Data Analysis 2

Maksuda Aktar Toma, Jo Charbonneau, Ryan Lalicker

November 3, 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. (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 which can be found at 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.

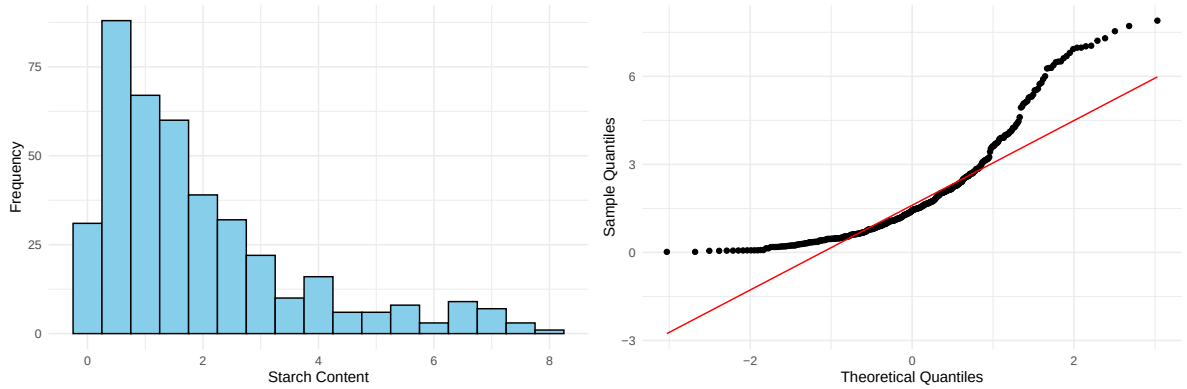
For starch contents across all measurements, the values range from about 0.019 to 7.898 with a median of roughly 1.430 and a mean of 1.925. The location of the median and mean with respect to the minimum and maximum is an early sign that the starch contents could be skewed and thus non-normal in distribution.

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.

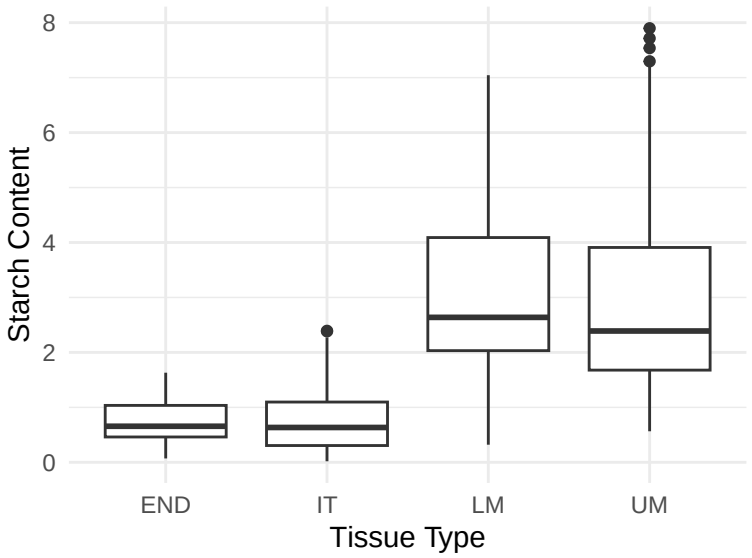


Figure 3: Boxplot of starch contents by tissue types

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.

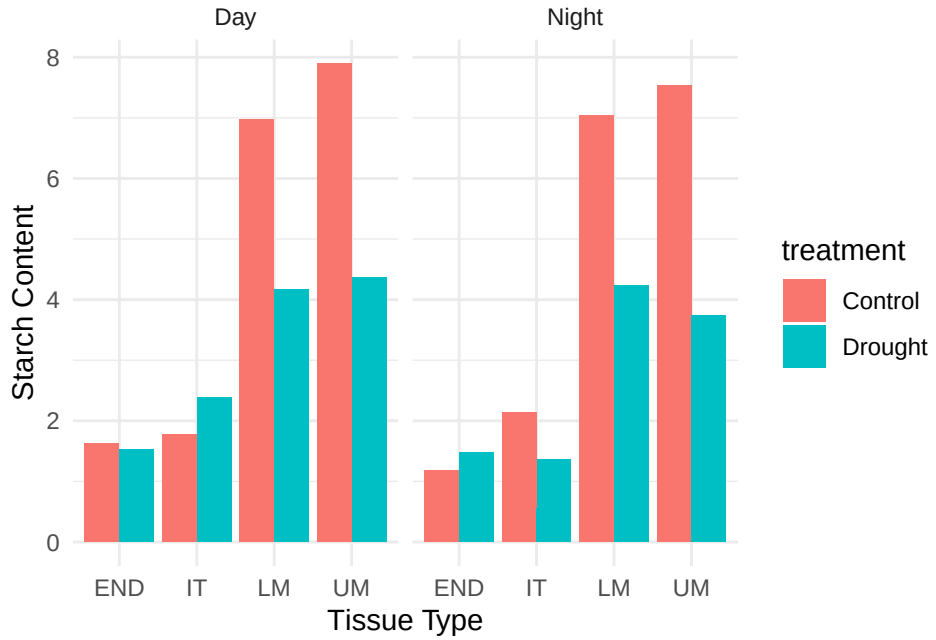


Figure 4: Barchat of starch content vs. tissue types, separating by treatment and day or night.

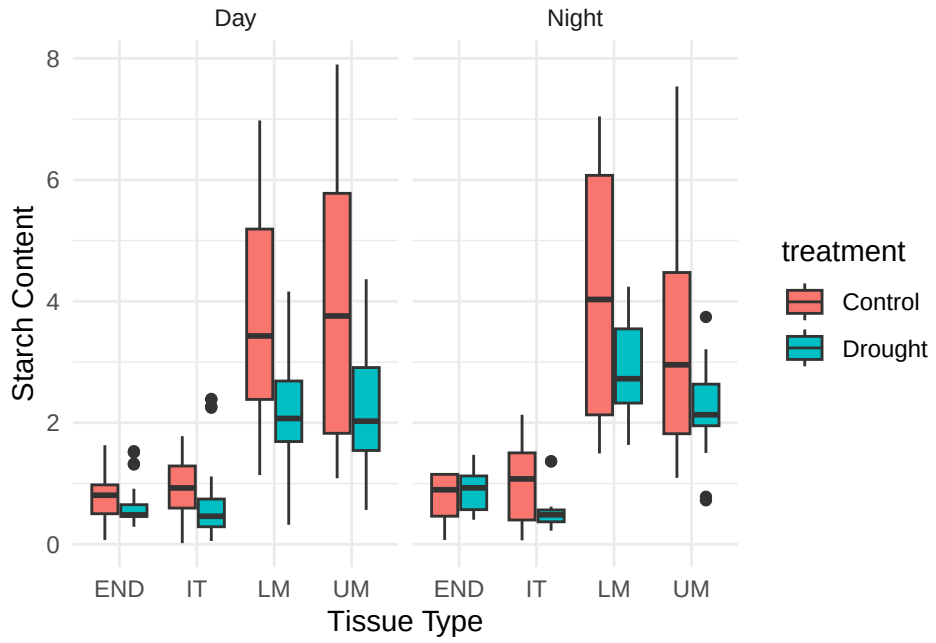


Figure 5: Boxplot of Starch Content by Tissue Type and Treatment.

In Figure 4 above 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. Once again, though, there does not seem to be a clear difference between day and night across all 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.

Model 1 : Mixed Effects 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.

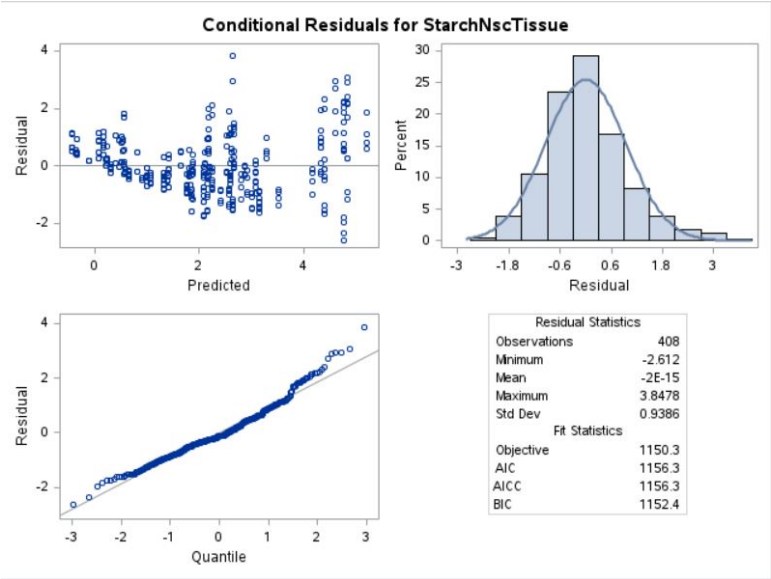


Figure 9: Residual plots and statistics for first proposed model.

While one could argue the homoscedasticity assumption is not definitely 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 `chamber`, `sample`, and `campagne`. Since each `chamber` represents the location of a certain `sample` and each `sample` is contained within a particular `campagne` we can say `chamber` is nested within `sample` which is nested within `campagne`. In the model below, u_l is once again the random effect for `campagne`, but $v_{m(l)}$ is the random effect of `sample` nested within `campagne` while $w_{n(l,m)}$ is the random effect of `chamber` nested within each `sample` within each `campagne`.

$$y_{ijklmn} = \mu + \tau_i + \alpha_j + \beta_k + (\tau\alpha)_{ij} + (\tau\beta)_{ik} + (\alpha\beta)_{jk} + (\tau\alpha\beta)_{ijk} + u_l + v_{m(l)} + w_{n(l,m)} + \epsilon_{ijklmn}$$

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 0.5207. Additionally the variance of `sample` nested within `campagne` has an estimated variance of 0.2477, which means both of these affect the starch content. The other nested structure seems to have little effect though.

Covariance Parameter Estimates	
Cov Parm	Estimate
campagne	0.5207
sample(campagne)	0.2477
chamb(campag*sample)	0.000819
Residual	0.9277

Fit Statistics	
-2 Res Log Likelihood	1151.9
AIC (Smaller is Better)	1159.9
AICC (Smaller is Better)	1160.0
BIC (Smaller is Better)	1154.6

Type 3 Tests of Fixed Effects				
Effect	Num DF	Den DF	F Value	Pr > F
treatment	1	386	4.38	0.0371
tissu	3	386	172.72	<.0001
treatment*tissu	3	386	13.06	<.0001
dayPeriod	1	386	2.93	0.0877
treatment*dayPeriod	1	386	0.17	0.6823
tissu*dayPeriod	3	386	2.14	0.0950
treatm*tissu*dayPeri	3	386	0.45	0.7153

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 larger than what we saw with the previous model. This could mean the nested approach is a slightly worse fit than before. The *Type 3 Tests of Fixed Effects* table shows which fixed effects and interactions are significant. The results are similar to before with all effects showing similar p-values. Only the treatment effect saw a slight increase in the p-value, but it is still significant at the 5% level. In the end, all fixed and interaction effects are significant at the same level as before. Once again it seems the treatment and tissue type are the primary factors in determining starch content.

Now let's consider the *Least Squares Means* table below. While some of the estimates have changed, with none being negative this time, we can see the only terms that saw a substantial

change in their p-values are the END and IT tissue types along with the drought effect. While the tissue types are still insignificant despite the decrease, drought has gone from marginally significant to significant at the 1% level. Note the order of effects is slightly different than in the first proposed model.

Least Squares Means

Effect	treatment	tissu	dayPeriod	Estimate	Standard Error	DF	t Value	Pr > |t|	Alpha	Lower	Upper
treatment	Control			2.2854	0.5725	386	3.99	<.0001	0.05	1.1598	3.4110
treatment	Drought			1.5169	0.5726	386	2.65	0.0084	0.05	0.3911	2.6427
tissu		END		0.7296	0.5496	386	1.33	0.1851	0.05	-0.3509	1.8101
tissu		IT		0.7420	0.5496	386	1.35	0.1778	0.05	-0.3385	1.8225
tissu		LM		3.2097	0.5496	386	5.84	<.0001	0.05	2.1292	4.2902
tissu		UM		2.9234	0.5496	386	5.32	<.0001	0.05	1.8429	4.0039
dayPeriod			Day	1.8129	0.5430	386	3.34	0.0009	0.05	0.7454	2.8805
dayPeriod			Night	1.9894	0.5465	386	3.64	0.0003	0.05	0.9149	3.0639

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

The *Differences of Least Squares Means* below follows the trend seen in the previous tables. Some estimates are slightly different than in the first model, but the adjusted p-values for multiple comparisons are significant at the same level as before.

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			0.7685	0.3673	386	2.09	0.0371	Tukey-Kramer	0.0371	0.05	0.04639	1.4905	0.04639	1.4905
tissu		END			IT		-0.01234	0.1453	386	-0.08	0.9324	Tukey-Kramer	0.9998	0.05	-0.2981	0.2734	-0.3874	0.3627
tissu		END			LM		-2.4800	0.1453	386	-17.06	<.0001	Tukey-Kramer	<.0001	0.05	-2.7658	-2.1943	-2.8551	-2.1050
tissu		END			UM		-2.1938	0.1453	386	-15.09	<.0001	Tukey-Kramer	<.0001	0.05	-2.4795	-1.9080	-2.5688	-1.8187
tissu		IT			LM		-2.4677	0.1453	386	-16.98	<.0001	Tukey-Kramer	<.0001	0.05	-2.7535	-2.1819	-2.8427	-2.0927
tissu		IT			UM		-2.1814	0.1453	386	-15.01	<.0001	Tukey-Kramer	<.0001	0.05	-2.4672	-1.8956	-2.5565	-1.8064
tissu		LM			UM		0.2863	0.1453	386	1.97	0.0496	Tukey-Kramer	0.2013	0.05	0.000505	0.5721	-0.08876	0.6613
dayPeriod			Day			Night	-0.1764	0.1031	386	-1.71	0.0877	Tukey-Kramer	0.0877	0.05	-0.3791	0.02622	-0.3791	0.02622

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.

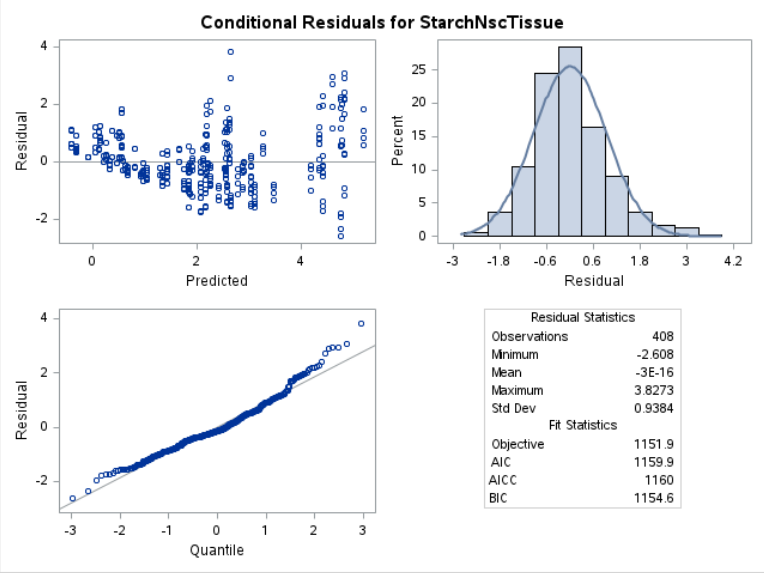


Figure 13: Residual plots and statistics for nested model.

We have seen the nested model does not improve upon some of the potential problems of the first model. It appears to be very similar and even has worse metrics in some cases, such as the AIC. This casts doubt on using this model over the original linear mixed model.

GLMM Model

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.

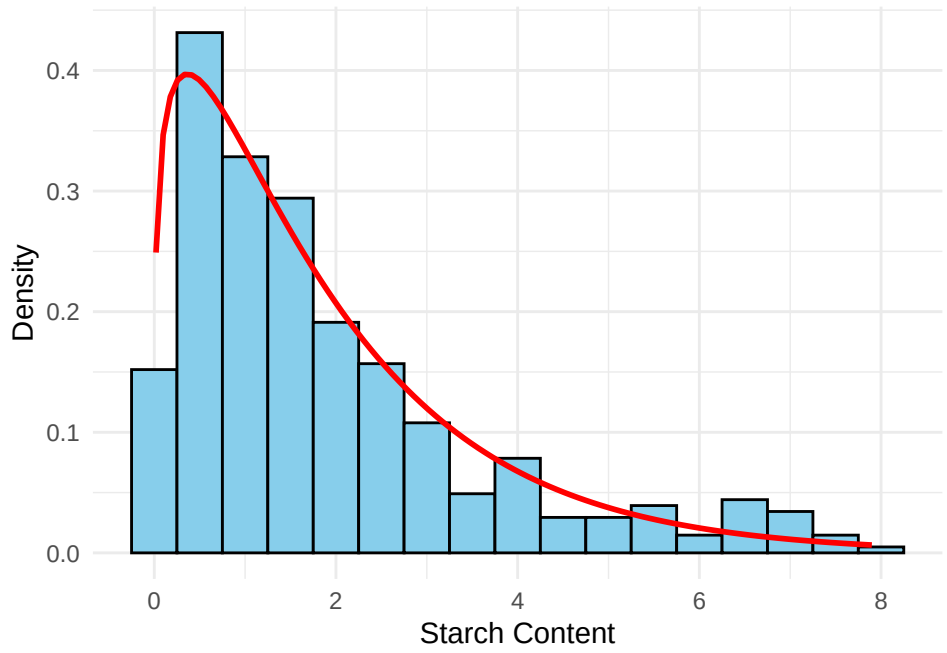


Figure 14: Histogram of starch content with overlay of gamma(1.2333, 0.6407) distribution.

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

This means that

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

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

$$\eta_{ijklmn} = \mu + \alpha_i + \tau_j + \beta_k + (\alpha\tau)_{ij} + (\alpha\beta)_{ik} + (\tau\beta)_{jk} + (\alpha\tau\beta)_{ijk} + u_l + v_m + w_n$$

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 η_{ijklmn}, not $E(y_{ijklmn})$.

Let's consider a part of the SAS output that is shown below. One potential concern is in the *Fit Statistics for Conditional Distribution* table. The value of 0.27 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 tissue type and the treatment by tissue type interaction are significant.

Fit Statistics	
-2 Log Likelihood	847.34
AIC (smaller is better)	887.34
AICC (smaller is better)	889.51
BIC (smaller is better)	861.20
CAIC (smaller is better)	881.20
HQIC (smaller is better)	832.68

Fit Statistics for Conditional Distribution	
-2 log L(StarchNscTissue | r. effects)	812.56
Pearson Chi-Square	108.96
Pearson Chi-Square / DF	0.27

Covariance Parameter Estimates		
Cov Parm	Estimate	Standard Error
campagne	0.008966	0.05038
sample	0.1116	0.08512
chamber	0.02969	.
Residual	0.2664	0.01805

Type III Tests of Fixed Effects				
Effect	Num DF	Den DF	F Value	Pr > F
tissu	3	386	217.81	<.0001
treatment	1	386	0.68	0.4095
tissu*treatment	3	386	4.60	0.0036
dayPeriod	1	386	0.90	0.3436
tissu*dayPeriod	3	386	1.67	0.1724
treatment*dayPeriod	1	386	1.62	0.2039
tissu*treatm*dayPeri	3	386	1.76	0.1551

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.

treatment Least Squares Means

treatment	Estimate	Standard Error	DF	t Value	Pr > |t|	Alpha	Lower	Upper
Control	0.4396	0.2033	386	2.16	0.0312	0.05	0.03987	0.8394
Drought	0.2154	0.2034	386	1.06	0.2902	0.05	-0.1845	0.6153

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	0.2242	0.2715	386	0.83	0.4095	0.4095	0.05	-0.3097	0.7581	-0.3097	0.7581

dayPeriod Least Squares Means

dayPeriod	Estimate	Standard Error	DF	t Value	Pr > |t|	Alpha	Lower	Upper
Day	0.3012	0.1521	386	1.98	0.0483	0.05	0.002198	0.6002
Night	0.3539	0.1557	386	2.27	0.0236	0.05	0.04768	0.6600

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

dayPeriod	_dayPeriod	Estimate	Standard Error	DF	t Value	Pr > |t|	Adj P	Alpha	Lower	Upper	Adj Lower	Adj Upper
Day	Night	-0.05266	0.05553	386	-0.95	0.3436	0.3436	0.05	-0.1618	0.05652	-0.1618	0.05652

tissu Least Squares Means

tissu	Estimate	Standard Error	DF	t Value	Pr > |t|	Alpha	Lower	Upper
END	-0.3593	0.1588	386	-2.26	0.0242	0.05	-0.6715	-0.04710
IT	-0.4292	0.1591	386	-2.70	0.0073	0.05	-0.7420	-0.1164
LM	1.1093	0.1589	386	6.98	<.0001	0.05	0.7969	1.4216
UM	0.9894	0.1589	386	6.23	<.0001	0.05	0.6770	1.3018

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	0.06992	0.07865	386	0.89	0.3746	0.8106	0.05	-0.08472	0.2246	-0.1330	0.2729
END	LM	-1.4686	0.07869	386	-18.66	<.0001	<.0001	0.05	-1.6233	-1.3139	-1.6716	-1.2656
END	UM	-1.3487	0.07885	386	-17.11	<.0001	<.0001	0.05	-1.5037	-1.1937	-1.5521	-1.1453
IT	LM	-1.5385	0.07968	386	-19.31	<.0001	<.0001	0.05	-1.6952	-1.3819	-1.7441	-1.3329
IT	UM	-1.4186	0.07958	386	-17.83	<.0001	<.0001	0.05	-1.5751	-1.2621	-1.6240	-1.2133
LM	UM	0.1199	0.07808	386	1.54	0.1255	0.4172	0.05	-0.03362	0.2734	-0.08156	0.3213

Interpretation:

Once again let's consider the plots generated by SAS for checking our assumptions. 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.

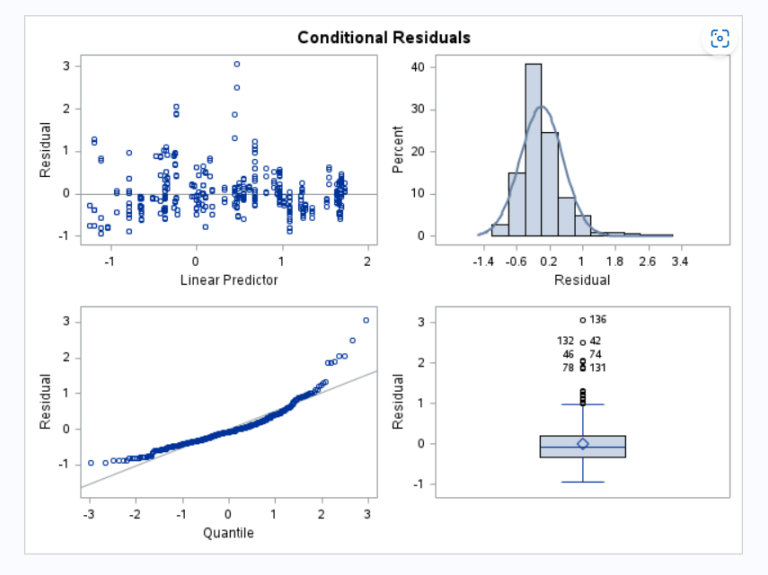


Figure 16: Residual plots and statistics for GLMM.

In this model we have seen the tissue type is highly significant. The interaction between it and the treatment is also a contributing factor in the model. Once again though, we have seen the period of the day is not significant. While the fit of the gamma distribution is not great, the improved AIC from the mixed model is promising.

Conclusion

we would like to fit the **Mixed Model** for this data set. As we can see from the fit statistics and diagnostic result, the mixed model gives us better fitting than the Nested and GLMM models. In the Hierarchical Nested Model, the AIC, BIC, and AICC are comparatively a little bit higher than the Mixed Model, and residual plots remain the same for both plots. Although the AIC, BICC, and AICC are lower in GLMM than in the Mixed Model, the assumptions hold better in the Mixed Model. So, it would be better to fit a **Mixed Model** to ignore unnecessary complexity in the model structure.

Summary

The study analyzed the effects of tissue type, treatment, and time of day on starch content in pine tissues across two locations, with minor data adjustments made for consistency. Exploratory analysis showed that LM and UM tissues had higher starch content

and control samples generally had higher values than drought samples, particularly in LM and UM. Three models were evaluated: Mixed Effects Model with Interactions, Hierarchical Nested Model, and GLMM. The Mixed Effects Model showed significant effects for tissue and treatment-tissue interaction, with residuals meeting normality assumptions better than the GLMM. Due to its balance of fit, interpretability, and simplicity, the Mixed Effects Model was recommended as the best approach.

Recommendation:
Since, the **significant Tissue*Treatment interaction** highlights the need for tissue-specific analysis in drought studies, as different tissue types respond uniquely to environmental stress. Future research 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 sample;
    model StarchNscTissue = treatment | tissu | dayPeriod;
    random campagne sample(campagne) chamber(sample*campagne);
    lsmeans treatment tissu dayPeriod / pdiff=all cl adjust=tukey;
run;

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


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