MSDS UNIT 3 HW 3

Conceptual questions.

1. State the definition of factor, levels, treatment, and main effects.

Factor- a categorical explanatory variable

Levels- various categories of the factor

Treatment- all possible level combinations of multiple factors in an experiment

Main effects- average values for each factor

Interactions- when effects of factors differ with the level of another factor

1. What are the necessary assumptions of a TWO WAY ANOVA analysis in order to perform hypothesis testing?
2. Errors are normally distributed with mean 0
3. Errors have constant variance
4. Errors are mutually independent
5. Suppose we have a two factor ANOVA model where each factor has only two levels (Let be the factors and call the levels ZERO and ONE. Treating the factors as dummy variables, write out a multiple linear regression model for the additive and nonadditive models. What does the intercept represent in your model? What regression coefficient should be used to test for each of the single factors in the additive model? What regression coefficient would be used to test if an interaction exists in the nonadditive model?

Additive Model is Nonadditive Model is

The intercept represents the estimated mean for the group observations that have levels 0 for each factor.

For the additive model. The regression coefficient represents the difference in means from transitioning from . So the appropriate test to see if there is a difference in the two levels for X1 would be to test if or not. Similarly for the second factor.

To test for an interaction for the nonadditive model, it is clear that the only difference in the two models is the inclusion of . So to test for an interaction, we would need to simply test if or not.

1. TRUE or FALSE? Suppose we run a two way ANOVA model using two factors A (High Mediam Low) and B (assembly line 1, 2, and 3) and there is plenty of data for each factor level combination. All 3 type-III sum of squares F-tests are significant. It is valid to directly compare High versus Low via a t-test regardless of assembly line.

FALSE. If an interaction is significant we must compare high versus low for each assembly line.

1. TRUE or FALSE? Since TWO WAY ANOVA is a special case of multiple regression, influential diagnostics such as leverage statstics should be examined in a 2Way ANOVA analysis.

FALSE. Per discussion in class, leverage values are only a concern for continuous predictors. Two way anova does not contain continuous explanatory variables.

1. TRUE or FALSE? Contrasts are specific hypothesis tests specified by the analyst to determine differences between specific factor levels or treatments in an ANOVA model.

TRUE.

1. The following boxplots summarize data for each factor combinations of Factor 1 (A,B) and Factor 2 (a,b,c) from a two way anova analysis. Using the median values of each boxplot, make a decision for each graph of whether or not an interaction potentially exists

The point of this exercise is to get you thinking about what it means to have a significant interaction viewed outside of simple means plot and looking for parallel lines.

FIGURE 1: This figure supports the addative model (no significant interaction). You can see that the change from a to b to c is the same for both levels A and B. Remember it’s the changes we are examining not necessarily the placement of the means.

FIGURE 2: Clearly nonadditive (significant interaction). There is no difference between c vs b for level A but there is a big difference in level B. So, the effects of c,b,a depend on what the level of the other currently is.

FIGURE 3: Nonaddative (significant interaction). The boxplots look sort of parallel because they are shifting in tandem but upon closer inspection the changes of a,b,c depend on what level of A, B you are looking at.

  

Analysis Question

1. Back in Unit 1 we considered a study in which 4 different fertilizers were tested for their yield (in mm of growth) on a local grass: Red Fescue (Just for fun .. this is a real mountain grass! http://fescue.com/info/creepingred.html#.WIu6rbYrKu4). To conduct their study they had enough money to run three replicates of each fertilizer. They knew the red fescue was a mountain grass so they went out to the mountains and carefully fertilized plots of land as you see in the diagram below. The data is contained in the ***growth3*** data set.



Given the information you have available, run a simple ONE WAY ANOVA to test for the effects (if any) of the fertilizers. **From a previous study, we know that the yields from each fertilizer are normally distributed with equal variances. For now assume that independence is not a concern here. You may assume the assumptions are met for all questions in this homework.**

Obtain the following Deliverables: 1. ANOVA table. 2. Means Plot (Interaction Plot from SAS) 2. Conclusion for the appropriate ANOVA test. 3. Confidence intervals with multiple comparison corrections for **SIGNIFICANT** differences (between Fertilizers). 4. SAS Code: proc glm or Proc mixed code. Your answer should fit in the given box.

|  |
| --- |
| **Since there is only one factor, an interaction plot isn’t necessary. From the ANOVA, there is not sufficient evidence to suggest that the mean growth of plants under different fertilizers is different (pvalue .1064 from an ANOVA).**  **proc** **glm** data= growth3 plots = (Diagnostics Residuals);  class fertilizer;  model growth = fertilizer;  **run**; |

1. Now assume that you get to thinking about it and realize that we may really be looking at three different environments here: Sunny, Wetlands and Mostly Shady. This data has been recorded in the ***growth4*** data set. Conduct a similar analysis as in 1 but now account for the environment variable (ENV).



Deliverables: 1. ANOVA table. 2. Means Plot (Interaction Plot from SAS) 2. Conclusion for the ANOVA (only step 6 is required.) 3. The researchers were specifically interested in (a) Inference on the main effect of the fertilizer. (b) If Fertilizer 4 performed significantly better in one environment than another. (c) Which fertilizer will perform best in the wetlands. Answer each QOI (question of interest) with a confidence interval and a 1 sentence conclusion. Your answer should fit in the space below (on this page.) 4. SAS Code: Proc glm or proc mixed statement. Answer should fit in the given box.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| |  |  | | --- | --- | | There is strong evidence from the ANOVA that at least one fertilizer and/or at at least one environment group has a deferent mean growth score than the others (pvalue < .0001). |  |  |  | | --- | | 3a.    There is strong evidence that, after accounting for the environments, there is a significant difference in mean growth between at least 1 pair of fertilizers. The table below indicates that each fertilizer has a mean growth that is significantly different that the others. For instance, the mean growth with Fertilizer 4 is estimated to be between 5.9mm and 4.1mm less than the mean growth with fertilizer 2 (after a Bonferroni adjustment).    **proc** **mixed** data= growth4 plots = (ResidualPanel);  class fertilizer env;  model growth = fertilizer env ;  lsmeans fertilizer env / CL adjust = bon;  **run**; |  |  | | --- | | 3b. If we model each Fertilizer / Environment combination separately (assume the non-addititve model), while fertilizer 4 performed best in the Wetlands, we only had one observation at this fertilizer / environment combination and thus we do not have an estimate of the standard deviation in which to run a test. For the same reason, we cannot construct a confidence interval for the difference between the mean growth with fertilizer 4 in the different environments.  **proc** **mixed** data= growth4 plots = (ResidualPanel);  class fertilizer env;  model growth = fertilizer | env ;  lsmeans fertilizer env / CL adjust = bon;  **run**;  However, if we assume that any differences between Fertilizers are consistent across Environments (the additive model) then we save degrees of freedom and have extra observations to estimate the standard error. In this case, we can simply use the confidence intervals from the table in 3a to provide evidence that Fertilizer 4 (which for the additive model means all the fertilizers) performed “better” in the Wetlands. We are 95% confidence that the growth for Fertilizer 4 in the Wetlands is between 9.3 and 10.7 mm more than in the Mostly Shady environment and between 7.1 and 8.4 mm more than in the Sunny environment. *Note the only difference in the code here is that the code below does not fit interaction parameters.*  **proc** **mixed** data= growth4 plots = (ResidualPanel);  class fertilizer env;  model growth = fertilizer env ;  lsmeans fertilizer env / CL adjust = bon;  **run**; |  |  | | --- | | 3c. Again, while fertilizer 4 performed best in the Wetlands, we only had one observation at each fertilizer / environment combination and thus we do not have an estimate of the standard deviation in which to run a test. For the same reason, we cannot construct a confidence interval for the difference between the mean growth with the different fertilizers in the Wetlands. |   **proc** **mixed** data= growth4 plots = (ResidualPanel);  class fertilizer env;  model growth = fertilizer | env ;  lsmeans fertilizer env / CL adjust = bon;  **run**;  Once again, on the other hand, if we assume that any differences between Fertilizers are consistent across Environments (the additive model) then we save degrees of freedom and have extra observations to estimate the standard error. In this case, we can simply use the confidence intervals from the table in 3a to provide evidence that in the Wetlands (which for the additive model means all the environments) Fertilizer 4 performed the best. We are 95% confidence that the mean growth for Fertilizer 4 in the Wetlands (as well as the other two environments) is between 8.4 and 10.2 mm more than that from Fertilizer 1, 3.4 to 5.2 mm more than that from Fertilizer 2, and 10.4 to 12.2 mm more than that from Fertilizer 3. *Note the only difference in the code here is that the code below does not fit interaction parameters.*  **proc** **mixed** data= growth4 plots = (ResidualPanel);  class fertilizer env;  model growth = fertilizer env ;  lsmeans fertilizer env / CL adjust = bon;  **run**; |

1. Consider the model you used in problem 2. Inspect the Means Plot (Interaction Plot). Does it look like there will be a significant interaction? Explain by interpreting what an interaction is and then comparing that to what you see in the plot.

|  |
| --- |
| An interaction in this case would be one in which the effect of the fertilizer depended on the environment that the fertilizer was used in. Since all the lines in the interaction plot above are close to parallel, there is not a lot of evidence of an interaction effect. |

1. Fit the full model that includes the interaction term? What do you notice? Why? Discuss.

|  |
| --- |
| Note that the degrees of freedom to estimate the error is ‘0’ and thus we do not have an estimate of the MSE and thus we cannot perform any tests or construct any confidence intervals. This underlines the importance of repeat observations at different treatments (fertilizer/environment combinations).  **proc** **mixed** data= growth4 plots = (ResidualPanel);  class fertilizer env;  model growth = fertilizer | env ;  lsmeans fertilizer env / CL adjust = bon;  **run**; |