Assginment 3 Report, DATA 400

Intructor: Dr. Spectrum Han

Yohen Thounaojam, 56112204

Introduction

This report is a part of Assignment 2 of DATA 400, 2021W2 at UBCO. It is composed of two parts (1 report each). Please refer to the Table of Contents for more information.

Other Information

- All R code used in this report can be found in the .Rmd file in the submission. Moreover, for the relevant portions of the report, the code has been *'echoed'* back into this PDF.

Question 1: *rice* data

Introduction

The data shown below in *Table 1.1* is from an experiment conducted on the mass of plants (ShootDryMass) as a function of fertilizer and plant variety.

The treatments given to the plants had a two-way structure, but the plants were also grown in blocks chosen so that conditions are as similar as possible within each block.

```
##    PlantNo Block RootDryMass ShootDryMass         trt   fert variety
## 30       6     1          23          129      NH4NO3 NH4NO3      wt
## 31       7     2          14           48      NH4NO3 NH4NO3      wt
## 32       8     2          14           60      NH4NO3 NH4NO3      wt
## 33       9     2          12           46      NH4NO3 NH4NO3      wt
## 34      10     2          23           74      NH4NO3 NH4NO3      wt
## 35      11     2          11           51      NH4NO3 NH4NO3      wt
## 36      12     2          20           64      NH4NO3 NH4NO3      wt
## 37       1     1           6            8 F10 +ANU843    F10  ANU843
## 38       2     1           4            6 F10 +ANU843    F10  ANU843
## 39       3     1           4            3 F10 +ANU843    F10  ANU843
## 40       4     1           7            1 F10 +ANU843    F10  ANU843
## 41       5     1           5            7 F10 +ANU843    F10  ANU843
## 42       6     1           6            5 F10 +ANU843    F10  ANU843
## 43       7     2           6           10 F10 +ANU843    F10  ANU843
## 44       8     2           5           17 F10 +ANU843    F10  ANU843
```

Table 1.1 15/72 rows of rice Data.

In the above table, each column is very self-explanatory. Each treatment combination occurred once per block. This kind of experiment is designed to take block effects into account but there should not be an interaction between block and treatment.

Objective

The goal of this analysis is to conduct a three-way ANOVA where there is an interaction between *variety* and *fert*, but no interaction with *Block*.

The questions we aim to answer are: 1. Are the treatment interactions significant? 2. Obtain a useful summary of the treatment effects.

Exploratory Data Analysis

Using the *summary* function, let us look at some EDA of the *rice* data.

```
##     PlantNo          Block      RootDryMass     ShootDryMass
##  Min.   : 1.00   Min.   :1.0   Min.   : 1.00   Min.   :  1.00
##  1st Qu.: 3.75   1st Qu.:1.0   1st Qu.: 7.00   1st Qu.: 35.00
##  Median : 6.50   Median :1.5   Median :12.50   Median : 58.00
##  Mean   : 6.50   Mean   :1.5   Mean   :18.07   Mean   : 59.56
##  3rd Qu.: 9.25   3rd Qu.:2.0   3rd Qu.:20.25   3rd Qu.: 80.50
##  Max.   :12.00   Max.   :2.0   Max.   :67.00   Max.   :134.00
##              trt         fert      variety
##  F10           :12   F10   :24   wt    :36
##  NH4Cl         :12   NH4Cl :24   ANU843:36
##  NH4NO3        :12   NH4NO3:24
##  F10 +ANU843   :12
##  NH4Cl +ANU843 :12
##  NH4NO3 +ANU843:12
```

Table 1.2 Summary of data in rice dataframe.

Method

As instructed, we will be performing a Three-Way ANOVA on *variety, fert* and *Block*.

```
rice_aov <- aov(ShootDryMass ~ fert + variety + Block + fert:variety, data=rice)
summary(rice_aov)
```

```
##              Df Sum Sq Mean Sq F value   Pr(>F)
## fert          2   7019    3509   10.85 8.63e-05 ***
## variety       1  22684   22684   70.10 6.37e-12 ***
## Block         1   3528    3528   10.90  0.00156 **
## fert:variety  2  38622   19311   59.68 1.93e-15 ***
## Residuals    65  21034     324
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Table 1.3 AOV of interaction between variety and fert, but no interaction with Block.

We see that there was a statistically significant interaction between **variety** and **fert** for the **ShootDryMass**

Results

To further understand the output we saw in detial, we will use the *Tukey honest significant differences* to perform multiple pairwise comparisons between *fert*.

```
TukeyHSD(rice_aov, "fert")
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = ShootDryMass ~ fert + variety + Block + fert:variety, data = rice)
##
## $fert
##                    diff       lwr       upr     p adj
## NH4Cl-F10    -9.416667 -21.87225  3.038916 0.1732576
## NH4NO3-F10   14.583333   2.12775 27.038916 0.0178092
## NH4NO3-NH4Cl 24.000000  11.54442 36.455583 0.0000547
```

Table 1.4 TukeyHSD of AOV from Table 1.3.

Here, we see that two differences in the *fert* variable is significant:

1. Between NH4NO3-F10 (adjusted p-value = 0.018)
2. Between NH4NO3-NH4Cl (adjusted p-value = 0.0001)

Since the experiment is a Balanced Complete Block Design (BCBD), we can summarize the AOV results below:

```
model.tables(rice_aov, "means", se = TRUE)
```

```
## Tables of means
## Grand mean
##
## 59.55556
##
##  fert
## fert
##    F10  NH4Cl NH4NO3
##  57.83  48.42  72.42
##
##  variety
## variety
##     wt ANU843
##  77.31  41.81
##
##  Block
## Block
##     1     2
## 66.56 52.56
##
##  fert:variety
##         variety
```

```
## fert     wt     ANU843
##   F10    108.33   7.33
##   NH4Cl   50.25  46.58
##   NH4NO3  73.33  71.50
##
## Standard errors for differences of means
##          fert variety fert:variety
##         5.193   4.240        7.344
## replic.    24      36           12
```

Table 1.5 Model Summary of AOV from Table 1.3

Discussions and Conclusion

A three-way ANOVA was performed to test the interaction between variety and fert, but no interaction with Block.

A very significant interaction was found between *variety* and *fert*. This means that both variables introduce significant variablibity on the mass of the plants. Further analysis of the *fert* variable shows us that interactions between NH4NO3 and F10, and, NH4NO3 and NH4Cl are significant showing evidence for a difference in the 4 types of *fert*.

Ex 5.2 from Chapter 5: *schooldays* data

Page 66

Introduction

The data shown below in *Table 2.1* is from a sociological study of Australian Aboriginal and white children. In this study, children of both sexes from the following four age groups and from two cultural groups were used: - Final grade in primary schools - First Grade in Secondary School - Second Grade in Secondary School - Third Grade in Secondary School

The response variable was the number of days absent from school.

```
## [1] TRUE
```

```
##              race gender school learner absent
## 70     aboriginal female     F3 average     10
## 71     aboriginal female     F3 average     14
## 72     aboriginal female     F3 average     21
## 73     aboriginal female     F3 average     36
## 74     aboriginal female     F3 average     40
## 75 non-aboriginal   male     F0    slow      6
## 76 non-aboriginal   male     F0    slow     17
## 77 non-aboriginal   male     F0    slow     67
## 78 non-aboriginal   male     F0 average      0
## 79 non-aboriginal   male     F0 average      0
```

Table 2.1 schooldays dataset.

Objective

As laid out in the exercise, the goal is to carry out an appropriate *Analysis of Variance* of the data.

Exploratory Data Analysis and Preparation

In the exercise, we are told that:

- The data is unbalanced.
- The response variable is a count.

Let us begin the analysis with a plot of the mean absent days for each of the four factors, shown in *Fig. 2.1.*

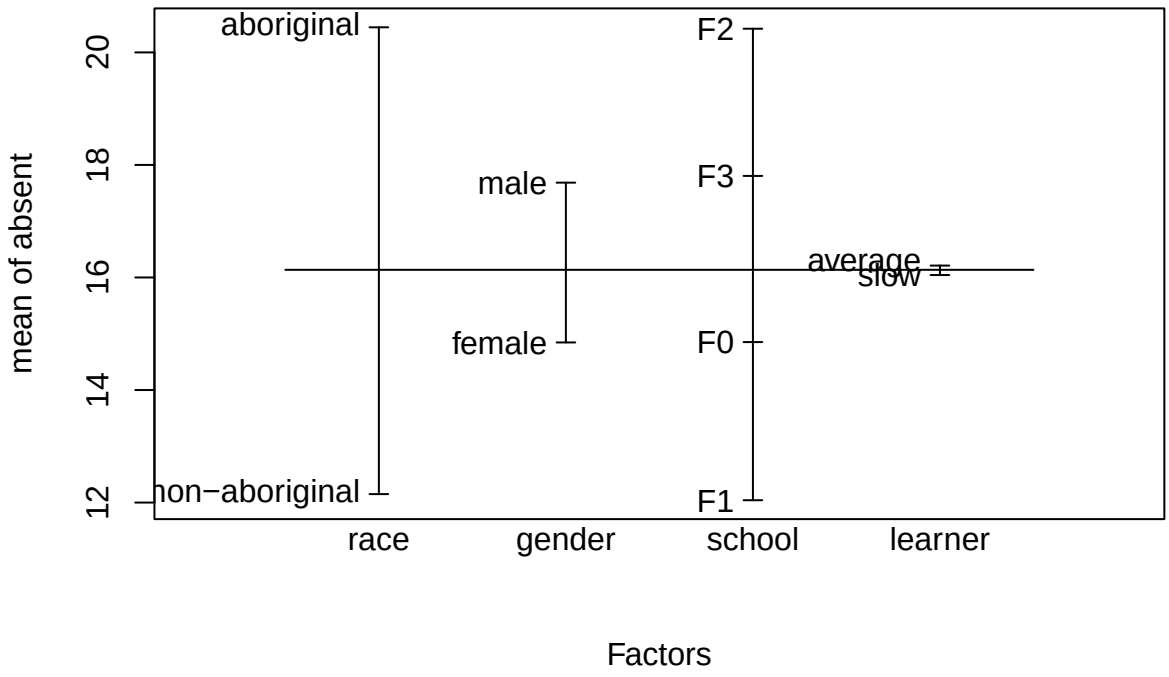


Fig 2.1 Plot of mean absent days for each factor: race, gender, school and learner.

From *Fig 2.1*, we can see that the differences in *Number of Days Absent* for its corresponding types are significant for **race** and **school**, but not so much for *gender* and negligible for *learner* type. Hence, we are going to perform AOV for the interactions between race, school and gender.

Method

We know that the unbalanced nature of the data brings some complications while performing an AOV since it is no longer possible to partition the variation in the data into non-overlapping sums of squares representing interactions. Let us derive a few Analyses of Variance tables.

```
summary(aov(absent ~ race*gender*school, data = schooldays))
```

```
##                     Df Sum Sq Mean Sq F value   Pr(>F)
## race                 1   2646  2645.7  12.606 0.000526 ***
## gender               1    339   338.9   1.615 0.205976
## school               3   1222   407.3   1.941 0.125883
## race:gender          1    174   173.9   0.829 0.364203
## race:school          3   3628  1209.4   5.762 0.000963 ***
## gender:school        3   1502   500.5   2.385 0.071889 .
## race:gender:school   3    233    77.8   0.371 0.774369
## Residuals          138  28963   209.9
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
summary(aov(absent ~ school*race*gender, data = schooldays))
```

```
##                     Df Sum Sq Mean Sq F value   Pr(>F)
## school               3   1661   553.7   2.638 0.052079 .
## race                 1   2310  2310.1  11.007 0.001161 **
## gender               1    235   235.3   1.121 0.291568
## school:race          3   3650  1216.8   5.798 0.000921 ***
## school:gender        3   1438   479.3   2.284 0.081710 .
## race:gender          1    215   215.2   1.025 0.313044
## school:race:gender   3    233    77.8   0.371 0.774369
## Residuals          138  28963   209.9
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
summary(aov(absent ~ gender*school*race, data = schooldays))
```

```
##                     Df Sum Sq Mean Sq F value   Pr(>F)
## gender               1    308   308.1   1.468 0.227758
## school               3   1548   515.9   2.458 0.065497 .
## race                 1   2351  2350.7  11.200 0.001054 **
## gender:school        3   1443   481.0   2.292 0.080873 .
## gender:race          1    243   243.5   1.160 0.283333
## school:race          3   3617  1205.7   5.745 0.000985 ***
## gender:school:race   3    233    77.8   0.371 0.774369
## Residuals          138  28963   209.9
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Table 2.2 AOV's taking factors in different orders.

As you can see in *Table 2.2*, there are differences in the sum of squares for certain factors and consequently, in the associated F-tests and p-values. This confirms the unbalanced nature of the data.

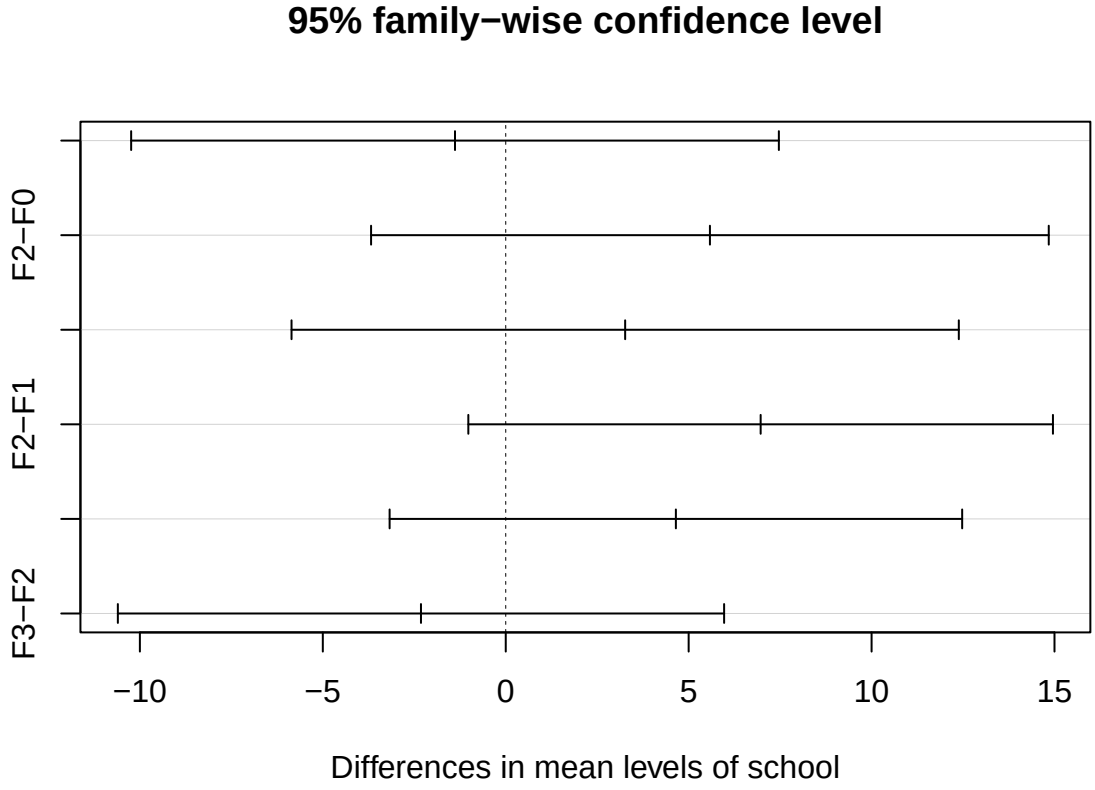
Moreover, we notice that there is always a consistent significant interaction ($p < 0.05$) between *race* and *school*; complementing our finding from the EDA previously done.

Results

As seen in the section above, the *unbalanced* nature of the data would not allow us to use just an ANOVA. Hence, we will be using Tukey's 'Honest Significant Difference' method followed by graphical representation of the multiple confidence intervals for *School* and *race*.

School Variable

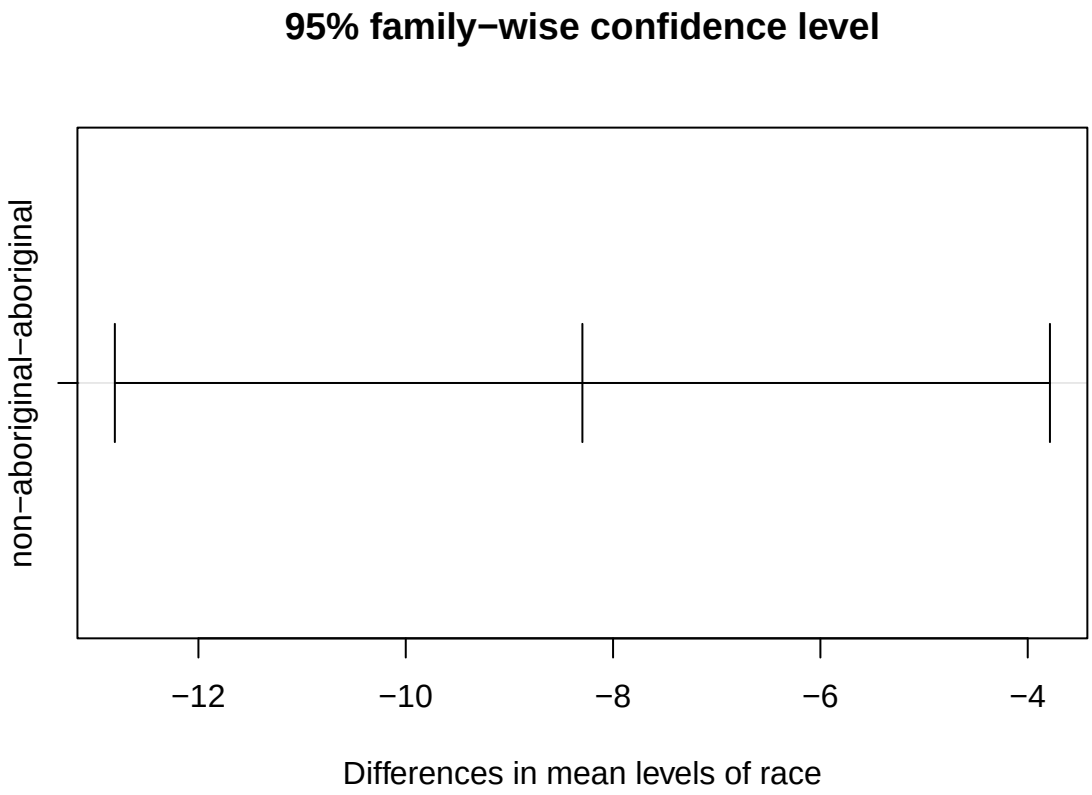
```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = absent ~ race * gender * school * learner, data = schooldays)
##
## $school
##            diff        lwr       upr     p adj
## F1-F0 -1.386454 -10.238346  7.465437 0.9769685
## F2-F0  5.581286  -3.680419 14.842991 0.3997252
## F3-F0  3.264633  -5.855236 12.384502 0.7875000
## F2-F1  6.967740  -1.022222 14.957703 0.1104385
## F3-F1  4.651087  -3.174022 12.476196 0.4121546
## F3-F2 -2.316653 -10.602516  5.969210 0.8855943
```



From the above results (Tukey and Plot), we see that there is high difference in mean values but none of them are significant enough. Next, let us look at the case for race.

Race Variable

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = absent ~ race * gender * school * learner, data = schooldays)
##
## $race
##                                 diff       lwr       upr     p adj
## non-aboriginal-aboriginal -8.295946 -12.80631 -3.785585 0.0003995
```



In the case of *race*, we see a high difference in means between the two recorded races. However, the more important take-away is that this is **highly significant** with a p-value of 0.0004. In the next section, we will explore ideas and reasoning behind the influence of race on the number of days absent.

Discussions and Conclusion

We conclude that there is a high dependence of the *'number of days absent'* on the race of the child (aboriginal or non-aboriginal). We also saw some dependence on the *school* variable.

Going back to *Fig. 2.1*, it makes sense that the *race* of the children introduces high variability in the data as we saw a high difference in mean absent days. This also supports reports of the Australian Institute of Health and Welfare report in 2018 on family, domestic and sexual violence in Australia; which states that **Aboriginal Australians had increased risk factors for family violence, such as poor housing and overcrowding, financial difficulties, low education and unemployment.**

The above claims (*Link: see News.com report*) can be a possible explanation to why aboriginal children are missing school more than their non-aboriginal counterparts.

Acknowledgements

Data Source: HSAUR3 | *Link: Documentation*

References

Question from the A Handbook of Statistical Analyses Using R | *Link*