf{a}_i = 1, \text{ and } \mathtt{Cov}(Z_i, Z_k) = 0, \text{ for } k \neq i. \end{gathered}$$

Since principal component analysis forcusses on maximizing the variance of the independent variables, the theorems for matrices are useful for understanding the properties of the principal components. Once such theorem is the maximization for quadratic forms. That is, given a positive definite matrix $\mathbf{B}$

$$\max_{\mathbf{x} \neq 0,\ \mathbf{x} \perp \mathbf{e}_1, \dots, \mathbf{e}_{k-1}} \frac{\mathbf{x}^{\mathrm{T}}\mathbf{B}\mathbf{x}}{\mathbf{x}^{\mathrm{T}}\mathbf{x}} = \lambda_k, \quad \text{and the maximization achieves when } \mathbf{x} = \mathbf{e}_k$$

where $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$ are the eigenvalues and $\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_p$ are the associated normalized eigenvectors of $\mathbf{B}$.

a) Using the maximization theorem for quadratic forms mentioned above, or otherwise, show that (i) $Z_i = \mathbf{e}_i^{\mathrm{T}}\mathbf{X}$, (ii) $\mathtt{Var}(Z_i) = \lambda_i$, (iii) $\sum_{k=1}^{p} \mathtt{Var}(X_k) = \sum_{i=1}^{p} \mathtt{Var}(Z_i)$, where e_i is the ith eigenvector of $\mathtt{Var}(X)$, corresponding to the ith largest eigenvalues of $\mathtt{Var}(X)$.

b) In practice, the linear model

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \epsilon,$$

can be replaced with

$$Y = \alpha_0 + \alpha_1 Z_1 + \cdots + \alpha_k Z_k + \epsilon, \quad k \leq p.$$

Explain how you would determine k.

c) An investigator calculate a linear regression model with 24 independent variables. He first centered and standardized the outcome Y and all covariates $X_1, \ldots, X_p$, then perform linear models. The resulting ANOVA table is shown in Table 1. The investigator then calculate the principal components for the independent variables (after centering and standardization), and ran another linear regression model using the first 3 principal components. The ANOVA table based on the principal components appears in Table 2. Which model would you recommend using? Can you formulate a hypothesis test to answer the question? What's your testing result?

Source	df	MS	F	Sig.
Model	24	76.20	9.47	< .001
Error	25	8.05		
Total	49			

Table 1: ANOVA Table for Linear Regression Model

Source	df	MS	F	Sig.
Model	3	570.20	82.13	< .001
Error	46	6.94		
Total	49			

Table 2: ANOVA Table for Principal Components Model

Problem 4. One market monitoring organization would like to compare the life time of two brands of bulbs, Brand A and Brand B. They design the experiment in this way. Let X_i and Y_i be the life time of ith bulb in Brand A and Brand B respectively, which can be approximated by independent random variables with exponential distributions with expectations λ and μ

respectively. They pair X_i and Y_i. In the ith experiment, instead of letting both two bulbs burn until they die out, they stop when one of the bulbs burn out, and record the burning time Z_i and indicator W_i of which one burns out. They repeat the experiment n times. Mathematically, Z_i and W_i can be defined as

$$Z_i = \min(X_i, Y_i) \text{ and } W_i = \begin{cases} 1 & \text{if } Z_i = X_i, \\ 0 & \text{if } Z_i = Y_i; \end{cases} \qquad i = 1, \ldots, n$$

a) Find closed form expressions for the maximum likelihood estimators of λ and μ. Note: Justify that your estimator is in fact the global maximizer of the likelihood. Use the data "bulb.txt" to compute the MLE.

b) Consider applying the EM algorithm with the complete data taken to be $(X_1, Y_1), \ldots, (X_n, Y_n)$. Show that the EM sequence is given by

$$\begin{aligned} \hat{\lambda}^{(k+1)} &= \frac{\sum_{i=1}^n W_i Z_i + \sum_{i=1}^n (1 - W_i)(Z_i + \lambda^{(k)})}{n} \\ \hat{\mu}^{(k+1)} &= \frac{\sum_{i=1}^n (1 - W_i) Z_i + \sum_{i=1}^n W_i (Z_i + \mu^{(k)})}{n} \end{aligned}$$

Use the data "bulb.txt" to find a solution, using the starting value $\mu^{(0)} = 1$ and $\lambda^{(0)} = 1$.

Note: For the model in Problem 4, the EM algorithm is not needed because a closed form expression for the MLE is available. But for other related models with censored data, no closed form expression is available and the EM algorithm is useful.

Problem 5. The incidence of a rare disease seems to be decreasing. In successive years, the number of new cases is $y_1, \ldots, y_n$. We assume that $y_1, \ldots, y_n$ are independent random variables from Poisson distributions with means $\theta, \theta^2, \ldots, \theta^n$ respectively.

a) Forumlate a likelihood ratio test for testing $\mathrm{H}_0 : \theta = 1$ versus $\mathrm{H}_a : \theta < 1$. For $(y_1, y_2) = (2, 0)$, would such test with size 0.20 test accept or reject H_0?

b) Describe a procedure for forming a level 0.95 one-sided confidence interval of the form $(0, \theta_u)$ [you do not need to come up with a closed form expression and can express that you would need to calculate the quantiles of certain distributions and do a numerical search to form the confidence interval]. Use your procedure to find (approximately) a realized confidence interval of the form $(0, \theta_u)$ for the sample $(y_1, y_2) = (2, 0)$ (you may want to write a computer program for this).