glm Wortkshop 1 Answers

Crispin Jordan

05/10/2020

These answers provide less detail than those you will have received for your 1-Factor glm homework. However, the process to analyze these data is comparable, so please feel free to refer to previous answer sets to find words of wisdom (or approximations thereof).

Question 1 - Pinecones

We begin by importing the data:

```
pine <- read.table("pinecones.csv",sep=',',header=TRUE)
```

Here, we have saved the dataset in the object, `pine`.

Let's get familiar with the data. We'll start by looking at the dataset:

```
pine
```

```
##             habitat conemass
## 1     island.absent      9.6
## 2     island.absent      9.4
## 3     island.absent      8.9
## 4     island.absent      8.8
## 5     island.absent      8.5
## 6     island.absent      8.2
## 7    island.present      6.8
## 8    island.present      6.6
## 9    island.present      6.0
## 10   island.present      5.7
## 11   island.present      5.3
## 12 mainland.present      6.7
## 13 mainland.present      6.4
## 14 mainland.present      6.2
## 15 mainland.present      5.7
## 16 mainland.present      5.6
```

We can see that the dataset has two columns: `habitat` and `conemass`. Let's get a summary of the data now:

```
summary(pine)
```

```
##    habitat             conemass
##  Length:16          Min.   :5.300
##  Class :character   1st Qu.:5.925
##  Mode  :character   Median :6.650
##                     Mean   :7.150
##                     3rd Qu.:8.575
##                     Max.   :9.600
```

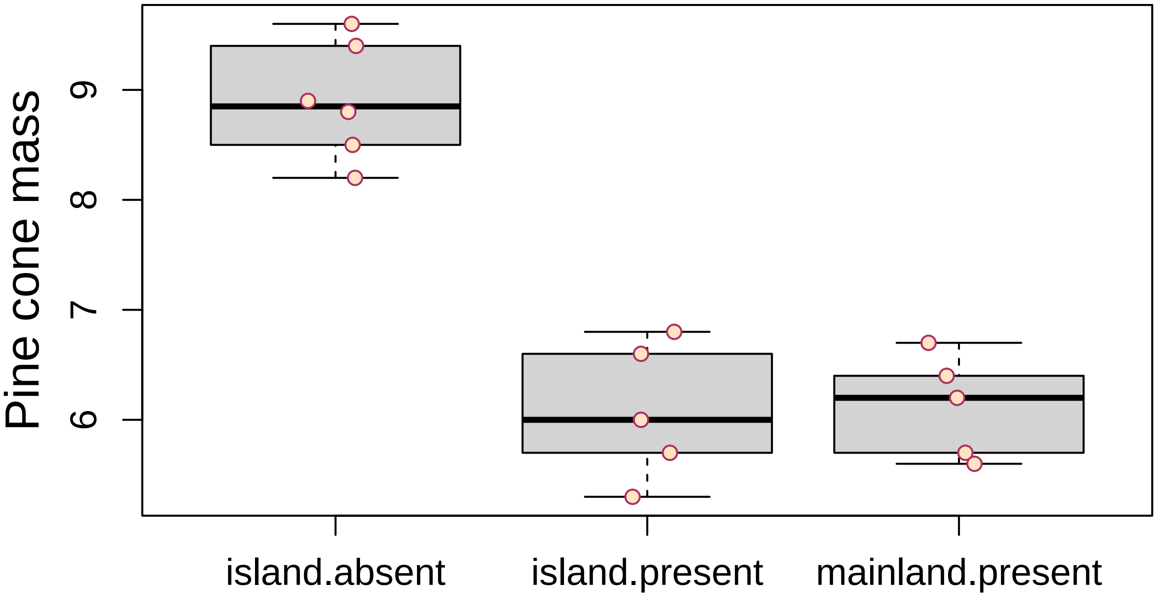
The column, `habitat`, has three levels: `island.absent`, `island.present`, and `mainland.present`; these have 6, 5 and 5 observations, respectively (not a very big dataset). The column, `conemass`, has an average value of 7.15, with a range of 5.3 to 9.6.

OK! The experiment aims to compare `conemass` among three different habitats (which are noted in the column, `habitat`). Let's ask ourselves, "Which column contains the *dependent* variable, and which contains the *independent* variable?".

Give our biological question, it makes most sense to say that `conemass` *depends* on `habitat`; i.e., `conemass` is the *dependent* variable, and the *independent* variable is `habitat`. This is consistent with asking whether `conemass` differs among `habitats`.

With this in mind, and now that we're familiar with the structure of our dataframe, we can begin our analysis by plotting the data:

```
boxplot(conemass ~ habitat, data=pine,cex.axis = 1.2, xlab = "", ylab = "")
stripchart(conemass ~ habitat, data=pine,
           vertical = TRUE, method = "jitter",
           pch = 21, col = "maroon", bg = "bisque",
           add = TRUE)
mtext("Pine cone mass",2,line=2.5,cex=1.5)
```



What can we conclude from this plot?

- There are no obvious outliers
- The boxplots are all fairly symmetrical, suggesting that the residuals will be normally distributed
- The breadth of the boxplots is also similar among the habitat types, suggesting that the data will meet the assumption of equal variance.
- The pinecone mass in the `island.absent` habitat appears to equal approximately 9.0, and is about 2.5 units larger than for `island.present` and `mainland.present`
- Pinecone mass does not appear to differ between `island.present` and `mainland.present`

These observations have prepared us for our analysis because we already have a good sense of what to expect: our data should meet the assumptions of the analysis, and we have expectations regarding the size of differences (or lack thereof) between the three treatments.

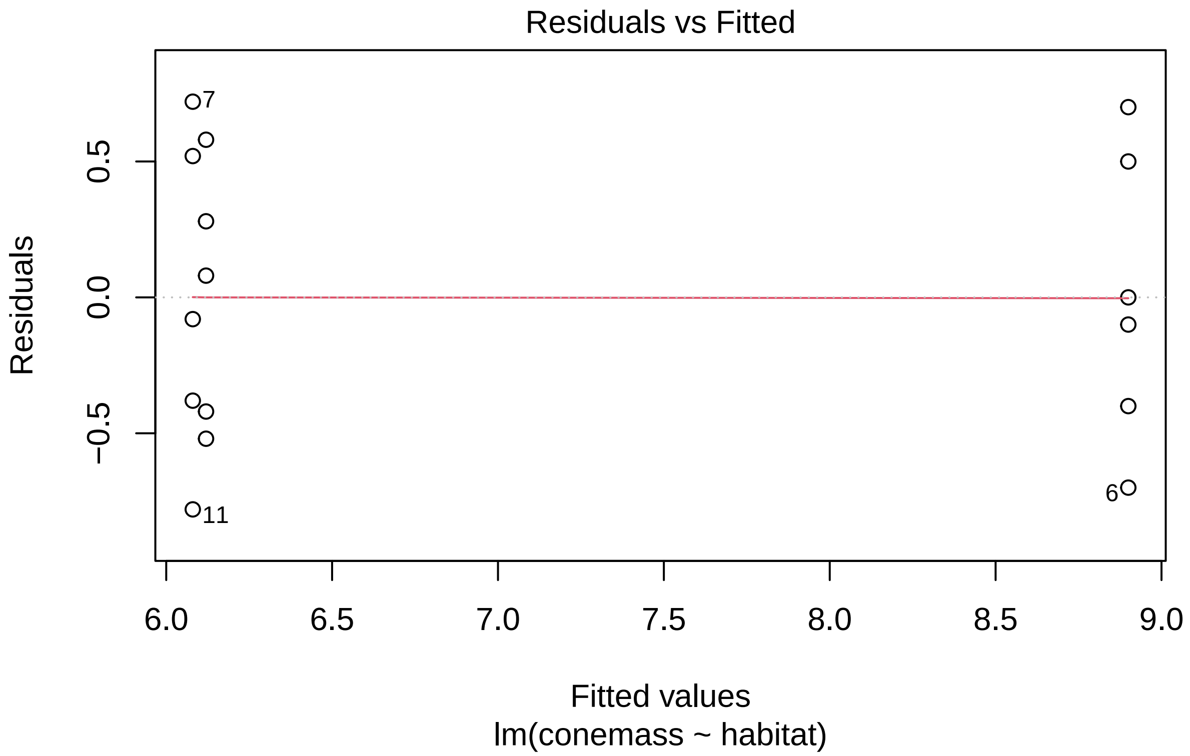
With this in mind, let's model our data. Remember that we decided that `conemass` was the *dependent* variable. Therefore, we can model the data as:

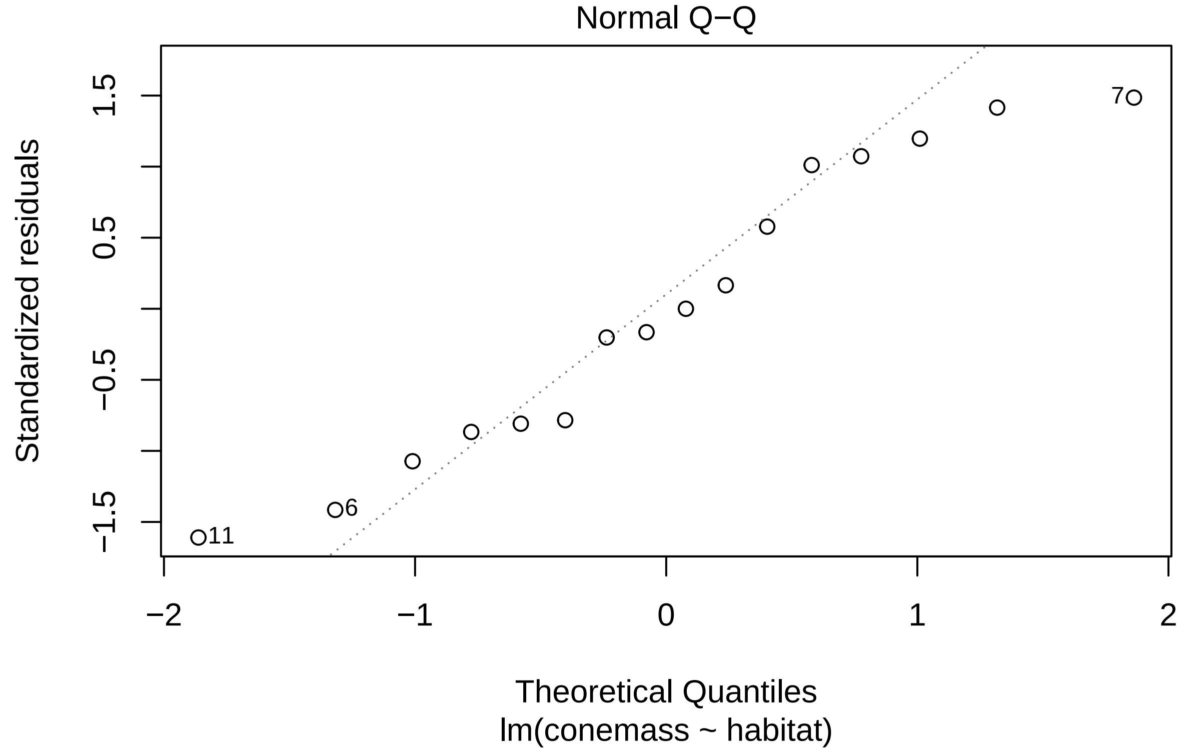
```r
p.lm <- lm(conemass ~ habitat, data=pine)
```

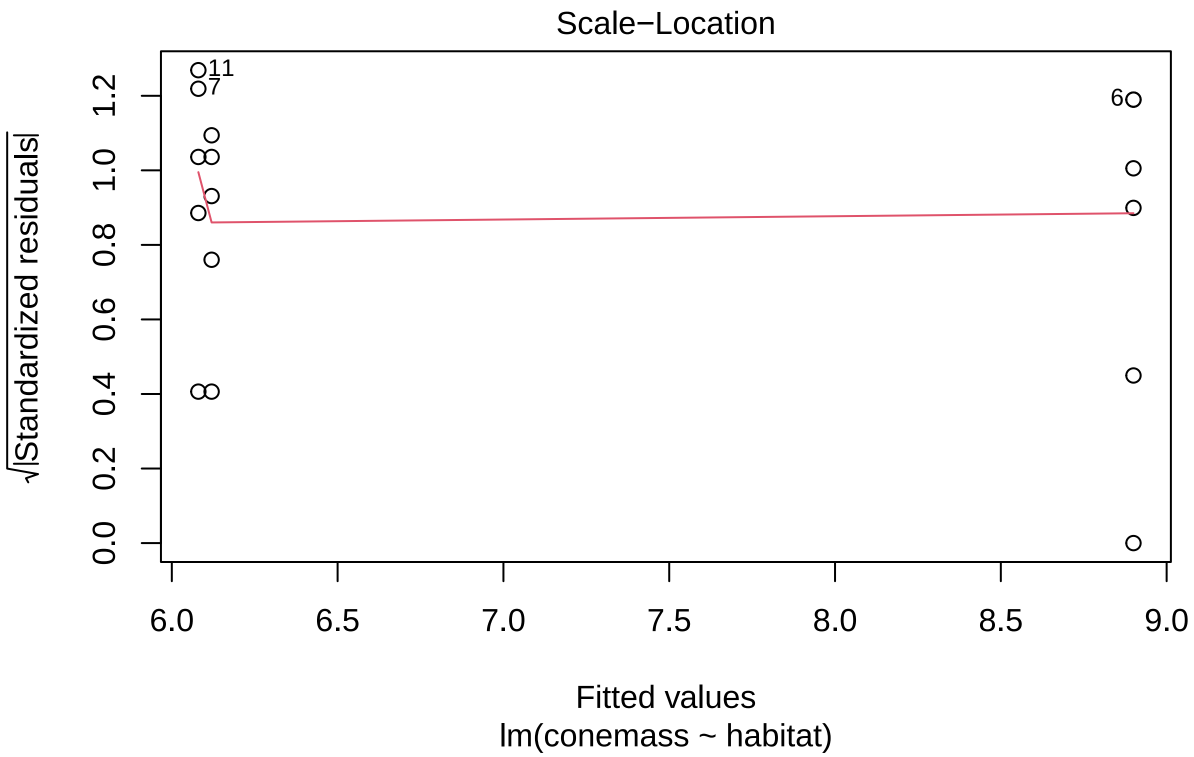
We saved the output of the model in the object, `p.lm`, which we will use for several purposes, below.

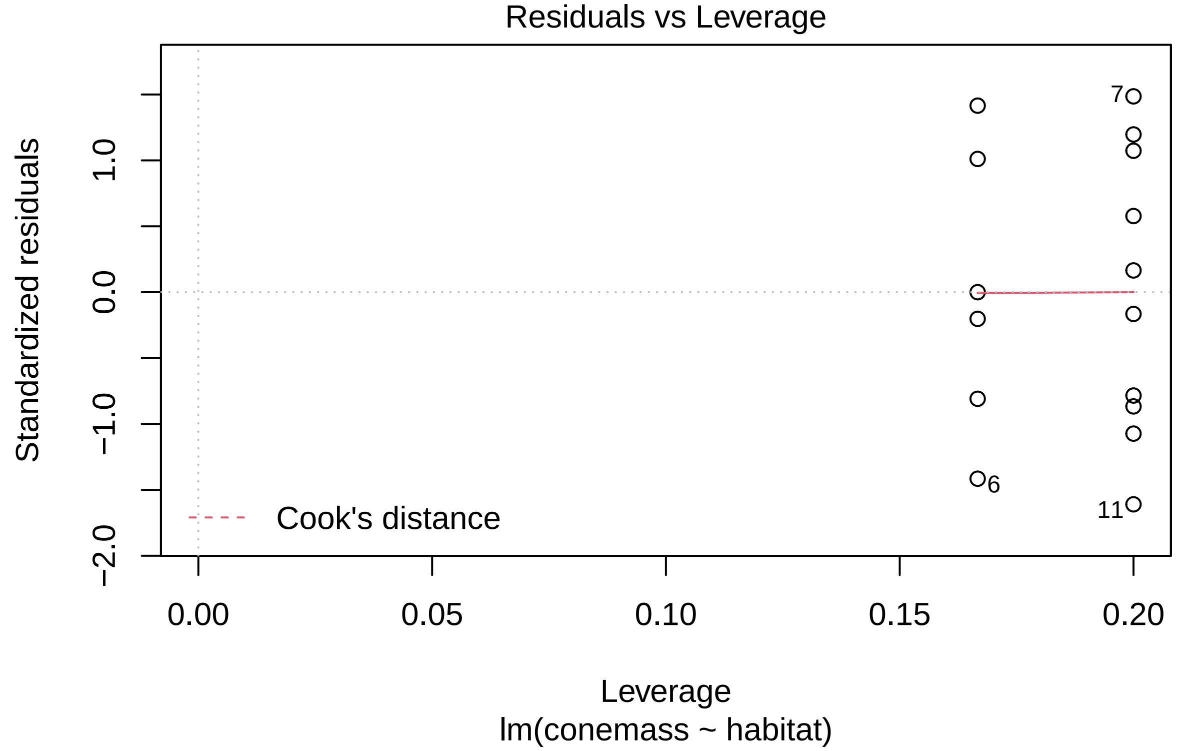
After running a model, we always check the assumptions before we proceed:

```r
plot(p.lm)
```









Let’s focus on the first three plots, above.

The first plot allows us to test the assumption of equal variance. We can see three columns of residuals, one at the right of the plot (at about 8.8 along the x-axis), and two columns very close toegther at the left (at about 6.05 and 6.10). These three columns represent the three treatments. The (vertical) distribution of residuals is very similar among the three columns, indicating that the data meet the assumption of equal variance.

The second plot allows a test of normally distributed residuals. This plot looks fine (we check whether the points fall along the dotted line): a few points fall far from the dotted line, but this is alright; it is also difficult to test normality with such a small dataset.

The third plot provides another perspective to test the assumption of equal variance. We see that the red line is relatively straight, which is consistent with equal variance. We also see that the points are relatively evenly distributed above and below the red line (also consistent with equal variance), except that the points below the red line on the right of the figure are a bit ‘stretched out’. Overall we’re happy that this plot suggests equal variance.

At this point, we’re satisfied that the data meet the assumptions of equal variance and normality. The paper from which these data came indicates that the data were randomly sampled and are independent. Therefore, our data meet the assumptions and we can check the results.

We’ll begin by looking at the summary of the `lm()` output:

```
summary(p.lm)
```

```
##
## Call:
## lm(formula = conemass ~ habitat, data = pine)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -0.780 -0.405 -0.040  0.505  0.720
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)               8.9000     0.2212  40.238 4.97e-15 ***
## habitatisland.present    -2.8200     0.3281  -8.596 1.01e-06 ***
## habitatmainland.present  -2.7800     0.3281  -8.474 1.18e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5418 on 13 degrees of freedom
## Multiple R-squared:  0.8851, Adjusted R-squared:  0.8675
## F-statistic: 50.09 on 2 and 13 DF,  p-value: 7.787e-07
```

Let’s begin by figuring out which level of `habitat` has been assigned to the `Intercept()`. In the second and third rows, we find both `island.present` and `mainland.present` listed; therefore `Intercept()` must refer to `island.absent` (this is also expected, as `island.absent` is the first in alphabetical order, which is how **R** assigns the `Intercept()`).

Let’s focus on the output for the coefficients. This output matches our expectations from our plot, above:

1. The `(Intercept)`, i.e., the mean value of `island.absent` (see `Estimate` column), equals 8.9; this nicely matches with our guess of about 9.0 when we examined the boxplot. The p-value for this mean (i.e., in column `Pr(>|t|)`) equals 4.97e-15 (i.e., 4.97 * 10^{-15}), indicating that the mean pinecone mass in `island.absent` habitats significantly differs from zero. This is not exciting, because we’d never expect pinecones to have a mass that equals zero.
2. Let’s look now at the `Estimate` values for `island.present` and `mainland.present` (i.e., the second and third rows of the Coefficients). These values equal -2.82 and -2.78, respectively, and indicate that the mean values of the `island.present` and `mainland.present` treatments are 2.82 and 2.78 units less than the mean value for `island.absent` (the `(Intercept)`) (i.e. the minus sign indicates that these means are smaller than the intercept). Again, these values match our expectations from the boxplot: there, we predicted that the means of these latter two treatments would be approximately 2.5 units smaller than that for `island.absent`. The p-values in these rows (1.01e-06 and 1.18e-06, respectively) are both very small, providing strong evidence that that these estimates of differences

from the (Intercept) differ from zero. i.e., the mean of these two habitat types differ from the mean of 'island.absent'. We will return to this when we examine the output of a pairs() function, below.

Consider what we have gained by examining the summary of our model, p.lm, as we did just above: we confirmed our intuition, and obtained information about the effect sizes for this experiment.

Now, we can look at an ANOVA table for our model output:

```
anova(p.lm)
```

```
## Analysis of Variance Table
##
## Response: conemass
##           Df Sum Sq Mean Sq F value    Pr(>F)
## habitat    2 29.404 14.7020  50.085 7.787e-07 ***
## Residuals 13  3.816  0.2935
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The output displays a p-value (Pr(>F)) that is very small, suggesting strong evidence that conemass depends on habitat; i.e., we can reject our null hypothesis that the mean conemass was equal among the levels of habitat. But, this result does not tell us which habitat types differ from which; this result also does not tell us about the effect sizes.

To determine which habitat types differ from which and to obtain effect sizes, we will perform a post-hoc (Tukey) test using two functions in the emmeans library.

```
library(emmeans)
p.emmeans <- emmeans(p.lm,"habitat")
p.emmeans
```

```
##  habitat          emmean    SE df lower.CL upper.CL
##  island.absent      8.90 0.221 13     8.42     9.38
##  island.present     6.08 0.242 13     5.56     6.60
##  mainland.present   6.12 0.242 13     5.60     6.64
##
## Confidence level used: 0.95
```

The output, above, indicates the mean values of the three habitat types, with their Standard Errors and 95% Confidence Intervals. (This is very useful when reporting results!). Generally, the function emmeans can compute estimated marginal means (EMMs) for specified factors or factor combinations in a linear model. In other words, emmeans can calculate the mean values for something, having taken into account the influence of other variables included in a model (in our current model, there is only one independent variable).

Note that we used the function emmeans to first calculate the mean values (or, EMMs) that interest us (i.e., the mean values that we wish to compare with one another); we will use these mean values in the pairs() function, below. Note as well how we specified the emmeans function: emmeans(p.lm,"habitat"). Within this statement, we first indicated the output of the analysis that we wished to use: this is p.lm, the object in which we saved the output from the function, lm(). Second, we indicated the *independent* variable in the model for which we want to estimate means for its three levels (island.absent, island.present, and mainland.present).

Before we continue, compare the means (emmean) from the p.emmeans output with the Estimate values of the Coefficients in the output from summary(p.lm), on the previous page. Convince yourself that these sets of output are consistent. (If you're unsure, ask for help.)

Now that we've calculated the mean values (EMMs) of conemass for the three levels of habitat, we can compare them using a Tukey test. We do this using the pairs() function (a part of the emmeans library).

We make comparisons among the pairs of means of `habitat` types, and save the results of these comparisons in the object we'll call `p.pairs`:

```
p.pairs <- pairs(p.emmeans)
p.pairs
```

```
##  contrast                          estimate    SE df t.ratio p.value
##  island.absent - island.present        2.82 0.328 13   8.596  <.0001
##  island.absent - mainland.present      2.78 0.328 13   8.474  <.0001
##  island.present - mainland.present    -0.04 0.343 13  -0.117  0.9925
##
## P value adjustment: tukey method for comparing a family of 3 estimates
```

There are two main conclusions that we can draw from this output. First, we can obtain p-values for comparisons of mean values among treatment pairs. Notice that each row has a different 'contrast', listed in the left-hand column. For example, the first row shows the results from comparing the mean value of `island.absent` treatment vs. the mean value of `island.present` treatment. The column `estimate` reveals the size of the difference between these means; this is a measure of **effect size**; the column `SE` displays the Standard Error for this effect size; i.e., `SE`, here, provides a sense of uncertainty (due to sampling error) in our `estimate` of the size of the difference between the mean of the `island.absent` treatment and the mean of the `island.present` treatment.

Before we proceed with looking at the p-values, compare the `estimate` values and `SE` values in the output of `p.pairs` to the `Estimate` and `Std. Error` in the output of, `summary(p.lm)`, above. (Focus on the comparisons between the `island.absent` treatment and the remaining treatments; remember that the mean of `island.absent` is assigned to the `(Intercept)` in the output of `summary(p.lm)`). What do you notice?

You should notice that the estimates and SE's are identical for the two outputs. If this output is identical, why did we not simply use the output from `summary(p.lm)` in order to make comparisons among treatments? Here are a few reasons:

1. The output from `summary(p.lm)` does not include all of the comparisons that we'd like. For instance, it does not include a comparison between the means values of the `island.present` and `mainland.present` treatments.
2. While the estimates and SE's are identical between the outputs, the p-values are not. Notice in the output from `p.pairs` the comment, `P value adjustment: tukey method for comparing a family of 3 estimates`. This implies that the `pairs()` function adjusted the p-values (and only the p-values) using the Tukey method to account for multiple comparisons. Recall that this is the reason for using Tukey test: **to adjust p-values in a manner that ensures that we do not inflate our Type 1 error rate due to making multiple comparisons**. This means that we should use the p-values from the output of `p.pairs` to help decide whether we have evidence that the mean values differ between treatment groups.

The p-values in `p.pairs` provide strong evidence that the `island.absent` treatment differs from the other two treatments (both p-values < 0.0001), but we have no evidence that the latter two treatments (`island.present` and `mainland.present`) differ from each other (p-value $= 0.9925$). Moreover, if we look at the `estimate` of the difference between `island.present` and `mainland.present`, we see that this effect size (-0.04) is very close to zero; this observation provides further evidence that there is very little difference in mean `conemass` between the `island.present` and `mainland.present` treatments.

As a final step in our analysis, we should obtain 95% Confidence Intervals for our estimates of effect size. We do this to provide an indication of the uncertainty in our estimates of effect size. We can use the `confint()` function to obtain these 95% CI's:

```
confint(p.pairs)
```

```
##  contrast                          estimate    SE df lower.CL upper.CL
##  island.absent - island.present        2.82 0.328 13    1.954    3.686
##  island.absent - mainland.present      2.78 0.328 13    1.914    3.646
##  island.present - mainland.present    -0.04 0.343 13   -0.945    0.865
##
## Confidence level used: 0.95
## Conf-level adjustment: tukey method for comparing a family of 3 estimates
```

This output provides 95% Confidence Intervals for the effect sizes; note, however, that these 95% Confidence Intervals have been adjusted to account for multiple comparisons, which makes the 95% CI's wider. Note that, if we wanted to obtain 95% CI's for our effect sizes *without* adjusting for multiple comparisons, we could calculate those, instead. Whether we report 95% CI's for effect sizes that are adjusted for multiple comparisions or not is a matter of some debate, and will be addressed in a future video.

What do these 95% CI's tell us? Loosely speaking, they indicate a range of plausible differences (effect sizes) in `conemass` between the `habitat` treatment combinations. Note that the lower 95% CI's (`Lower.CL`) for the effect size is approximately 1.9 when comparing `island.absent` to the remaining treatments. What does this mean, biologically? This means that we can say with reasonable confidence that the difference in mean `conemass` between the `island.absent` treatment and the remaining treatments is at least 1.9. If we consider that the mean `conemass` in the treatments where squirrels were present (`island.present` and `mainland.present`) is about 6.1 (see `p.emmeans` output), then 'island' sites that lack squirrels (`island.absent`) produce cones that at least ~31% larger (`1.9 / 6.1 = 0.311`). To me, this suggests a very big biological effect of squirrels on `conemass`; e.g., if `conemass` is directly proportional to seed production, then this result suggests that seed production will be much higher in the absence of squirrels.

To complete an analysis, we always want to have a biological interpretation of the results. I've indicated some of this interpretation in the paragraph immediately above: sites that do vs. do not have squirrels differ greatly in the mass of cones produced by trees. As trees produce larger cones in sites where squirrels are absent, this presumably suggests that squirrels impose natural selection that leads trees to produce smaller cones.

There's one further aspect to our biological interpretation: what can we learn from the lack of difference between the `island.present` and `mainland.present` treatments? These two treatments differed with respect to the type of forest that surrounded the study sites. The lack of difference (inferred both from the p-value and effect size) in `conemass` between *these* treatments suggests that cones in 'island' habitats are similar to those in 'mainland' habitats; i.e., there is nothing unusual about 'islands' with respect to `conemass`.

How might we report these results? We will not provide a complete example, but we'll provide some pieces.

First, you want to ensure that you plot the data and include a figure legend that explains how to interpret the plot.

You also want to state which assumptions were tested and how:

- Random sampling or allocation to treatment: we assumed this is met from the experimental design;
- Independence within treatments: we assumed this is met from the experimental design;
- Equal variance: the data meet this assumption based on visualizing residuals;
- Normally distributed residuals: the data meet this assumption based on visualizing residuals.

And, report your results…

A 1-Factor GLM provided strong evidence that `habitat` influences `conemass` (`F(2,13) = 50.085`; `p = 7.8e-07`). When squirrels were present, conemass was similar between `island` sites (mean `+/-` SE; 95% CI's: 6.08 `+/-` 0.242; 5.56 to 6.60) and `mainland` sites (6.12 `+/-` 0.242; 5.60 to 6.64) (Contrast estimate `+/-` SE: -0.04 `+/-` 0.343; t ratio = -0.117; Tukey Adjusted p-values and 95% CI's: p = 0.9925; 95% CI's: -0.945 to 0.865), suggesting surrounding habitat does not influence `conemass`.

On the other hand, `conemass` in island sites where squirrels were absent (8.90 `+/-` 0.221; 8.42 to 9.38) was greater than both in both island sites with squirrels present (Contrast estimate: 2.82 `+/-` 0.328; t ratio = 8.596; Tukey adjusted p-values and 95% CI's: `p < 0.0001`, 1.95 to 3.69) and mainland sites with squirrels present (2.78 `+/-` 0.328, t ratio = 8.474; `p < 0.0001`, 1.91 to 3.65). This result is consistent with the hypothesis that `conemass` is greater in the absence of squirrels.

Finally, we should include some kind of perspective on the size of the differences and their biological interpretation, as suggested, above.

Questions 2 - Eelgrass

Please note that answers to questions 2 & 3 will include less commentary than for Question 1. The general procedure is similar for all three questions, as are the things to look out for. We encourage you to explore your output for Questions 2 & 3 is the same detail as done for Question 1, but we will not repeat all explanations here.

In this question, Reusch et al. (2005) aimed to test whether genetic diversity of a species in a region affects the abundance of that species.

Let's begin by importing the data:

```
eg <- read.table("eelgrass.csv", header = TRUE, sep = ',')
```

Let's look at the data:

```
eg
```

```
##    treatmentGenotypes shoots
## 1          1_genotype     71
## 2          1_genotype     61
## 3          1_genotype     49
## 4          1_genotype     38
## 5          1_genotype     36
## 6          1_genotype     32
## 7          1_genotype     30
## 8          1_genotype     28
## 9          1_genotype     27
## 10         1_genotype     21
## 11         1_genotype     14
## 12         1_genotype     11
## 13        3_genotypes     67
## 14        3_genotypes     58
## 15        3_genotypes     53
## 16        3_genotypes     52
## 17        3_genotypes     47
## 18        3_genotypes     46
## 19        3_genotypes     41
## 20        3_genotypes     36
## 21        3_genotypes     35
## 22        3_genotypes     20
## 23        6_genotypes     86
## 24        6_genotypes     84
## 25        6_genotypes     69
## 26        6_genotypes     64
## 27        6_genotypes     62
## 28        6_genotypes     48
## 29        6_genotypes     45
## 30        6_genotypes     31
## 31        6_genotypes     47
## 32        6_genotypes     45
```

The dataset has two columns: `treatmentGenotypes` and `shoots`. Now let's obtain a summary:

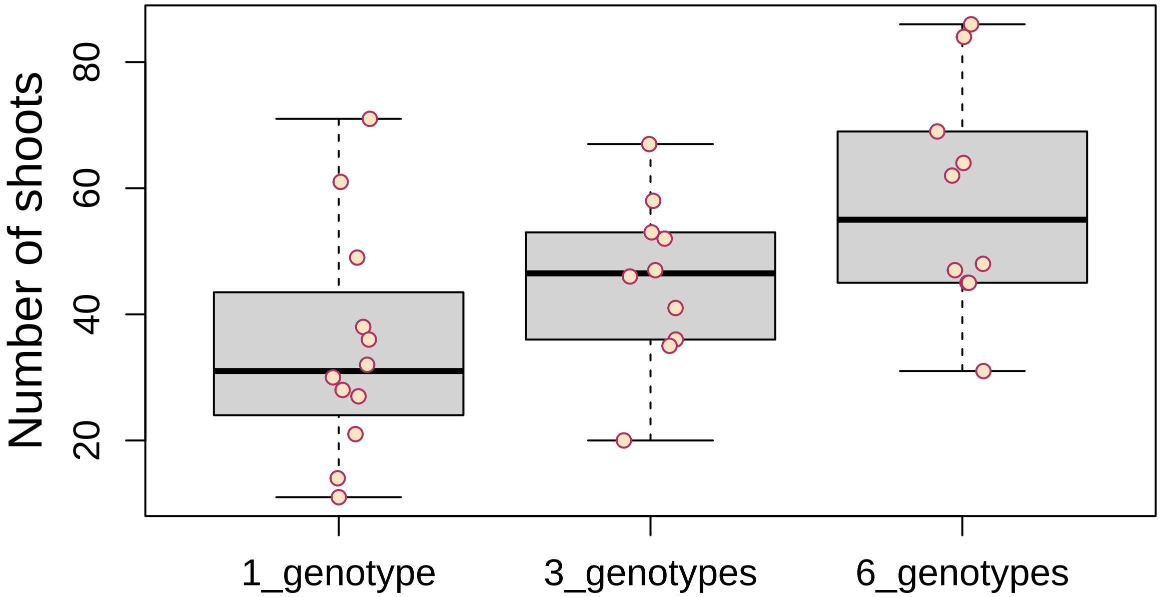
```
summary(eg)
```

```
##  treatmentGenotypes     shoots
##  Length:32          Min.   :11.00
##  Class :character   1st Qu.:31.75
##  Mode  :character   Median :45.50
##                     Mean   :45.44
##                     3rd Qu.:58.75
##                     Max.   :86.00
```

The column `treatmentGenotypes` has three levels, with either 1, 3 or 6 genotypes (these levels have 12, 10 and 10 datapoints, respectively). The mean number of `shoots` equals 45.44, with a range from 11.00 to 86.00.

Let's plot the data to get a sense of what to expect. Which variable will be the *dependent* variable? Our biological question is whether the number of genotypes (`treatmentGenotypes`) affects the number of shoots in a patch. This implies that `shoots` is the *dependent* variable, and `treatmentGenotypes` is the *independent* variable. With this in mind, we can plot the data as follows:

```
boxplot(shoots ~ treatmentGenotypes, data=eg, cex.axis = 1.2, xlab = "", ylab = "")
stripchart(shoots ~ treatmentGenotypes, data=eg,
           vertical = TRUE, method = "jitter",
           pch = 21, col = "maroon", bg = "bisque",
           add = TRUE)
mtext("Number of shoots",2,line=2.5,cex=1.5)
```



What do we learn from this plot?

- The boxplots are fairly symmetrical, implying that the data are likely normally distributed.
- There are no outliers to worry about.
- The breadth of the boxplots is fairly similar among the three treatments; this suggests that our data will meet the assumption of equal variance.
- `shoots` (y-axis) seems to increase as the number of genotypes (`treatmentGenotypes`) increases. Specifically, it looks like the mean number of shoots increases from about 35 for the first treatment (`1_genotype`), to about 45 (an increase of 10) for the second treatment; the average of the third treatment is about 25 larger than the first treatment.

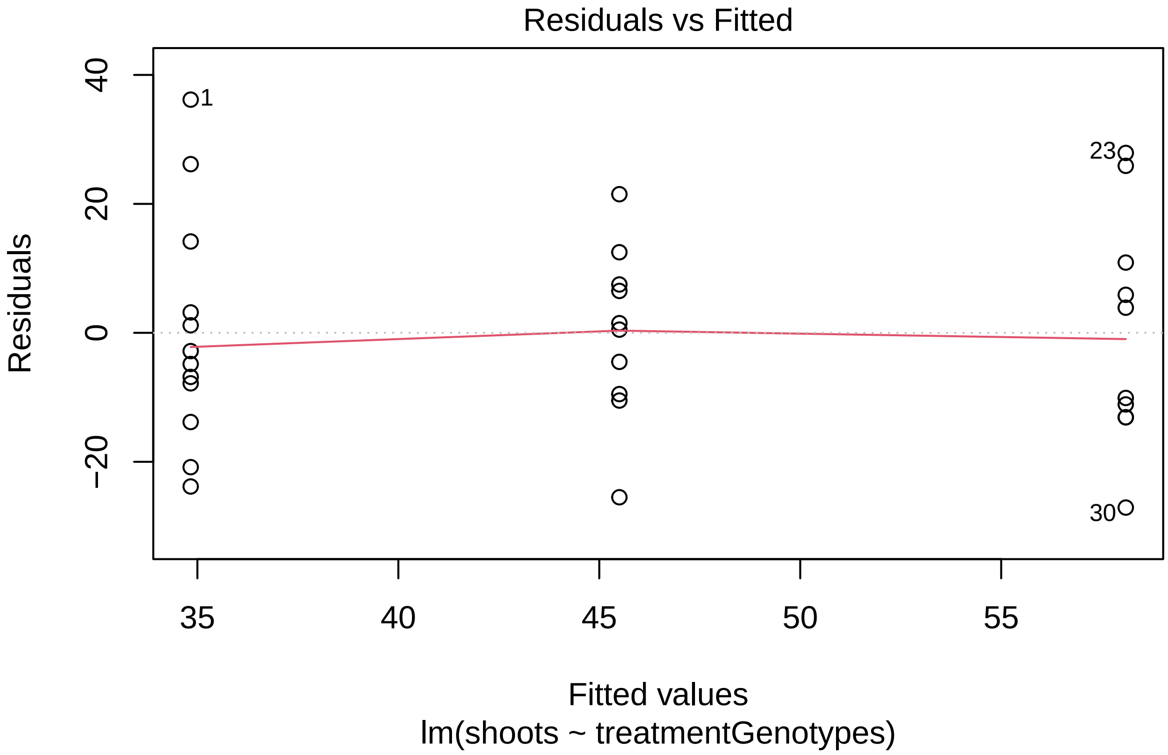
With these predictions in hand, let's make a model.

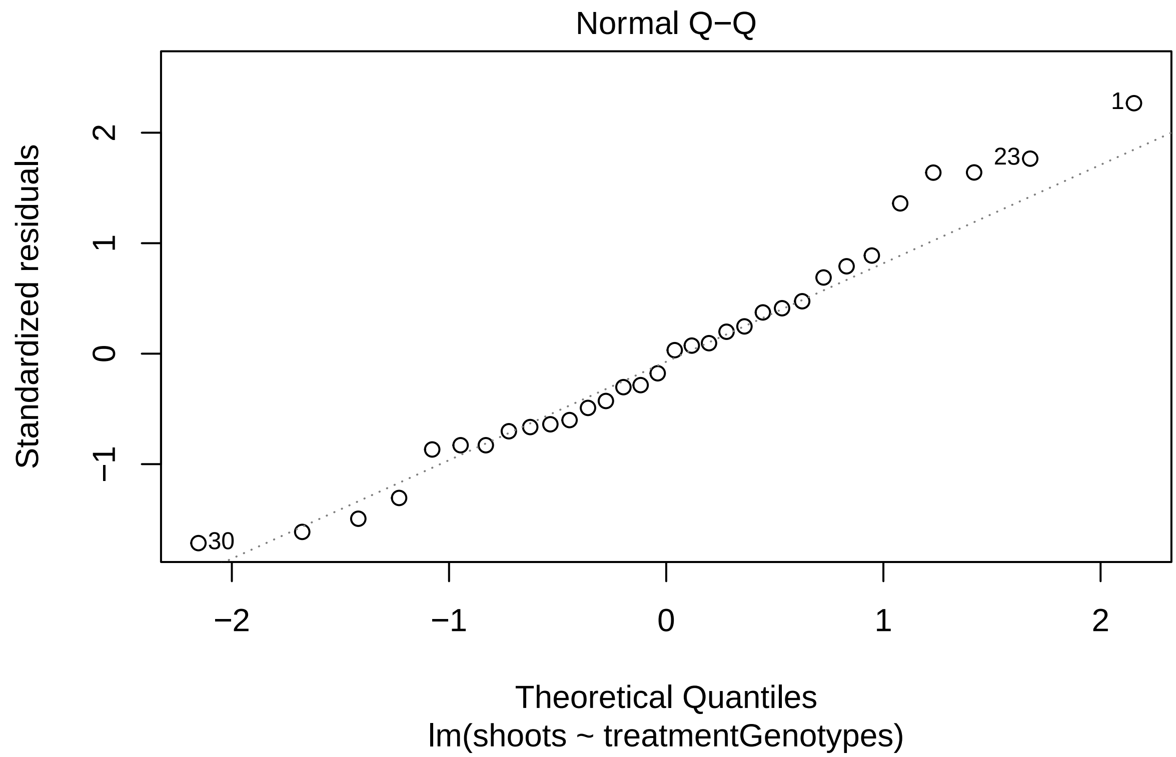
We've already decided that `shoots` is the dependent variable. Therefore, our model will be:

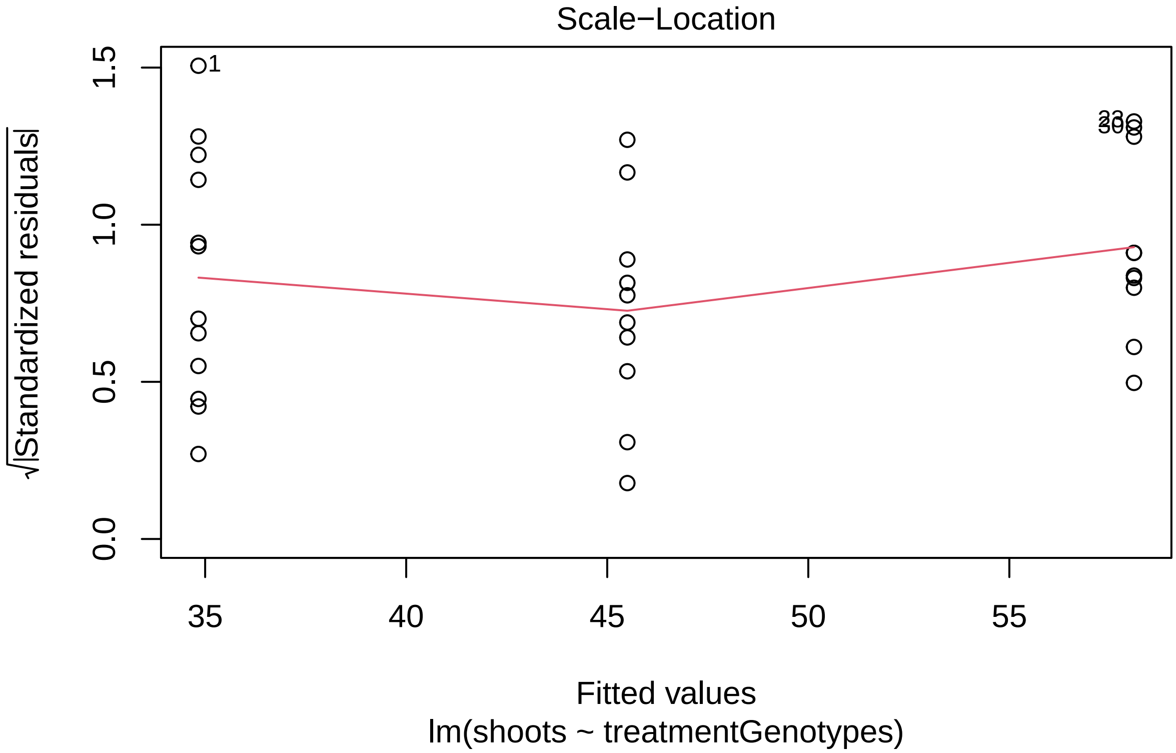
```
eg.lm <- lm(shoots ~ treatmentGenotypes, data = eg)
```

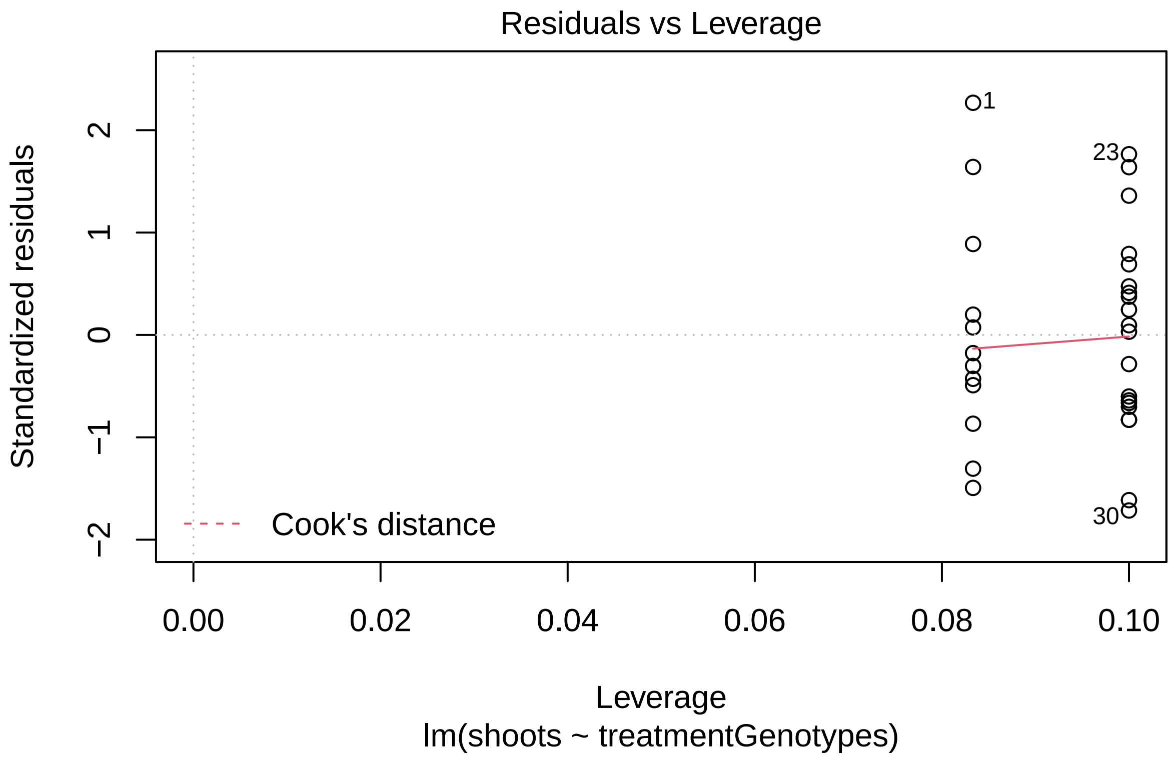
Let's check our assumptions:

```
plot(eg.lm)
```



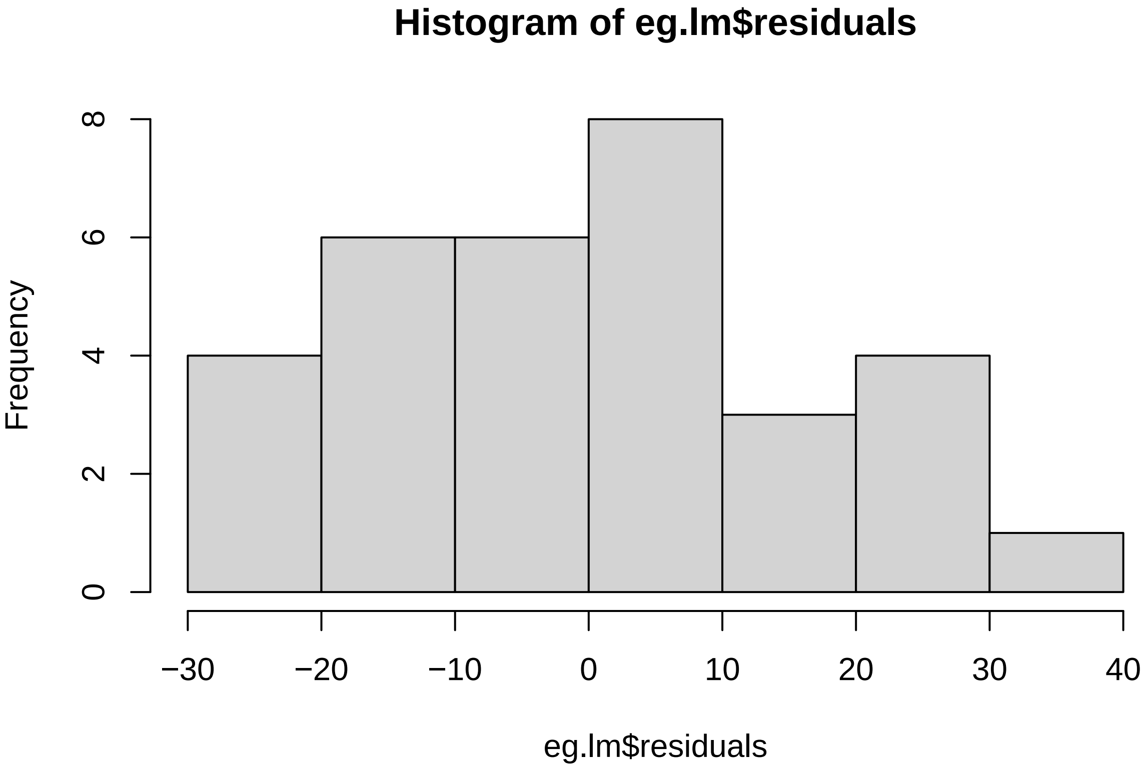

Normal Q−Q
Standardized residuals
2
1
0
−1
1
23
30
−2
−1
0
1
2
Theoretical Quantiles
lm(shoots ~ treatmentGenotypes)


Scale−Location
√|Standardized residuals|
1.5
1.0
0.5
0.0
1
23
30
35
40
45
50
55
Fitted values
lm(shoots ~ treatmentGenotypes)



The first plot indicates that the data meet the assumptions of equal variance. In the second plot, a number of data points fall off of the dotted line, but overall this qqplot indicates that the data meet the assumption of normality. We can also plot the data in this way, to see that the data adequately meet the assumption of normality:

```
hist(eg.lm$residuals)
```



The third residual plot is also consistent with equal variance.

Now that we’re satisfied that the data meet the assumptions, we can examine the model output. Let’s start by looking at the model summary to check that the model provides our expected results *(this is always a wise way to check whether we’ve thought our analysis through correctly)*:

```
summary(eg.lm)
```

```
##
## Call:
## lm(formula = shoots ~ treatmentGenotypes, data = eg)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -27.100 -10.650  -1.167   8.350  36.167
##
## Coefficients:
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                      34.833      4.809   7.243 5.64e-08 ***
## treatmentGenotypes3_genotypes   10.667      7.133   1.495  0.14563
## treatmentGenotypes6_genotypes   23.267      7.133   3.262  0.00283 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 16.66 on 29 degrees of freedom
## Multiple R-squared:  0.2684, Adjusted R-squared:  0.2179
## F-statistic: 5.319 on 2 and 29 DF,  p-value: 0.01077
```

Which treatment level of `treatmentGenotypes` is the `(Intercept)`? By process of elimination, we can see that the `(Intercept)` represents the mean value of treatment `1_genotype`.

Notice that the `Estimate` values match our expectations from our plot, above. Also, we see that the p-values indicate that the mean value of `shoots` differs between the treatments `1_genotypes` and `6_genotypes`. *(But remember that these p-values do not account for multiple comparisons.)*

With this in mind, let’s check the overall p-value of our model:

```
anova(eg.lm)
```

```
## Analysis of Variance Table
##
## Response: shoots
##                    Df Sum Sq Mean Sq F value  Pr(>F)
## treatmentGenotypes  2 2952.8 1476.40  5.3193 0.01077 *
## Residuals          29 8049.1  277.55
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The p-value for the overall model equals 0.01077. Overall, this represents moderate, or ‘suggestive’ evidence for an effect of Genotype diversity on number of shoots. Therefore we are somewhat justified to reject the Null hypothesis that the mean values of the three treatments all equal one another; i.e., we tentatively conclude that some mean values differ among the levels of `treatmentGenotype`. But, which treatments differ from which? We need a post-hoc test to answer this question and to estimate effect sizes; we’ll use a Tukey test, as implemented in ’emmeans‘. We’ll start by calculating the mean values of each treatment:

```
eg.emmeans <- emmeans(eg.lm, "treatmentGenotypes")
eg.emmeans
```

```
##  treatmentGenotypes emmean   SE df lower.CL upper.CL
##  1_genotype           34.8 4.81 29     25.0     44.7
##  3_genotypes          45.5 5.27 29     34.7     56.3
##  6_genotypes          58.1 5.27 29     47.3     68.9
##
## Confidence level used: 0.95
```

These results present the means of each treatment group, as well as SE's and 95% CI's for each group; a final report would include this information. As well, the next stage of our analysis will compare these means to one another, using the `pairs()` function:

```
eg.pairs <- pairs(eg.emmeans)
```

Let's look at the output of the `pairs()` function, which we stored in the object, `eg.pairs`:

```
eg.pairs
```

```
##  contrast                  estimate   SE df t.ratio p.value
##  1_genotype - 3_genotypes     -10.7 7.13 29 -1.495  0.3079
##  1_genotype - 6_genotypes     -23.3 7.13 29 -3.262  0.0077
##  3_genotypes - 6_genotypes    -12.6 7.45 29 -1.691  0.2256
##
## P value adjustment: tukey method for comparing a family of 3 estimates
```

The left-most column, `contrast`, indicates which sets of means are being compared (i.e., the first row compares the means from treatment `1_genotype` vs. `3_genotypes`). The column, `estimate`, provides the estimated difference between the means of the two treatments; notice that all the estimates are negative because in all cases a larger mean was subtracted from a smaller mean (also notice the similarity to the `Estimate` reported in `summary(eg.lm)`, above). `SE` indicates the standard error for the estimated difference between the means, and could be used to generate 95% CI's for these estimated differences. Finally, we see `Pr(>|t|)`, the p-values for these comparisons. Only one p-value provides reaonable evidence to a differene between treatments, which corresponds to the comparison between `1_genotype` and `6_genotypes`. (We might have guessed this when we plotted our data at the start of this analysis!). Overall, this stage of the analysis indicates that we do have evidence that changing genetic diversity affects the number of shoots; specifically, the treatment with highest diversity had more shoots than the treatment with the lowest diversity. The effect size indicates that increasing genetic diversity from 1 to 6 genotypes almost doubled the number of shoots.

Our reported results should include 95% CI's for the effect sizes, which are calculated like this:

```
confint(eg.pairs)
```

```
##  contrast                  estimate   SE df lower.CL upper.CL
##  1_genotype - 3_genotypes     -10.7 7.13 29    -28.3     6.95
##  1_genotype - 6_genotypes     -23.3 7.13 29    -40.9    -5.65
##  3_genotypes - 6_genotypes    -12.6 7.45 29    -31.0     5.80
##
## Confidence level used: 0.95
## Conf-level adjustment: tukey method for comparing a family of 3 estimates
```

At this stage we should be thinking clearly about the biological interpretation of these results. The effect size is important: the finding that shoot number almost doubled when increasing genetic diversity in this experiment (and this effect was had strong statistical support) implies that genetic diversity can strongly affect growth and/or survival.

We can say something more about the differences between the treatments. We find little evidence that the number of shoots produced in the `3_genotypes` treatment differs from either the `1_genotype` or `6_genotypes` treatments. However, the estimated effect sizes suggest potentially large effects: the number of shoots differs by about 50% when comparing `3_genotypes` to the remaining treatments. Thus, although we did not detect differences with strong statistical support, there is reason to believe that the differences could prove important with further study.

On that point, it is worth noting that, perhaps, we did not analyze the data in an optimal manner. The original biological question could have been asked as, "Does the number of shoots increase with the genetic diversity of plots?". This phrasing would suggest that we're really interested in a *relationship* between the number of genotypes and the numbers of shoots produced; e.g., is there a linear relationship between these variables (see our plot of these data)? This kind of analysis is called a *regression*, which we'll learn in the near future.

Reporting the results

Provide a figure, as described in Question 1.

Address the assumptions, as done in Question 1.

A 1-Factor GLM provides moderate evidence that genotype diversity affects shoot number (F(2,29) = 5.32; p = 0.011). Shoot number under six genotypes (mean `+/-` SE; 95% CI's: 58.1 `+/-` 5.27; 47.3 to 68.9) was approximately twice as great as when a single genotype occurred (34.8 `+/-` 4.81; 25.0 to 44.7) (Contrast estimate `+/-` SE: -23.3 `+/-` 7.13; t ratio = -3.262; Tukey adjusted p-value = 0.0077). However (Tukey adjusted) 95% CI's for this contrast (-40.9 to -5.65) suggest that shoot number might increase by as little as ~ 16% (i.e., `5.65/34.8`) to as much as 117%.

Shoot number in the three-genotype treatment (45.5 `+/-` 5.27; 34.7 to 56.3) tended to be intermediate between that of treatments with one genotype (Contrast 1 genotype - 3 genotypes estimate `+/-`SE: -10.7 `+/-` 7.13; t ratio = -1.495; Tukey adjusted p-value and 95% CI's: p = 0.3079; -28.3 to 6.95) and that with six genotypes (Contrast 3 genotype - 6 genotypes estimate: -12.6 `+/-` 7.45; t ratio = -1.691; Tukey adjusted p-value and 95% CI's: p = 0.2256; -31.0 to 5.80).

Question 3 - tsetse learning

This experiment has only two treatments. As a result we could analyze these data using a t-test. But, we'll use a 1-Factor general linear model just for practice.

Let's import the data:

```
tse <- read.table("tsetselearning.csv", header = TRUE, sep = ',')
```

And, let's look at the dataset:

```
tse
```

```
##    treatment proportionCow
## 1     lizard          0.66
## 2     lizard          0.58
## 3     lizard          0.52
## 4     lizard          0.37
## 5     lizard          0.35
## 6     lizard          0.34
## 7     lizard          0.29
## 8        cow          1.00
## 9        cow          0.98
## 10       cow          0.97
## 11       cow          0.96
## 12       cow          0.87
## 13       cow          0.83
```

It is a small dataset; only 13 datapoints. We have two variables: `proportionCow` (the proportion of flies that fed on cows) and `treatment`, which notes whether the flies previously fed on cows vs. lizards, before being offered cow.

We can summarise the data as follows:

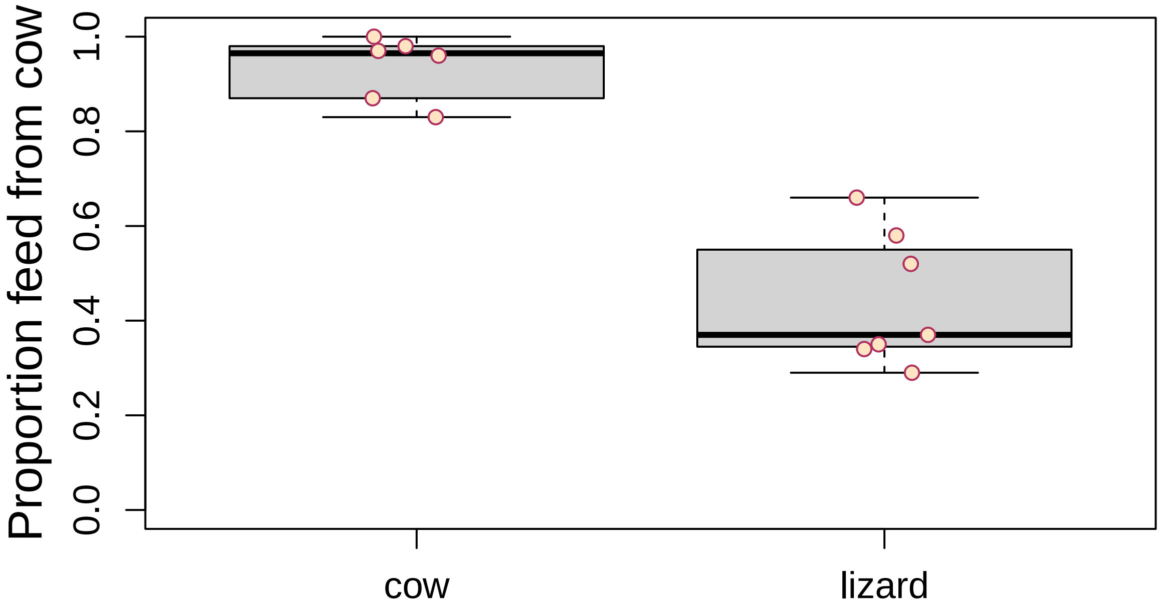
```
summary(tse)
```

```
##   treatment          proportionCow
##  Length:13          Min.   :0.2900
##  Class :character   1st Qu.:0.3700
##  Mode  :character   Median :0.6600
##                     Mean   :0.6708
##                     3rd Qu.:0.9600
##                     Max.   :1.0000
```

Given that the dataset is so small, our `summary(tse)` does not reveal much that we couldn't glean from looking at the complete dataset.

Let's begin our analysis by plotting the data. Which variable will be the *dependent* variable? The biological question is, "Does previous experience of feeding on cow vs. lizard affect a fly's decision to feed when presented with cow, later?" This question implies that `proportionCow` will be the dependent variable; we'll plot the data as:

```
boxplot(proportionCow ~ treatment, data=tse, cex.axis = 1.2, xlab = "", ylab = "", ylim = c(0,1))
stripchart(proportionCow ~ treatment, data=tse,
           vertical = TRUE, method = "jitter",
           pch = 21, col = "maroon", bg = "bisque",
           add = TRUE)
mtext("Proportion feed from cow",2,line=2.5,cex=1.5)
```



This plot reveals:

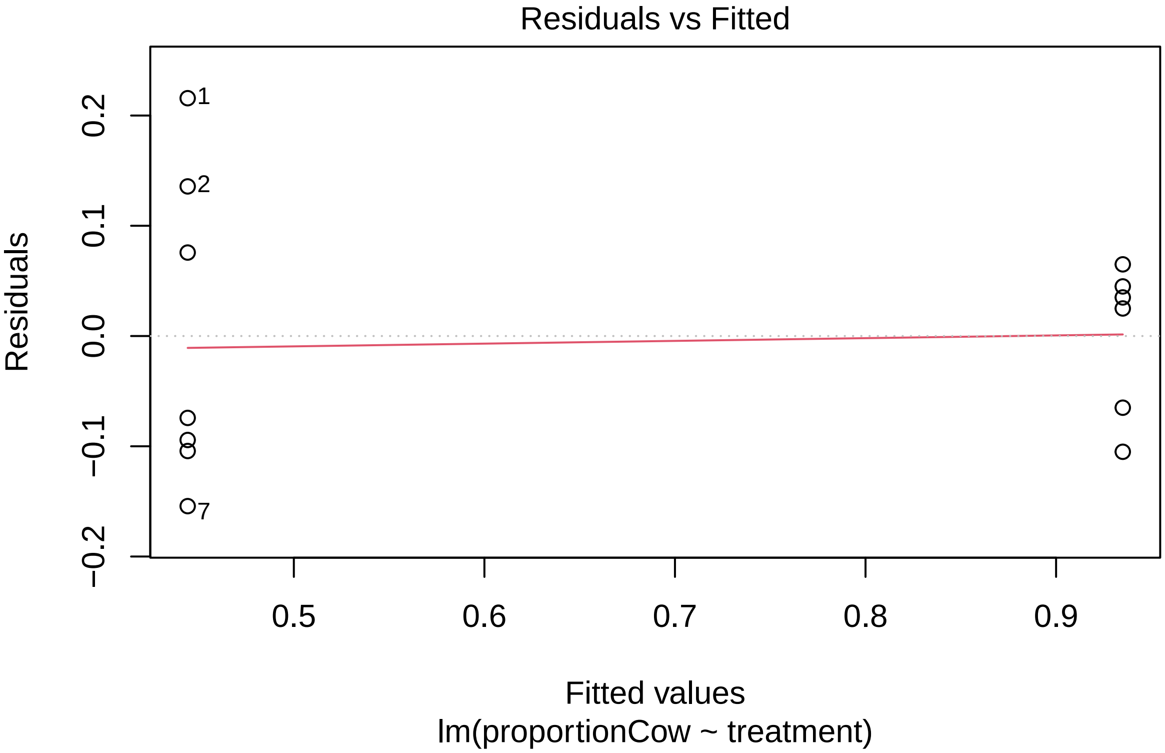
- The data may be normally distributed (largely symmetrical boxplots), but with so few datapoints it is hard to know with certainty.
- It is unclear whether the variance is equal between the treatments.
- We expect a strong difference between the treatments (note the lack of overlap in the distributions of the two treatments). This difference is likely to be about 0.5
- Our y-axis reminds us that the dependent variable is a proportion. Proportions are unlikely to be normally distributed, and we might be advised to try transforming the data... But we've not learned that yet! We'll learn to do so in future videos; for the moment we'll push onward.

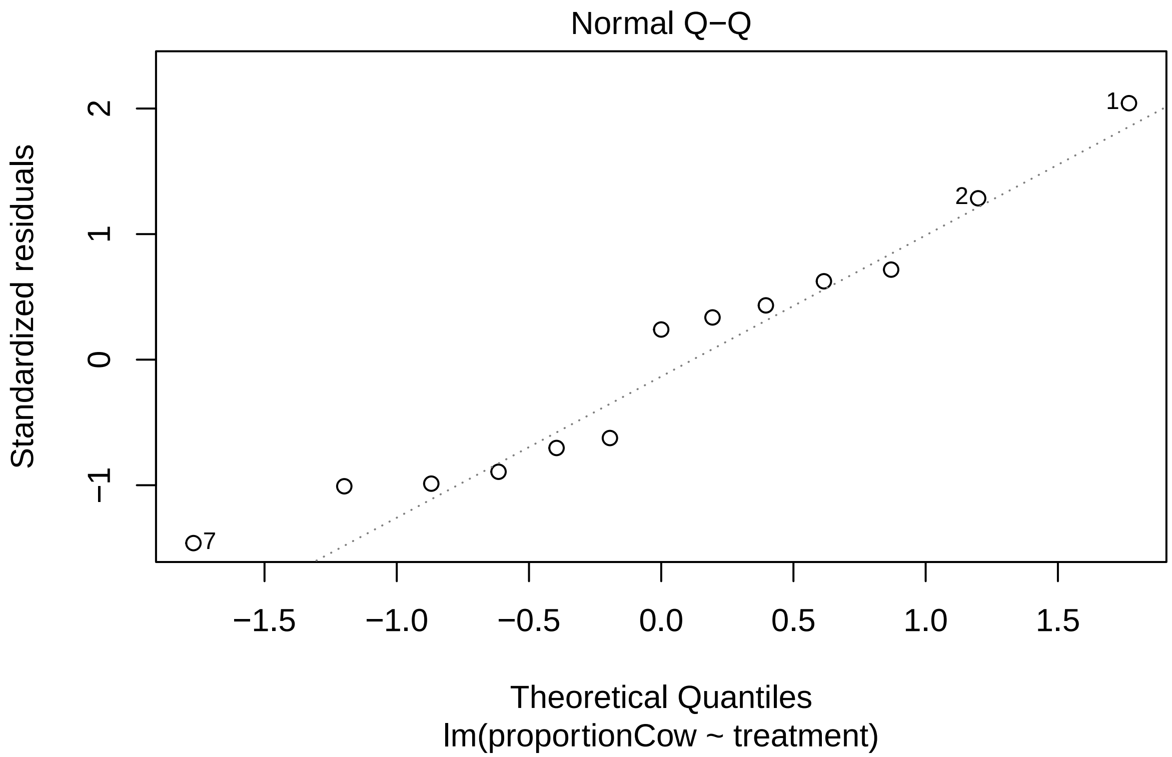
With these predictions in mind, and remembering that the dependent variable is `proportionCow`, we can model the data like this:

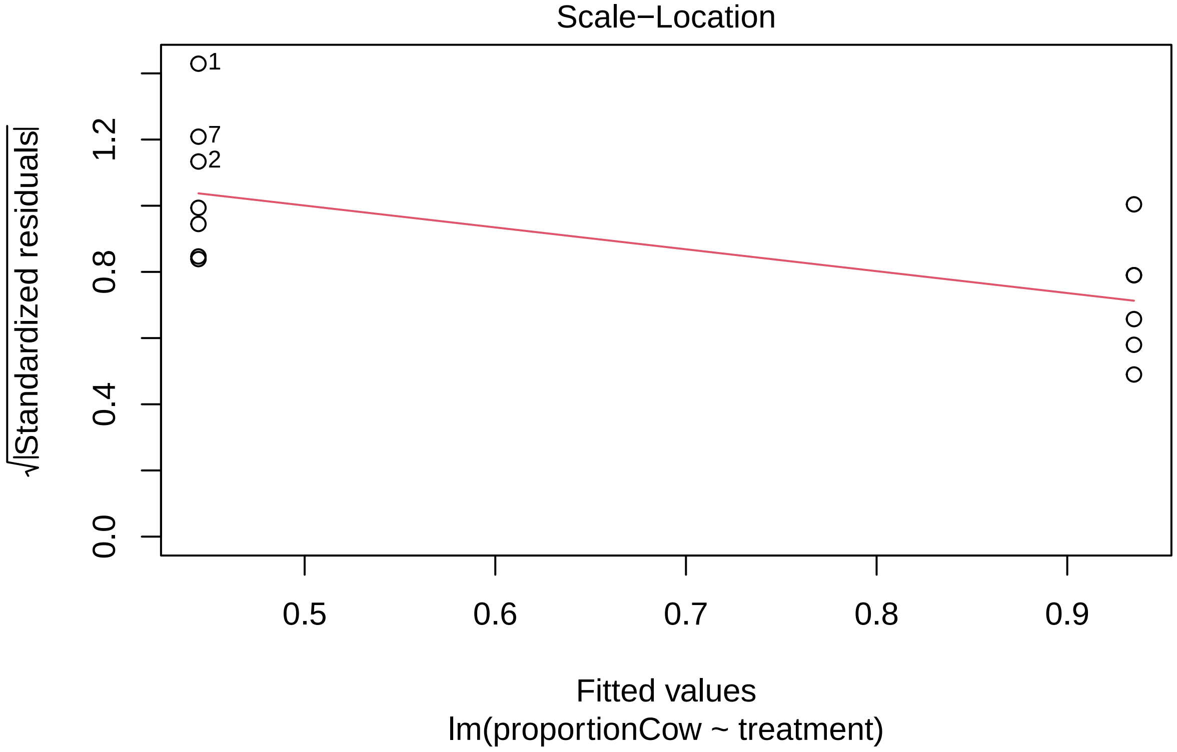
```
tse.lm <- lm(proportionCow ~ treatment, data=tse)
```

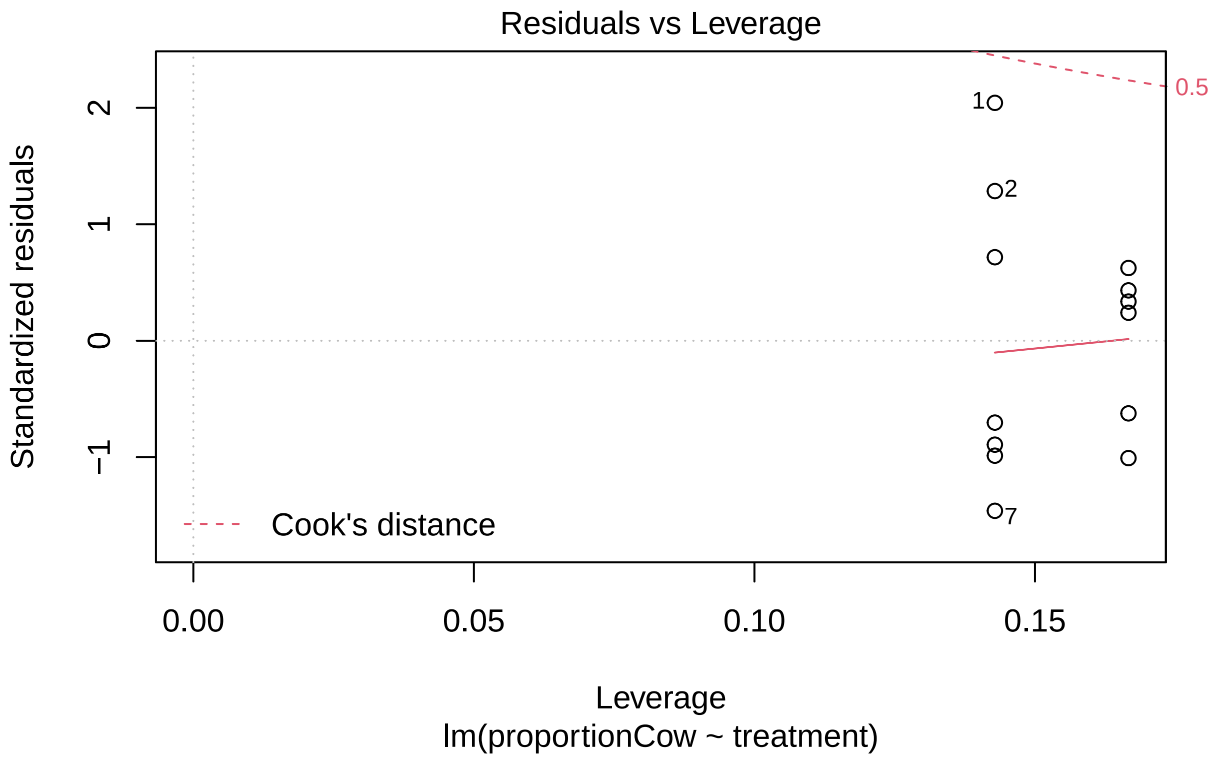
Let's check the assumptions:

```
plot(tse.lm)
```




Normal Q−Q
Standardized residuals
−1
0
1
2
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5
7
2
1
Theoretical Quantiles
lm(proportionCow ~ treatment)


Scale−Location
√|Standardized residuals|
0.0
0.4
0.8
1.2
0.5
0.6
0.7
0.8
0.9
1
7
2
Fitted values
lm(proportionCow ~ treatment)



From this plot we note:

1) from the first plot, it is difficult to tell whether the variances are (sufficiently) equal;
2) the second plot suggests that the data are reasonably normally distributed.
3) The third plot is worrying with respect to equal variance.

The potential for unequal variance is worrying. In future videos we'll learn how to deal with this (e.g., transform the data, or use a Welch's t-test). For now, however, just for the sake of practice, we'll continue our analysis but bear in mind that unequal variance might impact our results.

Let's look at the summary of our model output:

```
summary(tse.lm)
```

```
##
## Call:
## lm(formula = proportionCow ~ treatment, data = tse)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.15429 -0.09429  0.02500  0.06500  0.21571
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)      0.93500    0.04657  20.078 5.12e-10 ***
## treatmentlizard -0.49071    0.06346  -7.733 9.01e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1141 on 11 degrees of freedom
## Multiple R-squared:  0.8446, Adjusted R-squared:  0.8305
## F-statistic: 59.79 on 1 and 11 DF,  p-value: 9.011e-06
```

This output differs from that in Questions 1 & 2 in that, here, we have only two treatments. By process of elimination, we infer that the `(Intercept)` corresponds to the mean value of `proportionCow` for the treatment, cow. Its `Estimate` (0.935) matches our plot of the data, above.

The second row of the `Coefficients` indicates the difference between the mean `proportionCow` for the treatments cow vs. lizard. Its `Estimate` equals -0.49071, which matches our initial guess of about 0.5 when we plotted the data, above. This value (0.49071) is the effect size in this experiment, and we can use the `Std. Error` for this estimate as the standard error of our effect size (which we could convert to a 95% confidence interval, if we wished). The p-value for this difference (see `Pr(>|t|)`) equals 9.01e-06 (i.e., 0.00000901), which is very small. Hence, we have strong evidence to reject the Null hypothesis that the mean values of the cow and lizard treatments are equal.

How does this p-value compare to a p-value that we obtain using the `anova()` command?

```
anova(tse.lm)
```

```
## Analysis of Variance Table
##
## Response: proportionCow
##           Df  Sum Sq Mean Sq F value    Pr(>F)
## treatment  1 0.77797 0.77797  59.793 9.011e-06 ***
## Residuals 11 0.14312 0.01301
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We can see that the p-values are identical. This is because the analysis only has two treatments.

Do we need to perform a post-hoc test (e.g., Tukey test)? No. We only perform a post-hoc test when we wish to determine which treatments differ from which; we do not need to do this when an experiment has only two treatments because, in this case, it is obvious which treatments differ (it must be that the ‘cow’ treatment differs from the ‘lizard’ treatment). As we do not need to perform a post-hoc test, and we can obtain an estimate of the effect size and its SE from the output of `summary(tse.lm)`, we’ve got pretty much everything we need to report our results. But, we’re missing 95% CI’s for our effect size and estimates of ean mean (with SE and 95% CI). We’ll get them like this:

```
tse.emmeans <- emmeans(tse.lm, "treatment")
tse.emmeans
```

```
##  treatment emmean     SE df lower.CL upper.CL
##  cow        0.935 0.0466 11    0.833    1.037
##  lizard     0.444 0.0431 11    0.349    0.539
##
## Confidence level used: 0.95
```

```
confint(pairs(tse.emmeans))
```

```
##  contrast     estimate     SE df lower.CL upper.CL
##  cow - lizard    0.491 0.0635 11    0.351     0.63
##
## Confidence level used: 0.95
```

Or final step is to interpret the results in terms of biology. The experiment set out to test whether the tsetse fly learned from prior exposure to eating either cow or lizard. The results reveal that flies that previously fed on cows were highly likely to feed on cow again (about 93.5% of flies who had previously fed on cows decided to feed on cow again), but flies that has previously fed on lizard were much less likely to feed when presented with cow later on. Notably, this effect was large: only about half the flies that had previously fed on lizard decided to feed on cows. These results suggest that prior feeding experience strongly influences subsequent feeding behaviour in the tsetse fly. Do we call that learning? I don’t know. But it does suggest that the environment affects a tsetse fly’s behaviour.

Reporting

Plot your data with useful legend.

Discuss assumptions (we're worried… we should use methods we'll learn later in this course to address the problems).

A 1-Factor GLM revealed strong evidence that prior feeding experience affected feeding preference (F(1,11) = 59.793; p = 9.01e-06). The proportion of flies that fed on cows after being initially presented with lizards (mean `+/-` SE; 95% CI: 0.444 `+/-` 0.0431; 0.349 to 0.539) was lower than that observed when flies that had initially fed on cows (0.935 `+/-` 0.0466; 0.833 to 1.037) (contrast estimate `+/- SE`: 0.491 `+/-` 0.0635). Indeed, 95% CI's for this contrast suggest that the proportion of tse flies that subsequently fed on cow was between 0.351 to 0.63 greater when tse flies had initially fed on cow compared to when they had initially fed on lizard.