Data Analysis 2

Maksuda Aktar Toma, Jo Charbonneau, Ryan Lalicker

November 1, 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. 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

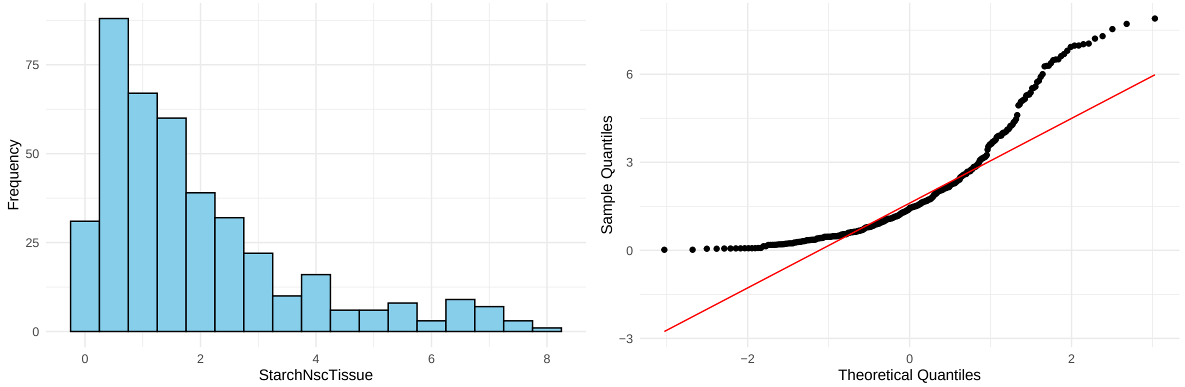
Figure 1: Summary statistics of starch content.

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

EXPLORE TRT AND TISSUE TYPES

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.

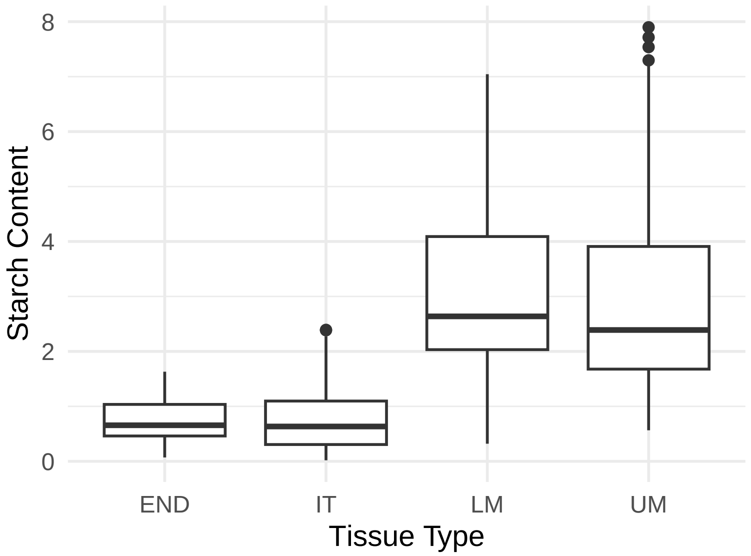


Figure 3: Boxplot of starch contents by tissue types

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.

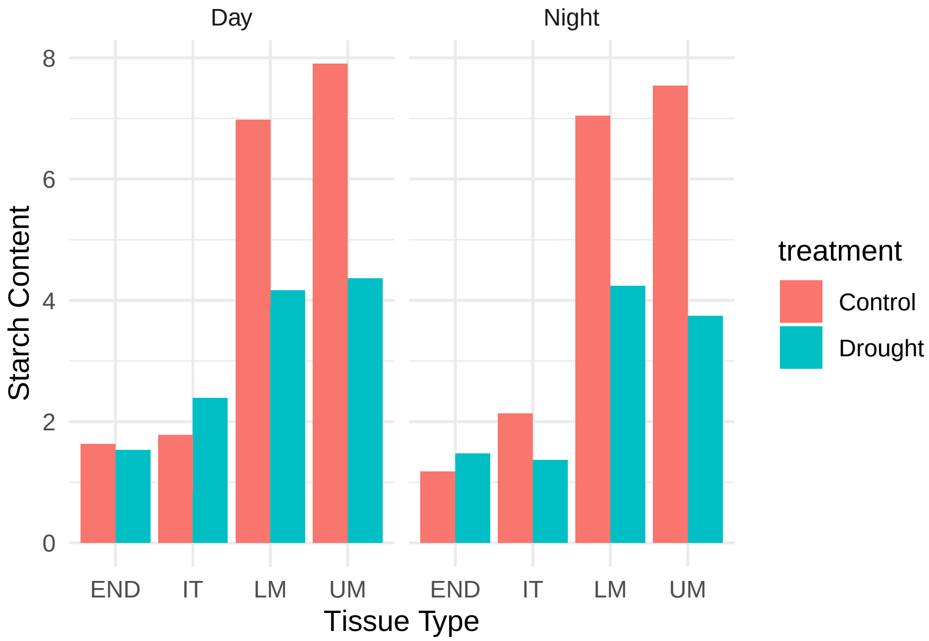


Figure 4: Barchat of starch content vs. tissue types, separating by treatment and day or night.

In the graph 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.

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. To use this, the residuals of the model must be approximately normally distributed.

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.)

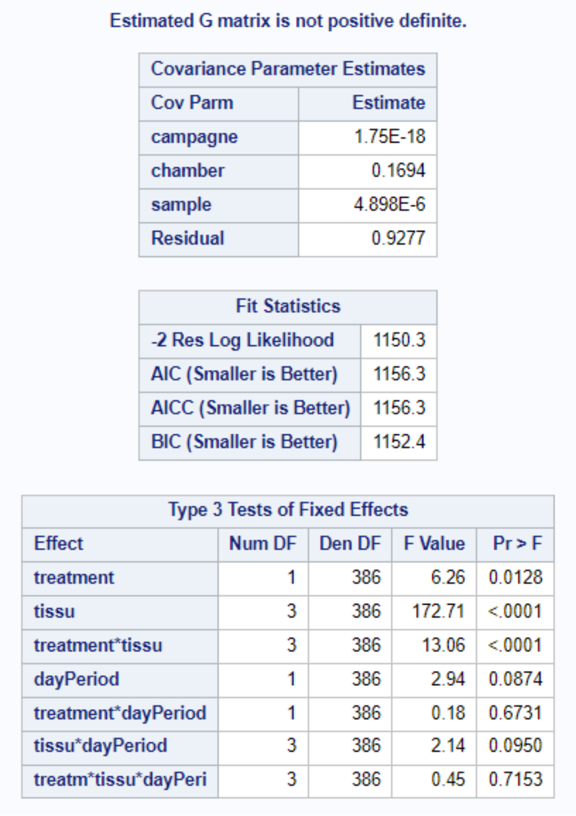
Model : Mixed Effects Model with Interactions

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

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

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

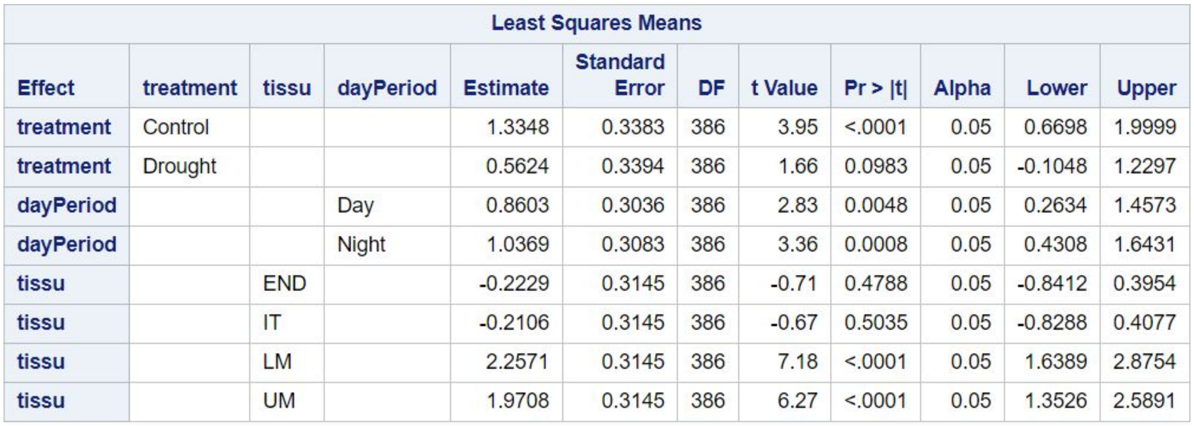
Fit Statistics	
-2 Res Log Likelihood	1150.3
AIC (Smaller is Better)	1156.3
AICC (Smaller is Better)	1156.3
BIC (Smaller is Better)	1152.4

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

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.

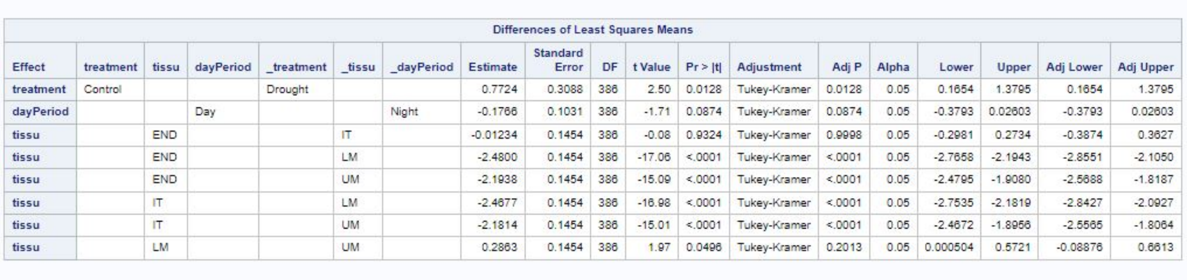


Least Squares Means											
Effect	treatment	tissu	dayPeriod	Estimate	Standard Error	DF	t Value	Pr > |t|	Alpha	Lower	Upper
treatment	Control			1.3348	0.3383	386	3.95	<.0001	0.05	0.6698	1.9999
treatment	Drought			0.5624	0.3394	386	1.66	0.0983	0.05	-0.1048	1.2297
dayPeriod			Day	0.8603	0.3036	386	2.83	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 6: *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 (((((CHECK THIS)))))). 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.

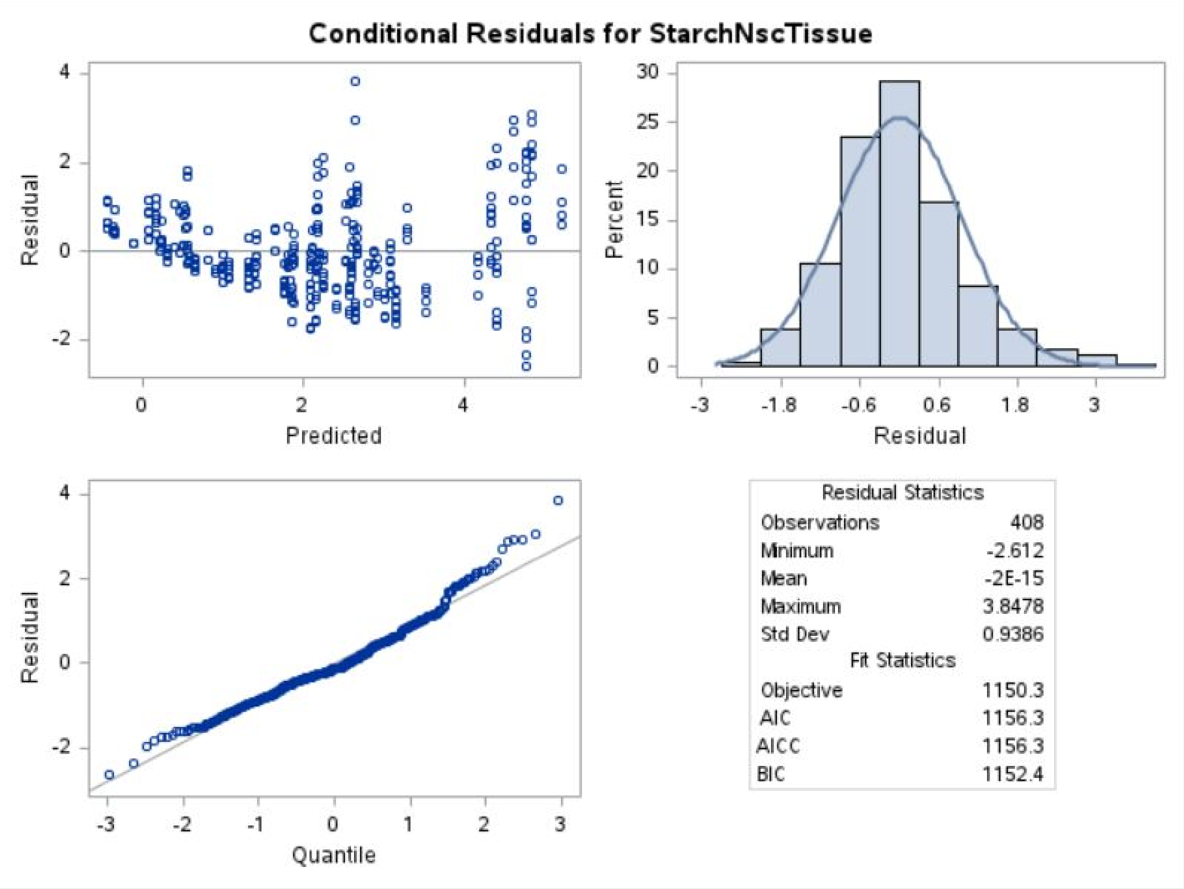


Differences of Least Squares Means																		
Effect	treatment	tissu	dayPeriod	_treatment	_tissu	_dayPeriod	Estimate	Standard Error	DF	t Value	Pr > |t|	Adjustment	Adj P	Alpha	Lower	Upper	Adj Lower	Adj Upper
treatment	Control			Drought			0.7724	0.3088	386	2.50	0.0128	Tukey-Kramer	0.0128	0.05	0.1654	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 7: *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.



Residual Statistics	
Observations	408
Minimum	-2.612
Mean	-2E-15
Maximum	3.8478
Std Dev	0.9386
Fit Statistics	
Objective	1150.3
AIC	1156.3
AICC	1156.3
BIC	1152.4

Figure 8: 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.

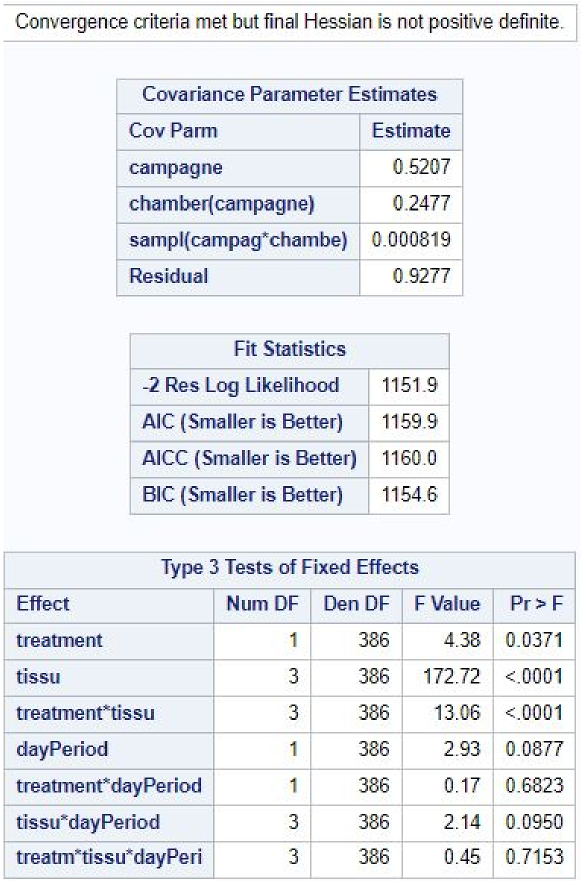
Where I (Ryan) stopped on 11/1.

Nested Model

CHAT GTP DOUBLE CHECK

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

Output



Convergence criteria met but final Hessian is not positive definite.

Covariance Parameter Estimates	
Cov Parm	Estimate
campagne	0.5207
chamber(campagne)	0.2477
sampl(campag*chambe)	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 9: Fig-1

Interpretation:

1. Covariance Parameter Estimates

- **campagne**: The estimated variance due to `campagne` is 0.5207, indicating that differences between locations (campagne) contribute to the overall variance in starch content.
- **chamber(campagne)**: The estimated variance due to `chamber` nested within `campagne` is 0.2477, suggesting that variation between chambers within each location also affects starch content.
- **sample(campagne*chamber)**: The estimated variance due to `sample` nested within `campagne` and `chamber` is 0.000819, which is relatively small, implying limited variability due to individual samples within chambers.
- **Residual**: The residual variance is 0.9277, which represents the unexplained variability after accounting for fixed effects and random effects.

2. Fit Statistics

The value of the AIC, AICC, and BIC are comparatively higher than the mixed effect model.

3. Type 3 Tests of Fixed Effects

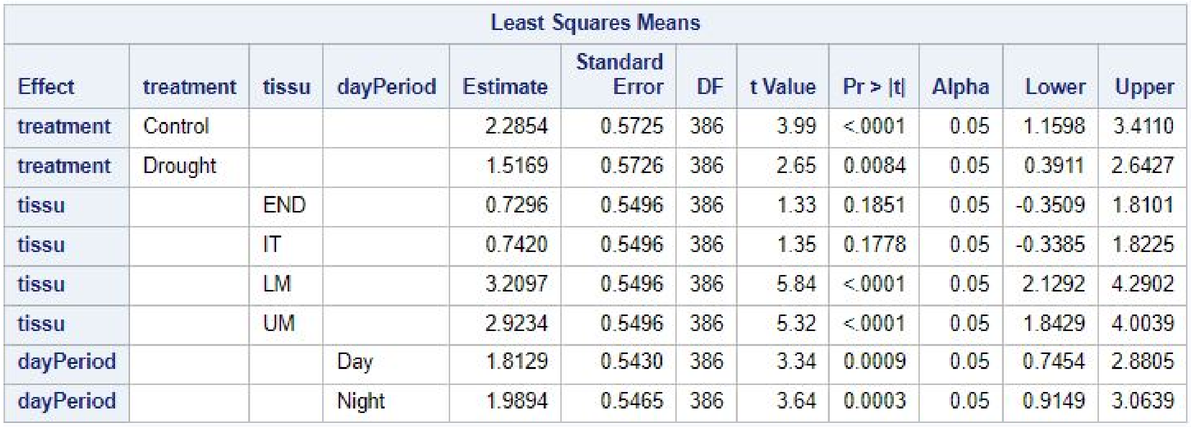
This table tests the significance of each fixed effect and their interactions.

- **Treatment** (`p = 0.0371`): Significant at the 0.05 level, indicating that the `treatment` effect (Control vs. Drought) has a statistically significant impact on starch content.
- **Tissu** (`p < 0.0001`): Highly significant, suggesting that different tissue types have a strong effect on starch content.
- **Treatment*Tissu Interaction** (`p < 0.0001`): Significant, indicating that the effect of treatment on starch content varies by tissue type.
- **DayPeriod** (`p = 0.0877`): Not significant at the 0.05 level, implying that the collection time (Day vs. Night) does not have a significant impact on starch content.
- **Treatment*DayPeriod Interaction** (`p = 0.6731`): Not significant, suggesting that there is no interaction between treatment and day period.
- **Tissu*DayPeriod Interaction** (`p = 0.0950`): Marginally significant, indicating a potential interaction between tissue type and day period, but it is not below the 0.05 significance level.
- **Treatment*Tissu*DayPeriod Interaction** (`p = 0.7153`): Not significant, indicating that there is no three-way interaction among treatment, tissue type, and day period.

Summary of Findings

1. **Significant Effects**: Treatment and tissue type are significant main effects, with a significant interaction between them. This suggests that starch content varies by treatment and tissue type, with the impact of treatment depending on the type of tissue.
2. **Non-Significant Effects**: DayPeriod does not have a significant effect, and there are no significant interactions involving DayPeriod with treatment or tissue type.
3. **Random Effects**: Variability due to `campagne` and `chamber` within `campagne` are notable, while the sample-level variability is minimal.

In conclusion, **treatment** and **tissue type** are the primary factors affecting starch content, with the interaction indicating that the effect of treatment depends on tissue type. The time of collection (Day vs. Night) does not significantly affect starch content in this 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

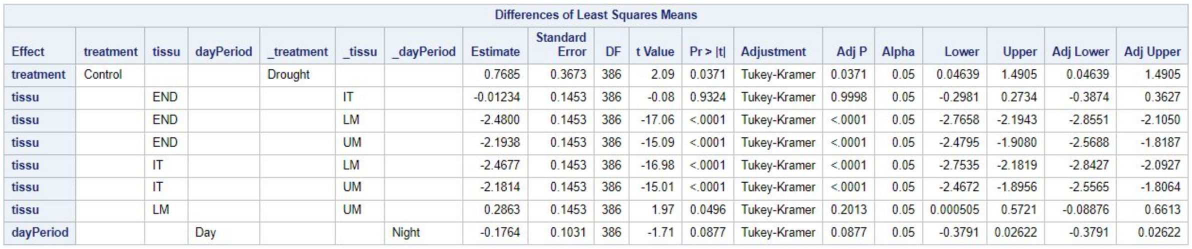
Figure 10: Fig-2

Interpretation:

Summary

1. **Treatment**: Control condition has higher starch content than Drought, and both are significantly different from zero.
2. **Tissue**: LM and UM tissues have significantly higher starch content compared to END and IT, which do not show significant starch content.
3. **DayPeriod**: Both Day and Night periods show significant starch content, with a slightly higher mean during the Night.

In summary, **treatment, tissue type, and collection time** all influence starch content, with the **Control treatment, LM and UM tissues, and Night period** showing higher values.



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

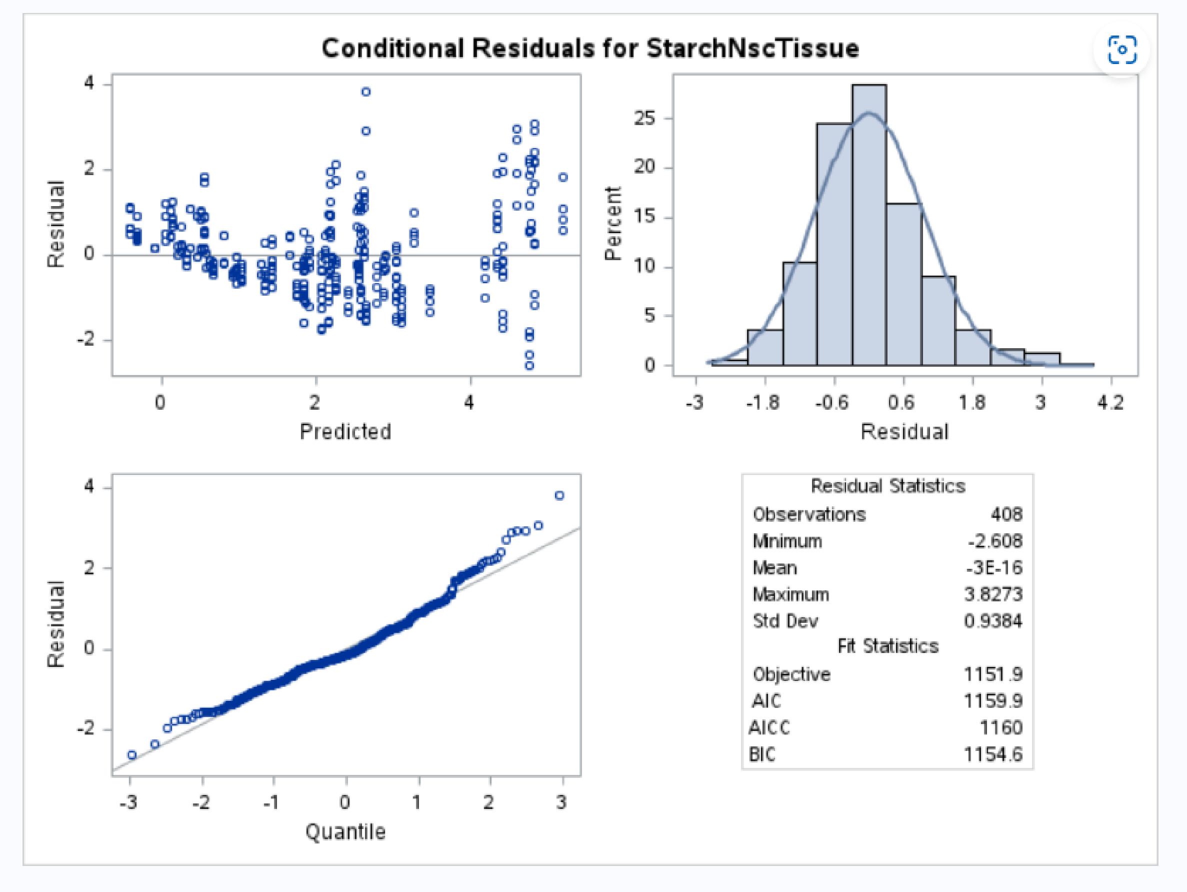
Figure 11: Fig-3

Interpretation:

Summary of Findings

1. **Treatment**: Control has a significantly higher starch content than Drought.
2. **Tissue**: LM and UM tissues have significantly higher starch content compared to END and IT tissues. However, there is no significant difference between END vs. IT or between LM vs. UM.
3. **DayPeriod**: No significant difference in starch content between Day and Night.

In summary, **treatment** and **tissue type** are the main drivers of differences in starch content, with **Control** and **LM/UM tissues** showing higher values. The **DayPeriod** does not significantly impact starch content



```
Residual Statistics
Observations        408
Minimum          -2.608
Mean              -3E-16
Maximum          3.8273
Std Dev          0.9384
Fit Statistics
Objective        1151.9
AIC              1159.9
AICC               1160
BIC              1154.6
```

Figure 12: Fig-4

##GLMM Model:

CHAT GTP DOUBLE CHECK

$$g(\mathbb{E}(y_{ijklm})) = \mu + \alpha_i + \tau_j + \beta_k + (\alpha\tau)_{ij} + (\alpha\beta)_{ik} + (\tau\beta)_{jk} + (\alpha\tau\beta)_{ijk} + u_l + v_m + w_n + \epsilon_{ijklm}$$

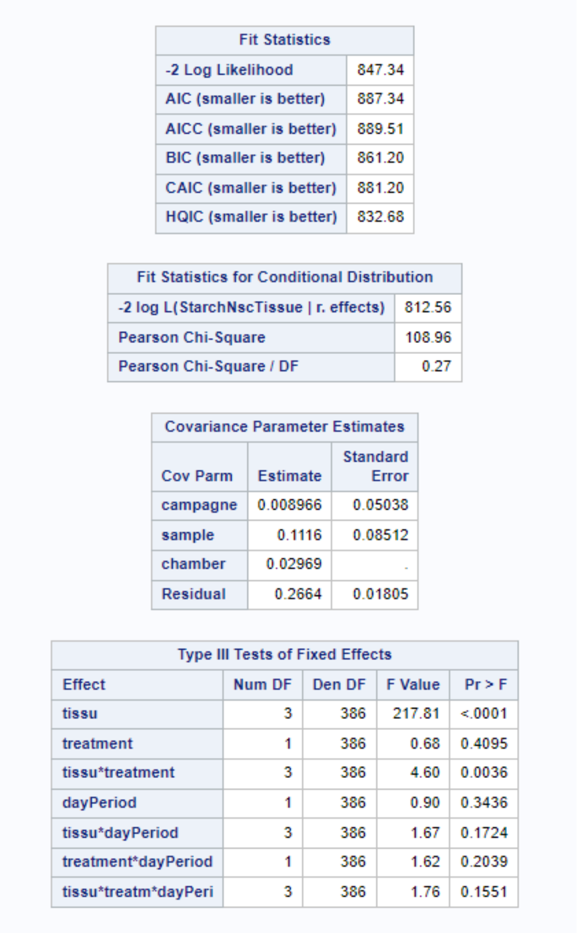
Summary

- **Normality**: Residuals are approximately normally distributed, as suggested by the histogram and Q-Q plot, though slight deviations are present in the tails.
- **Homoscedasticity**: There appears to be some heteroscedasticity, with residuals showing increasing variance for higher predicted values, as indicated in the residuals vs. predicted plot.

- **Model Fit**: The model appears reasonably well-fitted overall, though slight adjustments or transformations could be considered if the heteroscedasticity impacts model accuracy.

In conclusion, the model generally meets the assumptions of normality and homoscedasticity, but there are minor deviations that may warrant further investigation, particularly with the slight increase in residual variance at higher predicted values.

Output



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 13: Fig-1

Interpretation:

Here's an interpretation of each section in the provided output:

Fit Statistics

These statistics suggest that this model is a reasonable fit and can be compared with other models if needed to find the best balance of fit and simplicity.

Fit Statistics for Conditional Distribution

- **-2 log L(StarchNscTissue | r. effects)**: 812.56 – A measure of the fit of the conditional model, where lower values suggest better fit.
- **Pearson Chi-Square**: 108.96
- **Pearson Chi-Square / DF**: 0.27 – Values near 1 indicate a good fit. A value of 0.27 suggests possible overdispersion (less variation in residuals than expected under the model).

Covariance Parameter Estimates

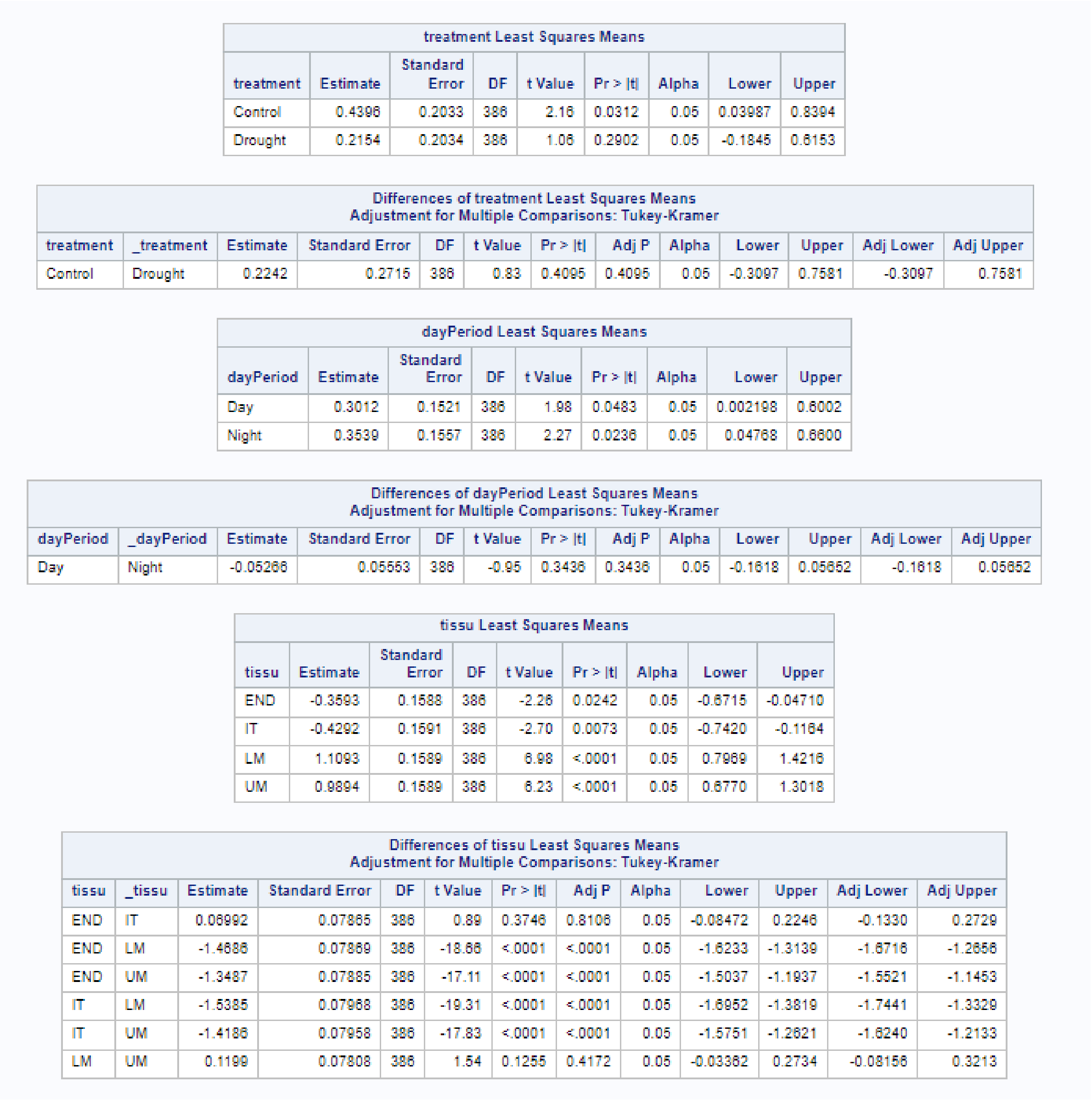
- **campagne**: Variance component of 0.008996, suggesting low variability attributed to differences between locations (campagne).
- **sample**: Variance component of 0.1116, indicating moderate variability between samples.
- **chamber**: Variance component of 0.02969, indicating minor variability between chambers.
- **Residual**: Variance component of 0.2664, representing the unexplained variability after accounting for the fixed effects and random effects.

The random effects `sample` and `chamber` show some variability, with `sample` contributing the most, whereas `campagne` has minimal variance. The residual variance is relatively small.

Summary of Findings

1. **Significant Effects**: Tissue type (`tissu`) has a strong effect on starch content, with a significant interaction between tissue type and treatment, meaning that the effect of treatment varies depending on the tissue type.
2. **Non-Significant Effects**: Treatment alone, day period, and most interactions involving day period do not significantly affect starch content.
3. **Random Effects**: The sample-level variance is notable, while location (`campagne`) and chamber-level variances are relatively small. The residual variance is moderate.

In summary, **tissue type is the primary factor** influencing starch content, with a significant interaction indicating that **treatment effects depend on the tissue type**. Day period and interactions involving day period are not significant in this model.



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:

Summary of Findings

1. **Treatment**: There is no significant difference between Control and Drought treatments on starch content, although the Control group alone shows a significant mean effect.
2. **DayPeriod**: Both Day and Night periods individually have significant effects, but there is no significant difference between them.
3. **Tissue (Tissu)**: LM and UM tissues have significantly higher starch content compared to END and IT. However, there is no significant difference between END vs. IT or between LM vs. UM.

In summary, **tissue type** is the primary factor influencing starch content, with **LM and UM showing higher values**. The **DayPeriod** and **Treatment** effects are individually significant, but the comparisons between levels do not show substantial differences.

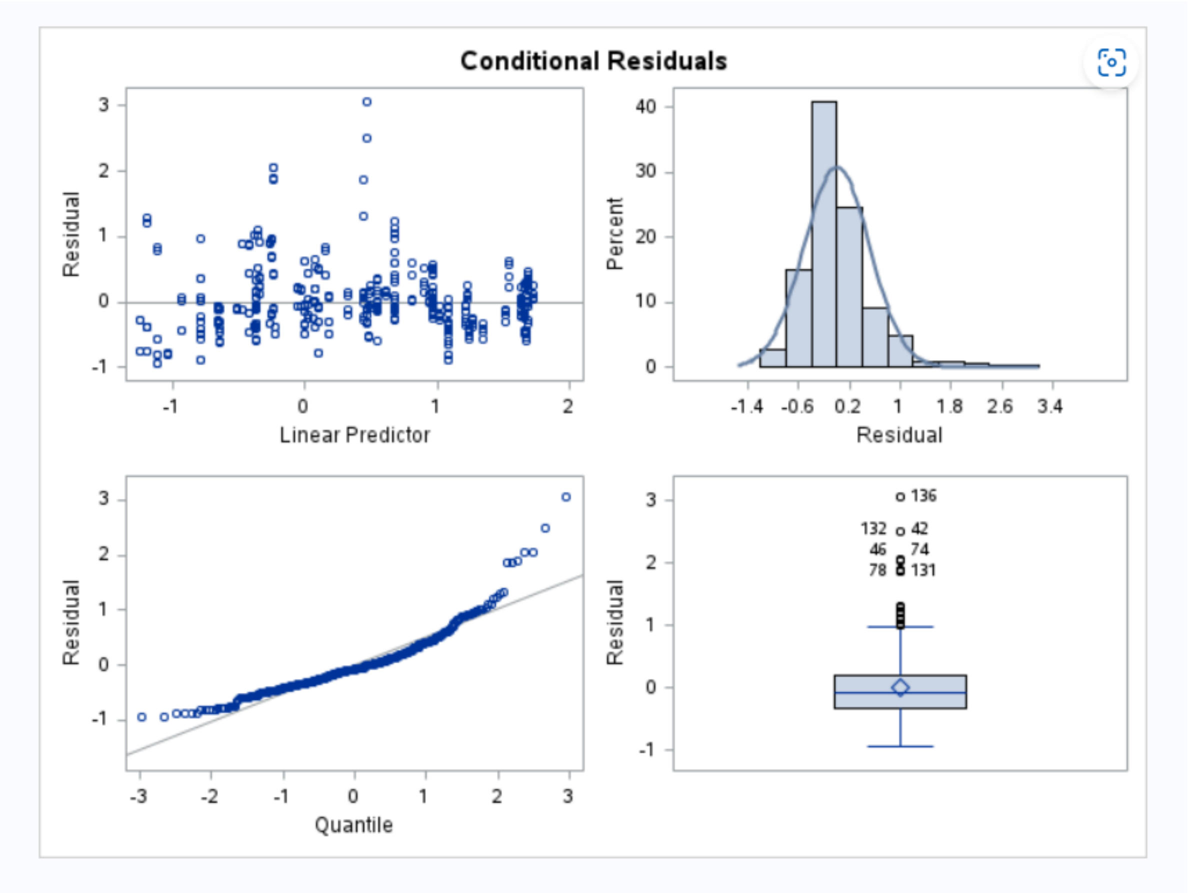


Figure 14: Fig-3

###Interpretation:

The **Conditional Residuals** plot provides diagnostic checks to evaluate the model's assumptions.

1. Residuals vs. Linear Predictor (Top Left)

This plot displays residuals against the linear predictor (fitted values). Ideally, residuals should be randomly scattered around zero with no discernible patterns.

- In this case, the residuals appear fairly well-scattered, but there is some minor clustering around zero, suggesting that the residuals are mostly unbiased but may have slight deviations. No obvious pattern indicates that the assumption of homoscedasticity (constant variance) is mostly met.

2. Histogram of Residuals (Top Right)

This histogram shows the distribution of residuals with an overlaid normal curve.

- The residuals appear approximately normally distributed, though there is some slight skewness, particularly on the right tail. This indicates that the normality assumption is reasonably met, but there may be a few outliers affecting the distribution.

3. Q-Q Plot of Residuals (Bottom Left)

The Q-Q plot compares the residuals to a theoretical normal distribution. Points should ideally lie along the straight line if the residuals are normally distributed.

- Most points fall along the line, indicating approximate normality, although there are deviations at the upper tail. This suggests that while most residuals are normally distributed, a few larger values deviate from normality, indicating possible outliers.

4. Boxplot of Residuals (Bottom Right)

The boxplot provides a summary of the residuals, showing the median, quartiles, and potential outliers.

- A few outliers are labeled and extend beyond the upper whisker. While the bulk of the residuals fall within a reasonable range, these outliers indicate that some data points do not fit the model as well as others.

Summary

- **Normality**: The residuals are approximately normally distributed, as indicated by the histogram and Q-Q plot, though there are minor deviations in the upper tail.
- **Homoscedasticity**: The residuals vs. linear predictor plot does not show any strong patterns, suggesting that the assumption of constant variance is reasonably met.
- **Outliers**: The boxplot and Q-Q plot show a few outliers, which may slightly affect the model fit but do not indicate severe violations of assumptions.

Overall, the model diagnostics suggest that the assumptions of normality and homoscedasticity are mostly met, with minor deviations due to a few outliers. The model appears to fit the data reasonably well, although addressing or investigating the outliers could further improve model performance.

##Best Model to fit Based on the above discussion, 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.

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.

Notes for US CHECK

Few notes- 1. Can we keep the model output from SAS but plots fro R? what do you think? 2. As tissue type and treatment are significant and we exactly know which one is mostly significant, can we make some plots or do anything else for this? 3. Can we do anything else for exploratory Analysis? 4. As Dayperiod is not significant wholly, we didn't do anything about time variable. what's your opinion on this?

Yet to be done- 1. Summary 2. Recommendation 3. References

Conclusion

GitHub page found here.

References

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

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.

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

```
/* 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;
    model StarchNscTissue = treatment|tissu|dayPeriod;
    lsmeans treatment dayPeriod tissu / pdiff=all cl adjust=tukey;
    random campagne chamber sample;
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;
```