Fall 2022 STA6703/EEL5934 SML Pretest

Directions

Please submit **ONE PDF** file including all your reports (answer + code + figures + comments; must be easily readable and have file size under a few megabytes) and **ONE R code script**. The R script is supplementary material to ensure that your code runs correctly. If you are using RMarkdown, please also include your `.Rmd` file. If not using R, the same considerations apply; i.e., please also include your scripts.

Place these two (or three) files in a folder, make a zip or rar archive, and submit the archive electronically via Dropbox file request at `tinyurl.com/nbliznyuk-submit-files` (on the landing page, enter your name so that we know it is you and email so that you get a confirmation).

Deadline: 26-Aug-2022 (Fri), 10:00 PM EST. This problem set should not take you more than a few hours to solve, hence the tight deadline to help you decide whether this course is appropriate for you. The problem set carries no grade or points towards the final course grade; however, students that can solve only 60% (tentative) or less of this assignment correctly will likely find this course unnecessarily challenging due to major gaps in their command of prerequisites.

The assignment must be completed individually. Any group work or discussion will be in violation of the UF Honor code.

Grading: each problem is worth 10 points and, whenever possible, all subproblems are equally weighted; problem 2: (3+3+4) split; problem 5: (4 + 3 + 3) split.

Typed problem 1 (some calculus-based probability)

Let $Y \sim Exponential(\lambda)$ be a model for the lifetime of an electr(on)ic component (e.g., a lightbulb), where λ is rate parameter (reciprocal of the scale parameter) and $Y_1, \ldots, Y_n$ are independent observations from the $Exponential(\lambda)$ distribution. Suppose $n = 5$ (known) and the true λ is 1 (generally, unknown but you'll need this to report the bias and variance of your estimators).

A useful fact (take it for granted without derivation): $S_n = Y_1 + \ldots + Y_n$ can be shown to have the probability density function $f(x) = \lambda^n x^{n-1} e^{-\lambda x}/(n-1)!$ for $x > 0$ and 0 otherwise. If $n = 5$, $(n-1)! = 24$.

You can solve the subproblems analytically or numerically. If solving numerically, your answers (all reported quantities) must be correct to 4 significant digits (at least).

1.1: Find bias and variance of the estimator $\widehat{\beta} = S_n/n$, (i.e., the sample mean) for estimation of $\beta = 1/\lambda$.

1.2*: Find bias and variance of the estimator $1/\widehat{\beta}$ for estimation of λ. In view of 1.a, would you expect $1/\widehat{\beta}$ to be unbiased for estimation of $\lambda = 1/\beta$? Briefly explain.

Typed problem 2 (basic linear algebra)

The goal here is to solve for x - by any means necessary (analytically or numerically) - the linear system (system of linear equations) $A \cdot x = b$, where b is the known right-hand side and A is the known matrix of equation coefficients.

2.1: Let A = [A1, A2, A3] and b be defined by the following table

A1	A2	A3	b
1	0	1	1
2	1	1	1
3	1	3	1

In R, the matrix A can be defined as

```
>> A = matrix(c(1,0,1, 2,1,1,3,1,3),nrow=3,byrow=TRUE);
```

Solve $A \cdot x = b$ (i.e., find x such that $A \cdot x = b$).

2.2: Suppose the A_{33} entry is now 2 but the rest of the A matrix is the same, i.e.,

A1	A2	A3	b
1	0	1	1
2	1	1	1
3	1	2	1

Can one solve the system $A \cdot x = b$ (with the same right-hand side b)? Briefly explain why or why not.

2.3*: Suppose the matrix of coefficients is the same as in 2.b., but the right-hand side b is general. Are there nonzero values of b for which there is a solution to the system $A \cdot x = b$? Is the solution unique? (If not, characterize all solutions.) Briefly explain why or why not.

Typed problem 3 (normal theory multiple linear regression)

Suppose one ran a multiple linear regression model with two covariates and an intercept, e.g.,

```
>> lm(Y ~ X1 + X2) # in R.
and obtained the following summary
---
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   1.2717     0.6961   1.827   0.0813 .
X1            1.6882     0.7044   2.397   0.0255 *
X2           -1.3406     1.0129  -1.324   0.1992
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.98 on 22 degrees of freedom
Multiple R-squared:  0.2331,    Adjusted R-squared:  0.1634
F-statistic: 3.344 on 2 and 22 DF,  p-value: 0.05396
```

Here, the number of observations n is 25, and the true errors can be assumed to be independent normal with a zero mean and the same variance (unknown). Call the regression coefficients β_0 (intercept), β_1 (for X1) and β_2 (for X2).

3.1: Test the null hypothesis H_0: $\beta_1 = 0$ and briefly report your findings.

3.2: Test the null hypothesis H_0: $\beta_1 = 1$ (stating the value of the test statistic and its reference distribution) and briefly report your findings. If this is not possible using the information provided, explain why and what additional information would be needed.

3.3: Test the null hypothesis H_0: $\beta_2 = 0$ and briefly report your findings. Can we conclude that there is no association between Y and X2? How can one ascertain this?

3.4: How would one test the null hypothesis of no association between Y and (jointly) X1 and X2? (I.e., here, H_0: $\beta_1 = \beta_2 = 0$.) Is the summary above sufficient to this end? Briefly explain.

3.5*: Suppose the covariate values for a new observation are $x_{new} = [1, 1]$. Can the above summary be used to test the null hypothesis about the mean of a new observation, H_0: $\beta_0 + \beta_1 \cdot 1 + \beta_2 \cdot 1 = 3$. Briefly explain why or why not. If not, what additional information would be required?

Typed problem 4 (basic programming)

Implement "by hand'' an R (or Python) function that outputs all subsets S_j of a set $S = \{s_1, \ldots, s_p\}$ that have exactly m elements, $1 \leq m \leq p$. To represent a subset of S, use a vector of inclusion indicators; i.e., represent S_j using a vector V_j of length p such that, for $k = 1, 2, \ldots, p$, $V_j[k] = 1$ if $s_k \in S_j$ and $V_j[k] = 0$ otherwise.

Here, "by hand" means providing your own implementation that does not depend on any of the existing R libraries (other than `base`). The same considerations apply if you choose to use Python (but you are allowed to use matrices or lists).

Use the following interface for the function:

```
getSubsets <- function(p,m) {
  # * inputs: p > 0 is the size of S; m >= 0 is the size of the desired subsets
  # * output: matrix M whose rows are vector representations of size m
    subsets of S (i.e., M[j,]=Vj), ordered lexicographically (i.e., from smallest to
    largest, as if your vector representations were words in a dictionary)
}
```

E.g., the call `getSubsets(4,3)` should produce the following matrix

```
[1,] 0 1 1 1
[2,] 1 0 1 1
[3,] 1 1 0 1
[4,] 1 1 1 0
```

but your function should work (albeit slowly) for all valid values of p and m (however, realistically it would be called only for small values of p, e.g., $p \leq 10$ or $p \leq 20$). Make sure you test it.

Notice that the above representation of sets looks like a binary representation of integers; e.g., the rows of the matrix M above correspond to integers 7, 11, 13, and 14. Write an auxiliary function that does this conversion for a general vector V_j; you'll need it for the online portion of this pretest. I.e., the decimal representation of V_j is $d_j = \sum_{k=1}^{p} V_j[k] \cdot 2^{p-k}$.

Hints: (1) Enumerate the entire set of 2^p subsets (of all sizes) of S first (using a loop with roughly p iterations), then identify and extract the subsets with exactly m elements (do this without loops). (2) If M is a matrix of all subsets of a set with k elements, how can one use it to obtain a matrix of all subsets of a set with $(k+1)$ elements?

This implementation is sensible when p is small and $m \in \{1, \ldots, p\}$. When p is large and m is very small (e.g., as in so called "sparse" models), alternative implementations are preferable (see below).

(Optional - not for submission): can you solve this problem for general values of p and m without ever generating a redundant subset thereby ignoring the above hint? E.g., if p=100 and m=1, then the function will generate only the 100 output vectors that represent valid subsets, and no invalid subsets are ever considered; however, these choices of p and m are not "hard-wired".

Typed problem 5 (dependence and prediction)

The goal of this problem is to use information about the dependence of two random variables (rvs), Y and X, to make predictions about the rv Y.

5.1. Suppose the rvs X and Y have expectations equal to zero, variances equal to 1 and correlation equal to ρ (known), where $|\rho| < 1$. Let $\widehat{Y} = a + bX$ be a predictor of Y that is linear in X, where a and b are known constants. Let $U = Y - \widehat{Y}$ be the prediction error incurred when predicting Y with $\widehat{Y}$. Find the values of a and b that simultaneously guarantee that (i) the predictor is unbiased, i.e., $E(U) = 0$ and (ii) the prediction error variance $Var(U)$ is minimized. Briefly discuss the effect of ρ on the prediction accuracy (variance).

For 5.2 and 5.3, assume the following setup:

Suppose X and Y have the following joint probability density function (pdf):

$$f(x, y) = c_2 \exp(-c_1(x^2 - 1.6xy + y^2)),$$

where $c_1 = 1/0.72$, $c_2 = 0.2652582$ (the normalizing constant), and $(x, y) \in \mathbb{R}^2$.

Find the marginal pdf of Y and the conditional pdf of Y given $X = x$; you'll need these below. *Hint: both of these will be univariate normal pdfs. Your goal is to fully characterize those by determining their parameters - means and variances.* You can do this analytically or numerically using deterministic numerical integration; ideally, both so that you know your answer is correct.

In what follows, suppose all of the above pdfs are known (i.e., need not be estimated).

5.2. If no information about the observed value of X is available, what is your best guess of Y and what are the bias and variance of the prediction error that it attains?

5.3. Suppose the event $X = 2$ is observed, and $E(Y|X = 2)$ is used to guess Y. What is the variance of the prediction error for this predictor?