Statistics 500 - Homework # 8, Fall 2020
Due: by noon Friday, 10/23/2020

Reading Assignment: Statistical Sleuth, Chapters 7, 8 (Simple Linear Regression) and 9 (Multiple Regression). PACQ, Review Appendices A and B and read Chapter 1

1. In a study of the effects of exposure to UV-B radiation on egg hatch rates for three species of frogs, eggs for three species of frogs were collected from two different locations (Three Creek and Sparks Lake). Thirty-six enclosures were constructed at leach location. Within each location, four enclosures were randomly assigned to each of the 9 combination of the two factors: frog species (*Hyla regilla*, *Rana cascade*, and *Bufo boreas*) and type of radiation filters (no filter, UV-B transmitting filter, and UV-B blocking filter). One hundred and fifty eggs for the designated frog species were placed in each enclosure. The response is the percentage of eggs that failed to hatch in each enclosure. Simple SAS code containing the data is posted as **frogeggs.sas** to get you started. The observed data (percentage of eggs that failed to hatch) are also displayed in the following tables:

<u>Data for Three Creek</u>

	Factor B (Frog Species)		
Factor A	*Hyla regilla*	Rana cascadae	Bufo Boreas
(Type of filter)	(j=1)	(j=2)	(j=3)
No filter (i=1)	6.0	38.7	42.0
	4.7	44.0	50.7
	0.7	30.0	32.7
	5.2	38.7	44.0
UV-B transmitting filter (i=2)	0.9	28.7	47.3
	6.7	32.7	22.0
	2.7	36.0	37.3
	0.7	40.7	43.3
UV-B blocking filter (i=3)	4.7	25.3	18.7
	0.7	18.7	17.3
	4.7	21.3	16.0
	0.7	16.7	4.7

<u>Data for Sparks Lake</u>

	Factor B (Frog Species)		
Factor A	*Hyla regilla*	Rana cascadae	Bufo Boreas
(Type of filter)	(j=1)	(j=2)	(j=3)
No filter (i=1)	1.5	36.7	54.0
	0.8	69.6	54.7
	2.9	39.3	48.0
	3.9	34.0	36.7
UV-B transmitting filter (i=2)	0.7	70.0	46.0
	2.1	54.0	46.7
	0.0	48.7	36.0
	1.4	51.3	35.3
UV-B blocking filter (i=3)	4.5	24.7	12.7
	0.0	25.3	17.3
	0.0	39.3	31.3
	0.0	32.7	17.3

a. What is the treatment design in this study and what is the experimental design in this study?

b. Consider the model $Y_{ijk\ell} = \mu + \alpha_i + \tau_j + (\alpha\tau)_{ij} + \beta_k + \varepsilon_{ijk\ell}$ where $\varepsilon_{ijkl} \sim N(0, \sigma_\varepsilon^2)$ are random errors, $\beta_k \sim N(0, \sigma_\beta^2)$ are random block effects corresponding to locations, and any random error is independent of any random block effect. SAS imposes the constraints $\alpha_3 = \tau_3 = (\alpha\tau)_{13} = (\alpha\tau)_{23} = (\alpha\tau)_{33} = (\alpha\tau)_{31} = (\alpha\tau)_{32} = 0$. With respect to the mean response (expected proportion of eggs that fail to hatch), how should the following parameters be interpreted?

(i) μ

(ii) α_1

(iii) τ_2

(iv) $(\alpha\tau)_{12}$

(v) $\mu + \alpha_1 + \tau_2 + (\alpha\tau)_{12}$

(vi) $(\alpha\tau)_{12} - (\alpha\tau)_{32} - (\alpha\tau)_{13} + (\alpha\tau)_{33}$

b. Complete the following ANOVA table:

Source of variation	d.f.	Sums of Squares	Mean Square	F	p-value
Locations					
Filters					
Species					
Filter × Species Interaction					
Error (residuals)					
Corrected Total					

c. Examine a profile plot of the treatment means (do not hand it in), plotting the sample means for the combinations of filters and frog species, averaging across locations, against the frog species. What does this plot suggest? Are your conclusions about interactions between types of filters and frog species supported by results in the ANOVA table?

d. Is there any evidence that the types of filter have different effects on egg hatch success? Explain.

e. Examine the plot of the residuals against the estimated mean responses. What does this plot indicate?

f. Examine a q-q normal probability plot of the residuals. What does this plot indicate?

g. If the residual plots suggest a need for a transformation, find which transformation of the responses is better, square root transformation or log transformation? Then compute a new ANOVA table for transformed responses using the transformation of your choice. If the residual plots did not indicate that a transformation is necessary, then there is no need to re-do the analysis. Write a summary of your conclusions for this analysis.

2. One factor that may explain the price of a diamond is the weight of the diamond. Data were collected for a sample of 46 diamonds, including the weight in grams (g) and the price (in Singapore dollars) of each diamond. These data are located in the file **diamonds.csv**.

a. For the simple linear regression model $Y_i = \beta_0 + \beta_1 x_i + \epsilon_i$, give the interpretation of the parameter values β_0, β_1, and σ^2 in the context of the response and explanatory variables.

b. Write the simple linear regression model for this problem in vector-matrix notation. Give the first 4 rows of the design matrix X.

c. Describe the scatterplot of the weight and price of the 48 diamonds in this sample. What do you notice about the relationship between these two values?

d. Calculate the sample correlation coefficient between the weight and price of the diamonds. How does the value of your correlation reinforce your description from part (c) above.

e. Give the equation for the least squares regression line to predict the price of a diamond from its weight.

f. Use the least squares regression line to predict the price of a diamond with a weight of 0.2 grams and the price of a diamond with a weight of 0.28 grams.

g. In the sample, one diamond had a weight of 0.2 grams and a price of $498 and another diamond had a weight of 0.28 grams and a price of $823. Find the residuals for both observations.

h. Give the ANOVA Table for this simple linear regression. Use the ANOVA Table to conduct a test of significance for the linear regression model.

i. Give the value of R^2 for this simple linear regression. Give an interpretation of this value.

j. Calculate a 95% confidence interval for the slope parameter in the simple linear regression model. Give an interpretation of this interval.

k. Obtain a confidence interval for the conditional mean price of all diamonds in the population with a weight of 0.2 grams. Give the interpretation of this interval.

l. Use the Scheffe method to obtain a formula for a simultaneous 95% confidence band for mean hardness, $E(Y \mid x) = \beta_0 + \beta_1 x$, between 0.12 and 0.35 grams. Display this confidence region on a plot along with the least squares estimate of the regression line.

m. Obtain a prediction interval for the price of a diamond in the population with a weight of 0.28 grams. Give the interpretation of this interval.

3. (optional, SAS code provided) The data shown in the table below are results from a study of amylace activity of malted wheat flour (Geddes, et al, 1941, *Cereal Chem* 18, 42-60.). Five factors, each at two levels, were examined:

Factor A: type of wheat (Amber durum (a), hard red spring (A))
Factor B: wheat protein content (low (b), high (B))
Factor C: wheat moisture content (40 percent (c), 44 percent (C))
Factor D: germination time (3 days (d), 5 days (D))
Factor E: kiln temperature (rising 100° F to 130° F (e), constant at 100° F (E))

Amylase activity was measured by the amount of malt from each flour that was required to produce 204.7ml of CO_2 . Measured amylase activity is reported in the following table in units of
$\mathrm{Y} = (0.6 + \log(\text{amount of malt})) \times 10^3$ SAS code containing the data is posted as **amylace.sas**.

	Yield		Yield		Yield		Yield
abcde	732	abcDe	200	abcdE	744	abcDE	253
Abcde	801	AbcDe	50	AbcdE	732	AbcDE	91
aBcde	717	aBcDe	292	aBcdE	713	aBcDE	265
ABcde	791	ABcDe	74	ABcdE	746	ABcDE	147
abCde	616	abCDe	62	abCdE	569	abCDE	80
AbCde	787	AbCDe	83	AbCdE	785	AbCDE	80
aBCde	540	aBCDe	97	aBCdE	486	aBCDE	102
ABCde	669	ABCDe	-9	ABCdE	544	ABCDE	-40

a. Use the posted SAS code to construct a normal probability plot of estimates of main effects and interaction contrasts, for which the estimate of every contrast has the same variance. Which effects appear to be large?

b. Use least squares estimation to fit the model that includes all main effects and all interaction effects that were identified as “non-zero” by the analysis in part (a). Include all main effects in this model, regardless of whether the plot suggests they are significant or not. The sum of sums of squares for the interaction contrasts that are not included in the model can be pooled to obtain a $\mathrm{MS}_{\mathrm{error}}$. Construct an ANOVA table and examine the results of F-tests for terms you kept in the model. State your conclusions.

c. Interpret any significant interactions.

d. Comment on the normal probability plot of the residuals for the model in part (b).

e. Comment on the plot of the residuals versus the estimated mean yields for the model in part (b).

f. Give a summary of the effects of the five factors on amylase activity. Keep in mind that low values of the response variable correspond to combinations of factors that produce 204.7 ml of CO2 with the least amount of wheat.