QE ID (given by Kisa)

August 2023 Qualifying Exam in Linear Models

August 4th, 2023

Instruction:

- This is a closed-book test.
- There are four questions; each has 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.
- Although the notations used in Q1, Q2, Q3, and Q4 are similar, they are independent, standalone problems.
- The total possible points is 100.

Q1 (40pts) Given a linear regression model:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \varepsilon_i,$$

where $\varepsilon_i \overset{iid}{\sim} N(0, \sigma^2)$ for $i = 1, \ldots, n$ and σ^2 is known. Denote the covariate vectors by $x_0 = (1, \ldots, 1)^\top$ and $x_j = (x_{1j}, \ldots, x_{nj})^\top$ for $j = 1, 2$. Assume that the covariate vectors satisfy that $x_0^\top x_1 = x_0^\top x_2 = 0$, $x_1^\top x_1 = x_2^\top x_2 = 1$, and $x_1^\top x_2 = 0$.

(a) (10pts) Under the constraint $\beta_1 = \beta_2 = \gamma$ where γ is unknown, derive the maximum likelihood estimator (MLE) for γ.

(b) (10pts) According to Q1(a), construct the uniformly most powerful level α test for $H_0 : \gamma \leq 0$ versus $H_1 : \gamma > 0$. Hint: show that the test statistic is $T = (\hat{\beta}_1 + \hat{\beta}_2)/2$, where $\hat{\beta}_1$ and $\hat{\beta}_2$ are the MLEs of β_1 and β_2, respectively, without constraints. And then show that the rejection region is $\{T : T > z_\alpha/\sqrt{2}\}$, where z_α is the αth upper quantile of the standard normal distribution.

(c) (10pts) Based on the testing procedure proposed in Q1(b), find the probability of rejecting H_0 when $\beta_1 > 0$ and $\beta_2 > 0$.

(d) (10pts) From Q1(b) and (c), find β_1 and β_2 over the region $B = \{(\beta_1, \beta_2) : \beta_1^2 + \beta_2^2 = 1, \beta_1 > 0, \beta_2 > 0\}$ that maximize the probability of rejecting H_0.

Q2 (20pts) Aerial observations Y_1, Y_2, Y_3, and Y_4 are made of angles $\theta_1, \theta_2, \theta_3, and\ \theta_4$, respectively, of a quadrilateral on the ground. If the observations are subject to independent normal errors with zero means and common variance σ^2.

(a) Derive a test statistic for the hypothesis that the quadrilateral is a parallelogram with $\theta_1 = \theta_3$ and $\theta_2 = \theta_4$.

(b) Construct a level α test for the hypothesis that the quadrilateral is a triangle.

Q3 (20pts) Consider a linear regression model

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

where $\mathbf{Y} = (y_1, \ldots, y_n)^\top$ is a $n \times 1$ response vector, $\mathbf{X} = (\boldsymbol{x}_1^\top, \ldots, \boldsymbol{x}_n^\top)^\top$ is a $n \times (p+1)$ and full-rank matrix with $\boldsymbol{x}_i = (1, x_{i1}, \ldots, x_{ip})$ for $i = 1, \ldots, n$. Parameter vector $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)^\top$ is a $(p+1) \times 1$ parameter vector with $p > 2$. Error vector $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)^\top$ follows a multivariate normal distribution with mean zero, variance-covariance matrix $\mathbf{I}_n\sigma^2$, and unknown $\sigma^2 > 0$.

(a) (10 pts) Construct a level α test for $\beta_1 = \cdots = \beta_p = 0$.

(b) (10 pts) Construct $(1-\alpha)100\%$ simultaneous prediction intervals at $\boldsymbol{x}_{0j} = (1, x_{0j1}, \ldots, x_{0jp})$ for $j = 1, \ldots, k$.

Q4 (20pts) Given a simple linear regression model:

$$Y_i = \beta_1 x_i + \varepsilon_i, \quad i = 1, \ldots, 2n-1, \tag{1}$$

where $x_i = (i/n) - 1$ and $\varepsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2)$ with known $\sigma^2 > 0$. We denote the MLE of β_1 by $\hat{\beta}_1$. However, the simple linear regression model in (1) is misspecified. Data were collected under the two-phase linear regression model:

$$Y_i = \begin{cases} \alpha_1 x_i, & \text{if } 1 \leq i \leq n, \\ -\alpha_1 x_i, & \text{if } n < i \leq 2n-1. \end{cases} \tag{2}$$

(a) When $\alpha_1 \neq 0$ in (2), Show that $E(\hat{\beta}_1) = E(\sum_{i=1}^{2n-1} x_i Y_i) = 0$ but $E[\sum_{i=1}^{2n-1} x_i Y_i \{I(x_i \leq 0) - I(x_i > 0)\}] \neq 0$, where I is the indicator function.

(b) Following Q4(a), find the best linear unbiased estimator (BLUE) for α_1.

You may use the following facts:

- $$\frac{d\mathbf{A}^\top \boldsymbol{\beta}}{d\boldsymbol{\beta}} = \mathbf{A}, \qquad\qquad \frac{d\boldsymbol{\beta}^\top \mathbf{A}\boldsymbol{\beta}}{d\boldsymbol{\beta}} = 2\mathbf{A}\boldsymbol{\beta} \quad (\mathbf{A} \text{ is symmetric.})$$

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

Qualifying Exam August 2023 — Statistical Methods

Instructions

- NOTE: You are not allowed to use the internet except for downloading the data, uploading your report, or using SAS OnDemand for Academics. To use it for any other purpose, ask the proctor.
- Log on to eLearning to download the data. 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 but adequate comments explaining each step) and outputs (ONLY relevant parts, highlighted wherever possible). Do not attach the parts of the output that are not used in answering the questions.
- Submit an electronic report in eLearning. Upload only one single PDF file with the whole report. DO NOT submit separate files for codes or outputs.
- **Write your QE ID number provided by Kisa 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 or share your exam with any one.**

Problems

1. Consider the life expectancy data (LifeExpectancy.csv) consisting of the life expectancy (Life.expectancy) and 14 other variables measured for different countries in a specific year. We would like to understand how life expectancy is related to other variables in this data set. [35 points]

   (a) (3 points) Examine carefully the pairwise relationships using appropriate tools. Based on these, are there any particular variable(s) that should not be considered for inclusion in a multiple linear regression model? If so, justify your choice.

   (b) (5 points) Fit a model to predict the life expectancy based on other variables (depending on your answer to the previous part, you may exclude one or more variables in this model). Test whether the variables whose p-values of t-tests are greater than 0.05 can be dropped all together from this model. Write down appropriate hypotheses, report the extra sum of squares, the F-statistic, and the p-value of this test. State your conclusion.

   (c) (7 points) Check all key assumptions of the model using appropriate tools and comment on each assumption. If an assumption is not met, attempt to remedy the situation. Also, perform collinearity and influential data diagnostics.

   (d) (4 points) Reduce the model you fit in party (c) using the stepwise selection method. Interpret the resulting model including coefficients and r-squared.

   (e) (4 points) What is an alternative interpretation of the coefficient of determination involving fitted values? Compute the coefficient of determination using that method and comment.

   (f) (6 points) Suppose a few observations are missing in this dataset. Two regression models of the life expectancy on the following two sets of predictors are fitted — (1) Status, HIV.AIDS, and Schooling; (2) Status and HIV.AIDS. The following three outputs are from the two models and another different model:

```
Output 1
Residual standard error: 4.219 on 116 degrees of freedom
Multiple R-squared:  0.769,Adjusted R-squared:  0.763
F-statistic: 128.7 on 3 and 116 DF,  p-value: < 2.2e-16
```

```
Output 2
Residual standard error: 5.818 on 117 degrees of freedom
Multiple R-squared:  0.5569,Adjusted R-squared:  0.5493
F-statistic: 73.53 on 2 and 117 DF,  p-value: < 2.2e-16

Output 3
Residual standard error: 4.195 on 115 degrees of freedom
Multiple R-squared:  0.7736,Adjusted R-squared:  0.7657
F-statistic: 98.24 on 4 and 115 DF,  p-value: < 2.2e-16
```

Which output(s) belongs to the two models and why? Compute the coefficient of partial determination of life expectancy with schooling when status and HIV.AIDS are already in the model. Interpret the coefficient.

(g) (6 points) As there are several outliers in the data, what method can be used to reduce their influence in the regression model? Write down the steps involved in fitting such a model (do not write code or fit the model).

2. Every year, *Fortune*, an American Business Magazine, publishes the *Fortune 500*, which ranks the top 500 corporations by revenue. This data set in "fortune_500.csv" includes the entire list of those corporations for year 2022, along with their `numbers of employees`, `revenue` (in millions), `profit` (in millions), and `assets` (in millions). In this study, our goal is to perform inference about the trimmed mean of `assets`. A trimmed mean is a statistical measure of the central location of a distribution, much like the mean and median. It involves the calculation of the mean after discarding observations at the high and low end in the sorted list. A typical type of trimmed means is α-trimmed mean ($\alpha \in (0, 0.5)$), which is defined as the average of the middle $(1-2\alpha)\times 100\%$ of the ordered data values, or the average of the remaining values after discarding the smallest and largest $\alpha \times 100\%$ values in the sample.

Important Note: For implementing bootstrap, you must write your own code and must NOT use any package in R (e.g., boot) or macro in SAS (e.g., %BOOT). [30 points]

(a) Make an appropriate histogram of `assets`, carefully choosing the bin width or number of bins. [3 points]

(b) Make a plot of α-trimmed means of `assets` against the values of $\alpha \in \{0.00, 0.05, 0.10, 0.15, \cdots, 0, 45\}$. Describe the relationship among the sample median, sample mean, and sample α-trimmed mean, and discuss why we need to use the α-trimmed mean, rather than the sample mean, to represent the center of the distribution of `assets` in this study. [4 points]

(c) Suppose we have absolutely no idea what is the underlying distribution of `assets`. Generate 1000 non-parametric bootstrap samples to estimate the standard error of its α-trimmed mean, where $\alpha = 0.20$. [9 points]

(d) Following (c), make a Q-Q plot of the bootstrap distribution of the 0.2-trimmed mean of `assets`. Comment the shape of the distribution. [5 points]

(e) Following (c), use two bootstrap methods for constructing confidence intervals: basic bootstrap and percentile bootstrap, to obtain 95% confidence intervals for the 0.2-trimmed mean of `assets`. Briefly discuss which one is the most appropriate in this case. [9 points]

3. In a study for Alzheimer's disease (AD), researchers recruited AD patients and randomly assigned them to the active treatment and placebo groups with an approximate 2:1 ratio, and treated them for 8 weeks. To assess whether the active treatment has effect in improving the cognitive abilities associated with learning and memory, they administered two cognitive tests, Test 1 and Test 2, before and after the 8-week period for each participant. The scores are integers between 0 and 15. The data are stored in "cognitive.csv". Use $\alpha = 0.05$ for all hypothesis tests and confidence intervals below.

(a) Comment on the pros and cons of the before-and-after design by comparing it to an alternative design that only conduct the test after the 8-week period. [5 points]

(b) To investigate whether there is an improvement in Test 1 scores, a naive strategy is to conduct a two sample paired t-test of Test 1 scores for the active treatment group, and another two sample paired t-test for the placebo group, followed by a comparison of the two confidence intervals. Conduct the naive analysis, plot the two confidence intervals on the same graph, and draw a conclusion. [5 points]

(c) Why is the analysis in part (b) a poor one? How do you improve it? Conduct your analysis and report your result. [10 points]

(d) Now we would like to explore the Treatment, Test and Time effects. Conduct an appropriate ANOVA analysis including the three factors. Do they interact with each other? Report your final fitted model. [10 points]

(e) Based on your model in (d), is there an improvement in the difference of Test 1 scores, defined as "After score" - "Before score", between the active treatment and placebo groups? Conduct a test to answer this question. [5 points]

STATISTICS Ph.D. QUALIFYING EXAM
STATISTICAL INFERENCE
August 2023

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

Problem 1. (10 points) Consider a sample of size n from Bernoulli(p) random variable X, and set $\bar{X} = n^{-1}\sum_{l=1}^{n} X_l$. Consider $p = 1/2$ and find the distribution that $Y_n := n[\bar{X}(1-\bar{X}) - 1/4]$ converges to in distribution. Hint: Simplify the random variable Y_n to a functional of $\bar{X} - 1/2$ and then use the sampling theory. Write down all theoretical results (laws) used.

Problem 2. (10 points) Consider a sample of size n from Uniform$(0, \theta)$. Find a complete sufficient statistic and prove your assertion. Hint: Write down all definitions and theoretical results used.

Problem 3. (20 points) Formulate and prove Basu's Theorem. Hint: Pay attention to subscripts, write down all definitions used.

Problem 4. (5 points) Consider a sample of size n from Normal$(\theta, 1)$, $\theta \in [0, \infty)$. Find the maximum likelihood estimate.

Problem 5. (5 point) Consider a sample of size n from X with the density $f^X(x|\theta) = \theta x^{-2} I(0 < \theta \leq x < \infty)$. Find the maximum likelihood estimate of θ.

Problem 6. (5 points) Let $f(x|\theta)$ be the logistic pdf,

$$f(x|\theta = \frac{e^{x-\theta}}{(1+e^{x-\theta})^2}, \ x \in (-\infty, \infty), \ \theta \in (-\infty, \infty).$$

Note that the corresponding logistic cdf is $F_\theta(x) = e^{x-\theta}/[1+e^{x-\theta}]$. Based on one observation, find the UMP size α test for $H_0: \ \theta \leq 0$ versus $H_1: \theta > 0$. Hint: Recall the notion of MLR (monotone likelihood ratio), the corresponding theory, and then apply it. As the bonus 3 points, what is the Type II error for $\theta = 1$?

Problem 7. (5 points) Pivoting an exponential density $f(x|\theta) = e^{-(x-\theta)}I(x > \theta)$ find a $1-\alpha$ confidence interval for θ.

Problem 8. (5 points) Consider a density $f_\theta(x)$ from an exponential family of distributions with θ being the natural parameter. Show that the likelihood equation has a unique root.

Problem 9. (20 points) Formulate and prove Cramer-Rao inequality for iid sample of size n.

Problem 10. (10 points) Formulate and prove Rao-Blackwell Theorem.

Problem 11. (5 points) Consider a sample of size n from X with Normal(θ, σ^2) distribution. Consider a Bayes approach with prior for θ being Normal(μ, τ^2). Here σ^2, μ, τ^2 are known. Consider testing $H_0 : \theta \leq \theta_0$ versus $\theta > \theta_0$, and find the Bayes test. For 2 extra points, comment on the case $\mu = \theta_0$.

STATISTICS Ph.D. QUALIFYING EXAM
Probability
August 2023

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. (10 points) Consider independent random variables $X_1, \ldots, X_n$ with finite expectation, and let $S_j = X_1 + \ldots + X_j$, $j = 1, \ldots, n$. Then for any $\epsilon > 0$ the following Kolmogorov's inequality holds,

$$\mathrm{P}(\max_{1 \le j \le n} |S_j - E\{S_j\}| \ge \epsilon) \le \mathrm{Var}(S_n)/\epsilon^2.$$

Prove this inequality.

Problem 2. (5 points) Definitions of a measurable function and Lebesgue integral.

Problem 3. (5 points) Formulate the monotone convergence theorem.

Problem 3. (5 points) Formulate Fatou's Lemma.

Problem 4. (5 points) Consider two sigma-fields satisfying $\mathcal{F}_1 \subset \mathcal{F}_2$. Simplify $E\{\{E\{Y|\mathcal{F}_\in\}|\mathcal{F}_1\}$ and prove your assertion. Hint: write down all used definitions.

Problem 5. (10 points) Prove that if for a sequence of measurable functions we have $f_n \to f$ almost uniformly, then $f_n \to f$ in measure and almost everywhere. Remark: write down all definitions of the considered converegences.

Problem 6. (5 points) Measure-Theoretical definition of the conditional probability. Hint: Begin with "Let $X : (\Omega, \mathcal{F}) \to (\Omega', \mathcal{F}')$ be a random object on ..." Prove uniqueness of the defined conditional probability. (10 points)

Problem 7. (10 points) Formulate and prove the second Borel-Cantelli Lemma. (10 points)

Problem 8. (10 points) Let $\{X_n, \mathcal{F}_n\}$ be a submartingale, g is a convex and increasing function from R to R. Suppose that $g(X_n)$ is integrable for all n. What can you say about $\{g(X_n), \mathcal{F}_n\}$? Note: Please give definitions of all notions mentioned in the problem, and present a proof.

Problem 9. (30 points) Formulate and prove Lindeberg's Central Limit Theorem.

Problem 10. (5 points) Let B_t be a Brownian motion. Give its definition and then find $E\{B_t B_{t+s}\}$. Prove your assertion.