Multi linear Regression: Cloud Data

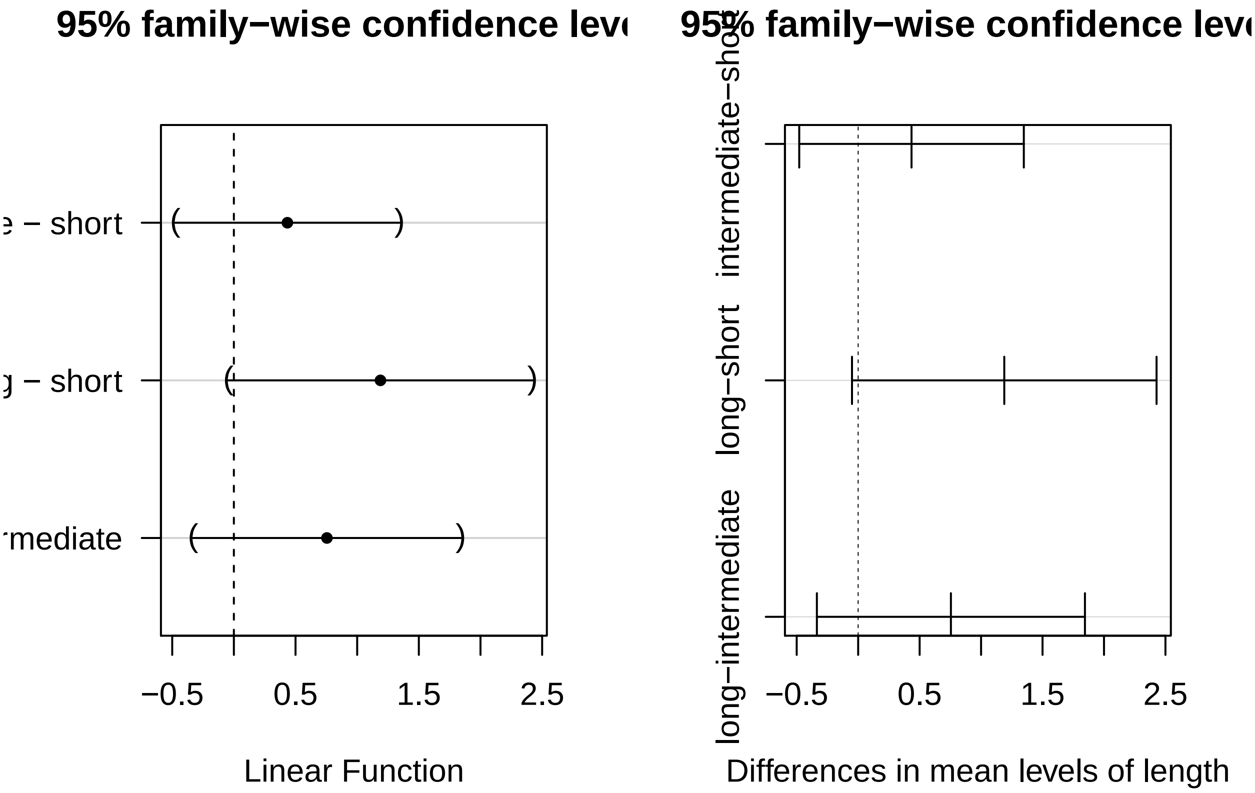
Yamuna Dhungana

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).

```
## 
##   Simultaneous Tests for General Linear Hypotheses
## 
## Multiple Comparisons of Means: Tukey Contrasts
## 
## 
## Fit: aov(formula = level ~ length, data = alpha_data)
## 
## Linear Hypotheses:
##                            Estimate Std. Error t value Pr(>|t|)  
## intermediate - short == 0    0.4342     0.3836   1.132   0.4924  
## long - short == 0            1.1888     0.5203   2.285   0.0614 .
## long - intermediate == 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)
```

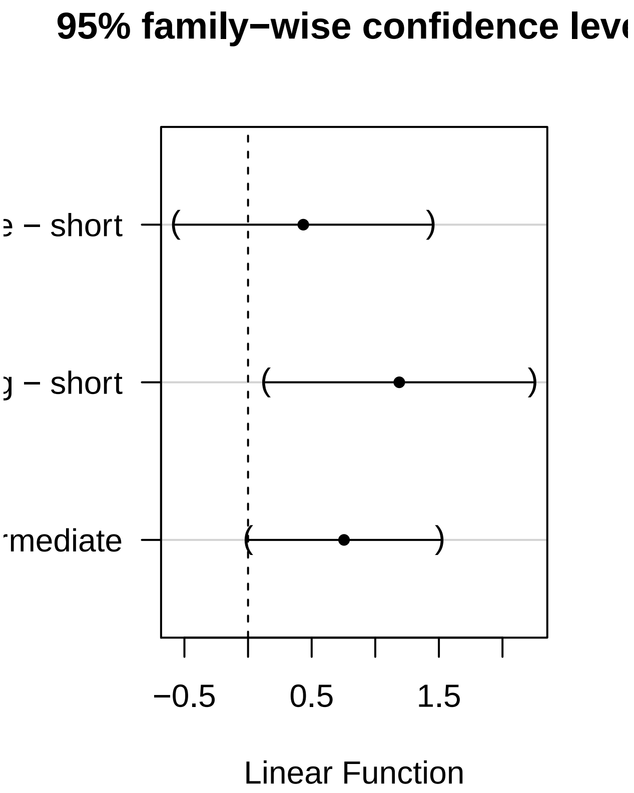
```
## 
##   Simultaneous Tests for General Linear Hypotheses
## 
## Multiple Comparisons of Means: Tukey Contrasts
## 
## 
## Fit: aov(formula = level ~ length, data = alpha_data)
## 
## Linear Hypotheses:
##                            Estimate Std. Error t value Pr(>|t|)  
## intermediate - short == 0    0.4342     0.4239   1.024   0.5594  
## long - short == 0            1.1888     0.4432   2.682   0.0227 *
## long - intermediate == 0     0.7546     0.3184   2.370   0.0501 .
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
## (Adjusted p values reported -- single-step method)
```

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



```
## intermediate - short          long - short  long - intermediate
##            0.4341523            1.1887500            0.7545977
```

```
##                       intermediate - short long - short long - intermediate
## intermediate - short            0.14717604    0.1041001         -0.04307591
## long - short                    0.10410012    0.2706603          0.16656020
## long - intermediate            -0.04307591    0.1665602          0.20963611
```



`Comparing the result with glht and TukeyHSD`

The data alpha in R is taken from the Coin library. Data alpha has two variables length and level.The variable length has three levels– short, intermediate, and long and the variable level is of alpha-synuclein mRNA. Various studies have linked alcohol dependence phenotypes to chromosome 4. One candidate gene is NACP (a non-amyloid component of plaques), coding for alpha-synuclein. B¨onsch et al. (2005) found longer alleles of NACP-REP1 in alcohol-dependent patients and report that the allele lengths show some association with levels of expressed alpha-synuclein mRNA in alcohol-dependent subjects. Here we are interested to find the comparison between the two test glht and TukeyHSD. We want to answer the question if there is any difference in the distribution of the expression levels among allele length groups. Firstly, the variables of the alpha data are converted into the data frame, then the model isc reated with anova. The model is then tested with the glht and TukeyHSD. With TukeyHSD, none of the pairwise comparison looks significant at 95% CI. Which means all the predictors have zero in their confidence intervals. As seen in the plot When 90% CI, there were significant values in long-short. Long-short is significant since it does not contain zero. which means there is some effect due the difference between long and short at 90% CI. Levels: long and short differ with Tukey's multiple comparison at 90%.Even though, Tukey HSD was adjusted for the heteroskedasticity and hence differs from the glht p-values. However, they both produced the similar results. Long - short is significant at 5% for glht.

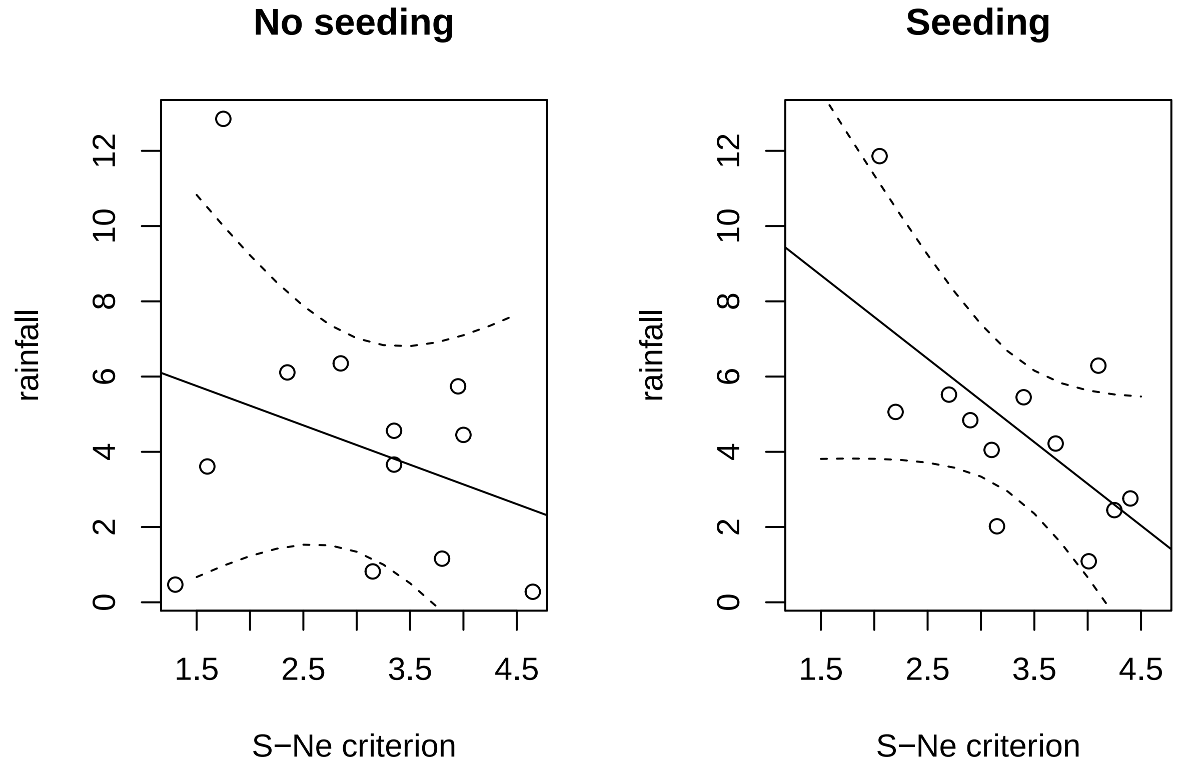
2. (Question 15.2 on pg. 296 in HSAUR, modified for clarity) Consider **clouds** data from **HSAUR3** package

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.

`clouds data from HSAUR3`

Weather modification, or cloud seeding, is the treatment of individual clouds or storm systems with various inorganic and organic materials in the hope of achieving an increase in rainfall. The data used in Cloud in R were collected in the summer of 1975 from an experiment to investigate the use of massive amounts of silver iodide (100 to 1000 grams per cloud) in cloud seeding to increase rainfall (Woodley et al., 1977). In the experiment, which was conducted in an area of Florida, 24 days were judged suitable for seeding on the basis that a measured suitability criterion, denoted S-Ne, was not less than 1.5. Here S is the 'seedability', the difference between the maximum height of a cloud if seeded and the same cloud if not seeded predicted by a suitable cloud model, and Ne is the number of this quantity biases the decision for experimentation against naturally rainy days. Consequently,optimal days for seeding are those on which seedability is large and the natural rainfall early in the day is small. On suitable days, a decision was taken at random as to whether to seed or not. For each day the following variables were measured: seeding: a factor indicating whether seeding action occurred (yes or no), time: number of days after the first day of the experiment, cloudcover: the percentage cloud cover in the experimental area, measured using radar, prewetness: the total rainfall in the target area one hour before seeding (in cubic metres ×107), echomotion: a factor showing whether the radar echo was moving or stationary, rainfall: the amount of rain in cubic metres ×107.

The objective in analysing these data is to see how rainfall is related to the explanatory variables and, in particular, to determine the effectiveness of seeding. The method to be used is multiple linear regression. Source: Brian S. Everitt, Torsten Hothorn - A handbook of statistical analyses using R-CRC (2010) A linear model was fitted with the dependent and independent variable—rainfall and S-Ne respectively. Around the linear line, the confidence band was also estimated. To estimate the confidence interval without increasing the type-1 error (False positive) requires the multiplication of the regression coefficients with a K- matrix. The linear model is fitted, setting the K matrix, and plotting both the regression line and confidence interval. It appears that the rainfall and S-ne values have an impact, So the graphical representation of it is shown. The scatter plot with the seeding and no seeding is plotted. The relationship of no seeding appears weaker between rainfall and S-ne values. Also, seeding has a stronger relationship. It appears that there is more uncertainty without seeding compared to the seeding.

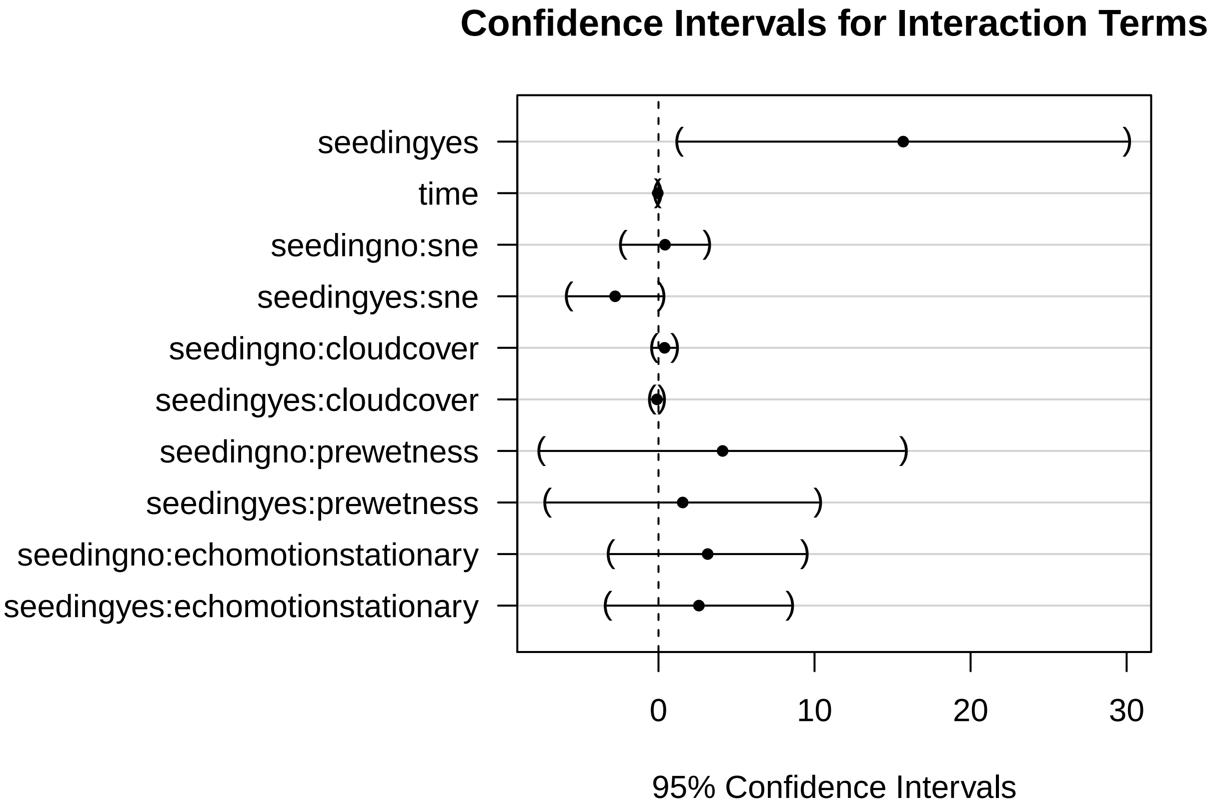


b. Consider the linear model fitted to the clouds data as summarized in Chapter 6, Figure 6.5. Set up a matrix K corresponding to the global null hypothesis that all interaction terms present in the model are zero. Test both the global hypothesis and all hypotheses corresponding to each of the interaction terms.

```
## 
## Call:
## lm(formula = clouds_formula, data = clouds)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -2.5259 -1.1486 -0.2704  1.0401  4.3913 
## 
## Coefficients:
##                                 Estimate Std. Error t value Pr(>|t|)   
## (Intercept)                     -0.34624    2.78773  -0.124  0.90306   
## seedingyes                      15.68293    4.44627   3.527  0.00372 **
## time                            -0.04497    0.02505  -1.795  0.09590 . 
## seedingno:sne                    0.41981    0.84453   0.497  0.62742   
## seedingyes:sne                  -2.77738    0.92837  -2.992  0.01040 * 
## seedingno:cloudcover             0.38786    0.21786   1.780  0.09839 . 
## seedingyes:cloudcover           -0.09839    0.11029  -0.892  0.38854   
## seedingno:prewetness             4.10834    3.60101   1.141  0.27450   
## seedingyes:prewetness            1.55127    2.69287   0.576  0.57441   
## seedingno:echomotionstationary   3.15281    1.93253   1.631  0.12677   
## seedingyes:echomotionstationary  2.59060    1.81726   1.426  0.17757   
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 2.205 on 13 degrees of freedom
## Multiple R-squared:  0.7158, Adjusted R-squared:  0.4972 
## F-statistic: 3.274 on 10 and 13 DF,  p-value: 0.02431
```

```
##                                 [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9]
## seedingyes                         0    1    0    0    0    0    0    0    0
## time                               0    0    1    0    0    0    0    0    0
## seedingno:sne                      0    0    0    1    0    0    0    0    0
## seedingyes:sne                     0    0    0    0    1    0    0    0    0
## seedingno:cloudcover               0    0    0    0    0    1    0    0    0
## seedingyes:cloudcover              0    0    0    0    0    0    1    0    0
## seedingno:prewetness               0    0    0    0    0    0    0    1    0
## seedingyes:prewetness              0    0    0    0    0    0    0    0    1
## seedingno:echomotionstationary     0    0    0    0    0    0    0    0    0
## seedingyes:echomotionstationary    0    0    0    0    0    0    0    0    0
##                                 [,10] [,11]
## seedingyes                          0     0
## time                                0     0
## seedingno:sne                       0     0
## seedingyes:sne                      0     0
## seedingno:cloudcover                0     0
## seedingyes:cloudcover               0     0
## seedingno:prewetness                0     0
## seedingyes:prewetness               0     0
## seedingno:echomotionstationary      1     0
## seedingyes:echomotionstationary     0     1
```

```
##
##   Simultaneous Tests for General Linear Hypotheses
##
## Fit: aov(formula = clouds_formula, data = clouds)
##
## Linear Hypotheses:
##                                      Estimate Std. Error t value Pr(>|t|)
## seedingyes == 0                      15.68293    4.44627   3.527   0.0293 *
## time == 0                            -0.04497    0.02505  -1.795   0.5009
## seedingno:sne == 0                    0.41981    0.84453   0.497   0.9992
## seedingyes:sne == 0                  -2.77738    0.92837  -2.992   0.0770 .
## seedingno:cloudcover == 0             0.38786    0.21786   1.780   0.5096
## seedingyes:cloudcover == 0           -0.09839    0.11029  -0.892   0.9656
## 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   0.6032
## seedingyes:echomotionstationary == 0  2.59060    1.81726   1.426   0.7331
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## (Adjusted p values reported -- single-step method)
```



The glht function was utilized to determine whether the interactions are equal to zero. The summary found that only one interaction term, seedingyes:sne, is not equal to zero, significant at the alpha<0.1 level. The corresponding graph shows how close the right-boundary of the confidence interval for seedingyes:sne is to 0, confirming the results from the summary. The graph also shows that seedingno:cloudcover and seedingyes:cloudcover also have boundaries close to 0. However, these confidence intervals have much smaller ranges which is likely why the summary is not finding the p-values anywhere near significant.

c. How does adjustment for multiple testing change which interactions are significant?

A summary of the original model finds the variables seedingyes, time, seedingyes:sne, and seedingno:cloudcover to be significant for fitting the model at the alpha<0.1 level. A summary of the cl_glht model, after removing the variables that were not interaction terms, only found seedingyes:sne to be significant at the alpha<0.1 level.

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?

```
##
## Call:
## glm(formula = formula, 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
```

```
##
##   Simultaneous Tests for General Linear Hypotheses
##
## Fit: glm(formula = formula, family = binomial(), data = womensrole)
##
## Linear Hypotheses:
##                             Estimate Std. Error z value Pr(>|z|)
## genderFemale == 0            0.90474    0.36007   2.513   0.0244 *
## education == 0              -0.23403    0.02019 -11.592   <0.001 ***
## genderFemale:education == 0 -0.08138    0.03109  -2.617   0.0177 *
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
## (Adjusted p values reported -- single-step method)
```

Varyfication of the result for womensrole dataset

The data womensrole is taken from a survey form 1917/1975 asking both female and male responders about their opinion on the satatemets: women should take care if running their homes and running the country up to men. The data in R is taken from the HASUR3 package. The dataset has total 42 observations with four variables. The gender variable has the two level male and female.The questions of interest here are whether the responses of men and women differ and how years of education affect the response.

The glht function was fitted as original model from chapter 7, excluding the intercept term. The original summary of the model found that gender, education, and the interaction gender:education were significant in fitting the model at the alpha<0.05 level. A summary of the glht function finds that gender, education, and gender:education are significant at the alpha<0.05 level and thus are not equal to 0. This makes them significant variables and the conclusions drawn in the original summary remain valid.