QE ID (given by Kisa)

January 2023 Qualifying Exam in Linear Models

January 9, 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 Kisa) 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 Kisa) is on each sheet.
- Although the notations used in all questions are similar, they are independent, standalone problems.
- The total possible points is 100.

Q1 Consider the model

$$y_i = \beta_0 \prod_{j=1}^{p-1} x_{ij}^{\beta_j} \varepsilon_i,$$

for $i = 1, \ldots, n$, where ε_i follows a log-normal distribution such that $\log \varepsilon_i \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)$ where $\sigma > 0$ is unknown.

(a) (10 points) Find the maximum likelihood estimator (MLE) for $(\beta_0, \beta_1, \ldots, \beta_{p-1})^\top$.

(b) (10 points) Construct a 95% confidence interval for β_0.

Q2 Consider the linear regression model

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

where $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^\top$, $\boldsymbol{X} = (\boldsymbol{1}_n, \boldsymbol{X}_1, \ldots, \boldsymbol{X}_{p-1})$ is a $n \times p$ and full rank covariate matrix, $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_{p-1})^\top$, and $\boldsymbol{\varepsilon} \sim \mathcal{N}_n(\boldsymbol{0}, \sigma^2 \boldsymbol{I}_n)$ with an unknown $\sigma > 0$. Denote $\widehat{\boldsymbol{\beta}}$ as the MLE of $\boldsymbol{\beta}$ and $\widehat{\boldsymbol{Y}} = (\hat{Y}_1, \ldots, \hat{Y}_n)$ as the vector of fitted values.

(a) (10 points) Find the joint distribution of $\widehat{\boldsymbol{Y}}$.

(b) (10 points) Show that $\widehat{\boldsymbol{\beta}}$ is independent of the residual sum of squares, i.e., $\sum_{i=1}^n (Y_i - \hat{Y}_i)^2$.

(c) (10 points) Construct **exact** 95% simultaneous **prediction** intervals for new responses when the covariate values are $\boldsymbol{X} = \boldsymbol{x}_{01}$ and $\boldsymbol{X} = \boldsymbol{x}_{02}$.

Q3 (20 points) Suppose $\beta_1 = \cdots = \beta_{p-1} = 0$ in Q2 and denote $\overline{Y} = n^{-1} \sum_{i=1}^n Y_i$. Find the distribution of the coefficient of determination $R^2 = \sum_{i=1}^n (\hat{Y}_i - \overline{Y})^2 / \sum_{i=1}^n (Y_i - \overline{Y})^2$ and prove that $E[R^2] = (p-1)/(n-1)$.

Q4 Consider the additive model

$$\boldsymbol{Y} = \boldsymbol{\mu} + \boldsymbol{\varepsilon},$$

where $\boldsymbol{\mu} \in \mathbb{R}^n$, $\boldsymbol{\varepsilon} \sim \mathcal{N}_n(\boldsymbol{0}, \sigma^2 \boldsymbol{I}_n)$ with a known $\sigma > 0$. While the true functional form for $\boldsymbol{\mu}$ is generally not available in practice, the following are two possible cases.

$$\mathcal{M}_1 : \boldsymbol{Y} = \boldsymbol{X}_1 \boldsymbol{\beta}_1 + \boldsymbol{\varepsilon},$$
$$\mathcal{M}_2 : \boldsymbol{Y} = \boldsymbol{X}_1 \boldsymbol{\beta}_1 + \boldsymbol{X}_2 \boldsymbol{\beta}_2 + \boldsymbol{\varepsilon},$$

where $\boldsymbol{X}_1$ is $n \times p_1$, $\boldsymbol{X}_2$ is $n \times p_2$, $\boldsymbol{X} = (\boldsymbol{X}_1, \boldsymbol{X}_2)$ is $n \times (p_1 + p_2)$ and has full rank, $\boldsymbol{\beta}_1$ is $p_1 \times 1$, and $\boldsymbol{\beta}_2$ is $p_2 \times 1$. Since $\boldsymbol{\mu}$ does not have to be $\boldsymbol{X}_1\boldsymbol{\beta}_1$ or $\boldsymbol{X}_1\boldsymbol{\beta}_1 + \boldsymbol{X}_2\boldsymbol{\beta}_2$, we consider the best approximations

$$\boldsymbol{\beta}_{11} = \arg\min_{\boldsymbol{\beta}_1} \|\boldsymbol{X}_1\boldsymbol{\beta}_1 - \boldsymbol{\mu}\|_2^2 \qquad \text{and} \qquad (\boldsymbol{\beta}_{21}^\top, \boldsymbol{\beta}_{22}^\top)^\top = \arg\min_{(\boldsymbol{\beta}_1^\top, \boldsymbol{\beta}_2^\top)^\top} \|\boldsymbol{X}_1\boldsymbol{\beta}_1 + \boldsymbol{X}_2\boldsymbol{\beta}_2 - \boldsymbol{\mu}\|_2^2.$$

Denote the MLE of $\boldsymbol{\beta}_1$ and $(\boldsymbol{\beta}_1^\top, \boldsymbol{\beta}_2^\top)^\top$ as $\widehat{\boldsymbol{\beta}}_{11}$ and $(\widehat{\boldsymbol{\beta}}_{21}^\top, \widehat{\boldsymbol{\beta}}_{21}^\top)^\top$ as in $\mathcal{M}_2$, respectively.

(a) (10 points) Show that $E[\widehat{\boldsymbol{\beta}}_{11}] = \boldsymbol{\beta}_{11}$ and $E[\widehat{\boldsymbol{\beta}}_{21}] = \boldsymbol{\beta}_{21}$.

(b) (10 points) If $\boldsymbol{X}_1^\top \boldsymbol{X}_1 = \boldsymbol{I}_{p_1}$, construct **exact** simultaneous 95% confidence intervals for $\boldsymbol{\beta}_{11}$.

(c) (10 points) If $\boldsymbol{X}_1^\top \boldsymbol{X}_2 = \boldsymbol{0}_{p_1 \times p_2}$, show that $\boldsymbol{\beta}_{11} = \boldsymbol{\beta}_{21}$ and $Cov(\widehat{\boldsymbol{\beta}}_{11}, \widehat{\boldsymbol{\beta}}_{22}) = \boldsymbol{0}_{p_1 \times p_2}$.

You may use the following facts:

- If $Z_1 \sim \chi^2_{\alpha_1}$, $Z_2 \sim \chi^2_{\alpha_2}$ are independent, then $Z_1/(Z_1 + Z_2) \sim Beta(\alpha_1, \alpha_2)$ with mean $\alpha_1/(\alpha_1 + \alpha_2)$.
- $$\frac{d\mathbf{A}^\top \boldsymbol{\beta}}{d\boldsymbol{\beta}} = \mathbf{A}, \qquad \frac{d\boldsymbol{\beta}^\top \mathbf{A} \boldsymbol{\beta}}{d\boldsymbol{\beta}} = 2\mathbf{A}\boldsymbol{\beta} \quad (\mathbf{A} \text{ is symmetric.})$$

Qualifying Exam January 2023 — Statistical Methods

Instructions

- NOTE: You are not allowed to use the internet except for downloading the data and uploading your report. 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. Criminologists are interested in the effect of punishment regimes on crime rates. The attached file `crime.csv` contains crime rates and data on 15 explanatory variables for 47 U.S. states, in which both the crime rates and the explanatory variables have been centered to mean zero and scaled to have variance one. The response variable Y denotes the rate of crimes in a particular category. See the description of the explanatory variables below.

	Variable name	Description
1	`M`	percentage of males aged 14–24
2	`So`	Indicator variable for a Southern state
3	`Ed`	Mean years of schooling
4	`Po1`	Police expenditure in the current year
5	`Po2`	Police expenditure in the previous year
6	`LF`	Labour force participation rate
7	`M.F`	Number of males per 1,000 females
8	`Pop`	State population
9	`NW`	Number of non-whites per 1,000 people
10	`U1`	Unemployment rate of urban males 14–24
11	`U2`	Unemployment rate of urban males 35–39
12	`GDP`	Gross domestic product per head
13	`Ineq`	Income inequality
14	`Prob`	Probability of imprisonment
15	`Time`	Average time served in state prisons

   (a) Fit a regression model $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, which includes all 15 explanatory variables (i.e., the full model). Describe the relationships between crime and the explanatory variables. Which variables are strongly predictive of crime rates? [3 points]

   (b) Check for assumptions of the full model. [6 points]

   (c) Based on the full model, check for influential points based on Cook's distance and exclude those points, if any, for all subsequent analysis. [3 points]

(d) Can you improve the full model by applying a transformation on crimes? If yes, directly investigate the utility of the transformation and show the transformed model has a better fit; if no, explain why and briefly describe a solution (You do not need to implement your strategy). [5 points]

(e) Explain what the R^2 is. Based on the full model, first use the generic function `fitted()` to extract the fitted values $\hat{\boldsymbol{y}}$ from the model, and then calculate the R^2 combining the given data $\boldsymbol{y}$. [6 points, a numerical answer without any explanation or derivation results in 0 points.]

(f) Using an exhaustive search, first determine the best models in terms of the R^2, and then determine the best models in terms of the adjusted R^2. Detail your strategy, write the code with sufficient comments, and use appropriate graphs to summarize your result. Between the models determined by the above two statistics, which model would you use and why? [6 points]

(g) Compare the full model and the best model in the predictive power as follows. Randomly split the data equally into a training set $\{\boldsymbol{y}_{\text{tr}}, \boldsymbol{X}_{\text{tr}}\}$ and a test set $\{\boldsymbol{y}_{\text{te}}, \boldsymbol{X}_{\text{te}}\}$. Using only the training set, obtain least squares regression coefficients $\hat{\boldsymbol{\beta}}$. Obtain predicted values for the test data by computing $\hat{\boldsymbol{y}}_{\text{te}} = \boldsymbol{X}_{\text{te}}\hat{\boldsymbol{\beta}}$. Plot $\hat{\boldsymbol{y}}_{\text{te}}$ and $\boldsymbol{y}_{\text{te}}$ and compute the prediction error $\frac{1}{n_{\text{te}}}\sum(y_{i,\text{te}} - \hat{y}_{i,\text{te}})^2$. Between the full model and the best model, show which model is better based on the prediction performance. [6 points]

2. Consider the full crime dataset including any observations you may have deleted in part (c) of the previous question. Fit a simple linear regresssion model of Y on Po2 using the ordinary least squares (OLS) method.

(a) Use the iteratively reweighted least squares (IWLS) approach to robust regression for dampening the influence of outlying cases. Use the weight function and Median Absolute Deviation (MAD for calculation of scaled residuals) as defined below:

$$w = \begin{cases} 1, & |u| \le 1.345, \\ \frac{1.345}{|u|}, & |u| > 1.345, \end{cases} \quad \text{where } u_i = \frac{e_i}{MAD}, \quad MAD = \frac{1}{0.6745}\,\text{median}\,\{|e_i - \text{median}\{e_i\}|\}.$$

The initial residuals can be obtained from the fitted OLS model. Carry out at least three iterations (excluding iteration 0). Compare the regression coefficient estimates obtained using the robust regression method with the ones obtained using the OLS method. Why MAD is used for scaling the residuals instead of MSE? [10 points]

(b) Use bootstrap method for estimating SE of the robust regression estimate b_1 and 95% confidence interval for β_1. Generate 1000 bootstrap samples using random X sampling, i.e., by sampling (X, Y) pairs. For each bootstrap sample, only one iteration of the IWLS procedure is to be used. Construct a histogram of bootstrap distribution of b_1. Does it appear to approximate a normal distribution? Report bootstrap estimate of SE of b_1 and CI for β_1 using the percentile method. [20 points]

Important Note: In both parts, you are required to write your own code for implementing robust regression and bootstrap and not use any package in R (e.g., boot) or macro in SAS (e.g., %BOOT).

3. A researcher wanted to assess the effects of Diet and Exercise on lowering the blood presure. She recruited 12 patients with high blood presure (BP). There were four treatments: (1) no diet and no exercises; (2) diet only; (3) exercises only; and (4) diet and exercises combined. Each participant was enrolled in all four trials. The order of the trials was randomized and time was allowed between trials so that effects of previous trials may have little impact on the subsequent one. Each trial lasted 9 weeks and a score of reduction in BP was measured at week 3, 6 and 9. You can find the data in "bp.csv". Use $\alpha = 0.05$ for all hypothesis tests below.

(a) Draw plots to explore the effects of Diet and Exercises on BP scores at all three times. Provide a summary of your observations and findings. [5 points]

(b) Conduct an appropriate ANOVA analysis. Carefully examine interaction effects. Verify the assumptions used with the model. Summarize your findings about the appropriateness of your model [10 points]

(c) Do Diet and Exercise affect the BP score? Do they interact with each other? Based on your model in part (b), conduct an appropriate tests and/or confidence intervals to answer this question. [10 points]

(d) Is there a time effect? Carefully interpret your findings. [10 points]

STATISTICS Ph.D. QUALIFYING EXAM
STATISTICAL INFERENCE
January 2023

General Instructions: Write your QE ID number (given by the department) on all answer sheets. Do NOT put your name, UTD ID, or any other identifying information on any of your answer sheets. You may use well-known results by stating them without proof. Show all work and justify all steps to get full credit. Simplify your answer 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 from a distribution with probability density

$$f(x) = \begin{cases} \frac{e^{-|x|}}{2(1-e^{-\theta})}, & -\theta \le x \le \theta; \\ 0, & \text{else;} \end{cases}$$

where $\theta > 0$ is an unknown parameter. Let $X_{(1)}$ and $X_{(n)}$ be the smallest and the largest order statistics, respectively.

   (a) (5 points) Find the sufficient statistic for θ.
   (b) (5 points) Find MLE of θ, $\widehat{\theta}_{MLE}$.
   (c) (3 points) Find $P(-a \le X_{(1)} \le X_{(n)} \le a)$ for any $0 < a \le \theta$.
   (d) (3 points) Is $\widehat{\theta}_{MLE}$ an unbiased estimator? Justify it.
   (e) (4 points) Show that $\widehat{\theta}_{MLE}$ is a consistent estimator for θ.

2. (20 points) Let $X_1, \ldots, X_n$ denote a random sample of size n from a Normal(μ, σ^2) distribution with probability density

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right].$$

   (a) (10 points) If μ is unknown, to test $H_0 : \sigma = 1$ against $H_1 : \sigma \neq 1$, construct the Likelihood ratio test of size α.
   (b) (10 points) To test $H_0 : (\mu, \sigma) = (0, 1)$ against $H_1 : (\mu, \sigma) = (1, 2)$, construct the uniformly most powerful test of size α.

3. (20 points) Suppose the random vector $Y = (Y_1, Y_2, Y_3, Y_4)$ follows a multinomial distribution with cell probabilities $(1/2 - \theta/2, \theta/4, \theta/4, 1/2)$. Here $\theta \in [0, 1]$ is an unknown parameter.

   (a) (6 points) If Y is observed, show that the MLE of θ is

$$\hat{\theta} = \frac{Y_2 + Y_3}{Y_1 + Y_2 + Y_3}.$$

   (b) (14 points) Suppose instead of Y, the observed data is $y_1 = 38$, $y_2 = 34$, and $y_3 + y_4 = 125$. Thus, we get to observe Y_1, Y_2, and the sum $Y_3 + Y_4$, but not Y_3 and Y_4 separately. Set up an EM algorithm which upon convergence gives the MLE of θ. Specifically, derive the E- and M-steps of the algorithm and show that at the end of an iteration of EM, θ_{old} — the current value of θ — will be updated as

$$\theta_{\text{new}} = \frac{34 + y_{3,\text{old}}}{72 + y_{3,\text{old}}}, \text{ where } y_{3,\text{old}} = \frac{125(\theta_{\text{old}}/4)}{1/2 + \theta_{\text{old}}/4}.$$

4. (20 points) Suppose Y_1 and Y_2 are two random variables, where

$$Y_i|U \sim \text{independent Poisson}\big(\exp(\beta_i + U)\big), i = 1, 2;\ U \sim N(0, \sigma^2).$$

Thus, given U, the random variables Y_1 and Y_2 follow conditionally independent Poisson distributions, and U itself follows a normal distribution.

(a) (6 points) Find $E(Y_i)$ and $var(Y_i)$, the marginal mean and variance of Y_i.

(b) (3 points) Use (a) to argue that the marginal distribution of Y_i is *not* Poisson.

(c) (5 points) Find $\text{cov}(Y_1, Y_2)$.

(d) (6 points) Find the joint probability mass function of Y_1 and Y_2. (Note: You may not have a closed-form expression for this function.)

5. (20 points) Let $X_1, \ldots, X_n$ denote a random sample of size $n\,(\geq 1)$ from a distribution with probability density

$$f(x) = \begin{cases} \frac{1}{\theta}\exp(-(x-\theta)/\theta), & x > \theta \\ 0, & \text{otherwise,} \end{cases}$$

where $\theta > 0$ is an unknown parameter.

(a) (5 points) Show that both $\overline{X}/\theta$ and $X_{(1)}/\theta$ are *pivotal quantities*, where $\overline{X}$ is the sample mean and $X_{(1)}$ is the smallest order statistic.

(b) (10 points) Obtain a $100(1-\alpha)\%$ confidence interval for θ based on each pivotal quantity in (a).

(c) (5 points) Explain which of the two confidence intervals in (b) you would recommend. Justify your answer.

STATISTICS Ph.D. QUALIFYING EXAM
Probability
January 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 a nonnegative random variable Y. Prove that

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

Problem 2. (5 points) Formulate the dominated convergence theorem.

Problem 3. (5 points) Let random variable $X = 0$ with probability 1. Using definition of Lebesgue integral, prove that the expectation of X is zero. (Hint: This is not a trivial problem. Recall several steps in definition of Lebesgue integral).

Problem 4. (10 points) Consider two sigma–fields $\mathcal{F}_1 \subset \mathcal{F}_2$. Simplify $E\{E\{Y|\mathcal{F}_1\}|\mathcal{F}_2\}$. Then prove your assertion and do not forget to give all definitions.

Problem 5. (10 points) Let $f_1, f_2, \ldots$ be real-valued Borel measurable and uniformly integrable functions on the probability space $(\Omega, \mathcal{F}, P)$. Show that

$$\int_{\Omega} (\liminf_n f_n) dP \leq \liminf_n \int_{\Omega} f_n dP.$$

(Hint: Give definitions of Borel measurable and uniformly integrable functions.)

Problem 6. (10 points) Write down measure-theoretical definition of the conditional probability. Then consider a discrete random variable X, suggest a formula for a conditional probability $P(B|X = x)$, and prove that this formula is correct according to the definition.

Problem 7. (5 points) Formulate and prove Borel-Cantelli Lemma.

Problem 8. (5 points) Let $\{X_n, \mathcal{F}_n\}$ be a martingale, g is a convex and increasing function from R to R. Suppose that $g(X_n)$ is integrable for all n. Prove that $\{g(X_n), \mathcal{F}_n\}$ is a submartingale. Note: Please give definitions of all notions mentioned in the problem.

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

Problem 10. (10 points) Formulate the law of iterated logarithm for Brownian motion. Hint: Do not forget to begin with the classical probability definition of the considered process.