Data Analysis 2

Maksuda Aktar Toma, Jo Charbonneau, Ryan Lalicker

November 4, 2024

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.

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: https://github.com/RyanLalicker/Data-Analysis-2-STAT-325-825.

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. As the column name may suggest, this will be considered the treatment,. 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.

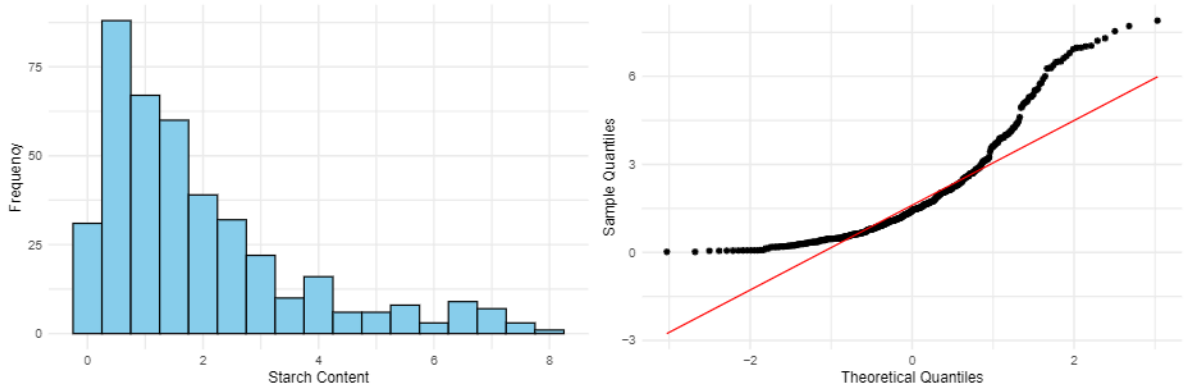
For starch contents across all measurements, the values range from about 0.019 to 7.898 with a median of roughly 1.430 and a mean of 1.925. The location of the median and mean with respect to the minimum and maximum is an early sign that the starch contents could be skewed and thus non-normal in distribution.

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.

As previously noted, the table above indicates the starch contents may be skewed and thus non-normal. This can be evaluated through a histogram and Q-Q plot. The histogram below supports our suspicion that the data is skewed and the Q-Q plot confirms the measure is non-normal. Note, all 408 measurements of starch content are used in the plots.



(a) Histogram of starch content.

(b) Q-Q plot of starch content.

Figure 2: Plots used to check normallity assumption.

Relationships among variables

Now let's see how some of the other variables relate to the starch content. First we can look at the four tissue types. To do this we will use the boxplot below. It appears the tissue types END and IT are similar to each other, as are LM and UM. The two pairs seem quite a bit different though as LM and UM have both far higher values than the other two. This indicates the tissue type could be significant.

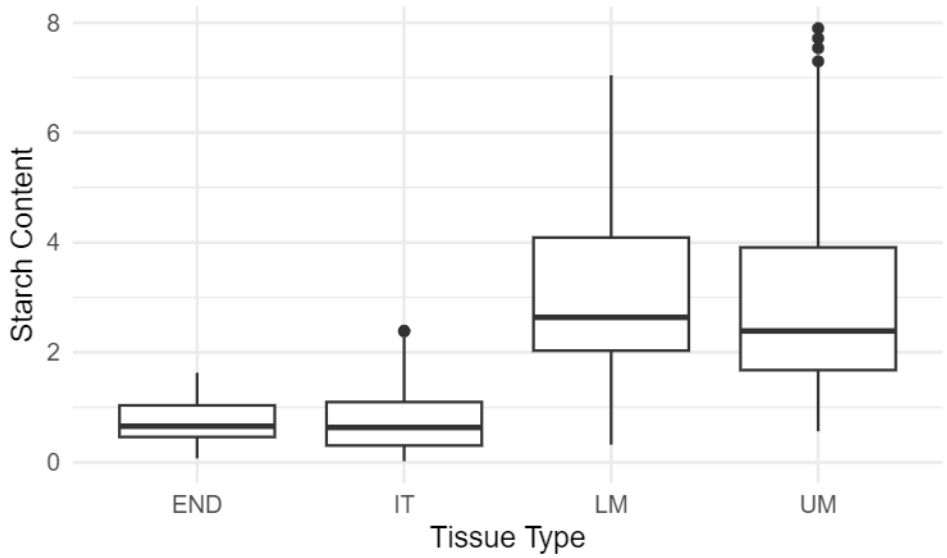


Figure 3: Boxplot of starch contents by tissue types

Another variable of that could have a major impact is the treatment. If some samples get more water than others it would make sense to see more growth. It is also possible that the time could impact the effect the water has on the starch content. Below is a bar chart that separates measurements first by day and night, and then by the treatment while still showing the differences in tissue type. Remember time points C and D are at night and the rest are during the day.

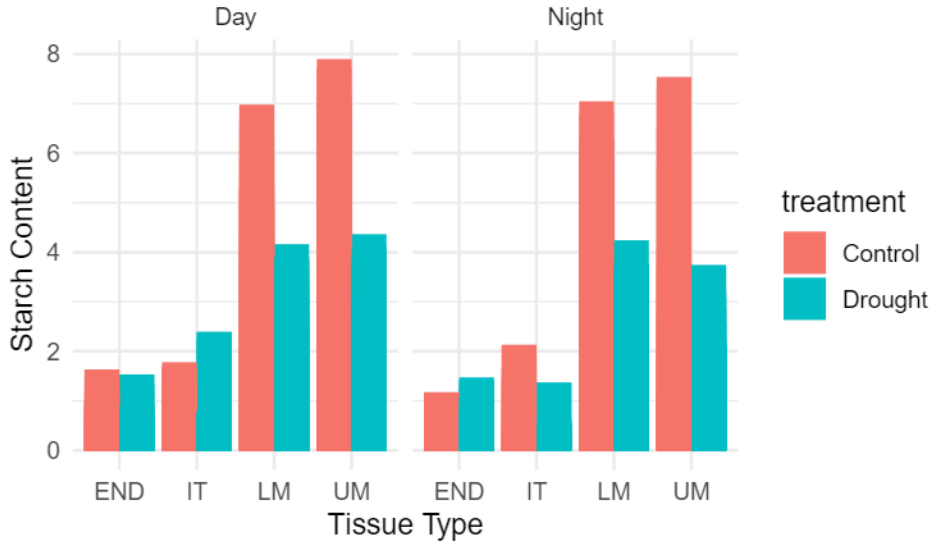


Figure 4: Barchat of starch content vs. tissue types, separating by treatment and day or night.

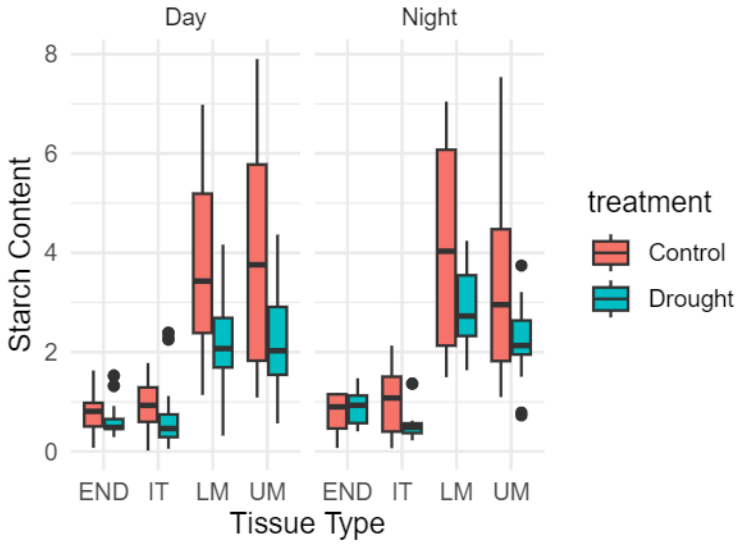


Figure 5: Boxplot of Starch Content by Tissue Type and Treatment.

In Figure 4, we can see the starch content for measurements with the tissue types LM and UM are higher when given the control treatment instead of the drought treatment. This is not as clear with the other two tissue types. Additionally, the effect day and night have on the starch contents are not clear, as we saw in the summary statistics table above.

In Figure 5 we can observe the groups with the control treatment tend to have more variance than those with the drought treatment. Additionally, there are a handful of outliers across the different tissue types.In summary, **starch content is influenced by tissue type and treatment**, with `LM` and `UM` tissues under Control treatment during the Day showing the highest levels.

Potential models

The replication mentioned previously suggests a mixed model approach is needed. This is due to the replication being a random effect. The simplest case of this type of model is a linear mixed model, but there generalized linear mixed models are also a possibility. Now we will explore few models to see which one fits better for this data set.

Linear Mixed Model

The first model we want to consider is a linear mixed model with fixed effects treatment, tissue type, and the period of the day, along with random effects for the larger location (`campagne`), the sample specific location (`chamber`), and the sample itself. Additionally, this model includes interaction terms for the fixed effects. This can be expressed as

$$y_{ijklmn} = \mu + \tau_i + \alpha_j + \beta_k + (\tau\alpha)_{ij} + (\tau\beta)_{ik} + (\alpha\beta)_{jk} + (\tau\alpha\beta)_{ijk} + u_l + v_m + w_n + \epsilon_{ijklmn}$$

where y_{ijklm} represents the starch content, μ is the overall mean, τ_i is the fixed effect for the ith treatment, α_j is the fixed effect for the jth tissue type, and β_k is the fixed effect for the period of the day. For the random effects u_l is the effect for the `campagne` variable, v_m is the effect for `chamber`, and w_n is the effect for the sample. The residuals are represented by ϵ_{ijklm}. The remaining terms represent the interaction between the fixed effects. For instance $(\tau\alpha)_{ij}$ is the interaction effect of the treatment and tissue type, while $(\tau\alpha\beta)_{ijk}$ represents the three-way interaction of all fixed effects in the model.

The model was applied in SAS and all code can be found in *Appendix B - SAS Code*. The figure below shows three tables that are a part of the SAS output. The *Fit Statistics* tables suggests we have a reasonably fitting model. Note these values can also be used for comparison later.

Estimated G matrix is not positive definite.

Covariance Parameter Estimates	
Cov Parm	Estimate
campagne	1.75E-18
chamber	0.1694
sample	4.898E-6
Residual	0.9277

Fit Statistics	
-2 Res Log Likelihood	1150.3
AIC (Smaller is Better)	1156.3
AICC (Smaller is Better)	1156.3
BIC (Smaller is Better)	1152.4

Type 3 Tests of Fixed Effects				
Effect	Num DF	Den DF	F Value	Pr > F
treatment	1	386	6.26	0.0128
tissu	3	386	172.71	<.0001
treatment*tissu	3	386	13.06	<.0001
dayPeriod	1	386	2.94	0.0874
treatment*dayPeriod	1	386	0.18	0.6731
tissu*dayPeriod	3	386	2.14	0.0950
treatm*tissu*dayPeri	3	386	0.45	0.7153

Figure 6: SAS output of *Covariance Parameter Estimates*, *Fit Statistics*, and *Type 3 Tests of Fixed Effects* for the first proposed model.

The first table in the figure above, the *Covariance Parameter Estimates*, show how much of the variance each random variable and the residuals are responsible for. We can see `campagne` and `sample` have almost no effect on the variance. The `chamber` does have a small effect on the total variance, indicating it plays a part in the starch content.

The *Type 3 Tests of Fixed Effects* reports what fixed effects are registering as significant. With p-values less than 0.0001 both the tissue and the treatment by tissue interaction are highly significant. The treatment effect on its own is still significant at a significance level of 5%. The day period and its interaction with the tissue type are marginally significant, but neither are at the 5% level. The remaining interactions are not significant either.

The *Least Squares Means* table below further investigates the fixed effects. We can see the estimate for each level of each variable in the `Estimate` column, as well as the p-value in the `Pr > |t|` column. As expected the estimated effect for the control treatment is greater than that of the drought treatment, and the LM and UM tissue types have larger estimates than the END and IT types. A somewhat surprising result is that the estimated coefficient for night is greater than that of day though not my much.

Least Squares Means											
Effect	treatment	tissu	dayPeriod	Estimate	Standard Error	DF	t Value	Pr > |t|	Alpha	Lower	Upper
treatment	Control			1.3348	0.3383	386	3.95	<.0001	0.05	0.6698	1.9999
treatment	Drought			0.5624	0.3394	386	1.66	0.0983	0.05	-0.1048	1.2297
dayPeriod			Day	0.8603	0.3036	386	2.83	0.0048	0.05	0.2634	1.4573
dayPeriod			Night	1.0369	0.3083	386	3.36	0.0008	0.05	0.4308	1.6431
tissu		END		-0.2229	0.3145	386	-0.71	0.4788	0.05	-0.8412	0.3954
tissu		IT		-0.2106	0.3145	386	-0.67	0.5035	0.05	-0.8288	0.4077
tissu		LM		2.2571	0.3145	386	7.18	<.0001	0.05	1.6389	2.8754
tissu		UM		1.9708	0.3145	386	6.27	<.0001	0.05	1.3526	2.5891

Figure 7: *Least Squares Means* table for the first proposed model.

In terms of significance, the control treatment is highly significant while the drought treatment is only marginally so. Similarly, the LM and UM tissue types are highly significant while IT and END are not at all. Both periods of day seem to be significant though.

The *Differences of Least Squares Means* table shows pairwise comparisons for the fixed effects in the model, with Tukey-Kramer adjustments for multiple comparisons. (Lane (2010)). This allows us to see whether changing the level is significant holding all else constant . Using the adjusted p-values, found in the `Adj P` column, we can see there are significant differences at the 5% between the treatment levels as well as most tissue types, with many being significant at lower levels. The lone exception to this in regards to the tissue levels is the difference between LM and UM. Additionally, the difference between day and night is only marginally significant.

Differences of Least Squares Means																	
Effect	treatment	tissu	dayPeriod	_treatment	_tissu	_dayPeriod	Estimate	Standard Error	DF	t Value	Pr > |t|	Adjustment	Adj P	Alpha	Lower	Upper	Adj Lower	Adj Upper
treatment	Control			Drought			0.7724	0.3088	386	2.50	0.0128	Tukey-Kramer	0.0128	0.05	0.1654	1.3795	0.1654	1.3795
dayPeriod			Day			Night	-0.1766	0.1031	386	-1.71	0.0874	Tukey-Kramer	0.0874	0.05	-0.3793	0.02603	-0.3793	0.02603
tissu		END			IT		-0.01234	0.1454	386	-0.08	0.9324	Tukey-Kramer	0.9998	0.05	-0.2981	0.2734	-0.3874	0.3627
tissu		END			LM		-2.4800	0.1454	386	-17.06	<.0001	Tukey-Kramer	<.0001	0.05	-2.7658	-2.1943	-2.8551	-2.1050
tissu		END			UM		-2.1938	0.1454	386	-15.09	<.0001	Tukey-Kramer	<.0001	0.05	-2.4795	-1.9080	-2.5688	-1.8187
tissu		IT			LM		-2.4677	0.1454	386	-16.98	<.0001	Tukey-Kramer	<.0001	0.05	-2.7535	-2.1819	-2.8427	-2.0927
tissu		IT			UM		-2.1814	0.1454	386	-15.01	<.0001	Tukey-Kramer	<.0001	0.05	-2.4672	-1.8956	-2.5565	-1.8064
tissu		LM			UM		0.2863	0.1454	386	1.97	0.0496	Tukey-Kramer	0.2013	0.05	0.000504	0.5721	-0.08876	0.6613

Figure 8: *Differences of Least Squares Means* table for the first proposed model.

Since we are working with mixed models, certain assumptions need to hold for us to trust the output above. One is that the residuals are both normally distributed and random, or homoscedastic. (Issa and Nadal (2011)). These can be checked graphically. The SAS figure below shows three graphs as well as statistics discussed above. The histogram, top right, and

Q-Q plot, bottom left, indicate the normality assumption holds. However, the top left graph presents an issue with the model. When residuals are random, this plot should be randomly scattered. In the figure below, there seems to be a fanning out pattern, which indicates homoscedasticity may be violated, meaning heteroskedasticity is present.

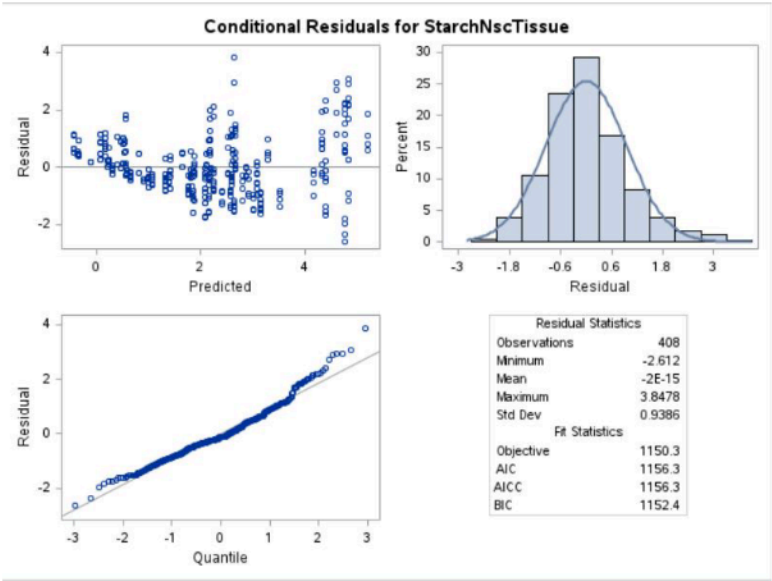


Figure 9: Residual plots and statistics for the first proposed model.

While one could argue the homoscedasticity assumption is not violated, the graphical evidence in enough for us to have questions regarding the model's viability. With that in mind, other models need to be considered.

Nested Model

The next model we want to consider is another linear mixed model. Many of the terms in the model below are the same as before in terms of notation. The additions to this model are the nested structure of the time points within the period of the day. In the model below many of the effects have the same notation as before, but $\beta_{k(l)}$ represents the effect of the lth time nested within the kth period of the day. These changes occur in the interaction terms as well producing the model

$$y_{ijklmno} = \mu + \tau_i + \alpha_j + \beta_{k(l)} + (\tau\alpha)_{ij} + (\tau\beta)_{i,k(l)} + (\alpha\beta)_{j,k(l)} + (\tau\alpha)_{i,jk(l)} + u_m + v_n + w_o + \epsilon_{ijklmno}$$

Now let's consider the same SAS tables and figures we saw in the first proposed model, this time for our nested model. In the *Covariance Parameter Estimates* table below we can see

very different results than previously. Here the estimated variance due to `campagne` has risen to 0.5561. Additionally, the `sample` has an estimated variance of 0.2463, which means both of these affect the starch content. The `chamber` seems to have little effect though.

Covariance Parameter Estimates	
Cov Parm	Estimate
sample	0.2463
campagne	0.5561
chamber	0.000305
Residual	0.8322

Fit Statistics	
-2 Res Log Likelihood	1076.3
AIC (Smaller is Better)	1084.3
AICC (Smaller is Better)	1084.4
BIC (Smaller is Better)	1084.6

Type 3 Tests of Fixed Effects				
Effect	Num DF	Den DF	F Value	Pr > F
treatment	1	5.02	4.52	0.0866
dayPeriod(time)	5	354	4.14	0.0011
treatm*dayPeri(time)	5	354	3.11	0.0092
tissu	3	354	218.53	<.0001
treatment*tissu	3	354	16.65	<.0001
dayPerio*tissu(time)	15	354	2.46	0.0020
trea*dayP*tiss(time)	15	354	0.92	0.5381

Figure 10: SAS output of *Covariance Parameter Estimates*, *Fit Statistics*, and *Type 3 Tests of Fixed Effects* for the proposed nested model.

The *Fit Statistics* shows values slightly smaller than what we saw with the previous model. This could mean the nested approach is a slightly better fit than before. The *Type 3 Tests of Fixed Effects* table shows which fixed effects and interactions are significant. Here we can see the treatment is marginally significant, but both tissue type and time nested within day are significant at the 1% level. Additionally, all two way interactions are significant, though the three-way interaction is not. It would seem that time nested within day is now a primary factory in determining starch content.

Least Squares Means

Effect	treatment	dayPeriod	tissu	time	Estimate	Standard Error	DF	t Value	Pr > |t|	Alpha	Lower	Upper
treatment	Control				2.2755	0.5863	1.22	3.88	0.1248	0.05	-2.6433	7.1943
treatment	Drought				1.5034	0.5866	1.22	2.56	0.1987	0.05	-3.4035	6.4104
dayPeriod(time)		Day		A	1.5970	0.5637	1.04	2.83	0.2074	0.05	-4.8844	8.0785
dayPeriod(time)		Day		B	1.8341	0.5682	1.08	3.23	0.1766	0.05	-4.2583	7.9265
dayPeriod(time)		Night		C	2.0106	0.5673	1.07	3.54	0.1618	0.05	-4.1542	8.1755
dayPeriod(time)		Night		D	1.9681	0.5673	1.07	3.47	0.1653	0.05	-4.1967	8.1330
dayPeriod(time)		Day		E	1.7188	0.5673	1.07	3.03	0.1894	0.05	-4.4461	7.8836
dayPeriod(time)		Day		F	2.2081	0.5682	1.08	3.89	0.1461	0.05	-3.8844	8.3005
tissu			END		0.7096	0.5633	1.04	1.26	0.4208	0.05	-5.8114	7.2305
tissu			IT		0.7259	0.5633	1.04	1.29	0.4137	0.05	-5.7951	7.2468
tissu			LM		3.1347	0.5633	1.04	5.57	0.1063	0.05	-3.3863	9.6556
tissu			UM		2.9877	0.5633	1.04	5.30	0.1116	0.05	-3.5332	9.5087

Figure 11: *Least Squares Mean* table for the nested model.

Now let's consider the *Least Squares Means* table. Not one level is considered significant on its own. There are some significant differences though, as seen in the *Differences of Least Squares Means* table below. We can see many of the tissue types have a significant difference as well as a handful of the time points such as A and D or E and F. To get a better view of this and other tables, you may want to consider using the code in *Appendix B - SAS Code* to replicate the results.

Differences of Least Squares Means

Effect	treatment	dayPeriod	tissu	time	_treatment	_dayPeriod	_tissu	_time	Estimate	Standard Error	DF	t Value	Pr > |t|	Adjustment	Adj P	Alpha	Lower	Upper	Adj Lower	Adj Upper
treatment	Control				Drought				0.7721	0.3630	5.02	2.13	0.0866	Tukey-Kramer	0.0866	0.05	-0.1601	1.7043	-0.1601	1.7043
dayPeriod(time)		Day		A		Day		B	-0.2371	0.1510	354	-1.57	0.1172	Tukey-Kramer	0.6187	0.05	-0.5341	0.05986	-0.6698	0.1956
dayPeriod(time)		Day		A		Night		C	-0.4136	0.1480	354	-2.79	0.0055	Tukey-Kramer	0.0607	0.05	-0.7047	-0.1225	-0.8377	0.01055
dayPeriod(time)		Day		A		Night		D	-0.3711	0.1480	354	-2.51	0.0126	Tukey-Kramer	0.1248	0.05	-0.6622	-0.07998	-0.7952	0.05305
dayPeriod(time)		Day		A		Day		E	-0.1217	0.1480	354	-0.82	0.4115	Tukey-Kramer	0.9633	0.05	-0.4128	0.1694	-0.5459	0.3024
dayPeriod(time)		Day		A		Day		F	-0.6110	0.1510	354	-4.05	<.0001	Tukey-Kramer	0.0009	0.05	-0.9080	-0.3141	-1.0437	-0.1784
dayPeriod(time)		Day		B		Night		C	-0.1765	0.1643	354	-1.07	0.2836	Tukey-Kramer	0.8915	0.05	-0.4997	0.1467	-0.6474	0.2944
dayPeriod(time)		Day		B		Night		D	-0.1340	0.1643	354	-0.82	0.4154	Tukey-Kramer	0.9646	0.05	-0.4572	0.1892	-0.6049	0.3369
dayPeriod(time)		Day		B		Day		E	0.1154	0.1643	354	0.70	0.4831	Tukey-Kramer	0.9816	0.05	-0.2078	0.4386	-0.3555	0.5863
dayPeriod(time)		Day		B		Day		F	-0.3739	0.1669	354	-2.24	0.0257	Tukey-Kramer	0.2221	0.05	-0.7022	-0.04564	-0.8522	0.1044
dayPeriod(time)		Night		C		Night		D	0.04250	0.1613	354	0.26	0.7923	Tukey-Kramer	0.9998	0.05	-0.2747	0.3597	-0.4196	0.5046
dayPeriod(time)		Night		C		Day		E	0.2919	0.1613	354	1.81	0.0712	Tukey-Kramer	0.4606	0.05	-0.02528	0.6090	-0.1702	0.7540
dayPeriod(time)		Night		C		Day		F	-0.1974	0.1643	354	-1.20	0.2304	Tukey-Kramer	0.8361	0.05	-0.5207	0.1258	-0.6684	0.2735
dayPeriod(time)		Night		D		Day		E	0.2494	0.1613	354	1.55	0.1229	Tukey-Kramer	0.6344	0.05	-0.06778	0.5665	-0.2127	0.7115
dayPeriod(time)		Night		D		Day		F	-0.2399	0.1643	354	-1.46	0.1452	Tukey-Kramer	0.6900	0.05	-0.5632	0.08328	-0.7109	0.2310
dayPeriod(time)		Day		E		Day		F	-0.4893	0.1643	354	-2.98	0.0031	Tukey-Kramer	0.0364	0.05	-0.8125	-0.1661	-0.9602	-0.01839
tissu			END				IT		-0.01630	0.1296	354	-0.13	0.9000	Tukey-Kramer	0.9993	0.05	-0.2711	0.2385	-0.3507	0.3181
tissu			END				LM		-2.4251	0.1296	354	-18.72	<.0001	Tukey-Kramer	<.0001	0.05	-2.6799	-2.1703	-2.7595	-2.0907
tissu			END				UM		-2.2781	0.1296	354	-17.58	<.0001	Tukey-Kramer	<.0001	0.05	-2.5330	-2.0233	-2.6126	-1.9437
tissu			IT				LM		-2.4088	0.1296	354	-18.59	<.0001	Tukey-Kramer	<.0001	0.05	-2.6636	-2.1540	-2.7432	-2.0744
tissu			IT				UM		-2.2618	0.1296	354	-17.46	<.0001	Tukey-Kramer	<.0001	0.05	-2.5166	-2.0070	-2.5963	-1.9274
tissu			LM				UM		0.1470	0.1296	354	1.13	0.2575	Tukey-Kramer	0.6686	0.05	-0.1079	0.4018	-0.1875	0.4814

Figure 12: *Differences of Least Squares Means* table for the nested model.

There had been some hope that the nested structure of the model may help with the potential homoscedasticity violation seen in the first linear mixed model proposed. In the SAS figure of three graphs below, we can see the problem persists in the top left graph. It is worth noting though that the normality assumption seems to hold still.

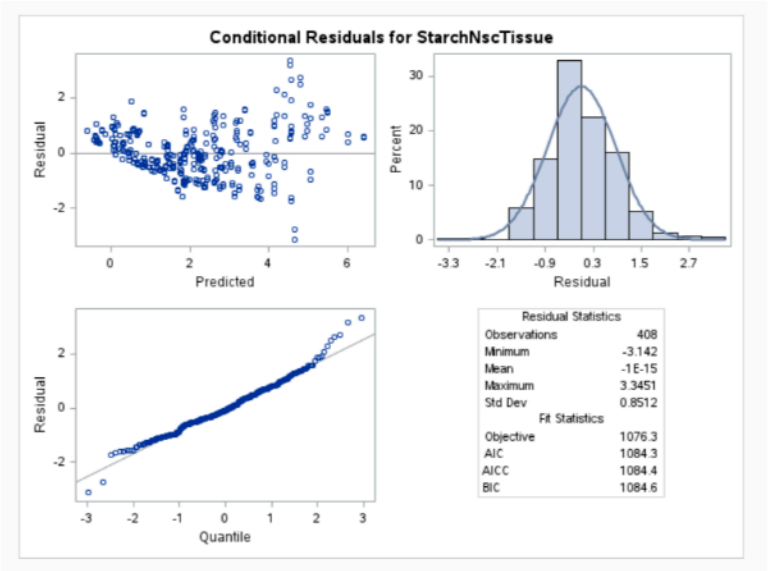


Figure 13: Residual plots and statistics for the nested model.

While the potential assumption violation still persists, this nested model does seem to be an improvement upon the original. For that reason we chose to keep the nested structure going forward.

GLMM Model with Nested Structure

For our third model, we want to consider a generalized linear mixed model, or GLMM, instead of the linear mixed models we've just looked at. This approach can be used on any response variable that follows a distribution belonging to an exponential family. For this approach, link functions are used to work with these different types of distributions. (Slavkovic (n.d.)).

To use this approach we should determine a distribution that fits the starch content. In Figure 2, we found the distribution has a skew. One distribution that could fit this shape is a gamma distribution. According to Casella and Berger (2001) and Hohenstein (2018), the gamma distribution requires some positive parameters α, β such that $E(X) = \alpha/\beta$ and $Var(X) = \alpha/\beta^2$ where $E(X)$ and $Var(X)$ represent the mean and variance of some variable X respectively. In our case X is the starch content. Using the formulas above it can be shown that if our response variable follows a gamma distribution, it would be with an α of roughly

1.2333 and a β of 0.6407. (Casella and Berger (2001); Hohenstein (2018)). Figure 14 shows a gamma distribution with these parameters on top of the histogram of starch content seen previously. We can see the data fits this distribution fairly well so we will proceed.

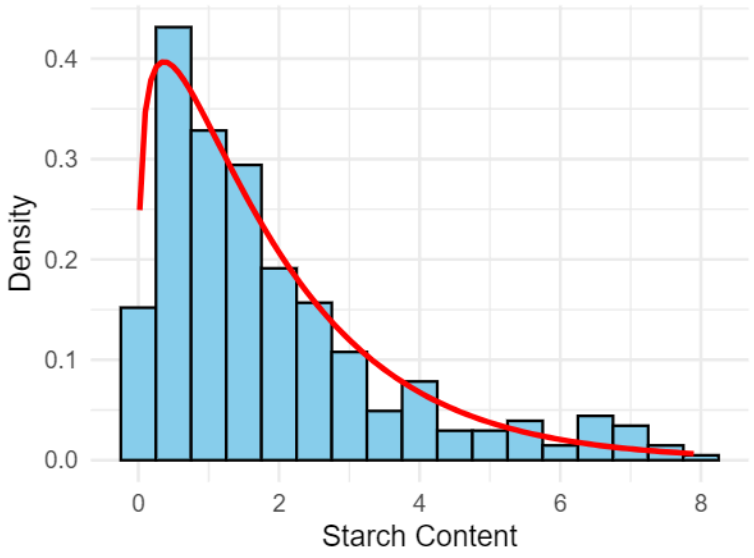


Figure 14: Histogram of starch content with overlay of gamma(1.2333, 0.6407) distribution.

Since we are using a gamma distribution in this GLMM, we need to use the appropriate link function. According to Newsom (2021), the canonical link function that is often used with a gamma GLMMs is the inverse function.

This means that

$$\mathbb{E}(y_{ijklmno})^{-1} = \eta_{ijklmno}$$

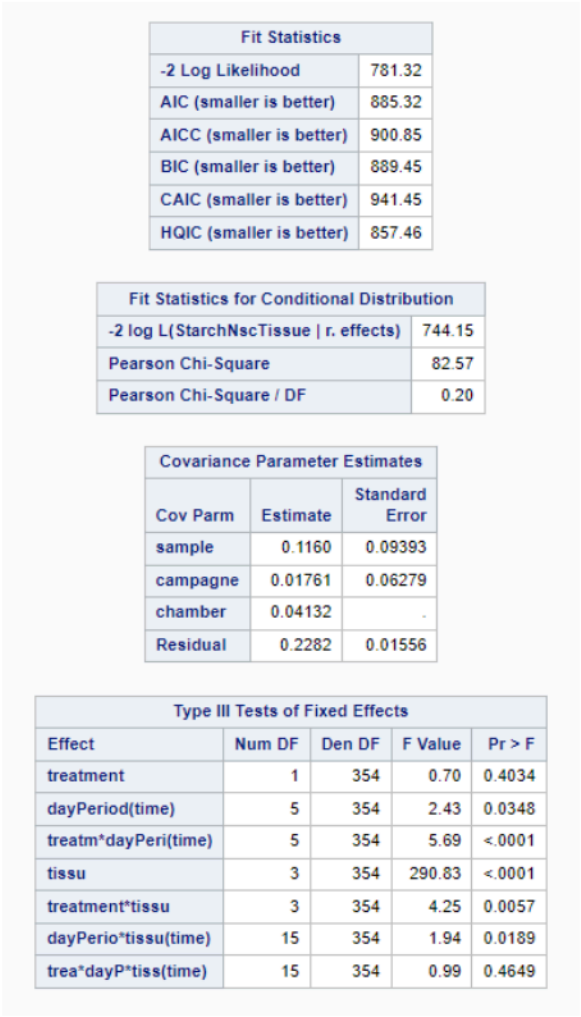
where $E(y_{ijklmno})$ is the mean of the starch content and $\eta_{ijklmno}$ represents the inverse of that mean. Using this new variable $\eta_{ijklmno}$ we can set up our model as

$$\eta_{ijklmno} = \mu + \tau_i + \alpha_j + \beta_{k(l)} + (\tau\alpha)_{ij} + (\tau\beta)_{i,k(l)} + (\alpha\beta)_{j,k(l)} + (\tau\alpha)_{i,jk(l)} + u_m + v_n + w_o$$

In this model all terms and types (fixed and random) match what was included in the first model since all nested effects were removed. As we move forward with analyzing SAS output from this model, it is important to keep in mind the formula above is modeling $\eta_{ijklmno}$, not $E(y_{ijklmno})$.

Let's consider a part of the SAS output that is shown below. The *Fit Statistics* table shows a lower AIC compared to our other models, but we need to be careful not to dig too deep into this since this model is actually looking at the inverse of starch contents. One potential

concern is in the *Fit Statistics for Conditional Distribution* table. The value of 0.20 for `Pearson Chi-Square / DF` is not outstanding since values close to 1 indicate a good fit. This implies there could be some overdispersion, which could contribute to the relatively low variance estimate for the residuals in the *Covariance Parameter Estimates* table. In terms of significance, only the treatment and the three-way interaction are not significant at the 5% level.



Fit Statistics	
-2 Log Likelihood	781.32
AIC (smaller is better)	885.32
AICC (smaller is better)	900.85
BIC (smaller is better)	889.45
CAIC (smaller is better)	941.45
HQIC (smaller is better)	857.46

Fit Statistics for Conditional Distribution	
-2 log L(StarchNscTissue | r. effects)	744.15
Pearson Chi-Square	82.57
Pearson Chi-Square / DF	0.20

Covariance Parameter Estimates		
Cov Parm	Estimate	Standard Error
sample	0.1160	0.09393
campagne	0.01761	0.06279
chamber	0.04132	.
Residual	0.2282	0.01556

Type III Tests of Fixed Effects				
Effect	Num DF	Den DF	F Value	Pr > F
treatment	1	354	0.70	0.4034
dayPeriod(time)	5	354	2.43	0.0348
treatm*dayPeri(time)	5	354	5.69	<.0001
tissu	3	354	290.83	<.0001
treatment*tissu	3	354	4.25	0.0057
dayPerio*tissu(time)	15	354	1.94	0.0189
trea*dayP*tiss(time)	15	354	0.99	0.4649

Figure 15: SAS output of *Covariance Parameter Estimates*, *Fit Statistics*, *Fit Statistics for Conditional Distribution*, and *Type 3 Tests of Fixed Effects* for GLMM.

The next figure shows two tables for each fixed effect which are the *Least Squares Means* and *Differences of treatment Least Squares Means* tables. For the treatments, the control level has a significant effect at the 5% level, but the difference between it and the drought level are not significant. Both day and night are significant levels of the day period, but again their

difference is not significant. All four tissue types and most of the differences are considered significant. The exceptions to this are the differences between END and IT types and LM and UM types.

Within the SAS output their are a series of tables looking at the least squares means and their differences for all levels of each fixed effect. These tables can be found in *Appendix C - Additional SAS Output*. The results are in-line with what we have seen previously and do not raise any concerns with this model. Once again please keep in mind the estimates shown in the tables correspond to an inverse model. This means that time A nested within day having an estimated effect of 0.1857 does not mean at time A the predicted starch content increases by 0.1857.

Now let's see the residual plots. Again, the Q-Q plot looks relatively normal and the histogram is decent other than it is not centered around zero, but this is not an issue since the normality assumption is not required for a GLMM. The residual versus linear predictor plot looks significantly better though as there does not seem to be any trend in the graph.

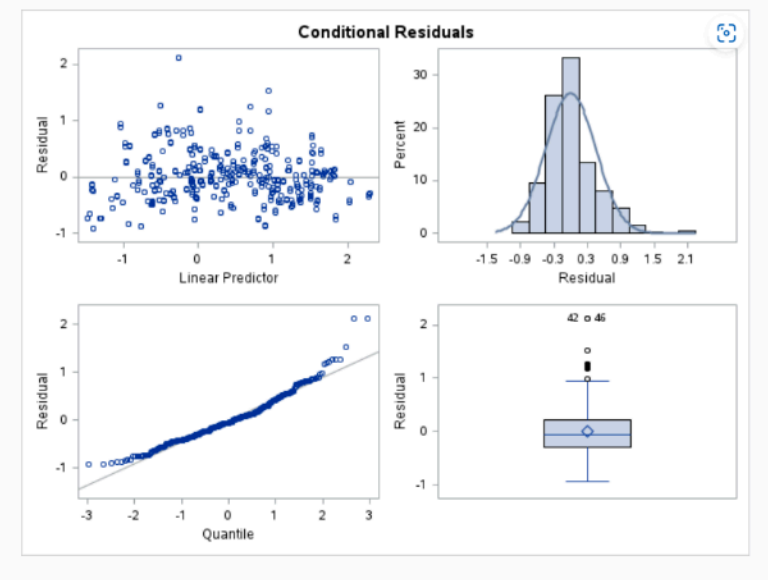


Figure 16: Residual plots and statistics for GLMM.

Conclusion

After considering several models, we have presented the three above to walk through our process of selecting a model. We began with a straightforward linear mixed model. We then adjusted some variables and got the linear mixed model that nested time within the day period. Finally, we used a GLMM to fit the highly skewed response variable better.

In the end, we believe the generalized linear mixed model with time nested within the day period is the best model. It satisfies the assumptions needed for a GLMM and performs at a level we are comfortable with. While this model does come with some risks in terms of complexity, we believe it is worth it in this case.

Recommendation

The significant tissue type by treatment interaction highlights the need for tissue-specific analysis in drought studies, as different tissue types respond uniquely to environmental stress. The client should focus on high-starch tissues (LM and UM), conducting separate analyses under various water conditions to better understand drought's impact on starch allocation and plant energy reserves. In summary, LM and UM tissues, with high starch levels under control conditions and significant reductions under drought, should be prioritized in drought management and resilience research.

References

Casella, George, and Roge Berger. 2001. *Statistical Inference*. Duxbury Resource Center. https://mybiostats.wordpress.com/wp-content/uploads/2015/03/casella-berger.pdf.

Hohenstein, Sven. 2018. "How to Find Alpha and Beta from a Gamma Distribution?" Cross Validated. https://stats.stackexchange.com/q/342644.

Issa, Marie-Anne, and Kevin L. Nadal. 2011. "Homoscedasticity." In *Encyclopedia of Child Behavior and Development*, edited by Sam Goldstein and Jack A. Naglieri, 752–52. Boston, MA: Springer US. https://doi.org/10.1007/978-0-387-79061-9_1382.

Lane, David Mark. 2010. "Tukey's Honestly Significant Difference (HSD)." *Encyclopedia of Research Design*. https://doi.org/https://doi.org/10.4135/9781412961288.n478.

Newsom. 2021. "Generalized Linear Models." *Psy 525/625 Categorical Data Analysis*. https://web.pdx.edu/~newsomj/cdaclass/ho_glm.pdf.

Slavkovic, Aleksandra. n.d. *Analysis of Discrete Data*. Penn State University. https://online.stat.psu.edu/stat504/lesson/6/6.1.

Appendix A - R Code

```
data <- read.csv("data.csv")

num_unique_tissu <- length(unique(data$tissu))
num_unique_tissu
num_unique_time <- length(unique(data$time))
num_unique_time
time_counts <- table(data$time)
time_counts
num_unique_dp <- length(unique(data$dayPeriod))
num_unique_dp
table(data$campagne)
num_unique_camp <- length(unique(data$campagne))


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


### Summary Statistics

library(knitr)
library(dplyr)

overall_summary <- data %>%
  summarize(
    Mean = mean(StarchNscTissue, na.rm = TRUE),
    Median = median(StarchNscTissue, na.rm = TRUE),
    SD = sd(StarchNscTissue, na.rm = TRUE),
    Min = min(StarchNscTissue, na.rm = TRUE),
    Max = max(StarchNscTissue, na.rm = TRUE),
    N = n()
  ) %>%
  mutate(Group = "Overall")

# By Location (campagne)
location_summary <- data %>%
  group_by(campagne) %>%
  summarize(
```

```
    Mean = mean(StarchNscTissue, na.rm = TRUE),
    Median = median(StarchNscTissue, na.rm = TRUE),
    SD = sd(StarchNscTissue, na.rm = TRUE),
    Min = min(StarchNscTissue, na.rm = TRUE),
    Max = max(StarchNscTissue, na.rm = TRUE),
    N = n()
  ) %>%
  mutate(Group = paste("campagne:", campagne))

# By DayPeriod
dayperiod_summary <- data %>%
  group_by(dayPeriod) %>%
  summarize(
    Mean = mean(StarchNscTissue, na.rm = TRUE),
    Median = median(StarchNscTissue, na.rm = TRUE),
    SD = sd(StarchNscTissue, na.rm = TRUE),
    Min = min(StarchNscTissue, na.rm = TRUE),
    Max = max(StarchNscTissue, na.rm = TRUE),
    N = n()
  ) %>%
  mutate(Group = paste("dayPeriod:", dayPeriod))

# Combine tables
combined_summary <- bind_rows(overall_summary, location_summary,
                              dayperiod_summary) %>%
  arrange(factor(Group, levels = c("Overall",
                                   unique(location_summary$Group),
                                   unique(dayperiod_summary$Group)))) %>%
  select(Group, N, Mean, Median, SD, Min, Max)

# Display the table
kable(combined_summary, format = "markdown",
      caption = "Summary statistics of starch content.")


## Normality check
library(ggplot2)

ggplot(data, aes(x = StarchNscTissue)) +
  geom_histogram(binwidth = 0.5, color = "black", fill = "skyblue") +
  labs(
    x = "Starch Content",
```

```
    y = "Frequency"
  ) +
  theme_minimal()

ggplot(data, aes(sample = StarchNscTissue)) +
  stat_qq() +
  stat_qq_line(color = "red") +
  labs(
    x = "Theoretical Quantiles",
    y = "Sample Quantiles"
  ) +
  theme_minimal()


ggplot(data, aes(x = tissu, y = StarchNscTissue)) +
  geom_boxplot() +
  labs(y = "Starch Content", x="Tissue Type") +
  theme_minimal()

ggplot(data, aes(x = tissu, y = StarchNscTissue, fill = treatment)) +
  geom_boxplot() +
  facet_wrap(~ dayPeriod) +
  labs(
    x = "Tissue Type",
    y = "Starch Content"
  ) +
  theme_minimal()


ggplot(data, aes(x = tissu, y = StarchNscTissue, fill = treatment)) +
  geom_bar(stat = "identity", position = "dodge") +
  facet_wrap(~ dayPeriod) +
  labs(
    x = "Tissue Type",
    y = "Starch Content"
  ) +
  theme_minimal()


starchmean <- mean(data$StarchNscTissue)
st_var <- var(data$StarchNscTissue)
```

```r
## Gamma(a,b)
# E(X) = a/b, Var(X)= a/(b^2)
# --> a = E(X)b --> Var(X)=E(X)/b
# --> b = E(X)/Var(X)
# --> a = [E(X)^2]/Var(X)
a <- starchmean^2/st_var
b <- starchmean/st_var


ggplot(data, aes(x = StarchNscTissue)) +
  geom_histogram(aes(y = ..density..), binwidth = 0.5,
                 color = "black", fill = "skyblue") +
  stat_function(fun = dgamma,
                args = list(shape = a, rate = b),
                color = "red", size = 1) +
  labs(
    x = "Starch Content",
    y = "Density"
  ) +
  theme_minimal()
```

Appendix B - SAS Code

```
/* Reading in csv file */
FILENAME REFFILE '<enter your file path';

PROC IMPORT DATAFILE=REFFILE
    DBMS=CSV
    OUT=data;
    GETNAMES=YES;
RUN;


/* Mixed Model*/
proc mixed data=data method=reml plots=(residualpanel);
    class treatment tissu dayPeriod campagne chamber time;
    model StarchNscTissue = treatment | tissu | dayPeriod;
    random campagne chamber sample;
    lsmeans treatment dayPeriod tissu / pdiff=all cl adjust=tukey;
run;

/* Hierarchial Nested Model*/
proc mixed data=data plots=(residualpanel);
    class campagne treatment dayPeriod tissu time sample chamber;
    model StarchNscTissue = treatment | dayPeriod(time) | tissu / ddfm=kr;
    lsmeans treatment dayPeriod(time) tissu/ pdiff=all cl adjust=tukey;
    random sample campagne chamber;
run;



/* GLMM Model */
proc glimmix data=data method=laplace plots=(residualpanel);
    class campagne treatment dayPeriod tissu time sample chamber;
    model StarchNscTissue = treatment | dayPeriod(time) | tissu / dist=gamma;
    random sample campagne chamber;
    lsmeans treatment dayPeriod(time) tissu / pdiff=all cl adjust=tukey;
run;
```

Appendix C - Additional SAS Output

Full *Least Squares Means* and *Differences of treatment Least Squares Means* tables for each fixed effect in GLMM.

treatment Least Squares Means

treatment	Estimate	Standard Error	DF	t Value	Pr > |t|	Alpha	Lower	Upper
Control	0.4224	0.2220	354	1.90	0.0578	0.05	-0.01413	0.8590
Drought	0.1843	0.2221	354	0.83	0.4072	0.05	-0.2526	0.6212

Differences of treatment Least Squares Means
Adjustment for Multiple Comparisons: Tukey-Kramer

treatment	_treatment	Estimate	Standard Error	DF	t Value	Pr > |t|	Adj P	Alpha	Lower	Upper	Adj Lower	Adj Upper
Control	Drought	0.2381	0.2846	354	0.84	0.4034	0.4034	0.05	-0.3217	0.7979	-0.3217	0.7979

dayPeriod(time) Least Squares Means

dayPeriod	time	Estimate	Standard Error	DF	t Value	Pr > |t|	Alpha	Lower	Upper
Day	A	0.1857	0.1759	354	1.06	0.2918	0.05	-0.1602	0.5316
Day	B	0.2468	0.1798	354	1.37	0.1708	0.05	-0.1069	0.6005
Night	C	0.3641	0.1791	354	2.03	0.0428	0.05	0.01194	0.7163
Night	D	0.3179	0.1790	354	1.78	0.0767	0.05	-0.03427	0.6700
Day	E	0.2691	0.1791	354	1.50	0.1338	0.05	-0.08307	0.6213
Day	F	0.4367	0.1801	354	2.42	0.0158	0.05	0.08247	0.7908

Differences of dayPeriod(time) Least Squares Means
Adjustment for Multiple Comparisons: Tukey-Kramer

dayPeriod	time	_dayPeriod	_time	Estimate	Standard Error	DF	t Value	Pr > |t|	Adj P	Alpha	Lower	Upper	Adj Lower	Adj Upper
Day	A	Day	B	-0.06112	0.07920	354	-0.77	0.4408	0.9721	0.05	-0.2169	0.09464	-0.2881	0.1658
Day	A	Night	C	-0.1784	0.07781	354	-2.29	0.0224	0.1997	0.05	-0.3314	-0.02538	-0.4014	0.04455
Day	A	Night	D	-0.1321	0.07740	354	-1.71	0.0887	0.5280	0.05	-0.2844	0.02009	-0.3539	0.08966
Day	A	Day	E	-0.08342	0.07780	354	-1.07	0.2843	0.8922	0.05	-0.2364	0.06959	-0.3064	0.1395
Day	A	Day	F	-0.2509	0.08039	354	-3.12	0.0019	0.0237	0.05	-0.4091	-0.09284	-0.4813	-0.02060
Day	B	Night	C	-0.1173	0.08641	354	-1.36	0.1755	0.7524	0.05	-0.2872	0.05265	-0.3649	0.1303
Day	B	Night	D	-0.07102	0.08625	354	-0.82	0.4108	0.9631	0.05	-0.2406	0.09860	-0.3182	0.1761
Day	B	Day	E	-0.02230	0.08620	354	-0.26	0.7960	0.9998	0.05	-0.1918	0.1472	-0.2693	0.2247
Day	B	Day	F	-0.1898	0.08861	354	-2.14	0.0328	0.2680	0.05	-0.3641	-0.01557	-0.4437	0.06406
Night	C	Night	D	0.04626	0.08465	354	0.55	0.5851	0.9942	0.05	-0.1202	0.2127	-0.1963	0.2888
Night	C	Day	E	0.09498	0.08473	354	1.12	0.2631	0.8725	0.05	-0.07166	0.2616	-0.1478	0.3378
Night	C	Day	F	-0.07255	0.08745	354	-0.83	0.4073	0.9619	0.05	-0.2445	0.09943	-0.3231	0.1780
Night	D	Day	E	0.04872	0.08474	354	0.57	0.5657	0.9926	0.05	-0.1179	0.2154	-0.1941	0.2915
Night	D	Day	F	-0.1188	0.08704	354	-1.36	0.1731	0.7480	0.05	-0.2900	0.05237	-0.3682	0.1306
Day	E	Day	F	-0.1675	0.08707	354	-1.92	0.0552	0.3891	0.05	-0.3388	0.003714	-0.4170	0.08196

tissu Least Squares Means								
tissu	Estimate	Standard Error	DF	t Value	Pr > |t|	Alpha	Lower	Upper
END	-0.3989	0.1755	354	-2.27	0.0236	0.05	-0.7440	-0.05379
IT	-0.4602	0.1758	354	-2.62	0.0092	0.05	-0.8059	-0.1146
LM	1.0761	0.1756	354	6.13	<.0001	0.05	0.7308	1.4215
UM	0.9965	0.1756	354	5.67	<.0001	0.05	0.6512	1.3419

Differences of tissu Least Squares Means Adjustment for Multiple Comparisons: Tukey-Kramer												
tissu	_tissu	Estimate	Standard Error	DF	t Value	Pr > |t|	Adj P	Alpha	Lower	Upper	Adj Lower	Adj Upper
END	IT	0.06137	0.06848	354	0.90	0.3708	0.8069	0.05	-0.07332	0.1961	-0.1154	0.2381
END	LM	-1.4750	0.06867	354	-21.48	<.0001	<.0001	0.05	-1.6101	-1.3400	-1.6523	-1.2978
END	UM	-1.3954	0.06871	354	-20.31	<.0001	<.0001	0.05	-1.5306	-1.2603	-1.5728	-1.2180
IT	LM	-1.5364	0.06988	354	-21.99	<.0001	<.0001	0.05	-1.6738	-1.3990	-1.7167	-1.3560
IT	UM	-1.4568	0.06976	354	-20.88	<.0001	<.0001	0.05	-1.5940	-1.3196	-1.6369	-1.2767
LM	UM	0.07959	0.06804	354	1.17	0.2429	0.6463	0.05	-0.05422	0.2134	-0.09604	0.2552