Homework 1

PubH 7406: Biostatistical Inference II – Jared D. Huling

Due: Tues, Feb 7, 2023 at the beginning of class

Homework 1 - ANOVA - 38 points

Instructions

Turn in the homework in the form of a PDF. It is fine to use existing functions to answer questions. **Some notes on how homeworks will be graded:** Simply providing output from statistical software is not sufficient and will not receive full points. Any output/results *must* be interpreted in the context of the real-world problem; accompanying explanations of the results are necessary to receive full credit. Use clearly-defined and explained statistical notation to accompany results. For example, any statistical tests should be accompanied by a formal statement of the null and alternative hypotheses with additional context explaining what the hypotheses mean in the context of the problem. Please follow the instructions on homeworks in the syllabus in order to receive full credit.

If you have any questions, please ask in the course Q&A on Canvas so that others can see any responses.

The Data

Fasting glucose levels Y (log of millimoles per liter) were recorded for 1000 unrelated non-diabetic individuals. Each individual has two measured *discrete* covariates, (G_1, G_2), taking values in $(0, 1, 2)$. We will study the relationship between the fasting glucose levels and (G_1, G_2). The questions will center around an ANOVA analysis of the data. Hint: if you are using `R`, you may need to convert G_1 and G_2 to factors.

Note that this is an unbalanced design with unequal sample sizes across levels. Conduct all hypothesis tests at the 0.05 significance level.

The data can be downloaded in R as follows:

```
FG2 <- read.table("https://jaredhuling.org/data/pubh7406/FG2.txt",
                  header = TRUE)
```

The Questions

1. **(5 points)** Provide preliminary visualizations of the data. Provide initial assessments about what can be seen from these visualizations in relation to an Analysis of Variance analysis of the data.

2. **(8 points)** Write down a two-factor ANOVA model using G_1 and G_2 and their interaction to explain the variation in the fasting glucose level. Define any notation you use and list all assumptions that are made by this model.

3. **(4 points)** Provide the ANOVA table associated with the model defined in the previous question and explain what the different sums of squares are.

4. **(8 points)** Provide visualizations that investigate the assumption of equal variances and the assumption of normality of the errors. Under the model assumed in question 2, conduct the Levene test to assess the equal variances assumption. Make sure to explicitly write the null and alternative hypotheses in clear statistical notation. Hint: the Levene test is available in the `car` package in `R`.

5. **(5 points)** Conduct a statistical test to determine whether the interaction of G_1 and G_2 has any impact on the mean response. What is the conclusion and how does that impact how you would model resting glucose level?

6. **(8 points)** Use the Tukey and Bonferroni tests to compare all pairwise comparisons of the 9 cell means. Describe and explain any differences in findings between the two approaches. What do the conclusions about the pairwise comparisons tell you about the relationship between G_1, G_2, and Y?