QE ID (given by Angie)

August 2021 Qualifying Exam in Linear Models

August 9, 2021

Instruction:

- This is a closed-book closed-notes test. There are 3 independent standalone questions, Q1, Q2, and Q3, each of which has multiple parts. The total number of points possible is 100.
- 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.
- Write your solutions on the blank sheets of paper that you prepare with. Begin each question on a new sheet and arrange your sheets in order, and number each sheet.
- On all answer sheets, write your QE ID number (given to you by Angie) and identify which question and part is being answered. **DO NOT** put your name, UTD ID, or any other identifying information on any of your answer sheets.

Q1. Denote by $\mathbf{Y}$ the vector of responses and by $\mathbf{X}$ the data matrix, where

$$\mathbf{Y} = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix}, \qquad \mathbf{X} = \begin{pmatrix} 1 & x_{11} & x_{12} \\ 1 & x_{21} & x_{22} \\ \vdots & \vdots & \vdots \\ 1 & x_{n1} & x_{n2} \end{pmatrix}.$$

Consider the following candidate linear regression models:

$$\begin{aligned} \mathcal{M}_0 : &\ \ Y_i = \beta_0 + \varepsilon_i; \\ \mathcal{M}_1 : &\ \ Y_i = \beta_0 + x_{i1}\beta_1 + \varepsilon_i; \\ \mathcal{M}_2 : &\ \ Y_i = \beta_0 + x_{i1}\beta_1 + x_{i2}\beta_2 + \varepsilon_i, \end{aligned}$$

for $i = 1, \ldots, n$, where $\varepsilon_i \overset{iid}{\sim} N(0, \sigma^2)$.

(a) Assume that $\mathcal{M}_1$ is the true model, where Y_i is collected from.

i. (10 pts) Find the least squared estimator for β_0 in $\mathcal{M}_0$ and derive its distribution.

ii. (10 pts) Find the least squared estimator for $(\beta_0, \beta_1, \beta_2)^{\mathrm{T}}$ in $\mathcal{M}_2$ and derive its joint distribution.

iii. (10 pts) Denote by $\hat{\beta}_{0j}$ the least squared estimator of the intercept β_0 in model $\mathcal{M}_j$ for $j = 0, 1, 2$. Note that $\beta_1 \neq 0$, sort $\hat{\beta}_{0j}$ according to their mean squared errors $E[(\hat{\beta}_{0j} - \beta_0)^2]$.

(b) (10 pts) Construct a level α test for the null hypothesis that $\mathcal{M}_0$ is the true model versus the alternative hypothesis that $\mathcal{M}_2$ is the true model.

(c) (10 pts) Use $\mathcal{M}_2$ to construct $(1 - \alpha)100\%$ simultaneous confidence intervals for m future observations

$$Y^{(i)} = \beta_0 + x_{i1}^0\beta_1 + x_{i2}^0\beta_2 + \varepsilon_i^0,$$

where $x_{ij}^0 \in \mathbb{R}$ for $j = 1, 2$, $i = 1, \ldots, m$, and $\varepsilon_i^0 \overset{\text{iid}}{\sim} N(0, \sigma^2)$.

Q2. Consider a linear regression model

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \varepsilon,$$

where $\mathbf{Y}$ is a $n \times 1$ response vector, $\mathbf{X}$ is a $n \times (p+1)$ data matrix, $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)^{\mathrm{T}}$ is a $(p+1) \times 1$ parameter vector with $p > 2$, and the error vector ε follows a multivariate normal distribution with mean zero and a known, positive definite variance-covariance matrix $\mathbf{V}$.

(a) (10 pts) Find the joint distribution of the least squares estimator $\hat{\boldsymbol{\beta}} = (\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p)^{\mathrm{T}}$ of $\boldsymbol{\beta}$.

(b) (10 pts) Construct $(1-\alpha)100\%$ simultaneous confidence intervals (or confidence region) for β_1 and β_2 with coverage probability **exactly** $1 - \alpha$.

(c) (10 pts) Construct a level α test for $\mathbf{A}\boldsymbol{\beta} = \mathbf{c}$ where $\mathbf{A}$ is a $d \times (p + 1)$ matrix with rank d. You may assume the inverse of squared matrices exist if necessary.

Q3. Consider a linear regression model

$$\mathbf{Y}_i = \mathbf{X}_i\boldsymbol{\beta} + \varepsilon_i,$$

where

$$\mathbf{Y}_i = \begin{pmatrix} Y_{i1} \\ Y_{i2} \end{pmatrix}, \qquad \mathbf{X}_i = \begin{pmatrix} X_{i11} & X_{i12} \\ X_{i21} & X_{i22} \end{pmatrix}, \qquad \beta = \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix}, \qquad \varepsilon_i = \begin{pmatrix} \varepsilon_{i1} \\ \varepsilon_{i2} \end{pmatrix},$$

with $\varepsilon_i \overset{\text{iid}}{\sim} N(\mathbf{0}, \sigma^2\mathbf{R}_0)$ for $1 \le i \le n$, $\mathbf{R}_0 = \begin{pmatrix} 1 & \rho_0 \\ \rho_0 & 1 \end{pmatrix}$ for $-1 < \rho_0 < 1$ with an unknown ρ_0. To estimate $\boldsymbol{\beta}$, one can solve the following estimating equation

$$\frac{1}{n}\sum_{i=1}^{n} \mathbf{X}_i^{\mathrm{T}}\mathbf{R}^{-1}(\mathbf{Y}_i - \mathbf{X}_i\boldsymbol{\beta}) = 0, \qquad (1)$$

where $\mathbf{R} = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$ is a working correlation matrix with $-1 < \rho < 1$ for a predetermined ρ. We do not have to impose $\rho = \rho_0$ here. Let

$$\mathbf{M}_n = \sum_{i=1}^{n} \mathbf{X}_i^{\mathrm{T}} \mathbf{R}^{-1} \mathbf{R}_0 \mathbf{R}^{-1} \mathbf{X}_i \qquad \text{and} \qquad \mathbf{H}_n = \sum_{i=1}^{n} \mathbf{X}_i^{\mathrm{T}} \mathbf{R}^{-1} \mathbf{X}_i.$$

(a) (10 pts) Find the estimator $\widehat{\boldsymbol{\beta}}$ by solving the estimating equation (1).

(b) (10 pts) Show that $\boldsymbol{\alpha}^{\mathrm{T}} \mathbf{M}_n^{-1/2} \mathbf{H}_n (\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})/\sigma$ follows the standard normal distribution where $\boldsymbol{\alpha} = (\alpha_1, \alpha_2)^{\mathrm{T}}$ with $\alpha_1^2 + \alpha_2^2 = 1$. You may assume the inverse of squared matrices exist if necessary.

Qualifying Exam August 2021 — Statistical Methods

Instructions

- NOTE: You are not allowed to use internet or email except for downloading the data and uploading the final report. To use these for any other purpose, ask the proctor through Blackboard Collaborate Ultra.
- Go to eLearning and download data files "bridge.txt" and "yield.csv". Let the proctor know if you have any problems with this step.
- You can use any software of your choice.
- Your report should clearly explain the steps, results, conclusions, and justification for the conclusions. Also, include your codes and outputs (ONLY relevant parts; highlighted wherever possible). Note that to to get credit, you have to PROPERLY ANNOTATE YOUR CODES so that it is easy to follow steps. Do NOT attach the parts of the output that were not used in answering questions.
- Write a report and convert it to a single PDF file. Submit it in eLearning by 12:30pm. DO NOT upload 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.**

Problems

1. Consider the one-sample t-test for the population mean. It is believed to be robust to the normality assumption of the parent population. We would like to investigate the extent of its robustness when the population is non-normal. Use Monte Carlo simulation (with 1000 replications for each scenario) to investigate how close is the the empirical type I error rate compared to its nominal level of significance when the parent population is (i) χ^2_1, (ii) Uniform(0, 2), and (iii) t_3. [30 points]

   (a) For each of the three distributions, use a sample size of 25 and conduct a two-sided t-test at 5% significance level to test the hypothesis that the population mean equals its true value. Comment about robustness of the test. [22 points]

   (b) Repeat the whole experiment when the sample size is 100. Are these results similar or different from those obtained with $n = 25$, and what is the reason? [8 points]

2. Before a bridge construction project begins, it will go through the design stage. Bridges are all different, and the time needed for engineers to design a bridge varies. For budgeting and scheduling purposes it is important to predict how long it will take to design a bridge. In the data in "bridge.txt", 45 bridge projects were complied with the following variables. [35 points]

#	Variable name	Description
1	`Time`	Design time in days
2	`DArea`	Deck area of bridge in $1,000$ square feet
3	`CCost`	Construction cost in $1,000$ USD
4	`Dwgs`	Number of structural drawings
5	`Length`	Length of bridge in feet
6	`Spans`	Number of spans

   (a) Make both numerical and graphical summaries of the data. Comment on anything that you find interesting. [2 points]

   (b) Fit the full model with `Time` as the response and `DArea`, `CCost`, `Dwgs`, `Length`, and `Spans` as predictors. Perform regression diagnostics (including multicollinearity) on this model. [7 points]

(c) Use at least two methods to check for influential points and their effects on the model. [4 points].

(d) Describe the procedure of, and then conduct a permutation test to see whether `Length` is statistically significant. Show all the code and report the p-value corresponding to this test. Perform at least $1,000$ iterations to obtain the p-value. [6 points].

(e) Plot 95% point-wise confidence intervals and simultaneous confidence band for the mean response and compare them. [4 points].

(f) Investigate the utility of a transformation on the response variable. Does it improve the model? If yes, work on the new model after transformation to answer question (g); otherwise carry forward the model described in (b) for the next question. [6 points]

(g) Using an exhaustive search, determine the best model using the adjusted R^2, Mallows's C_p, and Bayesian information criterion (BIC), respectively. Among the models determined by the above statistics, which one would you prefer and why? [6 points]

3. The file "yield.csv" contains data from an experiment to study the effects of manure (3 types: manure=1, 2, 3), irrigation (2 levels: 1=high, 2=low), and crop variety (2 kinds: variety=1, 2) on crop yield. Manure, derived from solid animal wastes, is used as organic fertilizer in agriculture. The experiment was conducted in two farm lands. In each farm land, researchers randomly assigned three large fields to apply one of the three types of manure. Then they divided each field into two sub-fields, one of which was chosen at random to be irrigated at high level, and the other at low level. Each sub-field was further split into two small plots, which were randomly assigned to the two varieties. [35 points]

   (a) What is the design? Write down the statistical model and the corresponding assumptions. [5 points]

   (b) Draw plots to explore the relationship between the yield and manure. [5 points]

   (c) Build an appropriate model. Verify the assumptions used with the model. Summarize your findings about the appropriateness of your model [10 points]

   (d) Examine the main effects (and possible interactions if appropriate) using $\alpha = 0.05$. Clearly specify the hypotheses, obtain appropriate test statistics, draw conclusions and carefully interpret your findings. [10 points]

   (e) What is the standard error of the estimated difference in average yield between manure 1 and manure 2? [5 points]

STATISTICS Ph.D. QUALIFYING EXAM
STATISTICAL INFERENCE
August 2021

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. 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. (20 points) Let $X_1, \dots, X_n$ be a random sample from a population with probability density function
$$f(x|\theta) = \frac{\theta}{2\Gamma(1/\theta)} \exp\{-|x|^\theta\}, \quad -\infty < x < \infty,$$
where $\theta \in \Theta = \{1, 2\}$ and Γ is the gamma function. Notice that θ has only two possible values. Also, recall that $\Gamma(1/2) = \sqrt{\pi}$.

    (a) (10 points) Find the maximum likelihood estimator of θ.

    (b) (10 points) Consider testing the hypotheses,
$$H_0 : \theta = 2 \text{ (i.e., the population distribution is } N(0, 1/\sqrt{2})\text{); versus}$$
$$H_1 : \theta = 1 \text{ (i.e., the population distribution is double exponential).}$$
Derive the likelihood ratio test for these hypotheses and show that it can be based on the test statistic $T = \sum_{i=1}^n (X_i^2 - |X_i|)$. Simplify the critical region as much as possible.

2. (20 points) Let $X_1, \dots, X_n$ be a random sample from a population with probability density function
$$f(x|\theta) = \begin{cases} \frac{2x}{\theta^2}, & \text{if } 0 \le x \le \theta, \\ 0, & \text{otherwise,} \end{cases}$$
where $\theta > 0$ is an unknown parameter. For a given $\theta_0 > 0$, suppose we wish to test $H_0 : \theta \le \theta_0$ versus $H_1 : \theta > \theta_0$, with a test of the form: reject H_0 if $X_{(n)} > c$, otherwise accept it, where $X_{(n)} = \max\{X_1, \dots, X_n\}$ and c is the critical point.

    (a) (5 points) Find the power function of this test.

    (b) (5 points) Find c so that this test has prescribed size α.

    (c) (5 points) Show that the test in (b) is a uniformly most powerful level α test.

    (d) (5 points) Invert the test in (b) to obtain an appropriate $1 - \alpha$ confidence set for θ.

3. (20 points) Let $X_1, \dots, X_n$ be a random sample from continuous probability distribution with density function $f(t - \mu)$, where both μ and f are unknown, but $f(t) = f(-t)$ for all t, and f is positive on the real line. Thus, $f(t - \mu)$ is symmetric about μ and is a member of the location family with μ as the location parameter. Let μ_0 be a known real number and $\theta = P(X_1 \ge \mu_0)$.

    (a) (5 points) Prove that testing $H_0 : \mu \le \mu_0$ versus $H_1 : \mu > \mu_0$ is equivalent to testing $H_0 : \theta \le 1/2$ versus $H_1 : \theta > 1/2$.

(b) (8 points) Let Y denote the total number of observations in the sample that are at least μ_0. Consider a test that rejects H_0 when $Y \geq k_\alpha$, otherwise accepts it. Identify k_α such that the test has size α (assuming that the size can be attained).

(c) (7 points) Deduce from (a) and (b) that $X_{(n-k_\alpha+1)}$, the $(n - k_\alpha + 1)$th order statistic of the sample, is a $1 - \alpha$ lower confidence bound for μ.

4. (20 points) Consider a single observation X from a distribution that depends on an unknown parameter θ. Suppose that θ itself is a random variable — it takes two possible values, θ_1 and θ_2, with equal probability. For a given θ, the probability mass function of X, $p(x|\theta)$ is given as:

	x	
θ	1	2
θ_1	0.8	0.2
θ_2	0.4	0.6

(a) (8 points) Derive the posterior distribution of θ.

(b) (5 points) Find the Bayes estimator of θ based on the squared error loss function.

(c) (7 points) Derive a Bayesian test for testing $H_0 : \theta = \theta_1$ versus $H_1 : \theta = \theta_2$. Suppose $X = 1$ is observed. Will you accept H_0 or reject it? Justify.

5. (20 points) Let $X_1, \dots, X_n$ be independent draws from a mixture distribution with two components whose probability density is $pf(x) + (1 - p)g(x)$, where f and g are known densities and the mixing proportion p is an unknown parameter. We would like to estimate p using the expectation-maximization (EM) algorithm. For this, think of $Z_1, \dots, Z_n$ as *missing data*, where for $i = 1, \dots, n$, Z_i is a binary indicator random variable indicating which component of the mixture X_i comes from. In other words,

$$X_i|(Z_i = 1) \sim f(x_i) \quad \text{and} \quad X_i|(Z_i = 0) \sim g(x_i),$$

and $P(Z_i = 1) = p$. Taking $(X_i, Z_i), i = 1, \dots, n$ as *complete data*, derive the E and M steps of the algorithm and show that the EM sequence for estimating p is given by

$$\hat{p}^{(r+1)} = \frac{1}{n} \sum_{i=1}^{n} \frac{\hat{p}^{(r)} f(x_i)}{\hat{p}^{(r)} f(x_i) + (1 - \hat{p}^{(r)}) g(x_i)}.$$

STATISTICS Ph.D. QUALIFYING EXAM

Probability

August 2021

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. Let A_n, $n = 1, 2, \ldots$ be subsets of a set Ω. Prove that $(\limsup_n A_n)^c = \liminf_n A_n^c$, or present an example which disproves this fact. (5 points)

Problem 2. Definition of the probability space $(\Omega, \mathcal{F}, P)$. Then explain why we use $\mathcal{F}$ and not a field (algebra) $\mathcal{A}$. (5 points)
Hint: Give definitions of all three components, recall Kolmogorov...

Problem 3. Definition of Lebesgue integral. (5 points)
Hint: Recall several steps in the definition.

Problem 4. Let $h_1, h_2, \ldots$ form an increasing sequence of nonnegative Borel measurable functions, and let $h(w) = \lim_{n\to\infty} h_n(w)$, $w \in \Omega$. What is a relationship between $\lim_{n\to\infty} \int_\Omega h_n d\mu$ and $\int_\Omega h d\mu$? (5 points)

Problem 5. Formulate and prove dominated convergence theorem. (10 points)

Problem 6. Measure-Theoretical definition of the conditional expectation of Y given X. Also prove uniqueness of the defined conditional expectation.(10 points)
Hint: Explain that Y is an extended random variable on a space... and $X : (\Omega, \mathcal{F}) \to (\Omega', \mathcal{F}')$, a random object. Then recall that you need to assume something about expectation of Y.

Problem 7. Suppose that $Y_n \geq 0$ and $Y_n \uparrow Y$ almost everywhere (a.e.). Prove that $E\{Y_n|X\} \uparrow E\{Y|X\}$ a.e. (5 points)
Hint: This is not a difficult assertion, but to get 5 points you need to present a carefully written proof with all necessary references.

Problem 8. Consider independent and not necessarily identically distributed RVs $X_1, \ldots, X_n$ with finite expectation, and let $S_j := X_1 + \ldots + X_j$, $j = 1, 2, \ldots, n$. Prove that for any $\epsilon > 0$

$$P\Big(\max_{1\leq j\leq n} |S_j - E\{S_j\}| \geq \epsilon\Big) \leq \epsilon^{-2}\mathrm{Var}(S_n).$$

(15 points)
Hint: This is a familiar inequality due to Kolmogorov. Please explain every step in the proof.

Problem 9. Formulate and prove Lindeberg's Central Limit Theorem. (35 points)

Problem 10. What is the difference (if any) between stable and infinitely divisible distributions? Give examples. (5 points)