# Homework 9, Snigdha Peddi

Exercises

Question 1. (Question 15.1 on pg. 295 in HSAUR, modified for clarity) Consider **alpha** dataset from the **coin** package. Compare the results when using **glht** and TukeyHSD (Refer to Chapter 5 for TukeyHSD).

INTRODUCTION

Alpha dataset from **coin** package has 97 observations and 2 variables *alength* and *elevel*.Allele length (alength variable) is a factor with levels "short","intermediate" and "long". Expression levels of alpha synuclein mRNA is presented in elevel variable.A One -way anova is performed and pairwise comparison is done using General Linear Hypothesis Test (glht) and Tukey HSD.

ANALYSIS AND DISCUSSION

The box plot of Allele length and Expression level shows increasing median values of expression level.
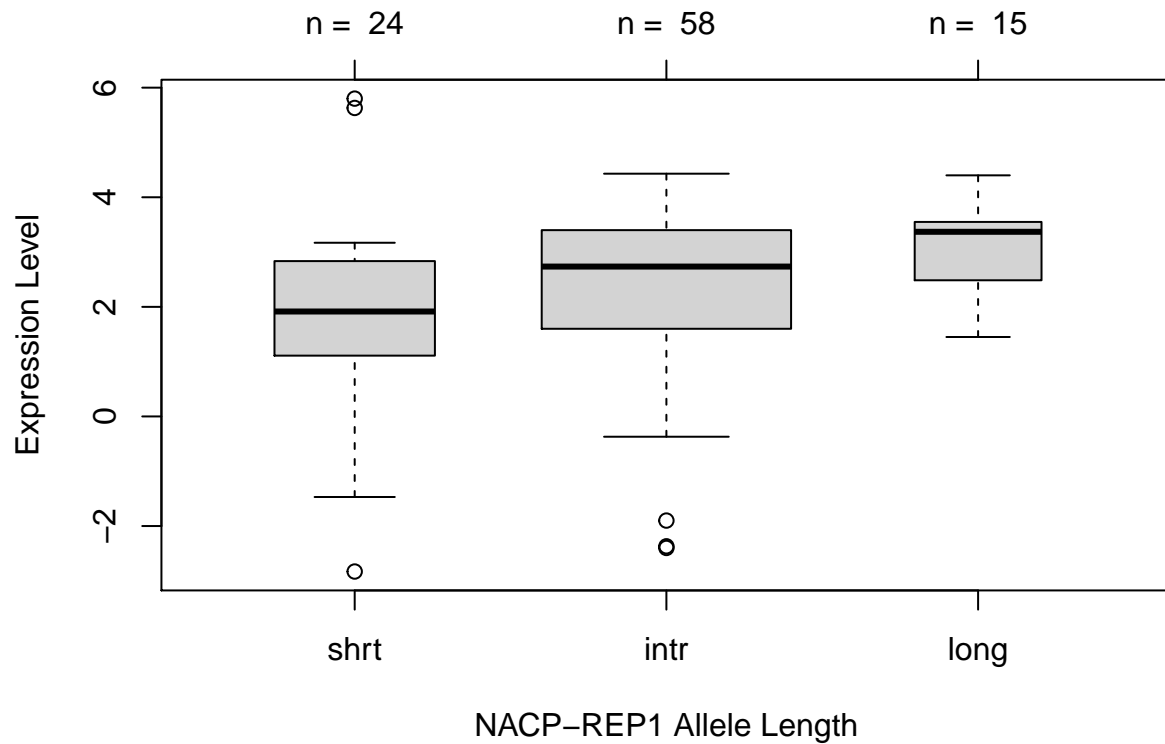
To check if there is any significant relation between the mean values of these levels a analysis of variance is performed.

```
    amod <- aov(elevel ~ alength, data = alpha)
```

A p value of 0.0786 indicate a marginal relation between expression level and allele length.

```
    amod_glht <- glht(amod, linfct = mcp(alength = "Tukey"))
```

To understand the significance of each level of the allele length a general linear hypothesis test is performed.A matrix of linear function is set up by Tukey method for factor allele length.The p values show that their is difference in expression between long and short lengths (0.0614) and and there is no difference in expression for these levels and intermediate level.

```
##             Df Sum Sq Mean Sq F value Pr(>F)
## alength      2  13.06   6.528   2.613 0.0786 .
## Residuals   94 234.85   2.498
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1


##
##   Simultaneous Tests for General Linear Hypotheses
##
## Multiple Comparisons of Means: Tukey Contrasts
##
##
## Fit: aov(formula = elevel ~ alength, data = alpha)
##
## Linear Hypotheses:
##                 Estimate Std. Error t value Pr(>|t|)
## intr - shrt == 0   0.4342     0.3836   1.132   0.4924
## long - shrt == 0   1.1888     0.5203   2.285   0.0614 .
## long - intr == 0   0.7546     0.4579   1.648   0.2270
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## (Adjusted p values reported -- single-step method)
```

Similarly, a pair wise comparison is done with Tukey's HSD (Honestly significant difference) test.
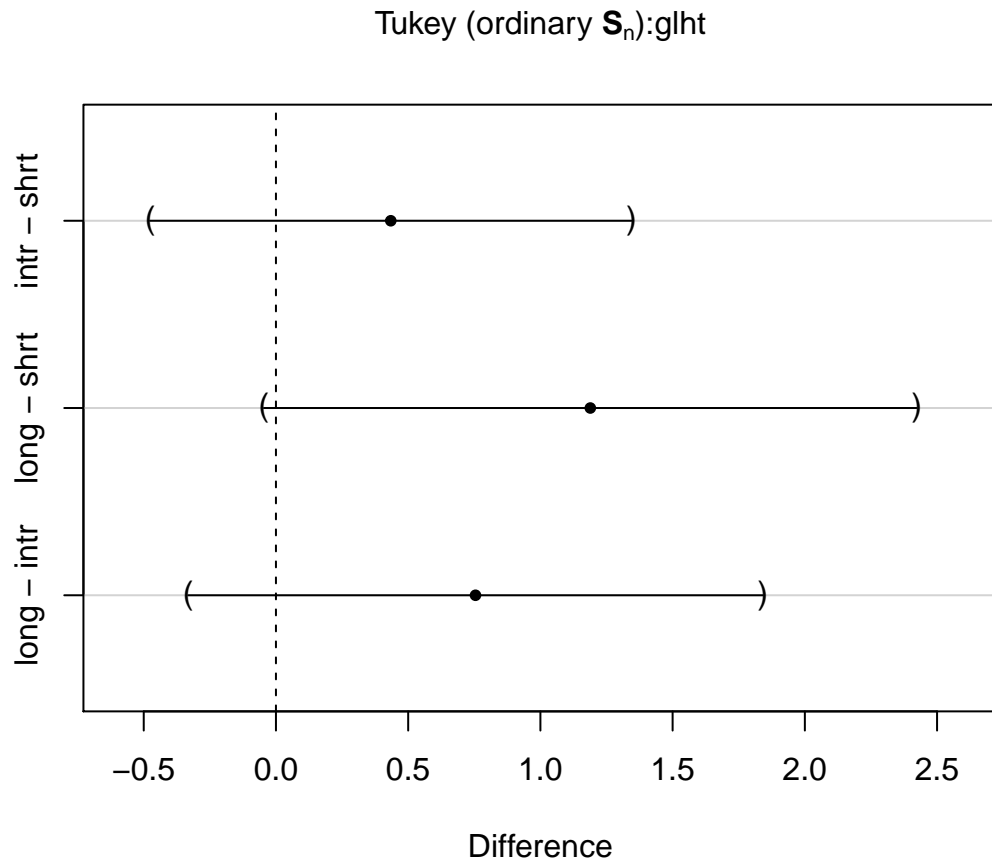
```
    TukeyHSD(amod)
```

The p values of TukeyHSD test are similar to the glht (a p value of 0.0628 for short-long alleles). This is clearly observed in the plot from both the methods.Both the test indicate that there is a difference in the expression level for short and long allele length and there is no difference between the short and intermediate alleles and intermediate and long alleles.

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = elevel ~ alength, data = alpha)
##
## $alength
```
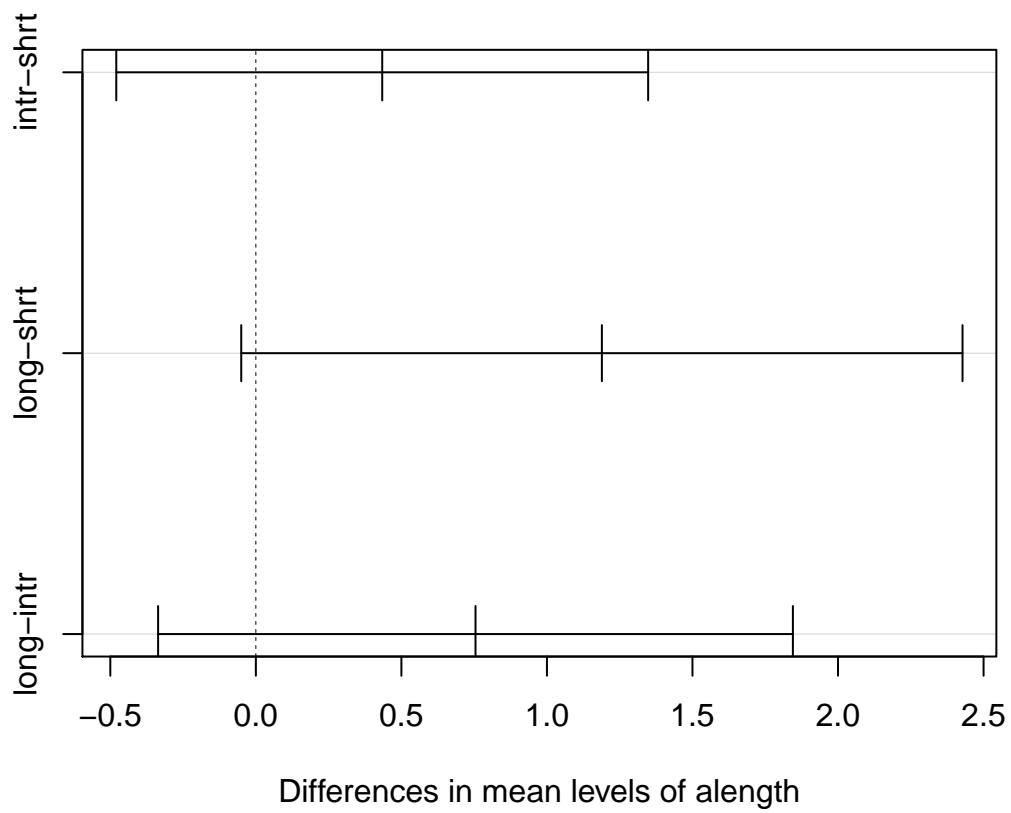
```
##                 diff          lwr       upr     p adj
## intr-shrt 0.4341523 -0.47943766 1.347742 0.4970962
## long-shrt 1.1887500 -0.05017513 2.427675 0.0628589
## long-intr 0.7545977 -0.33575201 1.844947 0.2307995
```

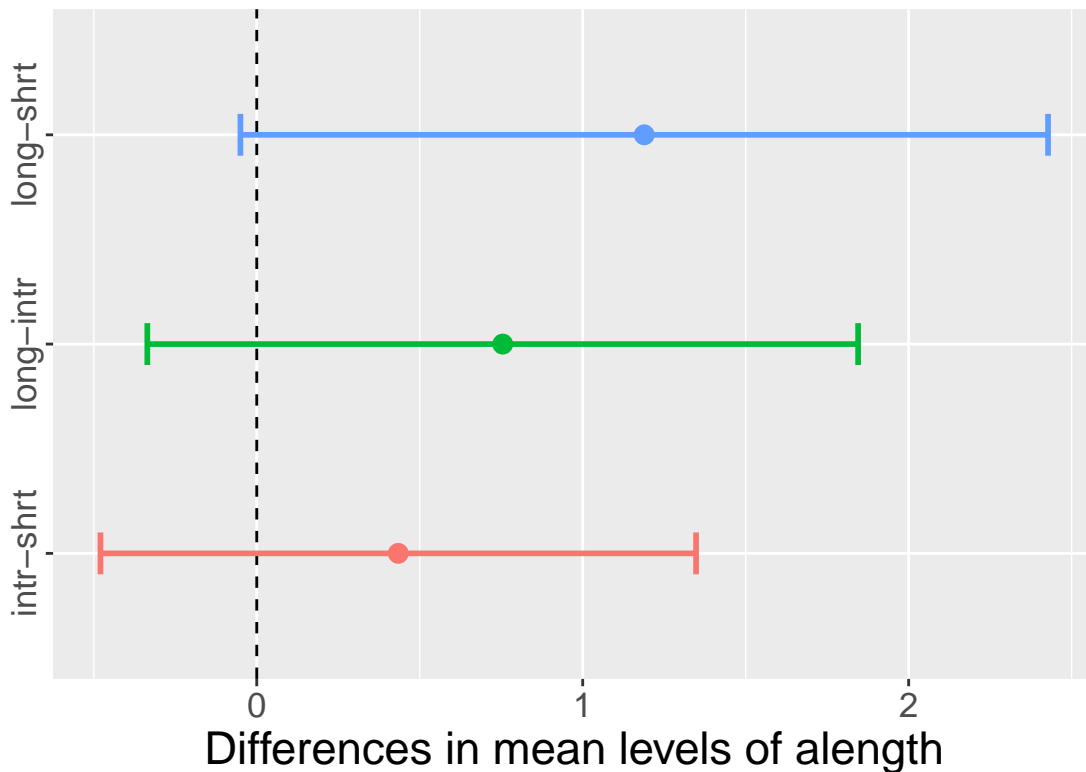**Difference in mean levels of allele length by General Linear Hypothesis test**

Tukey (ordinary $\mathbf{S}_n$):glht

95% family−wise confidence level

Differences in mean levels of alength

95% family–wise confidence level

long–shrt

long–intr

intr–shrt

Differences in mean levels of alength

REFERENCES

- Lecture code from Chapter 15, A Hand Book of Statistical Analyses Using R
- Lecture by Mike Marin, **ANOVA,ANOVA Multiple Comparision & Kruskal Wallis in R**, August 29,2013 (https://www.youtube.com/watch?v=lpdFr5SZR0Q)
- Lecture by Kyle Liu, **ANOVA(One and Two-way Between Subjects) and Tukey HSD in R**, March 23,2015 (https://www.youtube.com/watch?v=6-4mWkOgDtg)
- Lecture by Ole Forsberg, **Multiple Comparisons in R**, November 2,2014 (https://www.youtube.com/watch?v=nAs2Kq7d48s)
- ggiraphExtra documentation ,**ggHSD:Draw Tukey Honest Significant Differences Plot**,October 23,2020 (https://rdrr.io/cran/ggiraphExtra/man/ggHSD.html)

Question 2. (Question 15.2 on pg. 296 in HSAUR, modified for clarity) Consider **clouds** data from **HSAUR3** package

a. Read and write a report (no longer than one page) on the clouds data given in Chapter 15 section 15.3.3 from HSAUR Ed 3.
b. Consider the linear model fitted to the clouds data as summarized in Chapter 6, Figure 6.5. Set up a matrix K corresponding to the global null hypothesis that all interaction terms present in the model are zero. Test both the global hypothesis and all hypotheses corresponding to each of the interaction terms.
c. How does adjustment for multiple testing change which interactions are significant?

INTRODUCTION

Clouds data from **HSAUR3** package consists of experimental data investigating the use of massive amounts of Silver Iodide (100 to 1000 grams per cloud) in cloud seeding to increase rainfall.This data is collected from an area in Florida ,in the summer of 1975. 24 Days were judged suitable for seeding based on the measured suitability criterion. This data has 24 observations and 7 variables.These variables include *Seeding* which is factor indicating if seeding has occurred,*time* is the number of days after the first day of experiment, *sne* is the suitability criterion, *cloudcover* is percentage of experimental area covered with clouds, *prewetness* is the total rainfall in the target area one hour before seeding, *echomotion* is a factor showing if the radar echo is moving or stationary and *rainfall* is the amount of rain in cubic meters times 10ˆ7.

ANALYSIS AND RESULTS

Confidence band around regression line: The Clouds data is studied and presented by Brian S. Everitt and Torsten Hothorn in their book, A Handbook of Statistical Analyses using R. A Linear regression model is fit using rainfall as dependent variable and S-NE as independent variable .

$$rainfall = \beta_0 + \beta_1 sne_i + \epsilon_i \tag{1}$$

A confidence band for the predicted rainfall (around the regression line) is computed making sure that the type I error is controlled for all values of independent variable simultaneously.This can be formulated as a linear combination of the regression coefficients by multiplying with a matrix K to a grid of S-NE values. Here, values ranging from 1.5 to 4.5 with a step size of 0.25 is considered for S-Ne values.

$$\theta = (\beta_0, \beta_1) \tag{2}$$

$$K = \begin{bmatrix} 1 & 1.5 \\ 1 & 1.75 \\ 1 & 2.0 \\ . & . \\ . & . \\ 1 & 4.50 \end{bmatrix}$$

$$K\theta = (\beta_0 + \beta_1 1.50, \beta_0 + \beta_1 1.75.....\beta_0 + \beta_1 4.50) = v \tag{3}$$

Simultaneous confidence intervals formed from $k\theta$ forms a confidence band around the regression line. Above equation of v is used for the real values of S-Ne to compute the confidence band around the regression line of rainfall. A plot showing confidence bands for subset of data with and without seeding is presented below.
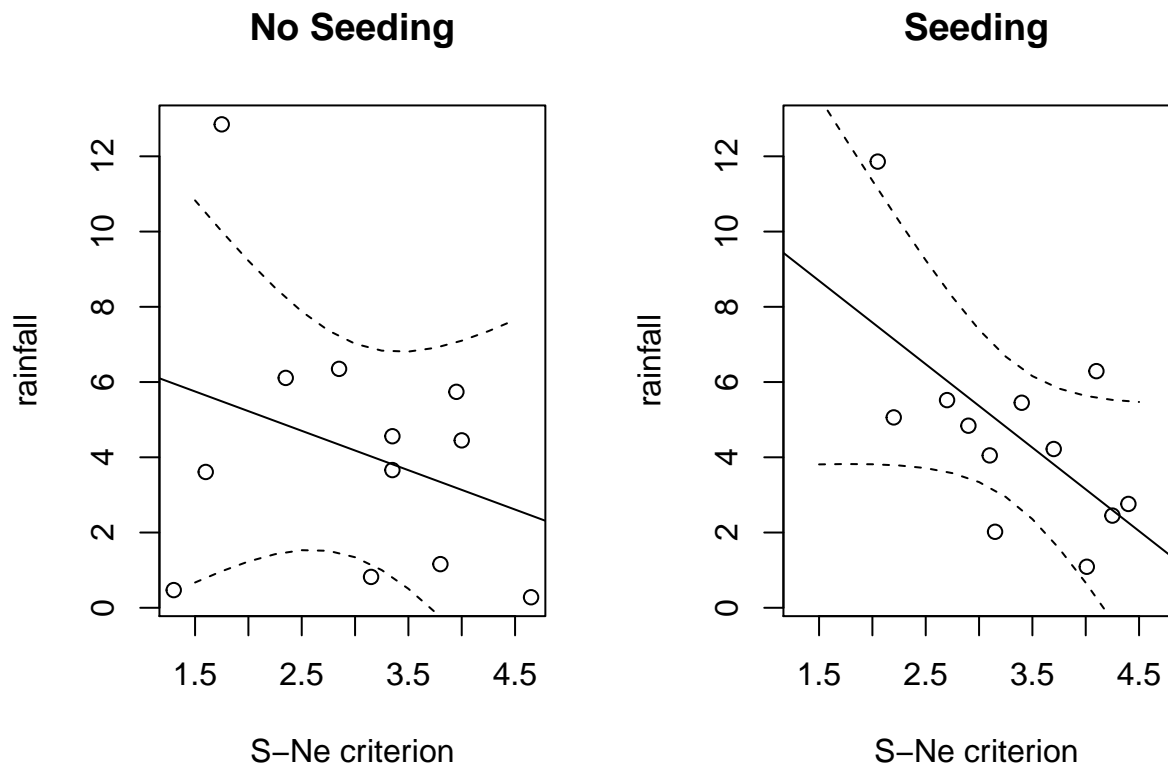
No Seeding   **Seeding**

From the plots it is clear that without seeding there is a high uncertainty for rainfall and with seeding the confidence band is tight indicating less uncertainty.

Global Null Hypothesis verses Hypothesis corresponding to each interaction term: Global null hypothesis is the first step to check if all the variables are significant and contributing at 95% confidence interval. For this purpose, two linear model with all the interaction terms and without any interaction terms were fit.

```
clouds_formula<- rainfall~seeding+seeding:(sne+cloudcover+prewetness+echomotion)+time
clouds_lm <-lm(clouds_formula,data=cloud)
```

```
clouds.mod <-lm(rainfall~.,data=clouds)
```

To verify whether the interaction terms are contributing to the model and effecting the rainfall an anova model is fit for the two linear models.

```
clouds.aov <-anova(clouds_lm,clouds.mod)
```

```
##
## Call:
## lm(formula = rainfall ~ ., data = clouds)
```

```
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.1820 -1.0170 -0.2685  0.7193  6.8074
##
## Coefficients:
##                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)            6.822352   2.451692   2.783   0.0128 *
## seedingyes             1.013201   1.203337   0.842   0.4115
## time                  -0.032125   0.028921  -1.111   0.2821
## sne                   -0.910860   0.751204  -1.213   0.2419
## cloudcover             0.005654   0.114899   0.049   0.9613
## prewetness             1.843790   2.758034   0.669   0.5128
## echomotionstationary   2.167987   1.579297   1.373   0.1877
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.836 on 17 degrees of freedom
## Multiple R-squared:  0.3849, Adjusted R-squared:  0.1678
## F-statistic: 1.773 on 6 and 17 DF,  p-value: 0.1647
```

A F-statistic value of 1.773 and a p-value of 0.1647 clearly indicate that the two models are not significantly different at 0.05% significance level and we accept null hypothesis. It indicates that the interaction terms do not have significant influence on the rainfall. A General Linear hypothesis test (glht) is performed to study the effect of individual interaction term simultaneously and to verify if the the model satisfies the global null hypothesis($\beta_0 = 0, \beta_1 = 0$ etc.).This is equal to linear hypothesis $K\beta = 0$, where $K = diag(11)$ in this case.

```
    new.mod <-glht(clouds_lm,linfct=K)
```

```
##
##   Simultaneous Tests for General Linear Hypotheses
##
## Fit: lm(formula = rainfall ~ seeding + seeding:(sne + cloudcover +
##     prewetness + echomotion) + time, data = clouds)
##
## Linear Hypotheses:
##                                  Estimate Std. Error t value Pr(>|t|)
## (Intercept) == 0                 -0.34624    2.78773  -0.124   1.0000
## seedingyes == 0                  15.68293    4.44627   3.527   0.0311 *
## time == 0                        -0.04497    0.02505  -1.795   0.5147
## seedingno:sne == 0                0.41981    0.84453   0.497   0.9994
## seedingyes:sne == 0              -2.77738    0.92837  -2.992   0.0815 .
## seedingno:cloudcover == 0         0.38786    0.21786   1.780   0.5234
## seedingyes:cloudcover == 0       -0.09839    0.11029  -0.892   0.9686
## seedingno:prewetness == 0         4.10834    3.60101   1.141   0.8932
## seedingyes:prewetness == 0        1.55127    2.69287   0.576   0.9981
## seedingno:echomotionstationary == 0   3.15281   1.93253   1.631   0.6169
## seedingyes:echomotionstationary == 0  2.59060   1.81726   1.426   0.7454
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## (Adjusted p values reported -- single-step method)
```

From the summary of the glht model it is clear that only *seedingyes* has significant effect on rainfall (p-value of 0.0315) and no other interaction terms are significant at 0.05% confidence interval.This rejects the Global Null hypothesis that all the coefficients are zero and do not have significant effect on the rainfall. Similar glht model is fit and multiplicity is adjusted for all the regression coefficients except for the intercept . The p-values indicate that only *seedingyes* has significant effect on rainfall and the Global null hypothesis is rejected.

```
##
##    Simultaneous Tests for General Linear Hypotheses
##
## Fit: lm(formula = rainfall ~ seeding + seeding:(sne + cloudcover +
##     prewetness + echomotion) + time, data = clouds)
##
## Linear Hypotheses:
##                                   Estimate Std. Error t value Pr(>|t|)
## seedingyes == 0                   15.68293    4.44627   3.527   0.0294 *
## time == 0                         -0.04497    0.02505  -1.795   0.5005
## seedingno:sne == 0                 0.41981    0.84453   0.497   0.9992
## seedingyes:sne == 0               -2.77738    0.92837  -2.992   0.0773 .
## seedingno:cloudcover == 0          0.38786    0.21786   1.780   0.5100
## seedingyes:cloudcover == 0        -0.09839    0.11029  -0.892   0.9657
## seedingno:prewetness == 0          4.10834    3.60101   1.141   0.8855
## seedingyes:prewetness == 0         1.55127    2.69287   0.576   0.9978
## seedingno:echomotionstationary == 0  3.15281    1.93253   1.631   0.6025
## seedingyes:echomotionstationary == 0 2.59060    1.81726   1.426   0.7330
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## (Adjusted p values reported -- single-step method)
```

Effect of adjustmants done for multiple testing on interaction terms: The p-values of the linear model with interaction terms are compared to the p-values of model adjusted for multiplicity. The table of p-values from linear model below shows that both the *seedingyes*(0.00372) and *seedingyes:sne*(0.01040) are significant at 95% confidence interval whereas the adjusted linear model shows that only *seedingyes*(0.0315) significant at 95% Confidence interval.Also, the pvalues of the multiplicity adjusted model are higher than the linear model.

p-values of linear model

```
##                                        p.value
## (Intercept)                        0.903055585
## seedingyes                         0.003715091
## time                               0.095897358
## seedingno:sne                      0.627420901
## seedingyes:sne                     0.010404243
## seedingno:cloudcover               0.098385030
## seedingyes:cloudcover              0.388538450
## seedingno:prewetness               0.274499083
## seedingyes:prewetness              0.574408155
## seedingno:echomotionstationary     0.126772119
## seedingyes:echomotionstationary    0.177565483
```

9

CONCLUSION

Cloud Seeding data is used to fit a linear model with S-Ne variable. This model is used to demonstrate the formulation to add confidence intervals to the regression line of rainfall and confidence intervals were plotted for subset of samples with and without seeding. Global Null hypothesis is tested for the linear model with interaction terms and without interaction terms. An Analysis of variance is performed between these models and a p-value of 0.1647 indicate that the Null hypothesis is accepted and there is no difference in the models.A General Linear hypothesis test (glht) is performed to study the effect of individual interaction term simultaneously and to verify if the the model satisfies the global null hypothesis. A comparison of p-values between this multiplicity adjusted model and the linear model with interaction terms have different significant variables at 0.05% significance level.Multiplicity adjusted model indicates only 'seedingyes' is significant but the linear model indicates both 'seedingyes' and 'seedingyes:sne' are significant.

REFERENCES

- Brian S. Everitt and Torsten Hothorn,*A Handbook of Statistical Analyses using R*,2010,Second Edition
- Lecture from Analytics University,*Significance of Global Null Hypothesis-using F-Statistics*, Oct 1),2015,(https://www.youtube.com/watch?v=095bOMlHK80)
- Lecture from The Pennsylvania State University, *Interpreting Output:summary(),anova(),aov(),and TukeyHSD()*,(https://online.stat.psu.edu/stat485/lesson/12/12.2)
- Torsten hothorn, *Additional multcomp Examples*,'Sep 23,2020',(https://cran.r-project.org/web/packages/multcomp/vignettes/multcomp-examples.pdf)

Question 3. (Question 15.3 on pg. 296 in HSAUR, modified for clarity) or the logistic regression model presented in Chapter 7 in Figure 7.7, perform a multiplicity adjusted test on all regression coefficients (except for the intercept) being zero. Do the conclusions drawn in Chapter 7 remain valid?

The summary of the binomial model indicates that education is highly significant and the gender is unimportant. With increase in number of years of education the probability of agreeing with the statement "Women should take care of running their homes and leave running the country up to men" is decreasing.

```
fm1<- cbind(agree,disagree)~gender+education
womensrole_glm_1 <-glm(fm1,data=womensrole,family=binomial())
```

```
##
## Call:
## glm(formula = fm1, family = binomial(), data = womensrole)
##
## Deviance Residuals:
##      Min       1Q    Median        3Q       Max
## -2.72544  -0.86302  -0.06525   0.84340   3.13315
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)    2.50937    0.18389  13.646   <2e-16 ***
## genderFemale  -0.01145    0.08415  -0.136    0.892
## education     -0.27062    0.01541 -17.560   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
```

```
##      Null deviance: 451.722  on 40  degrees of freedom
## Residual deviance:  64.007  on 38  degrees of freedom
## AIC: 208.07
##
## Number of Fisher Scoring iterations: 4
```

Another model is fit to check the influence of interaction between gender and education on the response. The summary of the model suggests that for Females, as the education increases the number of females agreeing with the statement decreases.

```
    fm2<- cbind(agree,disagree)~gender*education
    womensrole_glm_2 <-glm(fm2,data=womensrole,family=binomial())
```

```
##
## Call:
## glm(formula = fm2, family = binomial(), data = womensrole)
##
## Deviance Residuals:
##      Min        1Q     Median        3Q        Max
## -2.39097  -0.88062    0.01532    0.72783    2.45262
##
## Coefficients:
##                        Estimate Std. Error z value Pr(>|z|)
## (Intercept)             2.09820    0.23550   8.910  < 2e-16 ***
## genderFemale            0.90474    0.36007   2.513  0.01198 *
## education              -0.23403    0.02019 -11.592  < 2e-16 ***
## genderFemale:education -0.08138    0.03109  -2.617  0.00886 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##      Null deviance: 451.722  on 40  degrees of freedom
## Residual deviance:  57.103  on 37  degrees of freedom
## AIC: 203.16
##
## Number of Fisher Scoring iterations: 4
```

A multiplicity adjusted test is performed on all the regression coefficients using *glht* (general linear hypothesis test) function on glm model. A 3x4 diagonal matrix, is used as the linear function to test all the interaction terms simultaneously.

```
    womensrole.new.mod <-glht(womensrole_glm_2,linfct=K)
```

```
##
##    Simultaneous Tests for General Linear Hypotheses
##
## Fit: glm(formula = fm2, family = binomial(), data = womensrole)
##
## Linear Hypotheses:
```

11

```
##                            Estimate Std. Error z value Pr(>|z|)
## genderFemale == 0           0.90474    0.36007   2.513   0.0249 *
## education == 0             -0.23403    0.02019 -11.592   <0.001 ***
## genderFemale:education == 0 -0.08138    0.03109  -2.617   0.0180 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## (Adjusted p values reported -- single-step method)
```

Both the Generalized linear model and the multiplicity adjusted test indicate that all the variables and interaction terms are significant at 95% Confidence level.However, the pvalues of multiplicity adjusted test are higher than glm model (Ex: 0.0182>0.00886 for interaction term). And with one unit increase in years of education, 8% of Females agreeing to the statement "Women should take care of running their homes and leave running the country up to men" is decreasing.The conclusions drawn by Brian S.Everitt and Torsten Hothorn in their book, A Handbook of Statistical Analyses using R,chapter 7 are still valid for womensrole data when compared to multiplicity adjusted model.