School Days Analyses Report

Ryota Nishida

Executive Summary

This report summarizes a graphical exploratory analysis of the number of days absent from school during the school year for children of both sexes from four age groups (final grade in primary schools and first, second, and third in secondary school) and from two cultural groups. The children in each age group were classified as slow or average learners. The analysis indicates that only the main effect of race impacted the number of absents and the interaction between race and school impacted absents.

Data Description

The data comes from Table 5.4 in the book, "A Handbook of Statistical Analysis Using R"(Hothorn & Everitt, 2014 [1]). The data itself arises from a sociological study of Australian Aboriginal and white children reported by Quine (1975). The data frame consists of 154 observations on the following 5 variables.

- Race race of the child, a factor with levels aboriginal and non-aboriginal.
- Gender the gender of the child, a factor with levels female and male.
- School the school type, a factor with levels F0 (primary), F1 (first), F2 (second) and F3 (third form).
- Learner how good is the child in learning things, a factor with levels average and slow.
- Absent number of days absent from school.

Objectives

The schooldays data will be used to see how the average number of absents differ according to the different levels of the factors of race, gender,schooling, and learner type.It may also be interesting to see whether different combinations of the levels of the factors differently affect the average number of absents. We will for example, see whether a specific type of learner, combined with a specific level of schooling affect the number of absences.Then some conclusions about how different levels of factors affect average number of absences can be drawn from this study and applied to the larger population from which the sample was taken, giving us an idea about how absences are affected by the levels of factors such as race, gender, schooling, and learner type.

Methods

Four graphical techniques will be employed: A plot of univariate factors, plots of two way interaction effects of factors,a compound residual plot,and a 95 percent family wise confidence level. A plot of univariate factors is used for showing the changes in a dependent variable as a result of the effects of the levels of different factors. Plots of two-way interaction effects of factors are useful for plotting the mean (or other summary) of the response for two-way combinations of factors, thereby illustrating possible interactions. A compound residual plot is useful for showcasing multiple plots in a single frame. The compound residual plot will showcase: a residual vs fitted,scale location,constant leverage, and QQ plot. The 95 percent family wise confidence level

plot will help graphically display the results of the Tukey honest significant differences test. Which will help make the interpretations on multiple confidence intervals much easier.

Comparing the differences in the mean number of absences across the different levels for gender,race,schooling, and learner type can be accomplished by using the plot of univariate factors. The plot of two way interaction effects will be used to examine how different combinations of the levels of the factors differingly affect the average number of absents and it will allow us to visually locate the presence of any potential interaction effects. By using the compound residual plot, insights can be given about the distribution and linearity of the schooldays data set. Along with this, an understanding of the spread and scedasticity of the data set can also be gained in order to test out different assumptions. Through the use of the 95 percent family wise confidence level, assessments of the Tukey honest significant differences will be made more comprehensible.

Results and Discussion

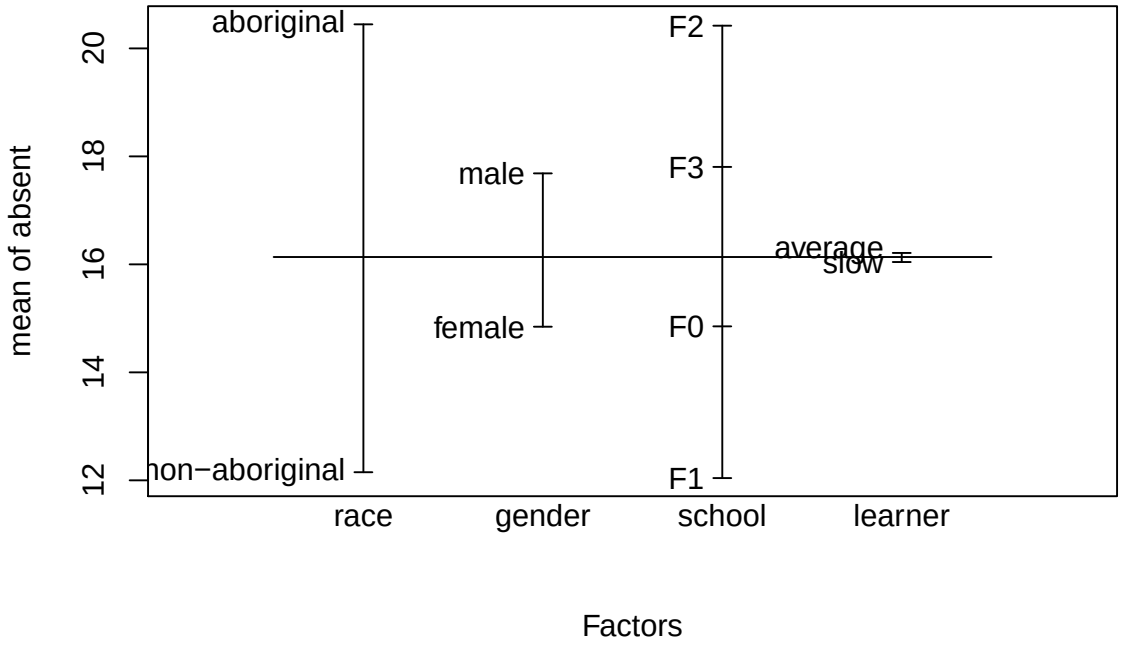

Figure 1: Plot of univariate factors

As can be seen in figure 1,the mean of the number of absents seems to largely differ mainly with race and school type, while gender differences and learner type seem to not influence the mean number of absences.

It may be interesting however to see whether different combinations of the levels of the factors differingly affect the average number of absents. We could for example see whether a specific type of learner, combined with a specific level of schooling affect the number of absences. Thus, let us plot the two-way interactions.

From figure 2 below, it seems that there is no interaction between race and gender when it comes to the mean number of absents. However, mean absents seem to decrease when the student is non-aboriginal independent of the gender. There also seems to be an interaction between race and learning type. That means that race may impact the average number of absents differently depending on which type of learner it is paired with.Gender also seems to impact mean absents differently depending on the interaction it has with learning type. When interacting with the varying levels of schooling, mean absents seem to impacted differently for race,gender, and learner.

```
##                          Df Sum Sq Mean Sq F value   Pr(>F)
## race                      1   2646  2645.7  13.258 0.000400 ***
## gender                    1    339   338.9   1.698 0.194985
```

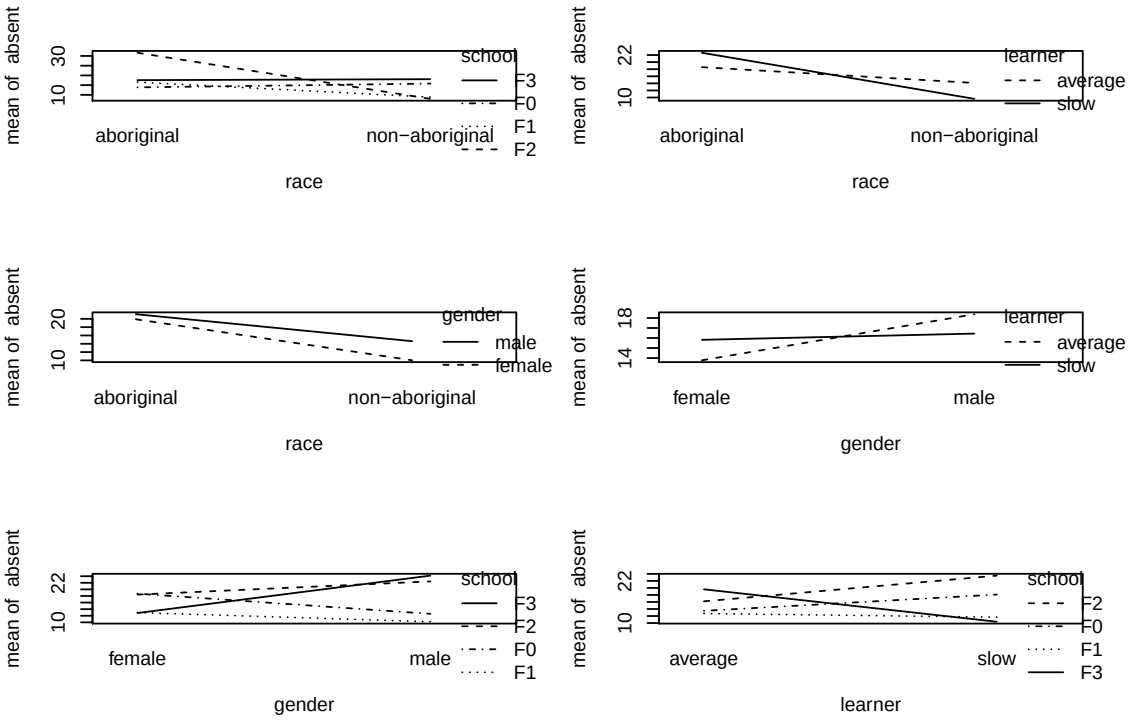

Figure 2: Plots of two way interaction effects of factors

```
## school                       3   1222   407.3   2.041 0.111672
## learner                      1     17    17.3   0.087 0.769087
## race:gender                  1    173   173.3   0.868 0.353295
## race:school                  3   3628  1209.2   6.059 0.000702 ***
## gender:school                3   1504   501.3   2.512 0.061747 .
## race:learner                 1     67    67.4   0.338 0.562291
## gender:learner               1      9     9.4   0.047 0.828522
## school:learner               3   1931   643.5   3.225 0.025011 *
## race:gender:school           3    206    68.7   0.344 0.793404
## race:gender:learner          1    388   387.6   1.942 0.165957
## race:school:learner          3   1419   473.0   2.370 0.073836 .
## gender:school:learner        3    609   203.1   1.018 0.387326
## race:gender:school:learner   3    202    67.4   0.338 0.798014
## Residuals                  122  24346   199.6
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

## Warning: not plotting observations with leverage one:
##   34, 62, 93, 112, 143, 144
```

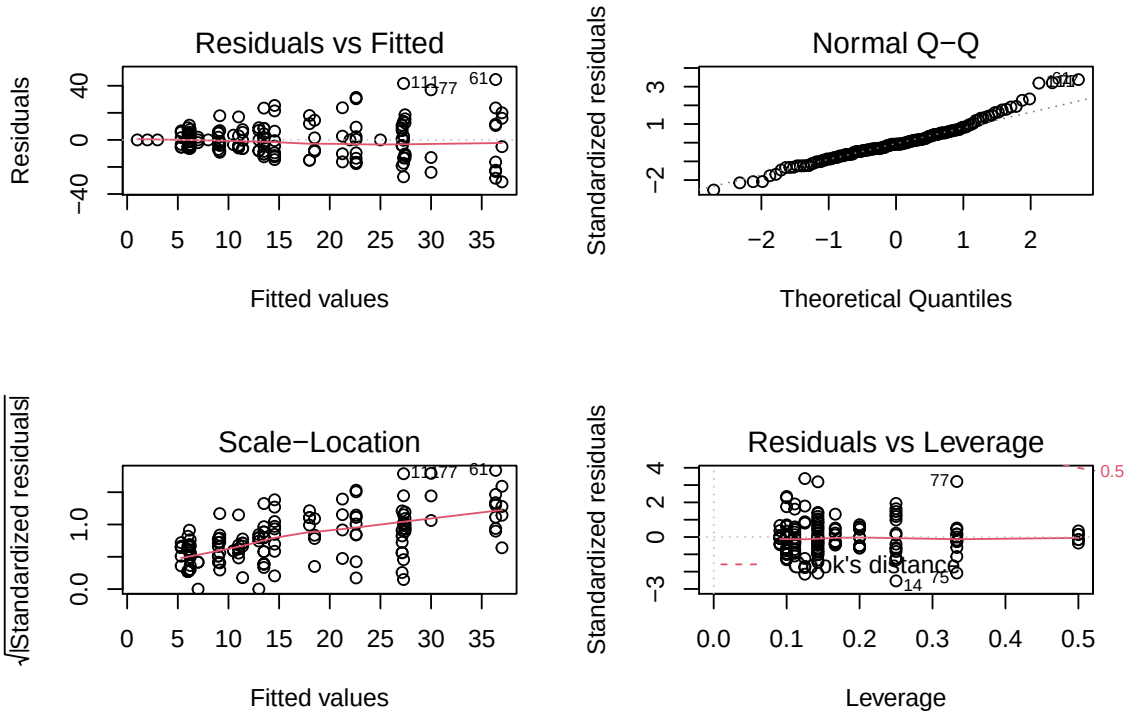

Figure 3: Compound residual plot

The above summary and figure 3 sheds more light on the data and the observations made from the graphs. Race seems to be the only main effect that effects the mean of absents. This is because it is the only main effect that is statistically significant at the 0.05 significance level. Only a few interactions seem to be statistically significant, these are the interactions between race with school, and school with learner. Furthermore, the QQ plot shows that the data is normally distributed and the residuals vs fitted plot shows that we can assume linearity due to the horizontal line at zero. The horizontal line at zero for the residuals vs leverage plot show that there are no influential observations that need to be removed. The scale location plot shows if residuals are spread equally along the ranges of predictors. It is good if a horizontal line with equally spread points is visible. In the above plot, this is the case, therefore Homogeneity of variance is assumed. As a result the anova regression will be used due to equal variance and normality being assumed.

```
##                  Df Sum Sq Mean Sq F value   Pr(>F)
## race              1   2646  2645.7  12.543 0.000538 ***
## school            3   1326   441.8   2.095 0.103583
## learner           1      4     3.8   0.018 0.893248
## race:school       3   3740  1246.6   5.910 0.000789 ***
## school:learner    3   1040   346.8   1.644 0.181891
## Residuals       142  29951   210.9
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

When the insignificant factors are removed from the anova test of the first regression, race and the interaction between race and school become the only significant effects as can be seen above.

```
##                Df Sum Sq Mean Sq F value   Pr(>F)
## race            1   2646  2645.7  12.461 0.000556 ***
## school          3   1326   441.8   2.081 0.105245
## race:school     3   3738  1246.1   5.869 0.000822 ***
## Residuals     146  30997   212.3
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

When the insignificant factors are removed from the anova test of the second regression, race and the interaction between race and school become the only significant effects as can be seen above.

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = absent ~ race + race * school, data = schooldays)
##
## $race
##                                 diff      lwr       upr     p adj
## non-aboriginal-aboriginal -8.295946 -12.9405 -3.651396 0.0005561
##
## $school
##            diff         lwr       upr     p adj
## F1-F0 -2.272485 -11.3819601  6.836990 0.9159894
## F2-F0  5.197258  -4.3339563 14.728472 0.4907290
## F3-F0  2.698227  -6.6870230 12.083478 0.8777444
## F2-F1  7.469743  -0.7527222 15.692208 0.0893594
## F3-F1  4.970712  -3.0821018 13.023527 0.3794697
## F3-F2 -2.499030 -11.0260063  6.027946 0.8715166
##
## $`race:school`
##                                           diff         lwr         upr
## non-aboriginal:F0-aboriginal:F0       1.9395604 -15.3199040 19.19902485
## aboriginal:F1-aboriginal:F0           2.7038462 -13.2605105 18.66820284
## non-aboriginal:F1-aboriginal:F0      -5.0247253 -20.0638431 10.01439259
## aboriginal:F2-aboriginal:F0          17.8038462   1.8394895 33.76820284
## non-aboriginal:F2-aboriginal:F0      -5.9017094 -22.2117125 10.40829374
## aboriginal:F3-aboriginal:F0           3.7252747 -12.0886352 19.53918468
## non-aboriginal:F3-aboriginal:F0       4.2038462 -11.7605105 20.16820284
## aboriginal:F1-non-aboriginal:F0       0.7642857 -14.8506992 16.37927064
## non-aboriginal:F1-non-aboriginal:F0  -6.9642857 -21.6320100  7.70343859
## aboriginal:F2-non-aboriginal:F0      15.8642857   0.2493008 31.47927064
## non-aboriginal:F2-non-aboriginal:F0  -7.8412698 -23.8094655  8.12692581
## aboriginal:F3-non-aboriginal:F0       1.7857143 -13.6754247 17.24685325
## non-aboriginal:F3-non-aboriginal:F0   2.2642857 -13.3506992 17.87927064
## non-aboriginal:F1-aboriginal:F1      -7.7285714 -20.8477829  5.39064002
## aboriginal:F2-aboriginal:F1          15.1000000   0.9296321 29.27036792
## non-aboriginal:F2-aboriginal:F1      -8.6055556 -23.1642246  5.95311352
## aboriginal:F3-aboriginal:F1           1.0214286 -12.9792282 15.02208537
## non-aboriginal:F3-aboriginal:F1       1.5000000 -12.6703679 15.67036792
## aboriginal:F2-non-aboriginal:F1      22.8285714   9.7093600 35.94778287
## non-aboriginal:F2-non-aboriginal:F1  -0.8769841 -14.4146806 12.66071239
## aboriginal:F3-non-aboriginal:F1       8.7500000  -4.1857169 21.68571693
## non-aboriginal:F3-non-aboriginal:F1   9.2285714  -3.8906400 22.34778287
## non-aboriginal:F2-aboriginal:F2     -23.7055556 -38.2642246 -9.14688648
## aboriginal:F3-aboriginal:F2         -14.0785714 -28.0792282 -0.07791463
## non-aboriginal:F3-aboriginal:F2     -13.6000000 -27.7703679  0.57036792
## aboriginal:F3-non-aboriginal:F2       9.6269841  -4.7665529 24.02052118
## non-aboriginal:F3-non-aboriginal:F2  10.1055556  -4.4531135 24.66422463
## non-aboriginal:F3-aboriginal:F3       0.4785714 -13.5220854 14.47922823
##                                         p adj
```

```
## non-aboriginal:F0-aboriginal:F0     0.9999706
## aboriginal:F1-aboriginal:F0         0.9995374
## non-aboriginal:F1-aboriginal:F0     0.9695362
## aboriginal:F2-aboriginal:F0         0.0174587
## non-aboriginal:F2-aboriginal:F0     0.9531213
## aboriginal:F3-aboriginal:F0         0.9961384
## non-aboriginal:F3-aboriginal:F0     0.9923451
## aboriginal:F1-non-aboriginal:F0     0.9999999
## non-aboriginal:F1-non-aboriginal:F0 0.8270213
## aboriginal:F2-non-aboriginal:F0     0.0435612
## non-aboriginal:F2-non-aboriginal:F0 0.8008800
## aboriginal:F3-non-aboriginal:F0     0.9999646
## non-aboriginal:F3-non-aboriginal:F0 0.9998353
## non-aboriginal:F1-aboriginal:F1     0.6131944
## aboriginal:F2-aboriginal:F1         0.0279099
## non-aboriginal:F2-aboriginal:F1     0.6090524
## aboriginal:F3-aboriginal:F1         0.9999985
## non-aboriginal:F3-aboriginal:F1     0.9999805
## aboriginal:F2-non-aboriginal:F1     0.0000090
## non-aboriginal:F2-non-aboriginal:F1 0.9999993
## aboriginal:F3-non-aboriginal:F1     0.4325838
## non-aboriginal:F3-non-aboriginal:F1 0.3801185
## non-aboriginal:F2-aboriginal:F2     0.0000424
## aboriginal:F3-aboriginal:F2         0.0476685
## non-aboriginal:F3-aboriginal:F2     0.0699405
## aboriginal:F3-non-aboriginal:F2     0.4477598
## non-aboriginal:F3-non-aboriginal:F2 0.3978905
## non-aboriginal:F3-aboriginal:F3     1.0000000
```

The Tukey honest significant differences test has been used above in order to allow a comparison of all pairs of levels of a factor whilst maintaining the nominal significance level at its specified value and producing adjusted confidence intervals for mean differences. For the factor of race, non aboriginal vs aboriginal seems to be significant. For the factor of schooling, none of the levels are significant. For the interactions between race and schooling, aboriginal:F2-aboriginal:F0, aboriginal:F2-non-aboriginal:F0, aboriginal:F2-aboriginal:F1,aboriginal:F2-non-aboriginal:F1,non-aboriginal:F2-aboriginal:F2,and aboriginal:F3-aboriginal:F2 are the only significant differences for the pairwise comparisons.

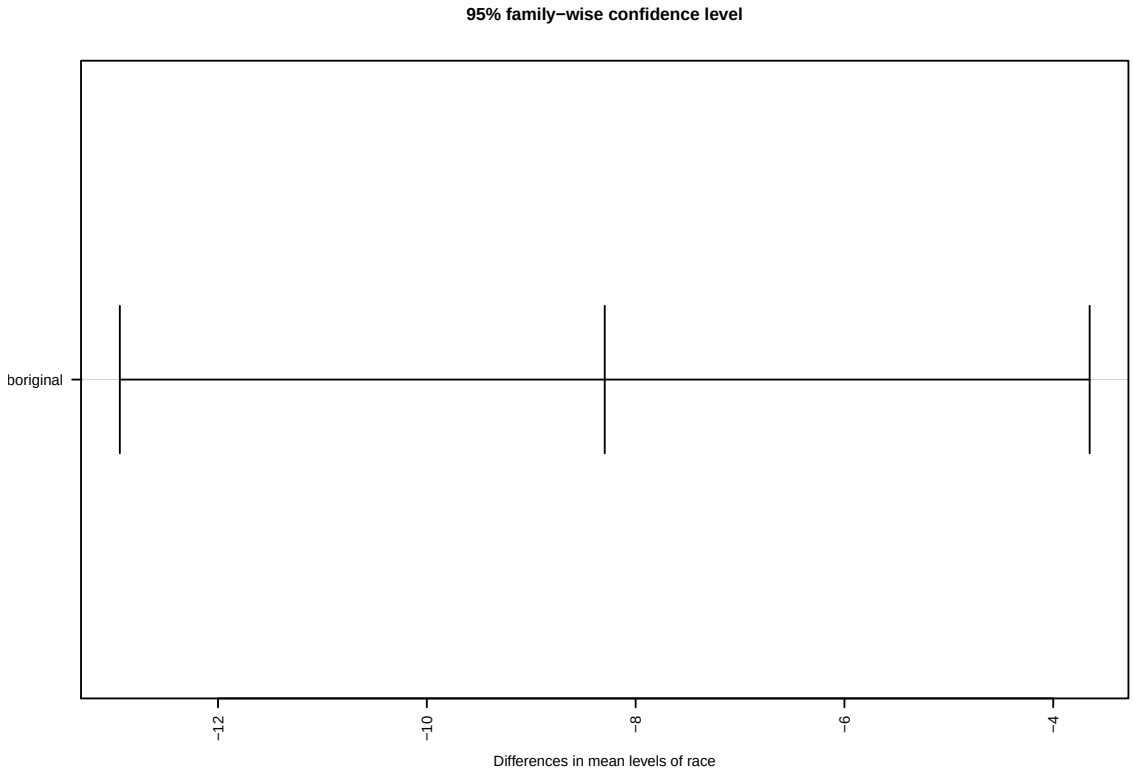

Figure 4: 95 percent family wise confidence level plot

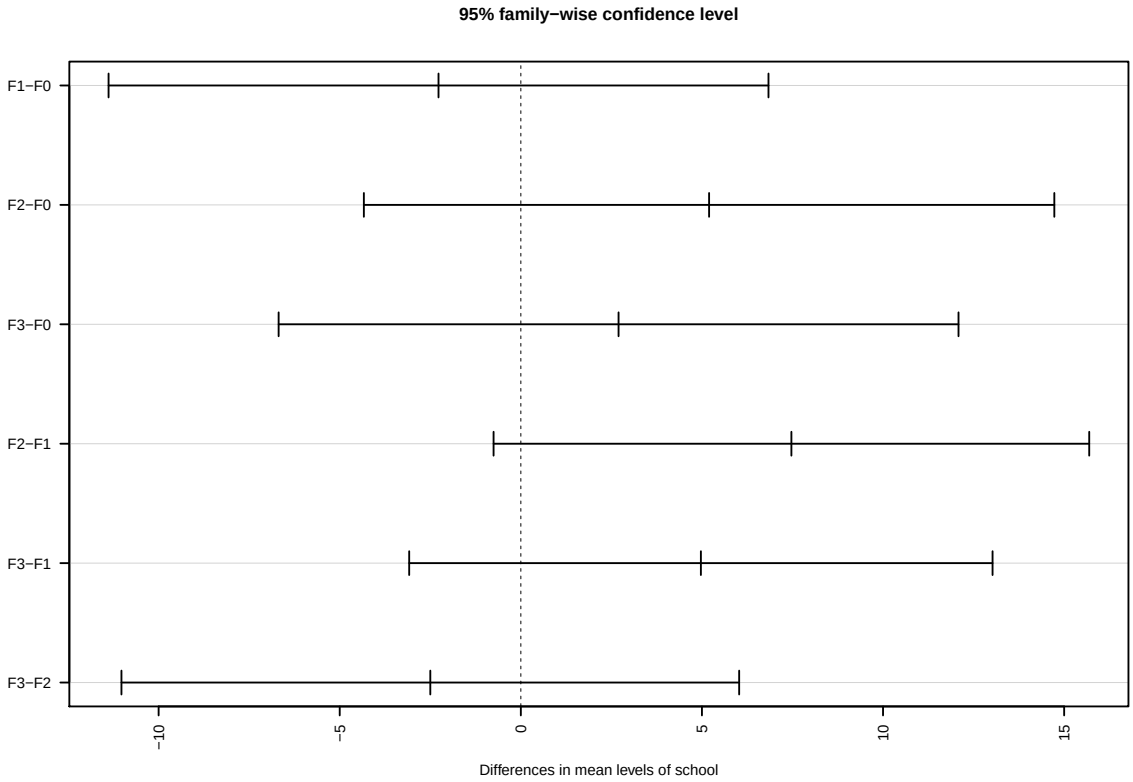

Figure 5: 95 percent family wise confidence level plot

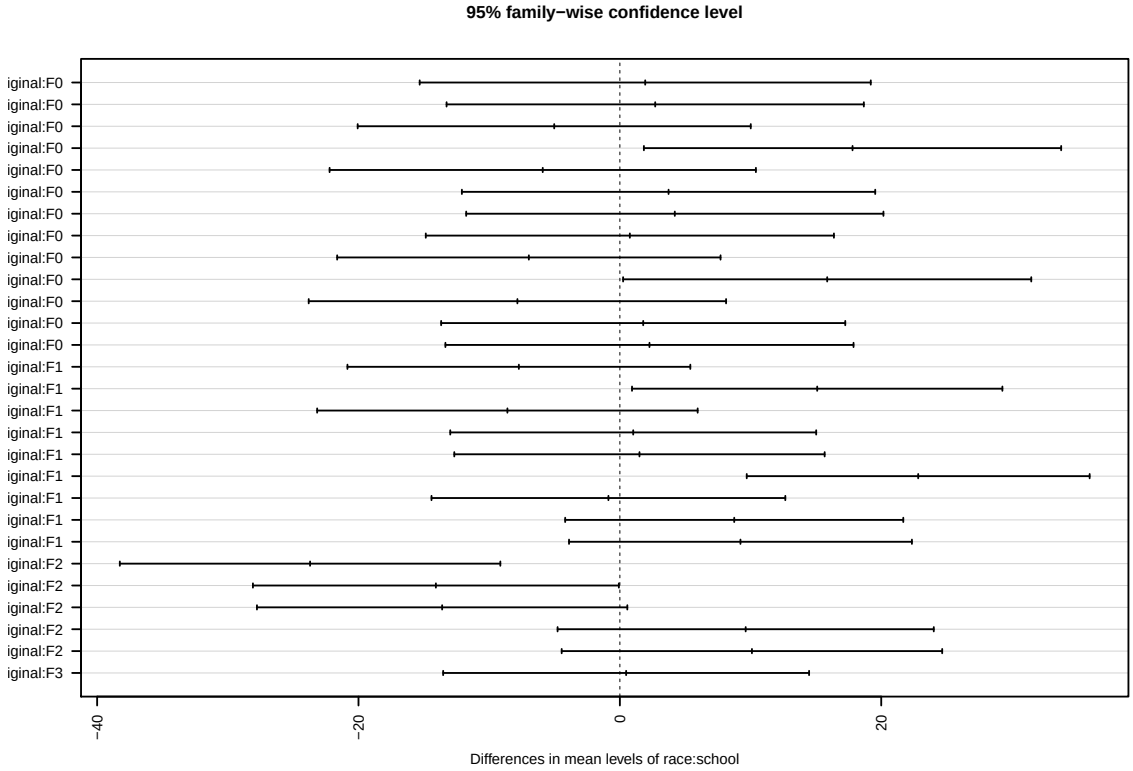

Figure 6: 95 percent family wise confidence level plot

The figures 4,5, and 6 are graphical representations of the Tukey honest significant differences test at the 95 percent confidence level. It is also useful as it makes things easier to understand visually and also gives more information such as the confidence intervals for the different comparisons.

```
## Tables of effects
##
##  race
##     aboriginal non-aboriginal
##           4.31         -3.986
## rep      74.00         80.000
##
##  school
##         F0     F1     F2     F3
##     -1.292 -3.565  3.905  1.406
## rep 27.000 48.000 38.000 41.000
##
##  race:school
##                 school
## race             F0     F1     F2     F3
##   aboriginal     -5.042 -0.032  7.542 -4.031
##   rep            13.000 20.000 20.000 21.000
##   non-aboriginal  4.682  0.023 -8.380  4.232
##   rep            14.000 28.000 18.000 20.000
```

The above table shows the effects for the different levels of interactions. For the significant comparison pairs, it helps quantify the extent of how the significant values impact the dependent variable.

Conclusion

The various analyses of variance used for un-balanced data all indicate that only the main effect of race and the interaction between race and school impacted the average number of absents.Detailed investigation of the race and race-aboriginal interaction effects by using the Tukey honest significant difference multiple comparison test suggests that the effects are largely produced by the difference between being an aboriginal or not, and the interactions between race(aboriginal and non-aboriginal) and the schooling levels of primary(F0),Form 1(F1),Form 2(F2), and Form 3(F3).The fact that race has an effect on the number of absents may suggest that there are socio-economic factors associated with different racial groups that have an impact on the number of absents.

Appendices

```
library(HSAUR3)
data("schooldays", package = "HSAUR3")
plot.design(schooldays)
```

```
data("schooldays")

sdays <- par(mfrow = c(3, 2))
with(schooldays, {
  interaction.plot(race,school,absent)
  interaction.plot(race, learner,absent)
  interaction.plot(race, gender,absent)
  interaction.plot(gender, learner,absent)
  interaction.plot(gender, school,absent)
  interaction.plot(learner, school,absent)
  }
)
```

```
par(sdays)

sdays1=aov(absent~race*gender*school*learner,data=schooldays)
summary(sdays1)


par(mfrow=c(2,2))
plot(sdays1)
```

```
par(mfrow=c(1,1))

sdays2=aov(absent~race+race*school+school*learner,data=schooldays)
summary(sdays2)

sdays3=aov(absent~race+race*school,data=schooldays)
summary(sdays3)

TukeyHSD(sdays3, conf.level=.95)

par(mfrow=c(1,1),cex=0.5)
plot(TukeyHSD(sdays3, conf.level=.95), las = 2,)
```

```
model.tables(sdays3)
```

References

[1] Hothorn, T and Everitt,B.(2014)*A Handbook of Statistical Analyses Using R* Florida:CRC Press.