Qualifying Exam April 2018 — 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 http://www.utdallas.edu/∼swati.biswas/QE and download two datasets cars0.csv and hyper.csv. Let the proctor know if you have any problems with this step.
- You can use `R` or SAS. 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, then 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 anyone other than Angie.**

1. Use the data in the file cars0.csv. Note that the first column is not data but instead gives model names. Treat cylinders as a categorical variable. Also, treat origin as a categorical variable with levels: 1 = US, 2 = Europe, 3 = Asia.

   (a) Fit an appropriate model to predict mpg based on the other variables. Since displacement, horsepower, and acceleration are engine attributes related to number of cylinders, include interactions between cylinders and each of those three variables in your model. [10 points]

   (b) Verify assumptions of your model. If any assumption is not satisfied, try to rectify it. [8 points]

   (c) Reduce the model, if possible, by removing unimportant predictor variables using the BIC criterion. Summarize the marginal effects of the remaining variables on mpg. [6 points]

   (d) Suppose you are interested in a car from this time period that has the following attributes: cylinders = 6, displacement = 240, horsepower = 120, weight = 2800, acceleration = 16.5, year = 80, origin = US. Construct a 95% confidence interval for the mean mpg of cars with those attributes. Construct a 95% prediction interval for the mpg of a particular car with those attributes. [6 points]

2. Conduct a Monte Carlo simulation study with 1,000 replications for comparing two tests of association based on Pearson product moment correlation ρ and Spearman's rank correlation coefficient ρ_s (both are implemented in `cor.test` in `R` and in PROC CORR in SAS). Make comparisons of (a) power and (b) type I error rates at 5% nominal level for two-sided tests under two distributional settings, when the data are sampled from (1) bivariate normal

and (2) bivariate central t with degrees of freedom (ν) = 3. You may set means to be 0, standard deviations to be 1, and correlation to be 0.8 wherever needed. Use sample size of 10. For sampling from bivariate normal and bivariate t, the following may be used: (i) In `R`, commands `rmvnorm` and `rmvt` from the package `mvtnorm` (ii) In SAS, PROC MODEL with SOLVE and ERRORMODEL statements. You may consult the help pages of these commands. Note that `rmvt` command requires the scale matrix Σ, which is NOT equal to the covariance matrix Cov, however, $Cov = (\nu/(\nu-2))\Sigma$ for $\nu > 2$. Compare the power and type I error rates of the two tests under the two distributional settings. In particular, comment on which test is preferable under which setting and why. [35 points]

3. "Hypertension data": This data set was from a study in which the effects of three possible treatments for hypertension were investigated. The details of the treatments are as follows:

#	Variable	Description
1	Drug	Medication: Drug X, Y, Z
2	Biofeed	Psychological feedback: Present (P), absent (A)
3	Diet	Special diet: Yes (Y), No (N)
4	bp	Blood pressure: continuous

Researchers are interested in finding which treatment(s) or treatment combination(s) are effective in treating the hypertension condition.

(a) Use graphical tools to check all possible two-way (i.e. first-order) interactions. Summarize your observations. [5 points]

(b) Visually check the three-way (i.e. second-order) interactions and summarize your results. [5 points]

(c) Write down the factor effect model of a three-way ANOVA. Check the model assumptions. [15 points]

(d) Conduct tests for the main effects and test whether any interaction exists (both using $\alpha = .05$). What are the appropriate conclusions on main effects? Is it meaningful to draw conclusions on two-way interactions? Explain your reasoning. [10 points]

STATISTICS Ph.D. QUALIFYING EXAM
STATISTICAL INFERENCE
April 2018

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. Formulate and prove Basu's Theorem. Please define all used notions.

Problem 2 Consider a sample of size n from Uniform$(-\theta, \theta)$ distribution. Find a minimal sufficient statistic and prove your assertion. Also, what are other (if any) nice properties of this statistic (like completeness) ?

Problem 3. Formulate all methods, that you know, of finding a UMVU estimate. Begin your answer with the definition of a UMVU estimate.

Problem 4 Consider a sample of size n from an exponential random variable with parameter (mean) $\lambda \geq 3$. Find the MLE estimate and then calculate its MSE (Mean Squared Error).

Problem 5 Consider a sample of size n from a Uniform$(0, \theta)$ random variable. Find the MLE estimate of θ^2, and then either calculate its MSE (Mean Squared Error) or explain why you cannot do that.

Problem 6 Suppose that a statistician believes that it is reasonable to use a Bayes approach for estimation of the variance of a Poisson distribution. The statistician feels good about using a Gamma(α, β) distribution as the prior for the variance.
a. Is that prior appropriate? Prove your assertion.
b. Calculate the Bayes estimator and compare it with the classical MLE estimator.
c. Suppose that it is known that the variance of the Poisson variable is not greater than a known constant σ_0^2. In this case, how would you modify the Bayesian approach to get a reasonable estimator?

Problem 7. Consider a sample of size n from a Bernoulli random variable with θ being the probability of success. Find the UMVU estimate of $\theta(1-\theta)$, and do not forget to explain why this is the UMVU estimate.

Problem 8. Prove the Neyman-Pearson Lemma about the most powerful test of a simple hypothesis versus another simple hypothesis.

Problem 9. Consider samples of sizes n_1 and n_2 from independent Normal random variables with variances σ_1^2 and σ_2^2. Consider testing the null hypothesis $\sigma_1^2/\sigma_2^2 = a$ versus the alternative hypothesis $\sigma_1^2/\sigma_2^2 > a$. Propose a test, explain your choice, and then write down an expression for the power function.

Problem 10. Consider a sample of size n from a Bernoulli random variable with mean θ. Suppose that you have a large sample size, say $n = 36$, and that you are interested in finding a one-sided confidence interval $[L, \infty)$ that covers the unknown θ with the probability $1 - \alpha$. Propose such a confidence interval and then justify your choice using any classical statistical approache.

QE ID (given by Angie)

April 2018 Qualifying Exam in Probability Theory

- This is a closed-book test.
- There are 4 questions; some have 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 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.*
- Total points = 100.

EXERCISE 1 *(20 points)*
Let $X_1, X_2, ...$ be a sequence of independent and identically distributed random variables with uniform distribution on $[0, 1]$. Prove that

$$\lim_{n\to\infty} (X_1 \cdot X_2 \cdot \ldots \cdot X_n)^{\frac{1}{n}}$$

exists with probability one and compute its value.

EXERCISE 2 *(30 points)*

(a) Consider a sequence of random variables $X_1, X_2, \ldots$ *such that* $X_n = 1$ *or* 0. *Assume* $P[X_1 = 1] \geq \alpha$ *and* $P[X_n = 1 | X_1, \ldots, X_{n-1}] \geq \alpha > 0$ *for n=2,3,....* *Show that*

1. $P[X_n = 1$ *for some* $n] = 1$.
2. $P[X_n = 1$ *infinitely often*$] = 1$.

(b) 1. *Let* $X_1, X_2, \ldots$ *be independent and identically distributed random variables with* $E[X_i] = 0$ *and* $E(X_i^2) = 1$ *for all i. Show that* $P(X_n > n$ *infinitely often*$) = 0$.

EXERCISE 3 *(30 points)* *Let ξ_n be independent and identically distributed random variables with $P(\xi_n = -1) = P(\xi_n = 1) = 1/2$.*

1. *Prove that the series $\sum_{n=1}^{\infty} e^{-n}\xi_n$ converges with probability one.*

2. *Prove that the distribution of $\xi = \sum_{n=1}^{\infty} e^{-n}\xi_n$ is singular, i.e., concentrated on a set of Lebesgue measure zero.*

EXERCISE 4 *(20 points) Let ξ_n be a sequence of independent random variables with ξ_n uniformly distributed on $[0, n^2]$. Find a_n and b_n such that $(\sum_{i=1}^{n} \xi_i - a_n)/b_n$ converges in distribution to a nondegenerate limit and identify the limit.*

QE ID (given by Angie)

April 2018 Qualifying Exam in Linear Models

- This is a closed-book, closed-notes test.
- There are 3 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 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 are 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.*
- Total points = 100.

EXERCISE 1 *(30 points) Consider the linear model* $Y_{ij} = \mu_i + \beta_1 z_{ij} + \beta_2 w_{ij} + \epsilon_{ij}$, $i = 1, \ldots, I$, $j = 1, \ldots, J$, *where the* ϵ_{ij} *are independently distributed as* $\mathcal{N}(0, \sigma^2)$ *random variables.*

1. *(15 points) Derive the least squares estimators of* β_1 *and* β_2, *say,* $\hat{\beta}_1$ *and* $\hat{\beta}_2$, *respectively.*

2. *(10 points) Derive the covariance matrix of the estimators* $\hat{\beta}_1$ *and* $\hat{\beta}_2$.

3. *(5 points) Under what conditions* $\hat{\beta}_1$ *and* $\hat{\beta}_2$ *are statistically independent?*

Important note: Simplify the expressions for the estimators and their covariance matrix as much as possible, otherwise no credit may be given. Ignorable hint: 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 $\mathbf{B}_{22} = \mathbf{A}_{22} - \mathbf{A}_{21}\mathbf{A}_{11}^{-1}\mathbf{A}_{12}$, $\mathbf{B}_{12} = \mathbf{A}_{11}^{-1}\mathbf{A}_{12}$, *and* $\mathbf{B}_{21} = \mathbf{A}_{21}\mathbf{A}_{11}^{-1}$.

EXERCISE 2 *(35 points) Consider the linear model* $Y_{ij} = \mu_i + \epsilon_{ij}$, $i = 1, \ldots, I$, $j = 1, \ldots, J$, *where the* ϵ_{ij} *are independently distributed as* $\mathcal{N}(0, \sigma^2)$ *random variables.*

1. *(15 points) Suppose* $I = 4$. *Derive the* F*-statistic for testing the null hypothesis that* $\mu_1 = \mu_2 = \mu_3$. *(Note: Simplify the expression for the statistic as much as possible, otherwise no credit may be given.)*

2. *(20 points) Suppose* $I = 2$. *Derive the* F*-statistic for testing the null hypothesis that* $\mu_1 = \mu_2$. *Show that this statistic is equal to the square of the usual* t*-statistic for testing the null hypothesis that the means of two normally distributed populations are equal assuming that their variances are equal but unknown.*

EXERCISE 3 *(35 points) Consider the usual linear model* $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, *where* $\mathbf{Y}$ *is a* $n \times 1$ *vector,* $\mathbf{X}$ *is a* $n \times p$ *matrix of full rank,* $\boldsymbol{\beta}$ *is a* $p \times 1$ *vector of regression coefficients, and* $\boldsymbol{\epsilon}$ *is a* $n \times 1$ *vector of random errors with* $E(\boldsymbol{\epsilon}) = \mathbf{0}$, $var(\boldsymbol{\epsilon}) = \sigma^2\mathbf{I}_n$. *Here* $\mathbf{I}_n$ *denotes a* $n \times n$ *identity matrix. Let* $\|\mathbf{a}\| = \sqrt{\mathbf{a}^T\mathbf{a}}$ *denote the Euclidean norm of a column vector* $\mathbf{a}$. *Next, let* $\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$ *denote the least squares estimator of* $\boldsymbol{\beta}$, *and also let* $(\lambda_1, \mathbf{P}_1), \ldots, (\lambda_p, \mathbf{P}_p)$ *denote the eigenvalue-eigenvector pairs of* $\mathbf{X}^T\mathbf{X}$ *that are obtained via its spectral decomposition. Consider the following estimator of* $\boldsymbol{\beta}$, *which is known as the* ridge estimator *of* $\boldsymbol{\beta}$:

$$\hat{\boldsymbol{\beta}}(k) = (\mathbf{X}^T\mathbf{X} + k\mathbf{I}_p)^{-1}\mathbf{X}^T\mathbf{Y},$$

where k *is a specified non-negative constant. When* $k = 0$, *the ridge estimator reduces to the least squares estimator* $\hat{\boldsymbol{\beta}}$.

1. *(5 points) Show that the ridge estimator can be written as* $\hat{\boldsymbol{\beta}}(k) = \mathbf{C}\hat{\boldsymbol{\beta}}$, *where* $\mathbf{C} = [\mathbf{I}_p + k(\mathbf{X}^T\mathbf{X})^{-1}]^{-1}$.

2. *(2 points) Deduce whether or not* $\hat{\boldsymbol{\beta}}(k)$ *is unbiased for* $\boldsymbol{\beta}$.

3. *(5 points) Show that the mean squared error (MSE) of* $\hat{\boldsymbol{\beta}}(k)$*, defined as* $E(\|\hat{\boldsymbol{\beta}}(k) - \boldsymbol{\beta}\|^2)$*, can be written as:*

$$MSE = E(\|\mathbf{C}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})\|^2) + \|(\mathbf{C} - \mathbf{I}_p)\boldsymbol{\beta}\|^2. \tag{1}$$

4. *(4 points) Interpret the two terms on the right hand side in equation* (1).

5. *(10 points) Show that the first term on the right hand side in equation* (1) *can be written as* $\sigma^2 \sum_{j=1}^{p} \{\lambda_j/(k+\lambda_j)^2\}$.

6. *(7 points) Show that the second term on the right hand side in equation* (1) *can be written as* $\sum_{j=1}^{p} \{\alpha_j^2 k^2/(k+\lambda_j)^2\}$*, where* $\alpha_j = \mathbf{P}_j^T \boldsymbol{\beta}$*,* $j = 1, \ldots, p$.

7. *(2 points) Obtain the resulting expression for the MSE of* $\hat{\boldsymbol{\beta}}(k)$ *given by equation* (1).