Variable and Model Selection

Dr Josh Hodge

Introduction

In the lecture, we covered a range of ideas:

1. Exploring and eradicating collinearity with variance-inflation factors
2. Model selection procedures through:

- Model specification and interpret regardless
- The hypothesis testing procedure via stepwise methods
- The Information Theoretic (IT) approach and the construction of a range of models

In today's practical, we are going to implement these procedures using a dataset from the ParkGrass Experiment.

```
require(usdm)
```

```
## Loading required package: usdm
```

```
## Loading required package: terra
```

```
## Warning: package 'terra' was built under R version 4.3.3
```

```
## terra 1.7.78
```

```
require(psych)
```

```
## Loading required package: psych
```

```
## Warning: package 'psych' was built under R version 4.3.3
```

```
##
## Attaching package: 'psych'
```

```
## The following objects are masked from 'package:terra':
##
##     describe, distance, rescale
```

```
require(lmerTest)
```

```
## Loading required package: lmerTest
```

```
## Loading required package: lme4
```

```
## Warning: package 'lme4' was built under R version 4.3.3
```

```
## Loading required package: Matrix

##
## Attaching package: 'lmerTest'

## The following object is masked from 'package:lme4':
##
##     lmer

## The following object is masked from 'package:stats':
##
##     step

require(sjPlot)

## Loading required package: sjPlot

## Warning: package 'sjPlot' was built under R version 4.3.3

## Learn more about sjPlot with 'browseVignettes("sjPlot")'.

parkgrass<- read.csv("parkgrass_ms.csv")
str(parkgrass)

## 'data.frame':    68 obs. of  14 variables:
##  $ Site              : chr  "10a" "10b" "10c" "10d" ...
##  $ CWM.Plant.Height  : num  0.463 0.478 0.441 0.437 0.878 ...
##  $ CWM.LDMC          : num  0.22 0.24 0.225 0.316 0.279 ...
##  $ CWM.SLA           : num  23.4 25.3 25 28 28.4 ...
##  $ CWM.Seed.Mass     : num  0.978 0.762 0.704 0.495 1.964 ...
##  $ CWM.Leaf.Thickness: num  0.217 0.192 0.204 0.115 0.136 ...
##  $ CWM.Leaf.N        : num  22.2 22.1 22 19 26.5 ...
##  $ CWM.C.N           : num  21.2 21.3 21.4 23.4 18 ...
##  $ SpRich            : int  14 9 8 3 10 10 8 1 9 8 ...
##  $ Harvest           : num  3.95 3.62 3.87 2.78 5.66 ...
##  $ Ammonium          : int  2 2 2 2 3 3 3 3 3 3 ...
##  $ Nitrate           : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ Minerals          : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ pH                : num  6.74 6.02 4.9 3.66 6.58 6.08 4.94 3.64
6.82 6.06 ...
```

The Park Grass Experiment

The Park Grass Experiment, begun in 1856, is the oldest ecological experiment in existence. Its value to ecology has changed and grown since it was founded to answer agricultural questions. It is a mosiac of sites that have received various fertiliser and liming treatments and the botanical composition studied. In recent times the experiment has shown *inter alia* how: plant species richness, biomass and pH are related; community composition responds to climatic perturbation and nutrient additions; soil is acidified and corrected by liming. The dataset you have has a whole range of measures from:

- Biodiversity indices: functional richness and functional evenness (measures of functional diversity), community-weighted measures of plant height (metres), leaf dry matter content (gg^-1), specific leaf area (mm^{2mg}-1), seed mass (grams), leaf thickness (mm), leaf nitrogen content (mgg^-1), leaf C:N ratio and Species Richness (count of species)
- Environmental measures: ammonium sulphate addition (0=nil, 1=low, 2=medium, 3=high), sodium nitrate addition (0=nil, 1=low and 2=high), mineral fertiliser addition (e.g. potassium, sodium etc.; 0=no, 1=yes), pH.
- Biomass measure: harvest (kilos).
- Site (site name)

We are going to use these measures to try and address the research question: *What is driving the provisioning of biomass (harvest)?*

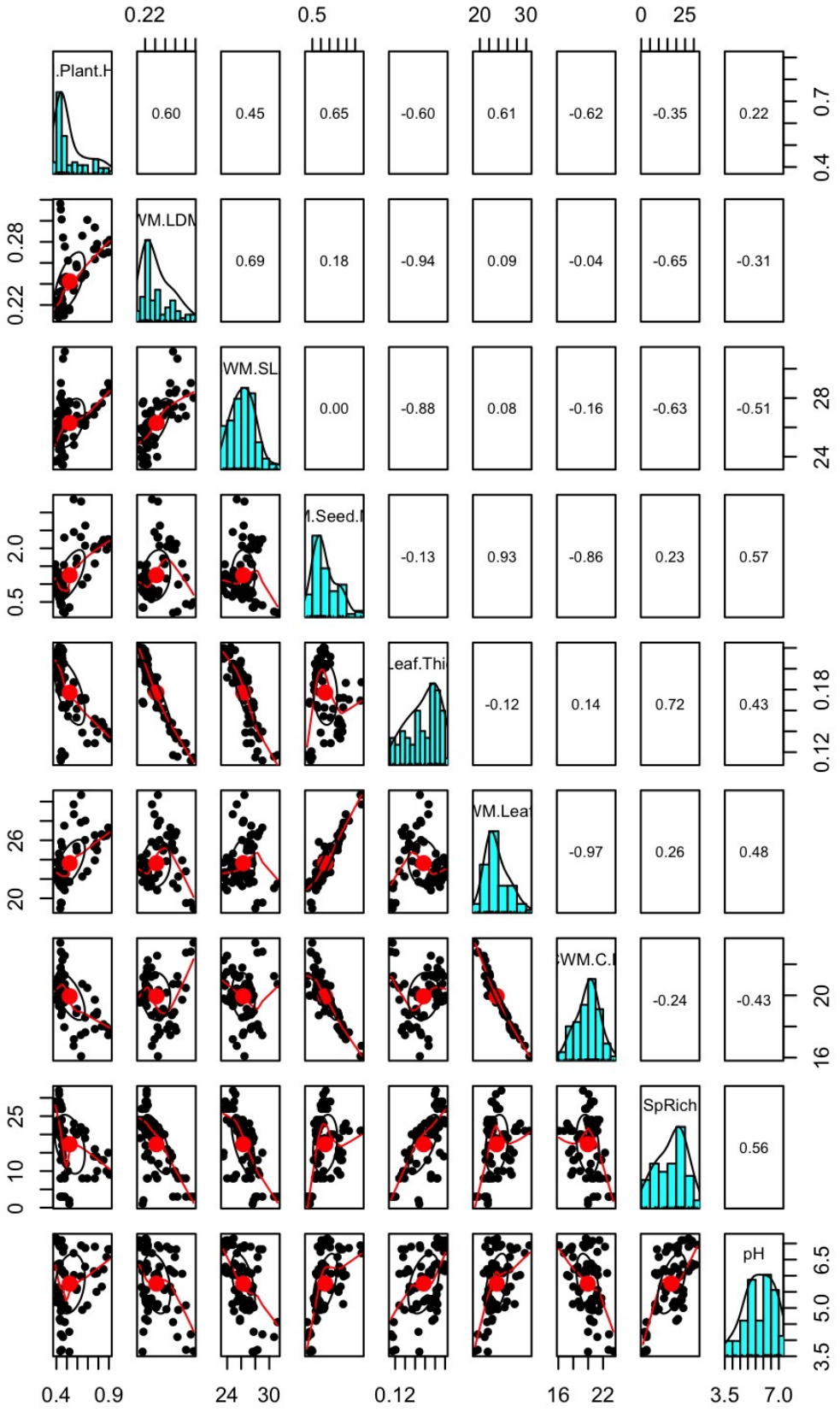
Exploring Covariates and Collinearity

The continuous covariates we are going to use for our modelling is everything except:

- Harvest (the response variable)

harvest (the response variable), ammonium sulphate addition (fixed factor), sodium nitrate addition (fixed factor) and mineral fertiliser addition (fixed factor). We can first explore collinearity through looking at the pairwise scatterplots and correlation.

```
pairs.panels(parkgrass[,-c(1,10,11,12,13)])
```


0.22
0.5
20 30
0 25
.Plant.H
0.60
0.45
0.65
-0.60
0.61
-0.62
-0.35
0.22
0.7
0.4
VM.LDM
0.69
0.18
-0.94
0.09
-0.04
-0.65
-0.31
0.28
0.22
WM.SL
0.00
-0.88
0.08
-0.16
-0.63
-0.51
28
24
.Seed.M
-0.13
0.93
-0.86
0.23
0.57
2.0
0.5
Leaf.Thi
-0.12
0.14
0.72
0.43
0.18
0.12
VM.Leaf
-0.97
0.26
0.48
26
20
WM.C.M
-0.24
-0.43
20
16
SpRich
0.56
25
10
0
pH
6.5
5.0
3.5
0.4 0.9
24 30
0.12
16 22
3.5 7.0

The pairs plot gives a good indication of the associations between the variables. There are strong correlations between:

- CWM.SLA and CWM.Leaf.Thickness; CWM.Seed.Mass and CWM.Leaf.N and CWM.C.N

Variance Inflation Factor

In the lecture, we explored the variance inflation factor and its implementation. In this walkthrough, we are going to sequentially remove the variable with the highest VIF until we reach the threshold of 3. You can choose to repeat this yourselves at the thresholds of 5 and 10 if you want to see how this affects your results.

```
vif(parkgrass[,-c(1,10,11,12,13)]) # let's remove CWM.Leaf.Thickness
and recalculate VIF
```

```
##            Variables       VIF
## 1   CWM.Plant.Height  6.426624
## 2           CWM.LDMC 43.722478
## 3            CWM.SLA 14.780998
## 4      CWM.Seed.Mass 11.889459
## 5 CWM.Leaf.Thickness 87.714537
## 6         CWM.Leaf.N 45.228890
## 7            CWM.C.N 47.746172
## 8             SpRich  4.306632
## 9                 pH  3.260847
```

```
vif(parkgrass[,-c(1,6,10,11,12,13)]) # let's remove CWM.Leaf.N
```

```
##          Variables       VIF
## 1 CWM.Plant.Height  6.347379
## 2         CWM.LDMC  6.196606
## 3          CWM.SLA  5.181288
## 4    CWM.Seed.Mass 11.870627
## 5       CWM.Leaf.N 44.003692
## 6          CWM.C.N 42.708801
## 7           SpRich  3.531756
## 8               pH  3.103518
```

```
vif(parkgrass[,-c(1,6,7,10,11,12,13)]) # let's remove CWM.Leaf.C.N
```

```
##          Variables       VIF
## 1 CWM.Plant.Height  5.461074
## 2         CWM.LDMC  5.235893
## 3          CWM.SLA  4.391994
## 4    CWM.Seed.Mass  8.888992
## 5          CWM.C.N 10.040921
## 6           SpRich  3.424583
## 7               pH  3.103137
```

```
vif(parkgrass[,-c(1,6,7,8,10,11,12,13)]) # let's remove
CWM.Plant.Height
```

```
##          Variables      VIF
## 1 CWM.Plant.Height 4.324883
## 2         CWM.LDMC 2.893797
## 3          CWM.SLA 2.765993
## 4    CWM.Seed.Mass 3.177948
## 5           SpRich 3.057156
## 6               pH 2.949539
```

```
vif(parkgrass[,-c(1,2,6,7,8,10,11,12,13)]) # # ok so all of our variable now have VIFs under 3! This is our final set of continuous covariates.
```

```
##       Variables      VIF
## 1      CWM.LDMC 2.708199
## 2       CWM.SLA 2.518032
## 3 CWM.Seed.Mass 2.006440
## 4        SpRich 2.493306
## 5            pH 2.517615
```

We have reduced down our variables using VIF and have our final variables to fit:

- Biodiversity measures: community-weighted means of leaf dry matter content, specific leaf area and seed mass; and species richness.
- Environmental measures: ammonium sulphate addition (0=nil, 1=low and 2=high), sodium nitrate addition (0=nil, 1=low, 2=medium, 3=high), mineral fertiliser addition (e.g. potassium, sodium etc.; 0=no, 1=yes), pH.

We are going to implement some model selection procedures. 1) fit your model and interpret regardless; 2) fit your maximal model and implement hypothesis testing and information criteria approaches (we'll do both) and finally 3) construct a range of models and analyse via the Information Theoretic approach.

The 'Do Nothing' Approach

Our initial research question was "what is driving the provisioning of biomass?" and we have a range of continuous covariates and fixed factors that could help us determine this question. We can choose to add in interactions but be mindful we only have 68 data points and don't want to fit too many and risk increasing our Type II errors. For this reason, we are going to fit a purely additive model. When you have multiple continuous explanatory variables on different scales, it is also best practice to z-standardize them so this is included in the code using the scale() function.

```
M1<- lm(Harvest~scale(CWM.LDMC)+scale(CWM.SLA)+scale(CWM.Seed.Mass)
+scale(SpRich)+factor(Ammonium)+factor(Nitrate)+factor(Minerals)
+scale(pH), data = parkgrass)
anova(M1)
```

```
## Analysis of Variance Table
##
## Response: Harvest
```

```
##                       Df  Sum Sq Mean Sq F value    Pr(>F)
## scale(CWM.LDMC)        1 12.1812 12.1812 37.7829 8.763e-08 ***
## scale(CWM.SLA)         1  1.2076  1.2076  3.7457 0.0580006 .
## scale(CWM.Seed.Mass)   1 29.7601 29.7601 92.3082 1.910e-13 ***
## scale(SpRich)          1 15.3586 15.3586 47.6384 5.023e-09 ***
## factor(Ammonium)       3  3.0249  1.0083  3.1275 0.0328008 *
## factor(Nitrate)        2  1.0728  0.5364  1.6638 0.1986393
## factor(Minerals)       1  4.8788  4.8788 15.1328 0.0002685 ***
## scale(pH)              1  1.0700  1.0700  3.3188 0.0738361 .
## Residuals             56 18.0544  0.3224
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Model Interpretation

From our analysis of variance, we are able to find out the variables that are having a significant effect. These are (in descending order of importance as inferred by the sum of squares - remember the sum of squares tells us how much variation in Harvest is explained by that variable): CWM.Seed.Mass, SpRich, CWM.LDMC; Minerals and Ammonium, suggesting that the botanical composition has a greater effect on harvest (overall) than the environmental measures. We can also interpret the summary table.

```
summary(M1)
```

```
##
## Call:
## lm(formula = Harvest ~ scale(CWM.LDMC) + scale(CWM.SLA) +
scale(CWM.Seed.Mass) +
##     scale(SpRich) + factor(Ammonium) + factor(Nitrate) +
factor(Minerals) +
##     scale(pH), data = parkgrass)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.70669 -0.27579  0.01743  0.30602  1.12966
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)            2.52318    0.27151   9.293 6.08e-13 ***
## scale(CWM.LDMC)       -0.01800    0.12105  -0.149 0.882358
## scale(CWM.SLA)        -0.26508    0.13406  -1.977 0.052933 .
## scale(CWM.Seed.Mass)   0.81608    0.12146   6.719 1.01e-08 ***
## scale(SpRich)         -0.02182    0.24521  -0.089 0.929411
## factor(Ammonium)1      0.62913    0.43211   1.456 0.150986
## factor(Ammonium)2      0.67383    0.45097   1.494 0.140742
## factor(Ammonium)3      1.26420    0.43045   2.937 0.004803 **
## factor(Nitrate)1       0.69639    0.24528   2.839 0.006292 **
## factor(Nitrate)2       0.79923    0.41320   1.934 0.058141 .
## factor(Minerals)1      0.85684    0.22645   3.784 0.000378 ***
## scale(pH)             -0.27422    0.15053  -1.822 0.073836 .
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5678 on 56 degrees of freedom
## Multiple R-squared:  0.7915, Adjusted R-squared:  0.7506
## F-statistic: 19.33 on 11 and 56 DF,  p-value: 2.941e-15
```

We need to remember that R has an annoying referencing system, and therefore the (Intercept) refers to plots with 'Nil' ammonium sulphate, sodium nitrate and mineral fertilisers. We can interpret however:

- As the community mean of seed mass increases: for every 1SD increase in community seed mass harvest increased by 0.82kg.
- Significant differences were found between 'Nil' ammonium sulphate and the high application condition with regards to harvest. Plots applied with high rates ammonium sulphate were found to harvest 1.26 kilos more of biomass.
- Significant differences were found between 'Nil' sodium nitrate and the low application condition with regards to harvest. Plots applied with low rates sodium nitrate were found to increase harvest 0.7 kilos.
- Significant differences were found between no mineral fertiliser addition and fertiliser addition. Plots applied with mineral fertilisers were found to increase harvest 0.86 kilos.

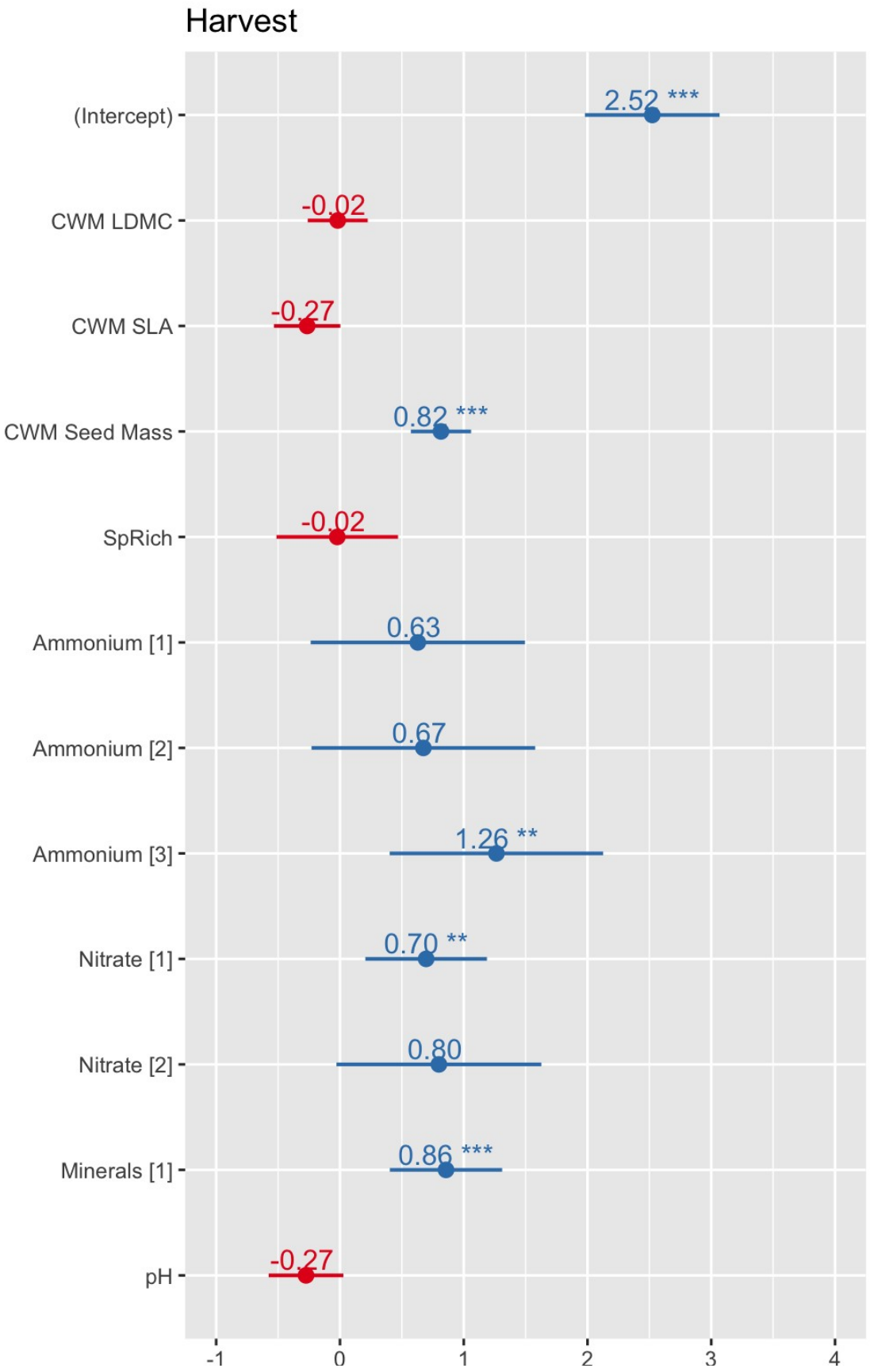
Finally, we can interpret that our model was able to explain 75.06% of variation (from the adjusted R^2) in harvest, which is good fitting model.

In the 'Do Nothing' approach, we simply interpret what comes from the model. We may decide to do some follow-up Tukey HSD analyses if we want to get into the intricacies of the fixed factors. But for now, we'll move on and look at the diagnostics of the model to validate it.

Plotting the Model

Models with lots of variables are notoriously hard to plot. We tend to present a forest plot and an example of one is shown here.

```
plot_model(M1, show.values = TRUE, show.intercept = TRUE)
```


Harvest
(Intercept)
CWM LDMC
CWM SLA
CWM Seed Mass
SpRich
Ammonium [1]
Ammonium [2]
Ammonium [3]
Nitrate [1]
Nitrate [2]
Minerals [1]
pH
2.52 ***
-0.02
-0.27
0.82 ***
-0.02
0.63
0.67
1.26 **
0.70 **
0.80
0.86 ***
-0.27
-1
0
1
2
3
4

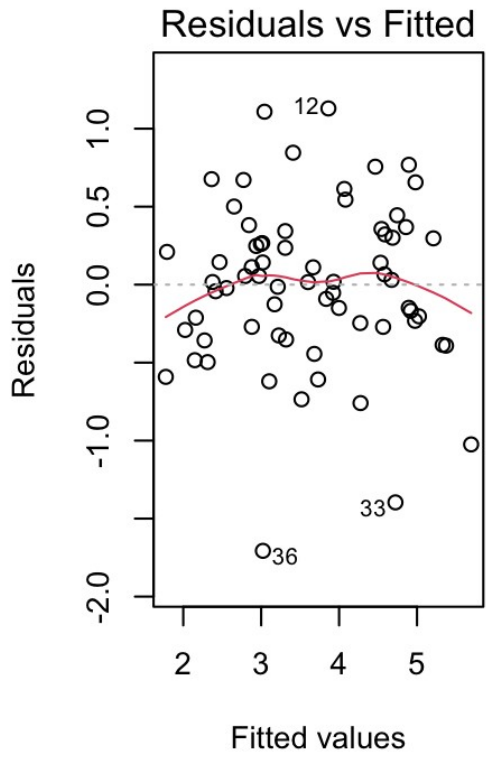
Model Validation

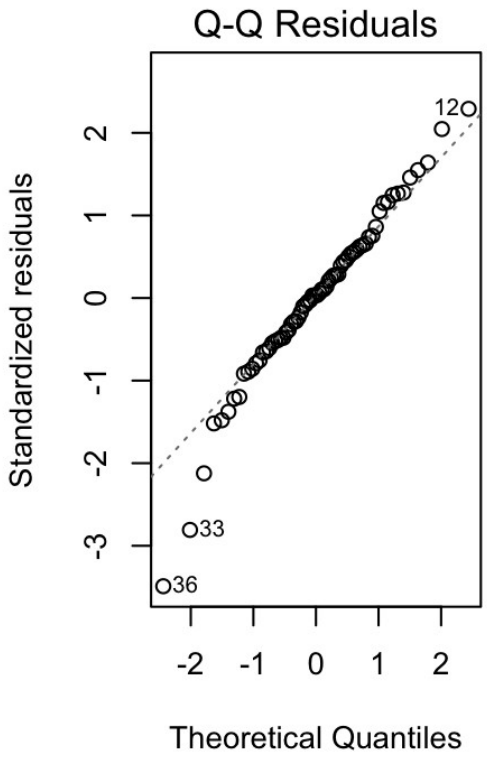
We can see quite clearly from the diagnostic plots that there are no violations of the assumptions.

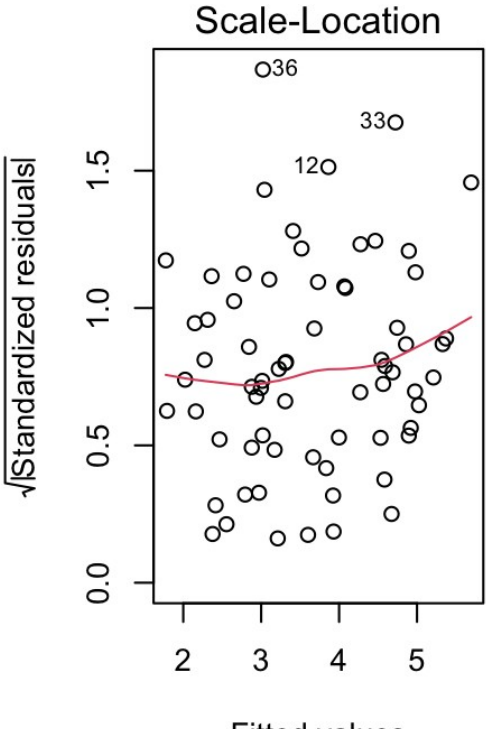
- The 'Residuals vs Fitted' and 'Scale-Location' plots indicate no violations of the homogeneity of variances as the points are evenly distributed ('Starry Night').
- The 'Normal Q-Q' has some minor deviations at the tails but the residuals are largely normal.
- The 'Residuals vs Leverage' plot indicates no outliers as they don't cross the Cook's distance of 1 dotted line.

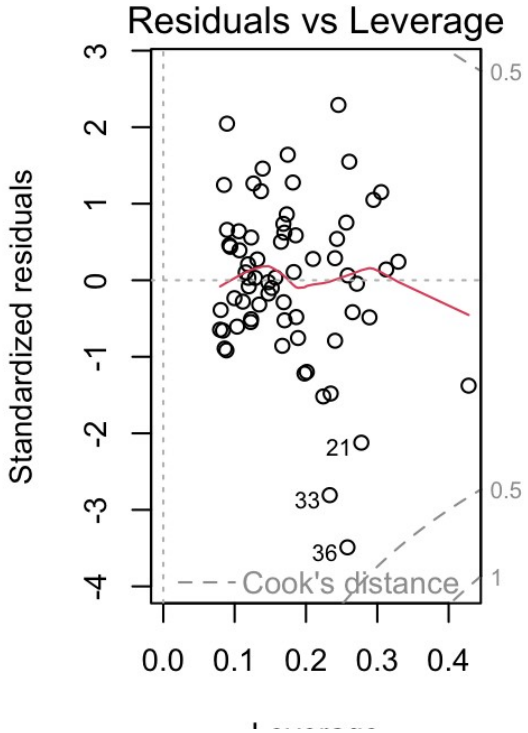
For this approach to model selection, the fitted model would be presented in its entirety - non-significant terms and all. But we may wish to reduce our model down to not include non-significant terms and variables and here is where hypothesis testing and information criteria approaches can come in - the classic model selection techniques. We'll first look at the hypothesis testing approach.

```
par(mfrow=c(2,2))
plot(M1)
```


Residuals vs Fitted
Residuals
Fitted values
12
33
36


Q-Q Residuals
Standardized residuals
Theoretical Quantiles
12
33
36


Scale-Location
√|Standardized residuals|
36
33
12


Residuals vs Leverage
Standardized residuals
Cook's distance
0.5
0.5
1
21
33
36

The Classic Approaches - Hypothesis Testing

Hopefully, you remember the two rules: 1) cannot remove insignificant terms or variables and 2) cannot remove a main effect without removing all interactions it is integral to. In this example, we only have main effects and no interactions so this should make our process a little simpler. First, we have to decide what variables can be dropped - we do this by investigating the summary output of our models.

```
summary(M1)
```

```
##
## Call:
## lm(formula = Harvest ~ scale(CWM.LDMC) + scale(CWM.SLA) +
scale(CWM.Seed.Mass) +
##     scale(SpRich) + factor(Ammonium) + factor(Nitrate) +
factor(Minerals) +
##     scale(pH), data = parkgrass)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.70669 -0.27579  0.01743  0.30602  1.12966
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)            2.52318    0.27151   9.293 6.08e-13 ***
## scale(CWM.LDMC)       -0.01800    0.12105  -0.149 0.882358
## scale(CWM.SLA)        -0.26508    0.13406  -1.977 0.052933 .
## scale(CWM.Seed.Mass)   0.81608    0.12146   6.719 1.01e-08 ***
## scale(SpRich)         -0.02182    0.24521  -0.089 0.929411
## factor(Ammonium)1      0.62913    0.43211   1.456 0.150986
## factor(Ammonium)2      0.67383    0.45097   1.494 0.140742
## factor(Ammonium)3      1.26420    0.43045   2.937 0.004803 **
## factor(Nitrate)1       0.69639    0.24528   2.839 0.006292 **
## factor(Nitrate)2       0.79923    0.41320   1.934 0.058141 .
## factor(Minerals)1      0.85684    0.22645   3.784 0.000378 ***
## scale(pH)             -0.27422    0.15053  -1.822 0.073836 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5678 on 56 degrees of freedom
## Multiple R-squared:  0.7915, Adjusted R-squared:  0.7506
## F-statistic: 19.33 on 11 and 56 DF,  p-value: 2.941e-15
```

From the summary table we could remove: CWM.LDMC, CWM.SLA, SpRich and pH. CWM.LDMC occurs has the highest p-value so we should drop this first. We can use a handy function to create a new model "M2" without rewriting the whole lm(Harvest...) - this function is *update*.

```
M2<- update(M1, .~.-scale(CWM.LDMC)) # we put a - sign before the name
of the term to tell R to remove that from the model
```

The next step is to compare M1 to M2 and see whether we have made our model significantly worse. We can do this comparison using the F-test for regular linear models or the log-likelihood test for GLMs or mixed models.

```
anova(M1,M2)
```

```
## Analysis of Variance Table
##
## Model 1: Harvest ~ scale(CWM.LDMC) + scale(CWM.SLA) +
scale(CWM.Seed.Mass) +
##     scale(SpRich) + factor(Ammonium) + factor(Nitrate) +
factor(Minerals) +
##     scale(pH)
## Model 2: Harvest ~ scale(CWM.SLA) + scale(CWM.Seed.Mass) +
scale(SpRich) +
##     factor(Ammonium) + factor(Nitrate) + factor(Minerals) +
scale(pH)
##   Res.Df    RSS Df  Sum of Sq      F Pr(>F)
## 1     56 18.054
## 2     57 18.061 -1 -0.0071248 0.0221 0.8824
```

In this step, we have not made our model significantly worse the RSS (residual sum of squares) has only increased from 18.05 to 18.06 and this is a negligible change. We therefore remove CWM.LDMC and M2 becomes our new model and let's investigate the summary table of M2 to see what can be removed next.

```
summary(M2) # let's remove SpRich
```

```
##
## Call:
## lm(formula = Harvest ~ scale(CWM.SLA) + scale(CWM.Seed.Mass) +
##     scale(SpRich) + factor(Ammonium) + factor(Nitrate) +
factor(Minerals) +
##     scale(pH), data = parkgrass)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.70200 -0.28066  0.01734  0.31011  1.15446
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)           2.52721    0.26782   9.436 3.02e-13 ***
## scale(CWM.SLA)       -0.27486    0.11579  -2.374 0.021002 *
## scale(CWM.Seed.Mass)  0.81009    0.11361   7.131 1.93e-09 ***
## scale(SpRich)        -0.01976    0.24271  -0.081 0.935398
## factor(Ammonium)1     0.62375    0.42688   1.461 0.149454
## factor(Ammonium)2     0.65743    0.43350   1.517 0.134907
## factor(Ammonium)3     1.25940    0.42553   2.960 0.004481 **
## factor(Nitrate)1      0.69697    0.24314   2.867 0.005804 **
## factor(Nitrate)2      0.79568    0.40895   1.946 0.056635 .
## factor(Minerals)1     0.85707    0.22449   3.818 0.000334 ***
```

```
## scale(pH)            -0.27256    0.14882  -1.832 0.072253 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5629 on 57 degrees of freedom
## Multiple R-squared:  0.7915, Adjusted R-squared:  0.7549
## F-statistic: 21.63 on 10 and 57 DF,  p-value: 6.173e-16
```

```
M3<- update(M2, .~.-scale(SpRich))
anova(M2, M3)
```

```
## Analysis of Variance Table
##
## Model 1: Harvest ~ scale(CWM.SLA) + scale(CWM.Seed.Mass) +
scale(SpRich) +
##     factor(Ammonium) + factor(Nitrate) + factor(Minerals) +
scale(pH)
## Model 2: Harvest ~ scale(CWM.SLA) + scale(CWM.Seed.Mass) +
factor(Ammonium) +
##     factor(Nitrate) + factor(Minerals) + scale(pH)
##   Res.Df    RSS Df  Sum of Sq      F Pr(>F)
## 1     57 18.061
## 2     58 18.064 -1 -0.0021003 0.0066 0.9354
```

By removing SpRich we have not made our model significantly worse and therefore its removal is statistically warranted. The change in RSS was negligible and no significant differences between the models were detected. M3, therefore, is our new model and we can examine its summary table to investigate whether other variables can be removed.

```
summary(M3) #the only variable we could remove is Ammonium so let's
try that anyway
```

```
##
## Call:
## lm(formula = Harvest ~ scale(CWM.SLA) + scale(CWM.Seed.Mass) +
##     factor(Ammonium) + factor(Nitrate) + factor(Minerals) +
scale(pH),
##     data = parkgrass)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.70078 -0.27801  0.01617  0.30127  1.16247
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)             2.5101     0.1648  15.227  < 2e-16 ***
## scale(CWM.SLA)         -0.2715     0.1071  -2.534 0.013989 *
## scale(CWM.Seed.Mass)    0.8113     0.1116   7.268 1.04e-09 ***
## factor(Ammonium)1       0.6456     0.3294   1.960 0.054856 .
## factor(Ammonium)2       0.6876     0.2230   3.083 0.003132 **
## factor(Ammonium)3       1.2876     0.2458   5.239 2.35e-06 ***
```

```
## factor(Nitrate)1       0.7032     0.2286   3.076 0.003194 **
## factor(Nitrate)2       0.8132     0.3446   2.360 0.021662 *
## factor(Minerals)1      0.8624     0.2130   4.049 0.000154 ***
## scale(pH)             -0.2798     0.1179  -2.373 0.020954 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5581 on 58 degrees of freedom
## Multiple R-squared:  0.7914, Adjusted R-squared:  0.7591
## F-statistic: 24.45 on 9 and 58 DF,  p-value: < 2.2e-16

M4<- update(M3, .~.-factor(Ammonium))
anova(M3, M4)

## Analysis of Variance Table
##
## Model 1: Harvest ~ scale(CWM.SLA) + scale(CWM.Seed.Mass) +
factor(Ammonium) +
##     factor(Nitrate) + factor(Minerals) + scale(pH)
## Model 2: Harvest ~ scale(CWM.SLA) + scale(CWM.Seed.Mass) +
factor(Nitrate) +
##     factor(Minerals) + scale(pH)
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1     58 18.064
## 2     61 27.211 -3   -9.1477 9.7907 2.545e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Removing Ammonium did make our model significantly worse and therefore M3 becomes our final model to interpretation. The RSS increased from 18.06 to 27.21 and this was statistically significant.

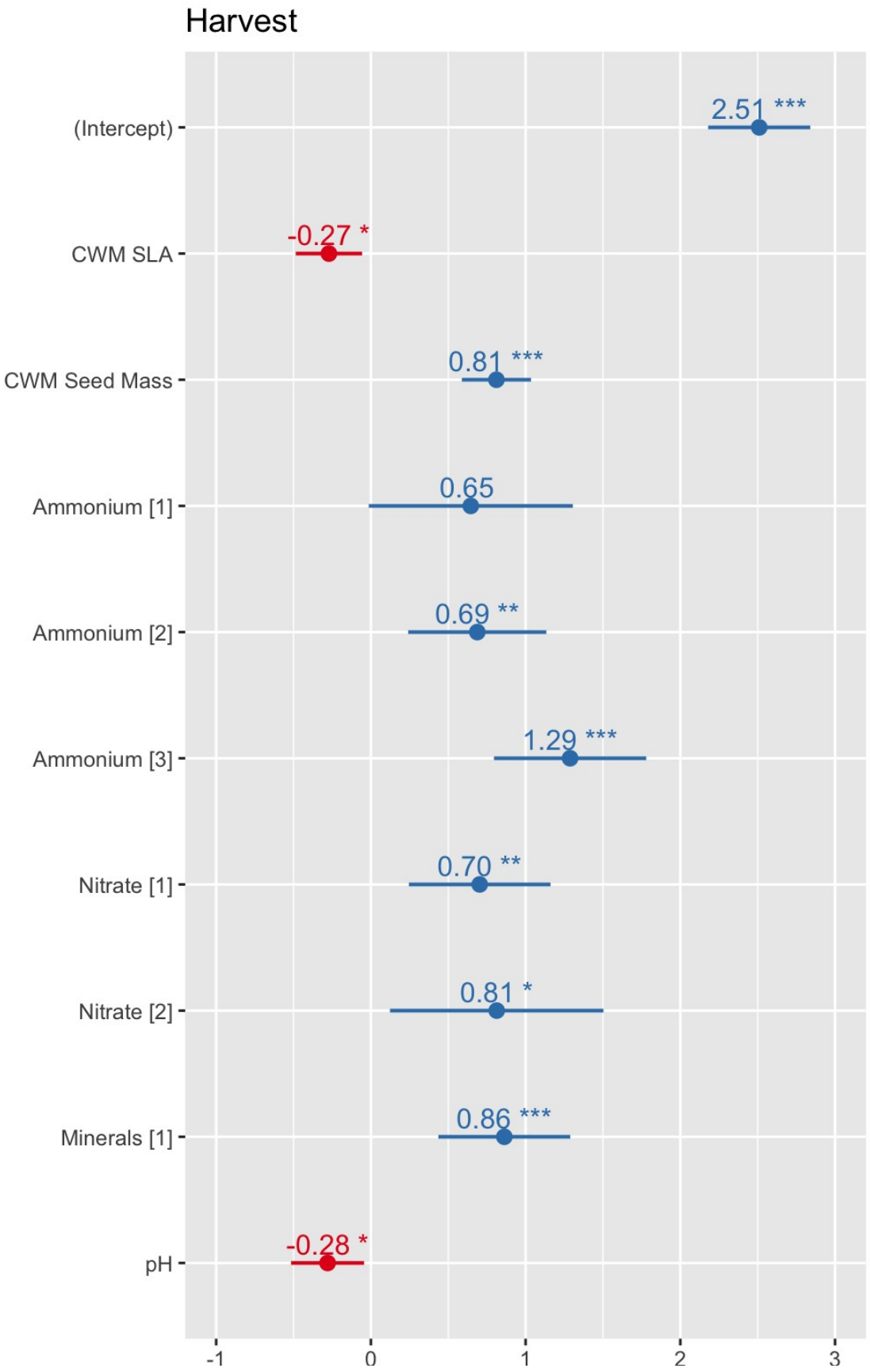
Model 3 Interpretation

```
summary(M3)

##
## Call:
## lm(formula = Harvest ~ scale(CWM.SLA) + scale(CWM.Seed.Mass) +
##     factor(Ammonium) + factor(Nitrate) + factor(Minerals) +
scale(pH),
##     data = parkgrass)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.70078 -0.27801  0.01617  0.30127  1.16247
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)             2.5101     0.1648  15.227  < 2e-16 ***
## scale(CWM.SLA)         -0.2715     0.1071  -2.534 0.013989 *
## scale(CWM.Seed.Mass)    0.8113     0.1116   7.268 1.04e-09 ***
```

```
## factor(Ammonium)1      0.6456     0.3294   1.960 0.054856 .
## factor(Ammonium)2      0.6876     0.2230   3.083 0.003132 **
## factor(Ammonium)3      1.2876     0.2458   5.239 2.35e-06 ***
## factor(Nitrate)1       0.7032     0.2286   3.076 0.003194 **
## factor(Nitrate)2       0.8132     0.3446   2.360 0.021662 *
## factor(Minerals)1      0.8624     0.2130   4.049 0.000154 ***
## scale(pH)             -0.2798     0.1179  -2.373 0.020954 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5581 on 58 degrees of freedom
## Multiple R-squared:  0.7914, Adjusted R-squared:  0.7591
## F-statistic: 24.45 on 9 and 58 DF,  p-value: < 2.2e-16

plot_model(M3, show.values = TRUE, show.intercept = TRUE)
```


Harvest
(Intercept)
CWM SLA
CWM Seed Mass
Ammonium [1]
Ammonium [2]
Ammonium [3]
Nitrate [1]
Nitrate [2]
Minerals [1]
pH
2.51 ***
-0.27 *
0.81 ***
0.65
0.69 **
1.29 ***
0.70 **
0.81 *
0.86 ***
-0.28 *
-1
0
1
2
3

I will leave the interpretation of this model to you. But you want to comment on:

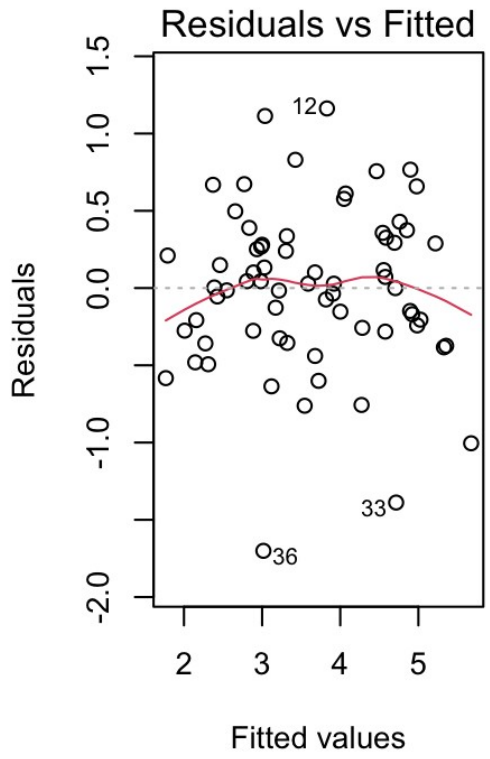
1. The effect sizes (slopes) of CWM.SLA, CWM.Seed.Mass and pH, how are these different to the 'Do Nothing' approach (see below table for help on this)?
2. What is the effect of ammonium addition on harvest? What are biomass means for each of the four conditions and how do they differ to the previous approach?
3. What is the effect of nitrate addition on harvest? What are biomass means for each of the four conditions and how do they differ to the previous approach?

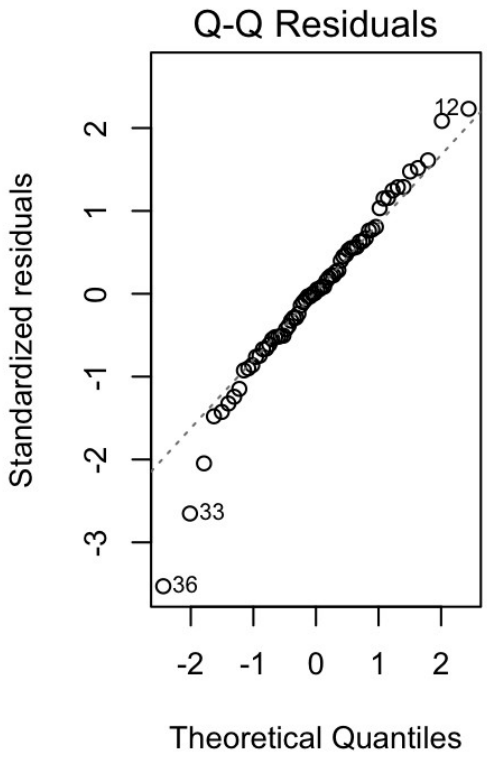
```
#tab_model(M1,M3) #the result of this code will appear in the "Viewer" where your plots are usually shown
```

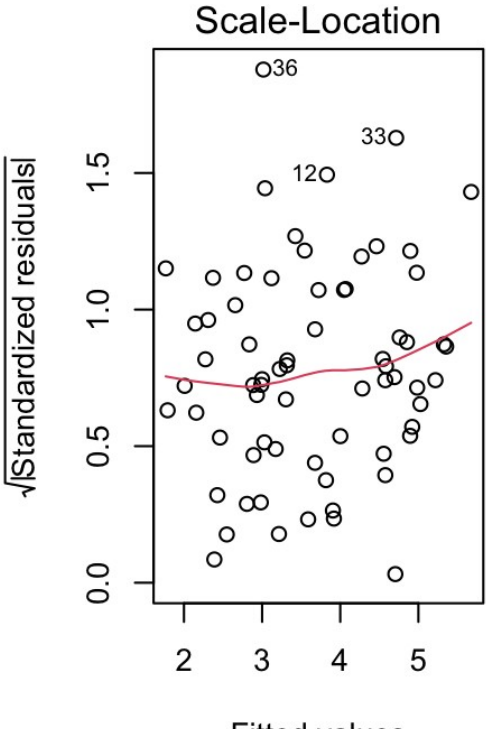
Model 3 Validation

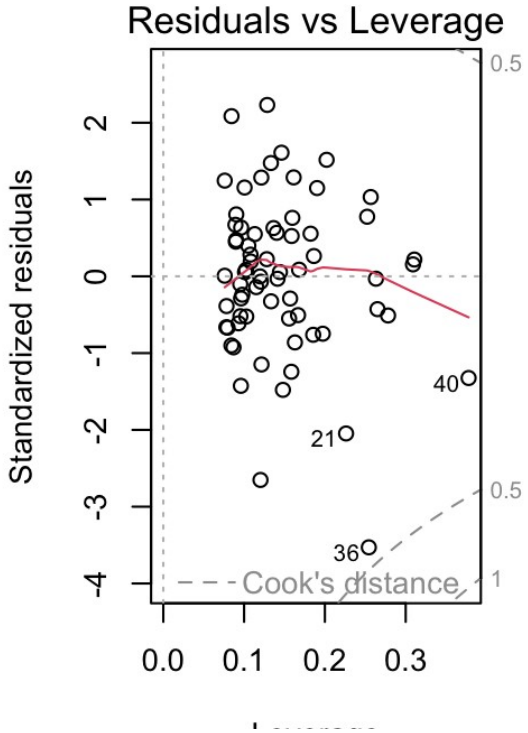
The final step is to validate our model by examining the diagnostic plots.

```
par(mfrow=c(2,2))
plot(M3)
```


Residuals vs Fitted
Residuals
Fitted values
12
33
36


Q-Q Residuals
Standardized residuals
Theoretical Quantiles
12
33
36


Scale-Location
√|Standardized residuals|
36
33
12


Residuals vs Leverage
Standardized residuals
40
21
36
Cook's distance
0.5
1

The Classic Approach - Information Criteria (AIC)

Unlike the hypothesis testing approach, model selection based on AIC can be automated in R using the *step* function.

```
M5<-step(M1, direction = "backward", scope = list(lower=~1,
upper=~scale(CWM.LDMC) + scale(CWM.SLA) + scale(CWM.Seed.Mass) +
scale(SpRich) + factor(Ammonium) +
    factor(Nitrate) + factor(Minerals) + scale(pH))) # scope here is
indicating the lower (null model) and upper (maximal model)

## Start:  AIC=-66.18
## Harvest ~ scale(CWM.LDMC) + scale(CWM.SLA) + scale(CWM.Seed.Mass) +
##     scale(SpRich) + factor(Ammonium) + factor(Nitrate) +
factor(Minerals) +
##     scale(pH)
##
##                         Df Sum of Sq    RSS     AIC
## - scale(SpRich)          1    0.0026 18.057 -68.167
## - scale(CWM.LDMC)        1    0.0071 18.061 -68.149
## <none>                               18.054 -66.176
## - scale(pH)              1    1.0700 19.124 -64.261
## - scale(CWM.SLA)         1    1.2606 19.315 -63.587
## - factor(Ammonium)       3    3.5715 21.626 -59.902
## - factor(Nitrate)        2    2.9455 21.000 -59.900
## - factor(Minerals)       1    4.6159 22.670 -52.695
## - scale(CWM.Seed.Mass)   1   14.5540 32.608 -27.976
##
## Step:  AIC=-68.17
## Harvest ~ scale(CWM.LDMC) + scale(CWM.SLA) + scale(CWM.Seed.Mass) +
##     factor(Ammonium) + factor(Nitrate) + factor(Minerals) +
scale(pH)
##
##                         Df Sum of Sq    RSS     AIC
## - scale(CWM.LDMC)        1    0.0067 18.064 -70.141
## <none>                               18.057 -68.167
## - scale(CWM.SLA)         1    1.3371 19.394 -65.309
## - scale(pH)              1    1.7518 19.809 -63.870
## - factor(Nitrate)        2    3.9013 21.958 -58.865
## - factor(Minerals)       1    5.1098 23.167 -53.222
## - factor(Ammonium)       3    8.5994 26.656 -47.681
## - scale(CWM.Seed.Mass)   1   14.7619 32.819 -29.538
##
## Step:  AIC=-70.14
## Harvest ~ scale(CWM.SLA) + scale(CWM.Seed.Mass) + factor(Ammonium)
+
##     factor(Nitrate) + factor(Minerals) + scale(pH)
##
##                         Df Sum of Sq    RSS     AIC
## <none>                               18.064 -70.141
## - scale(pH)              1    1.7545 19.818 -65.838
```

```
## - scale(CWM.SLA)         1    2.0003 20.064 -65.000
## - factor(Nitrate)        2    3.8952 21.959 -60.863
## - factor(Minerals)       1    5.1066 23.170 -55.212
## - factor(Ammonium)       3    9.1477 27.211 -48.280
## - scale(CWM.Seed.Mass)   1   16.4508 34.514 -28.113
```

We can see from the start that the maximal model has an AIC of -66.18 and a table that lists the removal of variables (these are ordered from lowest to highest). We can see that the model with SpRich removed has an AIC of -68.17, which is lower than the maximal model and therefore is removed in the first step. The results model has an AIC -68.17 and we can see the removal of CWM.LDMC will further improve the model and produce an AIC of -70.14. This produces our minimum adequate model and the coefficients are given in the final few lines. You'll notice this model is the same structure as the one determined through the hypothesis testing approach and therefore the interpretation would be the same.

The Information Theoretic (IT) Approach

In this scenario, the IT approach is rather preferable. Because we have a limited number of observations/data points, the IT approach can enable us to fit a range of models to test a range of plausible ecological hypotheses. We are should investigate these hypotheses using 10-15 models, here we are going to fit 10.

```
Model<- c("M1", "M2", "M3", "M4", "M5", "M6", "M7", "M8", "M9", "M10")
Code<- c("CWM.LDMC + CWM.SLA + CWM.Seed.Mass + SpRich +
factor(Ammonium) + factor(Nitrate) + factor(Minerals) + pH", "CWM.LDMC
+ CWM.SLA + CWM.Seed.Mass + factor(Ammonium) + factor(Nitrate) +
factor(Minerals) + pH", "SpRich + factor(Ammonium) + factor(Nitrate) +
factor(Minerals) + pH", "CWM.LDMC + CWM.SLA + factor(Ammonium) +
factor(Nitrate) + factor(Minerals) + pH", "CWM.Seed.Mass +
factor(Ammonium) + factor(Nitrate) + factor(Minerals) + pH", "SpRich +
factor(Ammonium) + factor(Nitrate) + factor(Minerals) + pH +
SpRich:factor(Ammonium) + SpRich:factor(Nitrate) +
SpRich:factor(Minerals)", "CWM.Seed.Mass + factor(Ammonium) +
factor(Nitrate) + factor(Minerals) + pH +
CWM.Seed.Mass:factor(Ammonium) + CWM.Seed.Mass:factor(Nitrate) +
CWM.Seed.Mass:factor(Minerals)","CWM.LDMC + factor(Ammonium) +
factor(Nitrate) + factor(Minerals) + pH + CWM.LDMC:factor(Ammonium) +
CWM.LDMC:factor(Nitrate) + CWM.LDMC:factor(Minerals)", "CWM.SLA +
factor(Ammonium) + factor(Nitrate) + factor(Minerals) + pH +
CWM.SLA:factor(Ammonium) + CWM.SLA:factor(Nitrate) +
CWM.SLA:factor(Minerals)", "factor(Ammonium) + factor(Nitrate) +
factor(Minerals) + pH + factor(Ammonium):factor(Minerals) +
factor(Nitrate):factor(Minerals) + factor(Ammonium):pH +
factor(Nitrate):pH + factor(Minerals):pH")
Description <- c("All variables model", "Plant traits and the
environment", "Species richness and the environment", "Leaf traits and
the environment", "Size traits and the environment", "Species richness
and the environment with selected interactions", "Seed mass and the
environment with selected interactions", "Leaf dry matter content and
the environment with selected interactions", "Specific leaf area and
```

```
the environment with selected interactions", "Only environmental
measures with selected interactions")
table<- data.frame(Model=Model, Code=Code, Description=Description)
knitr::kable(table, caption = "Models used for the IT Approach")
```

Models used for the IT Approach

Code	Description
CWM.LDMC + CWM.SLA + CWM.Seed.Mass + SpRich + factor(Ammonium) + factor(Nitrate) + factor(Minerals) + pH	All variables model
CWM.LDMC + CWM.SLA + CWM.Seed.Mass + factor(Ammonium) + factor(Nitrate) + factor(Minerals) + pH	Plant traits and the environment
SpRich + factor(Ammonium) + factor(Nitrate) + factor(Minerals) + pH	Species richness and the environment
CWM.LDMC + CWM.SLA + factor(Ammonium) + factor(Nitrate) + factor(Minerals) + pH	Leaf traits and the environment
CWM.Seed.Mass + factor(Ammonium) + factor(Nitrate) + factor(Minerals) + pH	Size traits and the environment
SpRich + factor(Ammonium) + factor(Nitrate) + factor(Minerals) + pH + SpRich:factor(Ammonium) + SpRich:factor(Nitrate) + SpRich:factor(Minerals)	Species richness and the environment with selected interactions
CWM.Seed.Mass + factor(Ammonium) + factor(Nitrate) + factor(Minerals) + pH + CWM.Seed.Mass:factor(Ammonium) + CWM.Seed.Mass:factor(Nitrate) + CWM.Seed.Mass:factor(Minerals)	Seed mass and the environment with selected interactions
CWM.LDMC + factor(Ammonium) + factor(Nitrate) + factor(Minerals) + pH + CWM.LDMC:factor(Ammonium) + CWM.LDMC:factor(Nitrate) + CWM.LDMC:factor(Minerals)	Leaf dry matter content and the environment with selected interactions
CWM.SLA + factor(Ammonium) + factor(Nitrate) + factor(Minerals) + pH + CWM.SLA:factor(Ammonium) + CWM.SLA:factor(Nitrate) + CWM.SLA:factor(Minerals)	Specific leaf area and the environment with selected interactions
factor(Ammonium) + factor(Nitrate) + factor(Minerals) + pH + factor(Ammonium):factor(Minerals) + factor(Nitrate):factor(Minerals) + factor(Ammonium):pH + factor(Nitrate):pH + factor(Minerals):pH	Only environmental measures with selected interactions

Model Fitting

Now we need to fit these ten models individual:

```
M1_IT<- lm(Harvest~scale(CWM.LDMC) + scale(CWM.SLA) +
scale(CWM.Seed.Mass) + scale(SpRich) + factor(Ammonium) +
factor(Nitrate) + factor(Minerals) + scale(pH), data = parkgrass)

M2_IT<- lm(Harvest~scale(CWM.LDMC) + scale(CWM.SLA) +
```

```
scale(CWM.Seed.Mass) + factor(Ammonium) + factor(Nitrate) +
factor(Minerals) + scale(pH), data = parkgrass)

M3_IT<- lm(Harvest~scale(SpRich) + factor(Ammonium) + factor(Nitrate)
+ factor(Minerals) + scale(pH), data = parkgrass)

M4_IT<- lm(Harvest~scale(CWM.LDMC) + scale(CWM.SLA) + factor(Ammonium)
+ factor(Nitrate) + factor(Minerals) + scale(pH), data = parkgrass)

M5_IT<- lm(Harvest~scale(CWM.Seed.Mass) + factor(Ammonium) +
factor(Nitrate) + factor(Minerals) + scale(pH), data = parkgrass)

M6_IT<- lm(Harvest~SpRich + factor(Ammonium) + factor(Nitrate) +
factor(Minerals) + scale(pH) + scale(SpRich):factor(Ammonium) +
scale(SpRich):factor(Nitrate) + scale(SpRich):factor(Minerals), data =
parkgrass)

M7_IT<- lm(Harvest~scale(CWM.Seed.Mass) + factor(Ammonium) +
factor(Nitrate) + factor(Minerals) + scale(pH) +
scale(CWM.Seed.Mass):factor(Ammonium) +
scale(CWM.Seed.Mass):factor(Nitrate) +
scale(CWM.Seed.Mass):factor(Minerals), data = parkgrass)

M8_IT<- lm(Harvest~scale(CWM.LDMC) + factor(Ammonium) +
factor(Nitrate) + factor(Minerals) + scale(pH) +
scale(CWM.LDMC):factor(Ammonium) + scale(CWM.LDMC):factor(Nitrate) +
scale(CWM.LDMC):factor(Minerals), data = parkgrass)

M9_IT<- lm(Harvest~scale(CWM.SLA) + factor(Ammonium) + factor(Nitrate)
+ factor(Minerals) + scale(pH) + scale(CWM.SLA):factor(Ammonium) +
scale(CWM.SLA):factor(Nitrate) + scale(CWM.SLA):factor(Minerals), data
= parkgrass)

M10_IT<- lm(Harvest~factor(Ammonium) + factor(Nitrate) +
factor(Minerals) + scale(pH) + factor(Ammonium):factor(Minerals) +
factor(Nitrate):factor(Minerals) + factor(Ammonium):scale(pH) +
factor(Nitrate):scale(pH) + factor(Minerals):scale(pH), data =
parkgrass)
```

Ok so now we have fitted these let's produce a table that will illustrate the AICs and Akaike Weights of each model.

```
AICs<- AIC(M1_IT, M2_IT, M3_IT, M4_IT, M5_IT, M6_IT, M7_IT, M8_IT,
M9_IT, M10_IT)
DoF<- AICs[,1]
AICsNum<- AICs[,2]
minAW<- min(AICsNum) # finding the minimum AIC
Delta <- AICsNum-minAW # finding the difference between each AIC and
the minimum AIC
RL <- exp(-0.5*Delta) # finding the relative likelihood of each
```

```
AIC/model
wi <- RL/sum(RL) # weighting these relative likelihoods
AkaikeWeights_table<- data.frame(Model=1:10,DoF=DoF,
AIC=round(AICsNum, digits = 2), AICDifferences= round(Delta, digits =
2), AkaikeWeights=round(wi, digits=2))
AkaikeWeights_table
```

```
##    Model DoF    AIC AICDifferences AkaikeWeights
## 1      1  13 128.80           3.06          0.11
## 2      2  12 126.81           1.06          0.31
## 3      3  10 166.24          40.50          0.00
## 4      4  11 165.44          39.69          0.00
## 5      5  10 129.98           4.23          0.06
## 6      6  16 168.97          43.23          0.00
## 7      7  16 125.74           0.00          0.52
## 8      8  16 165.08          39.34          0.00
## 9      9  16 168.14          42.39          0.00
## 10    10  16 176.84          51.09          0.00
```

Model 1, 2 5 and 7 have the highest Akaike Weights. The interpretation of these weights is as follows: "If we were to repeat the experiment a large number of times in 52% of cases model 7 would be the optimal model, and in 31% of the cases it would be model 2". Instead of presenting only the results of the best model, the IT approach advocates presenting and discussing the results of all the 'better' models (Akaike Weights that round to above 0). In this case, we may choose to present all model 1, 2, 5 and 7 to capture the 100% likelihood or we may choose to present 1, 2 and 7 to capture 94%. This is fairly subjective decision but we need to be mindful of the total likelihood of the models we present. Here, i'd present and interpret models 1, 2 and 7.

Models 1, 2 and 7 Interpretation

```
summary(M1_IT)
```

```
##
## Call:
## lm(formula = Harvest ~ scale(CWM.LDMC) + scale(CWM.SLA) +
scale(CWM.Seed.Mass) +
##     scale(SpRich) + factor(Ammonium) + factor(Nitrate) +
factor(Minerals) +
##     scale(pH), data = parkgrass)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.70669 -0.27579  0.01743  0.30602  1.12966
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)            2.52318    0.27151   9.293 6.08e-13 ***
## scale(CWM.LDMC)       -0.01800    0.12105  -0.149 0.882358
## scale(CWM.SLA)        -0.26508    0.13406  -1.977 0.052933 .
## scale(CWM.Seed.Mass)   0.81608    0.12146   6.719 1.01e-08 ***
## scale(SpRich)         -0.02182    0.24521  -0.089 0.929411
```

```
## factor(Ammonium)1     0.62913    0.43211   1.456 0.150986
## factor(Ammonium)2     0.67383    0.45097   1.494 0.140742
## factor(Ammonium)3     1.26420    0.43045   2.937 0.004803 **
## factor(Nitrate)1      0.69639    0.24528   2.839 0.006292 **
## factor(Nitrate)2      0.79923    0.41320   1.934 0.058141 .
## factor(Minerals)1     0.85684    0.22645   3.784 0.000378 ***
## scale(pH)            -0.27422    0.15053  -1.822 0.073836 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5678 on 56 degrees of freedom
## Multiple R-squared:  0.7915, Adjusted R-squared:  0.7506
## F-statistic: 19.33 on 11 and 56 DF,  p-value: 2.941e-15

summary(M2_IT)

##
## Call:
## lm(formula = Harvest ~ scale(CWM.LDMC) + scale(CWM.SLA) +
scale(CWM.Seed.Mass) +
##     factor(Ammonium) + factor(Nitrate) + factor(Minerals) +
scale(pH),
##     data = parkgrass)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.70518 -0.27759  0.02324  0.30166  1.13931
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)           2.50450    0.17070  14.672  < 2e-16 ***
## scale(CWM.LDMC)      -0.01739    0.11980  -0.145 0.885122
## scale(CWM.SLA)       -0.26168    0.12737  -2.054 0.044529 *
## scale(CWM.Seed.Mass)  0.81723    0.11972   6.826 6.19e-09 ***
## factor(Ammonium)1     0.65297    0.33615   1.943 0.057022 .
## factor(Ammonium)2     0.70648    0.25983   2.719 0.008661 **
## factor(Ammonium)3     1.29503    0.25318   5.115 3.83e-06 ***
## factor(Nitrate)1      0.70332    0.23054   3.051 0.003462 **
## factor(Nitrate)2      0.81842    0.34937   2.343 0.022669 *
## factor(Minerals)1     0.86269    0.21480   4.016 0.000175 ***
## scale(pH)            -0.28218    0.11999  -2.352 0.022172 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5628 on 57 degrees of freedom
## Multiple R-squared:  0.7915, Adjusted R-squared:  0.7549
## F-statistic: 21.64 on 10 and 57 DF,  p-value: 6.13e-16

summary(M7_IT)
```

```
##
## Call:
## lm(formula = Harvest ~ scale(CWM.Seed.Mass) + factor(Ammonium) +
##     factor(Nitrate) + factor(Minerals) + scale(pH) +
scale(CWM.Seed.Mass):factor(Ammonium) +
##     scale(CWM.Seed.Mass):factor(Nitrate) +
scale(CWM.Seed.Mass):factor(Minerals),
##     data = parkgrass)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.90125 -0.22457  0.03016  0.25485  1.46400
##
## Coefficients:
##                                      Estimate Std. Error t value
Pr(>|t|)
## (Intercept)                            2.7910     0.2184  12.782
< 2e-16 ***
## scale(CWM.Seed.Mass)                   1.3605     0.4319   3.150
0.00268 **
## factor(Ammonium)1                      1.6179     0.6240   2.593
0.01228 *
## factor(Ammonium)2                      0.7034     0.2329   3.020
0.00388 **
## factor(Ammonium)3                      0.9030     0.2110   4.280
7.88e-05 ***
## factor(Nitrate)1                       0.8009     0.2388   3.353
0.00148 **
## factor(Nitrate)2                      -0.6923     2.8796  -0.240
0.81092
## factor(Minerals)1                      0.7689     0.2633   2.920
0.00513 **
## scale(pH)                             -0.2763     0.1070  -2.581
0.01265 *
## scale(CWM.Seed.Mass):factor(Ammonium)1   1.0709     0.7394   1.448
0.15339
## scale(CWM.Seed.Mass):factor(Ammonium)2   0.5351     0.2466   2.170
0.03448 *
## scale(CWM.Seed.Mass):factor(Ammonium)3   0.4438     0.2160   2.054
0.04490 *
## scale(CWM.Seed.Mass):factor(Nitrate)1   -0.2866     0.2625  -1.092
0.27983
## scale(CWM.Seed.Mass):factor(Nitrate)2    0.9657     2.2040   0.438
0.66306
## scale(CWM.Seed.Mass):factor(Minerals)1  -0.7283     0.4266  -1.707
0.09365 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5461 on 53 degrees of freedom
```

```
## Multiple R-squared:  0.8175, Adjusted R-squared:  0.7693
## F-statistic: 16.96 on 14 and 53 DF,  p-value: 8.301e-15

#tab_model(M1_IT, M2_IT, M7_IT) #puts all three models in one table
```

Again, i'll leave the interpretation of these models to you. However, there are some important things to highlight (especially in the commonalities between these models), think about these questions:

- How is the effect of CWM.Seed.Mass different in between these models?
- How is the effect of Nitrate and Ammonium different between these models?
- How is the effect of pH different between these models?

Concluding Remarks

We have implemented different approaches to model selection that have provided different results and interpretations of the same initial research question. It is crucial to be mindful of these different approaches and understand their flexibility, especially with the IT approach. We can follow the same procedures with mixed models and generalised linear models. Recent developments in the IT approach has delved into model averaging - essentially averaging the coefficients of models (come and speak to me if you're interested in knowing more about this) or see the papers below:

- Multimodal inference in ecology and evolution: challenges and solutions (2011).
- Model averaging in ecology: a review of Bayesian, information-theoretic, and tactical approaches for predictive inference (2018).

Extra Tasks

Rather than providing you with random datasets, I want you to pick a dataset from a project you have worked on in the past. With model selection and model specification, it is always easier to use a dataset in an area of ecology you are more accustomed to. Practice going through the model selection procedures with another dataset of your choosing and consider how different are the final model(s) as suggested by the approaches.