Chapter 6 in Everitt and Hothorn (2010) Simple and Multiple Linear Regression

Start R.

If you were not able to edit Rprofile.site, load the HSAUR2 and Rcmdr either using the commands: library(HSAUR2);library(Rcmdr)
or from the R Console using the menu Packages > Load package ... > select HSAUR2 and Rcmdr > Ok

Estimating the Age of the Universe

The data are in the gamair package, so you may need to install that package from the R Console.

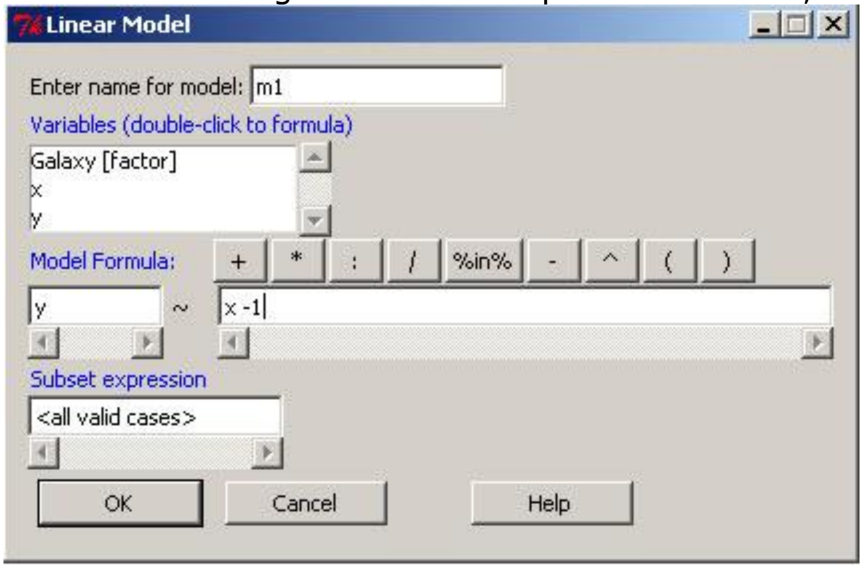
From the R Commander menu, select tools > Load package(s)... > gamair > click OK

From the R Commander menus select Data > Data in packages > Read data set from an attached package... > double click on gamair, select hubble, and click ok.

To see a description, from the R commander menu select Data > Active data set > Help on active data set (if available)

Click View data set to view it.

From the menu, select Statistics > Fit models > Linear model... > enter name m1, double click on y to add it to the left side of the model equation, double click on x to add it to the right side of the equation, enter -1 in the left side indicating that an intercept is not needed, and click Ok.



Linear Model

Enter name for model: m1

Variables (double-click to formula)

Galaxy [factor]
x
y

Model Formula: y ~ x -1

Subset expression

<all valid cases>

```
Call:
lm(formula = y ~ x - 1, data = hubble)
Residuals:
    Min      1Q  Median      3Q     Max
-736.49 -132.52  -19.00  172.18  557.98
Coefficients:
  Estimate Std. Error t value Pr(>|t|)
x   76.581      3.965   19.32 1.03e-15 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 258.9 on 23 degrees of freedom
Multiple R-squared: 0.9419,    Adjusted R-squared: 0.9394
F-statistic: 373.1 on 1 and 23 DF,  p-value: 1.032e-15
```

Using R as a calculator, enter the following statements into the Script Window and Submit them

```
mpc=3.09e19 # mega-parsec
ysec=60^2*24*365.25 # seconds per year
mpcYear= mpc/ysec
1/(coef(m1)/mpcYear)
```

12785935335 An estimated age of the universe of about 12.8 billion years

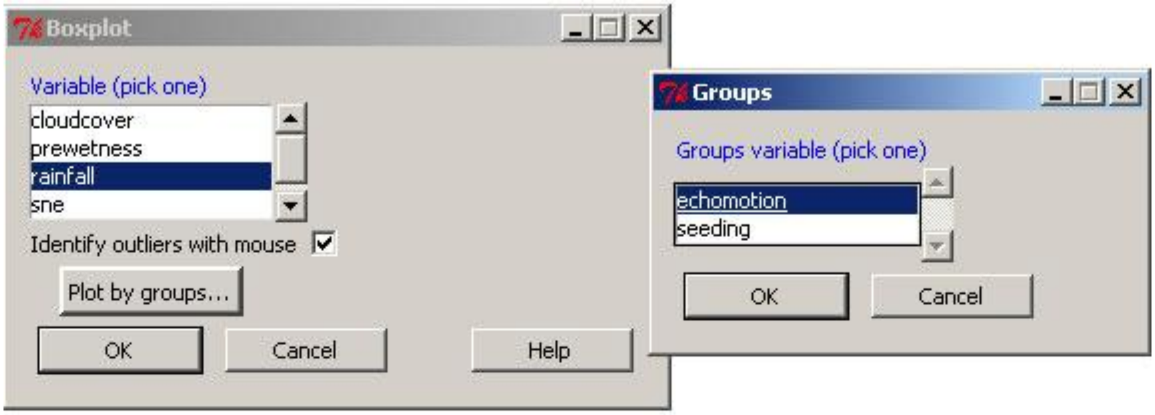
Cloud Seeding

From the R Commander menus select Data > Data in packages > Read data set from an attached package... >
Double click on HSAUR2 and select clouds, then click OK.

To see a description, from the R commander menu select Data > Active data set > Help on active data set (if available)

Click View data set to view it.

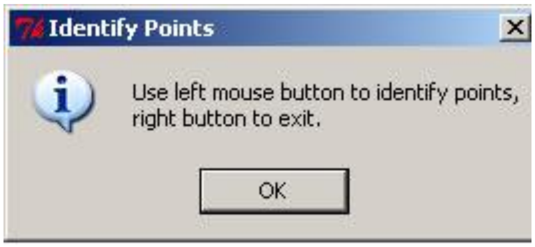
From the menu, select Graphs > Boxplot ... > select variable=rainfall, click Identify outliers, click Plot by groups... > select echomotion and click Ok > click Ok again.



Variable (pick one)
cloudcover
prewetness
rainfall
sne

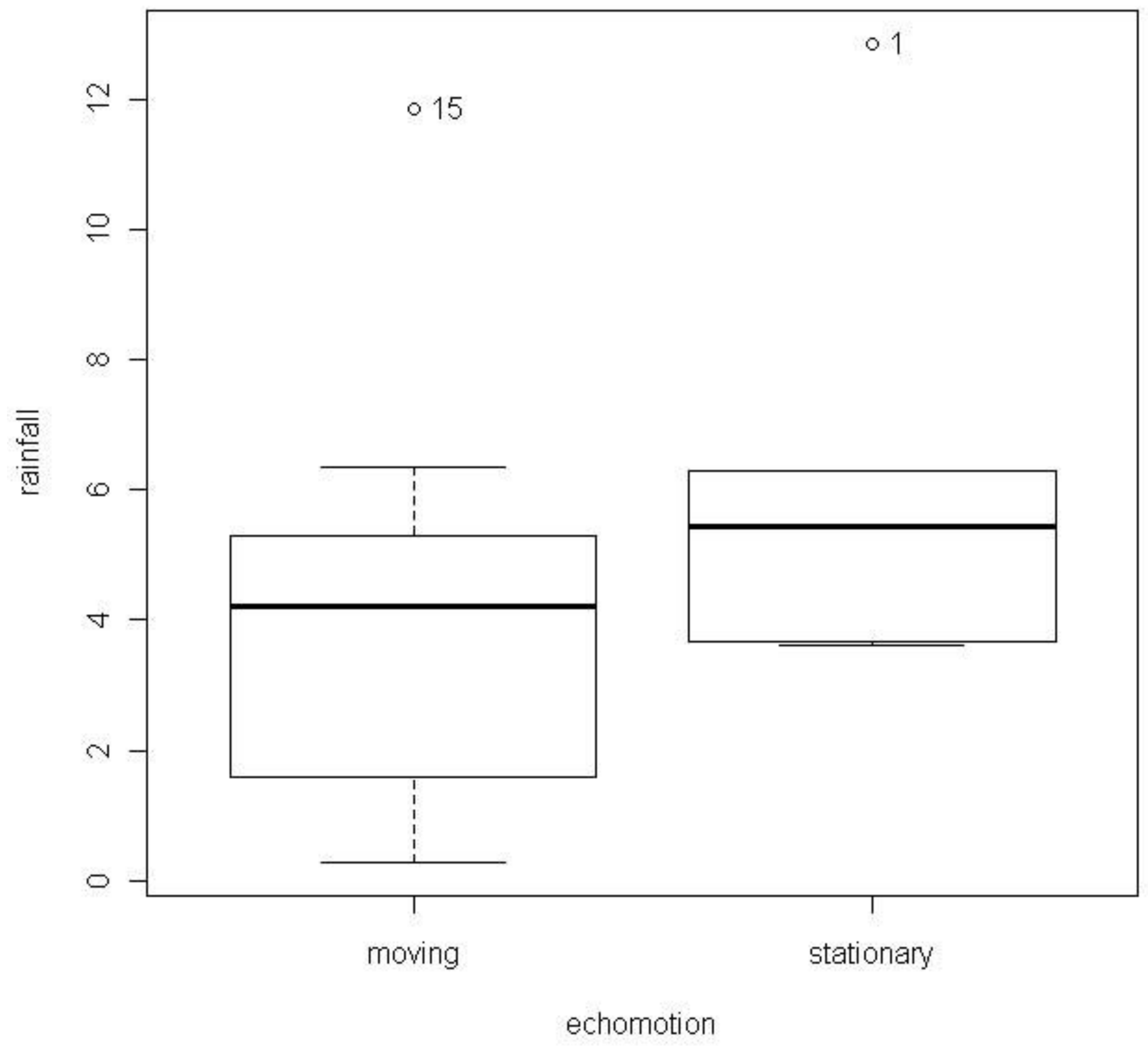
Identify outliers with mouse ☑

Groups variable (pick one)
echomotion
seeding



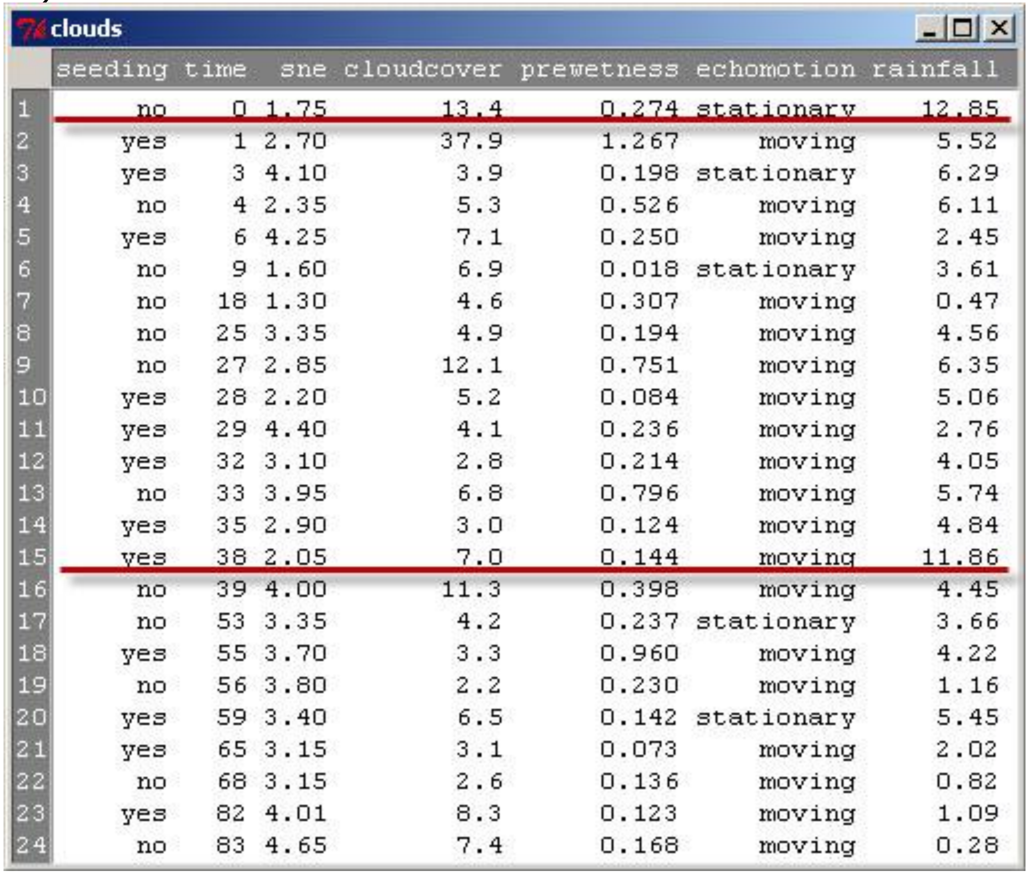
Use left mouse button to identify points, right button to exit.

Click on the outliers (circles) with the left mouse button to identify them and then click the right mouse button to exit. failure to exit will lock up R and it will have to be restarted.



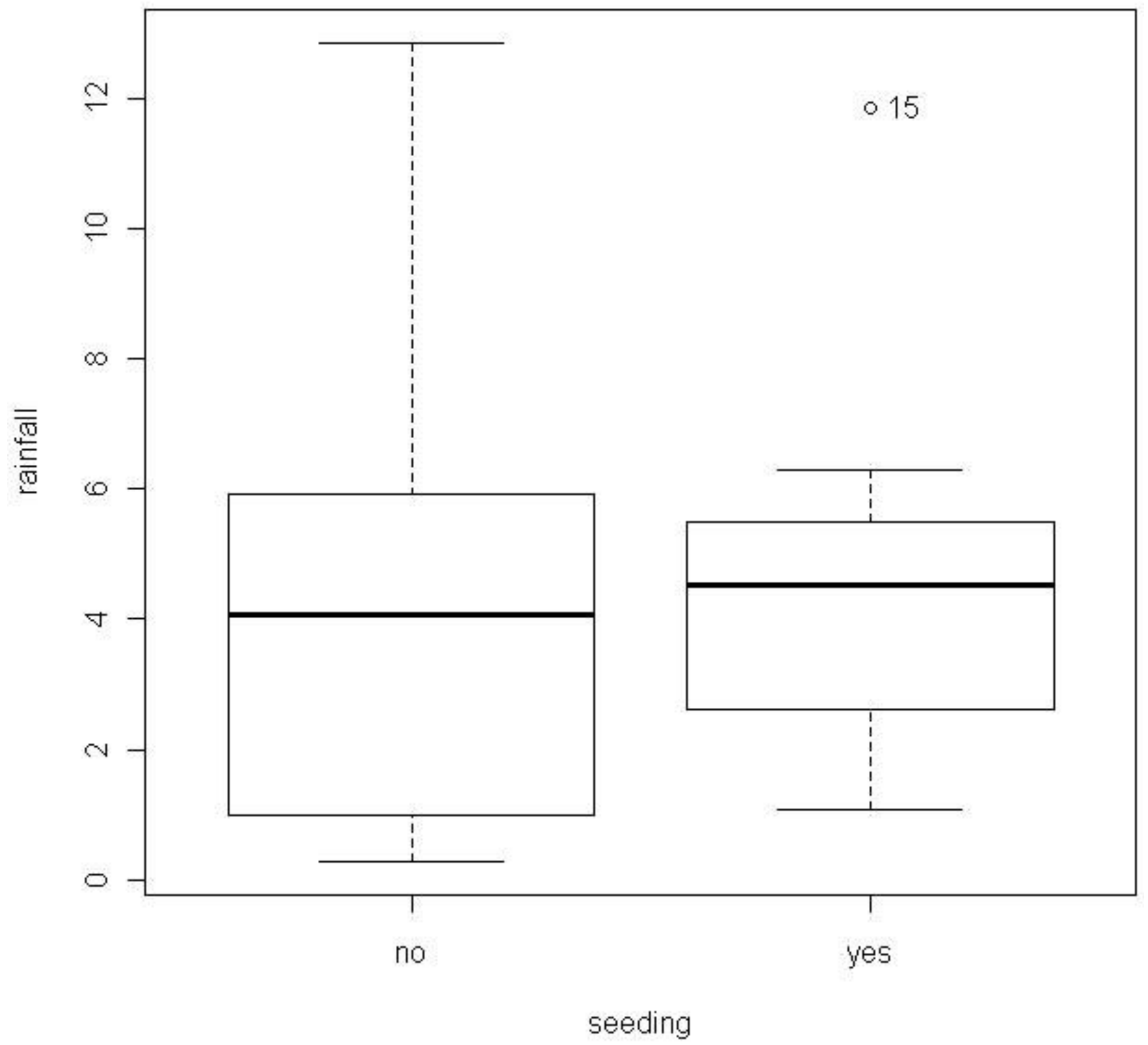
In the menu, click on View data set to see the outliers (observations 1 and

15).

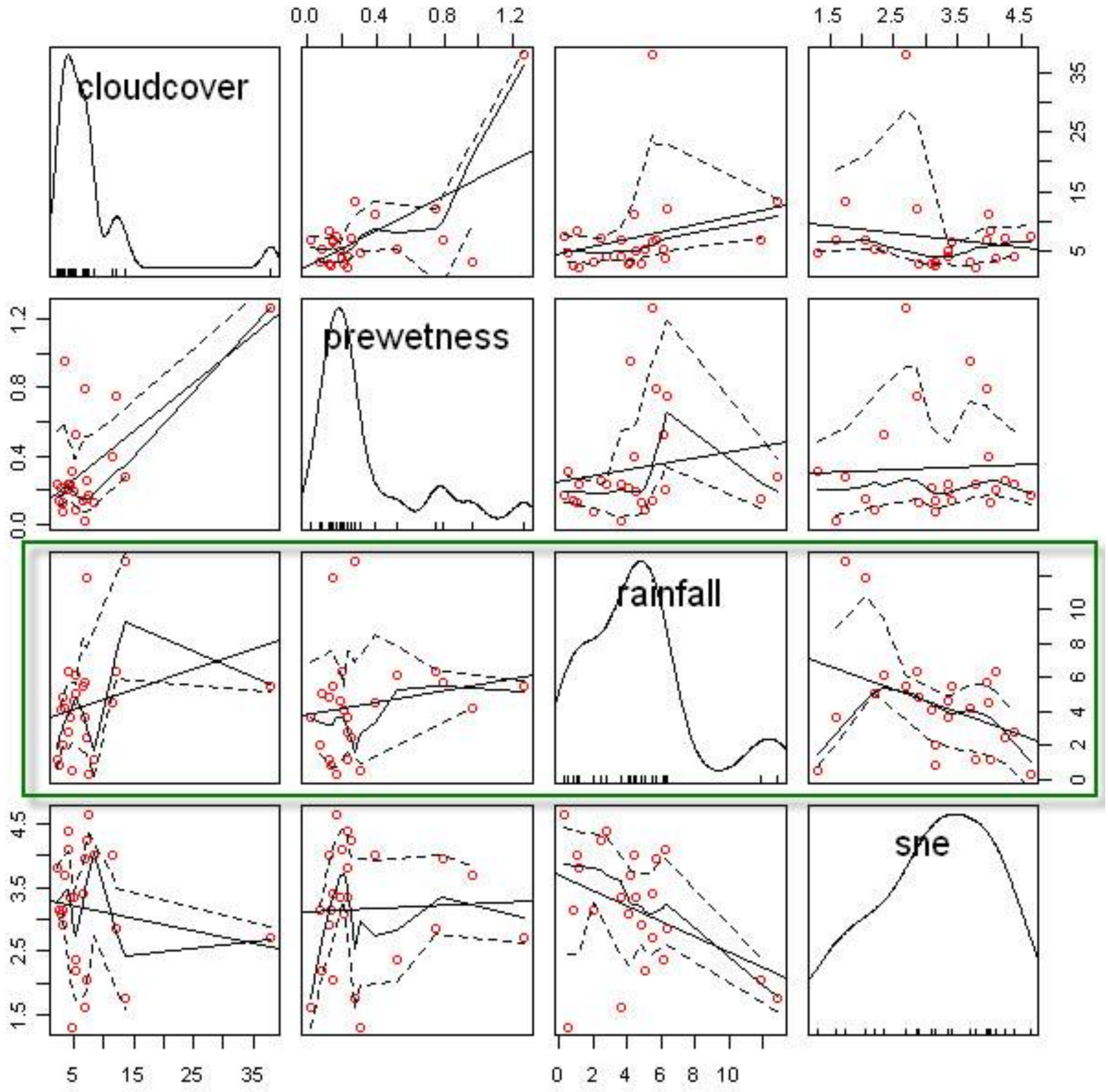


	seeding	time	sne	cloudcover	prewetness	echomotion	rainfall
1	no	0	1.75	13.4	0.274	stationary	12.85
2	yes	1	2.70	37.9	1.267	moving	5.52
3	yes	3	4.10	3.9	0.198	stationary	6.29
4	no	4	2.35	5.3	0.526	moving	6.11
5	yes	6	4.25	7.1	0.250	moving	2.45
6	no	9	1.60	6.9	0.018	stationary	3.61
7	no	18	1.30	4.6	0.307	moving	0.47
8	no	25	3.35	4.9	0.194	moving	4.56
9	no	27	2.85	12.1	0.751	moving	6.35
10	yes	28	2.20	5.2	0.084	moving	5.06
11	yes	29	4.40	4.1	0.236	moving	2.76
12	yes	32	3.10	2.8	0.214	moving	4.05
13	no	33	3.95	6.8	0.796	moving	5.74
14	yes	35	2.90	3.0	0.124	moving	4.84
15	yes	38	2.05	7.0	0.144	moving	11.86
16	no	39	4.00	11.3	0.398	moving	4.45
17	no	53	3.35	4.2	0.237	stationary	3.66
18	yes	55	3.70	3.3	0.960	moving	4.22
19	no	56	3.80	2.2	0.230	moving	1.16
20	yes	59	3.40	6.5	0.142	stationary	5.45
21	yes	65	3.15	3.1	0.073	moving	2.02
22	no	68	3.15	2.6	0.136	moving	0.82
23	yes	82	4.01	8.3	0.123	moving	1.09
24	no	83	4.65	7.4	0.168	moving	0.28

Repeat the process to see the boxplots for seeding.



From the menu, select Graphs > Scatterplot Matrix... > select all variables, check off Least-squares lines, Smooth lines, and Show spread and click Ok.



We are especially interested in the relationship to rainfall, but the relationships among the predictors is also of interest.

From the menu, select Statistics > Fit models > Linear model... > Enter the name m1 and the model formula
rainfall ~ seeding +seeding:sne +seeding:cloudcover +seeding:prewetness +seeding:echomotion +time

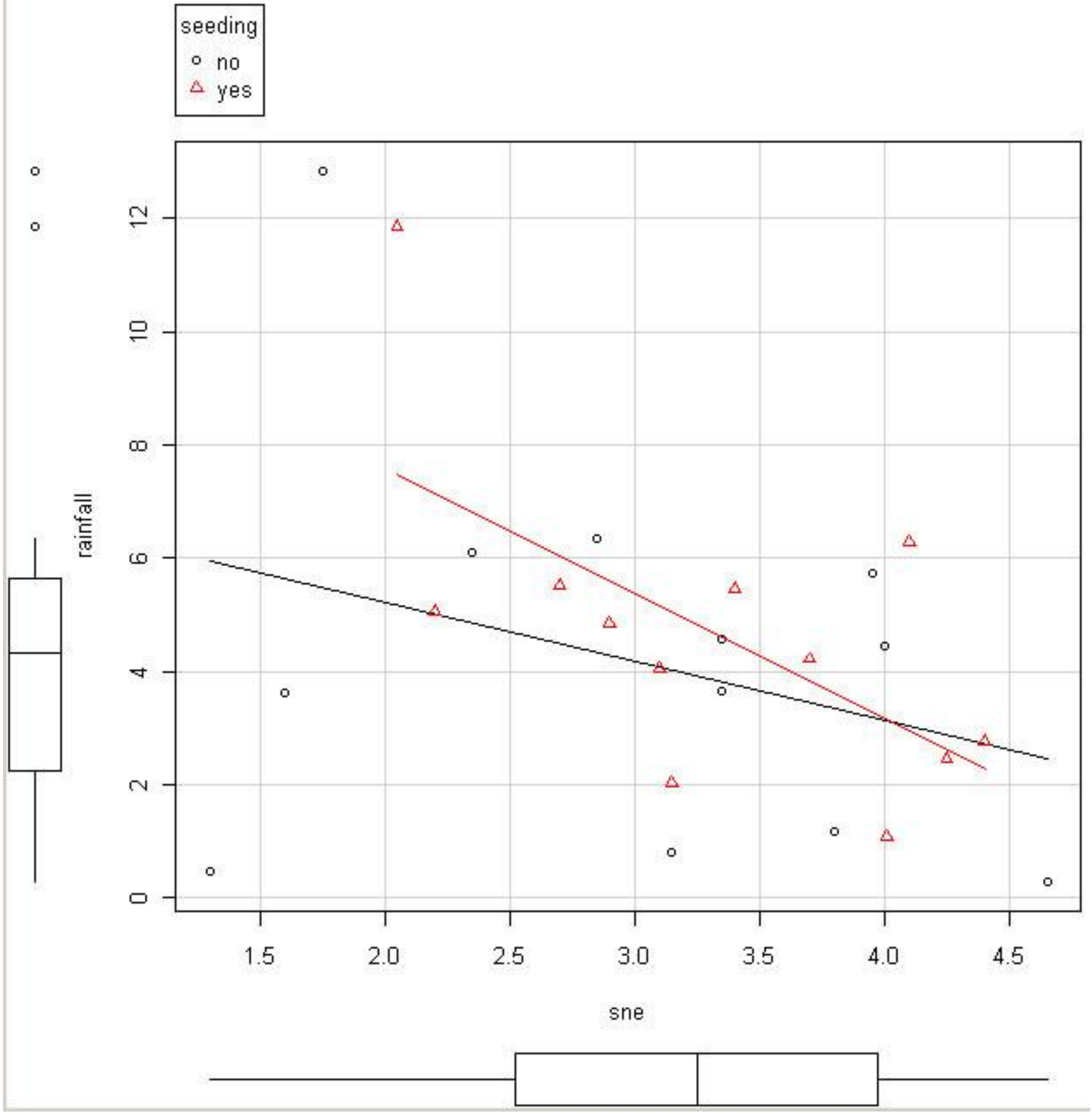
```
Call:
lm(formula = rainfall ~ seeding + seeding:sne + seeding:cloudcover +
    seeding:prewetness + seeding:echomotion + time, data = clouds)

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   
seeding[T.yes]                      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:echomotion[T.stationary]   3.15281    1.93253   1.631  0.12677   
seedingyes:echomotion[T.stationary]  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
```

From the menu, select Models > Hypotheses tests > anova table... > select type II tests and click Ok.

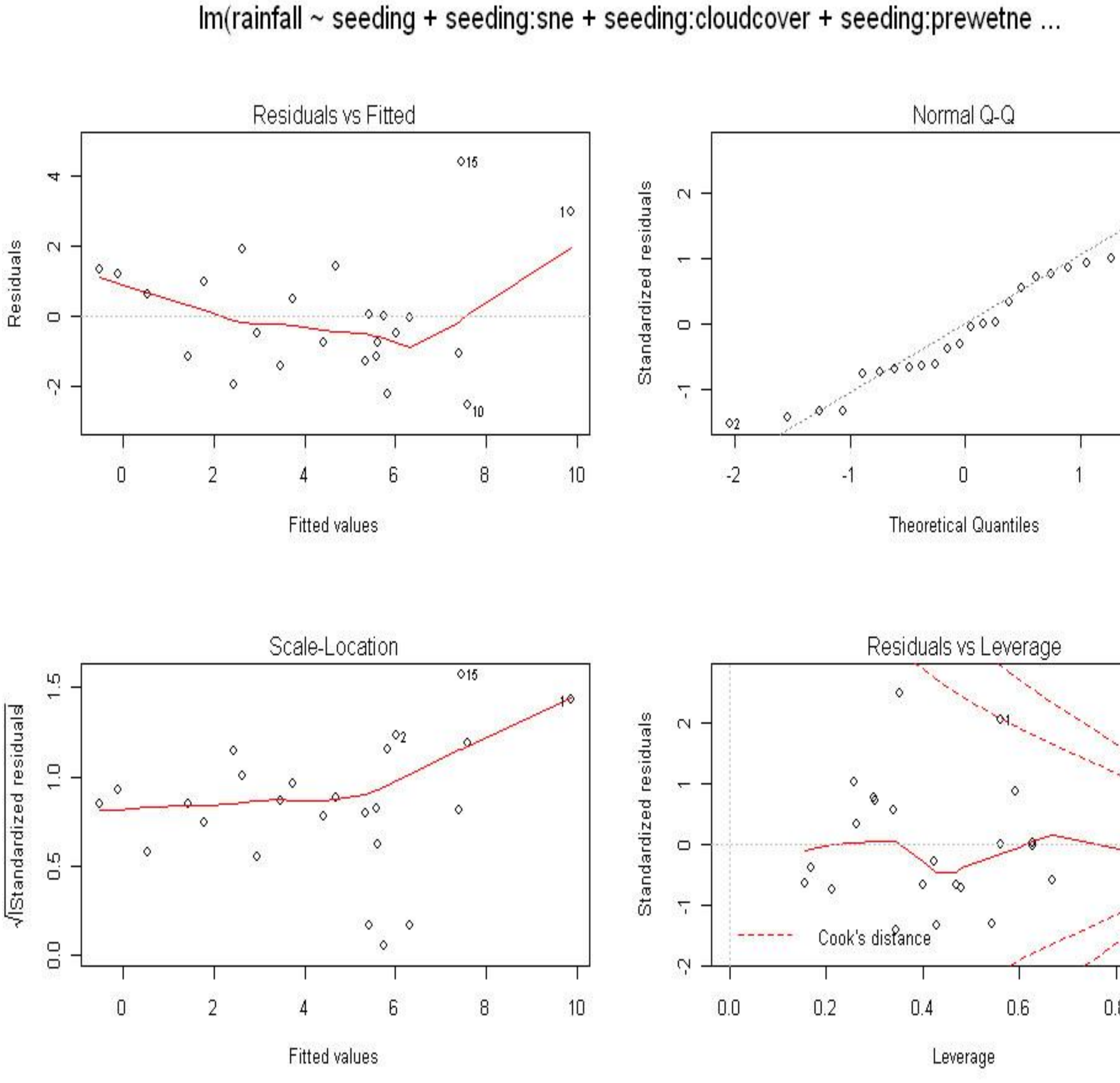
```
Anova Table (Type II tests)

Response: rainfall
                   Sum Sq Df F value  Pr(>F)  
seeding             1.843  1  0.3793 0.54863  
time               15.664  1  3.2227 0.09590 .
seeding:sne        44.441  2  4.5715 0.03138 *
seeding:cloudcover 19.782  2  2.0349 0.17027  
seeding:prewetness  7.894  2  0.8120 0.46526  
seeding:echomotion 22.672  2  2.3322 0.13630  
Residuals          63.189 13                  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Seeding:cloudcover interaction is significant.

From the menu, select Graphs > scatterplot... > select x-variable = sne, y-variable = rainfall, uncheck smooth line and show spread, click on Plot by groups > select seeding and click Ok > click Ok again.



From the menu, select Models > Graphs > Basic diagnostic plots



We see that points 1 and 15 stand out in the residuals vs fitted and the normal Q-Q plots. These are the same points we identified as outliers in the above boxplots. They may have a large influence on our analyses. Try running the analyses without these point to see how much it affects the results. If you delete points, be sure to say so in any reports, as unusual observations may be the most important points, telling you that something unusual is happening.

To remove these outliers (observations 1 and 15) enter the following line in the Script Window and Submit it.

clouds1=clouds[-c(1,15),]

Click on the Data set in the upper left and select clouds1
In the menu, click on View data set and note that observations 1 and 15 are not there.
Repeat the analysis using the clouds1 data frame.
From the menu, select Statistics > Fit models > Linear model... > Enter the name m2 and the model formula
rainfall ~ seeding +seeding:sne +seeding:cloudcover +seeding:prewetness +seeding:echomotion +time
From the menu, select Models > Hypotheses tests > anova table... > select type II tests and click Ok.

```
Anova Table (Type II tests)

Response: rainfall
                    Sum Sq Df F value   Pr(>F)   
seeding             1.4539  1  1.2266 0.291692   
time               13.0814  1 11.0366 0.006805 **
seeding:sne        11.8115  2  4.9826 0.028802 * 
seeding:cloudcover  3.5306  2  1.4894 0.267663   
seeding:prewetness 13.8648  2  5.8488 0.018611 * 
seeding:echomotion 16.2158  2  6.8405 0.011740 * 
Residuals          13.0380 11                    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Note the results are very different without these two points. This does NOT imply that the first analysis is wrong, only that it is very sensitive to the two outliers.