CS 7200 - Fall 2023 Olga Vitek

Homework 3

To facilitate grading, please use the homework solution template posted on Piazza.

Due on Canvas before midnight on Tuesday, November 14 2023.
Each part of the problems 5 points

1. *[Writing question].* You plan to analyze data from an experiment with a 2-factor, where the first factor has $a = 3$ levels, and the second factor has $b = 3$ levels, and 2 replicate observations per cell.

   (a) Write out the cell means model, and state the assumptions.

   (b) Provide the associated ANOVA table with columns Source, Some of Squares, and Degrees of Freedom, with the name or equation in each cell. State the null hypothesis that can be tested with this table. *[Hint: there is one.]*

   (c) Re-state the cell means model as a linear regression approach, and spell out the vector $\mathbf{Y}$, the design matrix $\mathbf{X}$, the vector of parameters μ, and vector of errors ε.

   (d) Write the vector of coefficients $\mathbf{C}$ associated with the parameters of the linear regression for the contrast $L = \mu_{12} - \mu_{13}$.

   (e) Re-write this model as a two-way factor effects model with zero sum constraints, and state the assumptions. Be sure to specify the distributional assumptions and the constraints.

   (f) Provide the associated ANOVA table with columns Source, Some of Squares, and Degrees of Freedom, with the name or equation in each cell. State the null hypotheses that can be tested with this table. *[Hint: there are multiple.]*

   (g) Re-state the factor effects model as a linear regression approach, and spell out the vector $\mathbf{Y}$, the design matrix $\mathbf{X}$, the vector of parameters μ, and vector of errors ε.

   (h) Write the vector of coefficients $\mathbf{C}$ associated with the parameters of the linear regression for the contrast $L = \mu_{12} - \mu_{13}$.

   (i) Re-write this model as a two-way factor effects model with reference (one-hot) constraints, and state the assumptions. Be sure to specify the distributional assumptions and the constraints.

   (j) Re-state the factor effects model as a linear regression approach, and spell out the vector $\mathbf{Y}$, the design matrix $\mathbf{X}$, the vector of parameters μ, and vector of errors ε.

   (k) Write the vector of coefficients $\mathbf{C}$ associated with the parameters of the linear regression for the contrast $L = \mu_{12} - \mu_{13}$.

2. *[Implementation question].* In this question, we will implement the linear regression-based estimation of parameters in a 2-factor experimental design above, with a reference constraint (μ_{11} as the reference). The implementations below must be done from scratch (i.e., you cannot use *lm* or other libraries for linear regression). The implementation

does not need to be general (i.e., it's enough to make it work on for this homework). Although linear regression libraries like *lm* are not allowed in your implementation, you are encouraged to use them to check your work. We recommend you use the dataset from the question 3 to test your implementation.

(a) Implement a linear regression-based estimation of parameters of this model, with a reference constraint (μ_{11} as the reference).

(b) Implement the estimation of contrasts, and of their standard errors.

3. *[Data analysis question].* Consider the dataset from KNNL problem 19.14 The dataset is available from this url

`http://users.stat.ufl.edu/~rrandles/sta4210/Rclassnotes/data/textdatasets/index.html`

(a) Briefly explain how to assign volunteers to treatments using an appropriate randomization, and why randomization is important.

(b) Visualize the dataset with a treatment means plot with Factor A as the x axis and Factor B as the line/point color. Based on the plot, comment on whether both factors and the interaction are likely to be present.

(c) Using your implementation in Question 2, test for the presence of interactions, at the confidence level of 95%. Please state the hypotheses and interpret your results.

(d) Using your implementation in Question 2, estimate μ_{23} with a 95% confidence interval and interpret the results.

(e) Using your implementation in Question 2, estimate $L = \mu_{12} - \mu_{13}$ with a 95% confidence interval and interpret the results.

4. *[Data analysis question].* Consider the dataset from KNNL problem 21.5. The dataset is available from this url

`http://users.stat.ufl.edu/~rrandles/sta4210/Rclassnotes/data/textdatasets/index.html`

(a) Explain why a randomized complete block design is useful in this problem.

(b) Visualize the dataset with a treatment means plot, and conduct the Tukey test for additivity. State the null and the alternative hypotheses, and the conclusion. You may use regression libraries like *lm* or *anova* for this question.

(c) Use a standard implementation of linear models in R (or any other language) to fit the *additive* model to the data. Interpret your results.

(d) Use a standard implementation of linear models in R (or any other language) to derive a confidence interval for the difference between the training methods 1 and 2. Interpret your results.

(e) Repeat (d), while ignoring the blocking (i.e., treat the data as if it came from a completely randomized experiment). Comment on the difference in the width of the confidence intervals.

(f) Repeat (d), while ignoring training method 3 (i.e., treat the data as if the experiment did not contain the third training). Comment on the difference in the width of the confidence intervals.