1. [15pts—3pts each part] Vinho Verde (literally “green wine” in Portuguese) is a wine produced in the Minho region of Portugal. A producer of this wine is interested in modeling the quality of the wine (rated on a scale of 0-10 with 0 being very bad and 10 being excellent) for two variants: red wine and white wine. The company has data from over 6000 wine samples and want to construct a model to predict wine quality. For each wine sample, multiple tasters provide a rating and the quality of wine reported in the data set for a sample is the median of the ratings. The data set has a total of 12 possible variables. For brevity the variables are labeled as x1 through x12.

   a. One option is to use the lm() function in R to model quality as a function of the explanatory variables. When using the lm() function, what distributional assumption is being made regarding the error term (and hence the response)? Briefly explain if it reasonable to assume that this is satisfied for this example. If there is additional information that you would need to assess this, indicate what is needed and why.

   b. In the output packet is some output from an all subsets regression using BIC, Mallow’s Cp, and adjusted R-squared as the criterion. Based on this output, which variables would you recommend including in your model to predict quality? Briefly explain your choice. Regardless of your answer to part a, you may assume that there are no concerns with assumption violations, outliers, or influential observations in the models under consideration.

   c. Another option is to use stepwise selection methods (e.g. forward selection, backward elimination, and stepwise selection) for model selection. Suppose that stepwise selection using BIC were used to select a model. Would the “best” model chosen using stepwise selection with BIC as the criterion be the same as the model chosen using all subsets regression with BIC as the criterion? Why or why not?

   d. After meeting with the company, a discussion ensues. The company would prefer to use a model with variables that are relatively cheap and easy to measure. They would prefer to use a model with x2, x4, x6, x10, x11, and x12. This model has an adjusted r-squared of 0.2891, Cp of 97.5, and a BIC of -2151. Further inspection indicates no issues with model assumptions, outliers, or influential observations. You need to recommend one model. Make a recommendation and explain your choice.

   e. Regardless of your answer in the previous parts, the model proposed by the company is considered (see output below for model m3). Using this model, provide an interpretation for the slope coefficients for x11 (alcohol—measured as percent of volume that is alcohol such that 8.4 means that the wine is 8.4% alcohol) and for x12 (wine variant) in context of the problem.
2. [38pts] In a study by Van Os et al. (2018), 12 Angus-Hereford cross heifers (each approximately one year old) were randomly assigned to consume either a high-roughage or low-roughage diet (six heifers per diet) which was placed in an open trough. For each heifer, the amount of time (in minutes) until the heifer started eating the hay in the open trough, called latency, was recorded. Van Os et al. (2018) hypothesize that heifers in the low roughage diet will show a shorter latency than heifers in the high-roughage diet.

a. According to the researchers, each heifer was individually housed in a randomly assigned pen and the set-up of each pen was identical. Briefly explain why the researchers would choose to do this.[2pts]

b. Based on this problem description which statistical method (e.g. simple linear regression, logistic regression, etc) do you recommend the researchers use to address their hypothesis? Briefly explain your choice in no more than two sentences.[2pts]

c. In the actual study, each heifer was measured for five days (with day labeled as -4, -3,..., 0). Does this change your answer to the previous part? If so, how? If not, why not?[3pts]

In the accompanying output packet is some output that may or may not be helpful in answering the remaining questions

d. In their analysis (replicated in the output packet), the researchers used a mixed model with day, treatment, and day*treatment as fixed effects and heifer as a random effect.

i. Briefly explain why the researchers would choose to use day and treatment as fixed effects and heifer as a random effect.[3pts]

ii. Write out the model of interest to the researchers, noting any relevant assumptions. Note that the researchers chose to (natural) log transform latency.[3pts]

iii. Derive the covariance between two log(latency) values on two different days for the same heifer.[4pts]

iv. Write out the covariance matrix of the responses (log(latency)) making sure to denote dimensions as needed. Note that the researchers used a homogeneous variance components covariance structure.[2pts]

v. What distribution does an individual value of log(latency) follow? Using known results from Casella and Berger, clearly identify the distribution and parameter(s) of the correct distribution. You do NOT need to derive the distribution—instead use relevant theorems.[3pts]

e. Based on the model fit by the researchers, does the variability across days (for a heifer) or the variability across heifers contribute more to the overall variability of log(latency)? Justify your answer.[2pts]

f. What is the estimated correlation between log(latency) values on two different days for the same heifer? Between log(latency) values on the same day for two different heifers?[2pts]

g. The researchers initially conduct a test of the treatment*day interaction.

i. Explain why the researchers may be interested in testing this hypothesis.[2pts]

ii. The degrees of freedom for this test are 1 and 33.4. The researchers are concerned because they were taught that degrees of freedom are always integer values. Why is this not the case here?[2pts]

iii. Summarize the results of this hypothesis test in one sentence in context of the problem.[2pts]

h. Report the following:

i. The estimated regression line for the low-roughage group.[1pt]

ii. The estimated regression line for the heifer with the ID 3088 (who was in the low-roughage group).[2pts]

iii. Explain why the estimated regression lines in the previous two parts are not the same.[3pts]
3. A researcher is interested in traffic flow and collects data during the same six days of two consecutive weeks. For each day, a researcher randomly selects a 15-minute interval and counts the number of vehicles that travel through a section of a highway. The number of vehicles observed are presented in the table below.

<table>
<thead>
<tr>
<th>Week</th>
<th>Monday</th>
<th>Tuesday</th>
<th>Wednesday</th>
<th>Thursday</th>
<th>Friday</th>
<th>Saturday</th>
</tr>
</thead>
<tbody>
<tr>
<td>1st</td>
<td>50</td>
<td>65</td>
<td>52</td>
<td>63</td>
<td>84</td>
<td>102</td>
</tr>
<tr>
<td>2nd</td>
<td>56</td>
<td>49</td>
<td>60</td>
<td>45</td>
<td>112</td>
<td>90</td>
</tr>
</tbody>
</table>

Let \( X_j \) represent the number of vehicles traveling through a randomly selected 15-minute interval on day \( j = 1 \) (Monday), \( 2 \) (Tuesday), \( 3 \) (Wednesday), \( 4 \) (Thursday) of week \( i = 1, 2 \).

Let \( Y_{ik} \) represent the number of vehicles traveling through a randomly selected 15-minute interval on day \( k = 1 \) (Friday), \( 2 \) (Saturday) of week \( i = 1, 2 \).

Assume that the number of vehicles follows a Poisson distribution. Assume further that the average number of vehicles traveling from Monday through Thursday is \( \lambda \) and for Friday and Saturday is \( \mu \). Assume the \( X_j \)'s and \( Y_{ik} \)'s are jointly independent random samples (as defined by Casella and Berger) from their respective distributions.

a. (3 pts) The \( X_j \)'s and \( Y_{ik} \)'s are assumed to be jointly independent random samples (as defined by Casella and Berger) from their respective distributions. Does this seem reasonable for this situation? Comment on the validity of each of these assumptions within the context of the scenario.

b. (7pts) Find a jointly sufficient statistic for \( (\lambda, \mu) \). Show all work and state any assumptions and theorems used.

c. (15pts) Derive the large-sample approximate likelihood ratio test (LRT) to test the hypotheses: \( H_0 : 2\lambda = \mu \) versus \( H_1 : 2\lambda \neq \mu \) at the 0.05 significance level. Do NOT carry out the test, but be sure to specify the test function, clearly defining the test statistic and the rejection region. Explain why a researcher might be interested in these results. (Note: You do not need to use the second derivative test to verify maximum likelihood estimators are maximums.)

d. (9 pts) Propose a point estimator that the researcher could use to obtain an interval estimate of \( 2\lambda / \mu \). Explain why the proposed point estimator is reasonable to use for this situation. Reference and use at least two different appropriate statistical criteria to justify your response.

e. Let \( C = \left( \sum_{j=1}^{4} \sum_{j=1}^{4} X_j + \sum_{j=1}^{2} \sum_{k=1}^{2} Y_{ik} \right) \).

i. (6pts) Use moment generating functions to prove \( C \sim \text{Poisson}(8\lambda + 4\mu) \). Show all work and state any assumptions and theorems used.

ii. (7pts) Find \( P(C > 800 | \lambda = 50, \mu = 100) \). Express your answer as a formula, with all of the numerical values clearly specified, but do NOT complete the calculation. Interpret in words what this probability represents in the context of the scenario.