QE ID (given by Angie)

January 2022 Qualifying Exam in Linear Models

January 3, 2022

Instruction:

- This is a closed-book closed-notes test. There are 4 independent standalone questions, Q1, Q2, Q3, and Q4, each of which may have multiple parts. The total number of points possible is 100.
- 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.
- Write your solutions on the blank sheets of paper that you prepared with. Begin each question on a new sheet, arrange your sheets in order, and number each sheet.
- On all answer sheets, write your QE ID number (given to you by Angie) and identify which question and part is being answered. **DO NOT** put your name, UTD ID, or any other identifying information on any of your answer sheets.

You may use the following facts:

- $$\frac{d\mathbf{A}^\top\boldsymbol{\beta}}{d\boldsymbol{\beta}} = \mathbf{A}, \qquad \frac{d\boldsymbol{\beta}^\top\mathbf{A}\boldsymbol{\beta}}{d\boldsymbol{\beta}} = 2\mathbf{A}\boldsymbol{\beta} \quad (\mathbf{A}\text{ is symmetric.})$$
- 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 $\mathbf{B}_{12} = \mathbf{B}_{21}^\top$.
- let $\boldsymbol{Z}$ be an $n \times 1$ vector of random variables, and let $\mathbf{A}$ be an $n \times n$ symmetric matrix. If $E(\boldsymbol{Z}) = \boldsymbol{\mu}$ and $Var(\boldsymbol{Z}) = \boldsymbol{\Sigma}$, then
$$E(\boldsymbol{Z}^\top\mathbf{A}\boldsymbol{Z}) = tr(\mathbf{A}\boldsymbol{\Sigma}) + \boldsymbol{\mu}^\top\mathbf{A}\boldsymbol{\mu},$$
where $tr(\mathbf{A}\boldsymbol{\Sigma})$ is the trace of matrix $\mathbf{A}\boldsymbol{\Sigma}$.

Q1 (10 pts) Given the regression model

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i, i = 1, \ldots, n,$$

where ϵ_i's are independent with $\mathrm{E}(\epsilon_i) = 0$ and $\mathrm{Var}(\epsilon_i) = \sigma^2 x_i^2$. Show that the weighted least squares estimates for (β_0, β_1) are equivalent to the ordinary least squares estimates for that under the alternative model

$$\frac{Y_i}{|x_i|} = \beta_1 + \frac{\beta_0}{|x_i|} + \zeta_i, i = 1, \ldots, n,$$

where ζ_i's are independent with $\mathrm{E}(\zeta_i) = 0$ and $\mathrm{Var}(\zeta_i) = \sigma^2$.

Q2 (20 pts) Let $Y_i = \beta_0 + \beta_1 x_{i1} + \cdots \beta_{p-1} x_{i,p-1} + \epsilon_i$, $i = 1, 2, \ldots, n$, where $\epsilon_i \overset{iid}{\sim} N(0, \sigma^2)$ for some $p > 1$. Prove that the F-statistic for testing the hypothesis $H_0 : \beta_q = \beta_{q+1} = \cdots = \beta_{p-1} = 0$ $(0 < q \leq p-1)$ is unchanged if a constant, say c, is subtracted from each Y_i.
(Hint: You may start from the special case of $q = 1$ and $p = 2$.)

Q3 (30 pts) Consider the regression model

$$Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \epsilon_i, i = 1, \ldots, n,$$

where ϵ_i's are independent with $\mathrm{E}(\epsilon_i) = 0$ and $\mathrm{Var}(\epsilon_i) = \sigma^2$, and are independent of the covariates x_{1i} and x_{2i}. A common method to consider when at least one of Y_i, x_{1i}, and x_{2i} is missing is the complete case method, which excludes incomplete observations in the analysis. Define the indicate variable Q_i such that $Q_i = 1$ if the ith row is complete (Y_i, x_{1i}, and x_{2i} are fully observed) and $Q_i = 0$ otherwise. Then the complete-data model is given by

$$Q_i Y_i = Q_i \beta_0 + Q_i \beta_1 x_{1i} + Q_i \beta_2 x_{2i} + Q_i \epsilon_i. \tag{1}$$

(a) (10 pts) Find the least squares estimator $\widehat{\boldsymbol{\beta}} = (\widehat{\beta}_0, \widehat{\beta}_1, \widehat{\beta}_2)^\top$ for $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2)^\top$ under Model (1). (Hint: Express Model (1) in matrix form: $\mathbf{QY} = \mathbf{QX}\boldsymbol{\beta} + \mathbf{Q}\boldsymbol{\epsilon}$, where $\boldsymbol{Q} = \mathrm{diag}(Q_1, \ldots, Q_n)$ is a $n \times n$ diagonal matrix.)

(b) (20 pts) Assume that ϵ_i and Q_i are independent,. Find the conditional expectations of $\widehat{\boldsymbol{\beta}}$ and the corresponding residual sum of squares given $\{Q_1, \ldots, Q_n, x_{11}, \ldots, x_{1n}, x_{21}, \ldots, x_{2n}\}$.
(Hint: The residual sum of squares is defined as

$$RSS = \sum_{i=1}^{n} Q_i (Y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_{1i} - \widehat{\beta}_2 x_{2i})^2.$$

$\{Q_1, \ldots, Q_n, x_{11}, \ldots, x_{1n}, x_{21}, \ldots, x_{2n}\}$ might be treated as constants in conditional expectations.)

Q4 (40 pts) Consider a linear regression model

$$Y_{ij} = \alpha_i + \beta x_{ij} + \epsilon_{ij},$$

where $\epsilon_{ij} \overset{iid}{\sim} N(0, \sigma^2)$ for $i = 1, 2$ and $j = 1, \ldots, n_i$. Denote the least squares estimator $\widehat{\boldsymbol{\theta}} = (\widehat{\alpha}_1, \widehat{\alpha}_2, \widehat{\beta})^\top$ for $\boldsymbol{\theta} = (\alpha_1, \alpha_2, \beta)^\top$ with the residual sum of squares

$$RSS = \sum_{i=1}^{2} \sum_{j=1}^{n_i} (Y_{ij} - \widehat{\alpha}_i - \widehat{\beta} x_{ij})^2.$$

Due to budget limitations, only half of the samples were sampled for the second group ($i = 2$), so the regression model is re-expressed as follow,

$$S_{ij} Y_{ij} = S_{ij} \alpha_i + S_{ij} \beta x_{ij} + S_{ij} \epsilon_{ij},$$

where $S_{1j} = 1$ for $j = 1, \ldots, n_1$, $S_{2j} = 1$ for $j = 1, \ldots n_2/2$, $S_{2j} = 0$ for $j = n_2/2 + 1, \ldots n_2$, and n_2 is an even number. The values of S_{2j} are predetermined, fixed, and independent of x_{2j} and ϵ_{2j}. The corresponding weighted least squares objective function has the following form

$$\sum_{i=1}^{2} \sum_{j=1}^{n_i} S_{ij} w_i (Y_{ij} - \alpha_i - \beta x_{ij})^2, \tag{2}$$

where $w_1 = 1$ and $w_2 = 2$ are weights that reflect the different sampling mechanisms. Let $\widehat{\boldsymbol{\theta}}^* = (\widehat{\alpha}_1^*, \widehat{\alpha}_2^*, \widehat{\beta}^*)^\top$ be the weighted least square estimator of $\boldsymbol{\theta}$ that minimizes (2).

(a) (10 pts) Find the weighted least square estimator $\widehat{\boldsymbol{\theta}}^* = (\widehat{\alpha}_1^*, \widehat{\alpha}_2^*, \widehat{\beta}^*)^\top$.

(b) (20 pts) Construct a simultaneous 95% confidence interval for the values $Y_{10}^\dagger = \alpha_1 + \beta x_{10} + \epsilon_1^\dagger$ and $Y_{20}^\dagger = \alpha_2 + \beta x_{20} + \epsilon_2^\dagger$ where x_{10} and x_{20} are fixed numbers, and $\epsilon_i^\dagger \overset{iid}{\sim} N(0, \sigma^2)$ for $i = 1, 2$.

(c) (10 pts) The residual sum of square from $\widehat{\boldsymbol{\theta}}^*$ is then denoted by

$$RSS^* = \sum_{i=1}^{2} \sum_{j=1}^{n_i} S_{ij} w_i (Y_{ij} - \widehat{\alpha}_i^* - \widehat{\beta}^* x_{ij})^2.$$

Argue if there is a $w_2 \geq 0$ that can satisfy $E(RSS) = E(RSS^*)$ when $w_1 = 1$.

Qualifying Exam January 2022 — Statistical Methods

Instructions

- NOTE: You are not allowed to use internet or email except for downloading the data and emailing the final report (optional). To use these for any other purpose, ask the proctor.
- Go to https://github.com/minchen01/QE to download datasets. Above the list of files, using the "Code" drop-down on the right side, click "Download ZIP". 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 comments explaining each step) and outputs (ONLY relevant parts; highlighted wherever possible). Do not attach the parts of the output that were not used in answering questions.
- Submit a report (written or typed), hard copy or by **email to Angie.Bustamante@utdallas.edu**. If you choose to email, attach only one single PDF file with the whole report. DO NOT email 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 other than Angie.**

Problems

1. Work on the diabetes data in a csv file "diabetes.csv". A population of 532 women in Richardson, Texas were tested for diabetes. Some information was collected for these women at the time of testing (see the table below). [35 points]

	Variable name	Description
1	`npreg`	Number of pregnancies
2	`glu`	Glucose level in mg/dL
3	`bp`	Blood pressure in mmHg
4	`bmi`	Body mass index in kg/m^2
5	`ped`	Diabetes pedigree function, which scores likelihood of diabetes based on family history
6	`age`	Age in year
7	`diabetes`	Testing outcome

(a) Make both numerical and graphical summaries of the data. Comment on anything that you find interesting. [6 points]

(b) Model the conditional distribution of glucose level (`glu`) as a linear combination of the other variables, excluding the variable `diabetes`. Perform regression diagnostics on this model and check for multicollinearity. [8 points]

(c) Following (b), investigate the utility of a transformation on the response variable `glu`. Does it improve the model? [3 points]

(d) Consider a logistic regression model for predicting diabetes as a function of the other variables, excluding the variable `glu`,

$$\begin{aligned} \Pr(\texttt{diabetes} = \text{Yes}) &= \exp(\theta_i)/(1+\exp(\theta_i)) \\ \theta_i &= \beta_0 + \gamma_1\beta_1\texttt{npreg} + \gamma_2\beta_2\texttt{bp} + \gamma_3\beta_3\texttt{bmi} + \gamma_4\beta_4\texttt{ped} + \gamma_5\beta_5\texttt{age}. \end{aligned}$$

In this model, each γ_j is either 0 or 1, indicating whether or not variable j is a predictor of diabetes. For example, if it were the case that $\boldsymbol{\gamma} = c(1,1,0,0,0)$, then $\theta_i = \beta_0 + \beta_1\texttt{npreg} + \beta_2\texttt{bp}$.

Note that there are $2^5 - 1 = 31$ sub-models available (excluding the null model represented by $\boldsymbol{\gamma} = c(0,0,0,0,0)$). Select the best sub-model (i.e. find the best $\boldsymbol{\gamma}$) using an appropriate criterion for model selection. Among the models determined by the above criterion, report the best sub-model in terms of $\boldsymbol{\gamma}$ and explain why it is the best sub-model. [12 points]

(e) Following (d), if the criterion is to select the best sub-model (i.e. find the best $\boldsymbol{\gamma}$) based on the prediction performance, what would you do? Just briefly discuss your strategy. You do not have to write code to do the actual model selection. [6 points]

2. Consider the diabetes data in Problem 1 under a simple logistic regression model for predicting diabetes status using one predictor, the pedigree function (`ped`). [30 points]

   (a) Report an estimate of the odds ratio θ, its large-sample SE and the corresponding Wald 95% confidence interval (CI). Interpret these numbers. [5 points]

   (b) Compute bootstrap (nonparametric) estimates of bias and standard error of $\hat{\theta}$, and 95% CI for θ using (i) Studentized method, (ii) basic bootstrap, and (iii) percentile method. In (i), you may use the same large-sample SE formula that you used in the previous part. Set the starting seed as 1111 and obtain 1000 bootstrap replications. Compare these three CIs. Also, compare them with the CI obtained in the previous part. [25 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).

3. Pervasive developmental disorders (`PDD`) in children are characterized by delays in the development of socialization and communication skills. The file "pdd.csv" stores data of 158 PDD children identified by the variable `ID`. The response variable `SCORE`, assessed for each child at ages 2, 3, 5, 9, and 13 years (`AGE`), was a numerical measure of the socialization skills. Initial language development (`LANG_DEV`) of each child was assessed when they were recruited and children were placed into one of three groups based on their initial language skills. The objective was to investigate the influence of the initial language proficiency (`LANG_DEV`) on the developmental trajectories of the socialization (`SCORE)` of children with PDD. Note that some children were not measured at all ages, and there are missing values for `SCORE`. Use $\alpha = 0.05$ for all hypothesis tests below.

   (a) Draw plots to explore the effects of `LANG_DEV` and `AGE`. Provide a summary of your observations and findings. [5 points]

   (b) Assume scores at different ages are independent. Treat `AGE` as a categorical variable and conduct an 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) Does language proficiency have an effect on socialization skills? Based on your model in part (b), conduct an appropriate test to answer this question. Also construct 95% confidence intervals for all pairwise differences of the levels of `LANG_DEV`. Carefully interpret your findings. [10 points]

   (d) Is it appropriate to assume that the scores of the same patient at different ages are independent, as you did in part (b)? Fit a model that has subjects effects in addition to `LANG_DEV` and `AGE`. Test whether or not the scores are independent at different ages . Then conduct a test on the main effect of `LANG_DEV` under this updated model. [10 points]

STATISTICS Ph.D. QUALIFYING EXAM
STATISTICAL INFERENCE

January 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. (25 points) Suppose the cell count random vector $Y = (Y_1, Y_2, Y_3, Y_4, Y_5)$ follows a multinomial distribution with parameters n and cell probability vector $\pi_\theta = (1/2, \theta/4, (1-\theta)/4, (1-\theta)/4, \theta/4)$, were $\theta \in (0,1)$ is an unknown parameter. Recall that a cell count vector $Y = (Y_1, \ldots, Y_k)$ follows a multinomial distribution with parameters n and cell probability vector $\pi = (\pi_1, \ldots, \pi_k)$ if its probability mass function is

$$p(y_1, \ldots, y_k) = \frac{n!}{\prod_{i=1}^k y_i!} \prod_{i=1}^k \pi_i^{y_i},$$

where $\sum_{i=1}^k \pi_i = 1$ and $\sum_{i=1}^k y_i = n$.

   (a) (5 points) Find the MLE of θ assuming that Y is observed.

   (b) (8 points) Find the conditional distribution of $Y_2|(Y_1 + Y_2 = m, Y_3 = y_3, Y_4 = y_4, Y_5 = y_5)$, where $m + y_3 + y_4 + y_5 = n$.

   (c) (12 points) Suppose instead of Y, the observed data is $(Y_1 + Y_2, Y_3, Y_4, Y_5)$. Thus, we get to observe Y_3, Y_4, Y_5, and the sum $Y_1 + Y_2$, but not Y_1 and Y_2 separately. Take Y_2 as *missing data* so that $(Y_1, \ldots, Y_5)$ is *complete data* and 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

$$\theta_{\text{new}} = \frac{y_{2,\text{old}} + y_5}{y_{2,\text{old}} + y_3 + y_4 + y_5}, \text{ where } y_{2,\text{old}} = (y_1 + y_2)\frac{(\theta_{\text{old}}/4)}{1/2 + \theta_{\text{old}}/4}.$$

2. (25 points) Let $X_1, \ldots, X_n$ be a random sample from a $N(\theta, a\theta^2)$ distribution, where $a > 0$ and $-\infty < \theta < \infty$ are unknown parameters. (Note: For this problem, there is no need to verify second-order conditions for maxima.)

   (a) (8 points) Find MLEs of a and θ.

   (b) (7 points) Find the MLE of θ assuming $a = a_0$ is known.

   (c) (5 points) Find the level α likelihood ratio test of $H_0 : a = a_0$ versus $H_1 : a \neq a_0$.

   (d) (5 points) Find the $1 - \alpha$ confidence set for a by inverting the test in (c).

3. (20 points) Let $X_1, \ldots, X_n$ denote a random sample of size n 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) (8 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) (6 points) Obtain a $100(1-\alpha)\%$ confidence interval for θ based on each pivotal quantity in (a).

(c) (6 points) Explain which of the two confidence intervals in (b) you would recommend. Justify your answer.

4. (30 points) For a given $-\infty < \mu < \infty$ and $\sigma^2 > 0$, let $X = (X_1, \ldots, X_n)$ represent a random sample from a $N(\mu, \sigma^2)$ distribution. Also let $\overline{X} = n^{-1}\sum_i X_i$ and $S^2 = (n-1)^{-1}\sum_i (X_i - \overline{X})^2$ be the sample mean and the sample variance. Assume that the prior distributions for μ and $\log\sigma$ are independent uniforms on $(-\infty, \infty)$, or equivalently, the joint prior probability density of (μ, σ^2) is,

$$\pi(\mu, \sigma^2) \propto (\sigma^2)^{-1}.$$

(a) (5 points) Show that the joint posterior density of (μ, σ^2) is,

$$\pi(\mu, \sigma^2|x) \propto (\sigma^2)^{-(n+2)/2} \exp\left[-\frac{1}{2\sigma^2}\{(n-1)s^2 + n(\overline{x} - \mu)^2\}\right].$$

(b) (5 points) Derive $\pi(\mu|x)$, i.e., the marginal posterior distribution of μ.

(c) (5 points) Show that the marginal posterior distribution of $(s/\sqrt{n})^{-1}(\mu - \overline{x})$ is a t-distribution with $(n-1)$ degrees of freedom. Recall that for $\nu = 1, 2, \ldots$, the pdf of a t_ν distribution is

$$f(y) = \frac{\Gamma((\nu+1)/2)}{\Gamma(\nu/2)\sqrt{\nu\pi}}\left(1 + \frac{y^2}{\nu}\right)^{-(\nu+1)/2}, \quad -\infty < y < \infty.$$

(d) (3 points) Derive an HPD interval for μ that has $(1-\alpha)$ posterior probability.

(e) (4 points) Consider testing the hypotheses: $H_0 : \mu \le 0$ versus $H_1 : \mu > 0$. Find posterior probability that the null hypothesis is true, $P(\mu \le 0|x)$.

(f) (4 points) Compare your answer in (d) with the usual classical two-sided $(1-\alpha)$ level *confidence interval* for μ based on t-distribution. What do you conclude?

(g) (4 points) Compare your answer in (e) with the *p-value* for the usual classical t-test of H_0 versus H_1. What do you conclude? [5 points]

PROBABILITY Ph.D. QUALIFYING EXAM

January 2022

General Instructions: Write your QE 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 Lebesgue integral. Hint: There are three steps in the definition. Then present an example of a function for which Riemann integral does not exist but Lebesgue does.

Problem 2. (5 points) Definition of a measure on a sigma field which is absolutely continuous with respect to another measure. Then formulate the Radon-Nikodim Theorem. Hint: Accurately introduce and define a measure space.

Problem 3. (10 points) Prove that X_n converges to X almost sure if and only if for any $\delta > 0$

$$P(\cup_{k=n}^{\infty}|X_k - X| \geq \delta) = o_n(1).$$

Problem 4. (10 points) Consider a probability space $(\Omega, \mathcal{F}, P)$. Let X be a random variable defined on this space and let $B \in \mathcal{F}$ be a fixed event. Present measure-theoretical definitions of X and of the conditional probability $P(B|X = x)$.

Problem 5. (5 points) Consider a probability space $(\Omega, \mathcal{F}, P)$. Suppose that X and Y are random variables with a joint density $f^{X,Y}(x, y)$. Suppose that $E\{Y\}$ exists. Write down an expression for $E\{Y|X = x\}$ and prove your assertion using measure-theoretical definition of conditional expectation.

Problem 6. (15 points) Let $\{X_n, \mathcal{F}_n\}$ be a submartingale. Suppose that

$\sup_n E\{X_n^+\} < \infty$. Prove that there is an integrable random variable X_∞ such that $X_n \to X_\infty$ almost everywhere. Hint: This is the famous submartingale convergence theorem. Give all definitions, and you may use known results, like “Optimal Skipping Theorem” and ”Upcrossing Theorem” without proof, only correctly formulate them.

Problem 7. (20 points) Formulate and prove the Lindeberg’s Central Limit Theorem.

Problem 8. (5 points) Definition of a stable distribution. Present two famous examples. Recall the general form of the symmetric stable characteristic function.

Problem 9. (10 points) Let $Z_0, Z_1, \ldots$ be iid standard normal random variables. Let $\{\varphi_j(x), j = 0, 1, \ldots\}$ be an orthonormal basis on $[0, 1]$. Define a stochastic process

$$Y(t) = \sum_{j=1}^{\infty} Z_j \int_0^t \varphi_j(x)dx, \ t \in [0, 1].$$

Prove that for any $0 \leq t_1 < t_2 < t_3 \leq 1$ the two variables $Y(t_2) - Y(t_1)$ and $Y(t_3) - Y(t_2)$ are uncorrelated.

Problem 10. (15 points) Consider a Brownian motion B_t and prove that almost sure

$$\lim_{t\to\infty} \sup \frac{B_t}{\sqrt{2t \log(\log(t))}} = 1.$$