BS805 Fall 2021 Week 3

Be sure to follow the *Assessment Guideline 1: Writing up Homework* at the end of the syllabus in preparing the homework for submission.

Homework assignments need to be uploaded to the blackboard website by 2 PM on the due date.

In each homework report, be sure to include an introductory and a summary paragraph. Also, include the relevant parts of your SAS code where appropriate in your answer for each question.

A study was conducted on the blood lead levels of children. The following variables for twenty-five children from the study have been entered in the saved SAS data set, *lead_s2021.sas7bdat*:

ID	ID Number (numeric, values 1-25)
DOB	Date of Birth (mmddyy8. format)
DAYBLD_A	Day of Blood Sample A (numeric, initial possible range: -1 to 31)
MTHBLD_A	Month of Blood Sample A (numeric, initial possible range: -1 to 12)
DAYBLD_B	Day of Blood Sample B (numeric, initial possible range: -1 to 31)
MTHBLD_B	Month of Blood Sample B (numeric, initial possible range: -1 to 12)
DAYBLD_C	Day of Blood Sample C (numeric, initial possible range: -1 to 31)
MTHBLD_C	Month of Blood Sample C (numeric, initial possible range: -1 to 12)
PBLEV_A	Blood Lead Level Sample A (numeric, possible range: 0.01 – 20.00)
PBLEV_B	Blood Lead Level Sample B (numeric, possible range: 0.01 – 20.00)
PBLEV_C	Blood Lead Level Sample C (numeric, possible range: 0.01 – 20.00)
SEX	Sex (character, ‘M’ or ‘F’)

All blood samples were drawn in 1990. However, during data entry the order of blood samples was scrambled so that blood sample A may not correspond to the first blood sample taken on a subject, it could be the first, second or third. The same ordering concern may apply to blood samples B and C as well. In addition, some of the months and days for the blood sampling were not written on the forms. At data entry, missing month and missing day values were each coded as -1 or 13 for month and -1 or 32 for day. Be sure to write your code to account for either possibility.

The team of investigators for this project has made the following decisions regarding the missing values. Any missing days should be set to 15 and any missing months set to 6. Any analyses that follow are to be done on this data set. **Be sure to implement the SAS syntax as indicated for each question.** For example, use SAS arrays and loops if the item states that these must be used.

A) Using this saved SAS data set, create a new, **temporary** SAS data set and performing the following:

1) use SAS arrays and looping to create a SAS date variable for each of the three blood samples and to address the missing data in accordance to the decisions of the investigators. Use arrays and a loop to recode the missing values for day and month;
2) use a SAS function to create a new variable for the highest, i.e., maximum, blood lead value for each child;
3) use SAS arrays and looping to identify the date on which this highest value was obtained and create a new variable for the date of the highest blood lead value;

4) create a variable for the age of the child in years when the largest blood lead value was obtained (rounded to two decimal places);
5) create a variable based on the age of the child in years when the largest lead value was obtained (call it, "*agecat*") that takes on three levels: for children less than 4 years old, *agecat* should equal 1; for children at least 4 years old, but less than 8, *agecat* should equal 2; and for children at least 8 years of age, *agecat* should be 3.;
6) Using this temporary SAS data set, print out the values for the variables for the ID, date of birth, date of the largest lead level, age at blood sample for the largest blood lead level, *agecat*, *sex*, and the largest blood lead level (**<u>Only print these requested variables</u>**). All dates should be formatted to use the **mmddyy10.** format on the output.
7) Use the MEANS procedure to provide the mean and standard deviation for the children's age in each sex group. Use the FREQ procedure to find the percentages in each age category (*agecat* variable) for each sex.

B) Because the amount of exposure to lead paint may affect the largest blood lead level, researchers are interested in seeing if the children at different age levels have different mean levels of lead in their blood. Because sex may also be involved in the blood lead levels, it is felt that sex should also be a factor in the analysis. Perform a two-factor analysis of variance with the dependent variable, the largest blood lead level, and the factors, *agecat* and *sex*, to test this hypothesis. To perform this analysis, follow the steps below. (For hypothesis testing use 0.05 as the significance level.)

1) Run a two-factor ANOVA with interaction. Include both main effects and their interaction in the model. Is this model significant? If the interaction is significant, stop here: **<u>main effects (significant or not) should not be removed from the model</u>** (It is difficult to interpret significant interactions when the main effects are not present in the model). Remember: do not interpret main effects in the presence of a significant interaction. In terms of hypothesis testing if the interaction is significant, report on the global null hypothesis for the model and for the interaction null hypothesis.
2) If the interaction is not significant, remove this term and run a two-factor ANOVA without interaction (a "main effects only" or additive model). In terms of hypothesis testing in this model, report on the global null hypothesis for the model and for each of the main effects along with the results of post-hoc tests if a main effect is statistically significant.

As noted above, your report on these analyses should state and explain the hypotheses of interest, what statistical methods were used, what terms have been included in the model, which have been removed from the model (if any), why they were removed, and noting which factors were statistically significant (if any). For hypothesis tests, all relevant test statistics, degrees of freedom, and p-values should be reported.

Be sure, however, to report on both hypothesis testing and estimation in your report. Results of estimation would include cell means and standard deviations for the interaction model, adjusted means and standard errors for the main effects only model (if this is the model to be interpreted), sample sizes, and confidence intervals (if relevant). Clearly note which levels (i.e., categories) of the factors are associated with increased maximum blood lead levels, that is, interpret the results and report fully the results of both the hypothesis tests and the relevant statistical estimates.