HW 8

2024-10-28

Stat 5000 Homework #8
Fall 2024 due Fri, November 1st @ 11:59 pm Name: Sam Olson
Collaborators: Sarah, Sabrina, **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 Ethanol	Air/Fuel Ratio		
	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

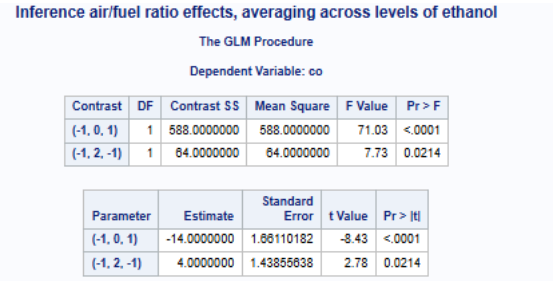
Parameter	Estimate	Standard Error	t Value	Pr > |t|
(-1, 0, 1)	10.0000000	1.66110182	6.02	0.0002
(-1, 2, -1)	5.0000000	1.43855638	3.48	0.0070

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 (Factor A)	Frog Species (Factor B)		
	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 (Factor A)	Frog Species (Factor B)		
	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?

Treatment Design: We have 2 factors each with 3 levels, so it's a 3×3 factorial arrangement (3 distinct Species by 3 distinct Filter Types)

Experimental Design: Randomized Complete Block Design (RCBD) with blocks defined by location (Two locations: Three Creek and Sparks Lake).

(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

The grand mean represents the overall average percentage of eggs that failed to hatch across all frog species, locations, and types of UV-B filters.

ii.

α_1

This is the effect of the first level of factor 1 (frog species), representing the difference in the percentage of failed hatches for Hyla regilla relative to the baseline species, Bufo boreas. It quantifies how the average hatching failure for Hyla regilla deviates from the baseline species when other factors remain constant.

iii.

τ_2

This is the effect of the second level of factor 2 (type of UV-B filter), specifically for the UV-B transmitting filter. It represents how much the percentage of failed hatches changes for enclosures with this filter type relative to the baseline filter type (UV-B blocking filter).

iv.

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

This parameter represents the interaction effect between the frog species Hyla regilla and the UV-B transmitting filter. It measures how much the combination of Hyla regilla and the UV-B transmitting filter influences the failure rate of egg hatching beyond what would be expected by considering the individual effects of the frog species and filter type alone.

v.

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

This expression provides the cell mean for the treatment combination of Hyla regilla (level 1 of factor 1) and the UV-B transmitting filter (level 2 of factor 2). It estimates the average percentage of eggs that failed to hatch under this specific combination, incorporating the grand mean, the effects of species and filter, and their interaction.

vi.

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

This expression measures the difference in interaction effects. It quantifies how much the interaction between Hyla regilla with the UV-B transmitting filter differs from other interaction effects, adjusted by the interactions involving the baseline species, Bufo boreas. Specifically, it contrasts the joint effects of Hyla regilla with UV-B transmitting against those of other frog species and filter combinations, which helps in understanding if certain species-filter pairings affect hatching success differently than expected under simple additive effects.

(c)

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

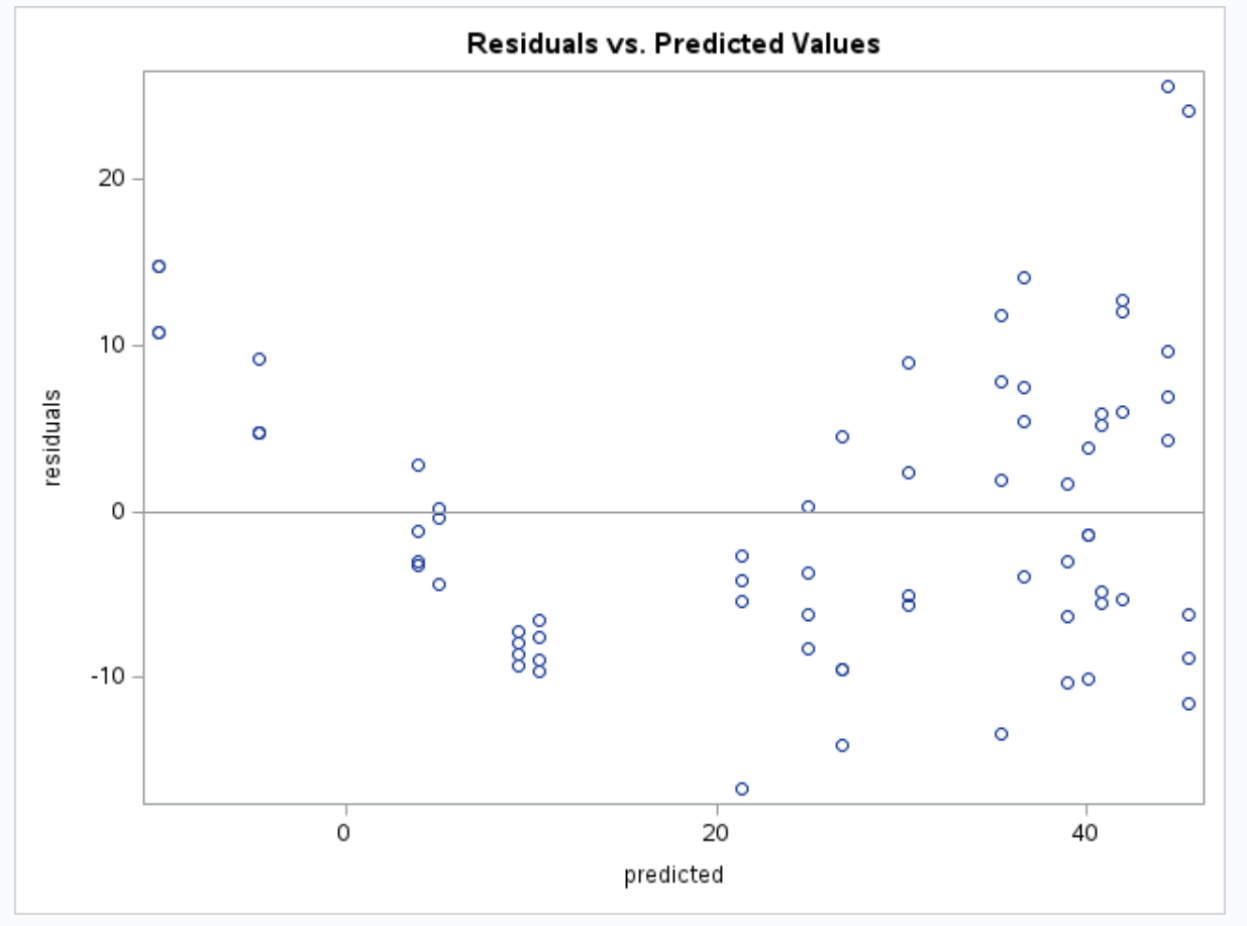


Figure 8: CocoMelon

For this assumption we review three residual plots: By fitted values, by levels of Factor 1 (Type of Filter), and by levels of Factor 2 (Frog Species). We are generally looking for a trend in these plots as evidence that our equal variances assumption is being violated. For small fitted values (left side of the fitted value plot), we do not observe any negative residuals, which is somewhat concerning. So there is some evidence that our equal variances assumption may be violated in this experiment, though the other two residual plots do not exhibit such a problematic trend.

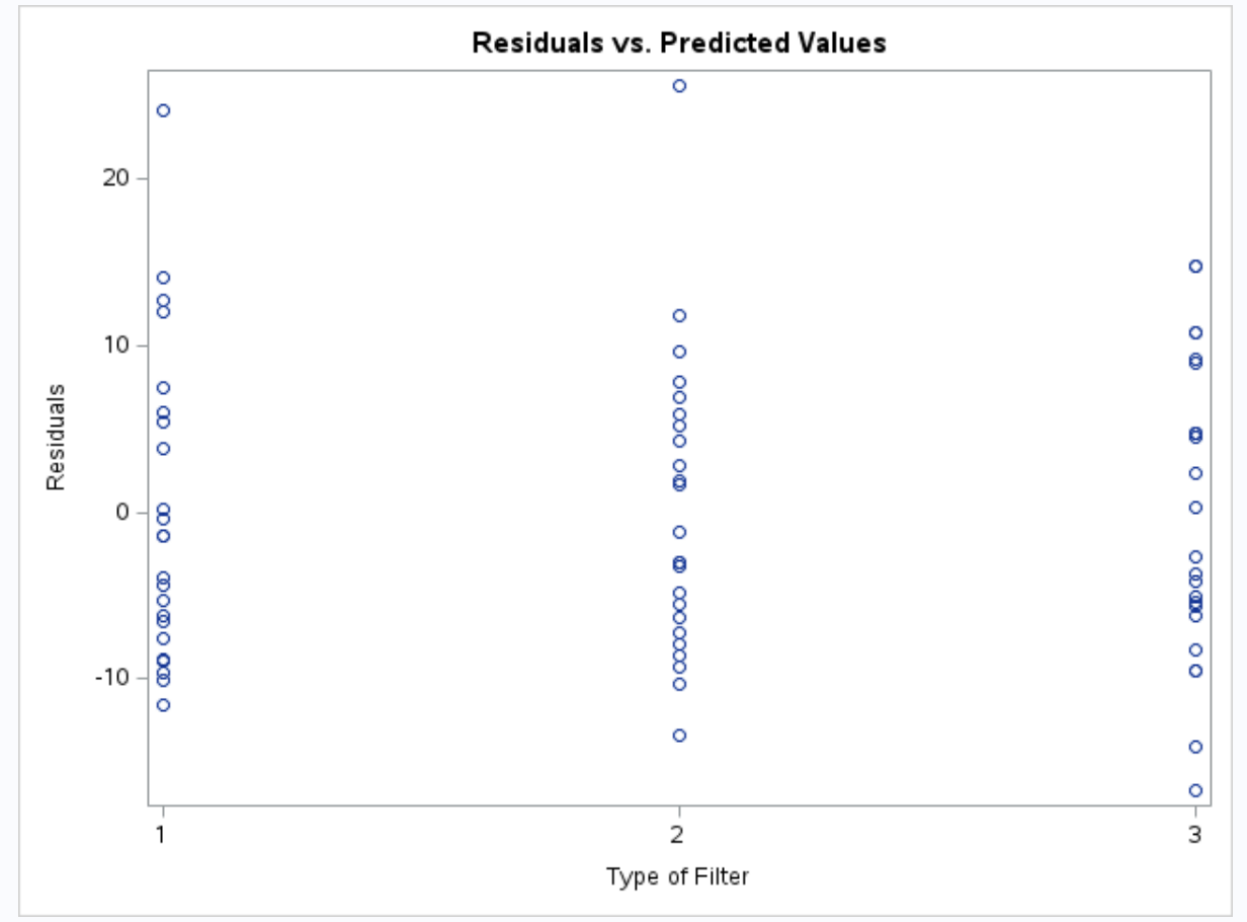


Figure 9: CocoMelon

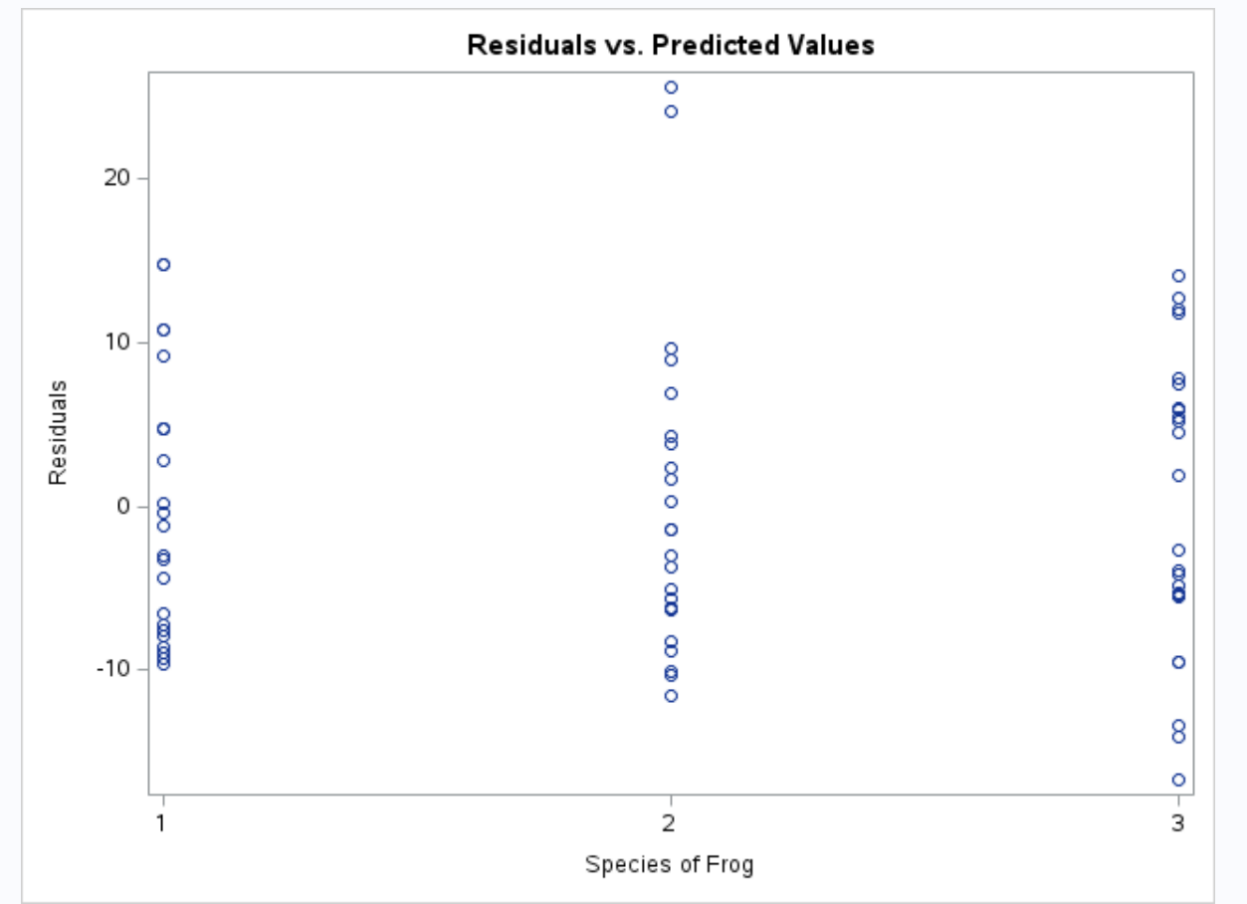


Figure 10: CocoMelon

(d)

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

To assess this assumption we look at: Residual Summary statistics: Unfortunately, we see that Mean of Residuals $\neq$ Median of Residuals Shapiro Wilk statistical test: Large p-value does not provide evidence to reject the null hypothesis that residuals are normally distributed Histogram of residuals: We observe residuals do exhibit a slight bell-shaped distribution but there is visually some evidence of non-normality in skewnewss Residual boxplot: Similar to the Histogram noted, the boxplot looks roughly normal, though there is not a symmetrical spread of residuals about the center QQ Plot: The residuals roughly align with the reference line, though there are deviations, particularly about the center (0 Normal Quartile), as well as some deviations near the tails. This is not especially problematic, but is nonetheless a cause of concern

Overall, we have some visual evidence to believe our residuals are not normally distributed, particularly with the histogram and QQ plot, though generally the above graphs and plots indicate the residuals are roughly normally distributed and that our normality assumption is not violated. We may certainly improve the normality of our residuals though, which we will do in the later parts of the problem.

The UNIVARIATE Procedure
Variable: residual

Moments			
N	71	Sum Weights	71
Mean	0	Sum Observations	0
Std Deviation	8.87351128	Variance	78.7392024
Skewness	0.63093889	Kurtosis	0.13908189
Uncorrected SS	5511.74416	Corrected SS	5511.74416
Coeff Variation	.	Std Error Mean	1.05309204

Basic Statistical Measures			
Location		Variability	
Mean	0.00000	Std Deviation	8.87351
Median	-1.41256	Variance	78.73920
Mode	4.72729	Range	42.32736
		Interquartile Range	12.40707

Tests for Location: Mu0=0				
Test	Statistic		p Value	
Student's t	t	0	Pr > |t|	1.0000
Sign	M	-3.5	Pr >= |M|	0.4767
Signed Rank	S	-68	Pr >= |S|	0.6997

Tests for Normality				
Test	Statistic		p Value	
Shapiro-Wilk	W	0.961954	Pr < W	0.0303
Kolmogorov-Smirnov	D	0.112618	Pr > D	0.0242
Cramer-von Mises	W-Sq	0.141514	Pr > W-Sq	0.0312

Figure 11: CocoMelon

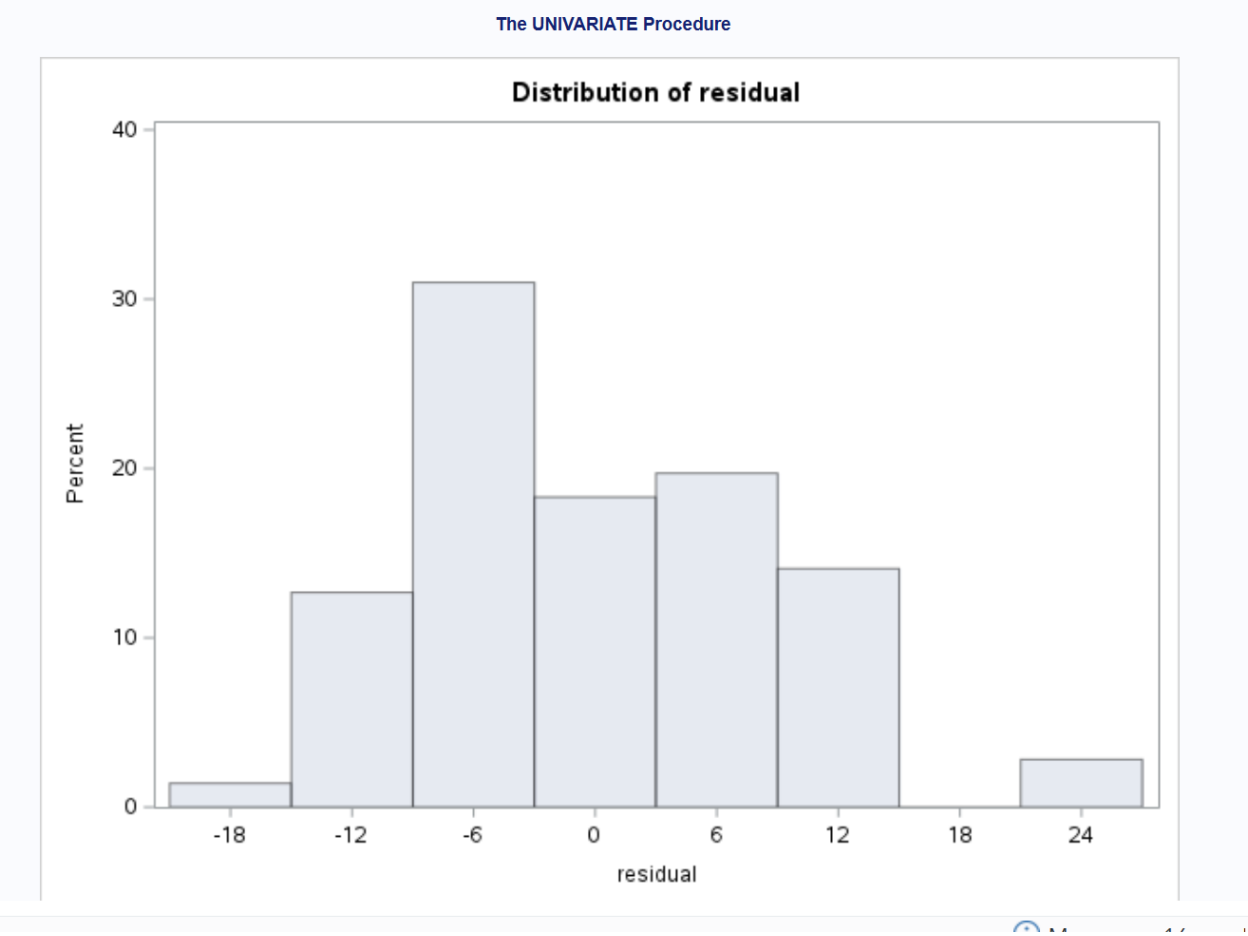


Figure 12: CocoMelon

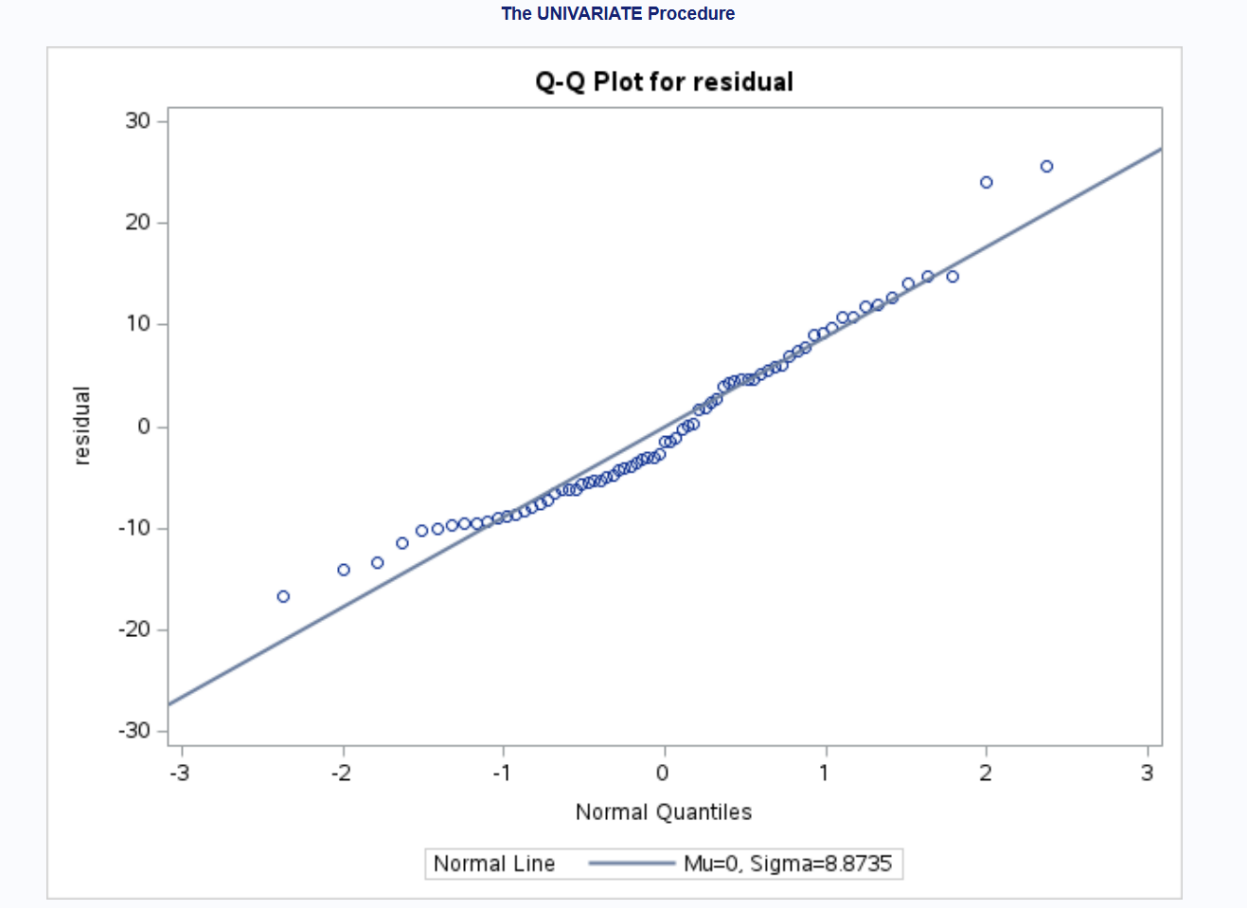


Figure 13: CocoMelon

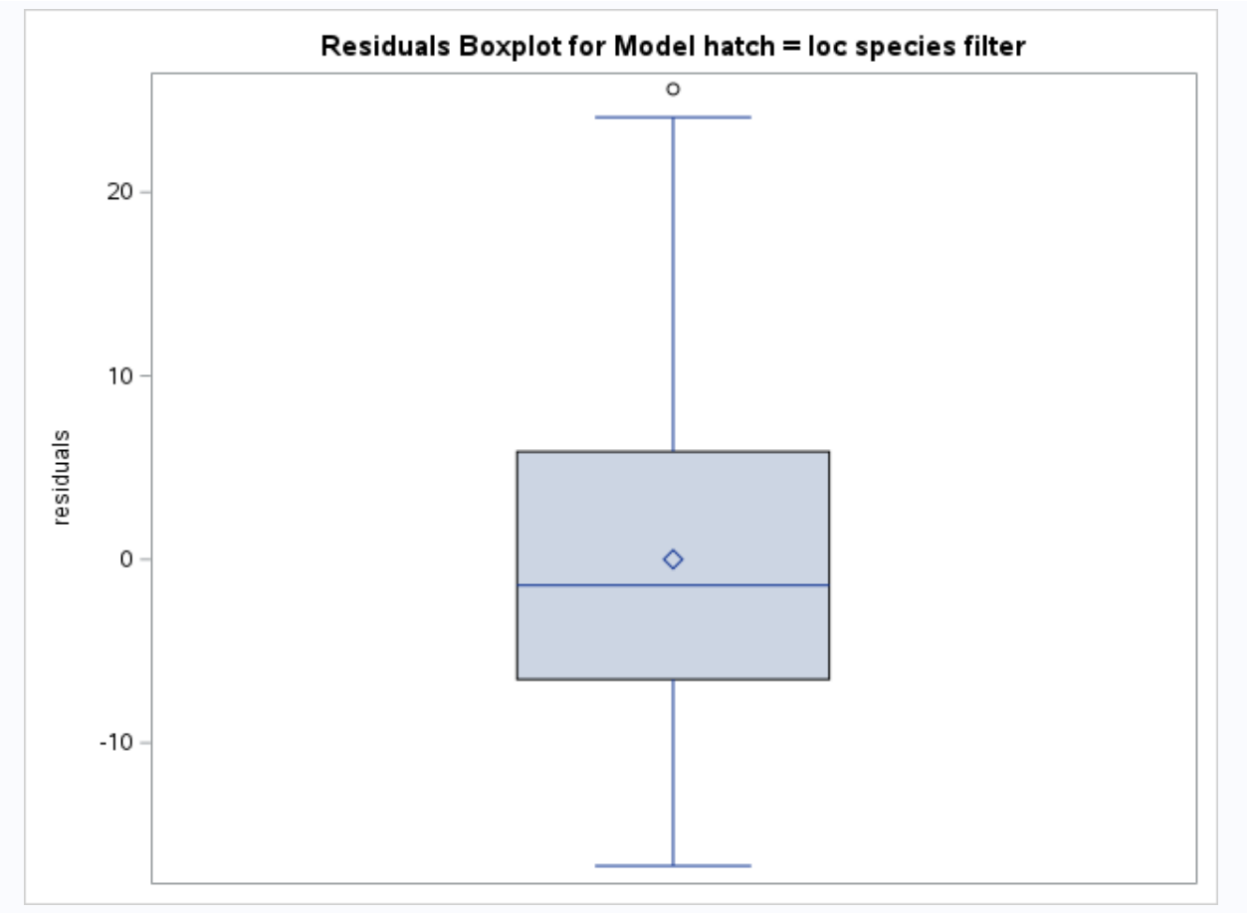


Figure 14: CocoMelon

(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.

Primarily compare the different transformations via "Visual" and "Statistical" Assessments. Graphs and plots are given below for these two categories.

Visual Assessment

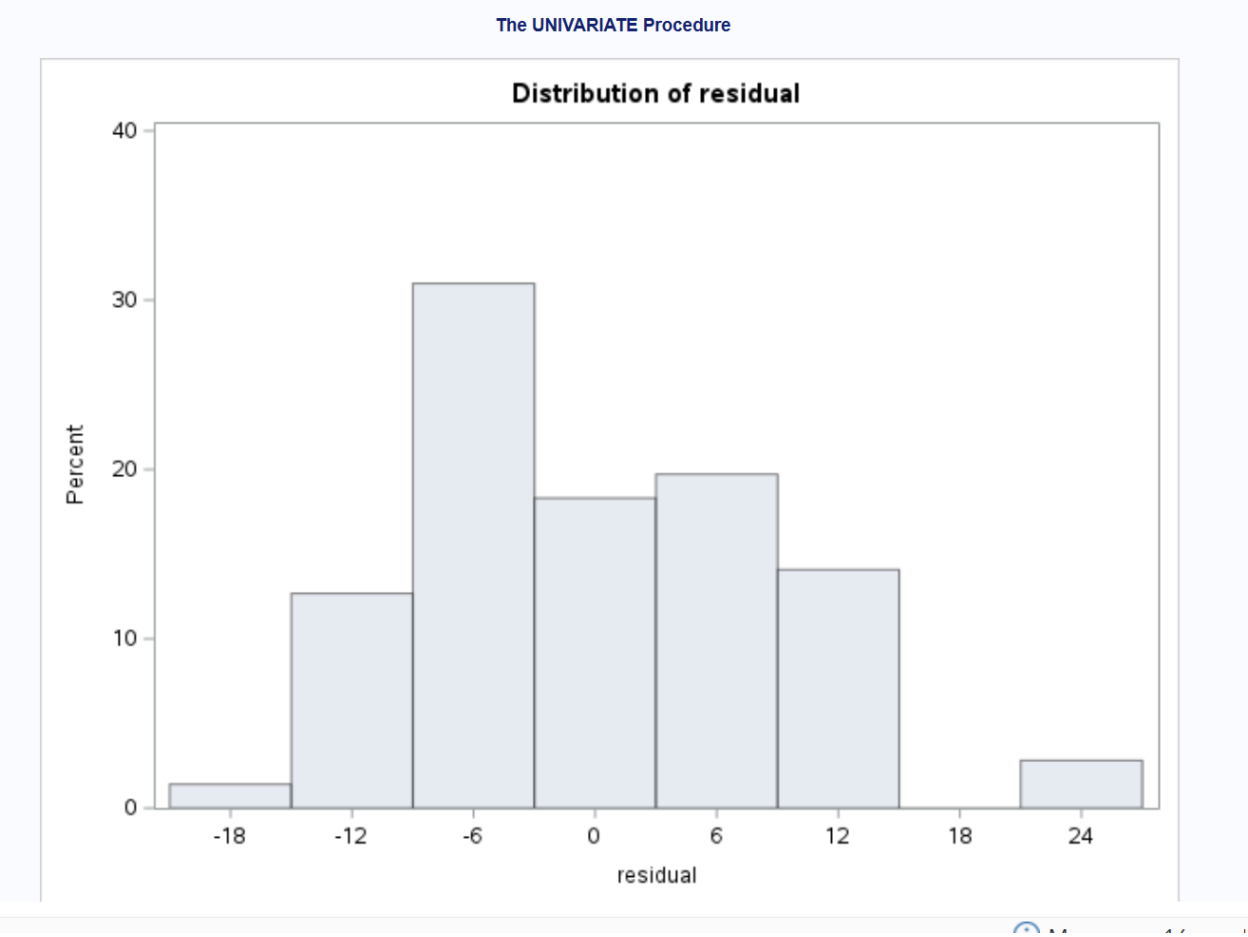


Figure 15: CocoMelon

Visually, the log transformation of residuals produces the most relatively normal looking histogram, in which we're looking for a histogram with symmetry near the center (lack of skewness). By comparison, the original and sqrt transformed residuals still have some skewness in the histogram, whereas the log transformed residuals appear to minimize this spread.

Statistical Tests

Via statistical tests, particularly the Shapiro-Wilk test, we are testing whether the residuals (original, sqrt, or log transformed) are not normally distributed; in all instances we observe rather large p-values such that

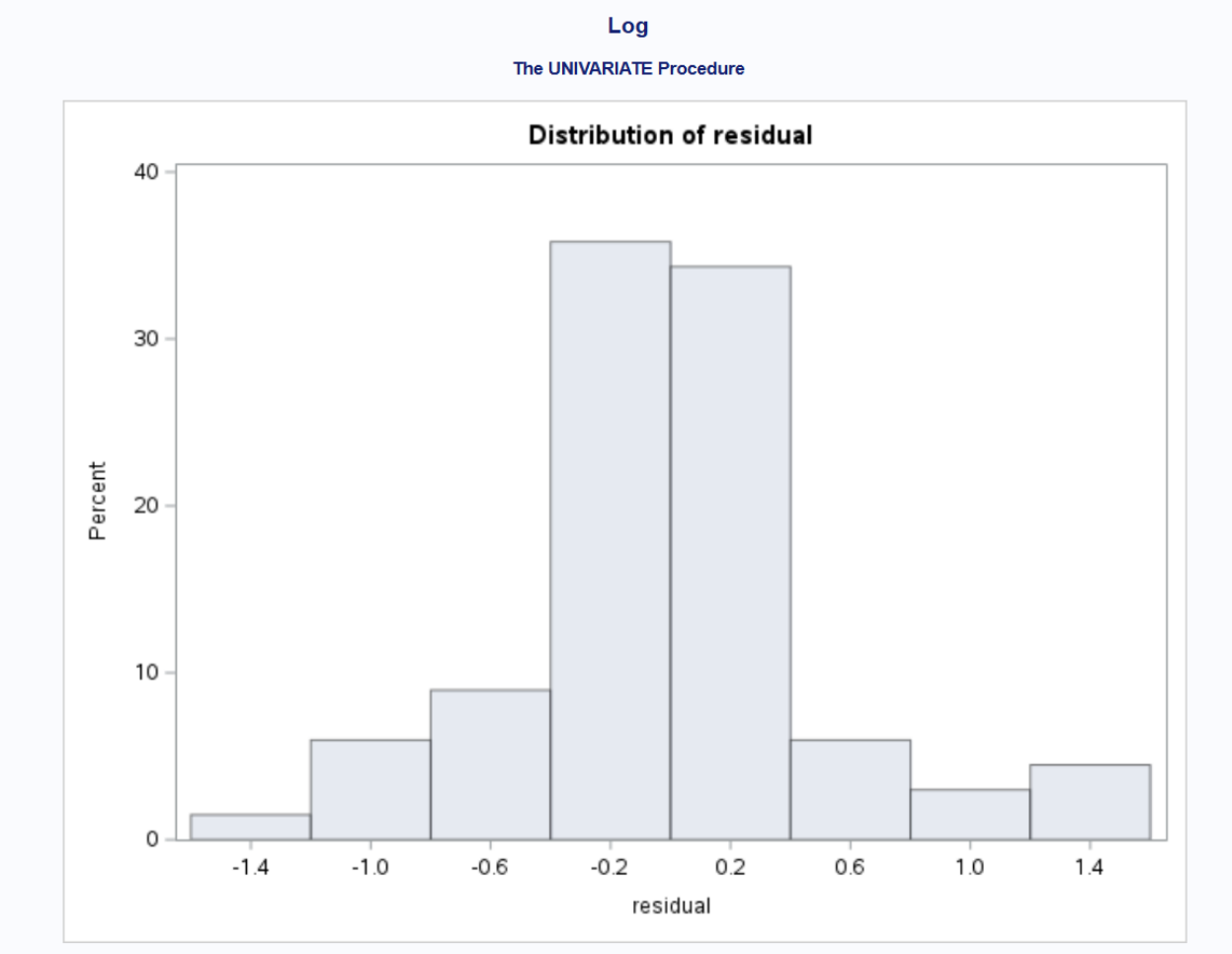


Figure 16: CocoMelon

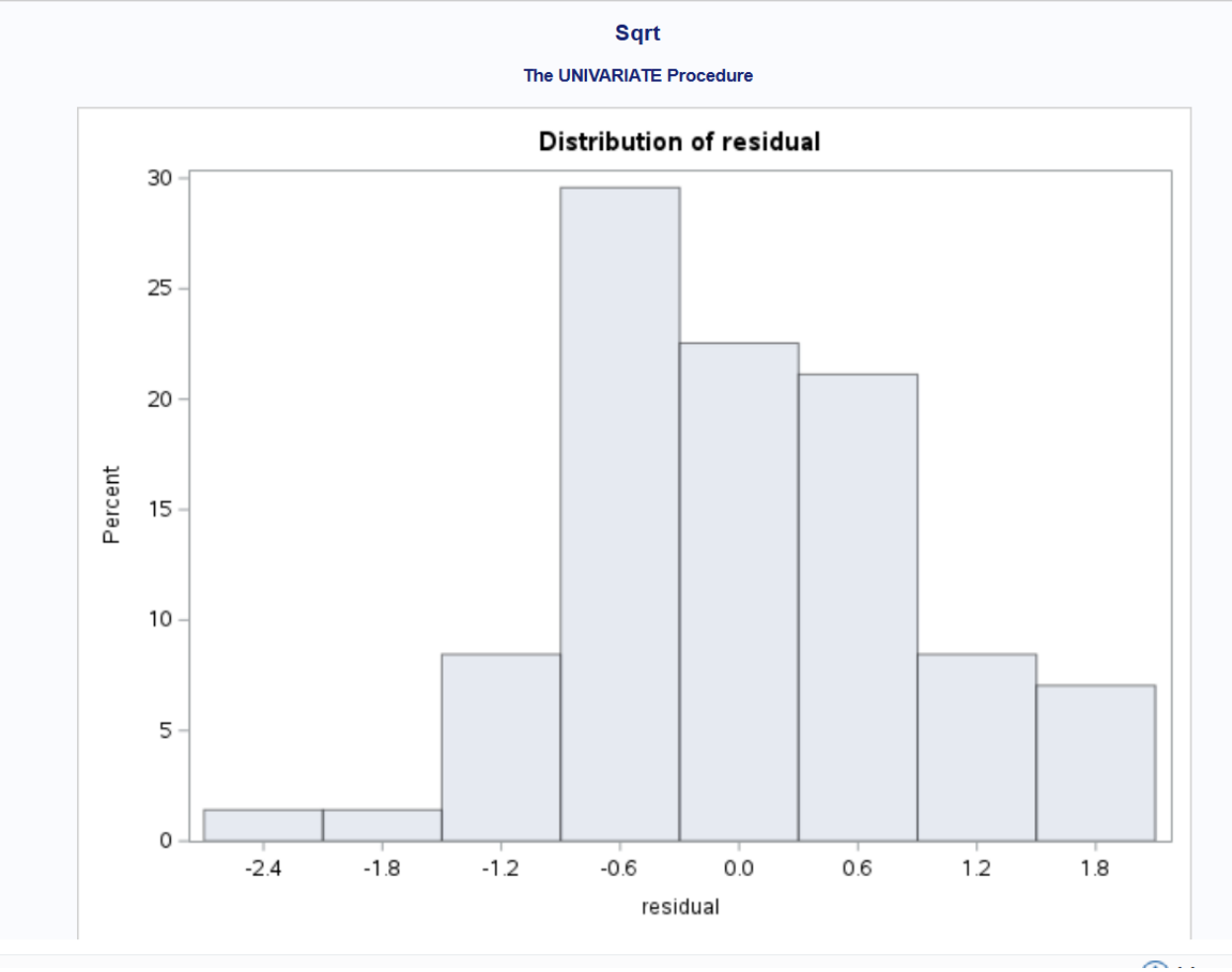


Figure 17: CocoMelon

The UNIVARIATE Procedure
Variable: residual

Moments			
N	71	Sum Weights	71
Mean	0	Sum Observations	0
Std Deviation	8.87351128	Variance	78.7392024
Skewness	0.63093889	Kurtosis	0.13908189
Uncorrected SS	5511.74416	Corrected SS	5511.74416
Coeff Variation	.	Std Error Mean	1.05309204

Basic Statistical Measures			
Location		Variability	
Mean	0.00000	Std Deviation	8.87351
Median	-1.41256	Variance	78.73920
Mode	4.72729	Range	42.32736
		Interquartile Range	12.40707

Tests for Location: Mu0=0				
Test	Statistic		p Value	
Student's t	t	0	Pr > |t|	1.0000
Sign	M	-3.5	Pr >= |M|	0.4767
Signed Rank	S	-68	Pr >= |S|	0.6997

Tests for Normality				
Test	Statistic		p Value	
Shapiro-Wilk	W	0.961954	Pr < W	0.0303
Kolmogorov-Smirnov	D	0.112618	Pr > D	0.0242
Cramer-von Mises	W-Sq	0.141514	Pr > W-Sq	0.0312

Figure 18: CocoMelon

Log

The UNIVARIATE Procedure
Variable: residual

Moments			
N	67	Sum Weights	67
Mean	0	Sum Observations	0
Std Deviation	0.54853428	Variance	0.30088986
Skewness	-0.0395734	Kurtosis	1.06667193
Uncorrected SS	19.8587305	Corrected SS	19.8587305
Coeff Variation	.	Std Error Mean	0.06701413

Basic Statistical Measures			
Location		Variability	
Mean	0.00000	Std Deviation	0.54853
Median	-0.01749	Variance	0.30089
Mode	-0.57654	Range	2.75735
		Interquartile Range	0.56271

Note: The mode displayed is the smallest of 4 modes with a count of 2.

Tests for Location: Mu0=0				
Test	Statistic		p Value	
Student's t	t	0	Pr > |t|	1.0000
Sign	M	-1.5	Pr >= |M|	0.8072
Signed Rank	S	23	Pr >= |S|	0.8870

Tests for Normality				
Test	Statistic		p Value	
Shapiro-Wilk	W	0.955582	Pr < W	0.0177
Kolmogorov-Smirnov	D	0.112395	Pr > D	0.0352
Cramer-von Mises	W-Sq	0.166558	Pr > W-Sq	0.0151
Anderson-Darling	A-Sq	1.094986	Pr > A-Sq	0.0071

Figure 19: CocoMelon

Sqrt

The UNIVARIATE Procedure
Variable: residual

Moments			
N	71	Sum Weights	71
Mean	0	Sum Observations	0
Std Deviation	0.86808311	Variance	0.75356828
Skewness	0.20788619	Kurtosis	0.19566928
Uncorrected SS	52.7497795	Corrected SS	52.7497795
Coeff Variation	.	Std Error Mean	0.10302251

Basic Statistical Measures			
Location		Variability	
Mean	0.00000	Std Deviation	0.86808
Median	-0.22800	Variance	0.75357
Mode	-0.44775	Range	4.38422
		Interquartile Range	1.19866

Tests for Location: Mu0=0				
Test	Statistic		p Value	
Student's t	t	0	Pr > |t|	1.0000
Sign	M	-4.5	Pr >= |M|	0.3425
Signed Rank	S	-30	Pr >= |S|	0.8649

Tests for Normality				
Test	Statistic		p Value	
Shapiro-Wilk	W	0.976037	Pr < W	0.1919
Kolmogorov-Smirnov	D	0.126022	Pr > D	<0.0100
Cramer-von Mises	W-Sq	0.124669	Pr > W-Sq	0.0510
Anderson-Darling	A-Sq	0.693885	Pr > A-Sq	0.0709

Figure 20: CocoMelon

we do not reject the null hypothesis in all three instances that the particular (original, sqrt, or log) residuals are normally distributed.

Overall, we observe that statistical tests are consistent across the residual types (original, sqrt, and log transformed), in that in all instances of Shapiro-Wilk we do not reject the null hypothesis that the (original/sqrt/log) residuals are normally distributed. So our choice ultimately comes down to the visual assessment, whereby we observe the most visually normal looking residuals from the log transformation, which becomes our choice for part (f).

(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.

The GLM Procedure

Dependent Variable: hatch

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	9	23410.64121	2601.18236	44.52	<.0001
Error	61	3563.69710	58.42126		
Corrected Total	70	26974.33831			

R-Square	Coeff Var	Root MSE	hatch Mean
0.867886	30.84636	7.643380	24.77887

Source	DF	Type I SS	Mean Square	F Value	Pr > F
loc	1	430.04971	430.04971	7.36	0.0087
species	2	17640.26700	8820.13350	150.97	<.0001
filter	2	3392.27743	1696.13872	29.03	<.0001
species*filter	4	1948.04707	487.01177	8.34	<.0001

Source	DF	Type III SS	Mean Square	F Value	Pr > F
loc	1	545.96147	545.96147	9.35	0.0033
species	2	17576.40224	8788.20112	150.43	<.0001
filter	2	3334.32401	1667.16200	28.54	<.0001
species*filter	4	1948.04707	487.01177	8.34	<.0001

Figure 21: CocoMelon

Parameter	Estimate		Standard Error	t Value	Pr > |t|
Intercept	19.68830645	B	2.85081596	6.91	<.0001
loc 1	-5.55161290	B	1.81603265	-3.06	0.0033
loc 2	0.00000000	B	.	.	.
species 1	-15.00000000	B	3.82169020	-3.92	0.0002
species 2	8.58750000	B	3.82169020	2.25	0.0283
species 3	0.00000000	B	.	.	.
filter 1	28.43750000	B	3.82169020	7.44	<.0001
filter 2	22.32500000	B	3.82169020	5.84	<.0001
filter 3	0.00000000	B	.	.	.
species*filter 1 1	-27.93225806	B	5.50188106	-5.08	<.0001
species*filter 1 2	-22.33750000	B	5.40468611	-4.13	0.0001
species*filter 1 3	0.00000000	B	.	.	.
species*filter 2 1	-12.56250000	B	5.40468611	-2.32	0.0235
species*filter 2 2	-2.56250000	B	5.40468611	-0.47	0.6371
species*filter 2 3	0.00000000	B	.	.	.
species*filter 3 1	0.00000000	B	.	.	.
species*filter 3 2	0.00000000	B	.	.	.
species*filter 3 3	0.00000000	B	.	.	.

Figure 22: CocoMelon

(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 23: 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.

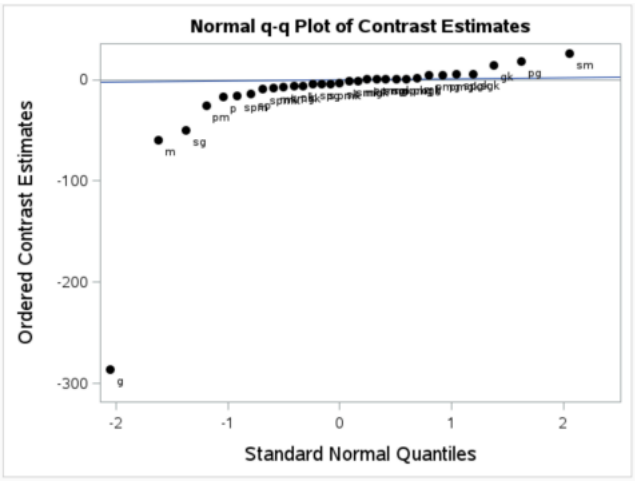


Figure 24: CocoMelon

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 25: 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 26: 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.

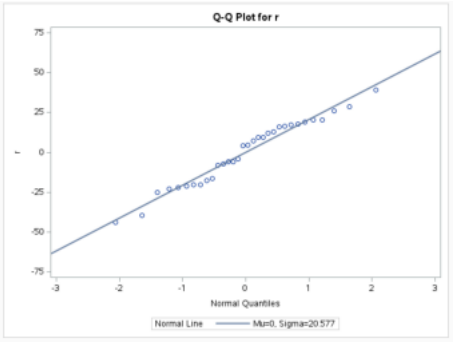


Figure 27: CocoMelon

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.

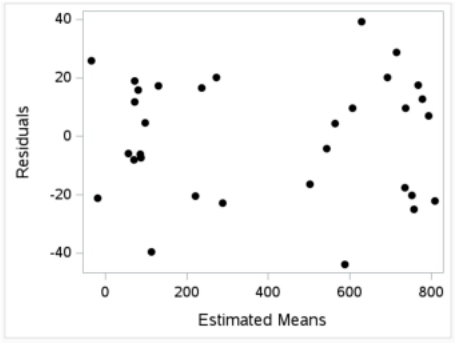


Figure 28: CocoMelon

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 29: 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.