Department of Statistics
University of Wisconsin, Madison
PhD Qualifying Exam Option B
August 25, 2020
12:30-4:30pm, Room 331 SMI

- There are a total of FOUR (4) problems in this exam. Please do all FOUR (4) problems.
- Each problem must be done in a separate exam book.
- Please turn in FOUR (4) exam books.
- Please write your code name and **NOT** your real name on each exam book.

1. Consider the negative binomial distribution, with probability mass function (pmf) indexed by parameters $\theta > 0$ and $k > 0$ given below:

$$\mathbb{P}\left(X = x \middle| \theta, k\right) = \frac{\Gamma(x+k)}{x!\Gamma(k)} \left(\frac{k}{k+\theta}\right)^k \left(\frac{\theta}{k+\theta}\right)^x, \quad x = 0, 1, \ldots$$

Let $X_1, \ldots, X_n$ be a random sample from this distribution. Please solve the following problems.

(a) Find the method of moments (MOM) estimators of θ and k. State a condition under which the MOM estimator $\widetilde{k}$ takes an allowable value of k.

For the remaining parts of this problem, assume the parameter $k > 0$ is known and $\theta > 0$ is the only parameter of interest.

(b) Show that the maximum likelihood estimator (MLE) of θ is given by $\widehat{\theta} = \frac{1}{n}\sum_{i=1}^n X_i$.

(c) Find the uniformly minimum variance unbiased estimator (UMVUE) of θ.

(d) A Bayesian approach to estimation in this situation is as follows: first, let $\eta = \frac{\theta}{\theta+k}$. Then, assume a Beta($\alpha$, β) prior for η. Using this approach, find the Bayes rule for θ under squared error loss.

Some useful facts:

(i) The *Gamma function* is defined by $\Gamma(\alpha) = \int_0^\infty t^{\alpha-1}e^{-t}dt$ for $\alpha > 0$. The Gamma function obeys the property $\Gamma(\alpha+1) = \alpha\Gamma(\alpha)$.

(ii) The Beta(α, β) distribution has probability density function (pdf) given by

$$f(x|\alpha, \beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}, \quad 0 < x < 1, \quad \alpha > 0, \beta > 0.$$

2. Assume that Z follows the standard normal $N(0,1)$ distribution, L is an integer random variable with the probability distribution $\mathrm{P}(\mathrm{L}=1)=\mathrm{P}(\mathrm{L}=2)=1/2$, and Z and L are independent. Define a random variable $Y=u_{\mathrm{L}}+Z$, i.e.,

$$Y=\begin{cases} u_1+Z, & \text{if L} = 1,\\ u_2+Z, & \text{if L} = 2,\end{cases}$$

where $u_1 \in (-\infty,\infty)$ and $u_2 \in (-\infty,\infty)$ are distinct non-random constants.

(a) Derive forms of the cumulative distributive function $F_Y(y)$ and probability density function $f_Y(y)$ of Y.

(b) Compute $\mathrm{E}(Y)$, $\mathrm{E}(Y^2)$ and $\mathrm{var}(Y)$.

(c) Consider the sampling distribution of $T_n = n^{-1}\sum_{i=1}^n c_i\,\{Y_i - \mathrm{E}(Y)\}$, where $Y_1,\ldots,Y_n \overset{\text{i.i.d.}}{\sim} Y$, with a sequence of known constants $c_1, c_2, \ldots$ such that $\lim_{n\to\infty}(\sum_{i=1}^n c_i^2)/n = a \in [0,\infty)$.

Show that T_n converges to 0 in probability as $n \to \infty$.

(d) For $c_i \equiv 1$, $i=1,\ldots,n$, $n \geq 1$, in part (c), derive the non-degenerate asymptotic distribution of T_n as $n \to \infty$.

3. This problem investigates fixed design linear regressions in high-dimensional settings. Suppose that we observe a sample of n observations, $(\mathbf{x}_i, y_i) \in \mathbb{R}^d \times \mathbb{R}$ for $i = 1, \ldots, n$. Let $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_n]^T \in \mathbb{R}^{n\times d}$ denote the design matrix and $\mathbf{y} = (y_1, \ldots, y_n)^T \in \mathbb{R}^n$ be the response vector. Consider a linear model with i.i.d. mean-zero Gaussian noise

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0},\ \mathbf{I}_{n\times n}), \tag{1}$$

where $\boldsymbol{\beta} \in \mathbb{R}^d$ is the unknown coefficient vector, $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)^T \in \mathbb{R}^n$ is the noise term, and $\mathbf{I}_{n\times n}$ is an n-by-n identity matrix. Many modern applications of this model are high-dimensional, in that the number of features d is comparable to, or even larger than, the sample size n. Assume that $d = n$, and $\mathbf{X}$ has orthonormal columns such that $\mathbf{X}^T\mathbf{X} = \mathbf{I}_{d\times d}$. Consider the following regularized estimator for $\boldsymbol{\beta}$,

$$\widehat{\boldsymbol{\beta}}_\lambda = \arg\min_{\boldsymbol{\beta}} \left\{ \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2 + \lambda\|\boldsymbol{\beta}\|^2 \right\}, \tag{2}$$

where λ is an unknown tuning parameter, $\|\cdot\|$ denotes the vector 2-norm; i.e., $\|\mathbf{a}\| = \left(\sum_{j=1}^d |a_j|^2\right)^{1/2}$ for a vector $\mathbf{a} = (a_1, \ldots, a_d)^T \in \mathbb{R}^d$.

(a) Let $\lambda = 0$.

i. Derive the expression for $\widehat{\boldsymbol{\beta}}_0$ by solving the optimization problem in (2) and find its distribution.

ii. Consider the prediction error for a new observation of the form $y_{\text{new}} = \mathbf{x}_{\text{new}}^T\boldsymbol{\beta} + \varepsilon_{\text{new}}$, where $\mathbf{x}_{\text{new}} \in \mathbb{R}^d$ is a fixed covariate vector and $\varepsilon_{\text{new}} \sim \mathcal{N}(0, 1)$ is a noise term independent of $\{\varepsilon_i\}_{i=1}^n$. Find the expected squared prediction error, $\mathbb{E}(y_{\text{new}} - \mathbf{x}_{\text{new}}^T\widehat{\boldsymbol{\beta}})^2$, for this new observation.

(b) Let $\lambda > 0$.

i. Derive the expression for ridge regression estimator $\widehat{\boldsymbol{\beta}}^{\text{ridge}}$ by solving the optimization problem in (2).

ii. Consider the prediction error for a new observation of the form $y_{\text{new}} = \mathbf{x}_{\text{new}}^T\boldsymbol{\beta} + \varepsilon_{\text{new}}$, where $\mathbf{x}_{\text{new}} \in \mathbb{R}^d$ is a fixed covariate vector and $\varepsilon_{\text{new}} \sim \mathcal{N}(0, 1)$ is a noise term independent of $\{\varepsilon_i\}_{i=1}^n$. Find the expected squared prediction error, $\mathbb{E}(y_{\text{new}} - \mathbf{x}_{\text{new}}^T\widehat{\boldsymbol{\beta}}^{\text{ridge}})^2$, for this new observation.

iii. For this part of the question only, assume that $\|\boldsymbol{\beta}\| = 1$. Derive the optimal λ that minimizes the mean squared error for the ridge estimator, $\mathbb{E}\|\widehat{\boldsymbol{\beta}}^{\text{ridge}} - \boldsymbol{\beta}\|^2$. Discuss how you would find λ in practice when $\|\boldsymbol{\beta}\|$ is unknown.

(c) Suppose that a prior distribution $\boldsymbol{\beta}^{\text{prior}} \sim \mathcal{N}(\mathbf{0}, \sigma^2\Phi)$ is imposed to the model (1), where σ^2 is an unknown variance parameter, and Φ is a known positive definite matrix. Furthermore, assume that $\boldsymbol{\beta}^{\text{prior}}$ and $\boldsymbol{\varepsilon}$ are independent.

i. Find the marginal distribution of $\mathbf{y}$.

ii. Derive the method-of-moments estimator for σ^2 based on the marginal distribution of $\mathbf{y}$.

4. An energy supplement company is developing two energy supplements, in liquid form labeled as A and B, specifically for female athletes. In order to evaluate the effectiveness of these supplements, they recruited 60 female runners aged between 21 and 40. In an experiment they named *run-consume-run*, they let 40 of the athletes run on treadmill until exhaustion. Then, during a three hour rest period, the research team randomly assigned each subject an energy supplement to consume. After the rest period, the subjects ran again until exhaustion and their energy expenditure on this second run period was evaluated on a scale of 0 to 100, with higher scores indicating better performance.

   In a modified version of the experiment named *consume-run-consume-run*, rest of the 60 athletes (the remaining 20) consumed one of the energy supplements prior to the first run and the other prior to the second run, in the rest period. The research team randomized the supplement order for each athlete and prior research indicated that subjects would have scored the same for supplements A and B regardless of the order of the trials performed.

   A dataset is constructed so that subjects 1 through 20 represent the athletes who performed the trial after consuming each of the energy supplements (*consume-run-consume-run*) and subjects 21 through 40 represent athletes who were on the *run-consume-run* regime and consumed the supplement A, whereas 41 through 60 were on the same *run-consume-run* regime with supplement B.

   They considered the following model for this data:

$$y_{ij} = \alpha_i + b_j + e_{ij}, \tag{3}$$

   where $i = A, B$ indexes the supplements and $j = 1, \cdots, 60$ indexes the athletes. Here, y_{ij} is the score for supplement i and subject j, α_i is the unknown mean score for supplement i, b_j is a random effect for athlete j, and e_{ij} is a random error. Furthermore, they assumed that e_{ij} are independent and identically distributed as $\mathcal{N}(0, \sigma_e^2)$ and $b_1, \cdots, b_{60}$ are independent and identically distributed as $\mathcal{N}(0, \sigma_b^2)$, and are independent of e_{ij}'s.

   (a) Let $d_1, \cdots, d_{20}$ represent the difference between scores of supplement A and B for subjects on the *consume-run-consume-run* regime. Derive the distribution of these differences based on the model in equation (3).

   (b) Derive a test statistic as a function of only the differences $d_1, \cdots, d_{20}$ from part (a) to test the null hypothesis that there is no difference between the two supplements. Specify the distribution of this test statistic under the null hypothesis of no difference between the two supplements.

   (c) Let $r_1, \cdots, r_{20}$ be the scores of the athletes who consumed supplement A on the *run-consume-run* regime and, similarly, $q_1, \cdots, q_{20}$ the scores for those who consumed supplement B on the *run-consume-run* regime. Given only these scores, derive a 95% confidence interval for $\alpha_1 - \alpha_2$ as a function of these scores.

   (d) Given both d's from part (a) and r's and q's from part (c), derive unbiased estimators of σ_b^2 and σ_e^2 as a function of these observations.

(e) Suppose you are given sample means of the d's, r's, and q's as $\bar{d} = \sum_{i=1}^{20} d_i/20$, $\bar{r} = \sum_{i=1}^{20} r_i/20$, and $\bar{q} = \sum_{i=1}^{20} q_i/20$, from parts (a) and (c). Assume further that σ_b^2 and σ_e^2 are known. Provide a simplified expression for the best linear unbiased estimator of $\alpha_1 - \alpha_2$ as a function of $\bar{d}$, $\bar{r}$, $\bar{q}$, σ_b^2 and σ_e^2.