Qualifying Exam January 2020 — 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/~mchen/QE and download data files "`cd4.csv`", "`divusa.csv`", and "`salary.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 any one other than Angie.**

Problems

1. [30points] The dataset `cd4.csv` contains CD4 counts (in hundreds) for 20 HIV-positive patients at baseline and after one year of treatment with an experimental anti-viral drug. We would like to perform inference on the *largest* eigenvalue θ of the covariance matrix of the two variables in the `cd4` data using nonparametric bootstrap (You may use `eigen` function in `R` or `eigval` function in `SAS` to compute eigenvalues). Set the starting random seed as 1111 and obtain 1000 bootstrap replications. Report the following:
   - a point estimate of θ
   - a histogram of bootstrap distribution of the point estimate
   - bootstrap estimates of bias and standard error of the point estimate
   - 95% bootstrap confidence intervals computed using three methods — (a) normal approximation (b) basic bootstrap, and (c) percentile bootstrap methods.

   Interpret the results and compare the various confidence intervals. **Important Note**: You are required to write your own code for implementing bootstrap and not use any package in R (e.g., boot) or macro in SAS (e.g., %BOOT).

2. [35 points] Work on fitting regression models on a data set in a comma-separated-values file "`divusa.csv`", which reports the divorce rate in the United States between 1920 and 1996. The variables are:

#	Variable name	Description
1	`divorce`	Divorce per 1,000 women aged 15+
2	`year`	The year from 1920 to 1996
3	`unemployed`	Unemployment rate
4	`femlab`	Percent female participation in labor force aged aged 16+
5	`marriage`	Marriages per 1,000 unmarried women aged 16+
6	`birth`	Births per 1,000 women aged 15 to 44
7	`military`	Military personnel per 1,000population

(a) Make a plot to investigate the relationship between `divorce` and `femlab`. Fit a simple linear regression model (Model 1) to predict `divorce` based on `femlab`. Is it a good model? Explain. [3 points]

(b) Continue to work on Model 1. What specific hypotheses are being tested with the t-statistics and their p-values in the software output for Model 1? A newspaper reported that if the percent of female participation in labor force increases 5%, then we would see 2 more divorces per $1,000$ women in a year. Perform a significance test to test the related hypothesis. (Show your work step-by-step and choose a significance level of 0.10). [4 points]

(c) Continue to work on Model 1. Find the confidence intervals for the mean divorce rate when `femlab` = 22 and `femlab` = 60. Create a plot to visualize your result when `femlab` ranges from 22 to 60. [4 points]

(d) Fit a multiple linear regression model (Model 2) to predict `divorce` based on all other variables in the data set. Check for assumptions of the linear models and check for influential points. [6 points]

(e) Investigate the utility of an appropriate transformation on the response variable in Model 2. Consider the new model after transformation (Model 3), can you compute the expected change in the mean divorce rate for one unit increase in one predictor variable holding the other predictors constant? Explain. [6 points]

(f) Find the "best" model by performing best subset selection based on some statistic (choose one from AIC, BIC, adjusted R^2, or Mallow's C_p) to predict `divorce` after the transformation determined in (e). Suppose this is Model 4, would R^2 go up, go down, or stay exactly the same, compared that of Model 3? Explain from a theoretical point of view. [8 points]

(g) Perform a statistical test that compares Model 3 and Model 4. Clearly state the hypotheses associated with this test and interpret the results. Among all the models (Model 1, 2, 3, and 4), which model would you finally suggest to use and why? [4 points]

3. [35 points] A university administrator wants to monitor the equity of faculty salary. She collected salary data, stored in "`salary.csv`", for Assistant Professors, Associate Professors and Professors. We will use it to assess salary differences between male and female professors. It has the following variables:

#	Variable name	Description
1	`salary`	annual salary in dollars
2	`gender`	Female/Male
3	`rank`	Asst. Prof. / Assoc. Prof / Full Prof.
4	`phd.yrs`	years since PhD.
5	`serv.yrs`	years of service.
6	`dept.type`	department type: A("theoretical") or B ("applied").

(a) Use graphical methods to explore the relationship between `salary` and `gender.` Is it plausible to perform a logarithm transformation of `salary`? [5 points]

(b) Build a model that only contains `gender` (Model 1). What conclusions do you draw in terms of gender equity in the faculty pay (conducting statistical tests using $\alpha = 0.05$)? [5 points]

(c) One argues that Model 1 is inadequate at best, and may be misleading at worst, because it fails to consider an important drawback associated with an observational study. What is it? Fit appropriate models to address this criticism. You may ignore interactive effects. [15 points]

(d) The administrator suspects that the effect of `phd.yrs` for males are different than that for females. If true, explain why this will be a form of gender inequity. Write down the formula of an appropriate model. What parameter(s) in your model is(are) important to answer this question? Fit the model and conduct a statistical test using $\alpha = 0.05$. Provide precise interpretations of the estimated effect(s) and the test result so that the administrator, a statistical layman, can easily understand your conclusions. [10 points]

STATISTICS Ph.D. QUALIFYING EXAM
STATISTICAL INFERENCE
January 2020

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

1. (10 points) Let $X_1, \ldots, X_n$ be independent and identically distributed random variables having the probability density

$$f(x|\theta) = \exp\left\{-\left(\frac{x-\mu}{\sigma}\right)^4 - \xi(\theta)\right\},$$

where $-\infty < x < \infty$, $\theta = (\mu, \sigma)$, $-\infty < \mu < \infty$, $\sigma > 0$, and ξ is a known function.

   (a) (3 points) Show that $f(x|\theta)$ is a member of exponential family.
   (b) (7 points) Show that the statistic $T = (\sum_i X_i, \sum_i X_i^2, \sum_i X_i^3, \sum_i X_i^4)$ is minimal sufficient for θ.

2. (25 points) Let $X_1, \ldots, X_n$, $n > 2$, be a random sample from a uniform distribution over the interval $(\theta_1 - \theta_2, \theta_1 + \theta_2)$, where $-\infty < \theta_1 < \infty$ and $\theta_2 > 0$. Let $X_{(1)}$ and $X_{(n)}$ be the smallest and the largest order statistics, respectively.

   (a) (5 points) Show that $(X_{(1)}, X_{(n)})$ is sufficient.
   (b) (5 points) Show that $(X_{(1)}, X_{(n)})$ is complete.
   (c) (5 points) Find $E\{X_{(1)}\}$.
   (d) (5 points) Find $E\{X_{(n)}\}$.
   (e) (5 points) Find the uniformly minimum variance unbiased estimators of θ_1 and θ_2.

3. (15 points) Let X be one observation from a distribution with the probability density

$$f(x|\theta) = \frac{1}{2\theta}\exp\{-|x|/\theta\}, \quad -\infty < x < \infty,$$

where $\theta > 0$ is unknown. Let $\tau(\theta) = 1/(1+\theta)$ be the parameter function of interest.

   (a) (5 points) Find the Cramer-Rao lower bound for an unbiased estimator of $\tau(\theta)$.

   (b) (5 points) Show that $\exp\{-|X|\}$ is the uniformly minimum variance unbiased estimator of $\tau(\theta)$.

   (c) (5 points) Find the variance of the estimator in (b). Does this variance attain the bound in (a)? Explain.

4. (15 points) Let $X_1, \ldots, X_n$ be a random sample from a uniform distribution on the interval $(\theta, \theta + |\theta|)$. Find the maximum likelihood estimator of θ when:

   (a) (4 points) $\theta \in (0, \infty)$.

   (b) (4 points) $\theta \in (-\infty, 0)$.

   (c) (7 points) $\theta \in (-\infty, \infty)$, $\theta \neq 0$.

5. (20 points) Let $X_1, \ldots, X_n$ be a random sample from a distribution with density

$$f(x|\theta) = \theta x^{\theta-1}, \quad x \in (0,1),$$

where $\theta > 0$ is an unknown parameter. Find a uniformly most powerful test of size α for testing $H_0 : \theta \leq \theta_0$ against $H_1 : \theta > \theta_0$, where θ_0 is a specified positive constant. Be sure to explicitly find the distribution of the test statistic.

6. (15 points) Let $X_1, \ldots, X_n$ be a random sample from a distribution with density

$$f(x|\theta) = \frac{a}{\theta}\left(\frac{x}{\theta}\right)^{a-1}, \quad x \in (0,\theta),$$

where $a \geq 1$ is known and $\theta > 0$ is unknown. Find an equal-tailed $100(1-\alpha)\%$ confidence interval for θ based on the largest observation $X_{(n)}$.

PROBABILITY Ph.D. QUALIFYING EXAM

January 2020

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.

Problem 1. Suppose that a function $f(x)$ is Riemann integrable on $[a, b]$. What else do you need to assume (if any) for the function to be integrable with respect to Lebesgue measure on $[a, b]$. Similarly, when are the two integrals equal?

Problem 2. Formulate and prove Chebyshev's inequality.

Problem 3. (a) Present the definition and a criteria (via probability of the union of related events) for convergence of the sequence of random variables X_n to X almost sure (you may also say a.e., with probability 1).
(b) Prove or present a counterexample that if $Z_1, Z_2, \ldots$ satisfy for some $r > 0$ the inequality $\sum_{k=1}^{\infty} E\{|Z_k|^r\} < \infty$, then the sample mean $\bar{Z}$ converges to zero almost sure. (Hint: The Borel-Cantelli Lemma may be useful to prove or disprove the assertion; if you use it - accurately formulate it.)

Problem 4. Consider a probability space $(\Omega, \mathcal{F}, P)$. Let X be a random variable defined on this space and let $B \in \mathcal{F}$ be a fixed event. Present measure-theoretical definitions of X and of the conditional probability $P(B|X = x)$.

Problem 5. Consider a probability space $(\Omega, \mathcal{F}, P)$. Let $\mathcal{G}_1 \subset \mathcal{G}_2$ be two sub sigma-fields of $\mathcal{F}$ (note that one is coarser than another) and Y is a random variable defined on the probability space.
(a) Simplify $E\{E\{Y|\mathcal{F}_2\}|\mathcal{F}_1\}$ and prove your assertion.
(b) Simplify $E\{E\{Y|\mathcal{F}_1\}|\mathcal{F}_2\}$ and prove your assertion.

Problem 6. Definition of a martingale and a submartingale. (Do not forget to begin with a sequence of random variables and a sequence of sub sigma-fields).

Problem 7. Let Y be an integrable random variable on $(\Omega, \mathcal{F}, P)$. Let= $\{\mathcal{F}_n\}$ be an increasing sequence of sub sigma-fields of $\mathcal{F}$. Set $X_n := E\{Y|\mathcal{F}\}$. Is $\{X_n, \mathcal{F}_n\}$ a martingale, submartingale, supermartingale, or none of the above? Prove your assertion.

Problem 8. Formulate (without proof) the Lindeberg's Central Limit Theorem.

Problem 9. Suppose that X_n converges to X in distribution, and Y_n converges in probability to a constant C. What can you say about convergence of X_nY_n? Prove your assertion. (You may notice that this is a famous Slutsky's Theorem).

Problem 10. Prove that if Y is a nonnegative random variable then

$$\sum_{n=1}^{\infty} P(Y \geq n) \leq E\{Y\} \leq 1 + \sum_{n=1}^{\infty} P(Y \geq n).$$