STAT 8120 – Applied Experimental Design

Lab 3 Report – February 13, 2020

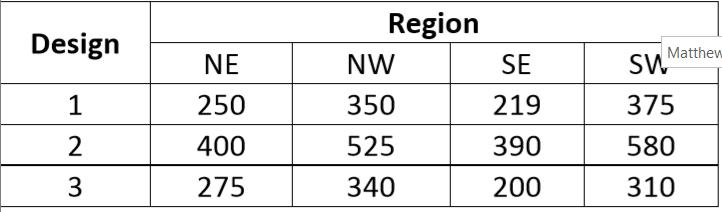
Connor Armstrong

*The purpose of this report is to fulfill the requirement for Module 4 Lab, according to the supplied lab documentation, S8120Lab3d051717.pdf. SAS and Minitab are utilized as analytical tools to address the questions in the lab document.*

***Setting:*** *Blocking is a fundamental tool used in Design of Experiments (DOE) to reduce experimental error. With smaller error, an experiment is more likely to detect treatment mean differences. A Randomized Complete Block Design (RCBD) is used when there is ONE NUISANCE FACTOR. Students can analyze the below problem and determine the advantage of blocking, if any, in this experiment. Students are encouraged to ask additional “what if” questions and simulate their own data that can then be analyzed.*

***Problem 4.8:*** *A consumer products company relies on direct mail marketing pieces as a major component of its advertising campaigns. The company has THREE different designs for a new brochure and wants to evaluate their effectiveness, as there are substantial differences in costs between the three designs. The company decides to test the three designs by mailing 5000 samples of each to potential customers in FOUR REGIONS of the country. Since there are known regional differences in the customer base, regions are considered as blocks. The number of responses is:*

***Table 1***



JUSTIFY all conclusions with good statistical analysis. Use Minitab for questions 1 and 2.

1) Do the designs differ? Perform a residual analysis.

A RCBD ANOVA was processed using Minitab. The results are below:

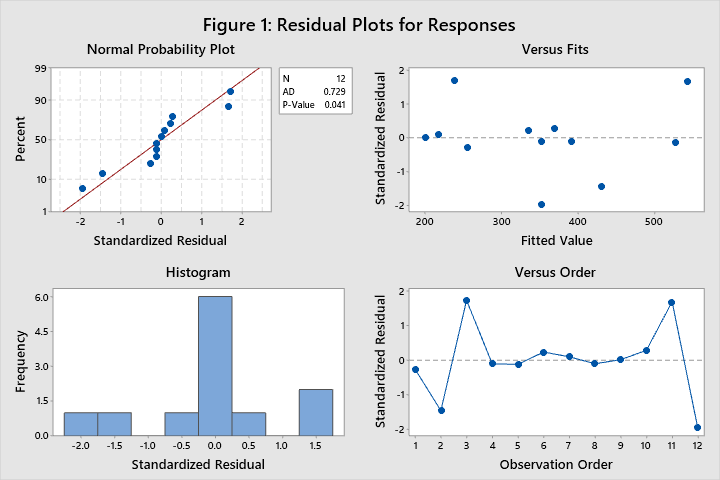
**Table 2 Analysis of Variance**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Source** | **DF** | **Adj SS** | **Adj MS** | **F-Value** | **P-Value** |
| Design | 2 | 90755 | 45377.6 | 50.15 | 0.000 |
| Region | 3 | 49036 | 16345.2 | 18.06 | 0.002 |
| Error | 6 | 5429 | 904.8 |  |  |
| Total | 11 | 145220 |  |  |  |

Having p-values below the standard significance level of 0.05, it is likely that both Design and Region influence the mean of the sampled population and that the means of the respective groups are not equivalent, subject to the validation of the assumptions.

Assumption 1, Normality: Referring to Figure 1, the data does not pass the A.D. test for normality with a p-value below 0.05 (0.041). It is likely that a transformation can be implemented to rectify this discrepancy. There is a “U” pattern in the versus fits plot in figure 1, which is indicative of a non-linear relationship warranting a transformation of the data.

The conclusion drawn from the ANOVA table above cannot be accepted given a violation of assumption 1, therefore there is no reason to continue validating the assumptions.



2) Justify a transformation of the data and repeat question 1. Which design is best?

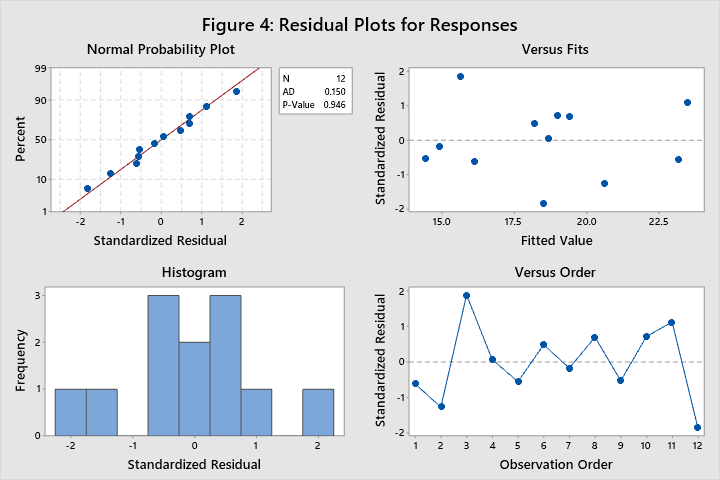
Having violated the normality assumption as discussed in question 1, and given a clear “U” pattern in the versus fits plot in figure 1, there is sufficient evidence to recommend that a transformation will likely improve the model.

The data was transformed using a square root transformation on the response variable. The p-value for the A.D. test in this case is 0.946, which is a strong indicator that the correct model was selected for the analysis. The ANOVA table for the the transformed data is below:

**Table 3: Analysis of Variance for Square Root Transformed Response**

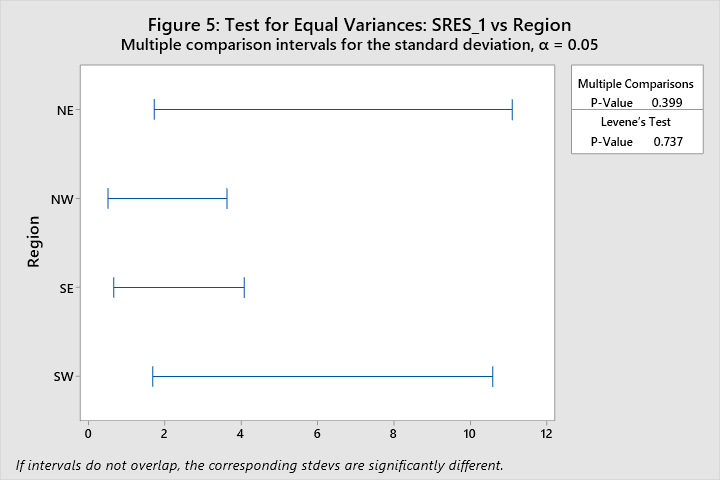
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Source** | **DF** | **Adj SS** | **Adj MS** | **F-Value** | **P-Value** |
| Design | 2 | 60.733 | 30.3666 | 60.47 | 0.000 |
| Region | 3 | 35.891 | 11.9638 | 23.82 | 0.001 |
| Error | 6 | 3.013 | 0.5022 |  |  |
| Total | 11 | 99.638 |  |  |  |

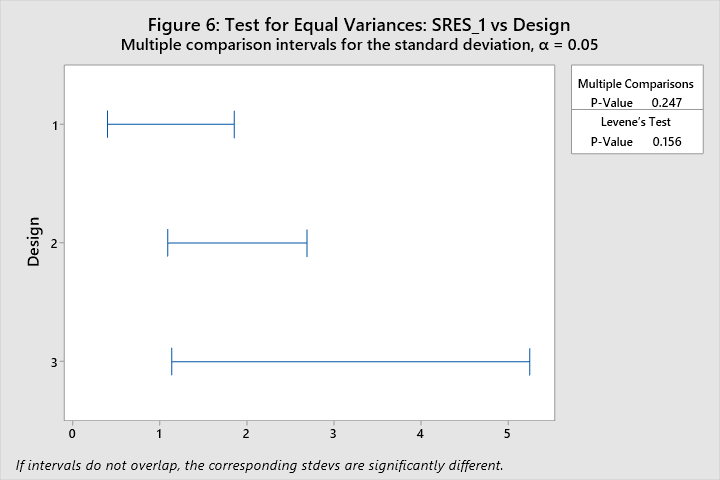
Having p-values well below the standard significance level of 0.05, it is likely that both design and region impact the mean response, subject to verification of the assumptions.



Assumption 1, Normality: Having a p-value of 0.946 for the A.D. test for normality, the data is likely normal.

Assumption 2, Homogeneity of Variance: viewing the variances associated with both region and design in figures 5 and 6, both pass Levene’s test of Homogeneity of Variance with p-values greater than the standard significance level of 0.05.





Assumption 3, Outliers: There are no outliers beyond +/- 3 standard deviations from the expected value as shown in the versus fits plot in figure 4.

Assumption 4, Independence: There is no run order information given in this dataset so Independence will remain unvalidated for this exercise.

3) Use SAS to repeat questions 2. Analyze the transformed data. Give SAS code, do not repeat any discussion.

Table 4

| **Class Level Information** | | |
| --- | --- | --- |
| **Class** | **Levels** | **Values** |
| **Region** | 4 | NE NW SE SW |
| **Design** | 3 | 1 2 3 |

Table 5

| **Source** | **DF** | **Sum of Squares** | **Mean Square** | **F Value** | **Pr > F** |
| --- | --- | --- | --- | --- | --- |
| **Model** | 5 | 96.62458259 | 19.32491652 | 38.48 | 0.0002 |
| **Error** | 6 | 3.01314672 | 0.50219112 |  |  |
| **Corrected Total** | 11 | 99.63772931 |  |  |  |

Table 6

| **R-Square** | **Coeff Var** | **Root MSE** | **rootres Mean** |
| --- | --- | --- | --- |
| 0.969759 | 3.827136 | 0.708654 | 18.51657 |

Table 7

| **Source** | **DF** | **Type I SS** | **Mean Square** | **F Value** | **Pr > F** |
| --- | --- | --- | --- | --- | --- |
| **Region** | 3 | 35.89140739 | 11.96380246 | 23.82 | 0.0010 |
| **Design** | 2 | 60.73317520 | 30.36658760 | 60.47 | 0.0001 |

Figure 7



Figure 8



Table 8

| **Tests for Normality** | | | | |
| --- | --- | --- | --- | --- |
| **Test** | **Statistic** | | **p Value** | |
| **Shapiro-Wilk** | **W** | 0.986474 | **Pr < W** | 0.9981 |
| **Kolmogorov-Smirnov** | **D** | 0.111638 | **Pr > D** | >0.1500 |
| **Cramer-von Mises** | **W-Sq** | 0.023606 | **Pr > W-Sq** | >0.2500 |
| **Anderson-Darling** | **A-Sq** | 0.150169 | **Pr > A-Sq** | >0.2500 |

Figure 9



Figure 10



Figure 11



SAS Code for Question 3

/\*

STAT 8120 - Module 4 Lab

\*/

libname mod4 "C:\Users\conno\OneDrive\Desktop\STAT 8120 - Applied Experimental Design\Module 4";

**run**;

**proc** **import** datafile = "C:\Users\conno\OneDrive\Desktop\STAT 8120 - Applied Experimental Design\Module 4\S8120Ch4Data122317.xlsx"

out = mod4.q3

DBMS = xlsx

Replace;

sheet = "Lab 3";

**run**;

**data** mod4.q3root;

set mod4.q3;

rootres = sqrt(responses);

**run**;

ods rtf;

ods graphics on;

**proc** **glm** data = mod4.q3root plots=diagnostics;

class region design;

model rootres = region design;

output out = stdres student = stdresidual;

Title "SAS RCBD for Example 4.1";

**run**;

Means region design / tukey;

**run**;

Contrast "Lowest vs. All Others" design **3** -**1** -**1** -**1**;

**run**;

**proc** **univariate** data = stdres normal;

var stdresidual;

qqplot stdresidual / normal(mu=est sigma=est);

histogram/normal;

**run**;

**proc** **sgplot** data = stdres;

scatter x=region y=stdresidual;

**run**;

ods graphics off;

ods rtf close;

**quit**;

4) What is the advantage of blocking for this experiment? Ignore the Region blocking factor and analyze the transformed data answering question 1. What is the % increase in experimental error (𝜎 ̂)?

The data from table 1 was analyzed without blocking by the region variable. The results are in the table 9 below:

**Table 9: Analysis of Variance**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Source** | **DF** | **Adj SS** | **Adj MS** | **F-Value** | **P-Value** |
| Design | 2 | 60.73 | 30.367 | 7.02 | 0.015 |
| Error | 9 | 38.90 | 4.323 |  |  |
| Total | 11 | 99.64 |  |  |  |

Knowing that (S8120SG3D022618.docx, Page 4), . Similarly, . Thus, by not including the blocking factor in the analysis, there is a [ increase in experimental error.