Qualifying Exam August 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 diabetes.csv and tox.csv. 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 Angie.Bustamante@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 QE ID number (given to you by Angie) on all answer sheets. DO NOT put your name, UTD ID, or any other identifying information on any of your answer sheets. DO NOT email your exam to anyone other than Angie.**

1. Consider the diabetes dataset containing the following variables:

#	Variable	Description
1	id	Subject ID
2	stab.glu	Stabilized Glucose
3	ratio	Cholesterol to HDL ratio
4	glyhb	Glycosolated Hemoglobin
5	location	Location (county) of subject (Buckingham or Louisa)
6	age	Age
7	gender	Gender
8	frame	Body frame (small, medium, or large)
9	bmi	Body mass index
10	whip	Waist to Hip ratio

   (a) Fit a model for predicting glycosolated hemoglobin using all other variables. Test if the variables location, gender, and frame can be dropped from this model. Write down the appropriate hypotheses, extra sum of squares, F-statistic value, degrees of freedom, and p-value for this test. State your conclusion. [6 points]

   (b) Find the best model(s) using adjusted R^2 and AIC criteria, and stepwise selection method. Interpret one of the best model(s) including coefficients and r-squared. [7 points]

   (c) For one of the best model(s) chosen above, check all key assumptions. If an assumption is not met, attempt to remedy the situation. Conduct diagnostics for checking collinearity and influential data/outliers. Based on these, comment on the appropriateness of the model. [10 points]

(d) For the above selected model, use iteratively reweighted least squares approach to robust regression for dampening the influence of outlying cases. Use Huber weight function. The initial residuals may be obtained from OLS model. Carry out at least three iterations (excluding iteration 0) and compare the residuals, weights, and regression coefficients across iterations including the ones from OLS. [12 points]

2. Use Monte Carlo (MC) simulation to estimate coverage probability of a confidence interval for variance. Suppose $X_1, \ldots, X_n$ $(n > 2)$ be a random sample from a distribution F_x and S^2 is the sample variance with denominator $n - 1$. Consider a one-sided 100(1-α)% confidence interval for variance given by $(0, (n-1)S^2/\chi^2_{\alpha,n-1})$, where $\chi^2_{\alpha,n-1}$ is the upper αth quantile of the χ^2_{n-1} distribution. Using 1000 MC replicates, estimate the coverage probability of 95% confidence interval in each of the following settings and comment on the results: [30 points]

   (a) F_x is $N(\mu = 0, \sigma = 2)$, $n = 20$.

   (b) F_x is χ^2_2, $n = 20$.

   (c) F_x is χ^2_2, $n = 100$.

3. Researchers want to know how exposure to a particular toxic chemical may affect the development of mice. In a study, the toxicant was administered at doses of 0, 100, 200, or 300 mg/kg/day to 94 pregnant mice. Following sacrifice, fetal weight was recorded for each live fetus. The dataset contains the following variables:

#	Variable	Description
1	id	Unique ID for each fetus
2	litterID	Unique ID for each litter (mother mouse)
3	dose	Dose of the toxicant
4	weight	Weight of each fetus (10^{-2} Oz)

In our analysis, we focus on factors that may affect the fetal weight.

   (a) Use graphical tools to explore the effects of dose and litter size on the fetal weight. Is the dose-response trend linear? If not, what will you do? Clearly state your conclusions. [10 points]

   (b) Fit a fixed-effect linear model including the dose and litter size effects. Is there any interaction? Conduct tests for the main effects (using $\alpha = .05$) and state your conclusions. [8 points]

   (c) In the above model we ignore the fact that fetuses belonging to the same litter may be more similar than those from different litters. To address this concern, fit an appropriate model containing the same main effects as in the previous problem. Conduct tests for the main effects (using $\alpha = .05$) and state your conclusions. Compare your estimates of the dose effect and its standard errors in the two models. Are they similar or different? State your findings and explain why they are similar (or different). [12 points]

   (d) Without fitting actual models, can you fit a fixed-effect model including the dose and litter effects? Briefly explain your reasoning. [5 points]

STATISTICS Ph.D. QUALIFYING EXAM
STATISTICAL INFERENCE

August 2017

General Instructions: Write your QE ID number (given to you by Angie) on all answer sheets. Do NOT put your name, UTD ID, or any other identifying information on any of your answer sheets. Show all work/proofs/references. Justify all arguments. Simply answers as much as possible. Please write neatly so that it is easy to read your solution. Total points = 100.

1. (15 points) State and prove Basu's theorem. (Write down all relevant definitions and assumptions and explain all steps in the proof, including where and how the assumptions are used.)

2. Let $X_1, \dots, X_n$ be a random sample from a Weibull distribution with density function

$$f_\theta(x) = \frac{a}{\theta} x^{a-1} \exp\{-x^a/\theta\} I(0 < x < \infty),$$

where $a > 0$ and $\theta > 0$ are unknown.

(a) (8 points) Find a pivotal quantity for (a, θ).

(b) (7 points) Construct a $100(1-\alpha)\%$ joint confidence set for (a, θ) based on the pivotal quantity found in part (a).

3. (10 points) Let $X_1, \dots, X_n$ be a random sample from an exponential distribution with density function

$$f_{a,\theta}(x) = \frac{1}{\theta} \exp\{-(x-a)/\theta\} I(a < x < \infty),$$

where $a > 0$ is unknown and $\theta > 0$ is known. Find a level α likelihood ratio test for testing $H_0 : a \leq a_0$ versus $H_1 : a > a_0$, where a_0 is a known positive constant.

4. Let $X_1, \dots, X_n$ be a random sample from an exponential distribution on $(0, \infty)$ with scale parameter 1. Suppose that we observe $T = X_1 + \dots + X_\theta$, where θ is an unknown positive integer. Consider estimation

of θ under the loss function $L(\theta, a) = (\theta - a)^2/\theta$ and a geometric distribution with mean $1/p$ as the prior distribution for θ, where $p \in (0, 1)$ is known.

(a) (5 points) Show that the posterior expected loss is

$$E[L(\theta, a)|T = t] = 1 + \xi - 2a + (1 - e^{-\xi})a^2/\xi,$$

where $\xi = (1 - p)t$.

(b) (5 points) Find the Bayes estimator of θ and show that its posterior expected loss is $1 - \xi \sum_{m=1}^{\infty} e^{-m\xi}$.

(c) (5 points) Find the marginal distribution of $(1 - p)T$, unconditional on θ.

(d) (5 points) Obtain an explicit expression for the Bayes risk of the Bayes estimator in part (b).

5. Let $X_1, \ldots, X_n$ be a random sample from a $N(\mu, \sigma^2)$ distribution, where $-\infty < \mu < \infty$ is unknown and $\sigma^2 > 0$ is known.

   (a) (7 points) Find the UMVUE of μ^3.

   (b) (8 points) Find the UMVUE of $P(X_1 \leq t)$ for a fixed real number t.

6. (10 points) Let X and Y be two random variables such that Y has a Binomial (N, π) distribution, and $X|Y = y$ has a Binomial (y, p) distribution. Suppose both $p \in (0, 1)$ and $\pi \in (0, 1)$ are unknown and N is known. Prove or disprove: (X, Y) is minimal sufficient for (p, π).

7. (15 points) Let $X_1, \ldots, X_n$ be a random sample from a uniform distribution on the interval $[0, 1]$ and let $R = X_{(n)} - X_{(1)}$ denote the sample range. Here $X_{(i)}$ is the ith order statistic. Derive the density function of R and show that the limiting distribution of $2n(1 - R)$ is a χ^2 distribution with 4 degrees of freedom.

QE ID

August 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.
- Write your QE ID number (given to you by Angie) on all answer sheets. Do NOT put your name, UTD ID, or any other identifying information on any of your answer sheets.
- 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 QE ID is on each sheet.*

QE ID

EXERCISE 1

Suppose that $X_1, X_2, \ldots$ are independent, identically distributed random variables defined on $\mathbb{R}$, with common distribution function F_X for which $F_X(x) < 1$ for all finite x. Let M_n be the maximum random variable defined for finite n by

$$M_n = \max(X_1, \ldots, X_n).$$

(a) Show that the sequence of random variables $\{M_n\}$ converges almost surely to infinity, that is, as $n \to \infty$

$$M_n \overset{a.s.}{\to} \infty.$$

Hint: use the Borel-Cantelli lemma.

(b) Now suppose that $F_X(x_U) = 1$ for some $x_U < \infty$. Find the almost sure limiting random variable for the sequence $\{M_n\}$.

Justify your answers.

QE ID

EXERCISE 2 *Let $\{X_n, n \geq 1\}$ denote a sequence of random variables. The next two questions are not related to each other.*

1. *Assume that there exists a constant θ and a sequence of constants a_n with $\lim_{n\to\infty} a_n \to 0$ such that*

$$\frac{X_n - \theta}{a_n} \overset{\mathcal{L}}{\to} N(0,1).$$

Prove that $X_n \overset{P}{\to} \theta$.

2. *Suppose that $X_1, \ldots, X_n$ are independent and identically distributed with $E(X_i) = \mu$ and finite variance $Var(X_i) = \sigma^2$. Show that*

$$\frac{1}{n\sigma^2} \sum_{i=1}^{n} E((X_i - \mu)^2 \mathcal{I}\{|X_i - \mu| \geq \epsilon\sigma\sqrt{n}\}) \to 0, n \to \infty.$$

for all $\epsilon > 0$.

____

QE ID

Exercise 3 *Let X and Y be random variables on some probability triple $(\Omega, \mathscr{F}, P)$. Suppose $E(X^4) < 1$, and that $P(m \leq X \leq z) = P(m \leq Y \leq z)$ for all integers m and all z in $\mathbb{R}$. Prove or disprove that we necessarily have $E(X^4) = E(Y^4)$.*

____

QE ID (given by Angie)

August 2017 Qualifying Exam in Linear Models

- This is a closed-book test.
- There are 4 questions; some have multiple parts.
- 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 are using 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.
- Write your QE ID number (given to you by Angie) on all answer sheets. DO NOT put your name, UTD ID, or any other identifying information on any of your answer sheets.
- 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 QE ID number (given by Angie) is on each sheet.*
- Total points = 100.

Exercise 1

(15 points) Suppose X_1, X_2 and X_3 are random variables with a common mean μ. Let $\mathbf{X} = (X_1, X_2, X_3)'$. Assume that its covariance matrix is

$$\mathbf{V} = Var(\mathbf{X}) = \sigma^2 \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & \frac{1}{4} \\ 0 & \frac{1}{4} & 1 \end{pmatrix}.$$

Compute the expectation of the quadratic form

$$Q = X_1^2 + 2X_1X_2 - 4X_2X_3 + X_3^2.$$

Exercise 2

Let $X_1, \dots, X_n$ be independent normal random variables with $E(X_i) = \mu$ and $Var(X_i) = \sigma_i^2$.

1. *(15 points) Find the distribution of the quadratic form $Q = \sum_{i=1}^{n}(X_i - \overline{X}.)^2$.*
2. *(5 points) Does the variance of Q in part (1) hold for distributions than the normal? Justify your answer in 1 - 2 sentences.*

Exercise 3

Suppose we want to predict a $q \times 1$ random vector $\mathbf{u} = (u_1, \ldots, u_q)'$ using an $m \times 1$ random vector $\mathbf{Y} = (Y_1, \ldots, Y_m)'$. The two vectors have a joint distribution with finite second-order moments. Let $\mathbf{V} = Var(\mathbf{Y})$, $\mathbf{D} = Var(\mathbf{u})$, and $\mathbf{C} = Cov(\mathbf{u}, \mathbf{Y})$. Assume that $\mathbf{V}$ is non-singular. Suppose the quality of a predictor $\tilde{\mathbf{u}}$, a $q \times 1$ function of $\mathbf{Y}$, is measured using a generalized mean squared prediction error

$$L = E\{(\tilde{\mathbf{u}} - \mathbf{u})'\mathbf{B}(\tilde{\mathbf{u}} - \mathbf{u})\},$$

where $\mathbf{B}$ is a known $q \times q$ positive definite symmetric matrix. We would like to restrict our attention to predictors $\tilde{\mathbf{u}}$ that are linear in $\mathbf{Y}$, that is, are of the form

$$\tilde{\mathbf{u}} = \mathbf{AY} + \mathbf{b},$$

where $\mathbf{A}$ is a $q \times m$ matrix and $\mathbf{b}$ is a $q \times 1$ vector. This goal of this exercise is to find the best linear predictor $\tilde{\mathbf{u}}$ obtained by minimizing the above criterion L with respect to $\mathbf{A}$ and $\mathbf{b}$.

1. *(10 points) Show that the criterion to be minimized can be written as*

$$L = \mathbf{b}'\mathbf{B}\mathbf{b} + E\{(\mathbf{AY} - \mathbf{u})'\mathbf{B}(\mathbf{AY} - \mathbf{u})\} + 2\mathbf{b}'\mathbf{B}E(\mathbf{AY} - \mathbf{u}).$$

2. *(5 points) Solve $\partial L/\partial \mathbf{b} = \mathbf{0}$ for $\mathbf{b}$ and show that the solution is*

$$\mathbf{b} = -E(\mathbf{AY} - \mathbf{u}).$$

3. *(10 points) Show that with $\mathbf{b}$ found in part (2) the criterion L can be expressed as*

$$trace\{\mathbf{B}\, Var(\mathbf{AY} - \mathbf{u})\}.$$

Deduce that minimizing this expression with respect to $\mathbf{A}$ is equivalent to minimizing trace$(\mathbf{BE})$, where $\mathbf{E}$ is a $q \times q$ matrix given by

$$\mathbf{E} = \mathbf{AVA}' - \mathbf{AC}' - \mathbf{CA}'.$$

Denoting the ith row of $\mathbf{A}$ as $\mathbf{a}_i'$ and the jth column of $\mathbf{C}'$ as $\mathbf{c}_j$, the (i,j)th element of $\mathbf{E}$ can be written as

$$\mathbf{a}_i'\mathbf{V}\mathbf{a}_j - \mathbf{a}_i'\mathbf{c}_j - \mathbf{c}_i'\mathbf{a_j}, \quad i, j = 1, \ldots, q.$$

4. *(10 points) Solve $\partial trace(\mathbf{BE})/\partial \mathbf{a}_i = \mathbf{0}$ for $\mathbf{a}_i$ and show that the solution is*

$$\mathbf{a}_i = \mathbf{V}^{-1}\mathbf{c}_i, \quad i = 1, \ldots, q.$$

The solutions can be written in the matrix form as

$$\mathbf{A} = \mathbf{C}\,\mathbf{V}^{-1}.$$

5. *(5 points) Deduce from parts (2) and (4) that the desired best linear predictor is*

$$\tilde{\mathbf{u}} = E(\mathbf{u}) + \mathbf{C}\,\mathbf{V}^{-1}\{\mathbf{Y} - E(\mathbf{Y})\}.$$

6. *(10 points) Show that covariance matrix of the prediction error $\tilde{\mathbf{u}} - \mathbf{u}$ of the best linear predictor is*

$$Var(\tilde{\mathbf{u}} - \mathbf{u}) = \mathbf{D} - \mathbf{C}\,\mathbf{V}^{-1}\mathbf{C}'.$$

Exercise 4

(15 points) Consider the linear model $\mathbf{Y} = \mathbf{X}\mathbf{b} + \boldsymbol{\epsilon}$*, where* $E(\boldsymbol{\epsilon}) = \mathbf{0}$ *and* $Var(\boldsymbol{\epsilon}) = \sigma^2\mathbf{I}$*. Assume that* $|\mathbf{X}'\mathbf{X}| = 0$*, where* $|\mathbf{A}|$ *denotes the determinant of the matrix* $\mathbf{A}$*. Let* $(\mathbf{X}'\mathbf{X})^-$ *denote any generalized inverse (g-inverse) of* $\mathbf{X}'\mathbf{X}$*. Let* $\widehat{\mathbf{b}}_0 = (\mathbf{X}'\mathbf{X})^-\mathbf{X}'\mathbf{Y}$ *denote a solution of the normal equations and assume that the linear combination* $\mathbf{c}'\mathbf{b}_0$ *is estimable. Carefully show that*

$$Var(\mathbf{c}'\widehat{\mathbf{b}}_0) = \sigma^2\mathbf{c}'(\mathbf{X}'\mathbf{X})^-\mathbf{c}.$$