QE ID (given by Angie)

August 2020 Qualifying Exam in Linear Models

August 3, 2020

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 you have prepared.
- 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 Assume that Y is a random variable and $\mathbf{X}$ is a random vector of length $p-1$ such that $(Y, \mathbf{X}^\top)^\top$ has a multivariate normal distribution:

$$\begin{pmatrix} Y \\ \mathbf{X} \end{pmatrix} \sim \mathcal{N}_{p+1}\left(\begin{pmatrix} \mu_Y \\ \mu_{\mathbf{X}} \end{pmatrix}, \begin{pmatrix} \sigma^2 & \mathbf{\Sigma_{YX}} \\ \mathbf{\Sigma_{XY}} & \mathbf{\Sigma_{XX}} \end{pmatrix} \right),$$

where $(\mu_Y, \mu_{\mathbf{X}}^\top)^\top \in \mathbb{R}^p$, $\sigma^2 > 0$, $\mathbf{\Sigma_{XX}}$ is a non-singular matrix, and $\mathbf{\Sigma_{XY}} = \mathbf{\Sigma_{YX}^\top}$ is a non-zero vector.

(a) (5 pts) Find a linear combination of $\mathbf{X}$, say $\mathbf{a}^\top\mathbf{X}$, such that $Y - \mathbf{a}^\top\mathbf{X}$ and $\mathbf{X}$ are independent.

(b) (10 pts) Derive the moment generating function of $(Y - \mathbf{a}^\top\mathbf{X})^2$.

Q2 Consider a linear model:

$$Y_i = \beta_0 + \sum_{j=1}^{p-1} \left\{ \left(\frac{x_{ij} - m_j}{M_j - m_j} \right) - \frac{1}{2} \right\} \beta_j + \varepsilon_i,$$

where M_j and m_j are the maximum and minimum values of $x_{1j}, \ldots, x_{nj}$ for $j = 1, 2, \ldots, p-1$, and $\varepsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2)$ with known $\sigma^2 > 0$.

(a) (10 pts) Derive the maximum likelihood estimator (MLE) for $(\beta_0, \beta_1, \ldots, \beta_{p-1})$.

(b) (10 pts) Given the MLE $(\hat\beta_0, \hat\beta_1, \ldots, \hat\beta_{p-1})$ from 2a and the predictor $\hat Y(\underline{\mathbf{x}}) = \hat\beta_0 + \sum_{j=1}^{p-1} x_j \hat\beta_j$, where $\underline{\mathbf{x}} = (x_1, \ldots, x_{p-1})^\top$, derive a $100(1-\alpha)\%$ confidence interval for $\hat Y(\underline{\mathbf{x}})$.

(c) (5 pts) Find the $\underline{\mathbf{x}}$ that gives the narrowest confidence interval in 2b.

Q3 Consider a linear model:

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i1} x_{i2} + \varepsilon_i, i = 1, 2, \ldots, n,$$

where $\varepsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2)$ with unknown $\sigma^2 > 0$. In the following questions, you are allowed to use the notations of the regression coefficients: $\hat\beta_0, \ldots, \hat\beta_3$ without the derivation of the regression coefficients via least squares method.

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

(b) (10 pts) Construct a level α test for

$$\begin{cases} H_0: & \beta_1 = \beta_2 = \beta_3 = 0; \\ H_1: & \beta_j \neq 0 \text{ for some } j = 1, 2, 3. \end{cases}$$

(c) (10 pts) Construct $(1-\alpha)100\%$ simultaneous prediction intervals for future values

$$Y^{(i)} = \beta_0 + \beta_1 x_{i1}^0 + \beta_2 x_{i2}^0 + \beta_3 x_{i1}^0 x_{i2}^0 + \varepsilon_i^0,$$

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

(d) (10 pts) Construct a $(1-\alpha)100\%$ confidence band for $(\beta_0 + \beta_2) + (\beta_1 + \beta_3)x$ for all $x \in \mathbb{R}$.

Q4 Consider a two-phase linear regression model:

$$Y_{ki} = x_{ki}\beta_k + \varepsilon_{ki}, k = 1, 2, i = 1, \ldots, n,$$

$$x_{11} < x_{12} < \cdots < x_{1n} < 0 < x_{21} < x_{22} < \cdots < x_{2n},$$

and $\varepsilon_{ki} \overset{\text{iid}}{\sim} N(0, \sigma^2)$ with known $\sigma^2 > 0$.

(a) (10 pts) Derive the MLE $(\hat\beta_1, \hat\beta_2)$ for (β_1, β_2) and provide the joint distribution of $(\hat\beta_1, \hat\beta_2)$.

(b) (10 pts) Provide a level α test for $H_0: \beta_1 \le \beta_2$ versus $H_1: \beta_1 > \beta_2$.

Qualifying Exam August 2020 — 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.
- Go to http://www.utdallas.edu/~mchen/QE and download data files "`oxygen_saturation.txt`", "`prostate.csv`", and "`fillings.txt`". 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. [30points] Consider the oxygen saturation data stored in `"oxygen_saturation.txt"`, which contains measurements of the blood oxygen levels, defined as the percent of oxygen-saturated hemoglobin relative to total hemoglobin, of 72 adults. The oxygen level of every subject is measured using two devices, an oxygen saturation monitor (`OSM`, method 1) and a pulse oximetry screener (`POS`, method 2). We are primarily interested in evaluating *the agreement* between the two methods for measuring the oxygen saturation.

   (a) Make a scatterplot of the paired data and superimpose the 45^o line. Based on the plot, comment on the extent of agreement between the two methods for measuring oxygen saturation (Hint: Think about when the two methods would agree perfectly). [3 points]

   (b) Let the random variables X_1 and X_2 represent the observations from the two methods for a randomly selected subject from the population. Further, let $\mu_j = E(X_j)$ and $\sigma_j^2 = Var(X_j)$, $j = 1, 2$, $\sigma_{12} = Cov(X_1, X_2)$, and $\rho = Corr(X_1, X_2)$. Note that perfect agreement implies that $P(X_1 = X_2) = 1$. Argue that perfect agreement corresponds to $\{\mu_1 = \mu_2, \sigma_1 = \sigma_2, \rho = 1\}$. [4 points]

   (c) Let θ be the *concordance correlation coefficient* (CCC) between the two methods. It is a measure of agreement between the methods and is defined as

   $$\theta = \frac{2\sigma_{12}}{(\mu_1 - \mu_2)^2 + \sigma_1^2 + \sigma_2^2} = \rho \frac{2\sigma_1\sigma_2}{(\mu_1 - \mu_2)^2 + \sigma_1^2 + \sigma_2^2}.$$

   Argue that perfect agreement implies $\theta = 1$. Provide a point estimate $\hat{\theta}$ of θ. [4 points]

   (d) Write your own code to compute (nonparametric) bootstrap estimates of bias and standard error of $\hat{\theta}$, and a 95% *lower confidence bound* for θ computed using the percentile method. Set the starting seed as 1111 and obtain 1000 bootstrap replications. Interpret the results. [15 points]
   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).

   (e) State your conclusion about the extent of agreement between the two methods. Would you say that the methods agree well enough to be used interchangeably in practice? [4 points]

2. [35 points] Consider the data in "`prostate.csv`" from a study of the prostate cancer. Prostate specific antigen, or PSA, is a protein produced by normal, as well as malignant cells of the prostate gland. The blood level of PSA is often elevated in men with prostate cancer. In this study researchers examined the correlation between the level of PSA and a number of clinical measures in men who were about to receive a radical prostatectomy, a surgical procedure for the partial or complete removal of the prostate. The data contain 97 samples and 8 variables listed as follows:

#	Variable name	Description
1	`cavol`	Cancer volume
2	`weight`	Prostate weight
3	`age`	Age
4	`bph`	Benign prostatic hyperplasia amount
5	`cp`	Capsular penetration
6	`gleason`	Gleason score
7	`pgg45`	Percentage Gleason scores 4 or 5
8	`psa`	Prostate specific antigen (PSA)

(a) Make a plot to investigate the relationship between `psa` and `cavol`. Fit a simple linear regression model (Model 1) to predict `psa` based on `cavol`. Is it a good model? Explain. [2 points]

(b) Investigate the utility of an appropriate transformation on the response variable in Model 1. Call the new model after transformation Model 2. Can you compute the expected change in the PSA corresponding to one unit increase in the cancer volume? Explain. [2 points]

(c) Continue to work on Model 2. What is the definition of R^2 in the software output for Model 2? Show your work step-by-step to compute this value. [2 points]

(d) Next fit a multiple linear regression model (call it Model 3) to predict `psa` (in the transformed scale using the transformation you found in (b)) based on `weight`, `age`, `bph`, `cp`, `gleason`, and `pgg45`. Check the assumptions of the linear models and identify potential influential points. Would R^2 go up, go down, or stay exactly the same, compared to that of Model 2? [5 points]

(e) Continue to work on Model 3, but train the model only with the training samples which are specified by the last column (`is_training`) in the data set. Find the confidence intervals for the mean PSA for those in the test set. What is the test error in terms of mean square error (MSE). Create a plot to visualize your result. [6 points]

(f) Continue to work on Model 3. It is often helpful to transform the predictor variable as well. In practice, the log transformations often work well for this purpose. Repeat (e) with log-transformed `weight`, `bph`, and `cp` and call it Model 4. Is it a better model than Model 3? For a robust answer, you will need to randomly split the data set into training and test sets of equal sizes, fit the models, repeat this procedure 10 times, and compare the test errors. [8 points]

(g) Find the "best" model (Model 5) by performing forward or backward step-wise regression to predict `psa` (in the transformed scale with the transformation determined in (b)). Is it a better model than Model 4? Explain from a practical point of view. [5 points]

(h) Perform a statistical test that compares Model 4 and Model 5 with all samples. Clearly state the hypotheses associated with this test and interpret the results. Explain which one is better from a theoretical point of view. [5 points]

3. [35 points] In conservative and restorative dentistry, as well as in orthodontics, gold alloyed together with other metals is widely used as fillings. This type of dental filling is generally considered the most durable, usually lasting 20 years or more. The alloy can vary in hardness depending on how the metal is treated during the manufactory process. The alloy type (`alloy`, 8 levels) and the condensation method (`condensation`, 3 levels) are two factors thought to influence the hardness (`hardness`). In addition, some dentists performing the filling are better at some types of fillings than others. Five

dentists (`dentist`) were selected at random, each preparing 24 fillings in random order, one for each of the combinations of alloy types and condensation methods. The hardness of fillings are measured with a continuous score (big numbers are better). The data are stored in "`fillings.txt`".

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

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

(c) Obtain the analysis of variance table. Test the main effects and possible interaction effects using $\alpha = 0.05$. Clearly specify the hypotheses, obtain appropriate test statistics, draw conclusions and carefully interpret your findings. [15 points]

(d) Construct confidence intervals of all pairwise differences in means of the main effect(s) that is(are) significant in part (c). Use the most efficient multiple comparison procedure with a 95% family-wise confidence level. Explain your results[5 points]

STATISTICS Ph.D. QUALIFYING EXAM
STATISTICAL INFERENCE
August 2020

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

   (b) (2 points) Use (a) to argue that the marginal distribution of Y_i is *not* Poisson.

   (c) (4 points) Find $cov(Y_1, Y_2)$.

   (d) (4 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.)

2. (25 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

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

where $\hat{p} = X/n$ is the sample proportion of successes, $\widehat{\text{SE}}(\hat{p}) = \sqrt{\hat{p}(1-\hat{p})/n}$ 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) (5 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) (10 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

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

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 $1 - \alpha$?

3. (15 points) Let $X_1, \ldots, X_n$ be a random sample from the triangular distribution with density function

$$f(x) = \begin{cases} 1 - |x|, & -1 \leq x \leq 1, \\ 0, & \text{otherwise.} \end{cases}$$

Let $X_{(i)}$ be the ith order statistic. Derive an explicit expression for $E\left(X_{(i)}\right)$.

4. (15 points) Let $X_1, \ldots, X_n$ be a random sample from a $N(\mu, \sigma^2)$ distribution with both parameters unknown. Assume that $n > 2$.

(a) (3 points) Find the uniformly minimum variance unbiased estimator (UMVUE) for μ.

(b) (7 points) Find the UMVUE for σ.

(c) (5 points) For a given probability $p \in (0, 1)$, find the UMVUE for the pth percentile of the population distribution.

5. (15 points) Let $X_1, \ldots, X_n$ be a random sample from the uniform distribution on the interval $(0, \theta)$, where $\theta > 0$ is unknown. Let the prior for θ be the log-normal distribution with parameter (μ_0, σ_0^2), where $\mu_0 \in \mathbb{R}$ and $\sigma_0 > 0$ are known constants. In other words, the prior for $\log(\theta)$ is $N(\mu_0, \sigma_0^2)$.

(a) (8 points) Find the posterior density of $\log(\theta)$.

(b) (7 points) Find the posterior mode of $\log(\theta)$.

6. (15 points) Let X_1 and X_2 follow independent exponential distributions with parameters θ_1 and θ_2, respectively. Define $\theta = \theta_1/\theta_2$. We would like to test $H_0 : \theta = 1$ against $H_1 : \theta \neq 1$.

   (a) (8 points) Show that the likelihood ratio statistic is $4F/(1+F)^2$, where $F = X_2/X_1$.

   (b) (7 points) Derive the rejection region for a size α likelihood ratio test.

STATISTICS Ph.D. QUALIFYING EXAM
Probability
August 2020

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. Definitions of a σ-field, a measurable function, and Lebesgue integral. (5 points)

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

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

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

Problem 5. 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. (10 points)

Problem 6. 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. Formulate and prove the second Borel-Cantelli Lemma. (10 points)

Problem 8. 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. Prove that $\{g(X_n), \mathcal{F}_n\}$ is also a submartingale. Note: Please give definitions of all notions mentioned in the problem. (10 points)

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

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