**Analyze the data sets below and write a brief report that includes a summary of findings, a graphical display and a section describing the methods used to answer the questions of interest.**

1. The data in the file wine.csv (in the datasets folder on Canvas) give the average wine consumption rates (in liters per person) and number of ischemic heart attack deaths (per 1000 men aged 55 to 64 years) for 18 industrialized countries.

Do these data suggest that heart disease death rates are associated with average wine consumption? If so, how can that be described?

Do any countries have substantially higher or lower death rates than others with similar wine consumption rates?

Analyze the data and write a brief report that includes a summary of findings, a graphical display and a section describing the methods used to answer the questions of interest.

1. Meadowfoam is a small plant that grows in Pacific Northwest and is domesticated for its seed oil. A study was set up to determine if meadowfoam can be made into a profitable crop. In a controlled growth chamber, the plant was grown at 6 different light intensities and two different timings of onset of light treatment. The outcome of interest is the number of flowers per plant which was measured by averaging numbers of flowers produced by 10 seedlings in each group. Growth was replicated at each combination of time and light intensity.

c. The research questions are: **What are effects of intensity and timing? Is there an interaction between the two factors?**

d. First create an analysis of variance using timing and the categorical form of the light intensity variable. Determine if there is an effect of each factor.

e. Then create an interaction between light intensity and timing by multiplying the two variables and test for the presence of an interaction.

f. Now repeat the process but using light intensity as a continuous variable.

g. Then perform F-tests to compare the four model you have created (light as continuous and categorical with and without the interaction)

h. Predict the number of flowers grown at each combination of light and timing for each of the four models.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Timing | Light | Model 1 | Model 2 | Model 3 | Model 4 |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |

i. Compare each prediction to the observed number of flowers and calculate the difference (observed – predicted). This is the residual. Calculate the residual mean squared error for each model by adding the squared residuals together and dividing by the number of residual degrees of freedom. This should equal the mean squared error in each ANOVA table.

j. Now plot the residuals vs. the predicted for each model and see if there are any patterns. If you see any, what might you do to remove them?

k. Finally, take the model you think describes the data the best and write a short report for your grandmother who would like to grow these flowers carefully explaining to her how she should best grow them and why. Note that your grandmother is curious about how much changes in light and timing might affect her flowers and how sensitive her results will be to the settings she makes.

Question 2

The goal of the following analyses is to develop a model that best models the data obtained from a study measuring flower production of Meadowfoam, a small plant that grows in the Pacific Northwest. Plants were grown at 6 different light intensities and two different timings of light exposure. The model will be used to determine the effects of light intensity and exposure time on flower growth. The analysis will also explore the relationship between light intensity and exposure to determine if there is an interaction effect between the 2 factors.

Exploratory Data Analysis

Before exploring and analyzing the data, two new variables were created from the original dataset. The “REPLICATE” variable is created as a categorical variable that takes on the value 1 or 2. This variable distinguishes experiments that were performed with the same time and intensity and will be used to determine if there is a significant difference in flower growth among experiments performed under the exact same conditions. The “INTENS\_CAT” variable is created as a categorical variable that takes on a discrete value between 1 and 6, depending on the value of “INTENS.”

Summary of Findings

Methods

\*\*T – TEST to determine significance of REPLICATE?\*\*

We will first perform an analysis of variance using timing and the categorical form of the light intensity variable as factors and number of flowers as the dependent variable.

Df Sum Sq Mean Sq F value Pr(>F)

TIME 1 887.0 887.0 21.914 0.000144 \*\*\*

INTENS\_CAT 1 2579.8 2579.8 63.739 1.2e-07 \*\*\*

REPLICATE 1 61.8 61.8 1.526 0.231038

Residuals 20 809.5 40.5

##part e

Next we will run a two factor analysis of variance with timing and categorical light intensity as the factors. An interaction term of the two factors will also be included in the model. When looking at the interaction term in the model, we see that it is insignificant with a p-value of 0.909. This p-value corresponds to a hypothesis test assessing the null hypothesis that there is no interaction between timing and light against the alternative that there is an interaction between the factors. With a highly insignificant p-value, the null hypothesis is not rejected and the insignificant interaction term should be dropped from the model.

First, we will assess the global hypothesis testing the null hypothesis that the mean flower growth is equal among all combinations of factor levels against the alternative that at least one factor level differs from the rest. At a significance level of α = 0.05, the resulting p-value of

Appendix