STATISTICS Ph.D. QUALIFYING EXAM
STATISTICAL INFERENCE

January 2019

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. Let $X_1, \ldots, X_n$ be a random sample from a location family. Show that $M - \bar{X}$ is an ancillary statistic. Here M is the sample median and $\bar{X}$ is the sample mean.

Problem 2. You have a sample from a distribution with density $f^{X|\theta}(x|\theta) = \theta x^{\theta-1} I(0 < x < 1) I(\theta > 0)$. Find a complete sufficient statistic for θ or show that it does not exist.

Problem 3. Let X be Normal with zero mean and variance σ^2. Is $|X|$ a sufficient statistic? Prove or disprove that assertion.

Problem 4. Consider a sample with the density

$$f^{X|\theta}(x|\theta) = \frac{\theta}{(1+x)^{1+\theta}} I(0 < x < \infty) I(\theta > 0).$$

Find a complete sufficient statistic or show that one does not exists.

Problem 5. Consider a sample of size n from the density $f^{X|\theta}(x|\theta) = 2\theta^2 x^{-3} I(0 < \theta \le x < \infty)$. Find the MLE estimator.

Problem 6. Consider a sample of size n from Normal(μ, σ^2) distribution with unknown $\mu \in (-\infty, \infty)$ and a known $\sigma^2 > 0$. Find the UMVUE of μ^3.

Problem 7. A random sample of size n is from a Pareto distribution with the pdf

$$f(x|\theta,\nu) = \theta\nu^{\theta}x^{-\theta-1}I(x \geq \nu > 0)I(\theta > 0).$$

Find the likelihood ratio test of $H_0 : \{\theta = 1$ and ν is unknown$\}$ versus $H_a : \{\theta \neq 1$ and ν is unknown $\}$.

Problem 8. Derive a confidence interval for a binomial probability of success θ by inverting an appropriate LRT test.

Problem 9. Formulate and prove Basu's Theorem. Please define all used notions.

Problem 10. Consider a sample from a distribution family $\{P_\theta, \theta \in \Omega\}$. Formulate all methods, that you know, of finding a UMVU estimate of $g(\theta)$ where $g(\cdot)$ is a given function. Begin your answer with the definition of a UMVU estimate.

____

QE ID (given by Angie)

January 2019 Qualifying Exam in Linear Models

- This is a closed-book test.
- There are 3 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.

Question 1 *(30 points) Consider the linear regression model with one covariate: $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$, $i = 1, \ldots, n$, where the ϵ_i are independent $N(0, \sigma^2/w_i)$ random variables for some fixed constant σ and known positive numbers w_i.*

i. *(15 points) Derive the weighted least squares (WLS) estimators of β_0 and β_1, say, $\hat{\beta}_0$ and $\hat{\beta}_1$, respectively, that minimize $\sum_{i=1}^n w_i(Y_i - \beta_0 - \beta_1 X_i)^2$ with respect to β_0 and β_1. Simplify the expressions for the estimators as much as possible.*

ii. *(7 points) Find the sampling distribution of the WLS estimator $\hat{\beta}_1$.*

iii. *(8 points) Construct a $100(1-\alpha)\%$ confidence interval for β_1 based on $\hat{\beta}_1$. Also provide a level-α test for $H_0 : \beta_1 = 0$ against $H_1 : \beta_1 \neq 0$.*

Question 2 *(30 points) Let*

$$Y_i = \theta_i + \epsilon_i, \; i = 1, 2, 3, 4,$$

where $\theta_1 + \theta_2 + \theta_3 + \theta_4 = 0$ and the ϵ_i are independent $N(0, \sigma^2)$ random variables. We would like to perform an F-test for testing $H_0 : \theta_1 = \theta_3$ against $H_1 : \theta_1 \neq \theta_3$.

i. *(20 points) Obtain an explicit expression for the F-statistic. Simplify the expression as much as possible.*

ii. *(5 points) What is the null distribution of the test statistic?*

iii. *(5 points) Provide the rejection region for a level-α F-test of the hypothesis.*

Question 3 *(40 points) Let $\mathbf{Y}$ be a $n \times 1$ vector of observations. It follows the linear model $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, where $\mathbf{X}$ is a $n \times p$ design matrix, $\boldsymbol{\beta}$ is a $p \times 1$ vector of unknown coefficients, and $\boldsymbol{\epsilon}$ is a $n \times 1$ vector of random errors. It is assumed that the model does not have any intercept term and $E(\boldsymbol{\epsilon}) = \mathbf{0}$ and var$(\boldsymbol{\epsilon}) = \sigma^2\mathbf{I}$. The Lasso procedure estimates $\boldsymbol{\beta}$ by minimizing the objective function*

$$\frac{1}{2}(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})^T(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) \text{ subject to } \sum_{j=1}^p |\beta_j| \leq c,$$

where $c \geq 0$ is a fixed tuning parameter. The equivalent Lagrangian form of the objective function is

$$Q(\boldsymbol{\beta}) = \frac{1}{2}(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})^T(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) + \lambda \sum_{j=1}^p |\beta_j|, \tag{1}$$

where $\lambda \geq 0$ is a fixed penalty parameter. The goal of this question is to explicitly derive the Lasso estimator of $\boldsymbol{\beta}$ for a fixed λ in two special cases. In what follows, the function sign(a) is defined through $|a| = a \cdot$ sign(a), where sign$(0) = 0$ by convention. Moreover, the positive truncation function a_+ is defined as

$$a_+ = \begin{cases} a, & \text{if } a > 0, \\ 0, & \text{otherwise.} \end{cases}$$

i. (20 points) Suppose $n = p = 1$ and the predictor $X = 1$. In this case, the objective function (1) *for a fixed λ is*

$$Q(\beta) = \frac{1}{2}(Y - \beta)^2 + \lambda|\beta|.$$

(a) (6 points) For $\beta \neq 0$, show that the Lasso estimator for β is obtained by solving for β in the normal equation

$$-Y + \beta + \lambda \cdot sign(\beta) = 0.$$

(b) (6 points) Argue that the Lasso estimator for β and Y must have the same sign, i.e., $sign(\beta) = sign(Y)$.

(c) (8 points) Show that the Lasso estimator for β is

$$\hat{\beta}(\lambda) = sign(Y) \cdot (|Y| - \lambda)_+.$$

ii. (12 points) Suppose $p = 1$ and $X = 1$ as before, but now we have n observations $Y_1, \ldots, Y_n$. In this case, the objective function (1) *reduces to*

$$Q(\beta) = \frac{1}{2}\sum_{i=1}^{n}(Y_i - \beta)^2 + \lambda|\beta|.$$

Use the solution of the previous part to show that the Lasso estimator of β is given by

$$\hat{\beta}(\lambda) = sign(\bar{Y}) \cdot (|\bar{Y}| - \lambda)_+.$$

iii. (8 points) Answer the following regarding the Lasso estimator $\hat{\beta}(\lambda)$. Justify all your answers.

(a) (2 points) What are its special cases when $\lambda = 0$ and $\lambda \to \infty$? Interpret the resulting estimators.

(b) (2 points) When is it zero? What does a zero estimate mean?

(c) (2 points) Is it unbiased?

(d) (2 points) Is it a linear estimator?

QE ID

QE ID

January 2019 Qualifying Exam in Probability Theory

- This is a closed-book test.
- There are 3 questions.
- 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 use 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.
- Identify each sheet with your name.
- 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 name is on each sheet.*

QE ID QE ID

Exercise 1

Assume $X_i \in \mathcal{L}^4$ have common expectation $E[X_i] = m$ and satisfy $M = \sup_n ||X||_4 < \infty$. If X_i are independent, is it true that

1. $\frac{S_n}{n} \to m$ *in probability?*
2. $\sum_{n=1}^{\infty} P\left[\left|\frac{S_n}{n} - m\right| \geq \epsilon\right]$ *converges for all $\epsilon > 0$?*

Justify your answers.

QE ID QE ID

EXERCISE 2 *A $\mathcal{L}^1$-process which is adapted to a filtration $\{A_n\}$ is called a submartingale if*

$$E[X_n|A_{n-1}] \geq X_{n-1}.$$

1. *If X is a martingale and u is convex such that $u(X_n) \in \mathcal{L}^1$, then show that $Y = u(X)$ is a submartingale.*

2. *If u is monotone and convex and X is a submartingale such that $u(X_n) \in \mathcal{L}^1$, then show that $u(X)$ is a submartingale.*

____________________ QE ID	____________________ QE ID

EXERCISE 3 *Let* $X_1, \ldots, X_n$ *he a random sample from a Poisson distribution with mean* $\lambda > 0$. *Let* $\bar{X}_n$ *be the sample average and let* $T_n = \sqrt{\bar{X}_n}$.

1. *Explain why and in what sense* T_n *converges to* $\sqrt{\lambda}$.
2. *Find the approximate distribution of* T_n *for large* n.

Qualifying Exam January 2019 — 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/~mchen/QE and download a dataset "helmet.txt". 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, 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 any one other than Angie.**

Problems

1. Stock price, $S(t)$, as a function of time t, is a random variable. The instantaneous stock price movements over a very small period dt, denoted by $dS(t)$, is often modeled as a diffusion process:

$$dS(t) = \mu S(t) \cdot dt + \sigma S(t)\epsilon \cdot \sqrt{dt},$$

where the parameters μ and σ^2 are known as the drift and volatility, respectively, and ϵ is a standard normal random variable. We consider the process with discrete times as follows. Let $S(i)$ be the closing stock price on day i for $i = 0, 1, 2, \cdots$. It can be shown that the above time evolution model yields

$$S(i+1) = S(i)\text{exp}\left(\mu - \frac{1}{2}\sigma^2 + \sigma \cdot Z(i+1)\right), \; i = 0, 1, 2, \cdots$$

where $Z(1)$, $Z(2)$, ... are independent standard normal random variables. [30 points]

   (a) Write a program that takes as input $\mu, \sigma^2, S(0)$ and t, to simulate $S(1)$, ..., $S(t)$ and to plot them in a graph. In your answer, include sample plots for at least two values of μ and two values of σ^2. Make sure that there are adequate explanations and descriptive captions for each figure. Write a paragraph that describes qualitatively what happens as μ increases/decreases and σ^2 increases/decreases. [10 points]

   (b) We will estimate $E[S(t)|S(0)]$ and $Cov[S(t-1), S(t)|S(0)]$ using simulation. Write a program, taking $\mu, \sigma^2, S(0)$ and t as its input, to do the following: (1) simulate $S(t-1)$ and $S(t)$ a number of times (at least 10,000); (2) estimate $E[S(t)|S(0)]$; (3) estimate $Cov[S(t-1), S(t)|S(0)]$; and (4) give a 95% confidence interval for each estimate. Use your program to estimate $E[S(100)|S(0) = 1]$ and $Cov[S(99), S(100)|S(0) = 1]$ for the following 2 settings of parameters: ($\mu = 0.05.\sigma^2 = 0.0025$) and ($\mu = 0.01.\sigma^2 = 0.01$). [10 points]

   (c) We would also like to estimate $\mathbb{P}(S(t) > S(0))$, the probability that the value of the stock at time t would exceed its beginning price, using simulation. How will $\mathbb{P}(S(t) > S(0))$ be affected by the value (or distribution) of $S(0)$? Construct a 95% confidence interval of the estimates of $\mathbb{P}(S(100) > S(0))$ under the two parameters given in part (b). [10 points]

2. The `MASS` library in **R** includes a large data frame called `Boston`. The response variable of interest in this data frame is `log(crim)`, log-transformed per capita crime rate. Use $\alpha = .10$ for any hypothesis tests performed. [35 points]

   (a) Fit a linear model to predict `crim` based on the independent variables:
   `medv`, the median value of owner-occupied homes in $1,000
   `chas`, Charles River dummy variable (= 1 if tract bounds river; 0 otherwise). Convert to factor.
   `nox`, nitrogen oxides concentration (parts per 10 million)
   `rm`, average number of rooms per dwelling
   `tax`, full-value property-tax rate per $10,000
   `lstat`, percent of the population classified as lower status.
   Include 2-way interactions between `chas` and each of the other independent variables. [7 points]

   (b) Obtain residual plots, including plots of the residuals versus each of the continuous independent variables and determine whether or not the assumptions for the linear model are reasonable. Perform any transformations that are necessary and refit. [7 points]

   (c) Determine if any of the interactions can be removed from the model. Refit the model after removing any unnecessary interactions. [7 points]

   (d) Interpret the remaining coefficients. [7 points]

   (e) Construct separate 95% confidence intervals for predicted `log(crim)` at the mean of each of the continuous independent variables for tracts that bound the Charles River and for tracts that do not bound the river. [7 points]

3. A manufacturer of football helmets wants to test several redesigned helmets in an effort to improve their protection from traumatic brain injuries. The helmet designs (`DESIGN`) for this study consists of 4 styles: Contemporary I (`DESIGN=1`), Contemporary II (`DESIGN=2`), Traditional I (`DESIGN=3`) and Traditional II (`DESIGN=4`). Each helmet receives 10 tests performed to both the front (`SIDE=1`) and rear (`SIDE=2`) sides. The dependent variable is attenuation of impact forces as measured by the Gadd Severity Index (`GSI`), with higher scores indicating a greater chance for cerebral injury. The data file “helmet.txt” contains 3 columns: `DESIGN`, `SIDE` and `GSI`, in that order. [35 points]

   (a) Consider an ANOVA analysis for this problem. Draw a plot to explore the interaction between `DESIGN` and `SIDE`. Does this plot suggest any interaction? [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 whether or not the interaction effect exists using $\alpha = 0.05$. Carefully interpret your findings. [10 points]

   (d) The manufacturer wants to know whether or not the `DESIGN`s are differing in their effects of impact forces. Construct confidence intervals of all pairwise differences in means of these 4 designs. Use the most efficient multiple comparison procedure with a 95% family-wise confidence level. [10 points]