Qualifying Exam April 2017 — Statistical Methods

Instructions:

- NOTE: You are not allowed to use internet or email except for downloading the data and emailing the final report (optional). To use these for any other purpose, ask the proctor.
- Go to http://www.utdallas.edu/∼swati.biswas/QE and download the datasets crime and insurance. Let the proctor know if you have any problems with this step.
- You can use any software of your choice. You can use the lab machines or your own laptop.
- Your report should clearly explain the steps, results, conclusions, and justification for the conclusions. Also, include the codes and outputs (ONLY relevant parts; highlighted wherever possible). Do not attach the parts of the output that were not used in answering questions.
- Submit a report, written or typed, hard copy or by email to swati.biswas@utdallas.edu. If you choose to email, then attach only one single PDF file with the whole report. DO NOT email separate files for codes or outputs.
- Write your name on the report. DO NOT write your ID.

1. Consider the crime data. We would like to understand how murder rate is related to the other variables in this data set.

   (a) (15 points) Fit a model to predict *murder.rate* based on the other variables. Check assumptions and perform any transformations needed to obtain a model that is reasonable with respect to the standard assumptions for linear models. Also, perform collinearity and influential data diagnostics. **Note**: since the sample size is small, don't include interactions with *region*.

   (b) (10 points) Reduce your model by removing any unimportant variables if such variables exist. Interpret the reduced model including coefficients and r-squared. Perform a statistical test that compares the full model to the reduced model. Clearly state the hypotheses associated with this test and interpret the results.

   (c) (5 points) Use your final model to obtain 95% confidence and 95% prediction intervals for the murder rates of states. Are there any states with higher murder rates than the upper bound of the confidence intervals? prediction intervals?

2. (35 points) Use Monte Carlo simulation to compare the power functions of two tests for normality — Shapiro Wilk test and skeweness test. Let $\mathscr{N}$ denote the family of univariate normal distributions. The hypotheses are

$$H_0 : F_X \in \mathscr{N} \text{ vs. } H_1 : F_X \notin \mathscr{N}.$$

   For this investigation, consider the following family of normal mixture of two components:

$$(1-\epsilon)N(\mu = 0, \sigma^2 = 1) + \epsilon N(\mu = 0, \sigma^2 = 100),\ 0 \leq \epsilon \leq 1.$$

   When $\epsilon = 0$ or $\epsilon = 1$, the distribution is normal (i.e., H_0 is true). If $0 < \epsilon < 1$, the distributions are non-normal. Note that ϵ represents the probability that a random draw

comes from the second component. We are interested in comparing the empirical power functions of the two tests by varying ϵ value. For this comparison, set the significance level $\alpha = 0.1$ and sample size $n = 30$. For a fixed ϵ value, generate at least 1,000 Monte Carlo samples and compute powers of the two tests. Vary ϵ from 0 to 1 (both included) in increments of 0.1 (i.e., a total of 11 ϵ values). Report the powers in a table and plot the two power curves as function of ϵ in one plot. Use these to make power comparisons between the two tests.

Note: Recall that skewness of a random variable X with mean μ and variance σ is defined as $\gamma_1 = E(X-\mu)^3/\sigma^3$ and can be estimated from a random sample $X_1, X_2, \ldots, X_n$ by

$$k_3 = \frac{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^3}{s^3},$$

where $\bar{X}$ = sample mean and s = sample standard deviation (with denominator of n). The null distribution of k_3 is asymptotically normal with mean 0 and variance $6/n$. However, for finite sample, it is better to use the exact variance of k_3 given by $6(n-2)/[(n+1)(n+3)]$.

3. A health insurance company collected information on 600 of its subscribers who has claims resulting from a certain disease. Data were obtained on total cost of services provided for these subscribers and the nature of the various services for a certain year. The dataset (Insurance.txt) is provided, and the following table provides information.

Variable Number	Variable Name	Description
1	Identification number	1-600
2	Total cost	Total cost of claims by subscriber (dollars)
3	Age	Age of subscriber(years)
4	Gender	Gender of subscriber: 1 male; 0 female
5	Interventions	# of interventions: 1 $(0-1)$; 2 $(2-4)$; 3 (≥ 5)
6	Drugs	# of tracked drugs prescribed
7	Comorbidities	# of complications that subscriber had during period: 1 (0-1); 2 (≥ 2)

(a) (10 points) In the dataset, the response of interest is the total cost (variable 2). Use three variables (Gender: V4; Interventions: V5; Comorbidities: V7) as well as their interactions to fit a regression model. Check assumptions and use appropriate transformation technique, if needed.

(b) (10 points) Report an unbalanced three-way ANOVA table of this data set for the response and factors used in part (a) (name the three factors as Factor A, Factor B, and Factor C, respectively for this part). Make sure your ANOVA table won't change with the order of the factors. Explain clearly the Sum of Squares for sources A and AB in your ANOVA table.

(c) (5 points) For the main effect Intervention, obtain confidence intervals for all pairwise comparisons with familywise confidence coefficient 90%.

(d) (10 points) We now wish to investigate the nature of the interaction effects. Plot the interaction plot for Intervention*Comorbidities interaction. Then estimate separately for subscribers with intervention levels 1, 2, and 3 how large is the difference in mean

costs for the two comordibilities levels. Employ a multiple comparison procedure with familywise confidence coefficient 90%.

STATISTICS Ph.D. QUALIFYING EXAM
STATISTICAL INFERENCE I and II
April 2017

General Instructions: Write your ID number on all answer sheets. Do not put your name on any of your answer sheets. Show all work/proofs/references. Please write neatly so it is easy to read your solution. Total points = 100.

Problem 1. Consider a random sample of size n from the pdf

$$f(x|\theta) = 2x\theta^{-k},\ 0 < x < \theta,\ \theta > 0.$$

Find k, and then either find a complete sufficient statistic or prove that it does not exist. (10 points)

Problem 2. Consider a random sample of size n from a distribution with the density $f(x|\theta) = \theta x^{\theta-1} I(0 \le x \le 1)$, $\theta > 0$.
a. Find the MLE for θ. (5 points)
b. Find the MSE of the MLE. (7 points)

Problem 3. Consider a random sample of size $n\,(\ge 3)$ from an exponential distribution with mean $1/\theta$.
(a) Find a sufficient statistic T for θ. What is its distribution? (5 points)
(b) Find the MLE $\hat{\theta}$ for θ. (5 points)
(c) Show that the MLE $\hat{\theta}$ from (b) is biased for θ. Find a multiple of $\hat{\theta}$ that is unbiased for θ. (5 points)
(d) Compute the Cramer-Rao lower bound for the variance of an unbiased estimator of θ. (5 points)
(e) Do you expect the bound in (d) to be attained? Justify your answer. (5 points)
(f) Is the unbiased estimator in (c) the best unbiased estimator for θ? Justify your answer. (3 points)

Problem 4. Consider a random sample of size n from Poisson(Λ) distribution. Assume that the prior distribution for Λ is a Gamma(α, β) distribution.
a. Prove that this is the conjugate prior. (3 points)

b. Find the posterior distribution of Λ. (5 points)
c. Find the posterior mean, and explain why this mean is of statistical interest. (5 points)

Problem 5. Consider a random sample of size n from a distribution with density $f(x|\theta) = (2\theta)^{-1} I(-\theta < x < \theta)$, $\theta > 0$. Find a best unbiased estimator of θ or prove that it does not exist. (10 points)

Problem 6. Consider a variable X with density $f(x) = e^{-x}$, $x > 0$. You get one observation of $Y = X^{\theta}$. You are asked to test $H_0 : \theta = 1$ versus $H_1 : \theta = 2$ on the basis of Y. Find the rejection region of the UMP level α test. (12 points)

Problem 7. Consider a single observation X from Beta(θ,1) distribution.
(a) Let $Y = -1/\ln(X)$. Evaluate the confidence coefficient of the set $[Y/2, Y]$ as a confidence interval for θ. (8 points)
(b) Find a pivotal quantity and use it to set up a confidence interval for θ having the same confidence coefficient as the interval in part (a). (7 points)

Name _____ First Name _____

April 2017 Qualifying Exam in Probability Theory

- This is a closed-book test.
- There are 3 questions.
- Answer each question as fully as possible.
- Show and justify all steps of your solutions.
- Refer clearly to any known results that you are using, **stating such results precisely**.
- Show how the assumptions of a result you use are satisfied in your application of the result.
- Indicate how the assumptions given in the question are used in the solution.
- Write your solutions on the blank sheets of paper that are provided.
- Identify each sheet with your name.
- On each sheet, identify which question and part is being answered.
- Begin each question on a new sheet.
- *When finished, arrange your sheets in order, number each sheet, and be sure that your name is on each sheet.*

Name _____ First Name _____

EXERCISE 1

Let $\{X_n, n \geq 1\}$ *be a sequence of independent and identically uniformly distributed random variables on* $(0, 1)$. *Define the sequence* Y_n *as*

$$Y_n = \min(X_1, \ldots, X_n).$$

(a) Does Y_n *converge to* 0 *in distribution?*

(b) Does Y_n *converge to* 0 *in probability?*

(c) Does Y_n *converge to* 0 *in mean?*

(d) Does Y_n *converge to* 0 *almost surely?*

Justify your answers.

Name _____ First Name _____

EXERCISE 2

1. *Let $X_1, X_2, \ldots,$ be independent and identically distributed random variables with $E(X_1) = 0$ and $Var(X_1) = \infty$. Show that*

$$P(\limsup_{n\to\infty}\{|X_n| \geq \sqrt{n}\}) = 1$$

2. *Let A_n be a sequence that satisfies $P(A_n) = \frac{1}{2^n}$. Show that*

$$P(\limsup_{n\to\infty} A_n) = 0.$$

3. *Let $X_n \sim Exp(1)$ be independent random variables and define the set*

$$\mathcal{A} = \left\{ \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} X_i < \infty \right\}.$$

Show that $\mathcal{A}$ is a tail event.

Name

First Name

EXERCISE 3 *Let* $\{Z_n\}$ *be independent, each with finite mean. Let* $X_0 = a$, *and* $X_n = a + Z_1 + \ldots + Z_n$ *for* $n \geq 1$, *and let* $\mathcal{F}_n = \sigma(X_0, X_1, \ldots, X_n)$ *be a* σ*-algebra generated by* $X_0, X_1, \ldots, X_n$. *Prove that* $E(X_{n+1}|\mathcal{F}_n) = X_n + E(Z_{n+1})$ *w.p.1.*

Name _____ First Name _____

April 2017 Qualifying Exam in Linear Models

- This is a closed-book test.
- There are 3 questions.
- Answer each question as fully as possible.
- Show and justify all steps of your solutions.
- Refer clearly to any known results that you are using, **stating such results precisely**.
- Show how the assumptions of a result you use are satisfied in your application of the result.
- Indicate how the assumptions given in the question are used in the solution.
- Write your solutions on the blank sheets of paper that are provided.
- Identify each sheet with your student ID. .
- On each sheet, identify which question and part is being answered.
- Begin each question on a new sheet.
- *When finished, arrange your sheets in order, number each sheet, and be sure that your student ID is on each sheet.*
- Total points = 100.

EXERCISE 1

(20 points) Aerial observations Y_{ij}*, where* $i = 1, 2, 3$ *and* $j = 1, \ldots, n$*, are made of angles* θ_1, θ_2 *and* θ_3*, respectively, of a triangle on the ground. Assuming that the observations are subject to independent normal errors with zero means and common variance* σ^2*, derive an appropriate estimator of the vector* $\boldsymbol{\theta} = (\theta_1, \theta_2, \theta_3)'$*.*

EXERCISE 2 *Let* $\mathbf{Y} = \mathbf{Xb} + \boldsymbol{\epsilon}$ *be a linear model with* $E(\boldsymbol{\epsilon}) = \mathbf{0}$ *and* $Var(\boldsymbol{\epsilon}) = \sigma^2\mathbf{I}$ *and* $|\mathbf{X}'\mathbf{X}| = 0$, *where* $|\mathbf{A}|$ *denotes the determinant of the matrix* $\mathbf{A}$. *Let* $(\mathbf{X}'\mathbf{X})^-$ *denote any generalized inverse (g-inverse) of* $\mathbf{X}'\mathbf{X}$. *Assume that the linear combination* $\mathbf{c}'\mathbf{b}$ *is estimable.*

1. *(10 points) Show that* $\hat{\mathbf{b}} = (\mathbf{X}'\mathbf{X})^-\mathbf{X}'\mathbf{Y}$ *is a solution of the normal equations.*
2. *(10 points) Show that* $\mathbf{c}'\mathbf{b}$ *is estimable if and only if*

$$\mathbf{c}'(\mathbf{X}'\mathbf{X})^-(\mathbf{X}'\mathbf{X}) = \mathbf{c}'.$$

3. *(10 points) Show that* $\mathbf{c}'\hat{\mathbf{b}}$ *does not depend on the specific choice of the g-inverse.*
4. *(10 points) Show that* $\mathbf{c}'\hat{\mathbf{b}}$ *is the best linear unbiased estimator (BLUE) of* $\mathbf{c}'\mathbf{b}$.
5. *(10 points) Show that the estimator*

$$\widehat{\sigma}^2 = \frac{1}{N - rank(\mathbf{X})}\mathbf{Y}'(\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^-\mathbf{X}')\mathbf{Y}$$

is an unbiased estimator of σ^2.

EXERCISE 3 *Let* $\mathbf{Y}$ *be a* $n \times 1$ *vector. Suppose that* $\mathbf{Y} \sim N(\mathbf{0}, \mathbf{\Sigma})$ *and let* $\mathbf{A}$ *denote a* $n \times n$ *nonnegative definite symmetric matrix.*

1. *(10 points) Show that the quadratic form* $Q = \mathbf{Y}'\mathbf{A}\mathbf{Y}$ *can be represented by*

$$Q = \sum_{i=1}^{n} \lambda_i C_i,$$

   where the C_i*'s are independently distributed as* χ^2_1 *variables.*

2. *(10 points) Use part (1.) to show that if* $\mathbf{A}\mathbf{\Sigma}$ *is idempotent, then* Q *follows a central* χ^2 *distribution. Determine the degrees of freedom.*

3. *(10 points) Show that* $E(Q) \leq \beta_{max} \sum_{i=1}^{n} \sigma_{ii}$*, where* β_{max} *is the largest eigenvalue of* $\mathbf{A}$ *and* σ_{ii} *is the* i*th diagonal element of* $\mathbf{\Sigma}$*.*