Homework 4

Brian Cervantes Alvarez

May 1, 2024

ST 553 Statistical Methods

Question 1

1.1

For the hypothesis test $H_0 : \mu_1 = \mu_2$ versus $H_a : \mu_1 \neq \mu_2$ at the level α, the test statistic is defined as:

$$Z = \frac{\bar{X}_1 - \bar{X}_2}{\sigma_{\bar{X}_1 - \bar{X}_2}}$$

where $\bar{X}_1$ and $\bar{X}_2$ are the sample means of the two independent normal distributions, both with variance σ^2. The standard deviation of the difference in sample means is given by:

$$\sigma_{\bar{X}_1 - \bar{X}_2} = \sqrt{\frac{\sigma^2}{n} + \frac{\sigma^2}{n}} = \sqrt{\frac{2\sigma^2}{n}}$$

Thus, the rejection region for this two-tailed test at the α significance level is where:

$$|Z| > Z_{\alpha/2}$$

1.2

Assuming $\mu_1 - \mu_2 = \delta > 0$, the test statistic under this alternative hypothesis becomes:

$$Z = \frac{(\bar{X}_1 - \bar{X}_2) - \delta}{\sqrt{\frac{2\sigma^2}{n}}}$$

This statistic follows a normal distribution with mean $-\frac{\delta}{\sqrt{\frac{2\sigma^2}{n}}}$ and standard deviation of 1, expressed as:

$$Z \sim N\left(-\frac{\delta}{\sqrt{\frac{2\sigma^2}{n}}}, 1\right)$$

The power of the test, or the probability of correctly rejecting H_0, is calculated as the probability that this test statistic falls in the rejection region defined for the null hypothesis:

$$\text{Power} = P(Z > Z_{\alpha/2}) = P\left(Z > Z_{\alpha/2} - \frac{\delta}{\sqrt{\frac{2\sigma^2}{n}}}\right)$$

where Z is now expressed relative to its new mean under the alternative hypothesis.

1.3

Given that $\beta = 1 - \text{Power}$, we rearrange to find:

$$Z_{\alpha/2} - \frac{\delta}{\sqrt{\frac{2\sigma^2}{n}}} = -Z_\beta$$

Solving for n gives:

$$\sqrt{n} = \frac{\delta}{\sqrt{2\sigma^2}(Z_{\alpha/2} + Z_\beta)}$$

$$n = \frac{2\sigma^2(Z_{\alpha/2} + Z_\beta)^2}{\delta^2}$$

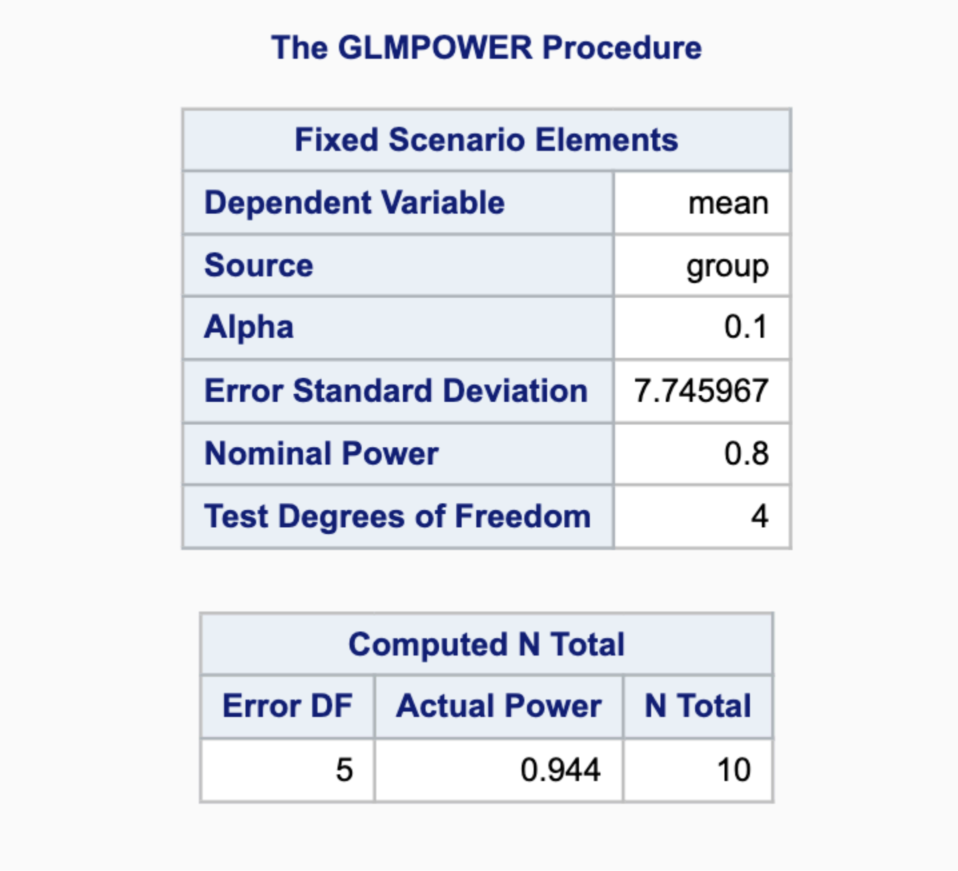
Therefore, the sample size n required to achieve the desired error rates α and β is:

$$n \geq \frac{2\sigma^2(Z_{\alpha/2} + Z_\beta)^2}{\delta^2}$$

Question 2

The total sample size needed to detect a minimum difference of 10 between at least two of the group means needs to be N = 10.

```
data Exemplary;
input group mean;
datalines;
1 -20
2 -10
3 0
4 10
5 20
;
run;
proc glmpower data=Exemplary;
   class group;
   model mean = group;
   power
      stddev = 7.7459666692
      alpha = 0.1
      ntotal = .
      power = 0.8;
run;
```



The GLMPOWER Procedure

Fixed Scenario Elements	
Dependent Variable	mean
Source	group
Alpha	0.1
Error Standard Deviation	7.745967
Nominal Power	0.8
Test Degrees of Freedom	4

Computed N Total		
Error DF	Actual Power	N Total
5	0.944	10

Question 3

3.1

Q-Q Plot of the Residuals

- **Purpose**: Evaluates the normality of residuals; deviations from a straight line suggest non-normal residuals.
- **Our tomato data reveals**: There appears to be non-constant variance present in the treatments. There is two patterns in the treatments that show a visible difference in their variance. It can be suggested that it doesn't fully follow normality.

Histogram of the Residuals

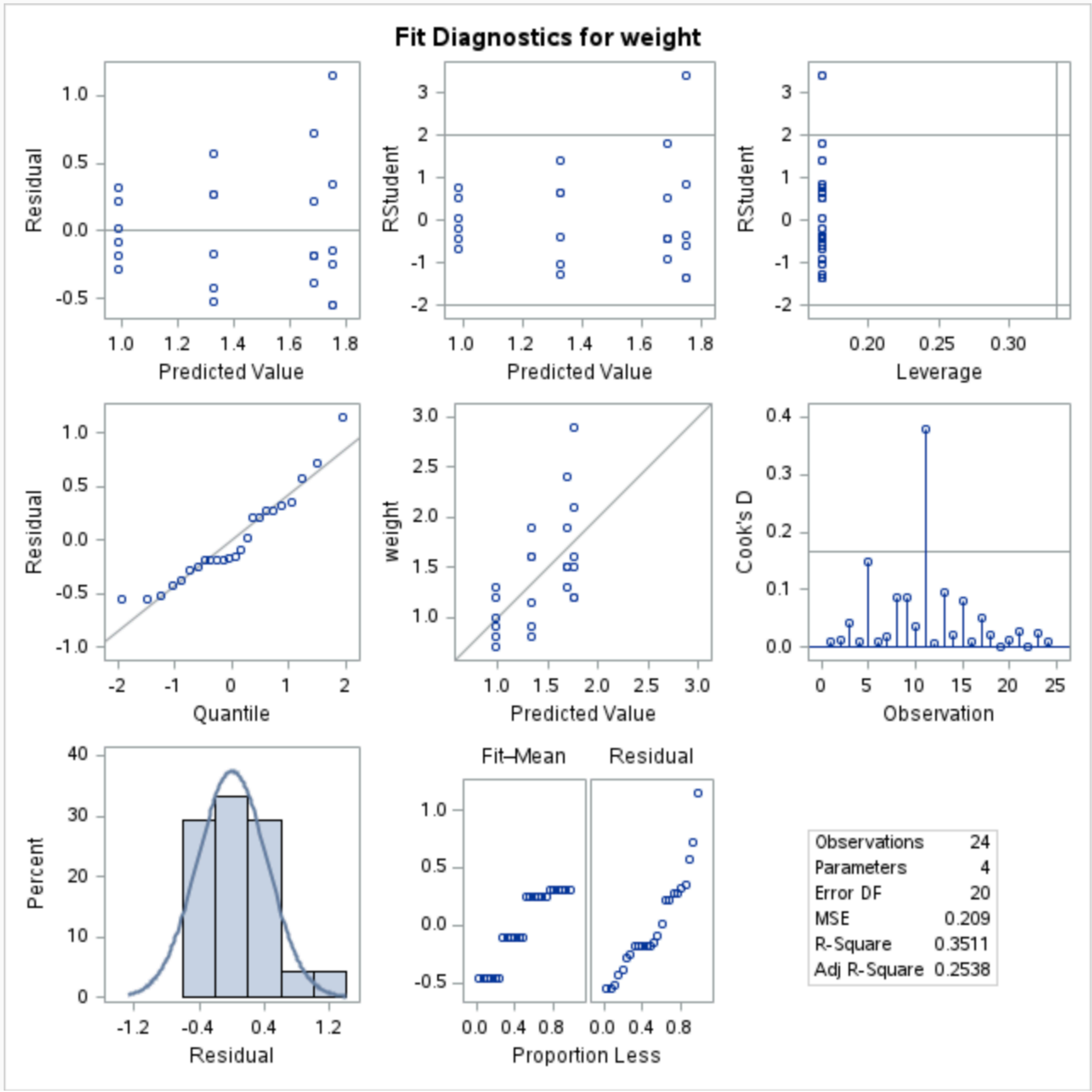
- **Purpose**: Illustrates the distribution of residuals to identify skewness and kurtosis, assessing normality.
- **Our tomato data reveals**: The histogram supports my previous assertions with the qq-plot where it's showing a skewness in the data. Specifically, it's right skewed, which means that the data does not fully follow normality.

Studentized Residuals vs. Predicted Values Plot

- **Purpose*: Tests for homoscedasticity of residuals; patterns may indicate non-constant variance.
- **Our tomato data reveals**: Non-constant variance is present among each treatment group. From trt 1 to trt 3, the spread is more pronounced vertically, again, supporting the consensus that the variance is not constant.

RF Plot

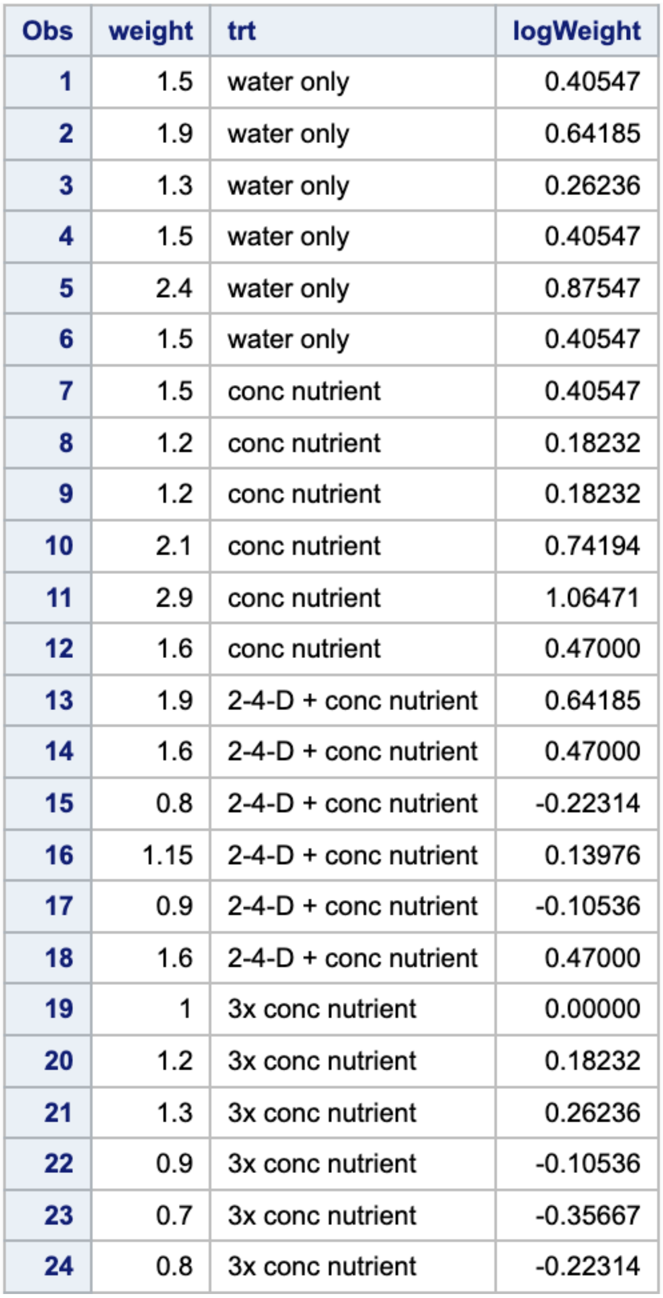
- **Purpose**: Assesses outliers and checks for constant variance of residuals across fitted values.
- **Our tomato data reveals**: There potentially exists one potential outlier, but whether it could be deemed worthy to removed would be left to the researcher.



Observations	24
Parameters	4
Error DF	20
MSE	0.209
R-Square	0.3511
Adj R-Square	0.2538

3.2

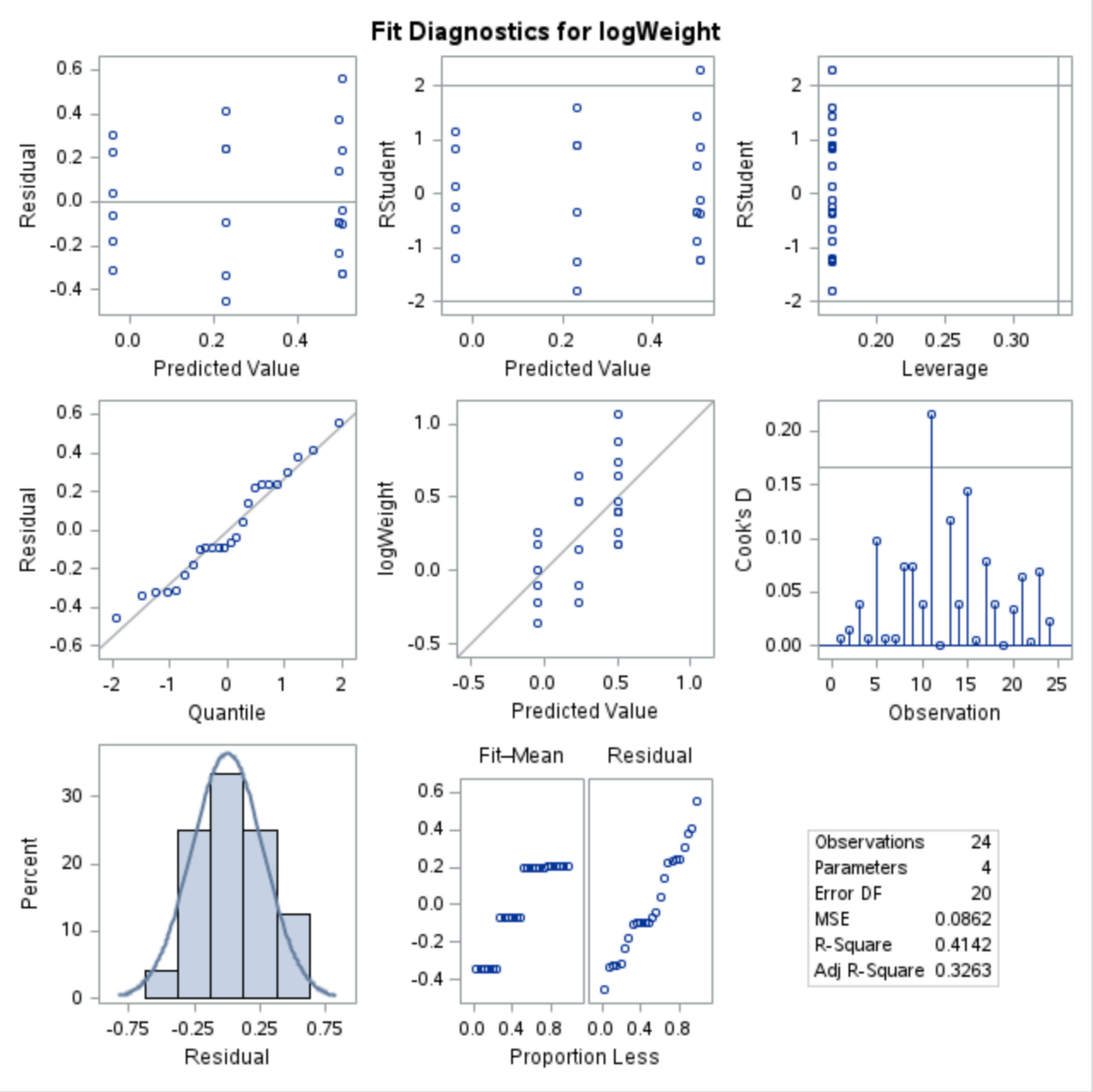
```
data tomato_log;
  set tomato;
  logWeight = log(weight);
run;
proc print data=tomato_log;
run;
```



Obs	weight	trt	logWeight
1	1.5	water only	0.40547
2	1.9	water only	0.64185
3	1.3	water only	0.26236
4	1.5	water only	0.40547
5	2.4	water only	0.87547
6	1.5	water only	0.40547
7	1.5	conc nutrient	0.40547
8	1.2	conc nutrient	0.18232
9	1.2	conc nutrient	0.18232
10	2.1	conc nutrient	0.74194
11	2.9	conc nutrient	1.06471
12	1.6	conc nutrient	0.47000
13	1.9	2-4-D + conc nutrient	0.64185
14	1.6	2-4-D + conc nutrient	0.47000
15	0.8	2-4-D + conc nutrient	-0.22314
16	1.15	2-4-D + conc nutrient	0.13976
17	0.9	2-4-D + conc nutrient	-0.10536
18	1.6	2-4-D + conc nutrient	0.47000
19	1	3x conc nutrient	0.00000
20	1.2	3x conc nutrient	0.18232
21	1.3	3x conc nutrient	0.26236
22	0.9	3x conc nutrient	-0.10536
23	0.7	3x conc nutrient	-0.35667
24	0.8	3x conc nutrient	-0.22314

3.3

Log transforming the tomato data resulted in a slight improvement in model fit. Examining each diagnostic plot reveals subtle yet positive changes due to the transformation. The Q-Q plot showed a marginal improvement toward normality. The histogram indicated a closer adherence to normal distribution, suggestive of log-normal characteristics. The studentized residuals plot displayed a tighter grouping across treatment groups, albeit some non-constant variance patterns persisted. Additionally, the RF plot did not identify any new outliers post-transformation.



Observations	24
Parameters	4
Error DF	20
MSE	0.0862
R-Square	0.4142
Adj R-Square	0.3263

3.4

In the instance where three tomato plants are grown in a single pot, the independence of each plant is compromised because they share resources like nutrients and space. This sharing violates the assumption that each plant grows under conditions unaffected by others, which is crucial for valid experimental design and analysis. As a consequence, the entire pot, rather than individual plants, becomes the experimental unit, requiring different analysis techniques to handle the intertwined growth outcomes.

In other words, this is a direct violation of independence

Question 4

4.1

The vector R can be defined in terms of Y and H as follows:

$$R = Y - HY$$

4.2

Y is $\sigma^2 I$ where I is the identity matrix. Given the matrix A, which is non-random and of appropriate dimension such that AY makes sense, then the variance of AY is calculated:

$$\text{var}(AY) = A\text{var}(Y)A'$$

Since $\text{var}(Y) = \sigma^2 I$, we can substitute and simplify the expression:

$$\text{var}(AY) = A\sigma^2 IA' = \sigma^2 AA'$$

Therefore, the variance-covariance matrix of $R = HY$ is:

$$\text{var}(R) = \sigma^2 HH'$$

4.3

We aim to calculate the matrix H and then determine the variance-covariance matrix of the residuals R, where there are $g = 3$ groups and $n = 2$ observations per group.

Calculation of H

In a balanced CRD:

$$H = \frac{1}{n} J_n$$

where J_n is an $n \times n$ matrix of ones. For $n = 2$, each block of H for a group is:

$$\frac{1}{2} J_2 = \frac{1}{2} \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix} = \begin{bmatrix} \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} \end{bmatrix}$$

H Matrix for the Entire Design

For $g = 3$ groups, the complete H matrix is:

$$H = \begin{bmatrix} \frac{1}{2} & \frac{1}{2} & 0 & 0 & 0 & 0 \\ \frac{1}{2} & \frac{1}{2} & 0 & 0 & 0 & 0 \\ 0 & 0 & \frac{1}{2} & \frac{1}{2} & 0 & 0 \\ 0 & 0 & \frac{1}{2} & \frac{1}{2} & 0 & 0 \\ 0 & 0 & 0 & 0 & \frac{1}{2} & \frac{1}{2} \\ 0 & 0 & 0 & 0 & \frac{1}{2} & \frac{1}{2} \end{bmatrix}$$

Variance-Covariance Matrix of R The variance-covariance matrix of R is:

$$\text{var}(R) = \sigma^2 H$$

Each diagonal element of var(R), representing the variance of the residuals for each observation, is $\sigma^2 \frac{1}{n}$ or $\sigma^2 \frac{1}{2}$ in this example. Thus, the k-th diagonal element of var(R) in any balanced CRD is:

$$\sigma^2 \frac{1}{n}$$