QE ID (given by Kisa)

August 2022 Qualifying Exam in Linear Models

August 8, 2022

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 Q1, Q2, Q3, and Q4 are similar, they are independent, standalone problems.
- The total possible points is 100.

Q1 (20pts) Suppose that $(Y_1, Y_2, Y_3)^\top \sim \mathcal{N}_3(\mathbf{0}, \mathbf{I}_3)$.

(a) (10pts) Find the moment generating function of $W = 2(Y_1Y_2 - Y_2Y_3 - Y_3Y_1)$.

(b) (10pts) Show that W has the same distribution as that of $2U_1 - U_2 - U_3$, where U_i's are independent and identically distributed χ^2_1 random variables.

Q2 (20pts) Given an additive error model:

$$Y_{ij} = \mu_i + \varepsilon_{ij},$$

where $\mu_i \in \mathbb{R}$ is unknown, $\varepsilon_{ij} \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)$ with known σ^2 for $1 \le j \le n_i$ and $1 \le i \le n$.

(a) (10pts) Find the MLE for $(\mu_1, \ldots, \mu_n)$ and provide its joint distribution.

(b) (10pts) If the covariate $x_i \in \mathbb{R}^p$ corresponding to data $Y_{i1}, \ldots, Y_{in_i}$ are collected, construct a level α test if $(\mu_1, \ldots, \mu_n)^\top$ lies in the column space of the $n \times p$ data matrix $(x_1, \cdots, x_n)^\top$.

Q3 (30pts) Let $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, where $\mathbf{X}$ is $n \times p$ of full rank p and $\boldsymbol{\varepsilon} \sim \mathcal{N}_n(\mathbf{0}, \sigma^2\mathbf{I_n})$ with an unknown $\sigma^2 > 0$. We apply the method of Lagrange multiplier to find the least squares estimator of $\boldsymbol{\beta}$ subject to linear restriction $\mathbf{A}\boldsymbol{\beta} = \mathbf{c}$, where $\mathbf{A}$ is a known $q \times p$ matrix and $\mathbf{c}$ is a known $q \times 1$ vector.

(a) (10pts) Consider the Lagrange function

$$\mathcal{L}(\boldsymbol{\beta}, \boldsymbol{\lambda}) = (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})^\top(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) + (\boldsymbol{\beta}^\top\mathbf{A}^\top - \mathbf{c}^\top)\boldsymbol{\lambda}.$$

Let $\hat{\boldsymbol{\beta}}_H$ and $\hat{\boldsymbol{\lambda}}_H$ be the solution that minimizes the Lagrange function, $\mathcal{L}(\boldsymbol{\beta}, \boldsymbol{\lambda})$, subject to $\mathbf{A}\boldsymbol{\beta} = \mathbf{c}$. Show that

$$\hat{\boldsymbol{\beta}}_H = \hat{\boldsymbol{\beta}} - \frac{1}{2}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{A}^\top\hat{\boldsymbol{\lambda}}_H \qquad \text{and} \qquad \hat{\boldsymbol{\lambda}}_H = -2[\mathbf{A}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{A}]^{-1}(\mathbf{c} - \mathbf{A}\hat{\boldsymbol{\beta}}),$$

where $\hat{\boldsymbol{\beta}}$ is the ordinary least square estimator.

(b) (10pts) Show that

$$(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}_H)^\top(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}_H) - (\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}})^\top(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}) = \sigma^2\hat{\boldsymbol{\lambda}}_H^\top(\mathrm{Var}[\hat{\boldsymbol{\lambda}}_H])^{-1}\hat{\boldsymbol{\lambda}}_H.$$

(c) (10pts) Construct a level α test for $\mathbf{A}\boldsymbol{\beta} = \mathbf{c}$ in terms of $\hat{\boldsymbol{\lambda}}_H$.

Q4 (30pts) Consider a linear model:

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \varepsilon_i,$$

where $\varepsilon_i \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2)$ with an unknown $\sigma^2 > 0$ for $1 \le i \le n$. Denote the covariate vectors $\mathbf{X}_j = (x_{1j}, \ldots, x_{nj})^\top$ for $j = 1, 2, 3$. Suppose that $\mathbf{X}_j$ is orthogonal with unit length, i.e., $\mathbf{X}_j^\top\mathbf{X}_k = 0$ and $\mathbf{X}_j^\top\mathbf{X}_j = 1$ for $1 \le j < k \le 3$.

(a) (10pts) Construct $(1-\alpha)100\%$ simultaneous confidence intervals (or confidence region) for β_1, β_2, and β_3 with coverage probability **exactly** $1 - \alpha$.

(b) (10pts) Construct a $(1-\alpha)100\%$ confidence band for $\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$ for all $(x_1, x_2, x_3)^\top \in \mathbb{R}^3$.

(c) (10pts) Construct $(1-\alpha)100\%$ simultaneous confidence intervals for a future observation

$$Y^{(i)} = \beta_0 + \beta_1 x^0_{i1} + \beta_2 x^0_{i2} + \beta_3 x^0_{i3} + \varepsilon^0_i,$$

where $x^0_{ij} \in \mathbb{R}$ and $\varepsilon_i \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2)$ for $j = 1, 2, 3$, $1 \le i \le m$.

You may use the following facts:

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

Qualifying Exam August 2022 — 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. [35 points] Evolutionary biologists are interested in the characteristics that enable a species to withstand the selective mechanisms of evolution. An interesting variable in this regard is the brain size. One might anticipate that bigger brains are smarter, but certain penalties are apparently associated with large brains, for example, the need for longer pregnancies and fewer offsprings. Although individual members of the large-brained species may have a better chance of surviving, the benefits for the species must be good enough to compensate for these penalties. In this example, you will examine how the size of brain is associated with several characteristics of the species.
   The original data set includes the average values of the brain weight, body weight, gestation length, and litter size, listed in the following table, for 96 species of mammals.

	Variable name	Description
1	`SPECIES`	Species name
2	`BRAIN`	Average of brain weight in gram
3	`BODY`	Average of body weight in kilogram
4	`GESTATION`	Average length of pregnancy in days
5	`LITTER`	Average number of offspring produced at one birth by an animal

   Due to file transfer errors, you receive incomplete data with 11 observations missing (see the CSV file "brain.csv"). However, your collaborator has already run the multiple linear regression model:

$$\texttt{BRAIN} = \beta_0 + \beta_1 \texttt{BODY} + \beta_2 \texttt{GESTATION} + \beta_3 \texttt{LITTER} + \text{Normal}(0, \sigma^2)$$

   on the complete data (with 96 observations) and obtained the `R` output shown at the top of the next page. Use this output to answer Questions (a) through (e). Note that you may receive no credit if giving numerical answers without any explanation or derivation.

   (a) Interpret the meaning of $\hat{\beta}_2 = 1.81$ in the context of the problem being addressed. [3 points]

```
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  -225.29     83.06    -2.71  0.00797 **
BODY            0.96      0.09    10.46  < 2e-16 ***
GESTATION       1.81      0.35     5.10 1.79e-06 ***
LITTER         27.65     17.41     1.59  0.11579
---
Residual standard error: 224.6 on 92 degrees of freedom
Multiple R-squared:  0.81,Adjusted R-squared:  0.8038
F-statistic:  --- on - and -- DF,  p-value: < 2.2e-16
```

(b) The average weight of humans is 65 kilogram, the average of length of pregnancy is nine months ($\approx$ 270 days) and the average number of offspring is one. What are the fitted value and residual of the average brain weight for humans? [4 points]

(c) Construct a 95% confidence interval for the regression coefficient associated with `GESTATION`. [3 points]

(d) Some output pertaining to the F-statistic is missing in the `R` output.

i. Calculate the F-statistic value as well as the corresponding degrees of freedom [2 points].

ii. Also write down the null hypothesis of this F-test, and interpret the conclusion of this test in the context of the problem being addressed. [4 points]

(e) The variance inflation factor (VIF) of the model is listed below.

BODY	GESTATION	LITTER
2.4927	2.8874	1.5878

i. Based on the information obtained from the complete data set, what can we say about the potential multicollinearity in the predictors? [2 points]

ii. Suppose we perform a multiple linear regression using `BODY` as the response variable, `GESTATION` and `LITTER` as the predictors, what will be the value of R^2? [2 points]

(f) Next you want to work on the incomplete data at hand and investigate the utility of a transformation on the response variable `BRAIN`. Does it improve the model? If yes, please use the new model to answer (g) and (h); otherwise, use the existing one. [5 points]

(g) Following (f), please use an exhaustive search to determine the best model in terms of adjusted R^2. Detail your strategy, write the code with adequate comments, and use appropriate graphs to summarize your result. [5 points]

(h) Following (f), please use an exhaustive search to determine the best model based on the prediction performance. Detail your strategy, write the code with adequate comments, and use appropriate graphs to summarize your result. [5 points]

2. [30 points] Use Monte Carlo simulation to compare the power functions of two tests for normality — the Shapiro Wilks test and Chi-square Goodness of Fit (GOF) test. Let $\mathscr{N}$ denote the family of univariate normal distributions. The hypotheses are

$$H_0 : F_X \in \mathscr{N} \text{ vs. } H_1 : F_X \notin \mathscr{N}.$$

For the Chi-square GOF test, use the following intervals: $(-\infty, -2], (-2, -1], (-1, 0], (0, 1], (1, 2], (2, \infty)$. In each iteration generate a sample of size $n = 25$ from t_k, the Student's t-distribution with k degrees of freedom (k fixed in advance), and apply the two tests. Use the significance level $\alpha = 0.1$. For a fixed k value, generate at least 1,000 Monte Carlo samples. Then change k to a different value and repeat the procedure. Carry out this procedure for a total of 10 different values of k.

(a) Choose the values of k judiciously and explain your choice.

(b) Report the powers in a table and plot the two power curves as functions of k in one plot. Use these to make power comparisons between the two tests.

3. [35 points] An animal study was conducted to test the allergic reactions of a skin care product (treatment). For comparison, one placebo was used as the control. Three animals were randomized to the control group and another three to the treatment group. Each animal was administered the products on the belly and on the back, and skin sensitivity scores were measured at both sites 10 minutes after applying the products. Duplicate tests were run. The data are tabulated below. Use $\alpha = 0.05$ in all statistical tests.

		Measurements	
Group	Animal	Belly	Back
Control	1	73, 70	90, 92
	2	66, 66	83, 83
	3	71, 70	77, 74
Treatment	4	77, 74	89, 88
	5	67, 70	70, 78
	6	76, 77	96, 96

(a) Write the mathematical formula of an appropriate ANOVA model to analyze the data. [5 points]

(b) Construct an ANOVA table, including the expected mean squares. [10 points]

(c) Estimate the main effect of the group. Construct 95% simultaneous confidence intervals for the marginal means of the treatment and placebo groups. Then test whether the treatment is different from the control. Clearly state your conclusions. [10 points]

(d) Is there a difference between the two sites – belly vs. back? Obtain a 95% confidence interval for the difference. [5 points]

(e) Estimate variance components where appropriate. Is there a notable variability among animals? Answer this question with an appropriate statistical test. [5 points]

STATISTICS Ph.D. QUALIFYING EXAM
STATISTICAL INFERENCE
August 2022

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

   (a) (3 points) Find $P(-a \leq X_{(1)} \leq X_{(n)} \leq a)$ for any $0 < a \leq \theta$.

   (b) (5 points) Show that $(X_{(1)}, X_{(n)})$ is sufficient.

   (c) (5 points) Find MLE of θ, $\widehat{\theta}_{MLE}$, and its expectation $\mathrm{E}(\widehat{\theta}_{MLE})$.

   (d) (3 points) Construct an unbiased estimator, $\widehat{\theta}_{unbiased}$, based on (c).

   (e) (4 points) Which of the two estimators in (c) and (d) is more efficient? Justify it.

2. (20 points) Let $X_1, \ldots, X_n$ denote a random sample of size n from a Gamma(k, θ) distribution with probability density

$$f(x) = \frac{1}{\Gamma(k)\theta^k}\, x^{k-1} \exp\left(-\frac{x}{\theta}\right),$$

   where $\theta > 0$ and $k > 0$ are parameters. Note that $\frac{1}{nk}\sum_{i=1}^{n} X_i \sim \text{Gamma}\left(nk, \frac{\theta}{nk}\right)$.

   (a) (10 points) If k is known, to test $H_0 : \theta = \theta_0$ against $H_1 : \theta \neq \theta_0$, construct the likelihood ratio test of size α. Be sure to explicitly find the distribution of the test statistic.

   (b) (10 points) If θ is known, to test $H_0 : k = 1$ against $H_1 : k = 2$, construct the uniformly most powerful test of size α.

3. (20 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) (6 points) $\theta \in (0, \infty)$.

   (b) (6 points) $\theta \in (-\infty, 0)$.

   (c) (8 points) $\theta \in (-\infty, \infty)$, $\theta \neq 0$.

4. (20 points) Suppose X follows a Binomial(n, p) distribution, with n known and p unknown. We want a $100(1-\alpha)\%$ confidence interval for p. The standard textbook interval is the large-sample interval

$$\mathrm{CI}_s = \hat{p} \pm z_{\alpha/2}\widehat{\mathrm{SE}}(\hat{p}),$$

   where $\hat{p} = X/n$ is the sample proportion of successes, $\widehat{\mathrm{SE}}(\hat{p}) = \sqrt{\hat{p}(1-\hat{p})/n}$ is the estimated standard error of $\hat{p}$, and $z_{\alpha/2}$ is the $100(1-\alpha/2)$th percentile of a standard normal distribution.

   (a) (4 points) Show that CI_s is obtained by inverting the acceptance region of the large-sample level α test of $H_0 : p = p_0$ against $H_1 : p \neq p_0$ using the test statistic $(\hat{p}-p)/\widehat{\mathrm{SE}}(\hat{p})$.

(b) (6 points) An alternative to CI_s, called the *Wilson interval*, can be obtained by using the null standard error $\sqrt{p_0(1-p_0)/n}$ instead of the estimated standard error $\sqrt{\hat{p}(1-\hat{p})/n}$ in the test statistic in (a). Show that the resulting interval is

$$\text{CI}_W = \tilde{p} \pm \frac{c\sqrt{n}}{n+c^2}\sqrt{\hat{p}(1-\hat{p}) + c^2/(4n)},$$

where $c = z_{\alpha/2}$ and

$$\tilde{p} = \frac{X + c^2/2}{n + c^2}.$$

(c) (10 points) Yet another alternative to CI_s, called the *Jeffreys interval*, is the $100(1-\alpha)\%$ posterior interval for p assuming a Beta$(1/2, 1/2)$ prior distribution for p. Obtain this interval. Do you expect the coverage probability of this interval to be exactly equal to $1-\alpha$?

5. (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) (5 points) Find UMVUE of $g(\mu)$.

   (b) (5 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) (5 points) Show that the variance in (b) is larger than the bound in (c) but their ratio converges to 1 as $n \to \infty$.

STATISTICS Ph.D. QUALIFYING EXAM
Probability
August 2022

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. (5 points). Let $A_1, A_2, \ldots$ be a sequence of sets, A_n^c denotes a complementary event of A_n. Express $(\limsup_n A_n)^c$ via sets A_n^c and prove your assertion.
Hint: Recall how to write $\limsup_n$ via classical set operations.

Problem 2. (5 points) Give definition of Borel sets on R.
Hint: To get 5 points you need to give definitions of all notions used.

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

Problem 4. (5 points). Formulate 3 Axioms of Kolmogorov. Then prove that probability of the null event is zero.

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 convergences, using set operations may be very useful here.

Problem 6. (10 points). Measure-Theoretical definition of the conditional expectation of Y given X. Also prove uniqueness of the defined conditional expectation.
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. (10 points) Formulate and prove the second Borel-Cantelli Lemma.

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 be said about $\{g(X_n), \mathcal{F}_n\}$? Prove your assertion.
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. (5 points). Let B_t be a Brownian motion. Give its probabilistic (not via Fourier extension) definition and then find $E\{B_t B_{t+s}\}$. Prove your assertion. (5 points)

Problem 11. (5 points). What is the difference (if any) between stable and infinitely divisible distributions? Give examples.

Qualifying Exam Optimization

Do any four out of five problems.

1/ Discuss the role of the Cauchy point in obtaining conditions for the global convergence of trust region methods.

2/ Compare the advantages and disadvantages of the SR1 and BFGS quasi-Newton methods.

3/ State the KKT theorem and discuss how Farkas' lemma is used in its proof.

4/ Describe the null-space method for the direct solution of the KKT system arising in equality constrained quadratic programs.

5/ Discuss the use of active set methods for solving convex quadratic programs containing equality and inequality constraints. What happens when convexity is not assumed?