QE ID (given by Angie)

January 2021 Qualifying Exam in Linear Models

January 4, 2021

Instruction:

- This is a closed-book test.
- There are three 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 blank sheets of paper.
- 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, and Q3 are similar, they are independent, standalone problems.
- The total possible points are 100.

Q1 Consider a regression model

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

where

$$\mathbf{Y} = \begin{pmatrix} Y_{11} & Y_{12} \\ Y_{21} & Y_{22} \\ \vdots & \vdots \\ Y_{n1} & Y_{n2} \end{pmatrix}, \quad \mathbf{X} = \begin{pmatrix} 1 & X_{11} & X_{12} \\ 1 & X_{21} & X_{22} \\ \vdots & \vdots & \vdots \\ 1 & X_{n1} & X_{n2} \end{pmatrix}, \quad \beta = \begin{pmatrix} \beta_{10} & \beta_{20} \\ \beta_{11} & \beta_{21} \\ \beta_{12} & \beta_{22} \end{pmatrix}, \quad \text{and} \quad \varepsilon = \begin{pmatrix} \varepsilon_{11} & \varepsilon_{12} \\ \varepsilon_{21} & \varepsilon_{22} \\ \vdots & \vdots \\ \varepsilon_{n1} & \varepsilon_{n2} \end{pmatrix}.$$

The elements ε_{ij} in ε are iid from $N(0, \sigma^2)$ with unknown $\sigma^2 > 0$. Define a loss function of β:

$$L(\beta) = \sum_{i=1}^{n} \sum_{j=1}^{2} (Y_{ij} - \beta_{j0} - X_{i1}\beta_{j1} - X_{i2}\beta_{j2})^2.$$

(a) (8 pts) Derive an estimator $\hat{\beta}$ of β minimizing $L(\beta)$.

(b) (8 pts) Show that $\hat{\beta}$ is the best linear unbiased estimator (BLUE).

(c) (7 pts) Find the joint distribution of the vectorized form of $\hat{\beta}$.

(d) (7 pts) Construct a level α test for $H_0 : \beta_{11} = \beta_{21}$ and $\beta_{12} = \beta_{22}$.

Q2 Consider a linear regression model

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

where $\varepsilon_{ij} \overset{\text{iid}}{\sim} N(0, \sigma^2)$ for $i = 1, \ldots, n_j$ and $j = 1, 2, 3$ and $\sigma^2 > 0$ is unknown. Define $\overline{Y}_{\cdot j} = \sum_{i=1}^{n_j} Y_{ij}/n_j$ and $\overline{Y}_{\cdot\cdot} = \sum_{j=1}^{3} \sum_{i=1}^{n_j} Y_{ij}/n$ where $n = \sum_{j=1}^{3} n_j$.

(a) (8 pts) Show that

$$\frac{\sum_{j=1}^{3} n_j (\overline{Y}_{\cdot j} - \overline{Y}_{\cdot\cdot})^2}{\sum_{j=1}^{3} \sum_{i=1}^{n_j} (Y_{ij} - \overline{Y}_{\cdot j})^2} \sim F_{2,n-3}$$

when $\mu_1 = \mu_2 = \mu_3$. [Hint: Find the least squares estimators of μ_1, μ_2, and μ_3 with and without $\mu_1 = \mu_2 = \mu_3$ constraint.]

(b) (8 pts) Construct a level α test for $H_0 : \mu_1 = \mu_2 = \mu_3$.

(c) (7 pts) Construct a $(1 - \alpha)\%$ Scheffé's simultaneous confidence intervals for all pairwise differences $\mu_{j_1} - \mu_{j_2}$ for $1 \leq j_1 < j_2 \leq 3$.

(d) (7 pts) When $\mu_1 \leq \mu_2 \leq \mu_3$ is acknowledged, construct a level α test for $H_0 : \mu_1 = \mu_2 = \mu_3$.

Q3 Consider a linear regression model

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

where Y is a $n \times 1$ response vector, $\mathbf{X}$ is a $n \times (p + 1)$ data matrix, β is a $(p + 1) \times 1$ parameter vector. The error vector ε follows a multivariate normal distribution with mean zero and semi-positive definite variance-covariance matrix $\sigma^2\mathbf{V}$ where $(\mathbf{V})_{ij} = \rho I(i \neq j) + I(i = j)$ for $1 \leq i, j \leq n$, where I is the indicator function.

(a) (6 pts) Show that $-1 \leq (n - 1)\rho$. [Hint: find the variance of $\varepsilon^{\mathrm{T}}\mathbf{1}_n$, where $\mathbf{1}_n$ is the vector with elements all equal to unity.]

(b) (7 pts) Find the mean and variance-covariance matrix of the least squares estimator $\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p)^{\mathrm{T}}$ of $\beta = (\beta_0, \beta_1, \ldots, \beta_p)^{\mathrm{T}}$ in terms of $\mathbf{X}$, σ^2, and ρ.

(c) (7 pts) Let $\nu_1(\rho)$ be the variance of $\hat{\beta}_1$. Specify ranges of ρ such that $\nu_1(\rho) > \nu_1(0)$.

Q4 Consider the linear regression model in Q3 with $\mathbf{V} = \mathbf{I}_n$ where $\mathbf{I}_n$ is the identity matrix. We further write $Y = (Y_1, \ldots, Y_n)^T$, $\mathbf{X} = (x_0, x_1, \ldots, x_p)$, where $x_j = (x_{1j}, \ldots, x_{nj})^{\mathrm{T}}$ for $j = 0, 1, \ldots, p$, $x_0 = \mathbf{1}_n$, and $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)^{\mathrm{T}}$. $\mathbf{1}_n$ is the vector with elements all equal to unity. Recall that the least squares estimator is denoted by $\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p)^{\mathrm{T}}$.

(a) (10 pts) You have been told that the first observation is contaminated such that

$$Y_1 = \mathbf{x}_1\beta + \varepsilon_1 + \delta_1,$$

where $\mathbf{x}_1 = (1, x_{11}, \ldots, x_{1p})^{\mathrm{T}}$, $\delta_1 \sim N(0, \sigma^2)$ and δ_1 is independent from the error vector ε. Propose a minimum variance unbiased estimator $\hat{\beta}^* = (\hat{\beta}_0^*, \hat{\beta}_1^*, \ldots, \hat{\beta}_p^*)^{\mathrm{T}}$ of β adjusting the contamination.

(b) (10 pts) Assume that x_j are standardized such that $x_j^{\mathrm{T}} x_0 = 0$ and $x_j^{\mathrm{T}} x_j = 1$ for $j = 1, \ldots, p$. If $x_i^{\mathrm{T}} x_j = 0$ for $i \neq j$ except $x_1^{\mathrm{T}} x_2 = r$, find r such that $Var(\hat{\beta}_1) = \infty$ and $Var(\hat{\beta}_2) = \infty$. [Hint: You may see the variances of $\hat{\beta}_1$ and $\hat{\beta}_2$ for $p = 2$. You may use the fact that 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 $\mathbf{B}_{12} = \mathbf{B}_{21}^{\mathrm{T}}$.]

Qualifying Exam January 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 "`blood_pressure.dat`", "`wine.txt`", and "`diamonds.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 12:20pm. 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 blood pressure data in "`blood_pressure.dat`" containing two columns, age and diastolic blood pressure, in that order. A simple linear regression model of blood pressure on age reveals (through diagnostics) that constant variance assumption is violated (you don't need to check it). Thus, we consider fitting a weighted least squares (WLS) regression model of diastolic blood pressure on age. [30 points]

   (a) Use iteratively weighted least squares (IWLS) method to fit the WLS model. Use weights based on residuals, i.e., follow these steps — (1) regress Y on X using OLS method and obtain the residuals, (2) estimate the standard deviation function by regressing the absolute residuals on X and then use the fitted standard deviation function to obtain weights, (3) use WLS to regress Y on X and obtain the estimate of the slope and residuals, and (4) repeat steps 2 and 3 until convergence of the estimate of slope (this will not require more than a few iterations). Report the final WLS estimate of slope (b_1), its estimated standard error (SE), and 95% confidence interval (CI) of slope (β_1). [10 points]

   (b) Next use the bootstrap method for estimating SE of WLS estimate b_1 and a 95% CI for β_1. For this, generate 1000 bootstrap samples using random X sampling, i.e., by re-sampling of (X, Y) pairs. For each bootstrap sample, follow these steps — (1) regress Y on X using OLS and obtain the residuals, (2) estimate the standard deviation function by regressing the absolute residuals on X and then use the fitted standard deviation function to obtain weights, and (3) use WLS to regress Y on X and obtain the bootstrap estimated regression coefficient b_1^*. Note that 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 three bootstrap methods — normal approximation, basic bootstrap, and percentile bootstrap. **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). [17 points]

   (c) Compare the three bootstrap CIs with each other as well as with the CI reported in part (a). [3 points]

2. Work on fitting regression models based on data in "wine.txt", which come from a study concerning Pinot Noir wine quality. It contains 38 samples and 7 variables, including: `Quality`, `Clarity`, `Aroma`, `Body`, `Flavor`, `Oakiness`, and `Region`. The goal is to identify variables that affect the quality of Pinot Noir, and can potentially be used to predict the quality of the wine. [35 points]

   (a) Make numerical and graphical summaries of the data. Comment on anything that you find interesting. [4 points]

   (b) Factorize the only categorical variable (i.e. `Region`) using an appropriate function in the programming software you choose (e.g. the function `factor` if you use `R`), and fit a linear regression model with `Quality` as the response variable and the other variables (including the factorized categorical variable) as the predictors. Check assumptions of the linear model. Check for influential points, if any, and exclude those points for all subsequent analysis. [9 points]

   (c) Investigate the utility of a transformation on the response variable. Does it improve the model? If yes, work on the new model after transformation for all subsequent subproblems; if no, carry forward the model described in (b) [4 points]

   (d) Using an exhaustive search to determine the best models (those you consider good candidates for detailed evaluation) using the adjusted R^2, Mallows's C_p, and Bayesian information criterion (BIC), respectively. Among the models determined with the above criteria, which model would you prefer and why? [10 points]

   (e) Check assumptions of the model that you have selected in (d). [4 points]

   (f) Examine all possible models (with `Quality` as the response variable) that have a two-way interactive term containing one qualitative predictor (i.e. `Region`) and one quantitative predictor. Compare the best of them with the model that you have selected in (d). [4 points]

3. The dataset "diamonds.csv" contains the prices and other attributes of diamonds. The variables are "four C's" – cut, carats, color and clarity - that are characteristics of the quality of diamond gemstones. The quality of the cut may be Fair, Good, Very Good, Premium, Ideal. The carats measure the weight of a diamond (one carat = 200 mg). Diamonds are graded on a color scale from D (colorless) to Z (light yellow). Colorless stones have scales from D to F, which are rare. Colors G-J are nearly colorless. The difference between two adjacent color levels, e.g. E and F, is quite small, especially to the untrained eye. The clarity of a diamond can be reduced by imperfections due to fractures/minerals, which interrupt the carbon crystal structure. The major grades of clarity are IF (internally flawless), VVS (very, very slightly imperfect), VS (very slightly imperfect), SI (slightly imperfect), and I (imperfect). Within the letter grades lower numbers are better, e.g. VS1 has greater clarity than VS2.

   The dataset only has three levels of color (E, F, G), two levels of cut (Fair and Good), and eight levels of clarity (I1, SI2, SI1, VS2, VS1, VVS2,VVS1, IF). Your goal is to explore the relations between price and the four C's based on the given data. [35 points]

   (a) First investigate cut and color. Draw a plot to explore the relationship between price and cut and color. Fit a model. Verify the assumptions and consider remedy measures if needed. Is the price of a diamond affected by its cut and color? Explain your findings. [13 points]

   (b) Next consider additional variables and fit an appropriate model. What, if any, of the fours C's can predict the price? Provide a brief discussion for comparison with your findings in part (a). [10 points]

   (c) Consider the following two hypotheses:

      i. Conditional on carat, the other three C's have no effect on the price.

      ii. The effect of carat on the price does not depend on any of the other C's.

      Translate (i) and (ii) into two hypotheses on the parameters of an appropriate model. Conduct the tests of the two hypotheses. What are the conclusions? [12 points]

STATISTICS Ph.D. QUALIFYING EXAM
STATISTICAL INFERENCE

January 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. (15 points) Let X be a continuous random variable whose probability distribution depends on a single unknown parameter $\theta \in \Theta$. Let $F(x|\theta)$ be the cumulative distribution function (cdf) of X and $Q_p(\theta)$ be the pth quantile of X for a specified $p \in (0, 1)$. Assume that the family of cdfs $\{F(x|\theta), \theta \in \Theta\}$ is *stochastically nondecreasing in* θ, i.e., for each x, $F(x|\theta)$ is a nonincreasing function of θ. Thus,

$$\theta_1 < \theta_2 \implies F(x|\theta_1) \geq F(x|\theta_2) \text{ for all } x.$$

   (a) (5 points) Show that $Q_p(\theta)$ is a nondecreasing function of θ.

   (b) (5 points) Suppose U is a $100(1-\alpha)\%$ upper confidence bound for θ. Show that $Q_p(U)$ is a $100(1-\alpha)\%$ upper confidence bound for $Q_p(\theta)$.

   (c) (2 points) What does (b) suggest regarding constructing confidence bounds for a quantile?

   (d) (3 points) As an example, give a distribution for X whose cdf is stochastically nondecreasing in θ.

2. (25 points) Suppose that $X \sim$ Binomial $(10, \theta)$ distribution, where $\theta \in \Theta = \{1/2, 2/3\}$ is the unknown parameter. Notice that the parameter space Θ has only two points. We are interested in testing $H_0 : \theta = 1/2$ against $H_1 : \theta = 2/3$, using classical and Bayesian approaches.

   (a) *Classical approach*:

      i. (6 points) Find the critical region of the most powerful level $\alpha = 0.06$ non-randomized test of the above hypotheses.

      ii. (2 points) Suppose $X = 8$ is observed. Will your test accept or reject H_0?

      iii. (4 points) Find the p-value of your test assuming $X = 8$ is observed.

   (b) *Bayesian approach*: Assume that the prior distribution of θ is such that the prior odds ratio $P(\theta = 1/2)/P(\theta = 2/3) = c$, where $c > 0$ is known.

      i. (6 points) Find the posterior odds ratio $P(\theta = 1/2\,|X = x)/P(\theta = 2/3\,|X = x)$ as a function of the observed data $X = x$ and the prior odds ratio c.

      ii. (7 points) Suppose $x = 8$ and $c = 1$. Find the posterior probability of H_0? Will you accept or reject H_0?

3. (15 points) Let $X \sim N(\mu, 2)$, $Y \sim N(2\mu, 1)$, and X and Y are independent. We would like to construct a confidence interval for μ based on the data (X, Y).

   (a) (5 points) Prove that $(1-c)X + cY/2 - \mu$ is a pivotal quantity for any given constant c.

(b) (5 points) Use the pivotal quantity in (a) to construct a two-sided equi-tailed 95% confidence interval for μ. Your answer should depend on the constant c.

(c) (5 points) What is the best choice of c for your interval?

4. (25 points) Let $X_1, \dots, X_n$ represent a random sample from a uniform distribution on $(0, \theta)$, where $\theta \in (0, 1)$ is the unknown parameter. We would like to estimate θ, with loss function $L(\theta, a) = (a - \theta)^2/\theta^2$. Assume that the prior distribution of θ is uniform over $(0, 1)$ and $T = \max\{X_1, \dots, X_n\}$.

   (a) (10 points) Find the posterior distribution of θ.

   (b) (15 points) Prove that the Bayes estimator of θ is given by
$$\frac{n+1}{n} T \frac{1 - T^n}{1 - T^{n+1}}.$$

5. (20 points) Consider a finite population (or universe) of N units denoted by the index set $\mathcal{U} = \{1, 2, \dots, N\}$. Let $\mathcal{S} \subset \mathcal{U}$ denote the index set of a sample from this population. The sample is selected by *probability sampling* so that $p(s)$ is the known probability of selecting the sample $\mathcal{S} = s$ with $\sum_s p(s) = 1$. For $k \in \mathcal{U}$, let I_k be the binary indicator of selection of unit k in the sample. Let
$$\pi_k = P(I_k = 1) \text{ and } \pi_{kl} = P(I_k = 1, I_l = 1),\ \ k, l \in \mathcal{U},$$
be respective probabilities that unit k and both units k and l are selected in the sample.

   A concrete example of probability sampling is *simple random sampling* in which case all subsets of size n from N units in the population are equally-likely. Thus, in this case, each possible sample $\mathcal{S} = s$ has probability $1/\binom{N}{n}$; and
$$\pi_k = \frac{n}{N},\ \pi_{kl} = \frac{n}{N}\frac{n-1}{N-1}.$$

   In what follows, we assume probability sampling, and not just its special case of simple random sampling. The probabilities π_k and π_{kl} are *known* since $p(s)$ is known. Assume that $\pi_k > 0$ for all $k \in \mathcal{U}$. Let $y_k, k \in \mathcal{U}$ denote the values of a variable y in the population. The target of inference is the population mean, $\bar{y}_U = \sum_{k \in \mathcal{U}} y_k/N$. Let $y_k, k \in \mathcal{S}$ denote the sample values. The values of y_k are observed only for $k \in \mathcal{S}$. Assuming N known, the standard estimator of $\bar{y}_U$ is the Horvitz-Thompson estimator,
$$\bar{y}_S = \frac{\sum_{k \in \mathcal{S}} w_k y_k}{N} = \frac{\sum_{k \in \mathcal{U}} w_k I_k y_k}{N},$$
where $w_k = 1/\pi_k$ is the weight of unit k. For inference here, the population values $y_1, \dots, y_N$ are viewed as fixed constants and the index set $\mathcal{S}$ of the sample is treated as random. Therefore, the sampling distribution of an estimator is induced by randomness in the selection indicators I_k.

   (a) (4 points) Show that $\bar{y}_S$ is unbiased, i.e., $E(\bar{y}_S) = \bar{y}_U$.

   (b) (7 points) Show that
$$\text{var}(\bar{y}_S) = \frac{1}{N^2} \sum_{k \in \mathcal{U}} \sum_{l \in \mathcal{U}} \Delta_{kl} \frac{y_k}{\pi_k} \frac{y_l}{\pi_l},$$
   where $\Delta_{kl} = \pi_{kl} - \pi_k \pi_l$.

   (c) (9 points) Find an unbiased estimator of $\text{var}(\bar{y}_S)$ in (b).

STATISTICS Ph.D. QUALIFYING EXAM
Probability
January 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. (5 points) Definition of the probability space $(\Omega, \mathcal{F}, P)$. Hint: Give definitions of all three components; do not forget about Kolmogorov.

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

Problem 3. (10 points) Definition of Lebesgue integral. Hint: Recall several steps in the definition.

Problem 4. (5 points) Formulate Radon-Nikodym Theorem. If in its assumptions you use some notions - define them.

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. In the proof you may use Fatou's lemma without proof, only formulate it.

Problem 6. (10 points) Measure-Theoretical definition of the conditional expectation. Hint: Begin with "Let Y be an extended random variable on $(\Omega, \mathcal{F}, P)$, and $X : (\Omega, \mathcal{F}) \to (\Omega', \mathcal{F}')$ be a random object. If $E\{Y\}$ exists, ..." Prove uniqueness of the defined conditional expectation.

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

Problem 8. (10 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. (5 points) Formulate Lindeberg's Central Limit Theorem (without proof).

Problem 10. (35 points) Formulate and prove the law of iterated logarithm for Brownian motion. Hint: Do not forget to begin with the classical probability definition of the considered process. Also, if you out of time, just formulate relations (like useful inequalities) that you use in the proof.