Qualifying Exam January 2018 — 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 four datasets contracts.csv, cputime, infection.dat, potato.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 your codes (with brief comments explaining each step) 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 contract.csv dataset containing data from Florida's Department of Transportation (DOT). These are data on contracts solicitated by DOT for constructing roads and highways. Sealed bids are submitted by the contractors, and the contractor with the lowest bid is awarded the road construction contract. Following variables are recorded:

#	Variable	Description
1	CONTRACT	Contract ID
2	COST	Low-bid contract cost (thousands of dollars)
3	logDOTEST	DOT engineer's cost estimate (log of thousands of dollars)
4	STATUS	Bid status (1 = Fixed, 0 = Competitive)
5	B2B1RAT	Ratio of 2nd-lowest bid to low bid
6	B3B1RAT	Ratio of 3rd-lowest bid to low bid
7	BHB1RAT	Ratio of highest bid to low bid
8	DISTRICT	Location of road (1 = South Florida, 0 = North Florida)
9	BTPRATIO	Ratio of number of bidders to number of plan holders
10	DAYSEST	DOT engineer's estimate of number of workdays required

   (a) Fit a model for predicting cost of a road construction, i.e., COST variable using other relevant variables. Ensure that the model does not contain unimportant variables. For any test(s) conducted in building this model, write down the appropriate hypotheses, test statistic value, p-value, and conclusion. Use 5% level of significance for any necessary tests. [7 points]

(b) For the above model, check all key assumptions. If an assumption is not met, attempt to remedy the situation. Also, conduct diagnostics for checking collinearity and influential data/outliers. Based on these, comment on the appropriateness of the final model. Also, interpret the regression coefficients in the final model. [12 points]

2. Consider the cputime data, which consists of cpu times used to accomplish a certain task by a computer in several runs. We would like to perform inference about the 3rd quartile (75th percentile) of cpu time (θ). An estimator of θ is $\hat{\theta}$ = sample 3rd quartile. Use nonparametric bootstrap with 1,000 resamples and report the following: [20 points]

   (a) histogram and Q-Q plot of the bootstrap distribution of $\hat{\theta}$ with comments about the shape of the distribution.

   (b) bias and standard error of $\hat{\theta}$.

   (c) 2.5th and 97.5th percentiles of the sampling distribution of $\hat{\theta}$.

   (d) 2.5th and 97.5th percentiles of the sampling distribution of $\hat{\theta} - \theta$.

   (e) 95% confidence interval for θ using three bootstrap methods: normal approximation, basic bootstrap, and percentile bootstrap.

3. The data file infection.dat contains data from a study which measured infection severity, treatment outcome (1=positive, 0=negative), and hospital (coded as 1, 2, 3).

   (a) Determine how the treatment outcome depends on infection severity and hospital. In particular, determine whether or not there is a significant interaction between severity and hospital. Interpret the final model. [8 points]

   (b) Compare the hospitals by constructing 95% confidence intervals for the probability of a positive treatment outcome for each of the hospitals when the severity is equal to the overall mean severity. Is one hospital better or worse than the others? [8 points]

   (c) Construct a graphic that plots curves representing the probability of a positive outcome as a function of severity for each hospital, with these curves superimposed on the same graph. Also superimpose curves representing 95% confidence bands (prediction of mean response) for each hospital on this graph. Use different line types or colors for each hospital and include an appropriate legend. [8 points]

4. "Potato flavor data": This data set contains measures of potato flavors of 160 observations. The 6 variables are listed as follows:

#	Variable	Description
1	area	Growing Area: 1, 2
2	temp	Temperature: 1=high, 2=low
3	size	Size: 1=Large, 2=Medium
4	time	Storage Time: 1=0 months, 2=2 mths 3=4mths, 4=6mths
5	cook	Cooking method: 1=Boil, 2=Steam, 3=Mash, 4=Bake Low 5=Bake High
6	flavor	Flavor score: continuous

Researchers are interested in factors associated with the flavor scores (flavor).

   (a) First, we examine the effects of cooking methods (cook) on the flavor. [10 points]

i. Test whether or not the average flavor scores are equal among these five cooking methods. Use $\alpha = 0.05$. Verify the assumptions used with the model.

ii. Obtain confidence intervals for all pairwise comparisons among the five methods at the family-wise confidence level of 95%. Interpret your result and state your conclusions.

(b) Next, we study the effects of growing area (area) and temperature (temp) in addition to the cook effect. [15 points]

i. Conduct an appropriate ANOVA and then check the assumptions.

ii. Conduct tests for the main effects (using $\alpha = .05$).

iii. Are there any significant interaction terms? Conduct tests for all two-way and three-way interaction effects (using $\alpha = .05$).

(c) Last, let us assume that the two areas are a random sample from a large pool of many locations. Assuming everything else is the same as in Question (4b), re-examine the factors "area", "cook" and "temp" in association with "flavor". [5 points]

i. Fit an appropriate model concerning all the main effects. Which variables are significant ($\alpha = 0.05$) for modeling the flavor?

Next we consider all factors: area, temp, size, time, and cook that may affect the flavor.

(d) Develop the best model you can. Test whether any interaction exists. Use variable selection and regression diagnostics methods. [7 points]

STATISTICS Ph.D. QUALIFYING EXAM
STATISTICAL INFERENCE
January 2018

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. Simplify answers as much as possible. Please write neatly so that it is easy to read your solution. Total points = 100.

1. (20 points) Let $X_1, \ldots, X_n$ be a random sample of size $n > 2$ from a uniform distribution over $(\theta - 1/2, \theta + 1/2)$, where $-\infty < \theta < \infty$ is unknown. Consider two estimators of θ, namely, the sample mean $\overline{X}$ and the mid-range $T = (X_{(1)} + X_{(n)})/2$, where $X_{(j)}$ is the jth order statistic.
   (a) (4 points) Find MSE of $\overline{X}$.
   (b) (10 points) Find MSE of T.
   (c) (6 points) Which of the two estimators would you recommend? Justify your answer.

2. (20 points) Let $X_1, \ldots, X_n$ be a random sample of size n from a $N(\mu, \sigma^2)$ distribution, where $-\infty < \mu < \infty$ is unknown and $\sigma > 0$ is known. We are interested in estimation of $g(\mu) = \exp(t\mu)$ for a fixed $t \neq 0$.
   (a) (6 points) Find UMVUE of $g(\mu)$.
   (b) (6 points) Find variance of the estimator in (a).
   (c) (5 points) Find the Cramer-Rao lower bound for the variance of an unbiased estimator of $g(\mu)$.
   (d) (3 points) Show that the variance in (b) is larger than the bound in (c) but their ratio converges to 1 as $n \to \infty$.

3. (15 points) Let $X_1, \ldots, X_n$ be a random sample of size n from a Bernoulli (θ) distribution, where $\theta \in [1/2, 3/4]$ is unknown. Find the maximum likelihood estimator of θ.

4. (20 points) Let F and G be cumulative distribution functions of two known continuous univariate probability distributions. Let X be a single observation from the distribution with cumulative distribution function $\theta F(x) + (1-\theta)G(x)$, where $\theta \in [0,1]$ is unknown. Find a UMP test of size α for testing $H_0 : \theta \leq \theta_0$ versus $H_1 : \theta > \theta_0$, where $\theta_0 \in [0,1]$ is known. Also find the power function of this test.

5. (25 points) Let $X_1, \ldots, X_n$ be a random sample of size n from a $N(\mu, \sigma^2)$ distribution, where both $-\infty < \mu < \infty$ and $\sigma > 0$ are unknown parameters. Consider the prior density

$$\pi(\mu, \sigma^2) = \pi_1(\mu|\sigma^2)\pi_2(\sigma^2),$$

where $\pi_1(\mu|\sigma^2)$ is the density of a $N(\mu_0, \sigma_0^2\sigma^2)$ distribution,

$$\pi_2(\sigma^2) = \frac{1}{\Gamma(a)b^a}\left(\frac{1}{\sigma^2}\right)^{a+1}\exp\{-1/(b\sigma^2)\}I_{(0,\infty)}(\sigma^2),$$

with μ_0, σ_0^2, a, and b as known hyperparameters.

   (a) (20 points) Find the marginal posterior distribution of μ.

   (b) (5 points) Find a $100(1-\alpha)\%$ HPD credible set for μ.

QE ID (given by Angie)

January 2018 Qualifying Exam in Probability Theory

- 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

(20 points) Let $\{X_1, X_2, \ldots\}$ be independent random variables where X_n has a Uniform distribution between 0 and n. Furthermore, let $\{Y_1, Y_2, \ldots\}$ be independent random variables, where Y_n has a Uniform distribution between 0 and n^2.
Which of the following events occur with probability 1? ***Justify your answers****.*

(a) $X_n < 1$ for only a finite number of n.
(b) $X_n < 1$ for infinitely many n
(c) $\liminf X_n < 1$
(d) $\liminf X_n > 1$
(e) $Y_n < 1$ for only a finite number of n.
(f) $Y_n < 1$ for infinitely many n.
(g) $\liminf Y_n < 1$
(h) $\liminf Y_n > 1$

EXERCISE 2 *(20 points)*

1. *Let $A_1, A_2, \ldots$ be any independent sequence of events and let*

$$S_x = \left\{ \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}_{A_i} \leq x \right\}.$$

   Show that for each $x \in \mathbb{R}$ we have $P(S_x) = 0$ or 1.

2. *Let $A_1, A_2, \ldots$ be independent events. Let Y be a random variable which is measurable with respect to $\sigma(A_n, A_{n+1}, \ldots)$ for each $n \in \mathbb{N}$. Prove that there is a real number a such that $P(Y = a) = 1$.*

EXERCISE 3 *(30 points)*

1. *Let $Z, Z_1, Z_2, \ldots$ be random variables. Suppose for each $\epsilon > 0$ we have $P(|Z_n - Z| \geq \epsilon \; i.o.) = 0$. Show that $Z_n \overset{a.s.}{\to} Z$.*

2. *Show the converse of the above, that is, if $Z_n \overset{a.s.}{\to} Z$, then for each $\epsilon > 0$ we have $P(|Z_n - Z| \geq \epsilon \; i.o.) = 0$.*

3. *Suppose $Z_n \overset{a.s.}{\to} Z$. Show that $Z_n \overset{p}{\to} Z$.*

4. *Consider the sample space $\mathbf{\Omega} = [0, 1]$ with probability measure $P([a, b]) = b - a$ for all $0 \leq a < b \leq 1$ and suppose*

$$Z_n(\omega) = \begin{cases} 1 & 0 \leq \omega < \frac{n+1}{2n} \\ 0 & otherwise \end{cases} \quad and \; Z(\omega) = \begin{cases} 1 & 0 < s \leq \frac{1}{2} \\ 0 & otherwise \end{cases}$$

   Does Z_n converge to Z

   (a) in probability?

   (b) almost surely?

   Justify your answers.

EXERCISE 4 *(30 points) Let $\mathbf{\Omega}$ denote a sample space, $\mathcal{A}$ an algebra on $\mathbf{\Omega}$ and a probability measure P on $\mathcal{A}$. Define the set*

$$\mathcal{M} = \left\{ \textit{sets } A \subseteq \mathbf{\Omega} \textit{ such that } P^*(E \cap A) + P^*(E \cap \overline{A}) = P^*(E) \; \forall E \subseteq \mathbf{\Omega} \right\}$$

Recall that

$$\begin{aligned} P^*(E) &= \inf_{\{A_n \in \mathcal{A}, E \subset \bigcup A_n\}} \sum_n P(A_n) \\ P_*(E) &= 1 - P^*(\overline{E}) \end{aligned}$$

1. *Let $A_1, A_2, \ldots \in \mathcal{M}$ be disjoint. Show that*

$$P^*\left(\bigcup_n A_n \right) = \sum_n P^*(A_n).$$

2. *Show that $P^*(\emptyset) = 0$.*

3. *Show that $P^*(\cdot)$ is monotone.*

4. *Show that $\mathcal{M}$ is an algebra.*

5. *Show that $\mathcal{M}$ is a monotone class.*

6. *Let $\mathbf{\Omega}$ be a set being equipped with a σ-algebra $\mathcal{B}$. Suppose μ and ν are two probability measures on $(\mathbf{\Omega}, \mathcal{A})$. Furthermore, let $\mathcal{F}$ be an algebra on $\mathbf{\Omega}$ such that $\sigma(\mathcal{F}) = \mathcal{A}$. Suppose μ and ν agree on $\mathcal{F}$. Show that $\mu = \nu$.*

QE ID (given by Angie)

January 2018 Qualifying Exam in Linear Models

- This is a closed-book test.
- There are 3 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 *(35 points) Suppose the true linear model is* $\mathbf{Y} = \mathbf{X}_1\boldsymbol{\beta}_1 + \boldsymbol{\epsilon}$*, but we fit the linear model* $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} = \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2 + \boldsymbol{\epsilon}$*, where* $\boldsymbol{\epsilon} \sim N_n(\mathbf{0}, \sigma^2\mathbf{I}_n)$ *and* $\mathbf{X}$*,* $\mathbf{X}_1$*, and* $\mathbf{X}_2$ *are full column rank matrices of dimensions* $n \times p$*,* $n \times k$*, and* $n \times (p-k)$*, respectively. Thus, we have* $\boldsymbol{\beta}_2 = \mathbf{0}$ *in the true model, and our least square estimator is* $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$*. Let* $\mathbf{e} = \mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}$ *be the residual,* $S^2 = \mathbf{e}'\mathbf{e}/(n-p)$ *be the unbiased estimator of* σ^2 *based on the fitted model, and* $\hat{\boldsymbol{\beta}}_1$ *consist of the first* k *elements of* $\hat{\boldsymbol{\beta}}$*.*

1. *(4 points) Show that the expectation of* $\hat{\boldsymbol{\beta}}_1$ *under the true model equals* $\boldsymbol{\beta}_1$*.*

2. *(3 points) Deduce that the expectation of* $\hat{\mathbf{Y}}$ *under the true model equals* $\mathbf{X}_1\boldsymbol{\beta}_1$*. What does this result imply?*

3. *(13 points) Show that for* $i = 1, \ldots, k$*, the* i*th diagonal element of* $(\mathbf{X}'\mathbf{X})^{-1}$ *is greater than the* i*th diagonal element of* $(\mathbf{X}_1'\mathbf{X}_1)^{-1}$*. What does this result imply? [Hint: You may use the fact that if all inverses exist,*

$$\begin{pmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{pmatrix}^{-1} = \begin{pmatrix} \mathbf{A}_{11}^{-1} + \mathbf{B}_{12}\mathbf{B}_{22}^{-1}\mathbf{B}_{21} & -\mathbf{B}_{12}\mathbf{B}_{22}^{-1} \\ -\mathbf{B}_{22}^{-1}\mathbf{B}_{21} & \mathbf{B}_{22}^{-1} \end{pmatrix},$$

   where $\mathbf{B}_{22} = \mathbf{A}_{22} - \mathbf{A}_{21}\mathbf{A}_{11}^{-1}\mathbf{A}_{12}$*,* $\mathbf{B}_{12} = \mathbf{A}_{11}^{-1}\mathbf{A}_{12}$*, and* $\mathbf{B}_{21} = \mathbf{A}_{21}\mathbf{A}_{11}^{-1}$*.]*

4. *(5 points) Show that the expectation of* S^2 *under the true model equals* σ^2*. What does this result imply?*

5. *(5 points) Find the covariance matrix of the residual* $\mathbf{e}$ *under the true model. What does this result imply?*

6. *(5 points) Suppose that* $\mathbf{x}_0' = (\mathbf{x}_{10}', \mathbf{x}_{20}')$*, where* $\mathbf{x}_0$*,* $\mathbf{x}_{10}$*, and* $\mathbf{x}_{20}$ *are* $p \times 1$*,* $k \times 1$*, and* $(p-k) \times 1$ *vectors, respectively. Show that* $\mathbf{x}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0 > \mathbf{x}_{10}'(\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{x}_{10}$*. What does this result imply?*

EXERCISE 2 *(25 points) Aerial observations* Y_1, Y_2, Y_3*, and* Y_4 *are made of angles* $\theta_1, \theta_2, \theta_3$*, and* θ_4 *of a quadrilateral on the ground. If the observations are subject to independent* $N(0, \sigma^2)$ *errors, derive an* F *statistic for testing the null hypothesis that the quadrilateral has* $\theta_1 = \theta_4$ *against the alternative that this is not the case. What is the null distribution of the test statistic?*

EXERCISE 3 *(40 points) Consider the Completely Randomized Factorial Design (ab) with random factors,i.e.*

$$Y_{ijk} \;=\; \mu + A_i + B_j + (AB)_{ij} + \epsilon_{ijk}, \quad 1 \le i \le a,\ 1 \le j \le b,\ 1 \le k \le n.$$

Let μ *be some fixed parameter and let the other variables be independent and normally distributed with expectation 0 and variances*

$$Var(A_i) = \sigma_A^2,\ Var(B_j) = \sigma_B^2,\ Var((AB)_{ij}) = \sigma_{AB}^2,\ Var(\epsilon_{ijk}) = \sigma_\epsilon^2.$$

1. *(5 points) Specify the model in matrix notation* $\mathbf{Y} = \mathbf{X}\mathbf{Z} + \boldsymbol{\epsilon}$ *with a random parameter vector* $\mathbf{Z}$ *of length* $a \cdot b$*.*

2. *(13 points) Determine the covariance structure of* $\mathbf{Z}$ *and* $\mathbf{Y}$*. Carry out the spectral decomposition of the covariance matrix of* $\mathbf{Y}$*.*

3. *(12 points) Determine the distributions of the quadratic forms* $\mathbf{Y}'\mathbf{T}_i\mathbf{Y}$ *and derive statistics to test the hypotheses on the variance components.*

4. *(10 points) Derive the ANOVA-table. What are the differences compared to the ANOVA-table of the CRF-ab with fixed factors?*