3.2: Assumptions and Diagnostics

Before we draw any conclusions about the significance of the model, we need to make sure we have a "valid" model. Like any other statistical procedure, the ANOVA has assumptions that must be met. Failure to meet these assumptions means any conclusions drawn from the model are not to be trusted.

Assumptions

So what are these assumptions being made to employ the ANOVA model? The ***errors*** are assumed to be independent and identically distributed (*iid*) with a normal distribution having a mean of 0 and unknown equal variance.

As the model residuals serve as estimates of the unknown error, diagnostic tests to check for validity of model assumptions are based on **residual plots**, and thus, the implementation of diagnostic tests is also called **Residual Analysis**.

Diagnostic Tests

Most useful is the residual vs. predicted value plot, which identifies the violations of zero mean and equal variance. Residuals are also plotted against the treatment levels to examine if the residual behavior differs among treatments.

The normality assumption is checked by using a normal probability plot.

Residual plots can help identify potential outliers, and the pattern of residuals vs. fitted values or treatments may suggest a transformation of the response variable.

Lesson 4: SLR Model Assumptions of STAT 501 online notes discuss various diagnostic procedures in more detail.

There are various statistical tests to check the validity of these assumptions, but some may not be that useful. For example, Bartlett's test for homogeneity is too sensitive and indicates that problems exist when they really don't. It turns out that the ANOVA is very robust and is not badly affected by minor violations of these assumptions. In practice, a good deal of common sense and the visual inspection of the residual plots are sufficient to determine if serious problems exist.

We will employ statistical software such as SAS to conduct the residual analysis. Here are common patterns that you may encounter in the residual analysis (i.e. plotting residuals, e, against the predicted values, $\hat{y}$).

Figure $3.2.1a$ shows the prototype plot when the ANOVA model is appropriate for data. The residuals are scattered randomly around mean zero and variability is constant (i.e. within the horizontal bands) for all groups.

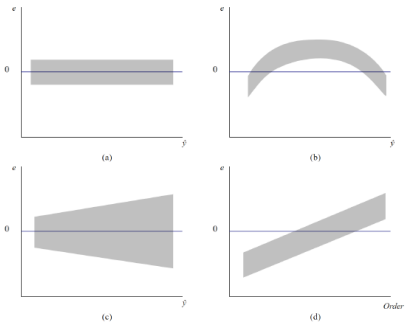

Figure 3.2.1: Common patterns in residual analysis.

Figure $3.2.1b$ suggests that although the variance is constant, there are some trends in the response that is not explained by a linear model. Using Figure $3.2.1c$, we can depict that the linear model is appropriate as the central trend in data is a line. However, the megaphone patterns in Figure $3.2.1c$ suggest that variance is not constant.

Alert!

A common problem encountered in ANOVA is when the variance of treatment levels is not equal (heterogeneity of variance). If the variance is increasing in proportion to the mean (panel (c) in Figure 3.2.1), a logarithmic transformation of Y can "stabilize" the variances. If the residuals vs. predicted values instead show a curvilinear trend (panel (b) in Figure 3.2.1), then a quadratic or other transformation may help. Since finding the correct transformation can be challenging, the Box-Cox method is often used to identify the appropriate transformation, given in terms of λ as shown below.

$$y_i^{(\lambda)} = \begin{cases} \frac{y_i^{\lambda}-1}{\lambda}, \text{ if } \lambda \neq 0, \\ \ln y_i, \text{ if } \lambda = 0 \end{cases} \tag{3.2.1}$$

Some λ values result some common transformations.

transformations.

λ	Y^{λ}	Transformation
2	Y^2	Square
1	Y^1	Original (No transform)
1/2	$\sqrt{Y}$	Square Root
0	$\log(Y)$	Logarithm
-1/2	$\frac{1}{\sqrt{Y}}$	Reciprocal Square Root
-1	$\frac{1}{Y}$	Reciprocal

Using Technology

Using Minitab

To run the Box-Cox procedure in Minitab, set up the data (Simulated Data), as a stacked format (a column with treatment (or trt combination) levels, and the second column with the response variable.

Treatment	Response Variable
A	12
A	23
A	34
B	45
B	56
B	67
C	14
C	25
C	36

Steps in Minitab

1. On the Minitab toolbar, choose **`Stat`** > **`Control Charts`** > **`Box-Cox Transformation`**

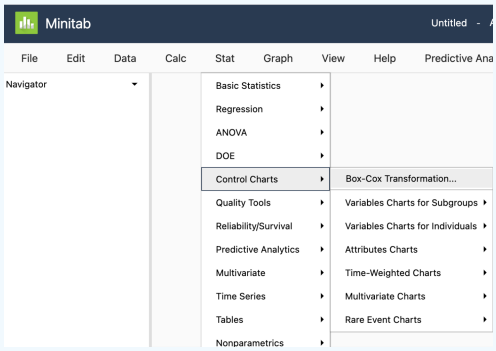

Figure 3.2.2: Selecting Box-Cox Transformation stat option.

2. Place "Response Variable" and "Treatment" in the boxes as shown below.

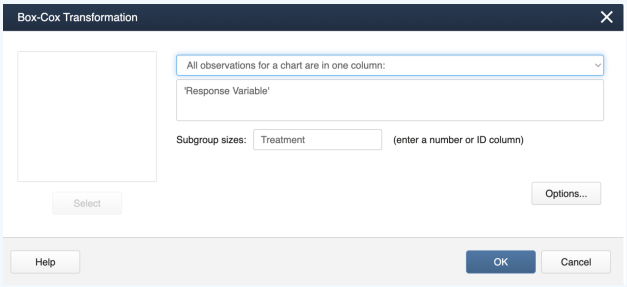

Figure 3.2.3: Inputting "Response Variable" and "Treatment" in pop-up window.

3. Click **`OK`** to finish. You will get an output like this:

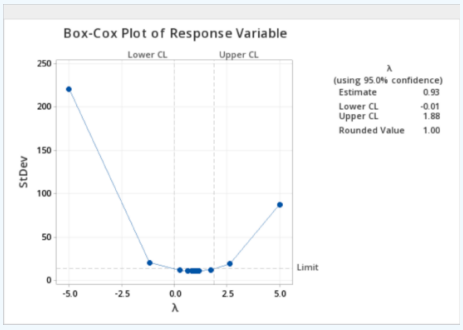

Figure 3.2.4: Minitab Box-Cox plot output.

In the upper right-hand box, the rounded value for λ is given from which the appropriate transformation of the response variable can be found using the chart above. Note, with a λ of 1, no transformation is recommended.

Using SAS

The Box-Cox procedure in SAS is more complicated in a general setting. It is done through the Transreg procedure, by obtaining the ANOVA solution with regression which first requires coding the treatment levels with effect coding discussed in Chapter 4.

However, for one-way ANOVA (ANOVA with only one factor) we can use the SAS *Transreg* procedure without much hassle.

Steps in SAS

Suppose we have SAS data as follows.

Obs	Treatment	ResponseVariable
1	A	12
2	A	23
3	A	34
4	B	45
5	B	56
6	B	67
7	C	14
8	C	25
9	C	36

We can use the following SAS commands to run the Box-Cox analysis.

```
proc transreg data=boxcoxSimData;
model boxcox(ResponseVariable)=class(Treatment);
run;
```

This would generate an output as follows, which suggests a transformation using $\lambda = 1$ (i.e. no transformation).

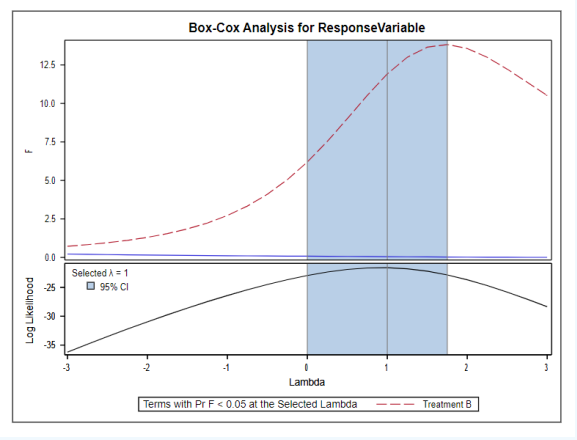

Figure 3.2.5: SAS Box-cox plot output.

Using R

Steps in R

Load the simulated data and perform the Box-Cox transformation. Note that simulated data are in the stacked format (a column with treatment levels and a column with the response variable)

```
setwd("~/path-to-folder/)
simulated_data<-read.table("simulated_data.txt",header=T)
attach(simulated_data)
library(AID)#Load package AID so that we can use the Box-Cox Procedure
boxcoxfr(Response_Variable,Treatment)#Box-Cox command for One-Way ANOVA
```

Output

Box-Cox power transformation	
data:	Response_Variable and Treatment
lambda.hat:	0.93

Shapiro-Wilk normality test for transformed data (alpha = 0.05)				
	Level	statistic	p.value	Normality
1	A	0.9998983	0.9807382	YES
2	B	0.9999840	0.9923681	YES
3	C	0.9999151	0.9824033	YES

Bartlett's homogeneity test for transformed data (alpha = 0.05)				
	Level	statistic	p.value	Homogeneity
1	All	0.008271728	0.9958727	YES

From the output, we can see that the lambda value for the transformation is 0.93 (the same value as Minitab suggested). Since this value is very close to 1 we can use $\lambda = 1$ (no transformation).

In addition, from the output, we can see that normality exists in all 3 levels (Shapiro-Wilk test) and we have the same variance (Bartlett's test).

Alternative:

We can use the command `boxcox` from package MASS

```
library(MASS)
Box_Cox_Plot<-boxcox(aov(Response_Variable~Treatment),lambda=seq(-3,3,0.01))
```

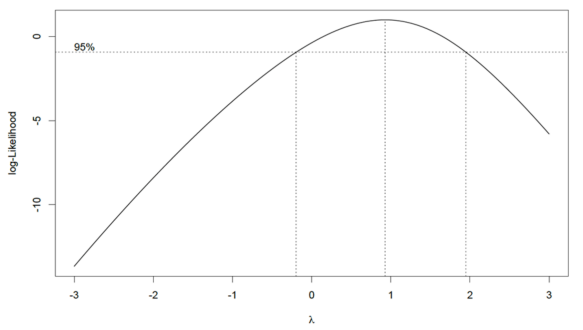

Figure 3.2.6: R-generated plot of log-likelihood vs λ.

From the plot, we can see the 95% CL. Since $\lambda = 1$ is within the interval there is no need for transformation.

```
#Exact lambda
lambda<-Box_Cox_Plot$x[which.max(Box_Cox_Plot$y)] #0.93
detach(simulated_data)
```