Data Analysis 2

Maksuda Aktar Toma, Jo Charbonneau, Ryan Lalicker

October 30, 2024

If we want an abstract it will go here. References are in the form Astley (1987) or (Astley 1987). For more information see here.

Introduction

Our clients conducted an experiment to determine the effect pine tissues, precipitation levels, time, and the interaction of these variables effects starch content. In total, 408 entries were recorded. The experiment was replicated at two locations as well and not all measurements within each replication were taken from the same sample location. (dont like that last line)

We intend to analysis the results of this data below. We will review the variables, fit multiple models, and make a suggestion to the client. The data set, `data.csv`, and all other files used in this project can be found on our Github page.

Exploring the Data

Variables

In the data set provided by the client there are four tissue types which are abbreviated as END, IT, LM, and UM. This can be found in the `tissu` column. The two precipitation levels, control and drought, are in the `treatment` column. The time component of the experiment is not simply one variable. The `time` column consists of six different times, with six being denoted by the first six letters of the alphabet. In addition to `time`, the column `dayPeriod` indicates whether the measurement was taken in the day or at night. Time points C and D appear to correspond to a `dayPeriod` of night, while all other time points are during the day. Note, the measurements for the starch contents can be found in the `StarchNscTissue` and each sample number can be found in the `sample` column.

The data set provided by the client also includes variables that indicate the physical location of where the measurement was taken within a sample. These are represented the columns `row`, `col`, and `chamber` with the latter being in the form `row-col` for each respective entry. The possible values of `row` and `col` range from one to four. Also, since the experiment was carried out at two locations which is represented by the `campagne` column.

Changes made to the variables in the original data set

Note there were a couple of problems with the original data set. Initially the `time` column included a seventh time, A’. Since this did not follow the format of the other time points and had substantially fewer occurrences in the data, we assumed this was a mistake. Therefore, we manually changed all occurrences of A’ to A.

The other potential issue was in the `chamber` column. As stated above this column should be a combination of `row` and `col`, but the original data set was treating it as a date. For example if one sample has the values `row` $= 1$ and `col` $= 4$, the result of `chamber` should be $1 - 4$. Instead the original data set was showing January 4th. We chose to manually change this to the correct format as well.

Summary Statistics

While some of the variables outlined above are numeric, most can be treated as categorical. The lone exception to this is the starch content. The table below shows some summary statistics for the starch content. This includes not only the summaries of all 408 measurements, but also the summaries based on the two values of `campagne` and `dayPeriod`.

Group	N	Mean	Median	SD	Min	Max
Overall	408	1.924902	1.429527	1.733284	0.0191182	7.898429
campagne: 1	184	1.340544	1.245685	1.008316	0.0191182	6.480553
campagne: 2	224	2.404911	1.677605	2.033619	0.2029488	7.898429
dayPeriod: Day	280	1.895429	1.357646	1.730086	0.0191182	7.898429
dayPeriod: Night	128	1.989375	1.483575	1.745326	0.0656625	7.537576

Figure 1: Summary statistics of starch content.

For starch contents across all measurements, the values range from about 0.019 to 7.898 with a median of roughly 1.430 and a mean of 1.925. The location of the median and mean with respect to the minimum and maximum is an early sign that the starch contents could be skewed and thus non-normal in distribution.

When comparing the two locations (`campagne`) where the experiment was replicated, we can see the 184 measurements from the first location seems to have lower values on average than the 224 measurements from location 2. There is a smaller difference in these metrics when comparing measurements taken in the day versus those taken in the night. Note over twice as many measurements were taken in the day.

To generate a table of summary statistics that account for more of the variables see *Appendix A - R Code*. That table is not included here due to its larger size.

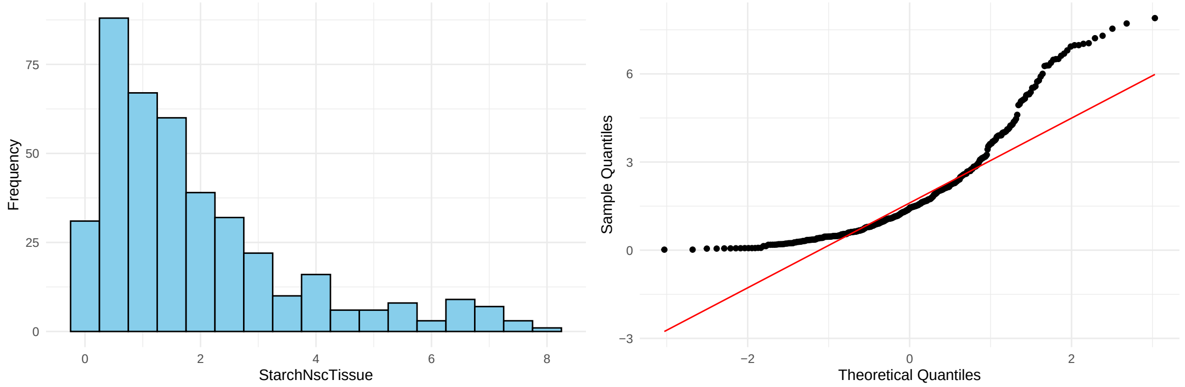
Relationships among variables

Potential models

The replication mentioned above suggests a mixed model approach is needed. This is due to the replication being a random effect. The simplest case of a this type of model is a linear mixed model. To use this, the starch measurements, which will be the dependent variable in whatever model we choose, must be approximately normally distributed.

Normality of Starch Content

As discussed in the *Summary Statistics* section, we suspect the starch content variable may be non-normal. There are a few ways to check this. One way is visually by using both a histogram and a Q-Q plot. For a histogram, shown in the leftward plot below, we would expect a symmetrical bell shape if the data is from a normal distribution. For the Q-Q plot, the second plot below, the points should fall along a straight line, which is indicated by the red line in the plot.



(a) Histogram of starch content.

(b) Q-Q plot of starch content.

Figure 2: Plots used to check normallity assumption.

As we can see, the normality assumption is not holding in either case. The histogram has a heavy right-skew and the points on the Q-Q plot do not follow a straight line.

To verify this result we can perform the Shapiro-Wilk test. (Kassambara (2024)). The null hypothesis of the test is that the data is normally distributed. However since the test returns a p-value of $2.2 * 10^{-16}$ we can safely reject the null hypothesis since this is well below any significance level (α) commonly used.

Since the normality assumption does not hold, we must consider a generalized linear mixed model which can work with non-normal dependent variables.

How explanatory variables can be used

(talk about nesting vs non-nesting methods I guess. Just introduce the idea before we actually make the models.)

Summary Statistics

Summary__Statistic

```
# A tibble: 48 x 10
# Groups:   tissu, treatment, dayPeriod [16]
   tissu treatment dayPeriod time  mean_Starch sd_Starch median_Starch
   <chr> <chr>     <chr>     <chr>       <dbl>     <dbl>         <dbl>
 1 END   Control   Day       A           0.800     0.427         0.766
 2 END   Control   Day       B           0.806     0.468         0.965
 3 END   Control   Day       E           0.736     0.598         0.562
 4 END   Control   Day       F           0.740     0.178         0.765
 5 END   Control   Night     C           0.824     0.478         1.03
 6 END   Control   Night     D           0.700     0.381         0.714
 7 END   Drought   Day       A           0.507     0.110         0.481
 8 END   Drought   Day       B           0.870     0.468         0.622
 9 END   Drought   Day       E           0.765     0.408         0.687
10 END   Drought   Day       F           0.449     0.157         0.479
# i 38 more rows
# i 3 more variables: min_Starch <dbl>, max_Starch <dbl>, n <int>
```

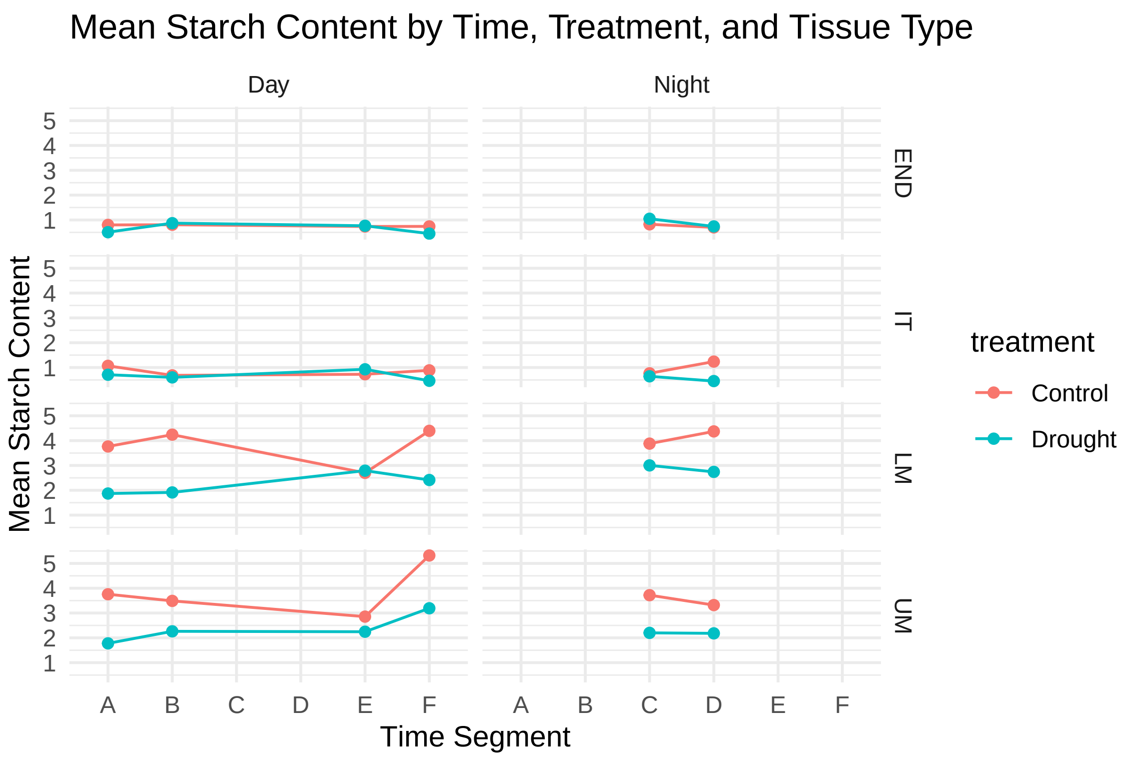

Mean Starch Content by Time, Treatment, and Tissue Type
Day
Night
END
IT
LM
UM
Mean Starch Content
1
2
3
4
5
A
B
C
D
E
F
Time Segment
treatment
Control
Drought

Figure 3: jjj

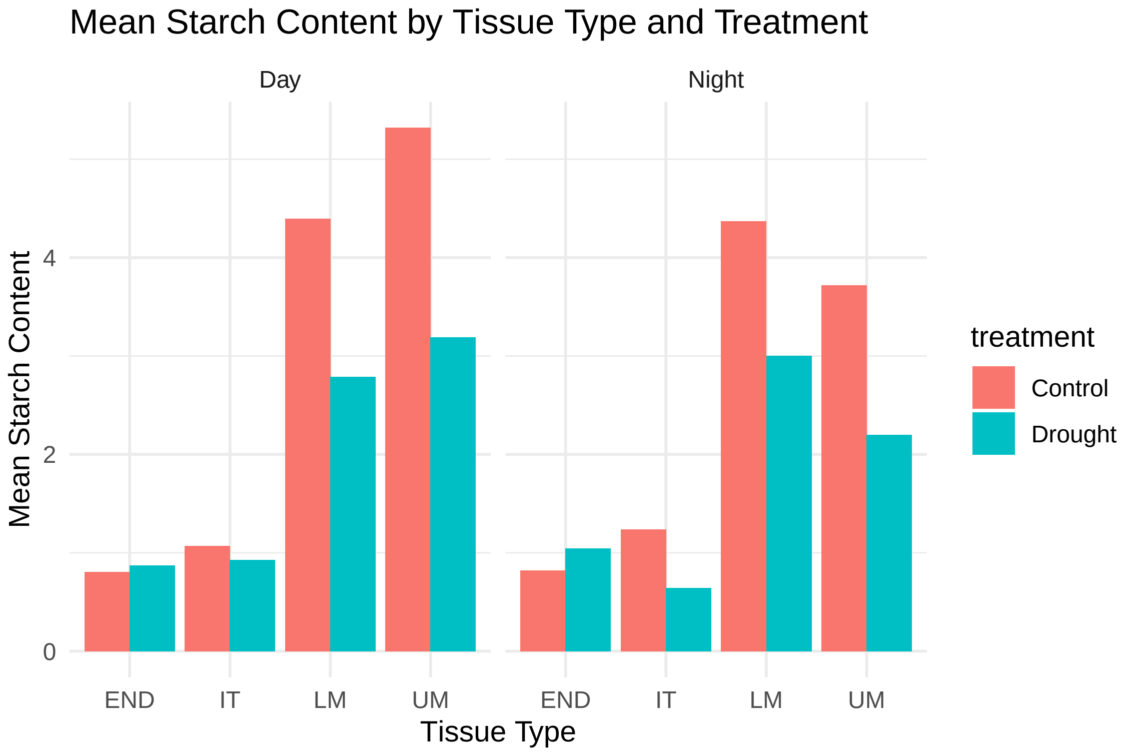

Mean Starch Content by Tissue Type and Treatment
Day
Night
Mean Starch Content
0
2
4
END
IT
LM
UM
END
IT
LM
UM
Tissue Type
treatment
Control
Drought

Figure 4: jjj

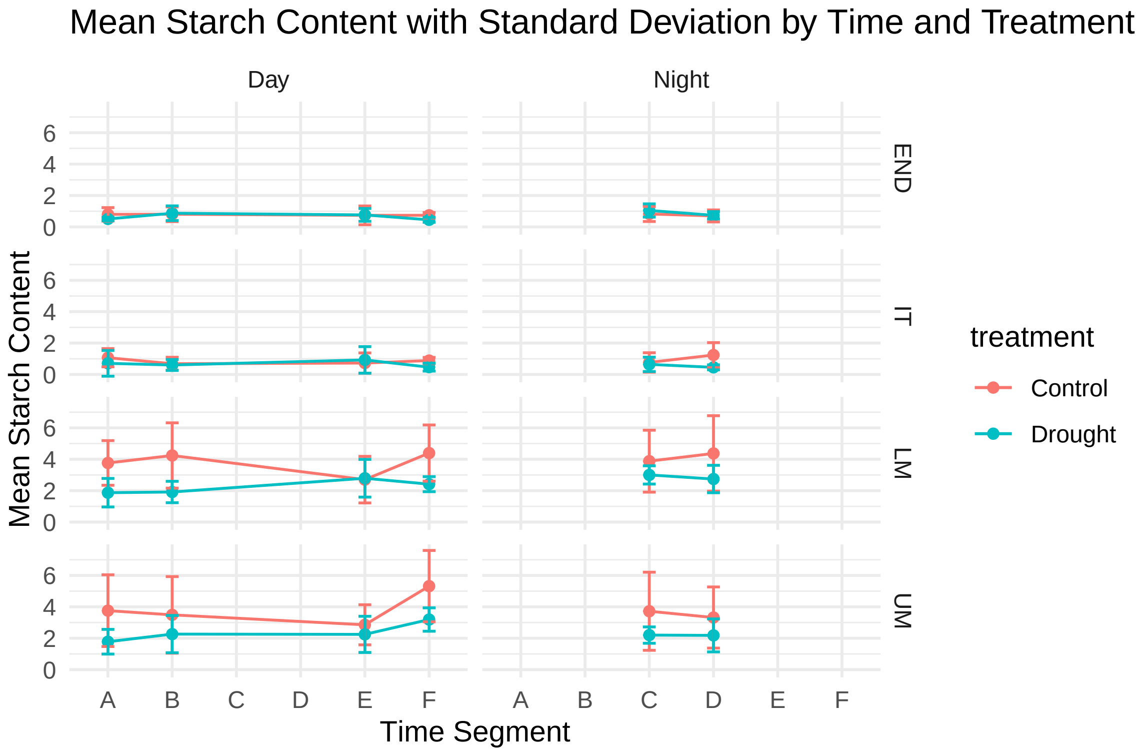

Mean Starch Content with Standard Deviation by Time and Treatment
Day
Night
END
IT
LM
UM
Mean Starch Content
0
2
4
6
A
B
C
D
E
F
Time Segment
treatment
Control
Drought

Figure 5: jjj

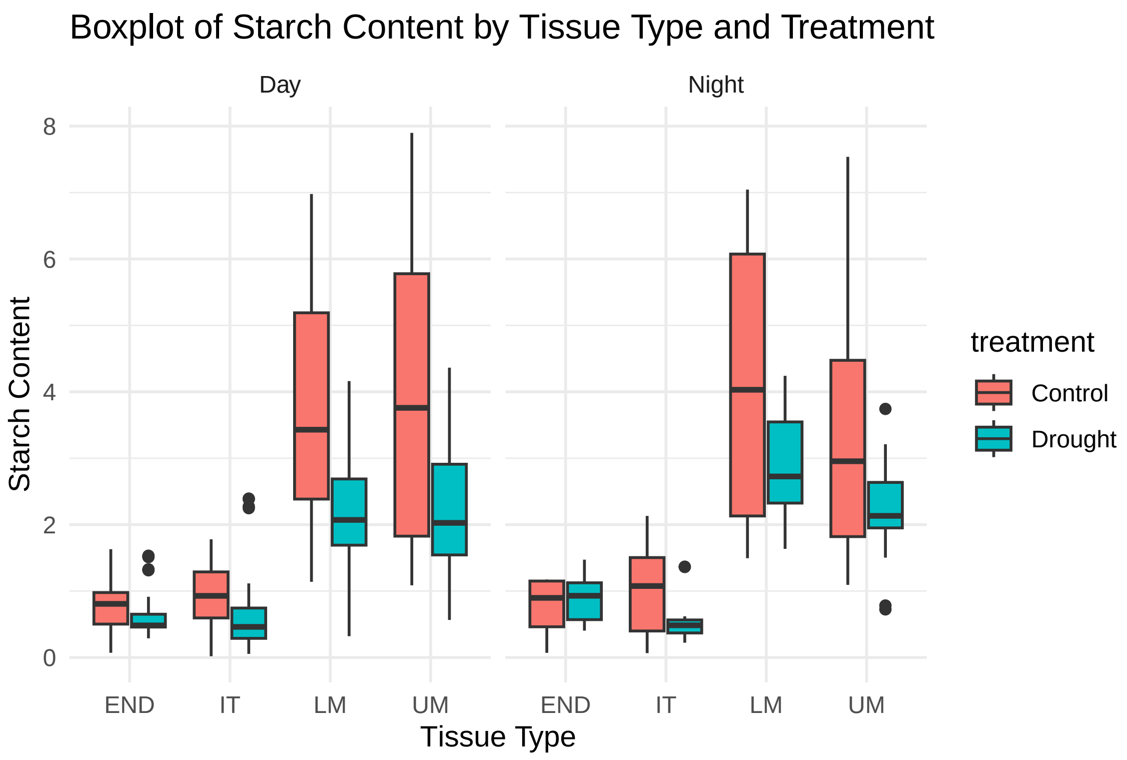

Boxplot of Starch Content by Tissue Type and Treatment
Day
Night
Starch Content
0
2
4
6
8
END
IT
LM
UM
END
IT
LM
UM
Tissue Type
treatment
Control
Drought

Figure 6: jjj

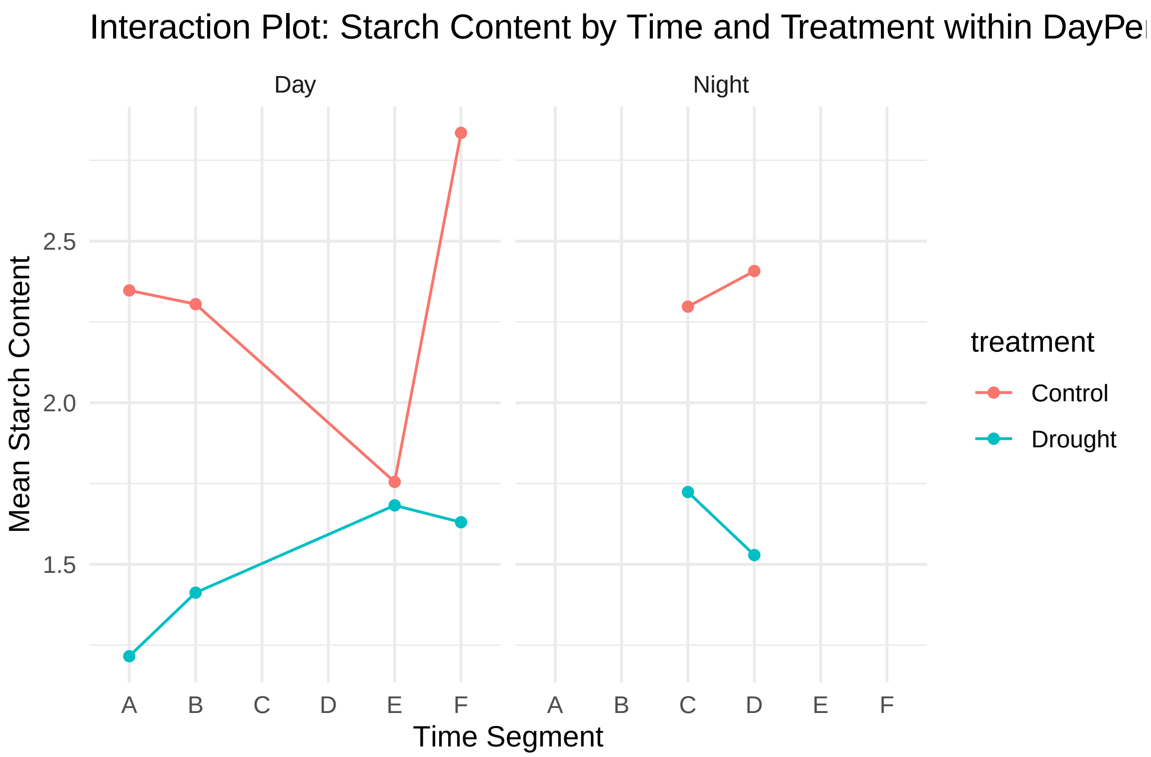

Interaction Plot: Starch Content by Time and Treatment within DayPe
Day
Night
Mean Starch Content
2.5
2.0
1.5
A
B
C
D
E
F
A
B
C
D
E
F
Time Segment
treatment
Control
Drought

Figure 7: jjj

Model : Mixed Effects Model with Interactions In this model, we include interactions between tissu, treatment, and dayPeriod to evaluate their combined effects on StarchNscTissue.

```
Linear mixed model fit by REML. t-tests use Satterthwaite's method [
lmerModLmerTest]
Formula: StarchNscTissue ~ tissu * treatment * dayPeriod + (1 | campagne) +
    (1 | sample) + (1 | chamber)
   Data: data

REML criterion at convergence: 1151.9

Scaled residuals:
    Min      1Q  Median      3Q     Max
-2.7075 -0.5735 -0.1460  0.5321  3.9737

Random effects:
 Groups   Name        Variance Std.Dev.
 chamber  (Intercept) 0.03069  0.1752
 sample   (Intercept) 0.21782  0.4667
 campagne (Intercept) 0.52077  0.7216
 Residual             0.92767  0.9632
Number of obs: 408, groups:  chamber, 8; sample, 8; campagne, 2

Fixed effects:
                                         Estimate Std. Error        df t value
(Intercept)                               0.67729    0.59020   1.40265   1.148
tissuIT                                   0.09316    0.22702 386.01503   0.410
tissuLM                                   3.00031    0.22702 386.01503  13.216
tissuUM                                   3.07044    0.22702 386.01503  13.525
treatmentDrought                         -0.09047    0.42136   8.85231  -0.215
dayPeriodNight                            0.08438    0.28949 386.03418   0.291
tissuIT:treatmentDrought                 -0.03511    0.32574 386.01503  -0.108
tissuLM:treatmentDrought                 -1.42355    0.32574 386.01503  -4.370
tissuUM:treatmentDrought                 -1.42214    0.32574 386.01503  -4.366
tissuIT:dayPeriodNight                    0.14820    0.40926 386.01503   0.362
tissuLM:dayPeriodNight                    0.36356    0.40926 386.01503   0.888
tissuUM:dayPeriodNight                   -0.31235    0.40926 386.01503  -0.763
treatmentDrought:dayPeriodNight           0.22150    0.41139 386.09586   0.538
tissuIT:treatmentDrought:dayPeriodNight  -0.54948    0.58140 386.01503  -0.945
tissuLM:treatmentDrought:dayPeriodNight   0.03887    0.58140 386.01503   0.067
tissuUM:treatmentDrought:dayPeriodNight  -0.03772    0.58140 386.01503  -0.065
```

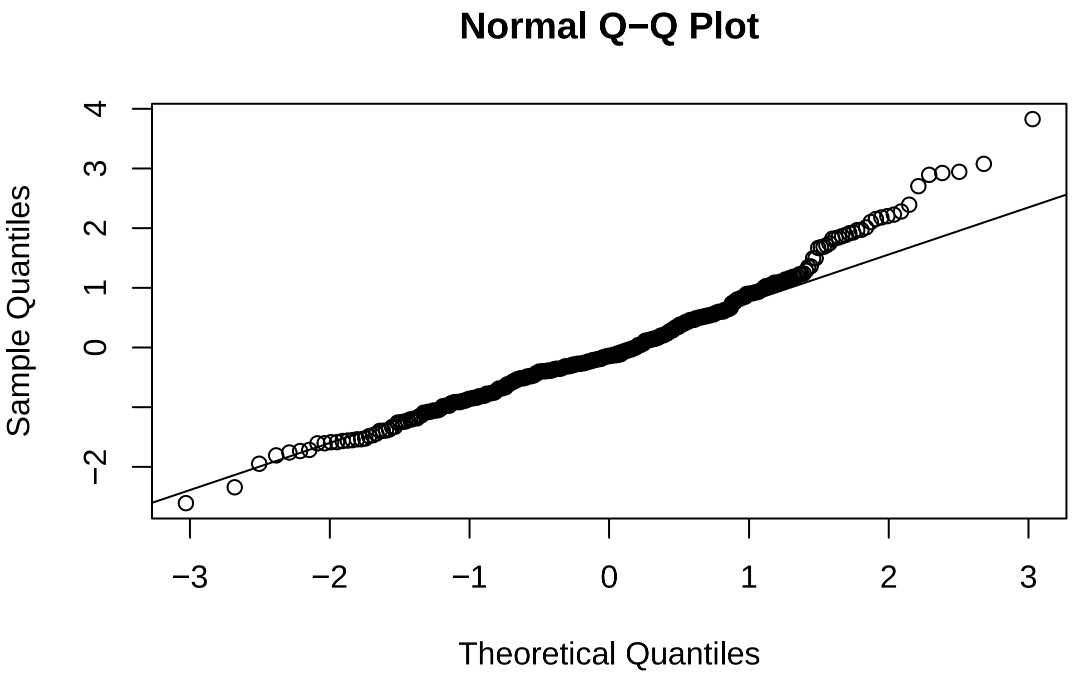
```
                                        Pr(>|t|)
(Intercept)                                0.410
tissuIT                                    0.682
tissuLM                                  < 2e-16 ***
tissuUM                                  < 2e-16 ***
treatmentDrought                           0.835
dayPeriodNight                             0.771
tissuIT:treatmentDrought                   0.914
tissuLM:treatmentDrought                1.60e-05 ***
tissuUM:treatmentDrought                1.63e-05 ***
tissuIT:dayPeriodNight                     0.717
tissuLM:dayPeriodNight                     0.375
tissuUM:dayPeriodNight                     0.446
treatmentDrought:dayPeriodNight            0.591
tissuIT:treatmentDrought:dayPeriodNight    0.345
tissuLM:treatmentDrought:dayPeriodNight    0.947
tissuUM:treatmentDrought:dayPeriodNight    0.948
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
optimizer (nloptwrap) convergence code: 0 (OK)
unable to evaluate scaled gradient
Model failed to converge: degenerate  Hessian with 1 negative eigenvalues
```

```
[1] 1191.869
```

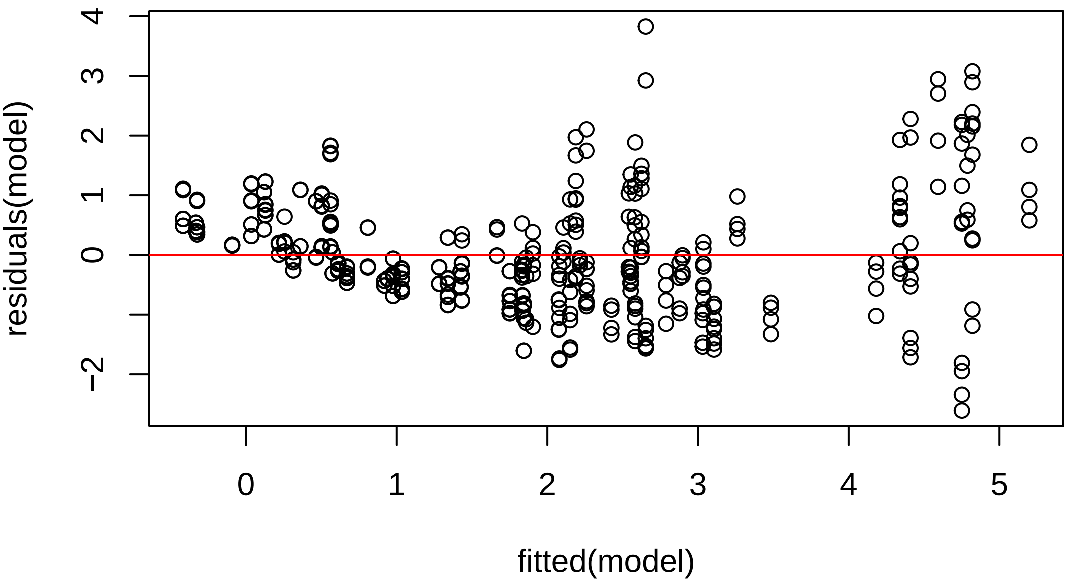
```
[1] 1272.094
```

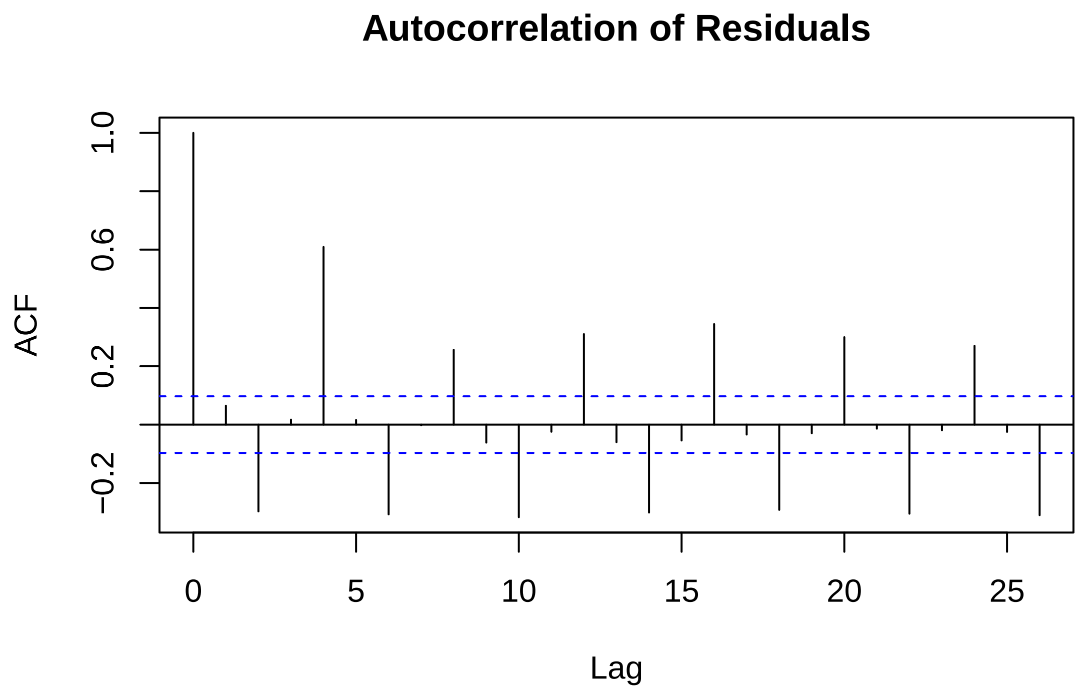
```
Type III Analysis of Variance Table with Satterthwaite's method
                        Sum Sq Mean Sq NumDF  DenDF  F value    Pr(>F)
tissu                   480.67 160.224     3 386.02 172.7160 < 2.2e-16 ***
treatment                 4.06   4.062     1   5.12   4.3784   0.08934 .
dayPeriod                 2.72   2.718     1 386.13   2.9302   0.08774 .
tissu:treatment          36.35  12.116     3 386.02  13.0608 3.848e-08 ***
tissu:dayPeriod           5.95   1.983     3 386.02   2.1380   0.09496 .
treatment:dayPeriod       0.16   0.156     1 386.33   0.1677   0.68235
tissu:treatment:dayPeriod   1.26   0.420     3 386.02   0.4530   0.71531
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Checking assumption



```
    Shapiro-Wilk normality test

data:  residuals(model)
W = 0.97114, p-value = 3.211e-07
```


4
3
2
1
0
−2
residuals(model)
0
1
2
3
4
5
fitted(model)



Model 3: Nested Model for DayPeriod and Time Effects In this model, dayPeriod is used as a broader time effect, with time nested within dayPeriod.

This model also includes campagne, sample, and chamber as random effects.

```
Linear mixed model fit by REML. t-tests use Satterthwaite's method [
lmerModLmerTest]
Formula: StarchNscTissue ~ tissu + treatment + dayPeriod + dayPeriod:time +
    (1 | campagne) + (1 | sample) + (1 | chamber)
   Data: data

REML criterion at convergence: 1192.6

Scaled residuals:
    Min      1Q  Median      3Q     Max
-2.2829 -0.6858 -0.0363  0.4039  3.8747

Random effects:
 Groups   Name        Variance Std.Dev.
 chamber  (Intercept) 0.0276   0.1661  
 sample   (Intercept) 0.2245   0.4738  
 campagne (Intercept) 0.5451   0.7383  
 Residual             1.0075   1.0037  
Number of obs: 408, groups:  chamber, 8; sample, 8; campagne, 2

Fixed effects:
                      Estimate Std. Error        df t value Pr(>|t|)    
(Intercept)            0.80373    0.59686   1.34648   1.347 0.360931    
tissuIT                0.03626    0.14055 392.01679   0.258 0.796569    
tissuLM                2.42265    0.14055 392.01679  17.237  < 2e-16 ***
tissuUM                2.26940    0.14055 392.01679  16.147  < 2e-16 ***
treatmentDrought      -0.77436    0.36888   5.01240  -2.099 0.089702 .  
dayPeriodNight         0.36950    0.16284 392.08694   2.269 0.023804 *  
dayPeriodDay:timeB     0.24285    0.16591 392.19489   1.464 0.144068    
dayPeriodNight:timeC   0.04250    0.17744 392.01679   0.240 0.810825    
dayPeriodDay:timeE     0.12013    0.16284 392.08694   0.738 0.461138    
dayPeriodDay:timeF     0.62719    0.16591 392.19489   3.780 0.000181 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Correlation of Fixed Effects:
            (Intr) tissIT tissLM tissUM trtmnD dyPrdN dyPD:B dyPN:C dyPD:E
tissuIT     -0.118                                                        
tissuLM     -0.118  0.500                                                 
tissuUM     -0.118  0.500  0.500                                          
trtmntDrght -0.309  0.000  0.000  0.000                                   
dayPerdNght -0.111  0.000  0.000  0.000 -0.001                            
dyPrdDy:tmB -0.110  0.000  0.000  0.000  0.006  0.396                     
dyPrdNght:C  0.000  0.000  0.000  0.000  0.000 -0.545  0.000              
dyPrdDy:tmE -0.111  0.000  0.000  0.000 -0.001  0.406  0.396  0.000       
dyPrdDy:tmF -0.110  0.000  0.000  0.000  0.006  0.396  0.390  0.000  0.396
fit warnings:
fixed-effect model matrix is rank deficient so dropping 6 columns / coefficients

[1] 1220.641

[1] 1276.799

Type III Analysis of Variance Table with Satterthwaite's method
               Sum Sq Mean Sq NumDF  DenDF  F value    Pr(>F)    
tissu          554.01 184.671     3 392.02 183.3018 < 2.2e-16 ***
```

```
treatment         4.44   4.440     1   5.01   4.4068  0.089702 .
dayPeriod         5.19   5.187     1 392.09   5.1489  0.023804 *
dayPeriod:time   15.20   3.801     4 392.11   3.7724  0.005036 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1


    Shapiro-Wilk normality test

data:  residuals(model3)
W = 0.96174, p-value = 8.022e-09
```

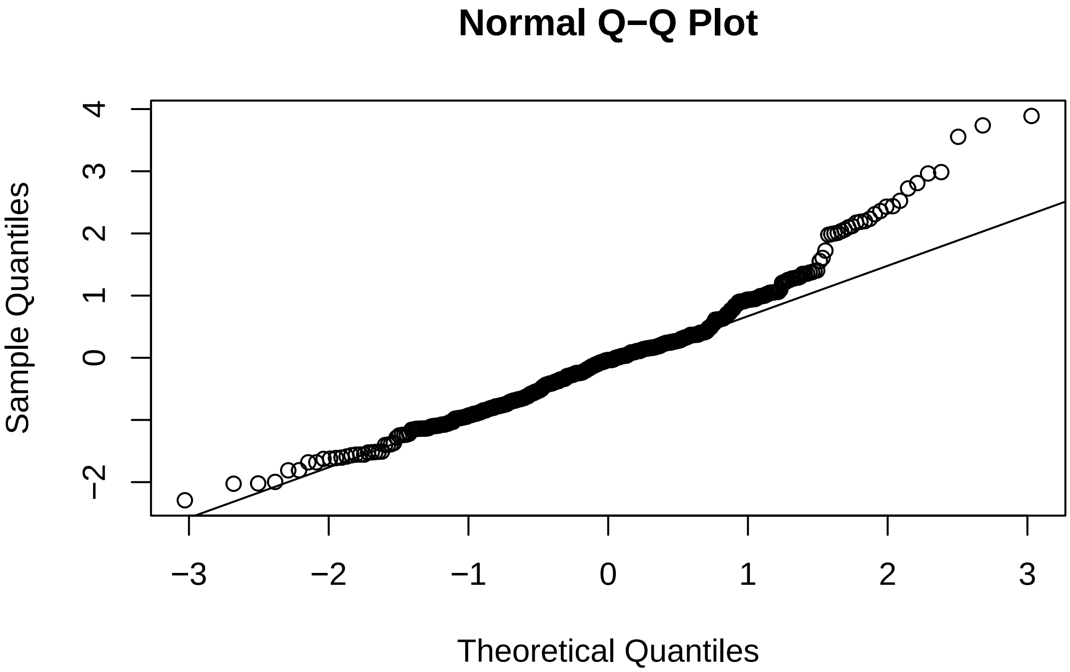


Figure 8

```
Call:
lm(formula = StarchNscTissue ~ treatment * tissu * dayPeriod +
    campagne, data = data)
```

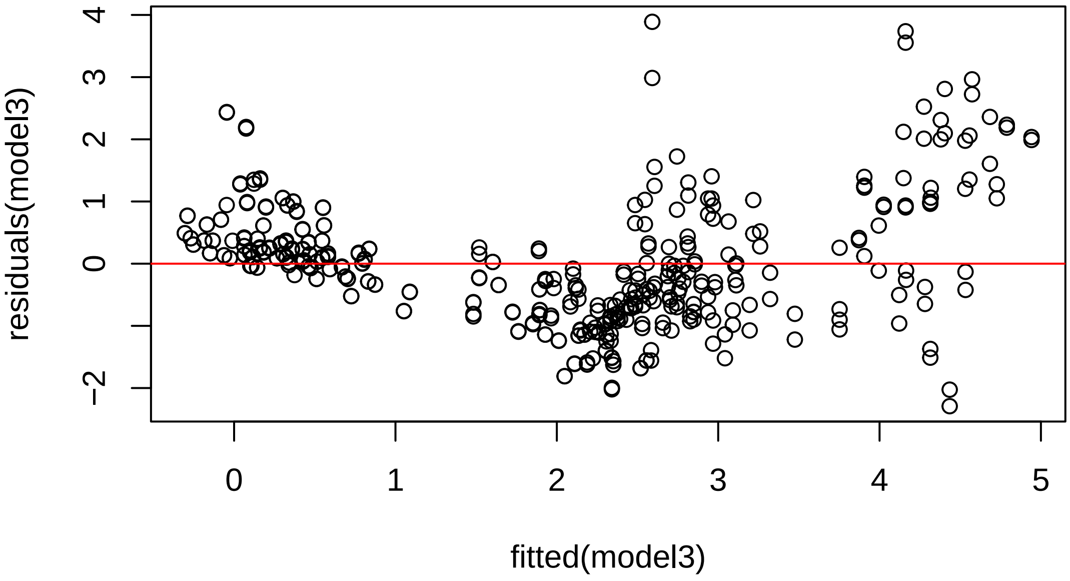


Figure 9

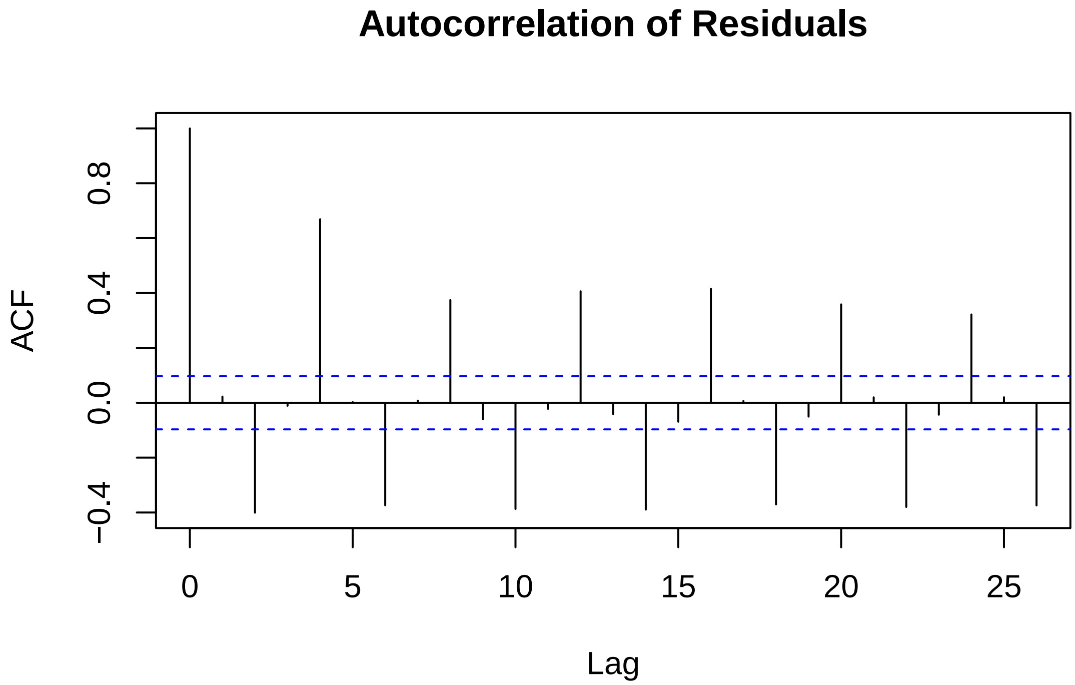
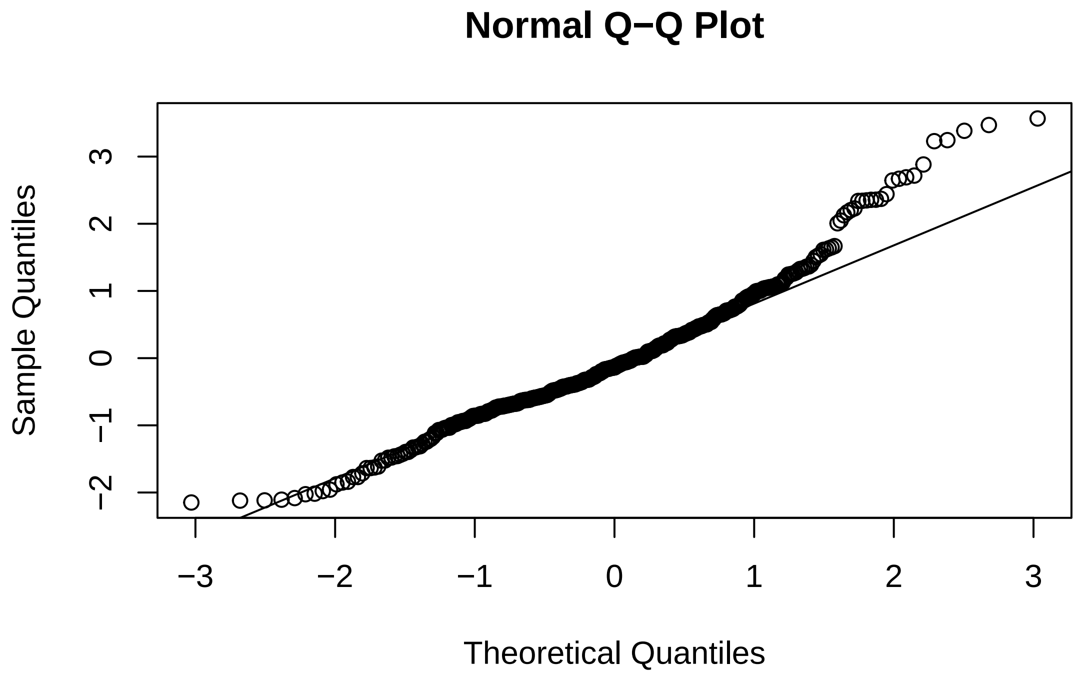

Autocorrelation of Residuals
ACF
0.8
0.4
0.0
−0.4
0
5
10
15
20
25
Lag

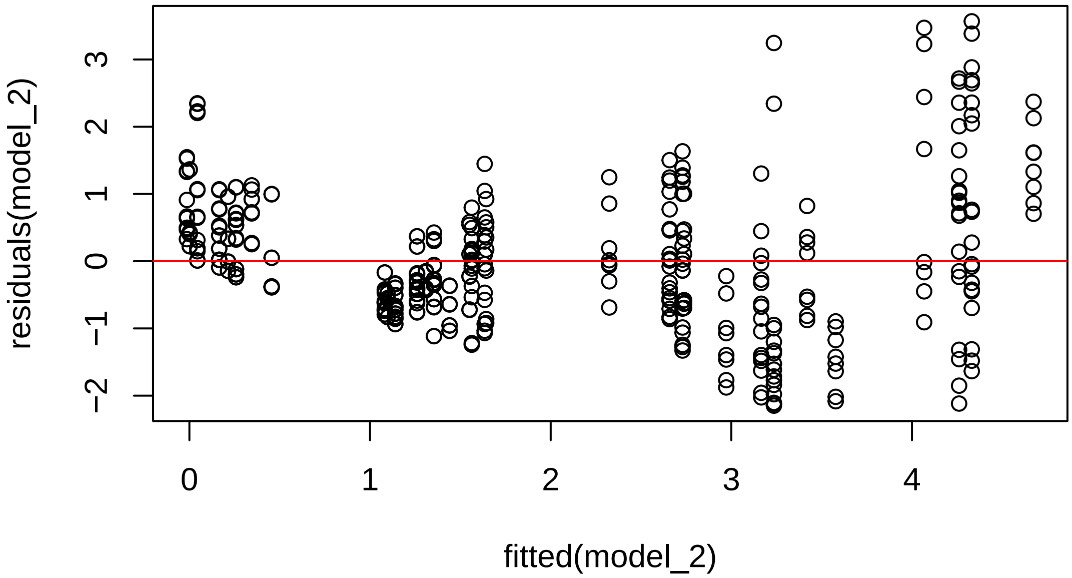
Figure 10

```
Residuals:
    Min      1Q  Median      3Q     Max 
-2.1490 -0.6395 -0.1383  0.5298  3.5674 

Coefficients:
                                        Estimate Std. Error t value Pr(>|t|)
(Intercept)                             -0.93022    0.23779  -3.912 0.000108
treatmentDrought                        -0.17885    0.24974  -0.716 0.474327
tissuIT                                  0.09316    0.24612   0.379 0.705243
tissuLM                                  3.00031    0.24612  12.190  < 2e-16
tissuUM                                  3.07044    0.24612  12.475  < 2e-16
dayPeriodNight                           0.04878    0.31380   0.155 0.876544
campagne                                 1.09540    0.10416  10.516  < 2e-16
treatmentDrought:tissuIT                -0.03511    0.35315  -0.099 0.920866
treatmentDrought:tissuLM                -1.42355    0.35315  -4.031 6.68e-05
treatmentDrought:tissuUM                -1.42214    0.35315  -4.027 6.79e-05
treatmentDrought:dayPeriodNight          0.30988    0.44572   0.695 0.487317
tissuIT:dayPeriodNight                   0.14820    0.44370   0.334 0.738553
tissuLM:dayPeriodNight                   0.36356    0.44370   0.819 0.413066
tissuUM:dayPeriodNight                  -0.31235    0.44370  -0.704 0.481870
treatmentDrought:tissuIT:dayPeriodNight -0.54948    0.63032  -0.872 0.383883
treatmentDrought:tissuLM:dayPeriodNight  0.03887    0.63032   0.062 0.950859
treatmentDrought:tissuUM:dayPeriodNight -0.03772    0.63032  -0.060 0.952308

(Intercept)                             ***
treatmentDrought                           
tissuIT                                    
tissuLM                                 ***
tissuUM                                 ***
dayPeriodNight                             
campagne                                ***
treatmentDrought:tissuIT                   
treatmentDrought:tissuLM                ***
treatmentDrought:tissuUM                ***
treatmentDrought:dayPeriodNight            
tissuIT:dayPeriodNight                     
tissuLM:dayPeriodNight                     
tissuUM:dayPeriodNight                     
treatmentDrought:tissuIT:dayPeriodNight    
treatmentDrought:tissuLM:dayPeriodNight    
treatmentDrought:tissuUM:dayPeriodNight    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 1.044 on 391 degrees of freedom
Multiple R-squared:  0.6513,	Adjusted R-squared:  0.6371 
F-statistic: 45.65 on 16 and 391 DF,  p-value: < 2.2e-16
```

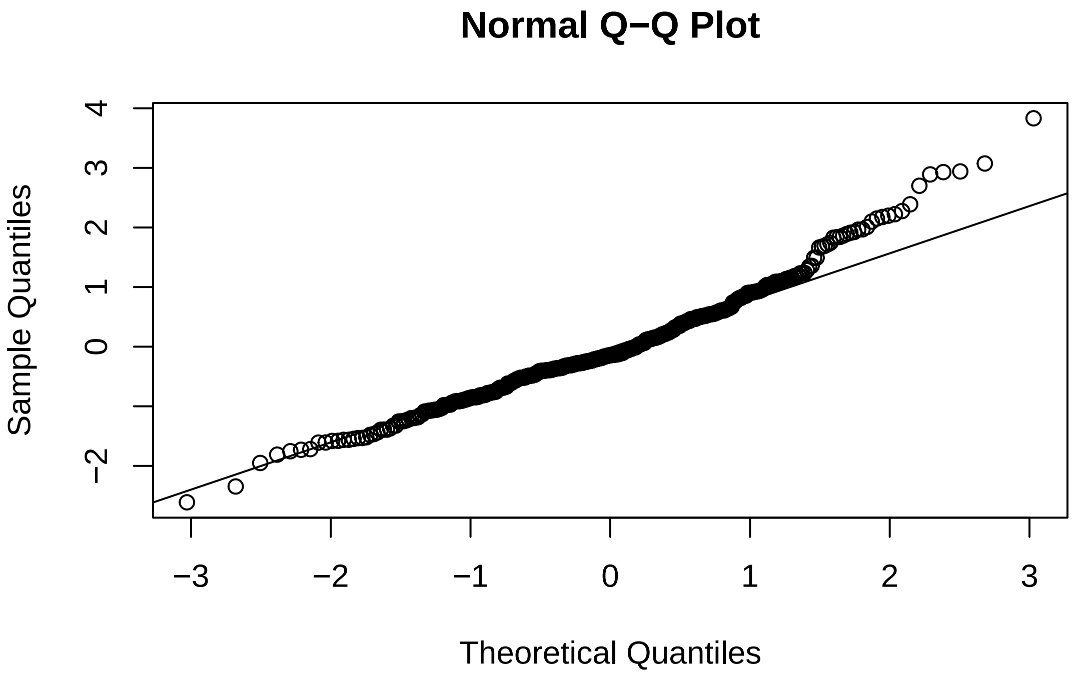


```
	Shapiro-Wilk normality test

data:  residuals(model_2)
W = 0.96501, p-value = 2.709e-08
```

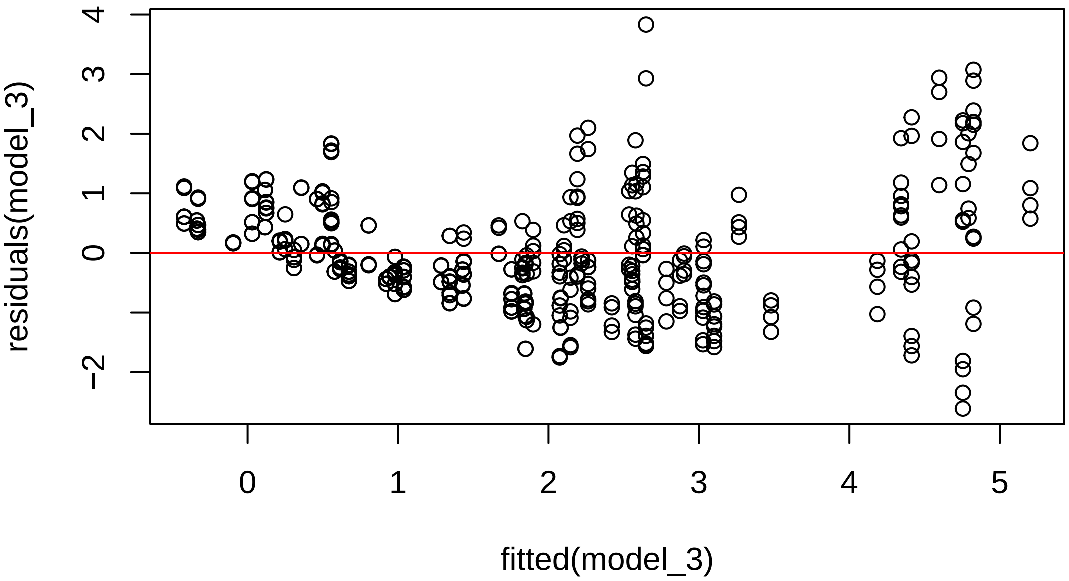


```
Linear mixed model fit by REML. t-tests use Satterthwaite's method [
lmerModLmerTest]
Formula: StarchNscTissue ~ treatment * tissu * dayPeriod + campagne +
    (1 | chamber)
   Data: data

REML criterion at convergence: 1148.9

Scaled residuals:
    Min      1Q  Median      3Q     Max
-2.7113 -0.5756 -0.1467  0.5350  3.9784

Random effects:
 Groups   Name        Variance Std.Dev.
 chamber  (Intercept) 0.2484   0.4984
 Residual             0.9277   0.9632
Number of obs: 408, groups:  chamber, 8

Fixed effects:
                                         Estimate Std. Error        df t value
```

```
(Intercept)                                -0.94922    0.62429   5.65354  -1.520
treatmentDrought                           -0.09081    0.42131   8.85361  -0.216
tissuIT                                     0.09316    0.22702 386.01184   0.410
tissuLM                                     3.00031    0.22702 386.01184  13.216
tissuUM                                     3.07044    0.22702 386.01184  13.525
dayPeriodNight                              0.08484    0.28949 386.02903   0.293
campagne                                    1.08403    0.36534   5.01449   2.967
treatmentDrought:tissuIT                   -0.03511    0.32574 386.01184  -0.108
treatmentDrought:tissuLM                   -1.42355    0.32574 386.01184  -4.370
treatmentDrought:tissuUM                   -1.42214    0.32574 386.01184  -4.366
treatmentDrought:dayPeriodNight             0.22184    0.41140 386.09222   0.539
tissuIT:dayPeriodNight                      0.14820    0.40926 386.01184   0.362
tissuLM:dayPeriodNight                      0.36356    0.40926 386.01184   0.888
tissuUM:dayPeriodNight                     -0.31235    0.40926 386.01184  -0.763
treatmentDrought:tissuIT:dayPeriodNight    -0.54948    0.58140 386.01184  -0.945
treatmentDrought:tissuLM:dayPeriodNight     0.03887    0.58140 386.01184   0.067
treatmentDrought:tissuUM:dayPeriodNight    -0.03772    0.58140 386.01184  -0.065
                                          Pr(>|t|)
(Intercept)                                 0.1822
treatmentDrought                            0.8342
tissuIT                                     0.6818
tissuLM                                    < 2e-16 ***
tissuUM                                    < 2e-16 ***
dayPeriodNight                              0.7696
campagne                                    0.0311 *
treatmentDrought:tissuIT                    0.9142
treatmentDrought:tissuLM                  1.60e-05 ***
treatmentDrought:tissuUM                  1.63e-05 ***
treatmentDrought:dayPeriodNight             0.5900
tissuIT:dayPeriodNight                      0.7175
tissuLM:dayPeriodNight                      0.3749
tissuUM:dayPeriodNight                      0.4458
treatmentDrought:tissuIT:dayPeriodNight     0.3452
treatmentDrought:tissuLM:dayPeriodNight     0.9467
treatmentDrought:tissuUM:dayPeriodNight     0.9483
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```



```

	Shapiro-Wilk normality test

data:  residuals(model_3)
W = 0.97131, p-value = 3.444e-07
```



Conclusion

GitHub page found here.

References

Astley, Rick. 1987. "Never Gonna GIve You Up." 1987. https://r.mtdv.me/videos/6QMWR9vBma.

Kassambara. 2024. "Normality Test in r." 2024. http://www.sthda.com/english/wiki/normality-test-in-r.

Appendix A - R Code

```r
## Prints code without running it

library(knitr)
data <- read.csv("data.csv")
knitr::kable(head(data), format = 'markdown')
```

Appendix B - SAS Code

```
data rptm_means;
input Inoculation_Method $ Thickness $ @@;
do Week=1 to 5 by 1;
    input mu @@;
    output;
end;
datalines;
Dry 1/4 4.2573 4.246 4.474 4.3327 4.0127
Dry 1/8 5.2907 4.9513 5.2013 5.2073 4.9713
Wet 1/4 5.4013 5.5727 5.55 5.4873 5.3807
Wet 1/8 5.56 5.7793 5.6313 5.7153 5.62
;
```