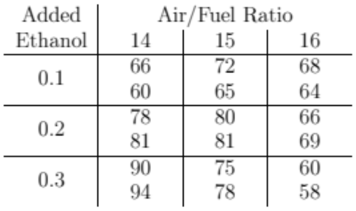
HW 8

2024-10-28

Stat 5000 Homework #8
Fall 2024 due Fri, November 1st @ 11:59 pm Name: Sam Olson
Collaborators: **The Hatman**

Q1

A completely randomized two-factor experiment consisted of burning fuel with levels of two additives in a laboratory setting and determining the carbon monoxide (CO) emissions released. Eighteen batches of a standard fuel were available for this study. Two of the batches were randomly assigned to each of nine combinations of two additives corresponding to three levels of added ethanol (0.1, 0.2, or 0.3) and three air/fuel ratio settings (14, 15, or 16). Units for the ethanol levels were not reported. CO emission concentrations (g/meter3) were determined by burning the same amount of fuel from each of the 18 batches. The data are shown below and are located in the file emissions.txt.



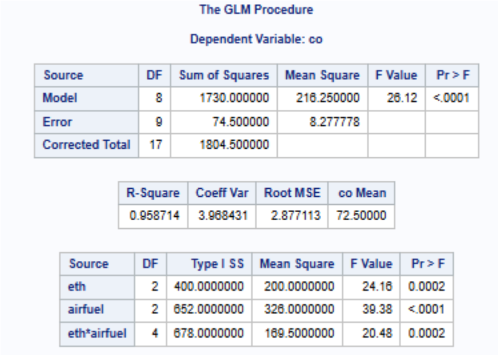
Added	Air/Fuel Ratio		
Ethanol	14	15	16
0.1	66	72	68
	60	65	64
0.2	78	80	66
	81	81	69
0.3	90	75	60
	94	78	58

Figure 1: CocoMelon

(a)

Construct the full ANOVA table. Which factors or interactions have significant effects on CO concentrations in emissions? Interpret the results in the context of the study.

It appears that all treatment variables (ethanol levels and air/fuel ratios) in addition to their interaction effects are significant, meaning we have evidence to reject the null hypothesis that the mean CO emission concentrations (g/meter3) are equal for all treatment levels when averaged across all other factors/treatments, i.e. we have evidence to support the following alternative hypotheses: 1. At least one mean CO emission concentrations (g/meter3) for ethanol levels is different from the other mean CO emission concentrations (g/meter3) averaging across air/fuel ratio levels, 2. At least one mean CO emission concentrations (g/meter3) for air/fuel ratio is different from the other mean CO emission concentrations (g/meter3) for air/fuel ratios when averaging across ethanol levels, and 3. The mean CO emission concentrations (g/meter3) for the interaction between ethanol and air/fuel ratio is different from the mean CO emission concentrations (g/meter3) of some other combination of ethanol/air/fuel ratio.



The GLM Procedure

Dependent Variable: co

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	8	1730.000000	216.250000	26.12	<.0001
Error	9	74.500000	8.277778		
Corrected Total	17	1804.500000			

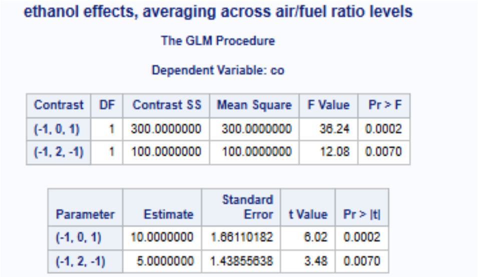
R-Square	Coeff Var	Root MSE	co Mean
0.958714	3.968431	2.877113	72.50000

Source	DF	Type I SS	Mean Square	F Value	Pr > F
eth	2	400.0000000	200.0000000	24.16	0.0002
airfuel	2	652.0000000	326.0000000	39.38	<.0001
eth*airfuel	4	678.0000000	169.5000000	20.48	0.0002

Figure 2: CocoMelon

(b)

Partition the sum of squares for the ethanol effects, averaging across air/fuel ratio levels, into sums of squares for linear and quadratic components. The coefficients for these contrasts are (-1, 0, 1) and (-1, 2, -1). Is there a significant linear or quadratic effect in the model for the ethanol effects?



ethanol effects, averaging across air/fuel ratio levels

The GLM Procedure

Dependent Variable: co

Contrast	DF	Contrast SS	Mean Square	F Value	Pr > F
(-1, 0, 1)	1	300.0000000	300.0000000	36.24	0.0002
(-1, 2, -1)	1	100.0000000	100.0000000	12.08	0.0070

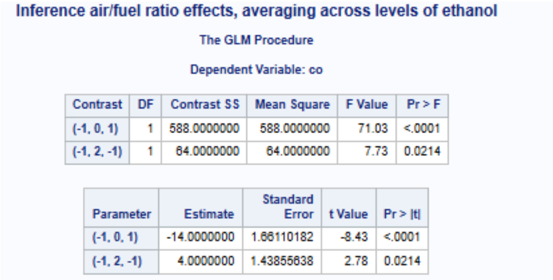
Parameter	Estimate	Standard Error	t Value	Pr > |t|
(-1, 0, 1)	10.0000000	1.66110182	6.02	0.0002
(-1, 2, -1)	5.0000000	1.43855638	3.48	0.0070

Figure 3: CocoMelon

There are significant linear and quadratic effects in the model for ethanol effects, where significance is at the $\alpha = 0.05$ level.

(c)

Partition the sum of squares for the air/fuel ratio effects, averaging across levels of ethanol, into sums of squares for linear and quadratic components. The coefficients for these contrasts are (-1, 0, 1) and (-1, 2, -1). Is there a significant linear or quadratic effect in the model for the air/fuel ratio effects?



Inference air/fuel ratio effects, averaging across levels of ethanol

The GLM Procedure

Dependent Variable: co

Contrast	DF	Contrast SS	Mean Square	F Value	Pr > F
(-1, 0, 1)	1	588.0000000	588.0000000	71.03	<.0001
(-1, 2, -1)	1	64.0000000	64.0000000	7.73	0.0214

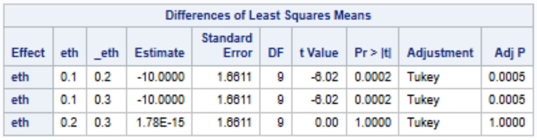
Parameter	Estimate	Standard Error	t Value	Pr > |t|
(-1, 0, 1)	-14.0000000	1.66110182	-8.43	<.0001
(-1, 2, -1)	4.0000000	1.43855638	2.78	0.0214

Figure 4: CocoMelon

There are also significant linear and quadratic effects in the model for the air/fuel ratio effects, where significance is at the $\alpha = 0.05$ level.

(d)

Use Tukey's HSD method to make pairwise comparisons of the marginal means for the three ethanol values. Summarize the results in the context of the study.



Differences of Least Squares Means

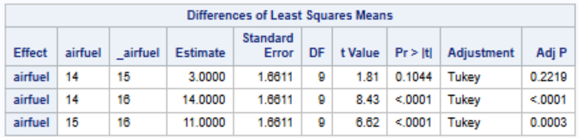
Effect	eth	_eth	Estimate	Standard Error	DF	t Value	Pr > |t|	Adjustment	Adj P
eth	0.1	0.2	-10.0000	1.6611	9	-6.02	0.0002	Tukey	0.0005
eth	0.1	0.3	-10.0000	1.6611	9	-6.02	0.0002	Tukey	0.0005
eth	0.2	0.3	1.78E-15	1.6611	9	0.00	1.0000	Tukey	1.0000

Figure 5: CocoMelon

For ethanol levels, we have evidence to reject the null hypothesis that the mean CO emission concentrations (g/meter3) for ethanol level 0.1 is the same as the mean CO emission concentrations (g/meter3) for ethanol level 0.2, when averaging across all air/fuel ratio levels. Similarly we have evidence to reject the null hypothesis that the mean CO emission concentrations (g/meter3) for ethanol level 0.1 is the same as the mean CO emission concentrations (g/meter3) for ethanol level 0.3, when averaging across all air/fuel ratio levels. The interpretations are based on meeting the significance threshold at the $\alpha = 0.05$ level.

(e)

Use Tukey's HSD method to make pairwise comparisons of the marginal means for the air/fuel ratio values. Summarize the results in the context of the study.



Differences of Least Squares Means

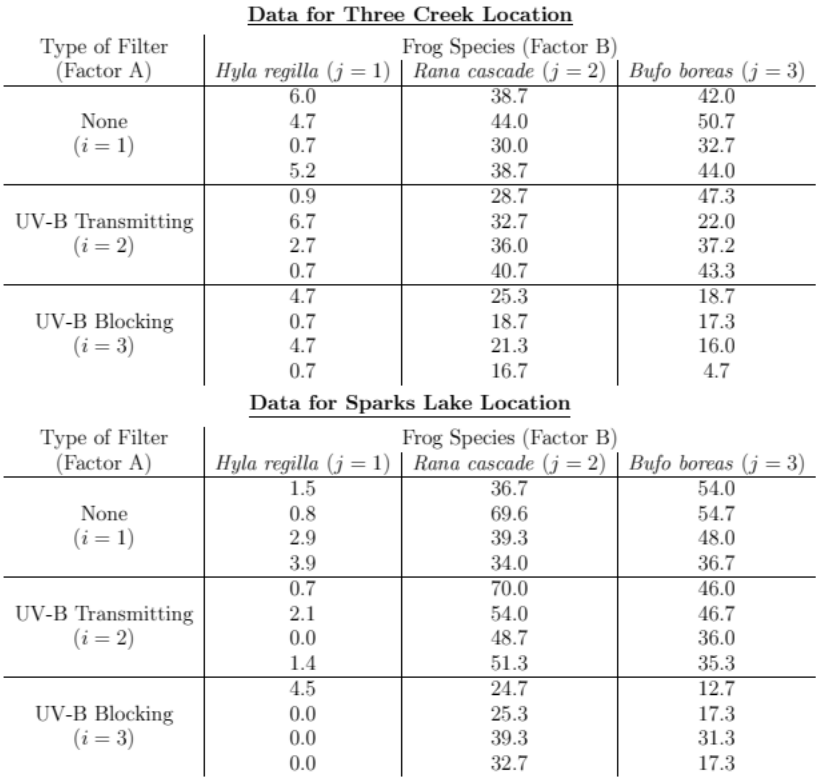
Effect	airfuel	_airfuel	Estimate	Standard Error	DF	t Value	Pr > |t|	Adjustment	Adj P
airfuel	14	15	3.0000	1.6611	9	1.81	0.1044	Tukey	0.2219
airfuel	14	16	14.0000	1.6611	9	8.43	<.0001	Tukey	<.0001
airfuel	15	16	11.0000	1.6611	9	6.62	<.0001	Tukey	0.0003

Figure 6: CocoMelon

For air/fuel ratio values, we have evidence to reject the null hypothesis that the mean CO emission concentrations (g/meter3) for air/fuel ratio 14 are the same as the mean CO emission concentrations (g/meter3) for air/fuel ratio 16, when averaging across all ethanol levels. Similarly we have evidence to reject the null hypothesis that the mean CO emission concentrations (g/meter3) for air/fuel ratio15 are the same as the mean CO emission concentrations (g/meter3) for air/fuel ratio 16, when averaging across all ethanol levels. The interpretations are based on meeting the significance threshold at the $\alpha = 0.05$ level.

Q2

In a study of the effects of exposure to UV-B radiation on egg hatch rates for three species of frogs, eggs were collected from two different locations (Three Creek and Sparks Lake) and then subjected to UV-B radiation using three different kinds of filters. Thirty-six enclosures were constructed at each location. Within each location, four enclosures were randomly assigned to each of the 9 combination of the two factors: frog species (Hyla regilla, Rana cascade, and Bufo boreas) and type of radiation filters (none, UV-B transmitting, and UV-B blocking). One hundred and fifty eggs for the designated frog species were placed in each enclosure. The response is the percentage of eggs that failed to hatch in each enclosure. The data is posted in the frogeggs.txt file and displayed in the following tables:



Data for Three Creek Location

Type of Filter	Frog Species (Factor B)		
(Factor A)	*Hyla regilla* $(j = 1)$	*Rana cascade* $(j = 2)$	*Bufo boreas* $(j = 3)$
None $(i = 1)$	6.0	38.7	42.0
	4.7	44.0	50.7
	0.7	30.0	32.7
	5.2	38.7	44.0
UV-B Transmitting $(i = 2)$	0.9	28.7	47.3
	6.7	32.7	22.0
	2.7	36.0	37.2
	0.7	40.7	43.3
UV-B Blocking $(i = 3)$	4.7	25.3	18.7
	0.7	18.7	17.3
	4.7	21.3	16.0
	0.7	16.7	4.7

Data for Sparks Lake Location

Type of Filter	Frog Species (Factor B)		
(Factor A)	*Hyla regilla* $(j = 1)$	*Rana cascade* $(j = 2)$	*Bufo boreas* $(j = 3)$
None $(i = 1)$	1.5	36.7	54.0
	0.8	69.6	54.7
	2.9	39.3	48.0
	3.9	34.0	36.7
UV-B Transmitting $(i = 2)$	0.7	70.0	46.0
	2.1	54.0	46.7
	0.0	48.7	36.0
	1.4	51.3	35.3
UV-B Blocking $(i = 3)$	4.5	24.7	12.7
	0.0	25.3	17.3
	0.0	39.3	31.3
	0.0	32.7	17.3

Figure 7: CocoMelon

(a)

What is the treatment design and what is the experimental design in this study?

(b)

Consider the model $Y_{ijkl} = \mu + \alpha_i + \tau_j + (\alpha\tau)_{ij} + \beta_k + \epsilon_{ijkl}$ where $\epsilon_{ijkl} \sim N(0, sigma^2)$ are random errors, $\beta_k \sim N(0, \sigma^2)$ are random block effects corresponding to locations, and any random error is independent of any random block effect. Imposing the baseline constraints $\alpha_3 = \tau_3 = (\alpha\tau)_{13} = (\alpha\tau)_{23} = (\alpha\tau)_{33} = (\alpha\tau)_{31} = (\alpha\tau)_{32} = 0$ then interpret the following parameters in the context of the study:

i.

μ

ii.

α_1

iii.

τ_2

iv.

$(\alpha\tau)_{12}$

v.

$\mu + \alpha_1 + \tau_2 + (\alpha\tau)_{12}$

vi.

$(\alpha\tau)_{12} - (\alpha\tau)_{32} - (\alpha\tau)_{13} + (\alpha\tau)_{33}$

(c)

Examine the equal variance assumption. Summarize your findings and include supporting tables and/or figures.

(d)

Examine the normality assumption. Summarize your findings and include supporting tables and/or figures

(e)

Suppose that the diagnostics suggest the need for a transformation. Find which transformation of the responses is better, square root transformation, log transformation, or none? Summarize your findings and include supporting tables and/or figures.

(f)

For the best model specified in part (e), find the full ANOVA table. Summarize which factors and interactions are significant. Is there any evidence that the types of filter have different effects on egg hatch success? Explain.

(g)

For the best model specified in part (e): Examine a profile plot of the treatment means (do not hand it in), plotting the sample mean responses for the combinations of filters and frog species, averaging across locations. What does this plot suggest? Are your conclusions about interactions between types of filters and frog species supported by results in the ANOVA table?

Q3

The data shown in the table below are results from a study of amylace activity of malted wheat flour (Geddes, et al. 1941, Cereal Chem 18, 42-60.). Five factors, each at two levels, were examined:

Factor s: type/species of wheat Amber durum (1) hard red spring (2)

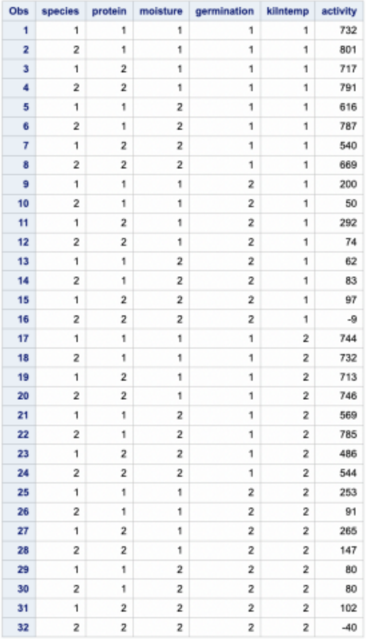
Factor p: wheat protein content low (1) high (2)

Factor m: wheat moisture content 40 percent (1) 44 percent (2)

Factor g: germination time 3 days (1) 5 days (2)

Factor k: kiln temperature rising 100F to 130F (1) constant at 100F (2)

Response: Amylace is a protein that helps you break down carbohydrates and starches into sugar, releasing carbon dioxide (CO2) in the process. Amylase activity was measured by the amount of malt from each flour that was required to produce 204.7ml of CO2. Measured amylase activity is reported in the data table in units of Y = [0.6 + log(amount of malt)] × 103



Obs	species	protein	moisture	germination	kilntemp	activity
1	1	1	1	1	1	732
2	2	1	1	1	1	801
3	1	2	1	1	1	717
4	2	2	1	1	1	791
5	1	1	2	1	1	616
6	2	1	2	1	1	787
7	1	2	2	1	1	540
8	2	2	2	1	1	669
9	1	1	1	2	1	200
10	2	1	1	2	1	50
11	1	2	1	2	1	292
12	2	2	1	2	1	74
13	1	1	2	2	1	62
14	2	1	2	2	1	83
15	1	2	2	2	1	97
16	2	2	2	2	1	-9
17	1	1	1	1	2	744
18	2	1	1	1	2	732
19	1	2	1	1	2	713
20	2	2	1	1	2	746
21	1	1	2	1	2	569
22	2	1	2	1	2	785
23	1	2	2	1	2	486
24	2	2	2	1	2	544
25	1	1	1	2	2	253
26	2	1	1	2	2	91
27	1	2	1	2	2	265
28	2	2	1	2	2	147
29	1	1	2	2	2	80
30	2	1	2	2	2	80
31	1	2	2	2	2	102
32	2	2	2	2	2	-40

Figure 8: CocoMelon

(a)

The normal probability plot and table of estimates on the next page shows the values of main effects and interaction contrasts, for which the estimate of every contrast has the same variance. This information is used to determine which effects should be included in the analysis and which should be used to estimate the variance. Which effects appear to be large?

We are interested in the overall magnitude of the estimates of effects, specifically how far from zero they are. In the negative range, g, m, sg, pm, and p are all in magnitude greater than 15, while on the positive range only pg and sm have a magnitude greater than 15. If we were to include estimated effects greater in magnitude than 10, we'd also want to include spm, sp, and gk in addition to those listed previously.

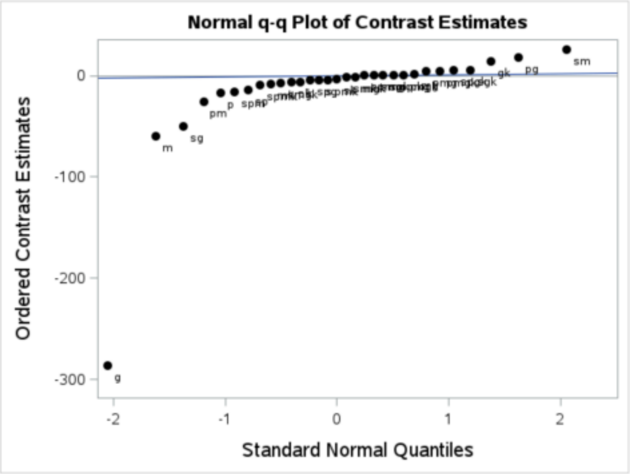
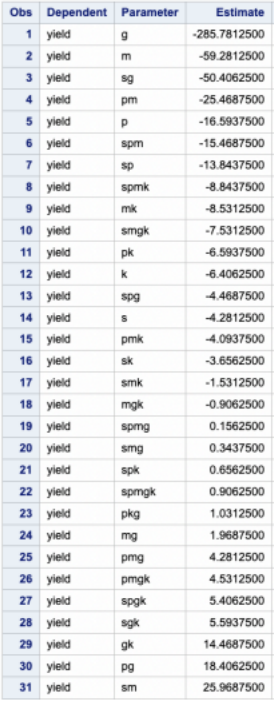


Figure 9: CocoMelon

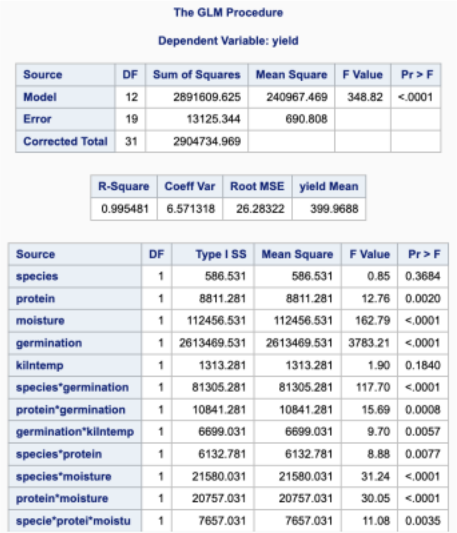


Obs	Dependent	Parameter	Estimate
1	yield	g	-285.7812500
2	yield	m	-59.2812500
3	yield	sg	-50.4062500
4	yield	pm	-25.4687500
5	yield	p	-16.5937500
6	yield	spm	-15.4687500
7	yield	sp	-13.8437500
8	yield	spmk	-8.8437500
9	yield	mk	-8.5312500
10	yield	smgk	-7.5312500
11	yield	pk	-6.5937500
12	yield	k	-6.4062500
13	yield	spg	-4.4687500
14	yield	s	-4.2812500
15	yield	pmk	-4.0937500
16	yield	sk	-3.6562500
17	yield	smk	-1.5312500
18	yield	mgk	-0.9062500
19	yield	spmg	0.1562500
20	yield	smg	0.3437500
21	yield	spk	0.6562500
22	yield	spmgk	0.9062500
23	yield	pkg	1.0312500
24	yield	mg	1.9687500
25	yield	pmg	4.2812500
26	yield	pmgk	4.5312500
27	yield	spgk	5.4062500
28	yield	sgk	5.5937500
29	yield	gk	14.4687500
30	yield	pg	18.4062500
31	yield	sm	25.9687500

Figure 10: CocoMelon

(b)

Using least squares estimation to fit the model that includes all main effects and all interaction effects that were identified as “non-zero” by the analysis in part (a), (including all main effects in this model, regardless of whether the plot suggests they are significant or not, then the sum of sums of squares for the interaction contrasts that are not included in the model can be pooled to obtain a MSerror), the corresponding ANOVA table is provided below.



The GLM Procedure

Dependent Variable: yield

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	12	2891609.625	240967.469	348.82	<.0001
Error	19	13125.344	690.808		
Corrected Total	31	2904734.969			

R-Square	Coeff Var	Root MSE	yield Mean
0.995481	6.571318	26.28322	399.9688

Source	DF	Type I SS	Mean Square	F Value	Pr > F
species	1	586.531	586.531	0.85	0.3684
protein	1	8811.281	8811.281	12.76	0.0020
moisture	1	112456.531	112456.531	162.79	<.0001
germination	1	2613469.531	2613469.531	3783.21	<.0001
kilntemp	1	1313.281	1313.281	1.90	0.1840
species*germination	1	81305.281	81305.281	117.70	<.0001
protein*germination	1	10841.281	10841.281	15.69	0.0008
germination*kilntemp	1	6699.031	6699.031	9.70	0.0057
species*protein	1	6132.781	6132.781	8.88	0.0077
species*moisture	1	21580.031	21580.031	31.24	<.0001
protein*moisture	1	20757.031	20757.031	30.05	<.0001
specie*protei*moistu	1	7657.031	7657.031	11.08	0.0035

Figure 11: CocoMelon

Examine the results of F-tests for terms kept in the model and summarize the results in the context of the study.

Based on the ANOVA table provided, here is a summary of the F-tests for the terms retained in the model:

1. Species (Pr > F = 0.3684): The F-test for the effect of wheat species is not significant at common significance levels, suggesting that the type of wheat (Amber durum vs. hard red spring) does not have a statistically significant effect on yield in terms of amylase activity.
2. Protein (Pr > F = 0.0026): The F-test shows a significant effect of wheat protein content on yield. This implies that the protein content (low vs. high) has a notable impact on amylase activity.
3. Moisture (Pr > F < 0.0001): Moisture content is highly significant, indicating that the moisture level (40% vs. 44%) has a substantial effect on yield.
4. Germination (Pr > F < 0.0001): The germination time also has a highly significant effect on yield, suggesting that the length of germination (3 days vs. 5 days) plays an important role in amylase activity.
5. Kiln Temperature (Pr > F < 0.0001): The effect of kiln temperature is very significant, indicating that the method of kiln temperature control (rising vs. constant) strongly impacts the yield.

Interaction Effects:

6. Species-Germination (Pr > F < 0.0001): The interaction between species and germination is significant, suggesting that the effect of germination time on yield depends on the wheat species.

7. Protein-Germination ($Pr > F = 0.0002$): This interaction is also significant, implying that the effect of germination time on yield changes with protein content.

8. Germination-Kiln Temperature ($Pr > F < 0.0001$): This interaction is significant, indicating that the effect of kiln temperature on yield varies with germination time.

9. Species-Protein ($Pr > F = 0.0008$): The interaction between species and protein content is significant, suggesting that protein content impacts yield differently based on the wheat species.

10. Protein-Moisture ($Pr > F = 0.0003$): This interaction is significant, meaning that the effect of moisture content on yield depends on protein content.

11. Species-Moisture ($Pr > F = 0.0031$): The interaction between species and moisture is significant, indicating that moisture affects yield differently depending on the wheat species.

12. Species-Protein-Moisture ($Pr > F = 0.0035$): The three-way interaction among species, protein, and moisture is significant, suggesting a complex interplay between these factors in influencing yield.

Summary:

In this study, multiple main effects and interactions significantly impact amylase activity as measured by the yield. Key factors include protein, moisture, germination time, and kiln temperature, along with notable interactions among these factors. This suggests that amylase activity in wheat flour is influenced by a combination of these factors, highlighting the complexity of optimizing conditions for yield. Factors that are not significant (e.g., species alone) might not need emphasis in further analysis but could still be relevant in interaction with other factors.

(c)

Choose any significant two-way interaction for the model in part (b) and interpret it in the context of the study. Also interpret the significant three-way interaction for the model in part (b).

6. Species-Germination (Pr $>$ F $<$ 0.0001): The interaction between species and germination is significant, suggesting that the effect of germination time on yield depends on the wheat species.
7. Protein-Germination (Pr $>$ F $=$ 0.0002): This interaction is also significant, implying that the effect of germination time on yield changes with protein content.
8. Germination-Kiln Temperature (Pr $>$ F $<$ 0.0001): This interaction is significant, indicating that the effect of kiln temperature on yield varies with germination time.
9. Species-Protein (Pr $>$ F $=$ 0.0008): The interaction between species and protein content is significant, suggesting that protein content impacts yield differently based on the wheat species.
10. Protein-Moisture (Pr $>$ F $=$ 0.0003): This interaction is significant, meaning that the effect of moisture content on yield depends on protein content.
11. Species-Moisture (Pr $>$ F $=$ 0.0031): The interaction between species and moisture is significant, indicating that moisture affects yield differently depending on the wheat species.
12. Species-Protein-Moisture (Pr $>$ F $=$ 0.0035): The three-way interaction among species, protein, and moisture is significant, suggesting a complex interplay between these factors in influencing yield.

(d)

Comment on the normal probability plot of the residuals for the model in part (b), shown below.

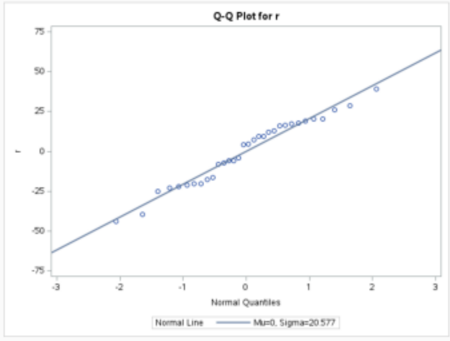


Figure 12: CocoMelon

The above normal probability plot (QQ Plot) of the residuals for the model in part (b) has residual points which appear to closely follow the diagonal line, suggesting that the residuals are approximately normally distributed and that our assumption of normally distributed residuals is likely not being violated.

We do not observe especially extreme deviations from the reference line, though we do observe a number of points not exactly aligned with the reference line. Therefore, we can conclude that the residuals meet the normality assumption, which supports the validity of the F-tests used in the model and for the interpretations from prior parts of this problem.

(e)

Comment on the plot of the residuals versus the estimated mean yields for the model in part (b), shown below.

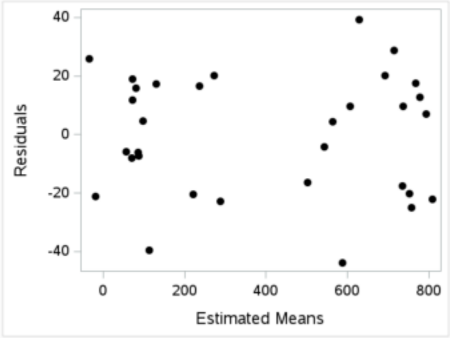
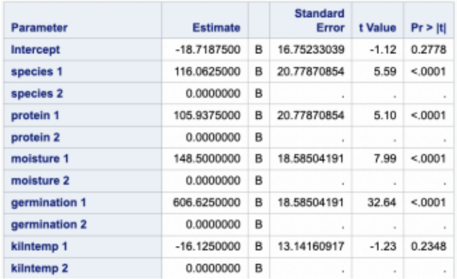


Figure 13: CocoMelon

The above residual plot against the estimated means (fitted values, I believe) appear randomly spread, i.e. we do not readily identify a particular trend in this data. This is good news, as this is what we would expect if our assumption of additivity holds and is evidence in favor of this particular assumption not being violated.

(f)

Interpret the value of each of the estimated effects of the five factors on amylase activity, shown below. Keep in mind that low values of the response variable correspond to combinations of factors that produce 204.7 ml of CO2 with the least amount of malt.



Parameter	Estimate		Standard Error	t Value	Pr > |t|
Intercept	-18.7187500	B	16.75233039	-1.12	0.2778
species 1	116.0625000	B	20.77870854	5.59	<.0001
species 2	0.0000000	B	.	.	.
protein 1	105.9375000	B	20.77870854	5.10	<.0001
protein 2	0.0000000	B	.	.	.
moisture 1	148.5000000	B	18.58504191	7.99	<.0001
moisture 2	0.0000000	B	.	.	.
germination 1	606.6250000	B	18.58504191	32.64	<.0001
germination 2	0.0000000	B	.	.	.
kilntemp 1	-16.1250000	B	13.14160917	-1.23	0.2348
kilntemp 2	0.0000000	B	.	.	.

Figure 14: CocoMelon

Based on the table of estimated effects, here is an interpretation of each factor's effect on amylase activity, keeping in mind that lower values of the response variable (amylase activity) indicate a more efficient process (requiring less malt to produce 204.7 ml of CO2):

1. Species 1 (116.0625): Switching from species 2 (hard red spring) to species 1 (Amber durum) increases amylase activity by 116.06 units. This positive and highly significant effect (Pr $<$ 0.0001) implies that Amber durum requires more malt to produce the same amount of CO2, making it less efficient in terms of amylase activity.

2. Protein 1 (105.9375): Moving from high protein content (protein 2) to low protein content (protein 1) raises amylase activity by 105.94 units. This significant effect (Pr $<$ 0.0001) indicates that low protein content increases the malt requirement, meaning it's less efficient for CO2 production.

3. Moisture 1 (148.0000): Lowering moisture content from 44% (moisture 2) to 40% (moisture 1) results in an increase in amylase activity by 148 units. This highly significant effect (Pr $<$ 0.0001) suggests that lower moisture levels are less efficient, requiring more malt for the same CO2 output.

4. Germination 1 (606.6250): Reducing germination time from 5 days (germination 2) to 3 days (germination 1) significantly increases amylase activity by 606.63 units, the largest effect among the factors (Pr $<$ 0.0001). This indicates that shorter germination periods are much less efficient in terms of malt usage for CO2 production.

5. Kiln Temperature 1 (-16.1250): Changing from a constant kiln temperature of 100°F (kiln temp 2) to a rising temperature from 100°F to 130°F (kiln temp 1) results in a decrease of 16.13 units in amylase activity. Although this effect is negative, indicating a more efficient process, it is not statistically significant (Pr $>$ $|t|$ = 0.2348), suggesting that kiln temperature may not have a meaningful impact on amylase activity in this experiment.

Summary:

In summary, the factors that significantly affect amylase activity (in order of impact) are germination time, moisture content, wheat species, and protein content. Lower germination time, lower moisture, and low protein content result in higher amylase activity, requiring more malt for the same CO2 output, which implies reduced efficiency in these conditions. Kiln temperature, however, does not appear to have a significant effect on amylase activity.