Name: _____

Department of Mathematics & Statistics
Statistics Methods & Computation Preliminary Examination

Start: August 10, 2023, 10:00 am
End: August 11, 2023, 12:00 pm (midday)

Directions:

1. This is a take-home exam on the contents of the following Methods/Computation courses: STAT 7650 Computational Statistics (CS), STAT 7840 Applied Multivariate Data Analysis (AMDA), and STAT 7020 Applied Regression Analysis (ARA).
2. The exam has 2 questions on each area (so in total 6 questions available), and you must pick one question from each area (therefore you will submit responses to 3 questions in total for this exam). You are not allowed to submit responses to more than one question for each area.
3. You are allowed to use a calculator and some problems will require the use of software. For the software, `R` is preferred, but you are free to use any software you have access to.
4. You are allowed to use any (text)book(s) and online resources, however your responses must be a result of your own work. Any sign of plagiarism or cheating will result in automatic failure.
5. Work any one of the two problems in each area. You may submit solutions for at most three problems.
6. You need to start each problem on a new page. Clearly label each problem and number each page and write your name on top right of each page.
7. To get full credit you need to properly document and explain your solutions.
8. Each problem is worth 20 points.

Please mark the three problems you are submitting for grading in the table below.

Problem	CS1	CS2	AMDA1	AMDA2	ARA1	ARA2
Submit for grading						
Score						

Computational Statistics (CS)

CS1. Consider the data on coal-mining disasters from 1851 to 1962 (see Exercise 6.4 in Givens & Hoeting) (data is provided in coal.csv in the email attachment). For these data, assume the model

$$X_j \sim \begin{cases} \text{Poisson}(\lambda_1), & \text{for } j = 1, \ldots, \theta \\ \text{Poisson}(\lambda_2), & \text{for } j = \theta + 1, \ldots, 112 \end{cases}$$

Assume $\lambda_i|\alpha \sim \text{Gamma}(3, \alpha)$ for $i = 1, 2$, where $\alpha \sim \text{Gamma}(10, 10)$, and assume θ follows a discrete uniform distribution over $\{1, \ldots, 111\}$. You will be estimating the posterior distribution of the model parameters via a Gibbs sampler.

(a) Derive the conditional distributions necessary to carry out Gibbs sampling for the change-point model.

(b) Implement your Gibbs sampler. Use a suite of convergence diagnostics (like trace plots and autocorrelation plots) for your samples of λ_1, λ_2, and θ to justify convergence and mixing of your sampler.

(c) Construct density histograms and a table of summary statistics for the approximate posterior distributions of θ, λ_1, and λ_2. Are symmetric HPD intervals appropriate for all of these parameters?.

Hints:

(I) Keep in mind that θ is an integer.

(II) Don't forget to interpret the results in the context of the problem.

CS2. Consider sampling from Gamma$(\theta, 1)$ distribution.

(a) If θ is an integer, how would you sample from this distribution using Exponential(1) random variables. Justify your answer.

(b) If θ is not an integer, then sampling from it is not as easy. Develop an rejecting sampling method for simulating from Gamma$(\theta, 1)$ for non-integer $\theta \geq 2$.

(c) Derive the acceptance rate for your rejection sampling method.

(d) Implement your rejection sampling method. Simulate 1000 values from Gamma$(5.5, 1)$ using your method and draw a histogram with the gamma density curve overlaid. How is the fit? Compare your empirical acceptance rate with the result in part (c)?

(I) In part (b), for non-integer θ, consider a gamma proposal with shape $[\theta]$, the integer part of θ, and scale b chosen so that the mean of the proposal is θ.)

(II) In part (d), you are not allowed to use `rgamma` or `qgamma`.

Applied Multivariate Data Analysis (AMDA)

AMDA1. Three different missile classes were designed in the Aerospace Engineering department and then for each class 1500 random flights were simulated (using their 6Dof flight simulator). For each flight, the maximum height reached (Apogee in thousands of feet), the downrange distance traveled (Distance in thousands of feet), and the amount of time the missile flew (TOF in seconds) were recorded. The data were randomly split into Training (500 flights each class), Validation (500 flights each class) and Testing (500 flights each class). The Test data does not contain the class labels.

1. Using the training data set, do numerical and graphical summary statistics to compare and contrast the three different classes. Among the varioius graphics include boxplots and a matrix plot. in the matrix plot, use different symbols for the three classes. Based on the tabular and graphical summaries, compare and contrast the three different classes.
2. Using the training data set, conduct a full MANOVA to compare the three classes and characterize the significant and/or non-significant differences.
3. Using the training data set, compute and interpret the principal components (based on the combined data ignoring classes), including the "factor loadings", and produce a scatter plot with the first component on the horizontal axis and the second component on the vertical axis and use different symbols for the different missile classes.
4. Using the training data set, for each class compute the sample mean vectors and sample var/cov matrices, $\bar{\mathbf{x}}_1$, $\bar{\mathbf{x}}_2$, and $\bar{\mathbf{x}}_3$, $\mathbf{S}_1$, $\mathbf{S}_2$, and $\mathbf{S}_3$, and compute the combined $\mathbf{S}_{\text{pooled}}$. Also, compute the generalized sample variances and total sample variances for each group and $\mathbf{S}_{\text{pooled}}$. Based on $\mathbf{S}_1$, $\mathbf{S}_2$ and $\mathbf{S}_3$, does the assumption $\mathbf{\Sigma}_1 = \mathbf{\Sigma}_2 = \mathbf{\Sigma}_3$ seem reasonable? Explain. Is there a way to test the common variance assumption? If so, do so.
5. Do a full linear discriminant analysis and classification assuming $\mathbf{\Sigma}_1 = \mathbf{\Sigma}_2 = \mathbf{\Sigma}_3$ and $\pi_1 = \pi_2 = \pi_3$. Compute the predicted class probabilities, allocate each observation to the appropriate class according to these probabilities, interpret the linear discriminants, produce the discriminant plot/histograms (different symbols for each class) and compute the confusion tables (Training and Validation). For the validation data, compute the sensitivity, specificity, false positive rate, false positive rate and the overall error rate.. Is there strong discrimination between the groups? Explain.
6. The test data set has 1500 observations but does not contain class labels. Use the model trained in (d) to predict the class for each of the fly-outs (summarize the results). Provide a comma delimited data set called "testout.csv" that contains the original test dataset, along with the class prediction probabiliites and the predicted classes.

AMDA2. 1200 missiles ($n = 1200$) were designed in the Aerospace Engineering department and from each they simulated a random flight (using their 6Dof flight simulator). For each flight, $p = 5$ different summary statistics were computed, the maximum thrust (Lbf/1000, thousands of foot pounds), maximum height reached (Apogee in thousands of feet), the downrange distance traveled (Distance in thousands of feet), the amount of time the missile flew (TOF in seconds) and weight of missile were recorded. The data was saved as a csv file called "UnsupervisedClass.csv". It turns our that the data is made up of many different classes of missles (muliple fly-outs per class), but the missile classes are unknown. The purpose of this project is to estimate the number of classes and group missiles into like classes (unsupervised classification/clustering).

1. Do numerical and graphical summary statistics as part of an exploratory analysis (include mean vectors and correlation matrix, as well as other statistics). Among the varioius graphics include boxplots and a matrix plot. etc. Explain any patterns that you discover.
2. Do the data seem to have come from a multivariate normal population ($p = 5$)? Explain using plots and statistical tests.
3. Do a full principal component analysis on this data based on the correlation matrix (scaled). Is there any indication that the data can be well represented in a lower dimension ($< p = 5$)? Explain. Do a scatter plot matrix of the first few principal components and interpret? Use the last principal component (the one with the smallest variance) to assess the multivariate normality assumption.
4. Use at least two unsupervised classification techniques to find natural groupings in the data. Use a k-means algorithym and at least one agglomerative method to come up with natural groupings in the data. Compare and contrast. In both cases, you must select and justify the number of clusters and groupings/clusters that you choose. Between the two or more methods pick one as your final set of clusters.
5. Provide scatter plots (or matrix plots) for the various combinations of variables (either raw/original and/or principal components) using different symbols or colors for each cluster.
6. Another way to assess cluster integrity (of your selected set of clusters) would be to do a Linear Discriminant Analysis treating the clusters as known classes and compute the confustion matrix. Also, the linear discrimants can tell you a lot of what differentiates the clusters in a meaningful way.

Department of Mathematics and Statistics
Auburn University

Applied Regression Analysis (ARA)

This exam assumes the data can be described by the linear model defined as follows:

$$\mathbf{y} = \mathbf{X}\beta + \varepsilon,$$

where $\mathbf{y} \in \mathbb{R}^n$ is the vector of responses, $\mathbf{X} \in \mathbb{R}^{n\times p}$ is the matrix of regressors, $\beta \in \mathbb{R}^p$ is the coefficient vector linking the regressors to the response, and $\varepsilon \in \mathbb{R}^n$ is the vector of *i.i.d.* errors.

ARA 1. Consider the `emeralds` data set where the interest lies in explaining/predicting the `price` of the emerald with respect to the other variables. Address the following points and adequately justify all your responses.

1. Perform a descriptive analysis of the data using the appropriate visual tools. Comment on the relationships in the data and discuss what implications these could have when fitting a linear model to the data.
2. Perform a multiple linear regression with `price` as the response and all other variables as the explanatory variables. Run and interpret all analyses and tests you feel are pertinent and necessary to obtain a conclusive and reliable analysis of this data. In particular, make sure you also address the following points:
   a) Use the R^2, R^2_{Adj} and an analysis-of-variance table to check the significance of your regression. Discuss the assumptions, advantages and limitations of these quantities and tests.
   b) Show numerically that the square of the simple correlation coefficient between the observed values y_i and the fitted values $\hat{y}_i$ equals R^2.
   c) Find a 95 % CI on the mean response when the `carat` is 0,75, the `cut` is premium, the `color` is G, `clarity` is $VS1$, `depth` is 62,5, `x` is 5.7, `y` is 4.7 and `z` is 6. Compare lengths of CIs when adding or removing the variable `carat`: what does this tell you about the impact of this variable on the response?
   d) Refit the model using only `carat` as the regressor. Run the pertinent tests and discuss your findings. Based on these analyses, are you satisfied with this model?
   e) Run a complete analysis of the residuals of the *full* model (i.e. all regressors). What assumptions need to be checked and are they satisfied for this model?
   f) Identify an appropriate transformed model for these data. Fit this model to the data and conduct the usual tests of model adequacy. Would a transformation improve this analysis? Why or why not? If yes, perform the transformation and repeat the full analysis.
   g) Perform a thorough influence analysis of these data. What conclusions do you draw from this analysis? What are the variance inflation factors? Why are these important?
   h) Apply ridge regression to the data (justifying your approach) and compare the result to those from the forward selection algorithm and stepwise regression. Discuss advantages and disadvantages of each in this comparison. Pick and justify the use of one of these procedures to obtain a final model and interpret your results.

i) Delete half the observations (chosen at random) to create the estimation data, and refit the regression model on this data. Have the regression coefficients changed dramatically? How well does this model predict the response for the deleted observations (i.e. the prediction data)?

j) Fit a model to the estimation data set using all possible regressions and select the final model based on criterion of your choice (justify it). Use this model to predict the responses for each observation in the prediction data set and comment on model adequacy.

3. Show that an equivalent way to perform the test for significance of regression in multiple linear regression is to base the test on R^2 as follows: To test $H_0 : \beta_1 = \beta_2 = \ldots = \beta_k$ versus H_1 : at least one $\beta_j \neq 0$, calculate

$$F_0 = \frac{R^2(n-p)}{k(1-R^2)}$$

and to reject H_0 if the computed value of F_0 exceeds $F_{\alpha,k,n-p}$, where $p = k+1$.

4. Show that $\text{Var}(\hat{\mathbf{y}}) = \sigma^2\mathbf{H}$.

5. Consider the multiple linear regression model $\mathbf{y} = \mathbf{X}\beta + \varepsilon$. Show that the least-squares estimator can be written as $\hat{\beta} = \beta + \mathbf{R}\varepsilon$ where $\mathbf{R} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$

6. Consider a correctly specified regression model with p terms, including the intercept. Make the usual assumptions about ε. Prove that

$$\sum_{i=1}^{n} \text{Var}(\hat{y}_i) = p\sigma^2$$

ARA 2. Consider the `expenses` data set where the interest lies in explaining/predicting the medical `charges` with respect to the other variables. Address the following points and adequately justify all your responses.

1. Perform a descriptive analysis of the data using the appropriate visual tools. Comment on the relationships in the data and discuss what implications these could have when fitting a linear model to the data.

2. Perform a multiple linear regression with `charges` as the response and all other variables as the explanatory variables. Run and interpret all analyses and tests you feel are pertinent and necessary to obtain a conclusive and reliable analysis of this data. In particular, make sure you also address the following points:

a) Using t statistics and confidence intervals for the coefficients, conclude on the roles of all variables in describing and predicting the response. Discuss the implications of selecting and interpreting variables using this approach.

b) Using the partial F test, determine the contribution of each variable to the model. How is this partial F statistic related to the t test for each variable?

c) Is multicollinearity a potential problem in this model? Find the condition number of $\mathbf{X}'\mathbf{X}$. Is there evidence of multicollinearity in these data? What can you say about the source of multicollinearity in these data? What are the possible consequences of multicollinearity on estimation and interpretation?

d) Run a complete analysis of the residuals of the *full* model (i.e. all regressors). What assumptions need to be checked and are they satisfied for this model?

e) Construct the partial regression plots for this model. Compare the plots with the plots of residuals versus regressors. Discuss the type of information provided by these plots. Does it seem that some variables currently in the model are not necessary?

f) Identify an appropriate transformed model for these data. Fit this model to the data and conduct the usual tests of model adequacy. Would a transformation improve this analysis? Why or why not? If yes, perform the transformation and repeat the full analysis.

g) Perform a thorough influence analysis of these data. What conclusions do you draw from this analysis? What are the variance inflation factors? Why are these important?

h) Estimate model parameters for the data using principal-component regression and provide a full justification and interpretation of the different aspects of this approach when run on this data (e.g. comparison of R^2 compared to basic least-squares, shrinkage size, etc.). Are there problems in applying principal-component regression to this data? If so, why?

i) Delete half the observations (chosen at random) to create the estimation data, and refit the regression model. Have the regression coefficients changed dramatically? How well does this model predict the response for the deleted observations (i.e. prediction data)?

j) Develop a regression model using the prediction data set. How do the estimates of the parameters in this model compare with those from the model developed from the estimation data? What does this imply about model validity?

3. Show that an alternate computing formula for the regression sum of squares in a linear regression model is

$$SS_R = \sum_{i=1}^{n} \hat{y}_i^2 - n\bar{y}^2$$

4. Show that the residuals from a linear regression model can be expressed as $\mathbf{e} = (\mathbf{I} - \mathbf{H})\varepsilon$.

5. Prove that R^2 is the square of the correlation between $\mathbf{y}$ and $\hat{\mathbf{y}}$.

6. For the multiple linear regression model, show that $SS_R(\beta) = \mathbf{y}'\mathbf{H}\mathbf{y}$.