STAT–351 Final Project. Due time: 10:05AM–12:05PM, on May 5 (Wednesday), 2021

Note: Please submit your report, **limited to 8 pages**, to Canvas of STAT-351, during 10:05AM–12:05PM, on May 5, 2021 (Wednesday). **No late work will be accepted**. In order to receive credit, you must show all work neatly.

Name: _____

SID: _____

Pledge: On my honor, I have neither given nor received unauthorized aid on this examination.

Signature: _____

1. Cuckoos are known to lay their eggs in the nests of other (host) birds. The eggs are then adopted and hatched by the host birds. Is there any relationship between the length of eggs and the host birds? What do we conclude? All data are lengths in millimeters. The data file "`Cuckoo_Egg_Lengths_data.pdf`" is uploaded at Canvas.

2. Assume that for a discrete random variable $L \sim \text{Bernoulli}(1/2)$, a random variable X has the conditional distribution,

$$\begin{aligned} X \mid \{L = 0\} &\sim N\Big(-2, \frac{1}{2^2}\Big), \\ X \mid \{L = 1\} &\sim N\Big(+2, \frac{1}{2^2}\Big). \end{aligned}$$

   Generate 100 random variates $\{X_i\}_{i=1}^{100}$ from the distribution of X. Use simulation studies to numerically estimate $\text{E}(X)$, $\text{median}(X)$, and $\text{var}(X)$.

3. A data set on ozone can be downloaded from the web site `http://web.stanford.edu/~hastie/ElemStatLearn/`. Let Y denote the ozone concentration level (given in the 1st column), X_1 denote "radiation" (given in the 2nd column), X_2 denote "temperature" (given in the 3rd column), and X_3 denote "'wind speed" (given in the 4th column).

   (a) Obtain the kernel regression fit to measurements of (X_1, Y).

   (b) Obtain the kernel regression fit to measurements of (X_2, Y).

   (c) Obtain the kernel regression fit to measurements of (X_3, Y).

   (d) Interpret the results obtained in parts (a), (b), and (c).

4. We wish to learn the relationship between ranks and variate values via an empirical study. Let $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} X$, where X has a continuous C.D.F. and $\text{E}(X^2) < \infty$. Let $R_1, \ldots, R_n$ be the ranks of $X_1, \ldots, X_n$, i.e., $R_i = \text{rank}(X_i)$, $i = 1, \ldots, n$. Please simulate data $X_1, \ldots, X_n$, with $n = 100$, and four types of distributions of X, $X \sim \text{Unif}(0, 1)$, $X \sim \chi^2_1$, $X \sim N(0, 1)$, $X \sim \text{Bin}(5, 0.9)$, respectively; for each distribution of X, compute the Pearson "product-moment" sample correlation coefficients $\widehat{\rho}_1$, $\widehat{\rho}_2$ and $\widehat{\rho}_3$ as follows:

(a) $\widehat{\rho}_1$ using $\{(X_i, R_i) : i = 1, \ldots, n\}$;

(b) $\widehat{\rho}_2$ using $\{(X_i, R_{i+1}) : i = 1, \ldots, n-1\}$;

(c) $\widehat{\rho}_3$ using $\{(X_i, R_{i-1}) : i = 2, \ldots, n\}$.

Please summarize your results and conclusion.

5. Let $Y_1 = 0$, $Y_2 = 2$, $Y_3 = -1$, and $Y_4 = 4$. For $t \in [1, 4]$, plot a curve $\mu(t)$, which is assumed to have two continuous derivatives and minimizes the penalized sum of squares,

$$\frac{1}{4}\sum_{i=1}^{4}\{Y_i - \mu(i)\}^2 + \lambda \int_1^4 \{\mu''(t)\}^2 \mathrm{d}t$$

for $\lambda = 0$, $\lambda = 1$, $\lambda = 2$, and $\lambda = 1000$ separately. (Warning: the MATLAB cubic spline minimizes a slightly different function.)