UNIVERSITY OF MASSACHUSETTS
Department of Mathematics and Statistics
Basic Exam - Applied Statistics
January 2021

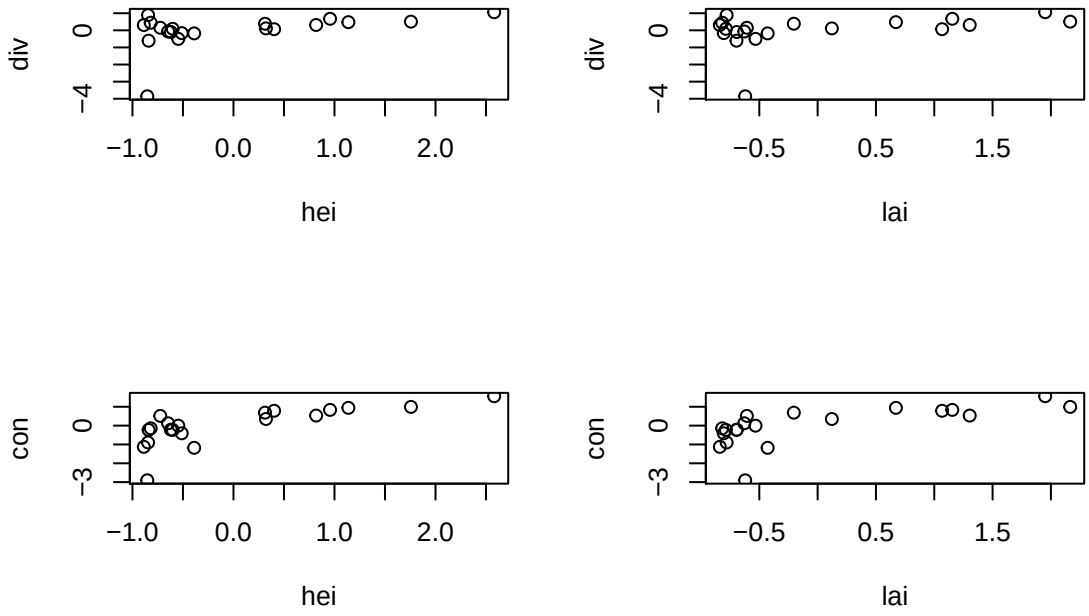
Work all problems. 60 points are needed to pass at the Masters Level and 75 to pass at the Ph.D. level. The number of points for each part of each question is listed inline with the part.

Question:	1	2	3	4	5	Total
Points:	25	30	15	10	20	100
Score:						

1. Consider a research project exploring the relationship between habitat and the coloration of birds, using the following measurements for for $i = 1, \ldots, 20$ species of birds:

   - $\texttt{div}_i$, the average color diversity of species i;
   - $\texttt{con}_i$, the average color contrast of species i;
   - $\texttt{hei}_i$, the average plant height in the habitat of species i;
   - $\texttt{lai}_i$, the leaf area index in the habitat of species i;
   - $\texttt{tem}_i$, the average temperature in the habitat of species i.

   Several exploratory plots of the data are provided below.



   Throughout, we will consider the following general linear regression model for this data.

$$\begin{pmatrix} \texttt{div}_i \\ \texttt{con}_i \end{pmatrix} \overset{indep.}{\sim} \text{Normal}\left(\begin{pmatrix} \beta_0 + \beta_1 \texttt{hei}_i + \beta_2 \texttt{lai}_i \\ \gamma_0 + \gamma_1 \texttt{hei}_i + \gamma_2 \texttt{lai}_i \end{pmatrix}, \boldsymbol{\Sigma} \right), \tag{1}$$

   where $\boldsymbol{\Sigma}$ is a 2×2 matrix with diagonal elements σ_{11} and σ_{22} that correspond to the variances of average color contrast and average color diversity respectively, and off-diagonal elements σ_{12} and σ_{21} that correspond to the covariance between average color contrast and average color diversity.

   (a) (5 points) Some `R` output for using `lm` to obtain estimates $\hat{\beta}_0$, $\hat{\beta}_1$, $\hat{\beta}_2$, $\hat{\gamma}_0$, $\hat{\gamma}_1$, and $\hat{\gamma}_2$ by regressing each response on an intercept and the two covariates separately is given below.

```
Response div :

Call:
lm(formula = div ~ hei + lai)

Residuals:
    Min      1Q  Median      3Q     Max
-3.4017 -0.1367  0.0654  0.2826  1.2568

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -8.478e-16  2.108e-01   0.000    1.000
hei          9.029e-01  6.712e-01   1.345    0.196
lai         -5.075e-01  6.712e-01  -0.756    0.460

Residual standard error: 0.9425 on 17 degrees of freedom
Multiple R-squared:  0.2051,Adjusted R-squared:  0.1116
F-statistic: 2.194 on 2 and 17 DF,  p-value: 0.1421


Response con :

Call:
lm(formula = con ~ hei + lai)

Residuals:
    Min      1Q  Median      3Q     Max
-2.2571 -0.2226  0.1329  0.3911  1.0513

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.064e-15  1.628e-01   0.000    1.000
hei          8.648e-01  5.186e-01   1.668    0.114
lai         -1.494e-01  5.186e-01  -0.288    0.777

Residual standard error: 0.7282 on 17 degrees of freedom
Multiple R-squared:  0.5256,Adjusted R-squared:  0.4697
F-statistic: 9.416 on 2 and 17 DF,  p-value: 0.001768
```

Based on this information alone, would you say that we can reject the null hypothesis $\beta_1 = 0$ at level $\alpha = 0.05$? Likewise, would you say that we can reject null hypothesis $\gamma_1 = 0$ at level $\alpha = 0.05$?

(b) (5 points) Suppose you have a collaborator on the research project, and they are interested in testing the hypothesis that the average change in color diversity given a one unit increase in average plant height, holding the average leaf area index constant and the average change in color contrast given a one unit increase in average plant height, holding the average leaf area index constant are both exactly equal to zero. Relate your response in the previous part to your collaborator's

question.

(c) (5 points) Your collaborator comes back to you and says "Maybe it would be better to test the null hypothesis $\beta_1 + \gamma_1 = 0$. I think we can do this from the output you provided! First, we can compute an estimate of $\beta_1 + \gamma_1$ by adding up $\hat{\beta}_1$ and $\hat{\gamma}_1$. Then we can compute the standard error of our estimate of $\beta_1 + \gamma_1$ by taking the square root of the squared standard errors of $\hat{\beta}_1$ and $\hat{\gamma}_1$. Then it's easy to construct an approximate 95% interval for the sum $\beta_1 + \gamma_1$, and if it does not contain 0 we reject the null! I tried it, and we get $0.9029 + 0.8648 = 1.7676$ for the sum, $\sqrt{0.6712^2 + 0.5186^2} = 0.8482$ for the standard error, and an 95% interval of $(0.1052, 3.4301)$ for the sum. This doesn't contain zero, so two species whose habitats differ in average plant height by one unit but have the same leaf area index will not have the same average color contrast and diversity on average!" Do you agree that this is correct? If you do not agree that this is correct, explain exactly what is incorrect.

(d) (5 points) An approximate 95% confidence interval for $\hat{\beta}_1 + \hat{\gamma}_1$ under the model given in (1) is $(-0.3875, 3.9227)$. Based on this information alone, can you conduct a level-0.05 test of the null hypothesis $\beta_1 + \gamma_1 = 0$? If yes, would you reject the null?

(e) (5 points) Again, relate your response in the previous part to your collaborator's original question described in (b). Be sure to explain any apparent conflicts between your collaborator's conclusion in (c).

2. Suppose that a person contacts you on behalf of a political candidate for mayor. They share with you some data on the results of a survey they conducted and they want to know your estimate of the number of people who are (A) requested a mail-in ballot and (2) did not receive a mail-in ballot.

   The town has a total of $100,000$ people. The person reports that $1,000$ phone numbers were called by telephone. Of these, $1,600$ calls were not answered, 300 were answered by a different person than was listed on the call list and thus were classified as "bad" numbers, and 100 calls were answered by the person listed on the call list and are "good" numbers. Of the 100 "good" calls 10 people reported that they requested a ballot and did not received a mail-in ballot.

   (a) (10 points) The fraction of the "good" calls that did not receive a mail-in ballot is $\hat{p} = 0.1$. Assume that the "good" call population is not only representative of the survey population of $1,000$ people, but also of the the town population of $100,000$ people. Using the finite sample Normal estimate, what is the estimate and approximate ($z_{\alpha/2} = 2.5$) confidence bounds for the number of people who did not receive a mail-in ballot? Show your work.

$$\hat{p} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n} \cdot \frac{N-n}{N-1}} \tag{2}$$

   (b) (10 points) Some of the calls in the survey population were "bad" calls. Assume that b is the fraction of the total survey population that were "bad" calls. Estimate the number of calls in the total survey population that could considered "bad".

   (c) (10 points) The person who has contacted you for your analysis has asked you to report how many people in the town requested a mail-in ballot and were not sent one. They also ask that you sign a legal document attesting to the validity of your estimate. An accurate estimate will gain you considerable recognition in the community. Explain in 1-2 paragraphs whether or not you will provide an estimate and why. If so, what your estimate and confidence limits are.

3. In this problem, you'll write a function to compute the Frobenius norm of a matrix. The Frobenius norm of a real-valued $m \times n$ matrix X is

$$\|X\|_F := \sqrt{\sum_{i=1}^{m} \sum_{j=1}^{n} x_{ij}}. \tag{3}$$

   (a) (5 points) In R or python write a function that takes one parameter `X` which is a list of lists of size $m \times n$. For example, in python you may have `X = [ [1, 2, 3], [4, 5, 6], [7, 8, 9] ]`. The function should return a scalar value. You may only use scalar product operations inside the function and you may not employ functions in special modules or libraries.

   (b) (5 points) Provide an analysis of the time complexity of the code that you wrote in the previous part to answer the question: how does the time complexity scale with the number of elements in the matrix `X`? Use Big O notation to express your results. Assume that the square root operation is $O(1)$ since the number of digits in the term under the square root is much smaller than the size of the matrix.

   (c) (5 points) Suppose that the matrix X is square ($m \times m$) and upper triangular (all elements below the diagonal are zero). What is the minimum time complexity of computing the Frobenius norm of this matrix? Justify your answer.

4. The following table summarizes the number of print magazine subscriptions that were renewed in two months, grouped by the sources of the subscriptions and also aggregated.

	Subscription Source					
Month	Gift	Previous Renewal	Direct Mail	Subscription Service	Catalog Agent	Overall
Jan						
Total	3,594	18,364	2,986	20,862	149	45,955
Renewals	2,918	14,488	1,783	4,343	13	23,545
Rate	.812	.789	.597	.208	.087	.512
Feb						
Total	884	5,140	2,224	864	45	9,157
Renewals	704	3,907	1,134	122	2	5,869
Rate	.796	.760	.510	.141	.044	.641

(a) (7 points) The overall renewal rate increases from January to February, but the rates decrease for all sources. Explain why that can happen.

(b) (3 points) If you worked at this magazine, would you be concerned with the overall rate or the rates for each source? Why?

This question is based on an article by Clifford Wagner in *The American Statistician*, February 1982, p46-48.

5. An experiment is conducted to examine the relationship between exercise and glucose metabolism. Thirty sedentary people are randomized into two groups of 15. The first group is the control group and the second group conducts a structured exercise program for 4 weeks. Each participant's glucose metabolism was measured twice, once after randomization and again four weeks later. Glucose metabolism is measured with a glucose tolerance test. That test consists of drinking a solution that is high in sugar while in a fasted state. The amount of glucose in the blood is measured 0, 30, 60, 90, and 120 minutes later. Those numbers are summarized into a single "area under the curve," and, roughly speaking, a lower number means better glucose metabolism. Let $y_{i,pre}$ and $y_{i,post}$ be those areas before and after the four weeks for participant i, and $x_i = C$ participant i is in the control group and $x_i = E$ if she's in the exercise group.

   (a) (5 points) Write down a reasonable regression model for these data.

   (b) (5 points) How would you use the model to examine if there is a relationship between exercise and glucose metabolism.

   (c) (5 points) What are three things you would do to assess the appropriateness of your model?

   (d) (5 points) Sketch a plot that you would make to summarize your results.