STATISTICS 210 – Fall 2018

Final Exam: Thursday December 13, 2018

- The exam is closed book, closed notes. You may use a calculator. Tables and supporting output are provided separately. Problem values are written to the left of each problem. The total number of points is equal to 160.

- Important reminders:
  - please write your student identification number on each page of your solutions;
  - show your work so that you can receive partial credit;
  - keep explanations brief;
  - budget your time carefully.

- Please do not write your solutions on the exam paper.

- Good luck!
1. **MESQUITE PLANTS** - Mesquite is a type of tree native to the southwestern U.S. and Mexico. A mesquite farm carried out a study to develop a model that could predict the net weight of harvestable mesquite leaves from quantities that could be easily measured before harvesting. There are 3 pages of associated output. The variables that appear on page 1 of the output are:

- `totwt` = total weight (grams) of harvested leaves
- `canopy1` = canopy diameter (meters) - short axis
- `canopy2` = canopy diameter (meters) - long axis
- `totht` = total height of plant (meters)
- `canht` = canopy height of plant (meters)
- `density` = plant density (stems per branch)
- `time` = indicator for harvest season (0 or 1)

The canopy is the part of the plant containing leaves. The total height of the plant includes the trunk of the tree and the leaves; the canopy height refers to the height of the portion containing the leaves.

Page 1 of the output contains summary statistics for the variables, a correlation matrix, and some output for regression Model 1 which is a regression of total weight on plant density, the two height variables, and the two canopy diameter variables. At the bottom of the page are residual plots for Model 1.

5. (a) We can tell that the distribution of total weight in this sample of trees is not a normal distribution because total weight is non-negative and the standard deviation is greater than the mean. This does not mean that the regression assumption of normality is violated. Explain.

5. (b) Provide an interpretation for the regression coefficient of "canopy1" in Model 1.

9. (c) Which of the regression coefficients are significantly different from zero at the .05 level?

7. (d) It is natural to wonder whether the population coefficients of canopy1 and canopy2 are equal. Can you test the hypothesis with the information provided? If so, test it; if not, tell how you could carry out this test.

6. (e) The output includes two residual plots for Model 1: a normal probability plot of the standardized residuals and a plot of the standardized residuals versus the fitted values. Address the validity of the regression assumptions based on these plots.

10. (f) The second page of the output contains case regression diagnostics for Model 1. A preliminary analysis suggests that observations 3, 27, 28, 35, 40 may deserve further attention. For each of these cases do a quick analysis - briefly describe what aspects of the observation (if any) are unusual.

5. (g) The third page of output contains results for Model 2, a regression using transformed variables. The logarithm of total weight has been regressed on the logarithms of the predictors. Tell what the resulting transformed regression model suggests about the relationship among the original untransformed variables.

8. (h) Suppose that we want to predict the total weight of harvestable leaves from a tree that has the same dimensions as the first tree in our sample. The fitted value from Model 2 (for log total weight) is 5.6946. A 95% confidence interval for the mean log total weight for a population of such trees is (5.5473, 5.8419). Based on this information, provide a 95% prediction interval for the total weight of harvestable leaves?
2. **CLOUD SEEDING** - Cloud seeding is a technique whereby chemicals are introduced into clouds on days with the potential for rain in order to increase rainfall. A randomized experiment was carried out in South Florida to assess the effectiveness of cloud seeding. Each day a weather prediction model was used to provide a predicted rainfall amount in millions of cubic meters (PRED). If that predicted amount was above a threshold, then that day was suitable for inclusion in the experiment. The day was then randomly assigned to one of three treatment groups (control – no chemicals used, treatment 1 – chemical compound 1 used, treatment 2 – chemical compound 2 used).

The questions below refer to the two output pages labelled "CLOUD SEEDING" (these appear on pages 4 and 5 of the supplied output). On the output TREAT1 is an indicator variable which is equal to 1 when treatment 1 was used and 0 otherwise and TREAT2 is an indicator variable which is equal to 1 when treatment 2 was used and 0 otherwise. The predicted rainfall amount is denoted as PRED and the response variable is RAIN (rainfall in the 24 hours following the seeding in millions of cubic meters). The top of the output contains some summary information on PRED and RAIN in the three groups.

10 (a) Model 1 is a regression of RAIN on PRED that allows for a separate regression line for each of the three treatment groups.
   i. Identify the fitted regression line for the treatment 2 group.
   ii. The overall regression is significant ($F=5.62, p = .0003$) but none of the predictors are significant. Explain how this can happen.

8 (b) The residual plot for Model 1 (not shown) suggested non-constant variance. As a result it was decided to use LOGRAIN (the logarithm of rainfall) as the response and LOGPRED (the logarithm of predicted rainfall) as a predictor. Models 2, 3, and 4 are regressions using the transformed variables. Model 2 fits a separate regression line for each of the three groups. Provide an interpretation for Model 3 and Model for 4.

12 (c) We would like to test whether Model 2 is required versus the null hypothesis that Model 3 is adequate. Specify the null hypothesis in terms of the parameters of Model 2 and perform a test of the null hypothesis. Report the $p$-value and state your conclusion.

5 (d) The $p$-value for TREAT:LOGPRED in Model 2 is .0776. Carefully and precisely explain what the $p$-value means. (This question is not asking about statistical significance; a correct answer will give a precise definition of this $p$-value.)

5 (e) Model 5 is a regression model that is equivalent to a traditional analysis of variance for these data. Note that Model 5 omits LOGPRED as a predictor even though Models 3 and 4 indicate that LOGPRED is a significant predictor. Despite this, Model 5 is still a valid approach to these data because treatment was randomly assigned on each day of the experiment. Explain.

5 (f) A reviewer of the study is concerned that there was an error in the randomization. Examine the table of means for the three treatment groups (back on the first page of the output for this problem) and comment on the reviewer's concern.

5 (g) The researcher who collected the data decides to use Model 2 because it has the highest $R^2$. Explain why this is usually a poor approach to choosing a model.

5 (h) You are concerned that this model has too many parameters and is overfit to these particular data. A new set of cloud seeding experiment data using the same chemicals is available to validate the model. How would you use the data from this new experiment to assess whether the model is overfit? Be specific.
3. MORTAR PRODUCTION - Mortar is a building material made by mixing water and plaster and then curing (hardening) the mixture. An experiment was carried out to compare the effectiveness of two types of water (soft and hard) and three curing times (3 days, 7 days, 28 days). Nine mortar samples were prepared using each of the six possible methods (e.g., soft water and 3 days). The nine samples created using each combination were evaluated for tensile strength (larger values indicate stronger mortar). The mean strengths for the six methods are provided in the table below. The overall mean is 50.

<table>
<thead>
<tr>
<th>Curing time</th>
<th>Water type</th>
<th>3 days</th>
<th>7 days</th>
<th>28 days</th>
</tr>
</thead>
<tbody>
<tr>
<td>hard</td>
<td>47</td>
<td>56</td>
<td>59</td>
<td></td>
</tr>
<tr>
<td>soft</td>
<td>36</td>
<td>44</td>
<td>58</td>
<td></td>
</tr>
</tbody>
</table>

(a) For analysis purposes we treat the curing time as a categorical variable. Let $Y_{ijk}$ denote the tensile strength of the $k$th mortar sample prepared with water type $i$ and curing time $j$. Write down a probability model for $Y_{ijk}$ that is appropriate for an analysis of variance for these data. Be sure to explain what each term represents.

(b) Explain how the analysis of variance model can be fit using linear regression. In other words, explain how you would define the matrix $X$ and the vector $\beta$ so that you could use the linear model $Y = X\beta + \epsilon$ to estimate the model you specified in (a).

(c) The analysis of variance summary table produced for these data is provided here.

<table>
<thead>
<tr>
<th>Source</th>
<th>d.f.</th>
<th>SS</th>
<th>MS</th>
</tr>
</thead>
<tbody>
<tr>
<td>water type</td>
<td>1</td>
<td>864</td>
<td></td>
</tr>
<tr>
<td>curing time</td>
<td>2</td>
<td>2601</td>
<td></td>
</tr>
<tr>
<td>interaction</td>
<td>2</td>
<td>333</td>
<td></td>
</tr>
<tr>
<td>error</td>
<td>48</td>
<td>8880</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>53</td>
<td>12678</td>
<td></td>
</tr>
</tbody>
</table>

i. Calculate and interpret the mean squared error.
ii. Which effects are significant at the .05 level.

(d) Consider the six treatment combinations as six separate groups.

i. Provide a 95% confidence interval for the contrast specified by the contrast weights $c = (-0.5, -0.5, 1, -0.5, -0.5, 1)$ where the first three weights are for the treatment combinations in the first row of the table and the next three are for the second row.

ii. Give a one sentence description of the resulting interval using language that you might use in the results section of a scientific paper.

(e) An alternative analysis of these data would use a regression model treating curing time as a continuous random variable and introducing an indicator variable for water type. Identify one advantage of the regression approach relative to the analysis of variance approach and one disadvantage.