Stat 500 **FINAL EXAM** **Fall 2020**

Due Date: Tuesday, November 24, at 2 pm. You may submit your paper any time before the due date and time to Canvas. **Late papers will not be accepted.**

Instructions: You must do your own work on this exam. You may not discuss this exam with anyone except the course instructor or the course TA. You may use SAS to perform calculations or construct graphs except for questions that require hand calculation. You should **not** submit pages of *unedited* computer output, but you may include specific tables or graphs in your report. Write concise answers that clearly describe the steps in your analysis and your conclusions. There are three problems, and the full score is 100 points.

Problem 1 (20 points):

The fuel that powers an automobile engine may corrode the gaskets when it comes in contact with the gaskets. (Gaskets are parts inside the engine.) The amount of corrosion may depend on the composition of the fuel, which consists of a mixture of gasoline, ethanol and methanol. The amount of corrosion may also depend on the material used to make the gaskets. To study this, data on corrosion were obtained on gaskets made from three different materials (called material A, material B and material C). This was done by constructing eight engines: two were constructed with gaskets made from material A, four were constructed with gaskets made from material B, and two were constructed with gaskets made from material C. The engines were run for an equal length of time using fuels with the levels of ethanol and methanol shown in the following table. Then, the gaskets were removed from the engines and the amount of corrosion was measured.

Table 1. Structure of the Data for Problem 1

Gasket Material	Percent ethanol (X_{1ij})	Percent methanol (X_{2ij})	Observed Corrosion (Y_{ij})
A	0	5	Y_{11}
A	5	5	Y_{12}
B	0	0	Y_{21}
B	0	0	Y_{22}
B	5	0	Y_{23}
B	5	0	Y_{24}
C	0	10	Y_{31}
C	5	10	Y_{32}

Suppose that the levels of corrosion (Y_{ij}) is modeled as follows:
$Y_{ij} = \mu + \alpha_i + \beta_1 X_{1ij} + \beta_2 X_{2ij} + \varepsilon_{ij}$ where $i = 1, 2, 3$ corresponds to the Gasket Material being A, B, C, and ε_{ij} are independent random errors with $\varepsilon_{ij} \sim N(0, \sigma^2)$.

Let $\boldsymbol{Y} = (Y_{11,}\, Y_{12}, Y_{21,}\, Y_{22}, Y_{23,}\, Y_{24}, Y_{31,}\, Y_{32})^T$, X_{1ij}, X_{2ij} are defined as in the column headings in Table 1, $\boldsymbol{\varepsilon} = (\varepsilon_{11,}\, \varepsilon_{12}, \varepsilon_{21,}\, \varepsilon_{22}, \varepsilon_{23,}\, \varepsilon_{24}, \varepsilon_{31,}\, \varepsilon_{32})^T$, and $\boldsymbol{\beta} = (\mu, \alpha_1, \alpha_2, \alpha_3, \beta_1, \beta_2)^T$ is the parameter vector.

a) (4 points) Write out the design matric **X**.
b) (2 points) Is this model a linear model?
c) (2 points) Is this model a Gauss-Markov model?
d) (2 points) What is the distribution of the vector $\boldsymbol{Y}$?
e) (2 points) What is the rank of matrix **X**?
f) (8 points) Determine which of the following are estimable. Justify your answer.

(i) $\mu + \alpha_1$ (ii) β_1 (iii) β_2 (iv) $\alpha_3 - \alpha_1 + 5\beta_2$

Problem 2 (40 points):

An experiment was conducted to study the effects of water and fertilizer amount on plant height. Twelve pots, each holding one plant, were available for this experiment. These pots were randomly assigned to three different levels of water: well-watered (level L1), half-watered (level L2, half of L1 amount), and water-stressed (level L3) such that four pots were assigned to each water level. Then, four amounts of nitrogen fertilizer (5, 10, 15, and 20 units) were randomly assigned to the four pots of each water level. Plant height (in cm) was recorded for each plant at the end of the experiment. Part of the data set is provided in Table 2. *Use hand calculation for this problem.*

Table 2: Plant Height Data

Pot	WaterLevel	FertilizerAmount	Height
1	L1	5	124
2	L1	10	130
3	L1	15	136
4	L1	20	143
.	.	.	.
.	.	.	.
.	.	.	.
12	L3	20	148

Let Y_k be the observed plant height for the *k*-th pot for *k=1, …, 12*. Let

$$X_{1k} = \begin{cases} 1 & \text{if WaterLevel is L2 for the } k\text{-th pot} \\ 0 & \text{if WaterLevel is not L2 for the } k\text{-th pot} \end{cases}$$

$$X_{2k} = \begin{cases} 1 & \text{if WaterLevel is L3 for the } k\text{-th pot} \\ 0 & \text{if WaterLevel is not L3 for the } k\text{-th pot} \end{cases}$$

and X_{3k} = FertilizerAmount (5, 10, 15, or 20) applied to the *k*-th pot.

A statistics graduate student would like to use multiple linear regression method to analyze the data. The first model that the student fit is: $Y_k = \beta_0 + \beta_1 X_{1k} + \beta_2 X_{2k} + \beta_3 X_{3k} + \varepsilon_k$ Model (1)

where ε_k are i.i.d. $N(0, \sigma^2)$. **Partial output from SAS for fitting Model (1) is on the next page.**

a) (2 points) What is the estimated expectation of plant height as a function of FertilizerAmount for plants from WaterLevel L3?
b) (3 points) Interpret the estimated value of parameter β_3 in the context of this study.
c) (4 points) Test for the effect of FertilizerAmount on plant height based on Model (1). Report your test statistic, degrees of freedom, and *p*-value. Give your conclusion in the context of this study.
d) (5 points) Based on the fitted regression model, the 95% confidence interval for the conditional mean height of plants with WaterLevel L2 and FertilizerAmount = 10 is (127.44, 133.16). Construct a 95% prediction interval for height of a plant with WaterLevel L2 and FertilizerAmount=10.

Partial SAS output for fitting Model (1)

Number of Observations Read	12
Number of Observations Used	12

Analysis of Variance					
Source	**DF**	**Sum of Squares**	**Mean Square**	**F Value**	**Pr > F**
Model					0.0001
Error			5.76250		
Corrected Total		529.00000			

Parameter Estimates					
Variable	**DF**	**Parameter Estimate**	**Standard Error**		
Intercept	1	119.75000	1.96002		
x1	1	-0.25000	1.69742		
x2	1	4.00000	1.69742		
x3	1	1.08000	0.12396		

Then, the student applied a second regression model to the same dataset:

$$Y_k = \beta_0 + \beta_1 X_{1k} + \beta_2 X_{2k} + \beta_3 X_{3k} + \beta_4 X_{1k} X_{3k} + \beta_5 X_{2k} X_{3k} + \varepsilon_k \qquad \text{Model (2)}$$

where ε_k are i.i.d, $N(0, \sigma^2)$. **Relevant SAS output from Model (2) is on pages 4 and 5**.

e) (3 points) What is the estimated expectation of plant height as a function of FertilizerAmount for plants from WaterLevel L3 based on Model (2)?

f) (4 points) Is there any significant difference in the slope of regression lines between WaterLevel L1 and WaterLevel L2? State your null and alternative hypotheses, report your test statistic, degrees of freedom, p-value, and your conclusion.

g) (4 points) Is there any significant difference in the slope of regression lines among water levels L1, L2, and L3? State your null and alternative hypotheses, report your test statistic, degrees of freedom, p-value, and your conclusion.

h) (4 points) The diagnostic plots generated from SAS while fitting Model (2) are on page 5. Based on these plots, give your conclusion regarding Model (2) assumptions.

i) (3 points) Among the assumptions that are needed for the multiple linear regression analysis for Model (2), which ones cannot be checked by the diagnostic plots? Discuss whether those assumptions are plausible or not for this experiment.

j) (4 points) Which regression model would you choose, Model (1) or Model (2)? Justify your answer with your choice of model selection criterion and its numerical values for the two models.

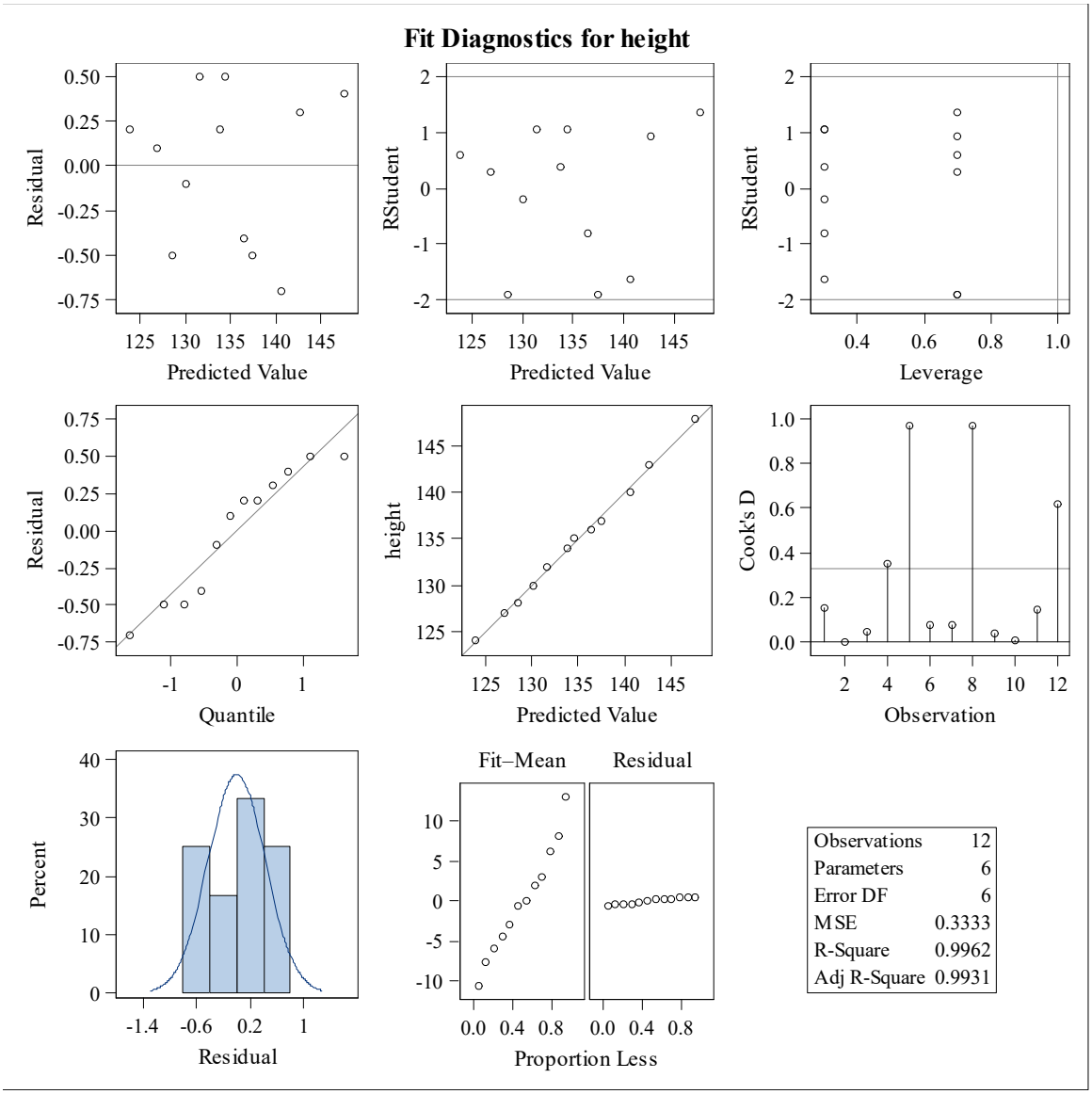
k) (4 points) Suppose that we have two pots for each combination of WaterLevel and FertilizerAmount and hence a total of 24 observations. We would like to check the lack of fit for Model (2). What are the degrees of freedom for the F statistic to test for lack of fit of Model (2)?

Partial SAS output for fitting Model (2)

Number of Observations Read	12
Number of Observations Used	12

Analysis of Variance					
Source	**DF**	**Sum of Squares**	**Mean Square**	**F Value**	**Pr > F**
Model				316.20	<.0001
Error			0.33333		
Corrected Total	11	529.00000			

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	117.50000	0.70711	166.17	<.0001
x1	1	8.00000	1.00000	8.00	0.0002
x2	1	2.50000	1.00000	2.50	0.0465
x3	1	1.26000	0.05164	24.40	<.0001
x1x3	1	-0.66000	0.07303	-9.04	0.0001
x2x3	1	0.12000	0.07303	1.64	0.1515



Problem 3 (40 points):

Insecticides for killing insects that live in soil are often applied to farmland or lawns in the form of granules over the ground. Some species of ground feeding birds may inadvertently pick up and ingest the granules. An experiment was performed by animal ecologists to determine if coloring the granules could help deter birds of a particular species (let's call it species A) from ingesting such granules. The researchers constructed a set of 27 pens, and each pen would host one bird at a time. Each bird could freely move around inside its pen. The birds were fed by scattering grain across the ground within each pen.

Bird of species A were randomly assigned to the pens. During the first two days the birds were in pens, only food was scattered across the ground inside each pen. During the next three days, granules were scattered across the ground along with the food in each pen. Nine types of granules, corresponding to three colors and three sizes, were used in this experiment and they were randomly assigned to pens so that each of the nine combinations was assigned to exactly three pens. The granules contained a substance that causes the level of a blood enzyme to increase when it is ingested by the birds. Larger intakes of this substance result in greater increases in the concentration of the blood enzyme.

A blood sample was taken from each bird just before it was placed in its pen. Then, a second blood sample was taken from each bird three days after the granules were introduced into the pens. The increase in the measured concentration of the blood enzyme between the first and second blood samples is the response for each bird.

After the second blood sample was taken, the birds were released from the pens, the pens were cleaned, and the experiment was repeated with a new set of 27 birds that were randomly picked from species A. The experiment was replicated eight times. Results could vary across replicates because changes in weather may affect the behavior of the birds. In very hot weather, for example, the birds may be less active and eat less.

The increase in the measured concentration of the blood enzyme between the first and second blood samples is the response recorded in the fifth column of the data file as shown in Table 3. The first column contains a bird identification number that is unique for each bird. The second column contains the replication number, coded as 1,2,3,4,5,6,7 or 8. The third column indicates the particle color (1=blue, 2=yellow, and 3=natural) and the last column indicates the granule size coded as 1=small (about 1mm in diameter), 2=medium (about 2mm in diameter) and 3=large (about 3mm in diameter). The data are posted as **granules.txt**. There is one line in the data file for each bird and the values of the variables are presented in the order shown in Table 3. This table also shows the first line and the last line of the data file. SAS code for entering and printing the data is posted as **granules_partial.sas**.

Table 3. Structure of the Data for Problem 3

Bird	Rep	Color	Size	Blood Enzyme Increase (Y)
1	1	1	1	79
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
216	8	3	3	380

The researchers wanted to examine (i) the main effects of color and size on increase in the measured concentration of the blood enzyme, (ii) whether the effects of color depend on size or not, (iii) in particular, whether the difference between blue and natural color depends on the size of granules or not, (iv) which treatment group(s) results in the smallest increase in the measured concentration of the blood enzyme, and (v) whether the replication effect should be considered when designing future experiments.

a) (4 points) Identify the design of this experiment and identify experimental units.
b) (6 points) Describe the major steps taken in your analysis. Comment on what you did to analyze these data and how it led you to the final analysis reported in the next section. This should not exceed one typewritten pages, excluding graphs and tables.
c) (10 points) A detailed report of your final analysis. You should give a description of the model that provides a basis for your analysis, *including model assumptions.* You should provide evidence (e.g., in the form of graphs, tests, diagnostic methods) to support that the final model and corresponding methods of analysis that you selected are appropriate for these data. Excluding graphs, this should not exceed two typewritten pages.
d) (20 points) Present your results and conclusions to answer each of the five questions the researchers were interested in. Show relevant calculations or SAS output. Describe your results in the context of the study.