DATA603 Statistical Modelling

## Part II: Model Adequacy Checking

## Basic Assumptions of One-Factor Designs and ANOVA

For a CRD model, those well-versed in regression analysis will find the assumptions familiar; the same so-called standard assumptions are common to both techniques. We ascribe no special meaning to the order in which we list them. Recall the statistical model . We assume the following three statements:

1. The are independent random variables for all , and uncorrelated.

This means that each error term, , is independent of each other error term.

1. Each is normally distributed. We can assume that (that is, the true mean of is zero)

This is equivalent to saying that if we consider all the data values in a specific column, they would (theoretically, or if we had an infinite number of them) be distributed according to a normal distribution, with a mean equal to whatever is the true mean of that column.

1. Each has the same variance, (a constant variance)

This says that the normal distribution of each respective column, though perhaps differing in mean, has the same variance. This assumption is often referred to as the assumption of constant variance, and sometimes **the assumption of homoscedasticity.**

Note!! We define the residual for observation in treatment as

## How to check Model Adequacy

**1. Plot of Residuals Versus Fitted Values**

If the model is correct and if the assumptions are satisfied, the residuals should be structureless; in particular they should be unrelated to any other variable including the predicted response.

A simple check is to **plot the residuals versus the fitted values** . This plot should not reveal any obvious pattern.

The plot also can check for **nonconstant variance**. Sometimes the variance of the observations increases as the magnitude of the observation increases. The residuals would get larger as gets larger, and the plot of residuals versus would look like an outward-opening funnel or megaphone. Nonconstant variance also arises in case where the data follow a non normal, skewed distribution because in skewed distributions the variance tends to be a function of the mean.

If the assumption of homogeneity of variances is violated, the test is only slightly affected in the balanced model.

**2. Statistical Tests for Equality of Variance** Although residual plots are frequently used to diagnose inequality of variance such as Breusch-Pagan test, Levene test, several statistical tests have also been proposed. Here we provide **the Bartlett’s test**

The procedure involves computing a statistic whose sampling distribution is closely approximated by the chi-square distribution with degrees of freedom.

The test statistic is

where and are some quantities. We reject when > . The p-value approach to decision making could be used.

**3. Plotting a normal probability plot of the residuals**

A check of the normallity assumption could be made by plotting a histrogram of the residuals. If the NID(0,) assumption (normality and independence) on errors is satisfied, then this plot should look like a sample from a normal distribution centered at zero. Unfortunally, with small samples, considerable fluctuation often occurs, so an extremely useful procedure is to construct a normal probability plot of the residuals. If the underlying error distribution is normal, this plot will resemble a straight line.

Note! There are also a variety of statistical tests for normality, including Shapiro-Wilk test (from Linear Regression). We could use this test as well.

From the broker study,check Model Adequacy (Normality Assumption, Constant Variance Assumption, Independent error term Assumption )

library(lmtest)

## Loading required package: zoo

##   
## Attaching package: 'zoo'

## The following objects are masked from 'package:base':  
##   
## as.Date, as.Date.numeric

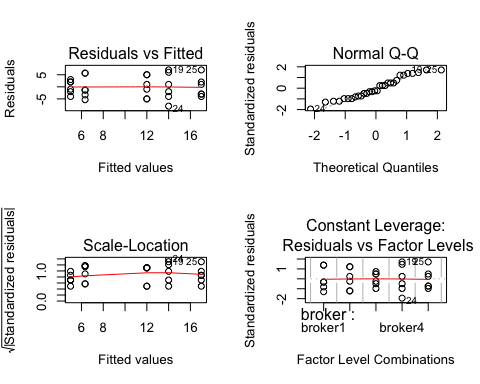
brokerstudy=read.csv("/Users/Ellsworth/Documents/School/DATA603/Lectures/Unit\ 3/brokerstudy.csv", header=TRUE)  
str(brokerstudy)#Read your data set and double check that dependent and indepent variables are correctly read by R

## 'data.frame': 30 obs. of 2 variables:  
## $ broker: Factor w/ 5 levels "broker1","broker2",..: 1 1 1 1 1 1 2 2 2 2 ...  
## $ price : int 12 3 5 1 12 5 7 17 13 11 ...

CRD<-aov(price~broker, data=brokerstudy) #Perform ANOVA for CRD  
summary(CRD)

## Df Sum Sq Mean Sq F value Pr(>F)   
## broker 4 622.1 155.53 7.695 0.000347 \*\*\*  
## Residuals 25 505.3 20.21   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

par(mfrow=c(2,2))  
plot(CRD)



bartlett.test(price~broker, data=brokerstudy)

##   
## Bartlett test of homogeneity of variances  
##   
## data: price by broker  
## Bartlett's K-squared = 2.4411, df = 4, p-value = 0.6552

shapiro.test(residuals(CRD))

##   
## Shapiro-Wilk normality test  
##   
## data: residuals(CRD)  
## W = 0.95177, p-value = 0.1885

bptest(CRD)

##   
## studentized Breusch-Pagan test  
##   
## data: CRD  
## BP = 5.3992, df = 4, p-value = 0.2487

## Inclass Practice Problem

From the MVPC experiment, check Model Adequacy (Normaillty Assumption, Constant Variance Assumption, Independent error term Assumption)

## Inclass Practice Problem

From the lifetime of AA battery experiment,check Model Adequacy (Normaillty Assumption, Constant Variance Assumption, Independent error term Assumption )

In the situations where assumptions are unjustified, the experimenter may wish to use an alternative procedure to the F -test analysis of the variance that does not depend on the assumptions. Such a procedure has been developed by Kruskal and Wallis (1952).

## Kruskal-Wallis Test

One way to avoid the distributional aspect of the standard assumptions (that is, the assumption of normality) is to perform what is called a nonparametric test. The Kruskal-Wallis test is used to test the null hypothesis that the treatments are identical against the alternative hypothesis that some of the treatments generate observations that are larger than others. Because the procedure is designed to be sensitive for testing differences in means, it is sometimes convenient to think of the Kruskal- Wallis test as a test for equality of treatment means. The Kruskal-Wallis test is a **nonparametric alternative** to the usual analysis of variance.

To perform a Kruskal-Wallis test, first rank the observations in ascending order and replace each observaton by its rank, say , with smallest observation having rank 1. In the case of ties, assign the average rank to each of the tied observations.

The test statistic is

We reject the null hypothesis when . The p-value approach could be used.

brokerstudy=read.csv("/Users/Ellsworth/Documents/School/DATA603/Lectures/Unit\ 3/brokerstudy.csv", header=TRUE)  
str(brokerstudy)#Read your data set and double check that dependent and indepent variables are correctly read by R

## 'data.frame': 30 obs. of 2 variables:  
## $ broker: Factor w/ 5 levels "broker1","broker2",..: 1 1 1 1 1 1 2 2 2 2 ...  
## $ price : int 12 3 5 1 12 5 7 17 13 11 ...

kruskal.test(price~broker, data=brokerstudy)

##   
## Kruskal-Wallis rank sum test  
##   
## data: price by broker  
## Kruskal-Wallis chi-squared = 16.518, df = 4, p-value = 0.002398

The output shows that H= 16.498 with the p-value= 0.002419 <. We can conclude that at least one average price is different.

If the Kruskal-Wallis test is significant, a post-hoc analysis can be performed to determine which levels of the independent variable differ from each other level. The most popular test for this is the **Dunn test**.

library(FSA)

## ## FSA v0.8.26. See citation('FSA') if used in publication.  
## ## Run fishR() for related website and fishR('IFAR') for related book.

brokerstudy=read.csv("/Users/Ellsworth/Documents/School/DATA603/Lectures/Unit\ 3/brokerstudy.csv", header=TRUE)  
DT = dunnTest(price~broker,data=brokerstudy,method="none")  
DT

## Dunn (1964) Kruskal-Wallis multiple comparison

## with no adjustment for p-values.

## Comparison Z P.unadj P.adj  
## 1 broker1 - broker2 -1.7253856 0.0844580131 0.0844580131  
## 2 broker1 - broker3 0.2629159 0.7926154057 0.7926154057  
## 3 broker2 - broker3 1.9883014 0.0467783609 0.0467783609  
## 4 broker1 - broker4 -2.2019206 0.0276709168 0.0276709168  
## 5 broker2 - broker4 -0.4765351 0.6336932376 0.6336932376  
## 6 broker3 - broker4 -2.4648365 0.0137075783 0.0137075783  
## 7 broker1 - broker5 -3.0728295 0.0021203959 0.0021203959  
## 8 broker2 - broker5 -1.3474440 0.1778372919 0.1778372919  
## 9 broker3 - broker5 -3.3357454 0.0008507103 0.0008507103  
## 10 broker4 - broker5 -0.8709089 0.3838038978 0.3838038978