BIO782P Introduction to General Linear Models Cribsheet

Rob Knell / Joe Parker

9 November 2017

Basic GLMs 1: Multi-factor ANOVA

Either:

Bushmeat example

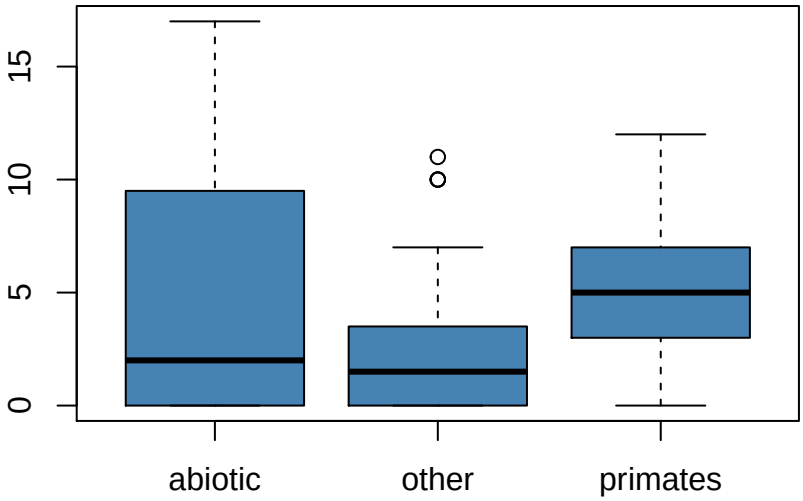
The dataset ***bushmeat.txt*** contains some of the data from a study published in 2013 by Effiom et. al. to investigate the effects of bushmeat hunting on vegetation regeneration in African Rainforests. The authors compared regeneration of experimentally cleared plots in sites where there is little hunting of primates and sites where primates are rare because of hunting. There are four variables in the data file. ***Dispersal*** is the method of seed dispersal (abiotic, other or primates), ***Hunting*** is whether the plot in question is hunted or not and ***Forest*** refers to the specific forest that the plot in question was located in. ***Number*** is the number of seedlings per category a year after the plot was cleared.

1. Load the data into R. Note that this is a csv (comma separated values) file rather than tab-separated text so you'll need to use the `read.csv()` function.
2. Check the data set using `str()`. Dispersal, Hunting and Forest should all be factors with 3,2 and 3 levels respectively. Number should be a numeric variable.
3. Draw a boxplot showing Number according to each level of Dispersal, and then another showing Number according to each level of Hunting. What do you see? Do you see anything that might make you cautious about analysing these data with a normal ANOVA?

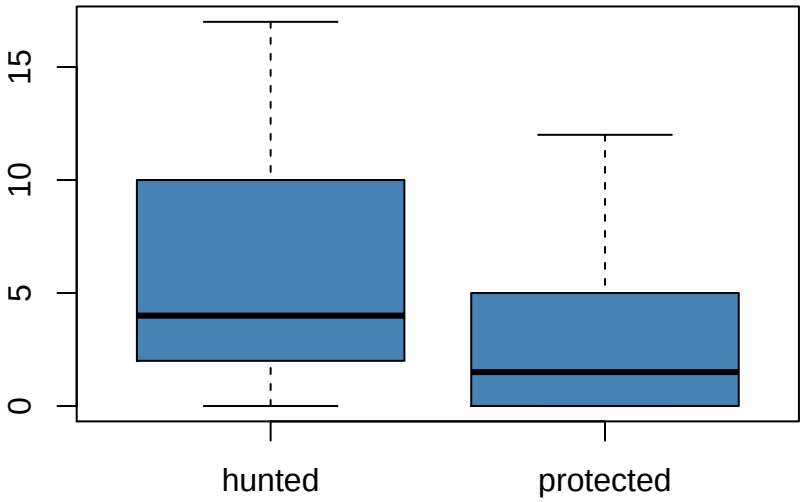
```
### Import data and save to object called "Bushmeat"

Bushmeat<-read.csv("Bushmeat.csv")

### Check data

str(Bushmeat)
```

```
## 'data.frame':    72 obs. of  4 variables:
##  $ Dispersal: Factor w/ 3 levels "abiotic","other",..: 1 2 3 1 2 3 1 2 3 1 ...
##  $ Hunting  : Factor w/ 2 levels "hunted","protected": 1 1 1 1 1 1 1 1 1 1 ...
##  $ Forest   : Factor w/ 3 levels "A","B","C": 1 1 1 1 1 1 1 1 1 1 ...
##  $ Number   : int  8 11 2 17 2 0 4 1 1 14 ...
```

```
attach(Bushmeat)

### Draw boxplot

boxplot(Number~Dispersal, col="steelblue")
```




```
boxplot(Number~Hunting, col="steelblue")
```



Boxplots are not especially symmetrical. This is because we are dealing with count data and so the distributions are bounded at zero. Also some factor levels have much wider distributions than others so we might have some issues with heteroscedasticity. This might affect our GLM but it's not clear whether we need to be concerned so for the moment we will fit a model but check the diagnostic plots carefully.

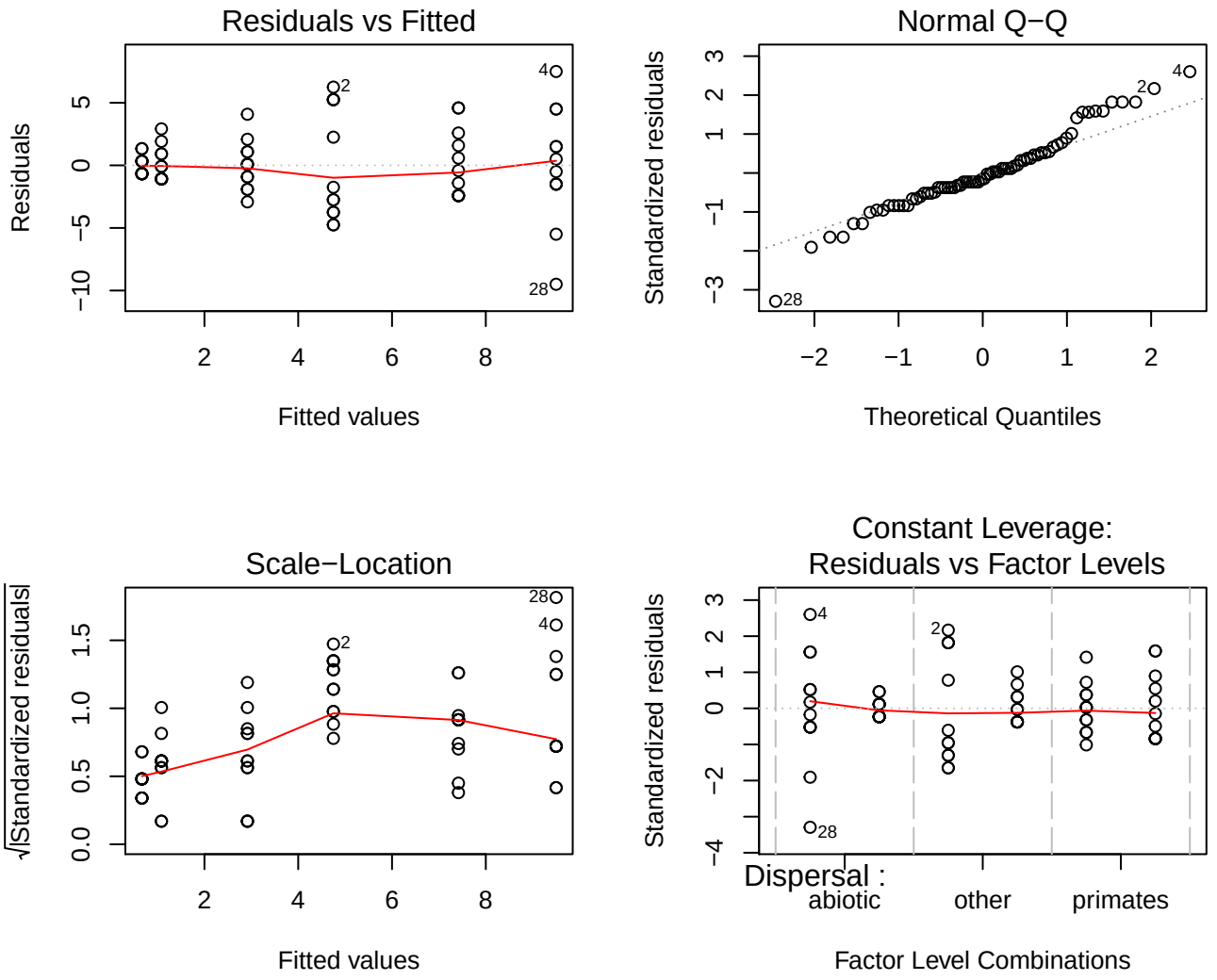
4. Fit a two-factor ANOVA to the data with Number as a response variable and Dispersal and Hunting as explanatory variables, plus the interaction between the two.
5. Is the interaction between the two significant? Check the diagnostic plots to see if the residuals are well-behaved. If they are not, then you might be able to correct the problem by transforming the Number variable. If you need to do this, re-fit the model and check the diagnostic plots again.

```
### Fit the model

model1<-lm(Number~Dispersal*Hunting)

### Check the significance of the interaction term

drop1(model1, test="F")
```

```
## Single term deletions
##
## Model:
## Number ~ Dispersal * Hunting
##                   Df Sum of Sq     RSS    AIC F value    Pr(>F)
## <none>                          598.67 164.50
## Dispersal:Hunting  2    542.33 1141.00 206.94  29.895 5.711e-10 ***
## ---
```

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

```
### Check diagnostic plots

par(mfrow=c(2,2)) ## plot in a 2x2 grid

plot(model1)
```



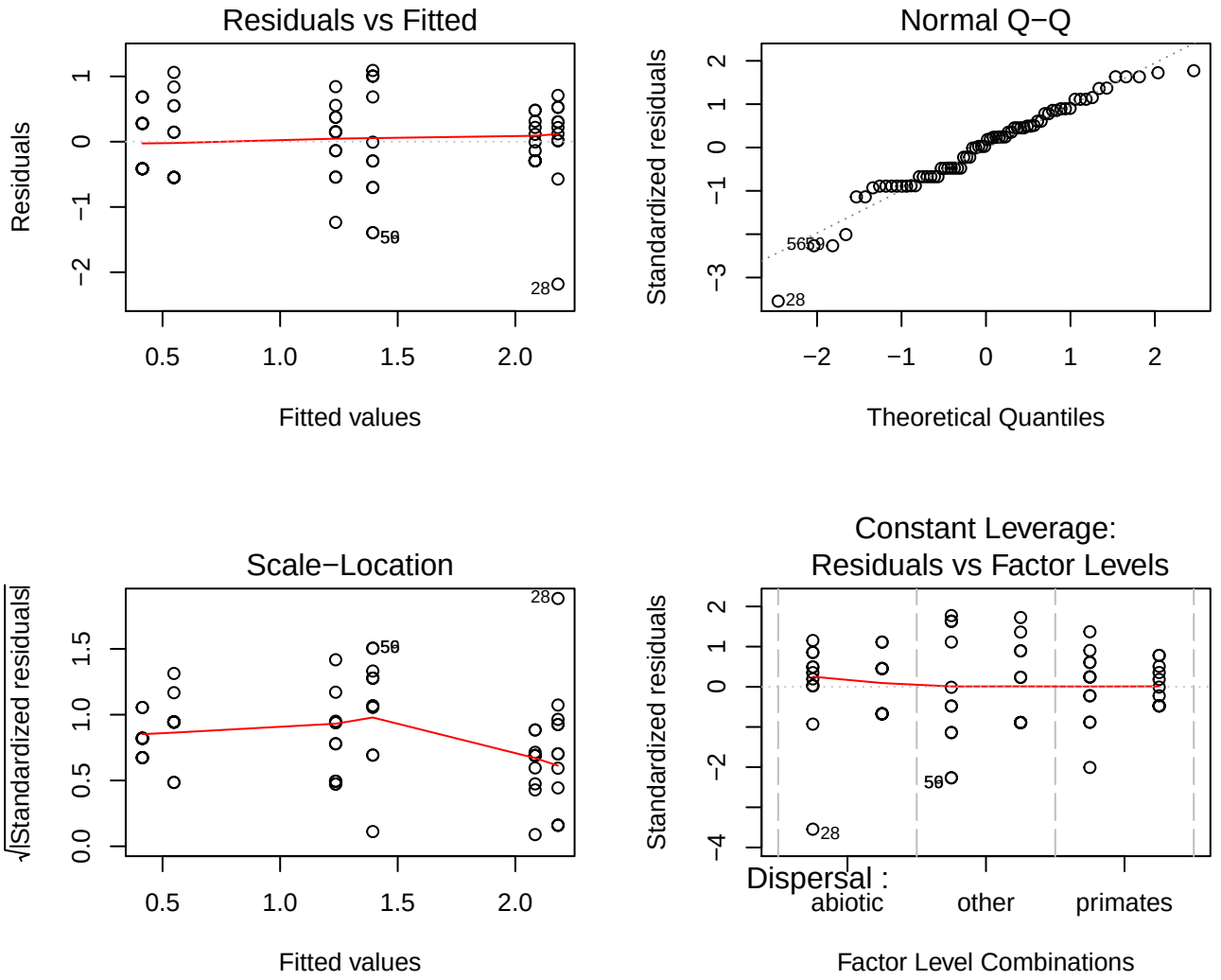

```
par(mfrow=c(1,1)) ## back to normal plotting
```

The residuals versus fitted values plot shows fairly clear indications of heteroscedasticity - the classic "fan shape" is there. We can try to correct this by log transforming our data, but since we have some zero values in number we have to add one to each data point prior to transformation.

```
### re-fit the model

model2<-lm(log(Number+1)~Dispersal*Hunting)

drop1(model2, test="F") ## Test that interaction term for significance again
```

```
## Single term deletions
##
## Model:
## log(Number + 1) ~ Dispersal * Hunting
##                   Df Sum of Sq    RSS     AIC F value   Pr(>F)
## <none>                          27.263 -57.922
## Dispersal:Hunting  2    21.121 48.384 -20.620  25.565 6.01e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
par(mfrow=c(2,2)) ## plot in a 2x2 grid

plot(model2) ## Check diagnostics again
```

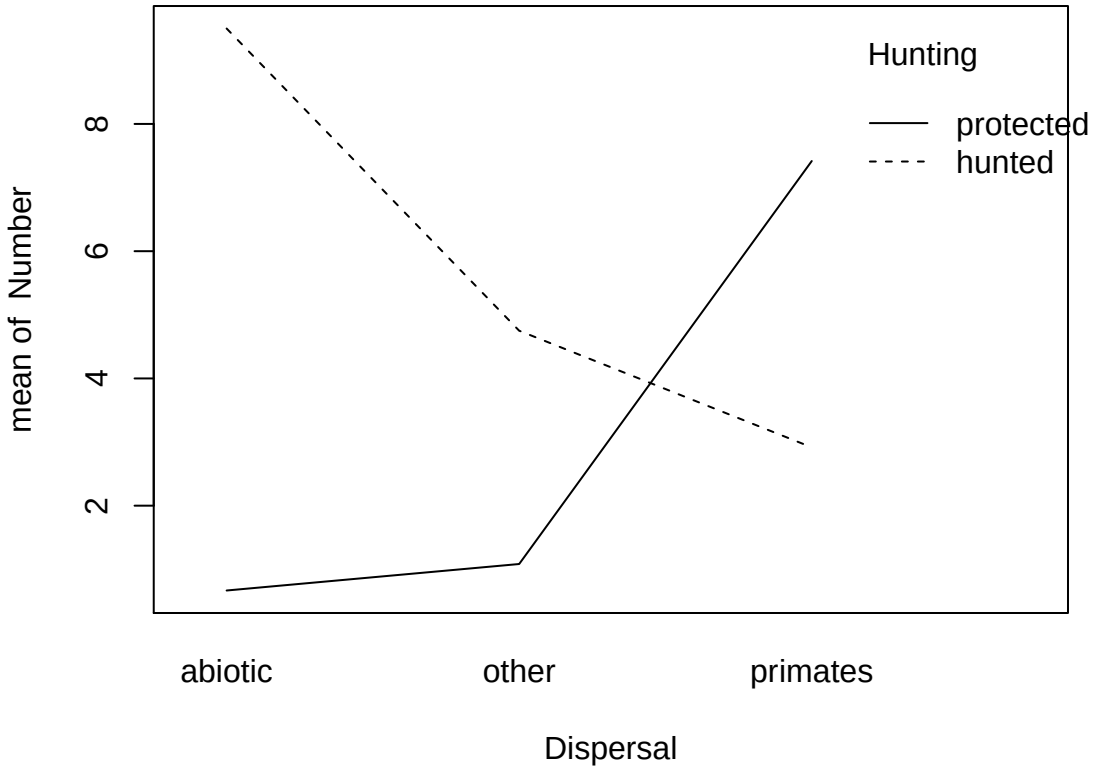



```
par(mfrow=c(1,1))
```

The diagnostic plots seem much improved. There are still some residuals that are a bit extreme but overall this is much better.

6. Examine the nature of the interaction between Dispersal and Hunting by using the `interaction.plot()` function. Read the help file to find out how it works. Use Dispersal as the x.factor, Hunting as the trace.factor and Number as the response. What do you see? What does this mean?

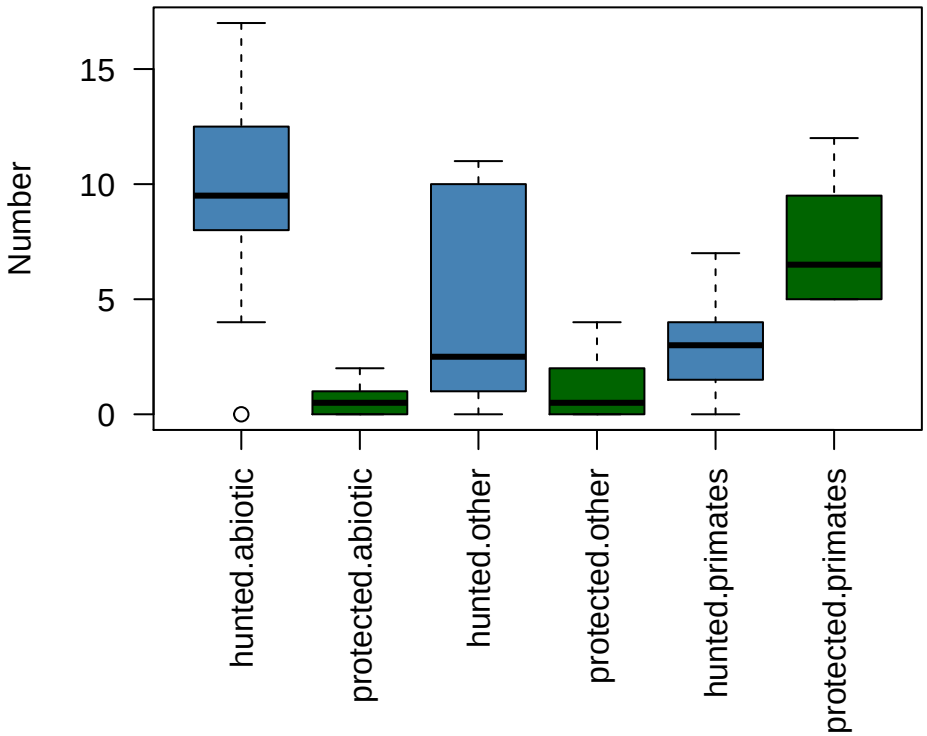
```
### Interaction plot

interaction.plot(Dispersal, Hunting, Number)
```



The effect of hunting depends on the dispersal method used by the plant, or alternatively the response of the plants to hunting is dependent on their dispersal method. Abiotic dispersal leads to plants that are much more common in hunted forests, whereas primate-dispersed plants are much more common in forests where there is no hunting. PLants classified as "other" are intermediate.

7. Another way of visualising these data is to draw a boxplot with the data categorised by more than one grouping factor. See if you can manage to do this. You can align the text on the x-axis to vertical using the `las=2` argument in the `boxplot` function call, but you'll also have to adjust the margins of the figure to make the x-axis text visible:

```
## Adjust margin sizes to fit the axis labels

par(mar=c(8,4,1,1))

boxplot(Number~interaction(Hunting, Dispersal), col=c("steelblue","darkgreen"), ylab="Number", las=2)
```



```
par(mar=c(5,4,4,2)+0.1)  ## Reset margins to defaults
```

Or:

Burying beetle gene expression example

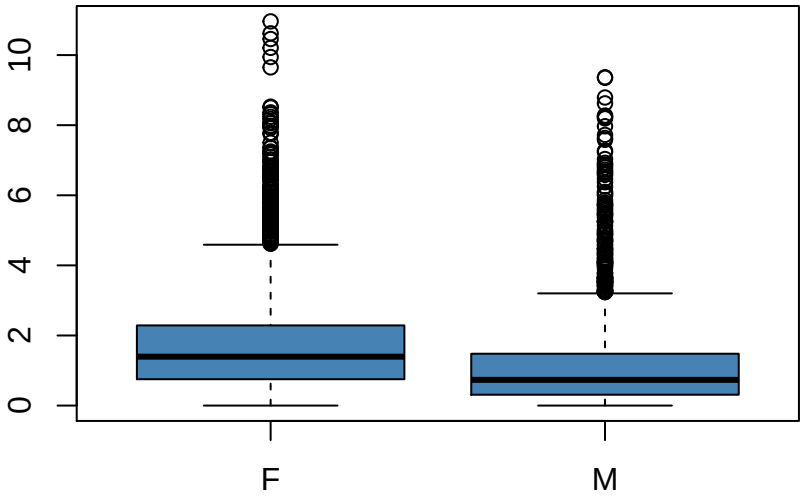
The dataset caring.csv contains a set of gene expression data for 867 genes (identified as a "caring" set of genes) from burying beetles ***Nicrophorus vespilloides***. Male and female beetles which were engaged in parental care of their offspring either as part of a pair of beetles or as a single parent had their transcriptomes sequenced and compared with transcriptomes from control beetles that were not engaged in caring for offspring. The data for the 867 pairs of genes are presented as the log2-fold change in expression in the beetles that were actively caring for their offspring.

1. Load the data into R. Note that this is a csv (comma separated values) file rather than tab-separated text so you'll need to use the `read.csv()` function.
2. Check the data set using `str()`. Sex and Biparental should be factors with 2 levels each. Log_2_fold_change should be a numeric variable.
3. The data as they are provided are signed but we want to analyse the absolute (unsigned) values since decreasing gene expression can be as important as increasing gene expression. You can convert your values to absolute ones using the abs() function.
4. Draw a boxplot showing the absolute value of Log_2_fold_change according to each level of Sex, and then another showing it according to each level of Biparental. What do you see? Do you see anything that might make you cautious about analysing these data with a normal ANOVA?

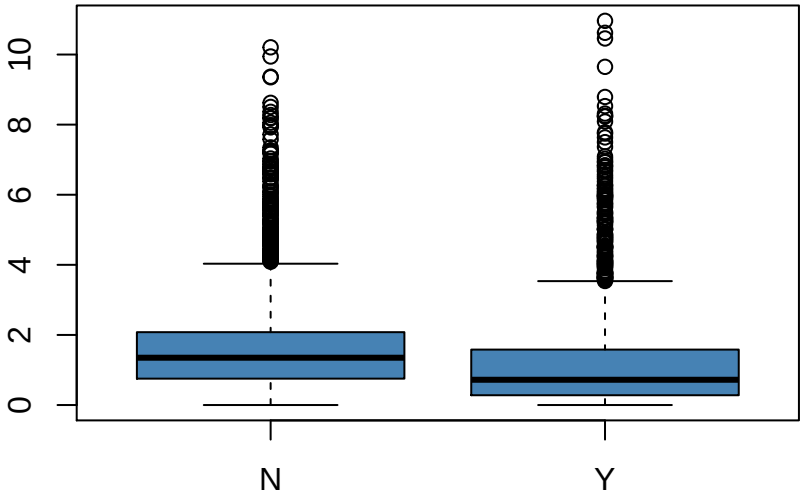
```
### Load the data into an object called "caring"

Caring<-read.csv("caring.csv")

### Check the sructure

str(Caring)
```

```
## 'data.frame':    3468 obs. of  3 variables:
##  $ Sex              : Factor w/ 2 levels "F","M": 1 1 1 1 1 1 1 1 1 1 ...
##  $ Biparental       : Factor w/ 2 levels "N","Y": 2 2 2 2 2 2 2 2 2 2 ...
##  $ Log_2_fold_change: num  -0.264 -0.08 0.107 0.03 -0.623 ...
```

```
### Attach the dataframe

attach(Caring)

### Set up new variable which is the absolute values

Change<-abs(Log_2_fold_change)

### Boxplots

boxplot(Change~Sex, col="steelblue")
```




```
boxplot(Change~Biparental, col="steelblue")
```



The response variable looks to be quite strongly positively skewed - the boxplots are asymmetrical and there are lots of data points outside the whiskers towards the positive but not towards the negative. Remember that these data are bounded at zero. We will fit the ANOVA and carefully look at the diagnostic plots

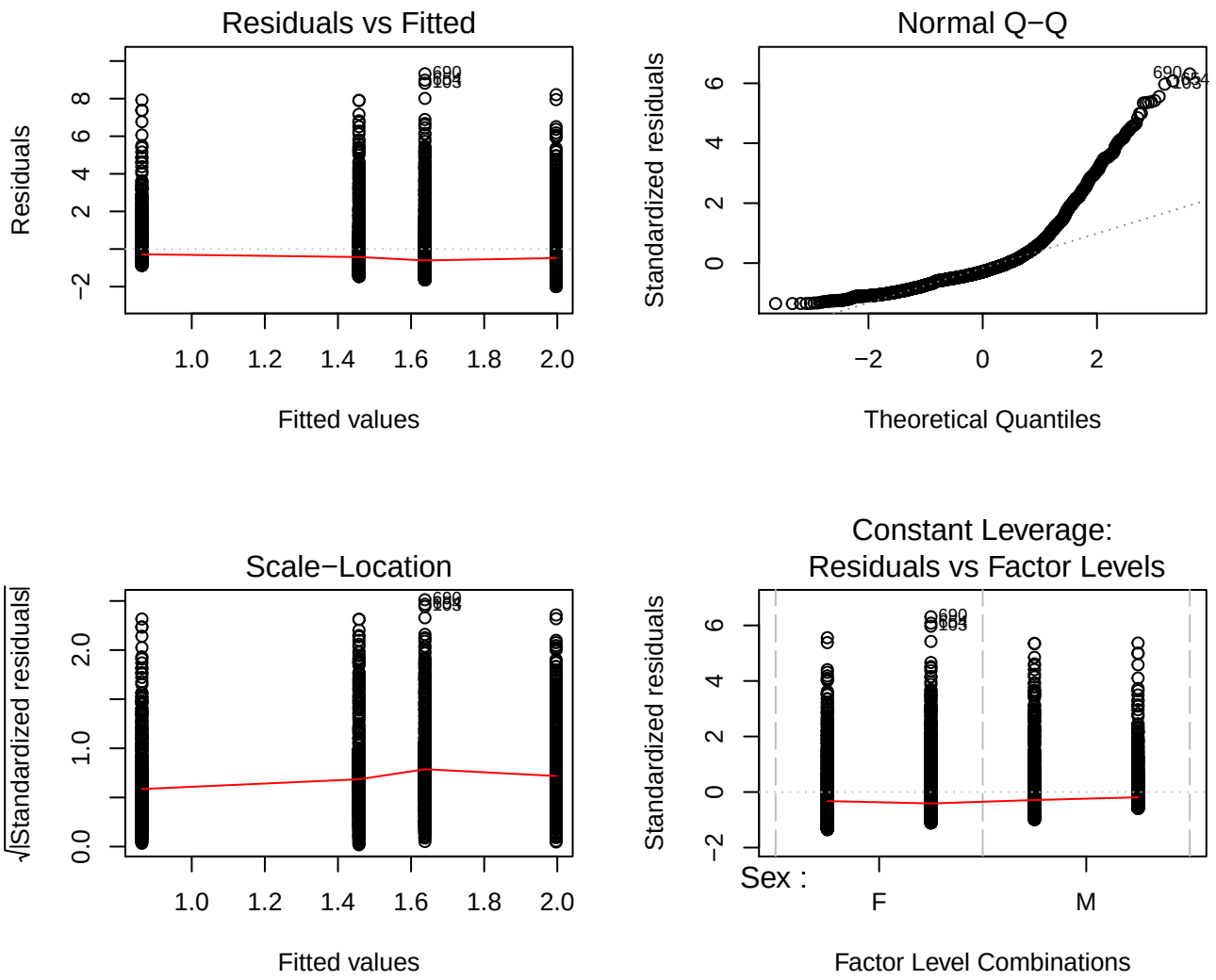
5. Fit a two-factor ANOVA to the data with the absolute (unsigned) value of Log_2_fold_change as a response variable and Sex and Biparental as explanatory variables, plus the interaction between the two.
6. Is the interaction between the two significant? Check the diagnostic plots to see if the residuals are well-behaved. If they are not, then you might be able to correct the problem by transforming the Number variable. If you need to do this, re-fit the model and check the diagnostic plots again.

```
### Fit the model
model1<-lm(Change~Sex*Biparental)


### Test for significant interaction
drop1(model1, test="F")
```

```
## Single term deletions
##
## Model:
## Change ~ Sex * Biparental
##                Df Sum of Sq    RSS    AIC F value  Pr(>F)
## <none>                      7566.8 2713.7
## Sex:Biparental  1    11.964 7578.8 2717.2  5.4768 0.01933 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
### Diagnostics
```

```
par(mfrow=c(2,2))
plot(model1)
```



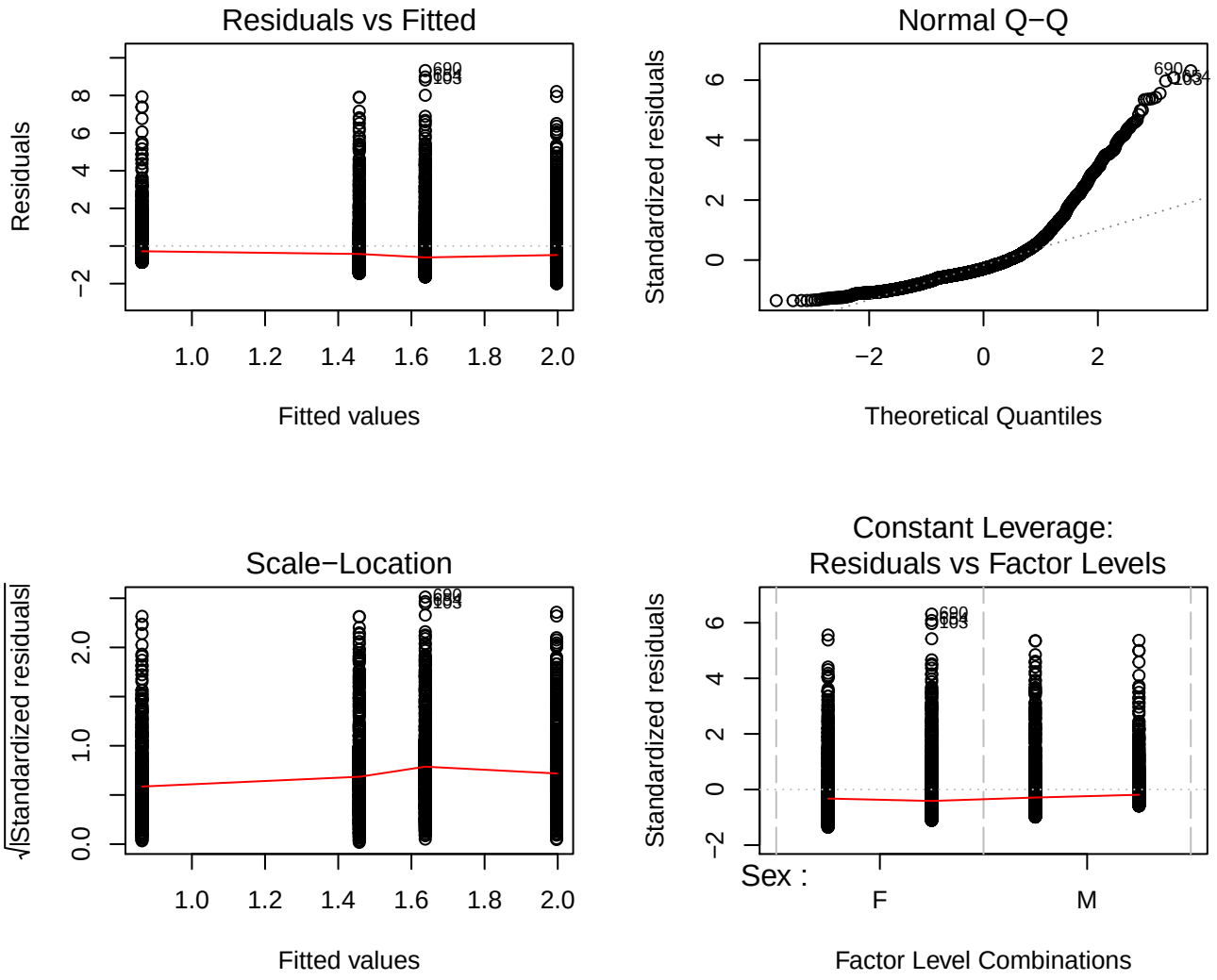

```
par(mfrow=c(1,1))
```

Very strong evidence of positive skew in the errors. We can try to correct for this with a log transformation.

```
### Model with log transformed data
model2<-lm(log(Change+1)~Sex*Biparental)

### Test for significant interaction
drop1(model2, test="F")
```

```
## Single term deletions
##
## Model:
## log(Change + 1) ~ Sex * Biparental
##                Df Sum of Sq    RSS     AIC F value  Pr(>F)
## <none>                      779.46 -5168.8
## Sex:Biparental  1    1.1765 780.63 -5165.6  5.2287 0.02228 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
### Diagnostics

par(mfrow=c(2,2))
plot(model1)
```

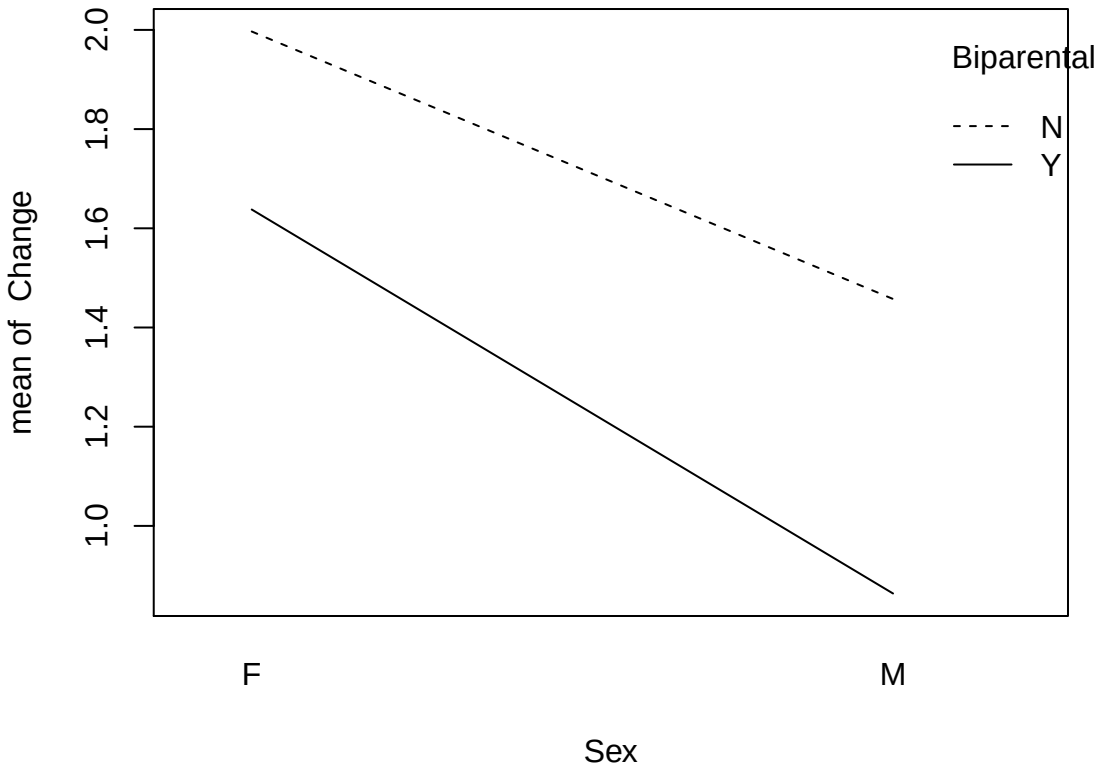



```
par(mfrow=c(1,1))
```

Still a fair amount of skew in the errors, however there is no heteroscedasticity and we have a large sample size and a balanced design so we will remember that under these circumstances the GLM is robust to departures from normality of errors and since the skew is not too extreme we'll accept our analysis.

7. Examine the nature of the interaction between Sex and Biparental by using the interaction.plot() function. Read the help file to find out how it works. What do you see? What does this mean?

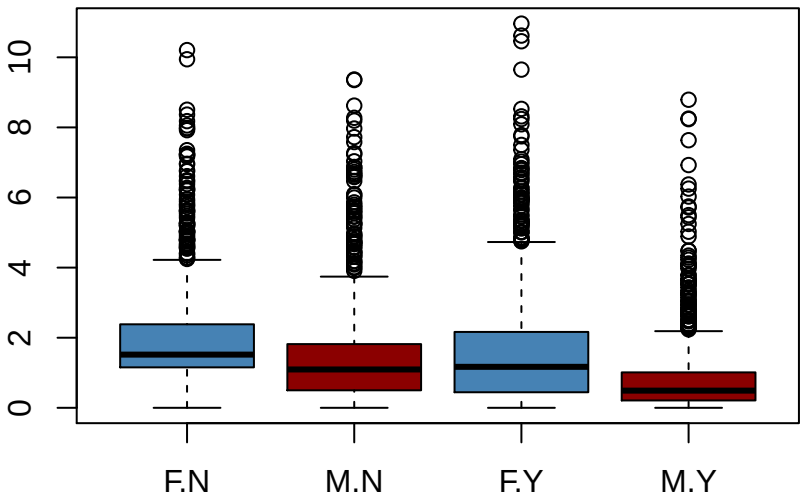
```
### Plot interaction plot

interaction.plot(Sex, Biparental, Change)
```



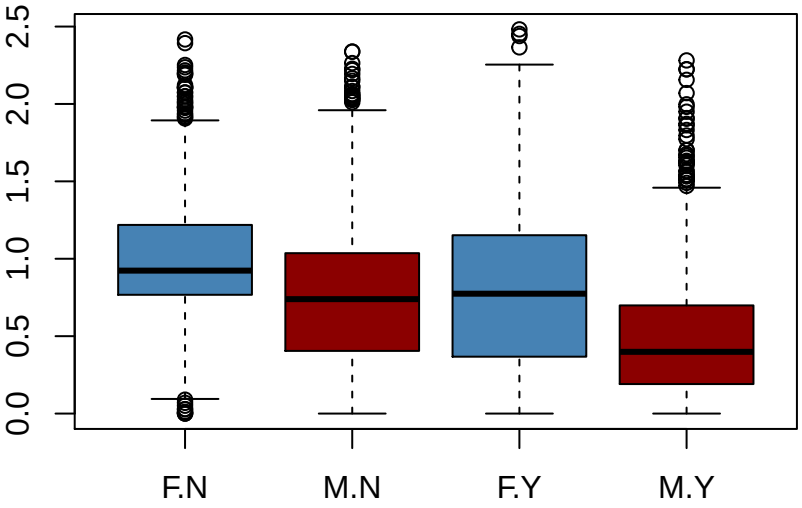
The degree of expression of the genes in the "caring" set changes more in females than in males, however when males are caring for offspring by themselves there is a bigger change in gene expression than in females which are caring by themselves. Males caring by themselves have a similar amount of change in gene expression to biparental females

8. Another way of visualising these data is to draw a boxplot with the data categorised by more than one grouping factor. See if you can manage to do this.

```
boxplot(Change~interaction(Sex, Biparental), col=c("steelblue","darkred"))
```




```
### Very hard to see patterns because of outliers - better to plot the log transformed data


boxplot(log(Change+1)~interaction(Sex, Biparental), col=c("steelblue","darkred")) ### Slightly better
```



Parker DJ, Cunningham CB, Walling CA, Stamper CE, Head ML, Roy-Zokan EM, McKinney EC, Ritchie MG, Moore AJ. 2015. Transcriptomes of parents identify parenting strategies and sexual conflict in a subsocial beetle. Nat Commun 6:8449.

Basic GLMs 2: A continuous explanatory variable and a factor

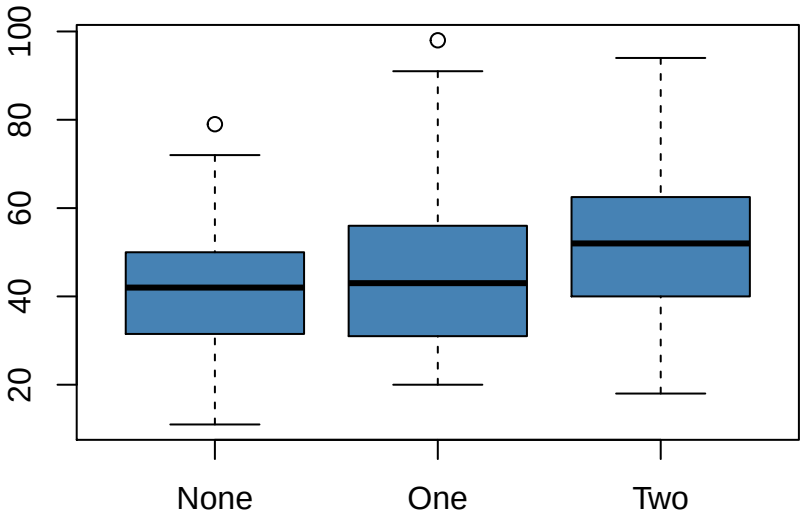
The deleterious effects of inbreeding are well known as problems in small populations. In a study published in 2014, van Bergen and colleagues described the results of an experiment to investigate the effects of inbreeding on flight performance and pheromone production in the butterfly ***Bicyclus anynana***. We will analyse one part of their data to look at how flight performance relates to thorax size and to inbreeding.

1. The file van_Bergen_Bicyclus.txt has data from this experiment. Save it as an object in R and check the data frame using `str()`. There should be three variables. Inbreeding is a factor with three levels indicating the number of generations of sib-matings in a butterfly's recent family history. There are three levels: none, one and two. Drythor is the dry weight of the butterfly's thorax, and FII is the flight inhibition index – the number of times the butterfly settled during a two-minute period when it was being stimulated to take off immediately once it settled. Butterflies that are less able to fly for long bouts have higher FII measurements.
2. For some exploratory data analysis, draw a boxplot of FII grouped by inbreeding level, and a scatterplot of FII against thorax weight.

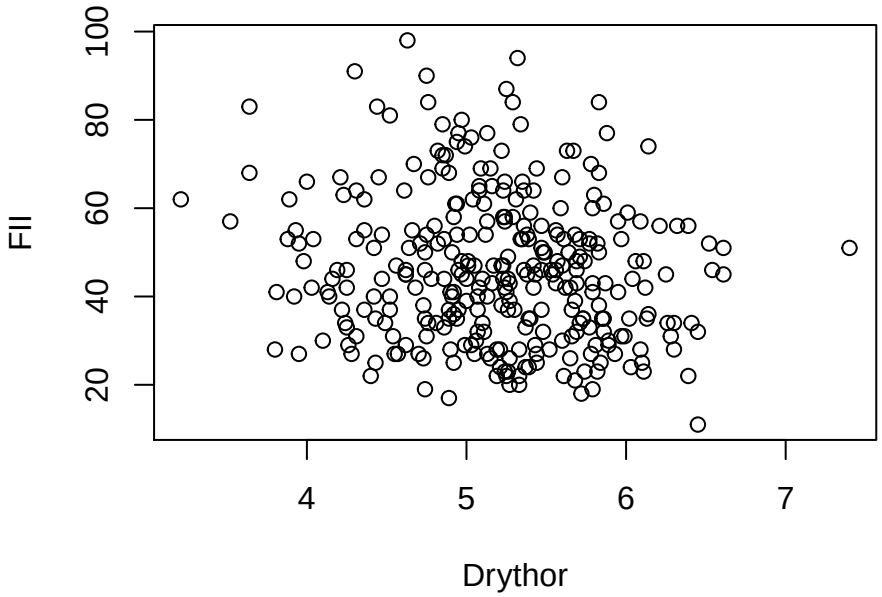
```
### Load data into object Bicyclus

Bicyclus<-read.table("van_Bergen_Bicyclus.txt", header=TRUE)

### Check structure
str(Bicyclus)
```

```
## 'data.frame':    313 obs. of  3 variables:
##  $ Inbreeding: Factor w/ 3 levels "None","One","Two": 1 1 1 1 1 1 1 1 1 1 ...
##  $ Drythor   : num  5.21 4.31 5.08 4.92 5.08 5.16 5.48 5.8 5.39 4.74 ...
##  $ FII       : int  24 53 64 25 42 43 50 63 24 19 ...
```

```
attach(Bicyclus)

### Exploratory plots

boxplot(FII~Inbreeding, col="steelblue")
```




```
plot(Drythor, FII)
```

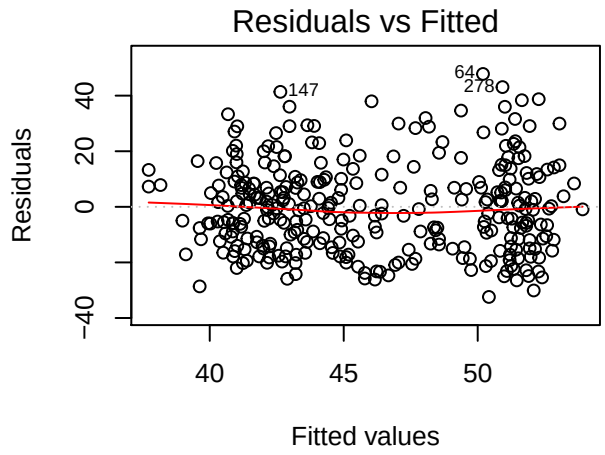


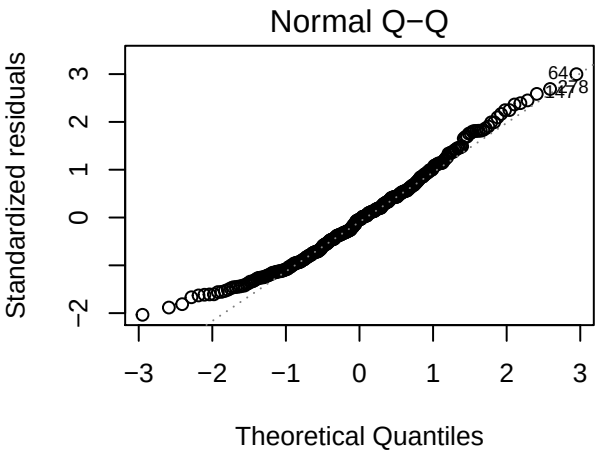
What do you see? Is there any reason to think these data might not be suitable for a standard parametric analysis?

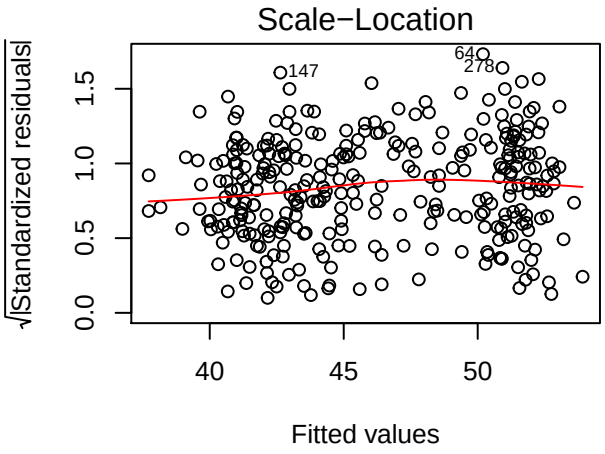
Both plots look fine - the boxplots are reasonably symmetrical and there doesn't seem to be any heteroscedasticity. It looks as though FII is related to inbreeding but we will have to wait for the analysis to find out. he scatterplot is very noisy but looks OK

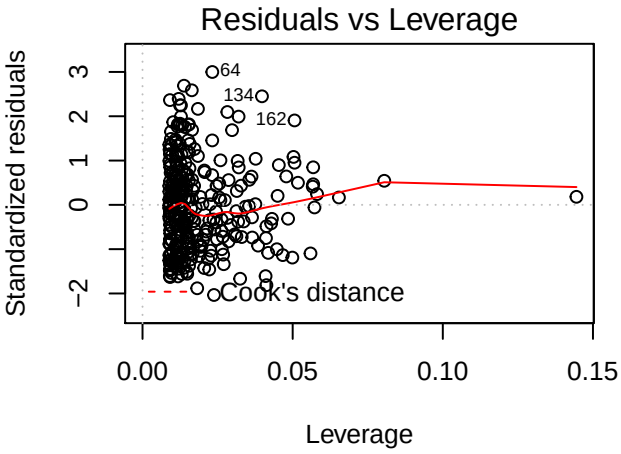
3. Fit a model with FII as the response variable and Inbreeding, Drythor and the interaction between the two as explanatory variables.
4. Check your diagnostic plots. How do they look? As with the previous exercise, if the residuals are not well behaved then you might be able to correct the problem by transforming the FII variable. If you need to do this, re-fit the model and check the diagnostic plots again.

```
### Fit model

model1<-lm(FII~Inbreeding * Drythor)


### Diagnostics

par(mfrow=c(2,2))
plot(model1)
```










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

Diagnostics all look OK, possibly some deviation from normality in the qq plot but not enough to worry about

5. This time we're going to use a deletion test to assess whether our interaction term is statistically significant. Use the drop1 function with test="F" to do this.
6. If the interaction term is not significant, re-fit the model without it and repeat the process until you have a minimal adequate model. Look at the summary table and try to work out what the coefficients mean.

```
### Deletion test

drop1(model1, test="F")
```

```
## Single term deletions
##
## Model:
## FII ~ Inbreeding * Drythor
##                    Df Sum of Sq   RSS    AIC F value Pr(>F)
## <none>                          79864 1746.6
## Inbreeding:Drythor  2    544.84 80409 1744.7  1.0472 0.3522
```

```
### Refit model without interaction and retest

model2<-lm(FII~Inbreeding + Drythor)

drop1(model2, test="F")
```

```
## Single term deletions
##
## Model:
## FII ~ Inbreeding + Drythor
```

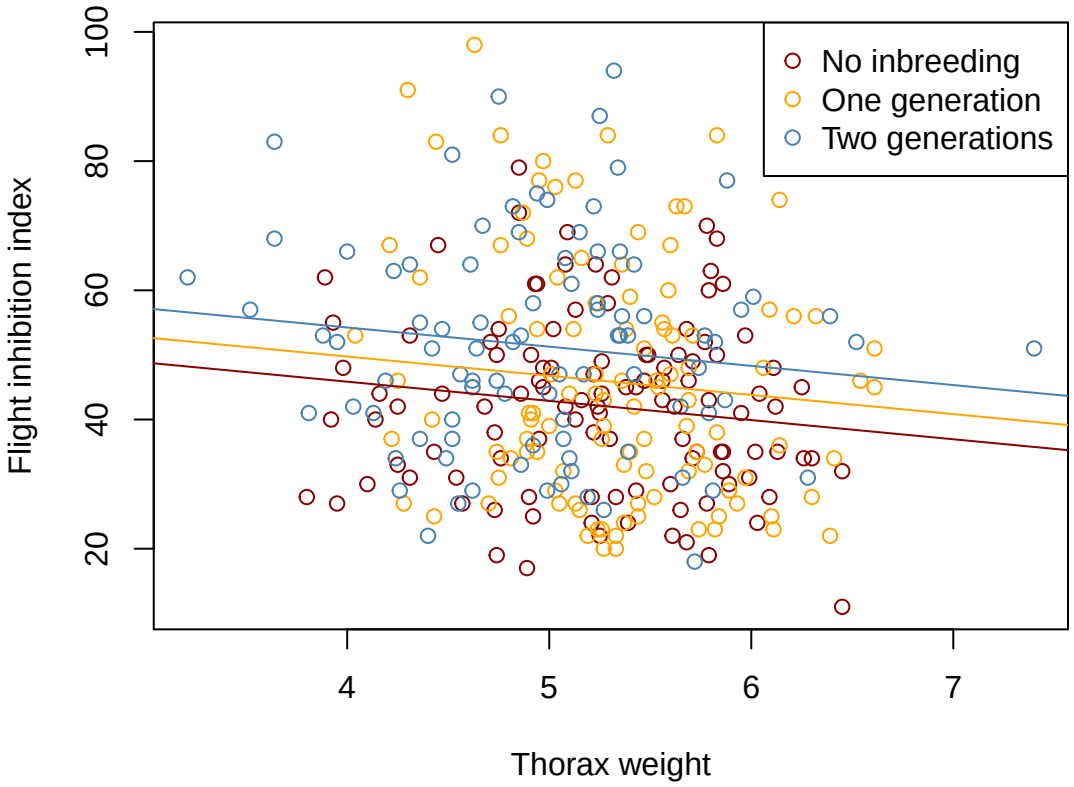
```
##            Df Sum of Sq   RSS    AIC F value   Pr(>F)
## <none>                  80409 1744.7
## Inbreeding  2    3431.9 83841 1753.8  6.5940 0.001569 **
## Drythor     1    1089.3 81499 1747.0  4.1861 0.041604 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
### Summary table

summary(model2)
```

```
##
## Call:
## lm(formula = FII ~ Inbreeding + Drythor)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -31.134 -12.724  -1.139   9.020  50.127
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)     57.735      7.772   7.429 1.08e-12 ***
## InbreedingOne    3.893      2.168   1.795 0.073603 .
## InbreedingTwo    8.392      2.313   3.628 0.000334 ***
## Drythor         -2.971      1.452  -2.046 0.041604 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 16.13 on 309 degrees of freedom
## Multiple R-squared:  0.06275,    Adjusted R-squared:  0.05365
## F-statistic: 6.896 on 3 and 309 DF,  p-value: 0.0001651
```

7. Plot the data with the fitted model. You might need to think about what the best approach might be for this.

```
### Plot empty graph

plot(Drythor, FII, xlab = "Thorax weight", ylab="Flight inhibition index", type="n")

### Draw in points with colour coding for the factor levels
points(Drythor[Inbreeding=="None"], FII[Inbreeding=="None"], col="darkred")
points(Drythor[Inbreeding=="One"], FII[Inbreeding=="One"], col="orange")
points(Drythor[Inbreeding=="Two"], FII[Inbreeding=="Two"], col="steelblue")

### Draw in fitted lines
abline(57.74, -2.971, col="darkred")
abline(57.74 + 3.893, -2.971, col="orange")
abline(57.74 + 8.392, -2.971, col="steelblue")

### Add legend
legend("topright", legend=c("No inbreeding", "One generation", "Two generations"),
       pch=1, col=c("darkred","orange","steelblue"))
```



8. Explain the model output in words.

Flight inhibition increases with inbreeding, with those butterflies that have experienced one generation of inbreeding settling, on average, 3.8 times more in a two minute period than outbred animals, and those that have experienced two generations of inbreeding settling an average of 8.4 times more than outbred butterflies every two minutes. Butterflies with heavier thoraxes tended to settle less, so a one unit increase in thorax dry weight corresponded to an average decrese in the number of settlings by a butterfly of 2.9. This effect of thorax weight was constant across all inbreeding treatments.

Bergen, E. van, Brakefield, P.M., Heuskin, S., Zwaan, B.J., and Nieberding, C.M. (2013). The scent of inbreeding: a male sex pheromone betrays inbred males. Proc. R. Soc. B 280, 20130102.