Department of Statistics
University of Wisconsin, Madison
PhD Qualifying Exam Option B
August 29, 2017
12:30-4:30pm, Room 133 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. Let $X \geq 0$ be a positive random variable such that $\mu = \mathbb{E}(X)$ and let X' be an independent, identically distributed copy of X. The *Gini coefficient* γ of the distribution of X is defined by

$$\gamma = \frac{1}{2}\frac{\mathbb{E}\left(|X - X'|\right)}{\mu}.$$

Note that γ is a unit-free measure of the dispersion of X, and $0 \leq \gamma \leq 1$.

This question is about statistical inference for the true value of γ in the Pareto(θ) distribution, whose pdf is given by

$$f_\theta(x) = \frac{\theta}{x^{\theta+1}}, \quad x > 1, \quad \theta > 1.$$

Please read the following problems very carefully, then answer them. Show your work in sufficient detail.

(a) For this problem, we will show that the value of γ for the Pareto(θ) distribution is given by

$$\gamma = \frac{1}{2\theta - 1}.$$

Please answer the following sub-parts.

i. For the Pareto(θ) distribution, calculate $\mu = \mathbb{E}_\theta(X)$.

ii. For the Pareto(θ) distribution, calculate $\frac{1}{2}\mathbb{E}_\theta\left(|X - X'|\right)$.
HINT: You may find the following identity useful. For real numbers a and b, we have $|a - b| = a + b - 2\min(a, b)$, where $\min(a, b)$ is the smaller of a and b.

iii. Use parts i. and ii. to obtain the value of γ.

(b) Let $X_1, \ldots, X_n$ be an i.i.d sample from the Pareto(θ) distribution. Derive an expression for $\widehat{\gamma}$, the maximum likelihood estimator of γ based on $X_1, \ldots, X_n$.

(c) Argue that $\widehat{\gamma}$ is an asymptotically normal estimator of γ, and show that its asymptotic variance is $\gamma^2(1 + \gamma)^2$.

(d) Consider testing $H_0 : \gamma \leq \gamma_0$ vs. $H_A : \gamma > \gamma_0$. Using the result of part (c), write down the rejection region of an asymptotic level-α Wald test of H_0 vs. H_A. Then invert this test into an asymptotic $1 - \alpha$ confidence bound for γ.
NOTE: In your solution to this problem, you may simply use the symbol $\widehat{\gamma}$ to represent the MLE instead of the expression you may, or may not, have derived in part (b).

2. This question is related to decision rules and contains two separate parts.

(a) (Part 1) Suppose that $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim}$ a cumulative distribution function F. Define

$$N_1 = \sum_{i=1}^{n} \mathrm{I}(X_i \leq 0), \qquad N_2 = \sum_{i=1}^{n} \mathrm{I}(X_i > 0).$$

Define $p_1 = \mathrm{P}(X_1 \leq 0)$ and $p_2 = 1 - p_1$, where $0 < p_1 < 1$. Let $S_n = \sum_{j=1}^{2} \frac{(N_j - np_j)^2}{np_j}$.

i. Derive and identify the limit distribution of S_n as $n \to \infty$.

ii. Derive and identify the limit distribution of $\sqrt{S_n}$ as $n \to \infty$.

iii. Suppose that we wish to test the null hypothesis $H_0 : p_1 = \pi_0$ versus the alternative hypothesis $H_1 : p_1 \neq \pi_0$, where the value $\pi_0 \in (0, 1)$ is a known constant. Develop an asymptotically level α test procedure.

(b) (Part 2) Suppose that $Y \in \{0, 1\}$ is a Bernoulli random variable and $\boldsymbol{X} = (X_1, \ldots, X_p)^T$ is a vector of p random covariates. Let $\phi(\boldsymbol{X}) : \boldsymbol{X} \mapsto \{0, 1\}$ denote a decision rule which maps $\boldsymbol{X}$ to the range $\{0, 1\}$ of Y. Here we use the 0-1 loss function (or "misclassification loss function"), i.e., $\ell(Y, \phi(\boldsymbol{X})) = \mathrm{I}\{Y \neq \phi(\boldsymbol{X})\}$, where $\mathrm{I}(\cdot)$ is the indicator function.

i. Calculate the risk $R(\phi) = \mathrm{E}_{(\boldsymbol{X},Y)}\{\ell(Y, \phi(\boldsymbol{X}))\}$.

ii. Calculate the conditional risk $R(\phi \mid \boldsymbol{X}) = \mathrm{E}\{\ell(Y, \phi(\boldsymbol{X})) \mid \boldsymbol{X}\}$.

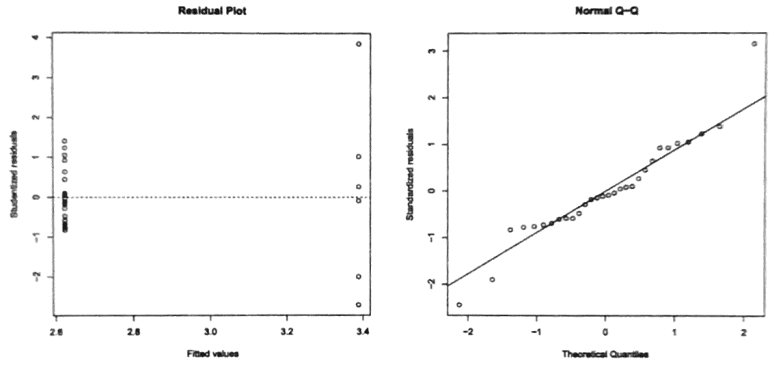
iii. Find the optimal rule ϕ_{opt} which minimizes $R(\phi)$ among all rules ϕ such that $\phi(\boldsymbol{X}) : \boldsymbol{X} \mapsto \{0, 1\}$.

3. A study was conducted to evaluate the extent to which red blood cells settle out of suspension in blood plasma, measured as an erythrocyte sedimentation rate (ESR) in millimeter per hour (mm/h), is related to a protein, called fibrinogen (FIBR), that is present in blood plasma. A random sample of $n = 30$ individuals was drawn from a pool of patients in a local hospital. An individual was classified as healthy if ESR is less than 20 mm/h or unhealthy if ESR is at least 20 mm/h. A binary variable BESR is defined to be 0 if ESR is less than 20 mm/h and 1 if ESR is at least 20 mm/h. FIBR was measured on each individual in units of grams/liter (gm/L).

   The data are organized such that the first $n_1 = 24$ individuals have healthy ESR (with BESR $= 0$) and the next $n_2 = 6$ patients have unhealthy ESR (with BESR $= 1$). The sample correlation between FIBR and BESR was computed to be 0.4788. A simple linear regression analysis was carried out based on the following *Model A*:

$$\text{FIBR}_i = \beta_0 + \beta_1 \text{BESR}_i + \epsilon_i,$$

   where $\epsilon_i \overset{\text{i.i.d.}}{\sim} N(0, \sigma^2)$ for $i = 1, \ldots, n$. Let $\boldsymbol{\beta} = (\beta_0, \beta_1)'$. Two model diagnostics plots are given in the figure below.



   Only for parts (a)–(c) will you need to use the data analysis results above.

   (a) Provide an interpretation of β_0 and β_1 and perform an appropriate test for $H_0 : \beta_1 = 0$ vs. $H_A : \beta_1 \neq 0$ under Model A.

   (b) Perform appropriate model diagnostics.

   (c) Critique the study design and the approach taken for the data analysis.

   (d) Let $\widehat{\boldsymbol{\beta}}$ denote the ordinary least squares (OLS) estimate of $\boldsymbol{\beta}$. Derive the analytical form of $\widehat{\boldsymbol{\beta}}$ and the distribution of $\widehat{\boldsymbol{\beta}}$, including the individual entries of vectors and matrices.

   (e) For testing $H_0 : \beta_1 = 0$ under Model A, derive an appropriate F test, including the degrees of freedom and the non-centrality parameter.

   (f) For this part only, consider a binary variable, denoted as BIN, that takes on the value of 0 for the first 29 individuals and 1 for the last individual. Consider a multiple linear regression *Model B*:

$$\text{FIBR}_i = \theta_0 + \theta_1 \text{BESR}_i + \theta_2 \text{BIN}_i + \epsilon_i,$$

where $\epsilon_i \overset{\text{i.i.d.}}{\sim} N(0, \sigma^2)$ for $i = 1, \ldots, n$. Let $\boldsymbol{\theta} = (\theta_0, \theta_1, \theta_2)'$. Provide an interpretation of θ_2. Derive the analytical form of the OLS estimate $\widehat{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}$ and the distribution of $\widehat{\boldsymbol{\theta}}$, including the individual entries of vectors and matrices.

4. Microbial technologies is a growing research area within agricultural biotechnology. There are growing efforts in reducing our reliance on synthetic fertilizers by the use of nitrogen use efficiency (NUE)-enhancing microbiomes. Recently, a UW-Madison Lab studying *A. Thaliana* carried out an experiment to study seed yield as a measure of NUE for two genotypes (varieties) of *A. Thaliana* exposed to three different compositions of microorganisms grown within shallow pots. The experiment was carried out in 4 different green houses across the UW campus. Each greenhouse contained three benches. Each bench had two pots with a single randomly assigned variety of the *A. Thaliana* plant in each pot. The experimenters made sure that both varieties were represented on each bench. The three microorganism compositions were randomly assigned to the three benches within each greenhouse, with one bench per composition. The researchers computed and recorded a composite measure of seed yield of each plant in each pot at the conclusion of the experiment.

   Let y_{ijk}, $i = 1, 2$, $j = 1, 2, 3$, and $k = 1, 2, 3, 4$, denote the seed yield measured for the plant with genotype i, grown with microorganism composition j in green house k. Consider the following model (Model I):

$$y_{ijk} = \mu + a_i + b_j + c_{ij} + \gamma_k + \tau_{jk} + \epsilon_{ijk}, \quad i = 1, 2, \quad j = 1, 2, 3, \quad k = 1, 2, 3, 4, \qquad (1)$$

   where $\gamma_k \sim \mathcal{N}(0, \sigma^2_\gamma)$, $\tau_{jk} \sim \mathcal{N}(0, \sigma^2_\tau)$, $\epsilon_{ijk} \sim \mathcal{N}(0, \sigma^2_\epsilon)$, and $\sigma^2_\gamma, \sigma^2_\tau, \sigma^2_\epsilon > 0$. All of these random terms are mutually independent, and the remaining terms in the model are unknown fixed parameters.

   (a) Under Model I, what is the correlation between the seed yields of two plants growing together on the same bench?

   (b) Rewrite Model I as: $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \mathbf{e}$ with

$$\begin{aligned}\mathbf{y} \;=\; & (y_{111}, y_{211}, y_{121}, y_{221}, y_{131}, y_{231}, y_{112}, y_{212}, y_{122}, y_{222}, y_{132}, y_{232}, \\ & y_{113}, y_{213}, y_{123}, y_{223}, y_{133}, y_{233}, y_{114}, y_{214}, y_{124}, y_{224}, y_{134}, y_{234})\,.\end{aligned}$$

   Explicitly specify $\mathbf{X}$, $\boldsymbol{\beta}$, $\mathbf{Z}$, and $\mathbf{u}$.

For the remaining parts below, consider the following `R` commands and output where `y` denotes the data vector $\mathbf{y}$ from part (b), `GH`, `MC`, and `GENO` are factors in `R` corresponding to the experimental factors greenhouse, microorganism composition, and genotype, respectively.

```
m <- lm(y ~ GH*MC*GENO)
anova(m)
Analysis of Variance Table
Response: y
          Df Sum Sq  Mean Sq
GH         3 113.3    37.8
MC         2 321.8   160.9
GENO       1   2.5     2.5
```

```
GH:MC       6 116.4    19.4
GH:GENO     3  11.7     3.9
MC:GENO     2  75.1    37.5
GH:MC:GENO  6  14.5     2.4
```

Answer the following questions based on the above ANOVA table for Model I.

(c) Using parameters of Model I, write down explicitly the null hypothesis of no microorganism composition effect.

(d) Test for microorganism composition main effect.

(e) Test for genotype main effect.

(f) Test for microorganism composition by genotype interaction effect.

(g) Given the overall objective of "reducing our reliance on synthetic fertilizers by the use of nitrogen use efficiency (NUE)-enhancing microbiomes", critique this designed experiment.

(h) The researchers are allowed to replace the three small benches with one large bench that can hold 6 pots within each green house. Then the two genotype and three microorganism composition combinations are randomly assigned to the pots within each green house by making sure each combination is represented once within each green house. Comment on this design compared to the original design and specify how your tests for parts (d), (e), and (f) would change. Be explicit and carry out the tests whenever you can.