BIOS 830: Homework 1

Due on February, 7 2017

January 23, 2017

Instructions: Students are encouraged to work together on this problem set. However, each student is expected to *independently* write up the assigned problems. Please provide the code (R/SAS) for Questions 2b and 9 in a clean and readable document; points will be deducted for "messy code". **Assignments are to be turned in at the beginning of lecture on the due date above. Any assignments not turned in at this time will be considered late.**

Question 1:
Wood and Bormann (1974) studied the effect of acid rain on trees. "Clean" precipitation has a pH in the 5.0 to 5.5 range, but observed precipitation pH in northern New Hampshire is often in the 3.0 to 4.0 range. The investigators were interested in examining whether acid rain is harming trees, and if so, does the amount of harm depend on the pH of the rain?

One of their experiments used 240 six-week-old yellow birch seedlings. These seedlings were divided into five groups of 48 at *random*, and the seedlings within each group received an acid mist treatment 6 hours a week for 17 weeks. The five treatments differed by mist pH: 4.7, 4.0, 3.3, 3.0, and 2.3; otherwise, the seedlings were treated identically. After the 17 weeks, the seedlings were weighed, and total plant (dry) weight was taken as response.

(a) Describe the study design being used to address the question of interest. What are the treatment groups and their corresponding sample sizes?

(b) What are the experimental and measurement units for this study?

(c) Write down an appropriate statistical model for these data and describe its assumptions in the context of this study. What is the hypothesis that the investigators are interested in testing?

Question 2:
The data given in "Hypertension.csv" were collected from a study comparing the effectiveness of two treatments (Treatment A and Treatment B) for treating hypertension. Briefly, 185 male patients with hypertension, ranging in age from 50 to 70 years old, were enrolled in the study and were randomly assigned to one of the two treatment groups. Treatment was administered for a total of 12-weeks, at which point blood pressure (mg/Hg) was measured

for each subject. The investigators are interested in determining whether there is a difference in blood-pressure between the two treatment groups following the 12-week treatment period.

(a) Conduct the appropriate parametric test to address the investigator's study objective and interpret the results in the context of this study. Explain the rationale for your choice of the parametric test? What are the assumptions of your test and do they hold, at least approximately, for the data at hand.

(b) Conduct a randomization-based test to address the investigator's study objective and interpret the results in the context of this study. Provide the R/SAS code and specific details on how you implemented this test.

(c) What do you conclude about the two treatments based on the results from parts (a-b)?

Question 3:
Mice were given one of four diets at random, and the response measure was liver weight as a percentage of body weight. The responses across the four diets are given below:

Treatment			
1	2	3	4
3.52	3.47	3.54	3.74
3.36	3.73	3.52	3.83
3.57	3.38	3.61	3.87
4.19	3.87	3.76	4.08
3.88	3.69	3.65	4.31
3.76	3.51	3.51	3.98
3.94	3.35		3.86
	3.64		3.71

(a) Compute the overall mean and treatment effects.

(b) Compute the ANOVA table for these data. What would you conclude about the four diets?

Question 4:
In an ANOVA framework, we have the following full model: $Y_{ij} = \mu_i + \epsilon_{ij}$, $i = 1, \ldots, g$, $j = 1, \ldots, n_i$, and $\epsilon_{ij} \sim N(0, \sigma^2)$, which are assumed to be independent. Experimental error (variability) is measured under this full model by the sum of squares for error:

$$SSE_F = \sum_i \sum_j \left(Y_{ij} - \bar{Y}_{i.}\right)^2 = \sum_i (n_i - 1) S_i^2$$

where $S_i^2 = \sum_j \left(Y_{ij} - \bar{Y}_{i.}\right)^2 / (n_i - 1)$.

(a) Show that S_i^2 is an unbiased estimator of the error variance σ^2

(b) Derive the expectation of SSE_F.

Question 5:
(a) Show that the least squares estimator of μ is $\bar{Y}_{..}$ for the linear model $Y_{ij} = \mu + \epsilon_{ij}$ $(i = 1, 2, \ldots, g, j = 1, 2, \ldots, n_i)$, where ϵ_{ij}'s are independent normal random variables with mean zero and variance σ^2. (This is the reduced model for the one-way ANOVA test).

(b) For the model above, find an unbiased estimator for σ^2. (Hint: first calculate $E(SSE_R)$, where $SSE_R \equiv SSE_0$, represents the sums of squared error under the reduced model)

Question 6:
Show that each sums of squares in the ANOVA table can be rewritten using the regression parameterization:

Source of Variation	Degrees of Freedom	Sum of Squares	SS in Regression Notation
Treatments	$g-1$	$\sum n_i(\bar{Y}_{i.} - \bar{Y}_{..})^2$	$\boldsymbol{Y}^T\left[\boldsymbol{H} - \frac{1}{N}\boldsymbol{J}\right]\boldsymbol{Y}$
Error	$N-g$	$\sum\sum\left(Y_{ij} - \bar{Y}_{i.}\right)^2$	$\boldsymbol{Y}^T[\boldsymbol{I} - \boldsymbol{H}]\boldsymbol{Y}$
Total	$N-1$	$\sum\sum\left(Y_{ij} - \bar{Y}_{..}\right)^2$	$\boldsymbol{Y}^T\left[\boldsymbol{I} - \frac{1}{N}\boldsymbol{J}\right]\boldsymbol{Y}$

Some useful properties of $\mathbf{H}$ and $(\mathbf{I} - \mathbf{H})$ matrices:

- *Symmetric*: $\mathbf{H} = \mathbf{H}'$ and $(\mathbf{I} - \mathbf{H}) = (\mathbf{I} - \mathbf{H})'$.
- *Idempotent*: $\mathbf{HH} = \mathbf{H}$ and $(\mathbf{I} - \mathbf{H})(\mathbf{I} - \mathbf{H}) = (\mathbf{I} - \mathbf{H})$.

Question 7:
We observe random variables Y_{ij}, $i = 1, 2, \ldots, g$, and $j = 1, 2, \ldots, n_i$ where Y_{ij}'s are independent and distribution via $Y_{ij} \sim N(\mu_i, \sigma^2)$ and the total sample size is given by N, where $N = \sum_i n_i$. Suppose we are interested in testing the hypothesis that $H_0 : \sum_i c_i\mu_i = 0$ within a one-way ANOVA framework.

(a) What is the distribution of X, where $X = \frac{\sum_i c_i\bar{Y}_{i.} - \sum_i c_i\mu_i}{\sqrt{\sigma^2\sum_i c_i^2/n_i}}$ and $\bar{Y}_{i.}, i = 1, 2, \ldots, g$ are the least-squares estimators of $\mu_1, \mu_2, \ldots, \mu_g$, respectively?

(b) What is the distribution of W, where $W = SSE/\sigma^2$ and SSE denotes the sum of squared error.

Question 8: Consider the one sample t-test of the null hypothesis that $H_0 : \mu = 0$ against the two-sided alternative that $H_1 : \mu \neq 0$. The test of this hypothesis is based on the following test statistic:

$$t = \frac{\bar{X}}{s/\sqrt{n}} \tag{1}$$

When the observations are independent, the standard error of $\bar{X}$ is $\sigma/\sqrt{n}$ and is estimated by denominator of the test statistic above.

Suppose now that $X_1, X_2, \ldots, X_n$ are sequentially collected over time, such that adjacent observations have correlation ρ; observations one step apart have correlation ρ^2; and so on. More formally, this can be written as $Corr(X_i, X_j) = \rho^r, |i-j| = r$ where $i, j = 1, 2, \ldots, n$. This is called a first order autoregressive process AR(1) with correlation ρ.

(a) Assuming that that variance of X_i is equal to σ^2 for all $i = 1, 2, \ldots, n$, show that:

$$\text{standard error of } \bar{X} \approx \sqrt{\frac{1+\rho}{1-\rho}}\frac{\sigma}{\sqrt{n}}$$

(b) How would the Type 1 error rate be affected if $\rho > 0$ assuming we were to use the one-sample t-test for these data?

(c) How would the Type 1 error rate be affected if $\rho < 0$?

Question 9: Similar to Question 8, assume that $X_1, X_2, \ldots, X_n$ are sequentially collected over time, such that:

$$Corr(X_i, X_j) = \rho^r, |i-j| = r, \text{ where } i, j = 1, 2, \ldots, n$$

Conduct a simulation study to estimate the Type 1 error rate for the one-sample t-test for the values of ρ given in the table below.

ρ	$\sqrt{\frac{1+\rho}{1-\rho}}$	Type 1 error
-0.50	0.58	
-0.25	0.77	
-0.10	0.90	
0.00	1.00	
0.10	1.11	
0.25	1.29	
0.50	1.73	

For your simulation study, assume that $n = 10,000$ and $X_i \sim N(0, 25)$.