Understanding Analysis of Variance

Best Practice

Authored by: Cory Natoli

21 December 2017

Revised 24 September 2018

The goal of the STAT COE is to assist in developing rigorous, defensible test strategies to more effectively quantify and characterize system performance and provide information that reduces risk. This and other COE products are available at www.afit.edu/STAT.

STAT Center of Excellence
2950 Hobson Way – Wright-Patterson AFB, OH 45433

Table of Contents

Revision 1, 24 Sep 2018: Formatting and minor typographical/grammatical edits.

Executive Summary

An important step in the DOE process is determining which factors truly affect the (numeric) response variable. A solution to this problem is through the use of analysis of variance (ANOVA). ANOVA is a procedure that uses hypothesis testing to determine whether the factor effects of two or more factors are the same. This paper seeks to explain the basic statistical theory behind one-way ANOVA, as well as detail the process and how to utilize ANOVA conceptually. Code is provided to perform ANOVA in R and JMP.

Keywords: ANOVA, DOE, statistically significant, hypothesis testing, R, JMP

Introduction

Analysis of Variance (ANOVA) is a common technique for analyzing the statistical significance of a number of factors in a model. The overall goal of ANOVA is to select a model that only contains terms that add valuable insight in determining the value of the response, or in other words, a model that only includes statistically significant terms. Assuming the data comes from a designed experiment, the analysis will answer the question "Does this (do these) factor(s) cause a statistically significant difference in the response variable?" An important step in the design of experiments (DOE) process is determining which variables truly affect the response, and ANOVA allows us to do this when the response is continuous. In order to understand the concepts of ANOVA, this paper begins with an introduction to ANOVA when there is only one factor (one-way ANOVA), and then connects the concepts to multifactor ANOVA. ANOVA can only be used if the response variable is quantitative; it will not work for qualitative responses. All factors in the model are treated as if they are qualitative (categorical) variables. The results tell us whether or not there is a difference in the average response at the various levels of the factor. If the data do not fulfill these requirements (i.e., the response is qualitative, the goal is a prediction, etc.), it is possible that regression is better suited for the analysis as it is a more general method.

Before diving into the concepts, we first establish the notation associated with ANOVA.

- $\bar{x}$ is the mean of a sample of data, $x_1, x_2, \ldots, x_n$; the bar simply denotes "arithmetic average."
- $\hat{x}$ is the estimated value for x; the hat denotes that it is a predicted value from data.
- $x.$ is the value of the x values summed over all levels of the factor in which the "." replaces.

Formulas are provided for all of the values discussed to further increase conceptual understanding, but the values can be calculated using statistical software (examples shown in detail later).

Method

One-Way ANOVA

Motivating example

We use a simple example to demonstrate the one-way ANOVA scenario. Suppose there is a new missile being tested and we wish to characterize the miss distance. In this scenario, the miss distance is the response variable. We determine there is only one factor, distance to the target, that affects the miss distance. Therefore, we wish to determine if this solitary factor has a statistically significant effect on the miss distance.

Table 1. Generic data table for one factor experiment

Factor Level	Observations				Totals	Averages
1	y_{11}	y_{12}	$\cdots$	y_{1n}	$y_{1.}$	$\bar{y}_{1.}$
2	y_{21}	y_{22}	$\cdots$	y_{2n}	$y_{2.}$	$\bar{y}_{2.}$
$\vdots$	$\vdots$	$\vdots$	$\cdots$	$\vdots$	$\vdots$	$\vdots$
v	y_{v1}	y_{v2}	$\cdots$	y_{vn}	$y_{v.}$	$\bar{y}_{v.}$
					$y_{..}$	$\bar{y}_{..}$

Model

In order to conduct ANOVA, one must first establish the mathematical model. The model for one-way ANOVA is as follows:

$$Y_{it} = \mu + \tau_i + \epsilon_{it} \text{ where } 1 \le t \le r_i\,, 1 \le i \le v\,, \epsilon \sim iid\ N(0, \sigma^2)$$

- Y_{it} is the t^{th} observation at the i^{th} level of the factor
- r_i is the number of observations in factor level i
- μ is the overall mean of the response
- τ_i is the effect of the ith factor level on the response
- ϵ_{it} is the random error, assumed to be independently and identically normally distributed with mean 0 and constant variance σ^2 ($iid\ N(0, \sigma^2)$)
- v is the number of factor levels being tested

The factor effect τ_i measures the change in the response due to the i^{th} factor level. These factor effects are parameters in the model that must be estimated using the test data.

Model Building

The model includes a single term for the factor for which we wish to determine the significance through the use of ANOVA. Since there is just one factor of interest, this is the simplest case of ANOVA. If the term is significant, we leave the model as written above and it can be used as the empirical model for the system. However, if the term is determined not to be significant, we can reduce the model to

$$Y_{it} = \mu + \epsilon_{it}$$

In other words, the model to predict the response no longer includes a factor effect – the response is simply the mean plus some random error. This paper begins by introducing the concepts that allow a decision to be made on the significance of a single factor and then extends the process to the more realistic scenario where a system is characterized by many factors.

Assumptions

ANOVA requires a set of assumptions to be met or the results will not be valid. The assumptions are:

- Error terms are independent, normally distributed with mean 0 and variance σ^2
- All levels of each factor follow a normal distribution
- Variance, σ^2, is the same across all levels of each factor
- Observations are independent

These are the same assumptions that are necessary for building a regression model. More information on these can be found in "The Model Building Process Part 1: Checking model Assumptions" best practice (Burke 2017).

Hypotheses

If the factor in the experiment has v levels, we can use ANOVA to test for equality of factor effects. The null hypothesis states that the factor has no effect on the response as indicated by the fact that all factor terms in the model vanish:

$$H_0\!: \{\tau_1 = \tau_2 = \cdots = \tau_v = 0\}$$

The alternative hypothesis states that the factor has an effect on the response because at least one factor term in the model is non-zero:

$$H_a\!: \{Not\ all\ \tau_i\ equal\ 0\}$$

The concept of ANOVA revolves around comparing the "within" factor variance with the "between" factor mean variances. This comparison is made by partitioning the total variability of the response into two important components: sum of squares for error (SSE) and sum of squares for factors (SSR). SSR measures the difference between factor level means; SSE, the difference within a factor level to the factor level mean can only be due to random error (Montgomery, 2017).

The total sum of squares (SST) measures the total variability in the response and is defined as

$$SST = \sum_{i=1}^{v} \sum_{t=1}^{r_i} (y_{it} - \bar{y}_{..})^2$$

SST can be decomposed into SSE and SSR. SSE, the "within" factor variance, is calculated as

$$SSE = \sum_{i=1}^{v}\sum_{t=1}^{r_i}(y_{it} - \bar{y}_{i\cdot})^2$$

where $\bar{y}_{i.} = \frac{\sum_{t=1}^{r_i} y_{it}}{r_i}$. SSE is also called the error sum of squares.

SSR, the “between” factor variance, is the value “used in describing how well a model… represents the data being modeled” (*Explained sum of squares 2017*). SSR is calculated as

$$SSR = \sum_{i=1}^{v} r_i(\bar{y}_{i.} - \bar{y}_{..})^2$$

where $\bar{y}_{..} = \frac{\sum_{i=1}^{v}\sum_{t=1}^{r_i} y_{it}}{n} = \frac{y_{..}}{n}$. SSR is the variability between the means of each factor level.

Figure 1 shows a visual decomposition of these sums of squares. In the figure, the blue line represents the predicted values for a given model. The black dots are the actual data points. SSE is calculated from the difference between the actual point and the predicted value, SSR is calculated from the difference between the predicted value and the grand mean, and the sum of squares total, SST, is the sum of SSE and SSR. The estimated values for μ and τ are the values that minimize the SSE.

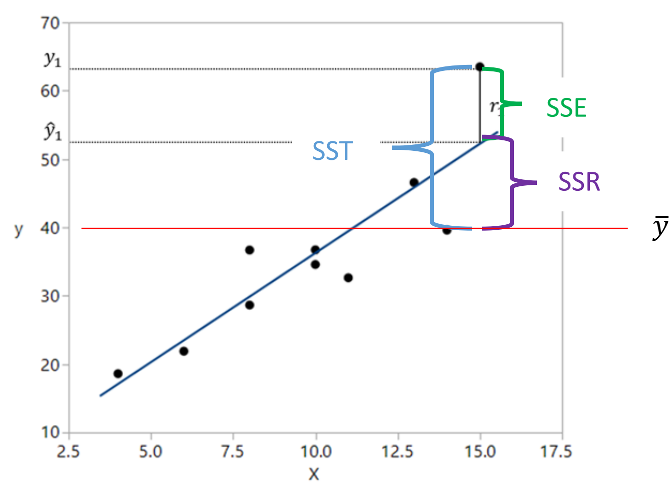


Figure 1: Decomposing Sum of Squares

Test Statistic

SSE and SSR are used to conduct the hypothesis test on the factor effects. The ANOVA framework uses an F-statistic to compare the means of the different factor levels. The idea behind the F-statistic is to compare the ratio of the between factor variability vs. the within factor variability ($F = \frac{"between"\ factor\ variance}{\text{"within" factor variance}}$) in order to make a decision on a difference in means.

Figure 2 shows the difference between the “between” and “within” variances, as well as a number of different scenarios for large and small “between” factor variance and “within” factor variance. The “between” variance is the variance between the three different curves. The “within” variance is the variance within a single curve. Essentially, if the “between” variance is much larger than the “within” variance, the factor is considered statistically significant. Recall, ANOVA seeks to determine a difference in means at each level of a factor. If the factor level impacts the mean, then that factor is statistically significant. The image in the top right shows the most obvious case of this instance. When the difference in the means is large, this implies the “between” group variation is large. The variance within groups is also small, so we can clearly identify the effect on the response caused by that factor. Conversely, the bottom left demonstrates a scenario with a larger “within” factor variance and a small “between” factor variance, as the means are close together and there is a lot of overlap in the distributions. This may be a scenario in which the factor is not statistically significant.

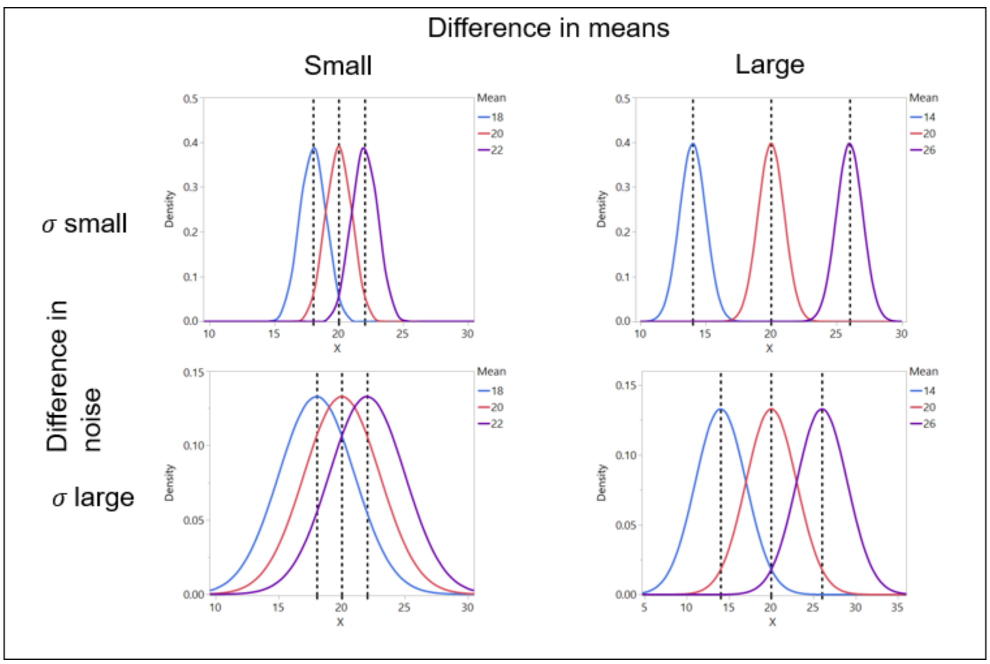


Figure 2: Conceptual ANOVA

Decisions are hardest to make when there is large variance within factors or the means between the factor levels are close. If there is a large variance within factors, there is a lot of overlap between the distributions and it makes determining significance much more difficult. Also if the means are close, it would follow that stating the means are different is difficult.

Decision

A large value for the F statistic suggests evidence against the null hypothesis. The natural decision is to then reject the null hypothesis. Conversely, a small value for the F statistic suggests the data support the null hypothesis and the decision is to fail to reject the null hypothesis.

$$Reject\ H_0\ if\ \frac{SSR/(v-1)}{SSE/(n-v)} > F_{\alpha,v-1,n-v}$$

Where $F_{\alpha,v-1,n-v}$ represents a cut-off value at α-level of significance. We can also make a decision based on the p-value. Figure 3 shows how a p-value is calculated for an F-distribution. The p-value is the probability of observing the test statistic or something more extreme given the null hypothesis is true. The probability shaded in the figure represents the p-value. A p-value smaller than the significance level (α level, set prior to the experiment, often 0.1 or 0.05) provides evidence against the null hypothesis and leads to concluding the alternative hypothesis is true (at least one factor effect is nonzero). If the p-value is larger than α, the decision is to fail to reject the null hypothesis.

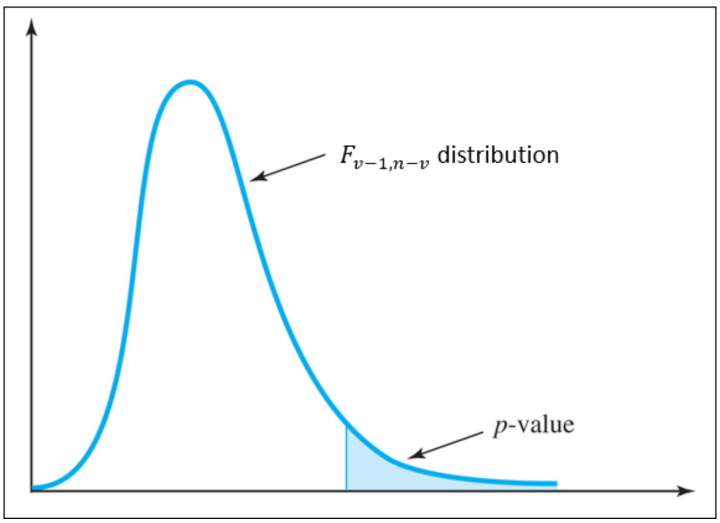


Figure 3. F distribution

Most commonly, calculations associated with ANOVA are presented in an ANOVA table. Table 2 shows a generic ANOVA table.

Table 2. Generic ANOVA Table

Source	Degrees of Freedom	Sum of Squares	Mean Square	F-Ratio	P-Value
Factors	$v-1$	SSR	MSR = SSR/$v-1$	MSR/MSE	Calculated
Error	$n-v$	SSE	MSE = SSE/ $n-v$		
Total	$n-1$	SST			

- n is the total sample size and is equal to $\sum_{i=1}^{v} r_i$
- P-value is calculated by software (see http://stattrek.com/online-calculator/f-distribution.aspx for technical details)

Multifactor ANOVA

Motivating example

Continuing with the missile example, suppose we determine there are two factors that may affect the miss distance. Now, instead of simply looking at the distance from the target, we are also interested in the height of the launch. Our model must be modified to reflect this change.

Model

Once again, the model is the foundation of all calculations. Terms are added to the one-factor model to accommodate the main effects of additional factors, and more terms may be added to include interactions between factors. After adding the main effect term for the new factor and a term for factor interaction, our new two-factor model is

$$Y_{ijt} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \epsilon_{ijt} \; where \; 1 \le t \le r_i \,, 1 \le i \le a \,, 1 \le j \le b \,, \epsilon \sim iid \; N(0, \sigma^2)$$

- Y_{ijt} is the t[th] observation at the i[th] factor A level and j[th] factor B level.
- μ is the overall mean
- α_i is the factor effect of factor A at level I; $\sum_{i=1}^{a} \alpha_i = 0$
- β_j is the factor effect of factor B at level j; $\sum_{j=1}^{b} \beta_j = 0$
- $(\alpha\beta)_{ij}$ represents the interaction effect of factors A and B; $\sum_i (\alpha\beta)_{ij} = 0 \,, j = 1, \ldots, b$; $\sum_j (\alpha\beta)_{ij}, i = 1, \ldots, a$
- ϵ_{ijt} is the random error

This model can be extended to three or more factors by including the additional corresponding terms.

$Y_{ijk\ldots t} = \mu$ Overall mean

$+\alpha_i + \beta_j + \gamma_k + \cdots$ These factors make up the main effects model

$+ (\alpha\beta)_{ij} + (\alpha\gamma)_{ik} + (\beta\gamma)_{jk} + \cdots$ 2-factor interactions

$+ (\alpha\beta\gamma)_{ijk} \ldots$ 3-factor interactions

… Higher-order interactions (dependent on number of factors)

$$+ \epsilon_{ijt} \; where \; 1 \le t \le r_i \,, 1 \le i \le a \,, 1 \le j \le b, 1 \le k \le c \,, \epsilon \sim iid \; N(0, \sigma^2)$$

Similar to the one-way ANOVA scenario, the model includes terms for the factors, or factors that we wish to determine significance for through the use of ANOVA. In this situation, we are able to perform a hypothesis test on each term, both for main effects and interactions, to determine significance. In order to maintain model hierarchy, we never want to remove a lower order term that is significant in a higher order interaction term. For example, we would not remove α_i if the model contained $(\alpha\beta)$, the interaction between A and B. The idea is to build the best fitting, and easiest to interpret, model possible. We test each term and reduce the model as necessary until this goal is achieved.

For conceptual understanding, this paper focuses on the two-factor model.

Hypotheses

The hypotheses for multifactor ANOVA have the same simple interpretation, but are done for all the main effects and interaction effects. The hypotheses for the main effect terms are as follows:

For factor A – $H_0: Factor\ A\ has\ NO\ effect\ on\ the\ response$

For factor B – $H_0: Factor\ B\ has\ NO\ effect\ on\ the\ response$

Both share the same alternative hypothesis of $H_a: The\ factor\ HAS\ an\ effect\ on\ the\ response.$

The hypothesis for an interaction term is as follows:

$$H_0: \text{The interaction effect AB has NO effect on the response}$$

$$H_a: \text{The interaction effect AB HAS an effect on the response}$$

Decision

0nce again, there is a decision rule based off the F-distribution that is

$$\text{Reject } \mathrm{H}_0 \text{ if } \frac{SSE_{factor}/df_{factor}}{SSE/df_E} \sim F_{df_{factor}, df_E} > F_{(a-1)(b-1)(r-1)ab,\alpha}$$

Where df is the degrees of freedom. Degrees of freedom are the number of independent elements in that sum of squares (Montgomery, 2017). As before, the F-ratios will be converted into p-values to assist the decision. Once again, we will use software to calculate the ANOVA table. The two-factor ANOVA table is shown in Table 3.

Table 3. ANOVA table for a two-factor study

Source	Degrees of Freedom	Sum of Squares	Mean Square	F-Ratio	P-Value
Factor A	$a-1$	SSA	$MSA = SSA/(a-1)$	MSA/MSE	Calculated
Factor B	$b-1$	SSB	$MSB = SSB/(b-1)$	MSB/MSE	Calculated
AB	$(a-1)(b-1)$	SSAB	MSAB = SSAB/$(a-1)(b-1)$	MSAB/MSE	Calculated
Error	$n-ab$	SSE	MSE = SSE/$(n-ab)$		
Total	$n-1$	SST			

- P-value will be calculated by software or see http://stattrek.com/online-calculator/f-distribution.aspx for technical details

After the ANOVA process, the model can be simplified by removing terms that are not statistically significant. However, a lower order term cannot be removed from the model if it is included in a higher order term that will be retained. For example, if Factor B has a large p-value which indicates it is not significant, but AB has a very small p-value that indicates it is significant, then Factor B cannot be removed from the model.

Example

Consider a sensor assessment study with two factors, slant range and look down angle, which may impact the target location error (TLE). This is an example of a two-factor ANOVA.

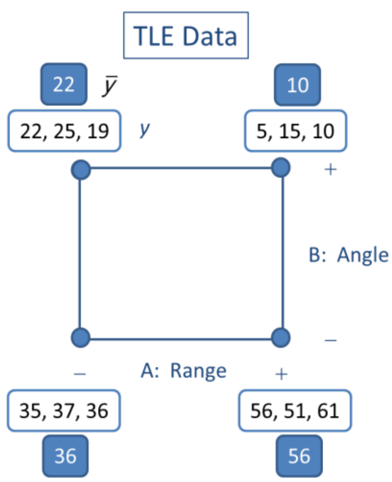


Figure 4: summary of data for a TLE test with two factors, slant range and look down angle

The dataset has 12 observations consisting of 4 observations at different combinations of look down angle and slant range with 2 additional replicates. The model we wish to analyze is:

$$Y_{ijt} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \epsilon_{ijt} \ where\ 1 \le t \le 3, 1 \le i \le 2\ , 1 \le j \le 2\ , \epsilon \sim iid\ N(0, \sigma^2)$$

- μ is the overall mean

- α_i is the factor effect of slant range at level i
- β_j is the factor effect of look down angle at level j
- $(\alpha\beta)$ is the interaction of slant range and look down angle
- ϵ_{it} is the random error

We will use both R and JMP to analyze the data at an alpha=0.05 level.

R Code

- Load the Data. This example loads a comma separated values file with the data.

```
TLE.example <- read.csv("~/Documents/TLE example.csv")
```

- Check that the data has loaded properly

```
TLE.example

## Angle Range TLE
## 1 1 -1 22
## 2 1 -1 25
## 3 1 -1 19
## 4 1 1 5
## 5 1 1 15
## 6 1 1 10
## 7 -1 1 56
## 8 -1 1 51
## 9 -1 1 61
## 10 -1 -1 35
## 11 -1 -1 37
## 12 -1 -1 36
```

- Attach the dataset to the variables, otherwise all variables must have TLE.example$VariableName. This step is optional but helps keep the code cleaner and saves time.

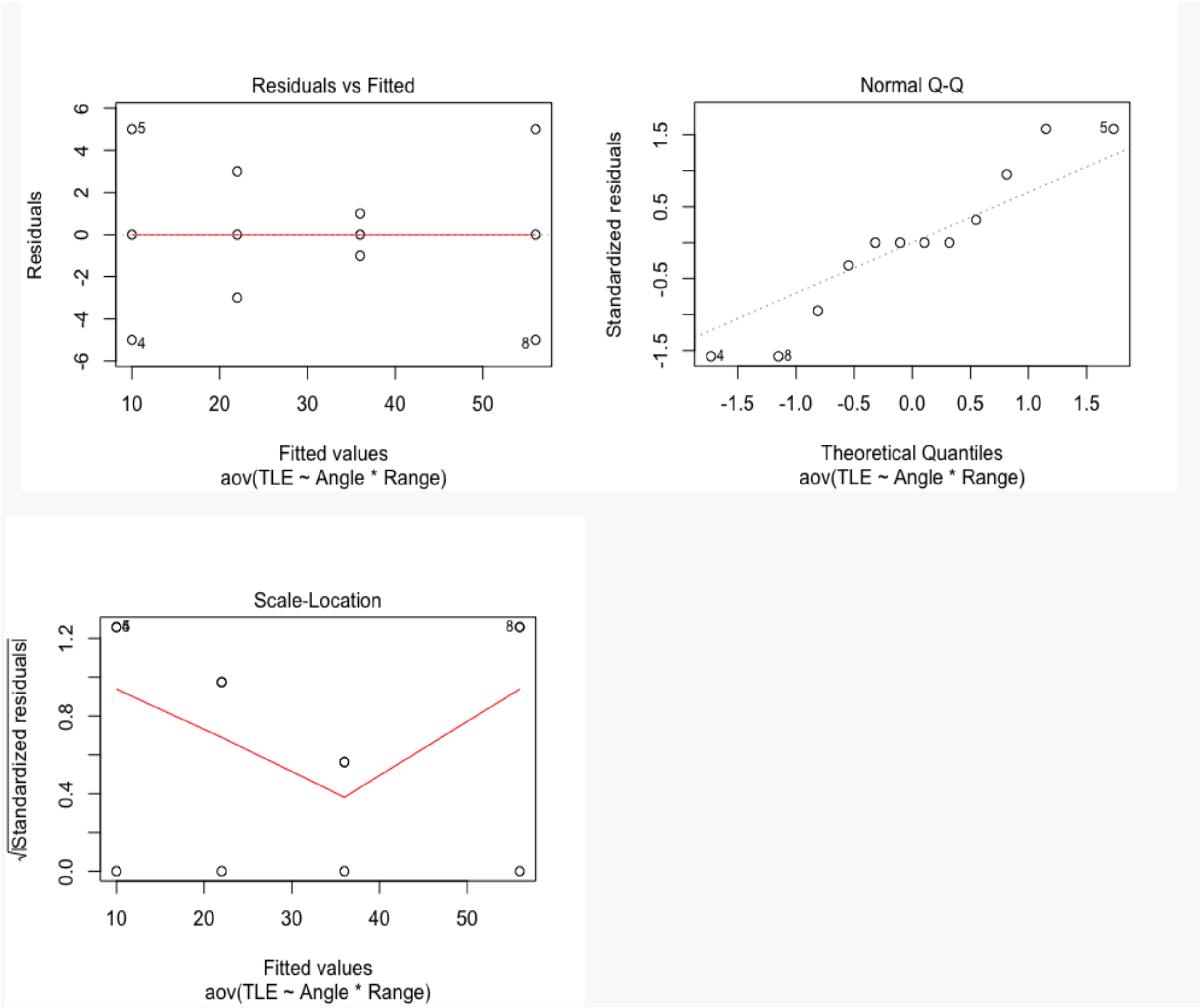
```
attach(TLE.example)
```

- Fit the model. This can also be coded as TLEfit<-aov(TLE~Angle + Range + Angle:Range, data=TLE.example)

```
TLEfit <- aov(TLE ~ Angle*Range, data=TLE.example)
```

- Check the assumptions

```
plot(TLEfit)
```



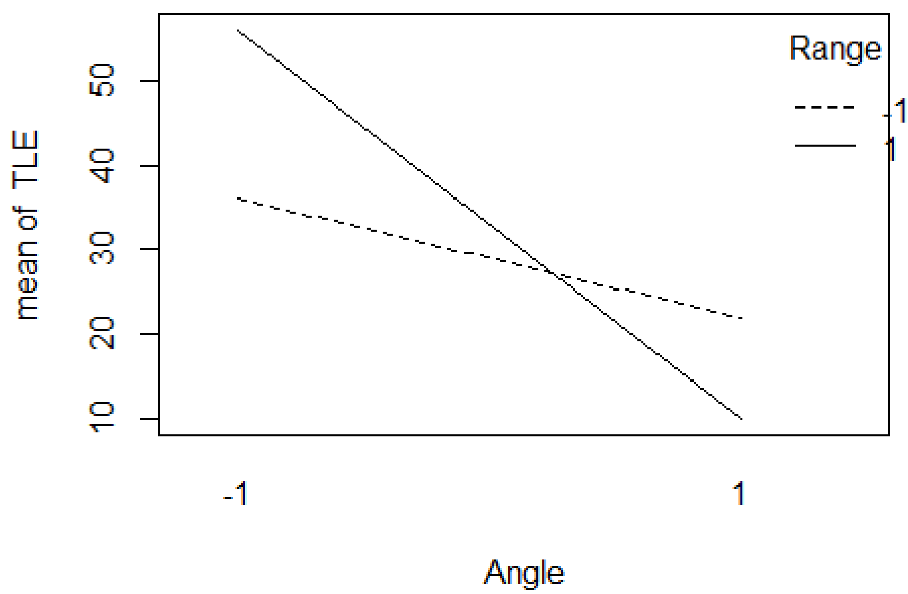
- Create the ANOVA table

```
anova(TLEfit)

## Analysis of Variance Table
##
## Response: TLE
## Df Sum Sq Mean Sq F value Pr(>F)
## Angle 1 2700 2700 180.0 9.123e-07 ***
## Range 1 48 48 3.2 0.1114
## Angle:Range 1 768 768 51.2 9.658e-05 ***
## Residuals 8 120 15
## ---
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We can note the p-values [Pr(>F) in R] to determine which factors are significant in the model. Both Angle and Angle*Range, the two-factor interaction, are significant with small P-values, while Range is not significant. However, Range should not be removed from the model since the higher order interaction Angle*Range is significant.

- Show the coefficients and the interaction plot

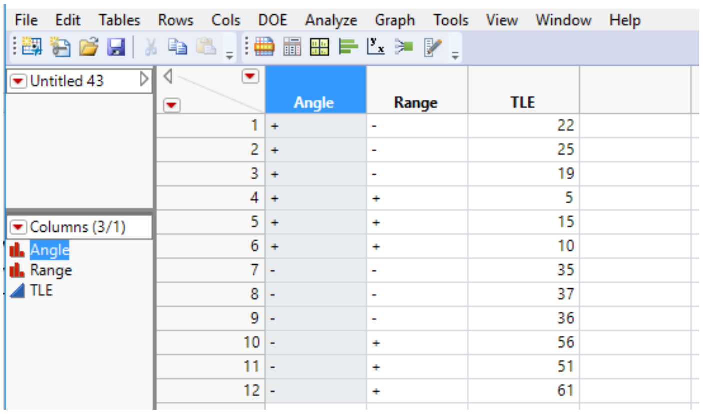
```
TLEfit$coefficients

## (Intercept) Angle Range Angle:Range
## 31 -15 2 -8

interaction.plot(Angle, Range, TLE)
```



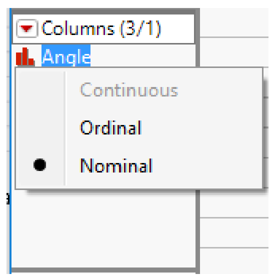
From the ANOVA table, we would then conclude that our final model should be $Y_{ijt} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \epsilon_{ijt}$. The interaction plot is another way to show that keeping the interaction term is necessary as the lines are not parallel. The coefficients allow us to fully construct the model, $\hat{Y} = 31 - 15(Angle) + 2(Range) - 8(Angle * Range)$. Recall, these are in coded -1 and 1 units.

JMP

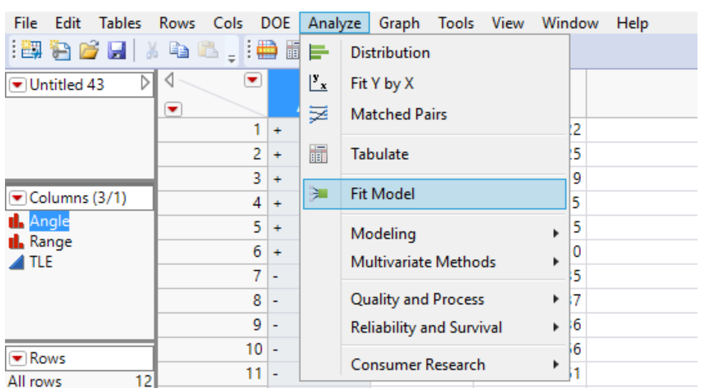
1. Enter the data into JMP



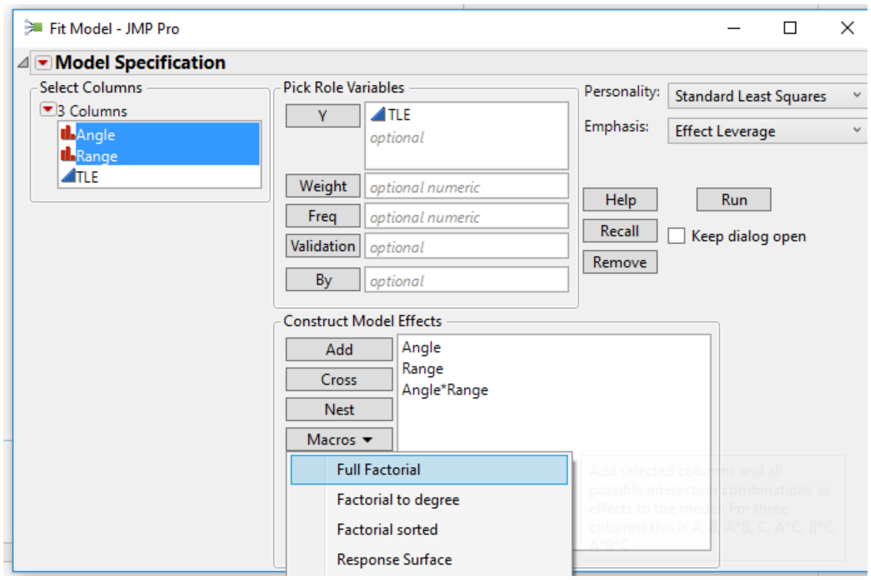
2. Check that Angle and Range are nominal variables by left clicking the symbol next to the variable.



3. Select "Analyze -> Fit model"



4. Enter TLE into the Y by selecting the variable and then clicking "Y". Enter the full factorial of the responses into construct model effects by highlighting Angle and Range, then selecting "Macros" and selecting the full factorial option.



5. Select "Run"
6. Observe the ANOVA, Parameter estimates, and effects test tabs for important p-values. The ANOVA table p-value tells if there is a difference in means due to any factor. The other two tabs further break it down to show which factor is causing the difference.

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	3	3516.0000	1172.00	78.1333
Error	8	120.0000	15.00	**Prob > F**
C. Total	11	3636.0000		<.0001 *

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	31	1.118034	27.73	<.0001 *
Angle	-15	1.118034	-13.42	<.0001 *
Range	2	1.118034	1.79	0.1114
Angle*Range	-8	1.118034	-7.16	<.0001 *

Effect Tests

Source	Nparm	DF	Sum of Squares	F Ratio	Prob > F
Angle	1	1	2700.0000	180.0000	<.0001 *
Range	1	1	48.0000	3.2000	0.1114
Angle*Range	1	1	768.0000	51.2000	<.0001 *

7. *Optional* If needed, remove any non-significant terms using the effect summary. Highlight the term that needs to be removed and select "Remove" at the bottom of the window.

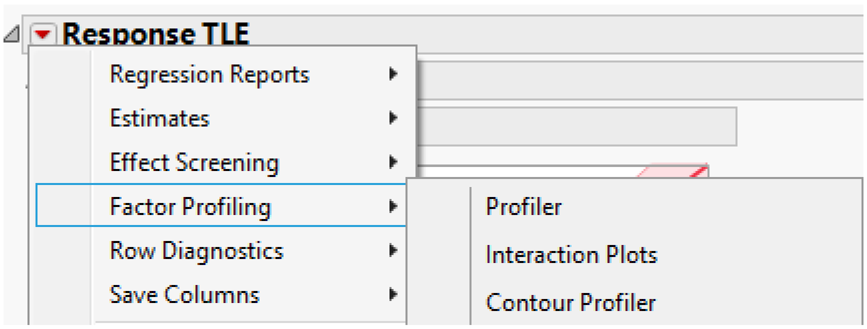
Effect Summary

Source	LogWorth	PValue
Angle	6.040	0.00000
Angle*Range	4.015	0.00010
Range	0.953	0.11143 ^

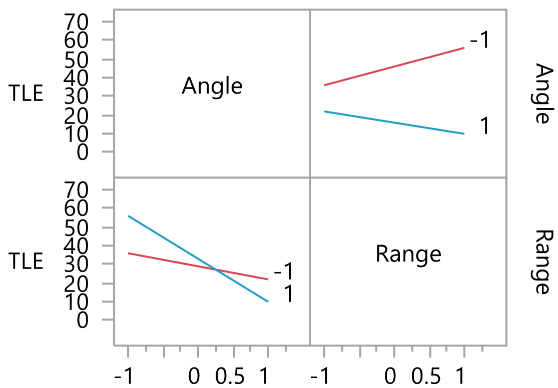
Remove Add Edit ☐ FDR ('^' denotes effects with containing effects above them)

*Remember, do not remove Range because it is also included in the higher order interaction of Angle*Range.

8. Observe the Interaction plot. Select "red drop down arrow -> Factor profiling -> Interaction plots"



Interaction Profiles



The conclusions we draw using JMP are the same as those from using R.

Conclusion

ANOVA is a powerful tool used to determine which factors are significant in affecting the response. The overall goal of ANOVA is to select a model that only contains significant terms. ANOVA can work for a single factor or be extended to multiple factors. A number of hypotheses have been presented that can be analyzed using ANOVA. Through the use of ANOVA, one can obtain a model that accurately represents the data without including terms that do not.

References

Burke, Sarah. "The Model Building Process Part 1: Checking model Assumptions." Scientific Test and Analysis Techniques Center of Excellence (STAT COE), 12 Oct. 2017.

Dean, Angela & Voss, Daniel *Design and Analysis of Experiments*. New York: Springer Science + Business Media Inc., 1999.

Montgomery, Douglas C. *Design and Analysis of Experiments.* 9th ed., John Wiley & Sons, Inc., 2017