UNIVERSITY OF MASSACHUSETTS
Department of Mathematics and Statistics
Basic Exam - Applied Statistics
January 2022

Work all problems. 60 points are needed to pass at the Masters Level and 75 to pass at the Ph.D. level. The number of points for each part of each question is listed inline with the part.

Question:	1	2	3	4	5	Total
Points:	20	20	20	15	25	100
Score:						

1. For 50 high schools in Montana, researchers examined the relationship between

   - Y, a school performance measure (based on test outcomes for a random group of recent graduates of the school), measured on a scale from -100 to 100 points

   and predictors X_1 and X_2, where

   - X_1 = a continuous variable that represents the quality of teachers in the school (centred such that $X_1 = 0$ for a school with teachers of average quality),
   - X_2 is an indicator variable that refers to school size, where $X_2 = 1$ for a large school, and $X_2 = 0$ for a small school (based on number of students).

   Results of a fitted regression model with X_1, X_2 and an interaction between X_1 and X_2 are below.

```
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.2279     0.1093   2.086   0.0383 *
X1            2.9695     0.2016  14.728  < 2e-16 ***
X2           -1.1965     0.1519  -7.877 2.22e-13 ***
X1:X2        -2.0134     0.2650  -7.599 1.19e-12 ***
```

   (a) (2 points) Write down the expression for the FITTED regression lines for the two types of schools (no b's, use numbers).

   (b) (2 points) Interpret estimates b_0 and b_3. Words only, no Y's or X's in your interpretation.

   (c) (4 points) Assume modeling assumptions hold true. After fitting the regression model, the researchers concluded that the null hypothesis that $\beta_2 \geq 0$ was rejected at significance level $\alpha = 0.05$ against the alternative hypothesis that $\beta_2 < 0$. One of the researchers concluded: "At significance level $\alpha = 0.05$, we reject the null hypothesis." Do you agree with this assessment? How would you interpret the test result in terms of school performance?

   (d) (2 points) If model assumptions for the model above hold true, do we have statistical evidence that large schools are doing worse than small school with respect to students' test scores, regardless of teacher quality? Do not do any calculations but motivate your answer.

   (e) (4 points) Suppose that in an alternative classification system, schools are classified into three groups: very small, medium and very large schools (thus some of the small and large schools from the former classification end up in the medium category). Write down one general expression for $E(Y)$ (that applies to schools of any size) that includes the new categorization, such that mean test score is estimated using

   - the new categorization for school size,
   - X_1
   - the interaction between X_1 and the new categorization of school size.

   Be sure to be clear on which categories are associated with which X values.

   (f) (6 points) How can the researchers decide if they should present the results from the first model vs. the second? Suggest an exploratory approach (using residual plots) AND a quantitative approach (using a test or some other measure of model fit).

2. The computation of a determinant is needed when solving a set of linear equations using Cramer's Rule or when doing a transformation of variables using the Jacobian of a function.

   For a 2×2 matrix, A, the determinant can be computed as

$$|A| = \begin{vmatrix} a & b \\ c & d \end{vmatrix} = ad - bc. \tag{1}$$

   For a 3×3 matrix, A, the determinant can be computed as

$$|A| = \begin{vmatrix} a & b & c \\ d & e & f \\ g & h & i \end{vmatrix} = a \begin{vmatrix} e & f \\ h & i \end{vmatrix} - b \begin{vmatrix} d & f \\ g & i \end{vmatrix} + c \begin{vmatrix} d & e \\ g & h \end{vmatrix}. \tag{2}$$

   For a general $n \times n$ matrix, the determinant is computed as the sum of the signed minors of any row or column of the matrix scaled by the elements in that row or column. In the case of the 3×3 matrix, we selected the first row. For a square matrix $n \times n$ A with elements a_{ij}, we can select the first row ($i = 1$) containing the vector $[a_{11}, \ldots, a_{1n}]$. The determinant is $|A| = \sum_{j=1}^{n} a_{1j}(-1)^{1+j} M_{1j}$, where M_{ij} is the minor. The minor is the determinant of a smaller square matrix, cut down from A by removing the i-th row and the j-th column.

   (a) (8 points) Write a function in python or R to compute the determinant of a 2×2 matrix. In python, the function should take as a parameter a nested list A and return a scalar and there should be an equivalent representation in R. The function name should be `det2`. For example, $A = [[1, 2], [3, 4]]$.

   (b) (8 points) Suppose, now, that you need to write a function to compute the determinant for an $n \times n$ matrix. You must write a recursive function to do so using the 2×2 case as the base case. For an $n \times n$ matrix, how many 2×2 determinants must be computed? Show your derivation and justify your answer.

   (c) (4 points) Given your above derivation, what is the computational complexity of this recursive algorithm for computing the determinant? Is the algorithm NP-hard (not a polynomial) or not and why?

3. A common way to sample from a distribution is splice sampling. The idea is to draw a sample from a uniform distribution and then use the inverse cumulative distribution function to identify a sample from the corresponding distribution. Recall, the CDF is upper semi-continuous and non-decreasing.

   Consider the following complex distribution. A person flips a fair coin and if the coin is heads, they roll a fair die and report the number of pips on the die. Otherwise, if the coin turns up tails, they roll a 3-sided fair die with sides labeled $\{1, 2, 3\}$ and report the result. The sample space of this experiment is $\{1, 2, 3, 4, 5, 6\}$.

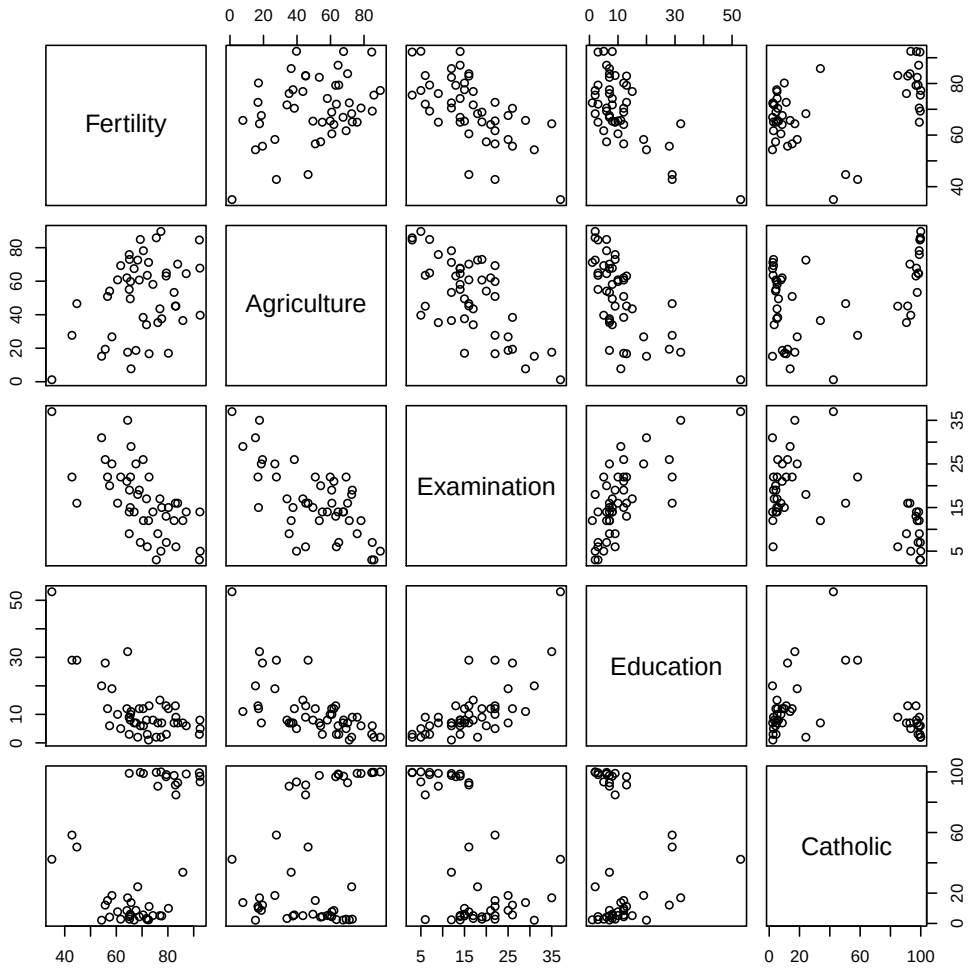
   (a) (8 points) Write a *generative* function (without using slice sampling) to draw and return a single sample according to the algorithm described above. The `numpy` function `numpy.random.choice(a=a, p=p)` returns a single sample from a 1-d array `a` (e.g. ['H', 'T'] or ['1', '2', '3', '4', '5', '6']) according to the probabilities in the 1-d array `p`. You may approximate the probabilities as needed. The `R` function `sample(x=x, size=1, prob=p)` returns a single sample from a vector `a` according to the probabilities in the vector `p`. You may approximate the probabilities as needed.

   (b) (4 points) Compute the probabilities of all of the outcomes of the complex distribution and write them as a table.

   (c) (8 points) Write a *function that uses slice sampling* to return a sample from the complex distribution. The python function for drawing from a uniform distribution is `numpy.random.uniform()`. The R function for drawing from a uniform distribution is `runif(1)`.

4. Below is a pairs plot of 5 variables related to Fertility (rates of childbirth) in 47 provinces of Switzerland in 1888. Suppose you are interested in using the data to model Fertility as a function of these other variables, most of which were measured on men drafted into the military:

Agriculture	% of males who work in agriculture
Examination	% of students receiving the highest score on an examination
Education	% of draftees with beyond primary school education
Catholic	% Catholic (as opposed to 'Protestant')



(a) (5 points) Do you notice any possible outliers in these data? If so, for the point you are most concerned about, briefly describe why it is an outlier. Briefly describe how you would decide what to do about this point.

(b) (5 points) Does the relationship between Catholic and Fertility look linear? If not, describe a different way you could include the Catholic variable in a linear regression for Fertility.

(c) (5 points) Describe one other feature in these plots that you would want to consider when fitting your model. What would you do about this issue when fitting and checking your model?

5. In the last few years, the world has experienced an epidemic of SARS-COV-2 (COVID-19). Data have been very important in combating SARS-COV-2. In this problem we will consider hypothetical data that might give us even more information about SARS-COV-2. Suppose we have data on the COVID-19 history of every person in Massachusetts in June 2020, then again in June 2021. We have the following variables available for each person in Massachusetts at each of these times:

Ever_Positive	Ever had COVID-19 (0=no, 1=yes)
Ever_Hospital	Ever hospitalized for COVID-19? (0=no, 1=yes)
Ever_Symptomatic	Had symptoms of COVID-19 since February 2020? (0=no, 1=yes)
Test_pos	Ever tested positive for COVID-19? (0=no, 1=yes)
Positive	Currently has an active COVID-19 infection? (0=no, 1=yes)
Hospital	Currently hospitalized COVID-19? (0=no, 1=yes)
Symptomatic	Currently has symptoms of COVID-19? (0=no, 1=yes)
Vaccinated	Received COVID-19 vaccination? 3 levels: full, partial, none
Health_work	A healthcare worker? (0=no, 1=yes)
Essential	An 'essential worker' as classified by Massachusetts? (0=no, 1=yes)
Age	Age in years in June 2020
Gender	3 levels: Male, Female, Other
Condition	Have a pre-existing condition relevant to COVID-19 risk? (0=no, 1=yes)

(a) (5 points) We fit a logistic regression model the outcome variable `y=Ever_Hospital` with the explanatory variables: `(intercept)` and `age`. Write the mathematical form of the logistic regression model. Make sure to be clear how the explanatory variables and coefficients relate to the probability of being hospitalized.

(b) (5 points) Do you expect the coefficient of `age` in the previous part to be positive or negative, significant or not significant? Give an interpretation of this coefficient.

(c) Suppose the model in (a) is fitted twice: in June 2020 then again in June 2021. If the value of the coefficient of `age` stays the same, how would you expect the intercept to change between the two fittings?

(d) (5 points) Suppose you are given these data (2 datasets: June 2020 and June 2021). What would you want to learn from these data using logistic regression? Briefly describe a questions you would want to answer from these data, then describe how you would use logistic regression to answer it.

(e) (5 points) Consider a plot you would like to see to help you understand the issue you explored in the previous question. Sketch one possible version of this plot and describe what you would learn from seeing it.

(f) (5 points) Many parts of this hypothetical dataset are impossible to actually collect. Is it possible to answer your question in part (d) with data it is possible to collect? Briefly describe any limitations of practically available data sources with respect to answering your question of interest.