Chemometrics

MA4605

Week 5. Lecture 10. ANOVA-assumptions

October 4, 2011

One way ANOVA tests the claim that the means are equal for k independent groups

$H_0 : \mu_1 = \mu_2 = ...\mu_k$

H_a: not all the means are equal (at least one mean is different).

ANOVA model

- $Y_{ij} = \mu + \tau_j + \epsilon_{ij}$
- There are k levels of the factor (i.e. k groups) and j=1,2... k
- The number of observations at the j^{th} factor/group is n_j
- The total number of observations is n = $\sum(n_j)$
- Y_{ij} is the i^{th} observation of group j (response variable)
- μ is the overall population mean.
- τ_j is the effect of treatment/group j
- $\epsilon_{i,j}$ are the random independent errors

Assumptions

Assumptions

- Each population from which a sample is taken is assumed to be normal.

Assumptions

- Each population from which a sample is taken is assumed to be normal.
- Each sample is randomly selected and independent.

Assumptions

- Each population from which a sample is taken is assumed to be normal.
- Each sample is randomly selected and independent.
- The populations have approximately equal variances(standard deviations).

Check the ANOVA assumptions

The residuals are the basic building block for checking the suitability of the working ANOVA model and validating the assumptions on which it is based.
In general the residuals can be defined as

Residuals = DATA - FITTED VALUES

In the case of the One-Way ANOVA the fitted values are simply the group means:

- $\widehat{\mu}_1 = 55$
- $\widehat{\mu}_2 = 68$
- $\widehat{\mu}_3 = 57$
- $\widehat{\mu}_4 = 69$

The residuals can be constructed as

$\widehat{\epsilon}_{i,j} = Y_{i,j} - \widehat{\mu}_j$

Oxygen demand	Fitted Values	Residuals
48	55	-7
54	55	-1
57	55	2
54	55	-1
62	55	7
73	68	5
63	68	-5
66	68	-2
64	68	-4
74	68	6
51	57	-6
63	57	6
61	57	4
54	57	-3
56	57	-1
70	69	3
68	69	-1
71	69	2

The residuals can also be obtained from *R* with

$> \ model <- aov(y \sim group, det)$

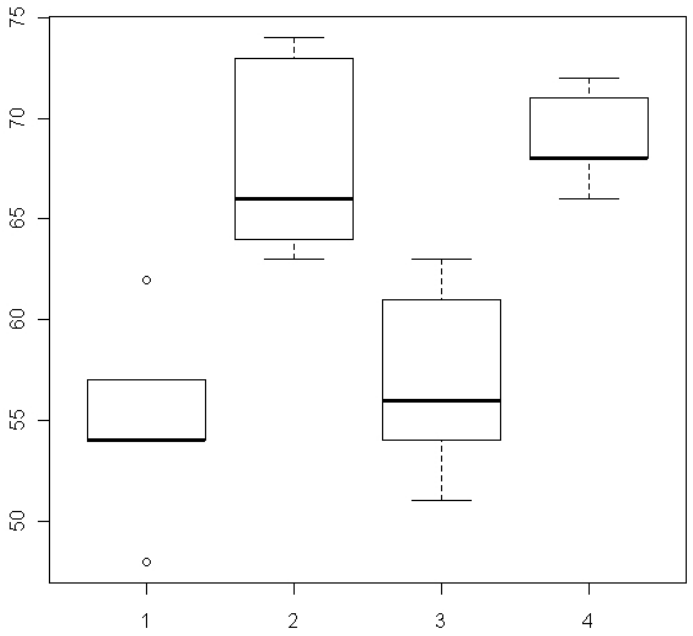
$> \ resid <- residuals(model)$

$> \ resid$

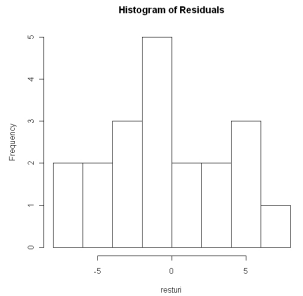
1	2	3	4	5	6	7	8	9	10
-7	-1	2	-1	7	5	-5	-2	-4	6
11	12	13	14	15	16	17	18	19	20
-6	6	4	-3	-1	3	-1	2	-1	-3

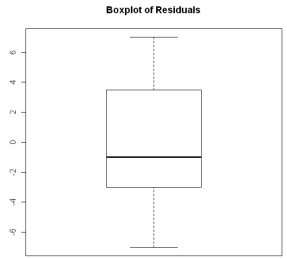
Before running the analysis of variance, you should graph the means and standard errors. Illustrate the data with parallel boxplots (one for each treatment).

$> \; plot(group, y)$



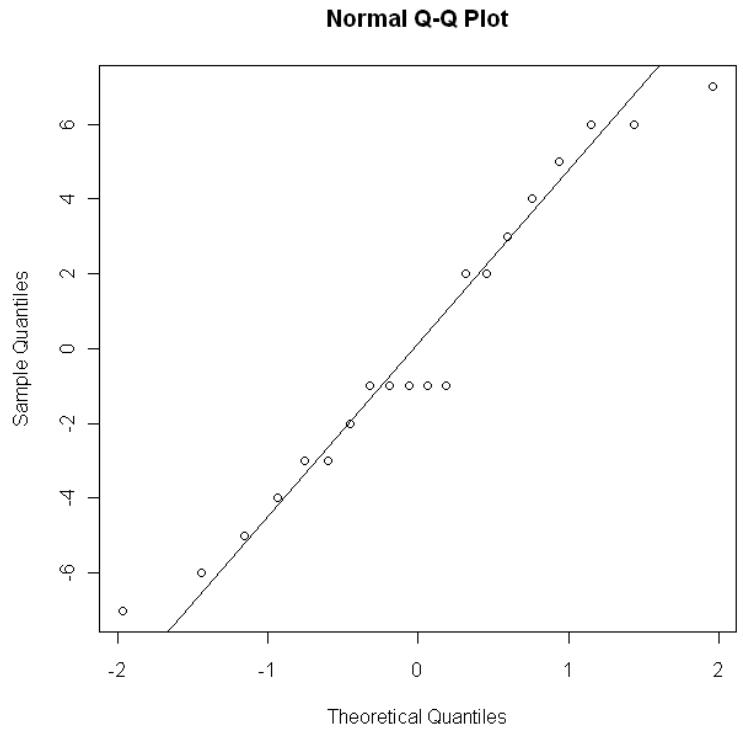
We check the assumption of normality by looking at the following plots:





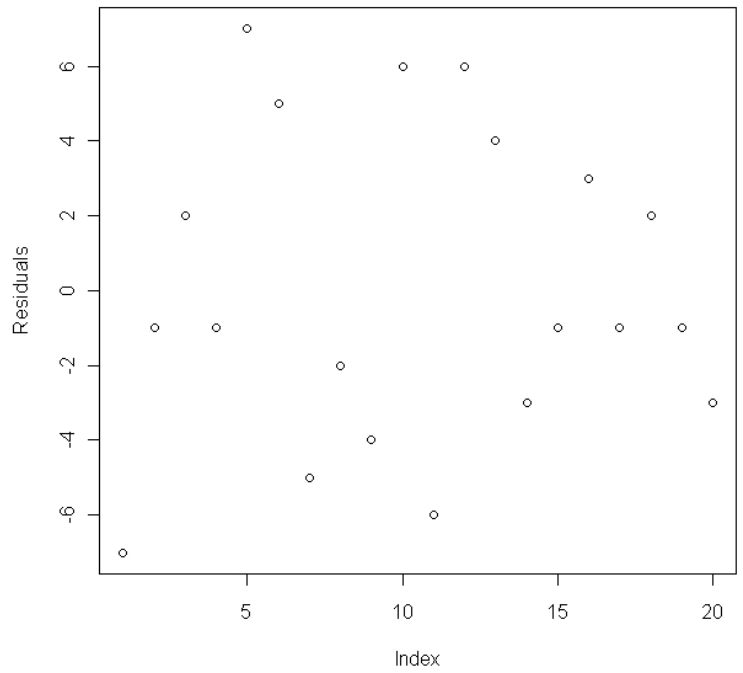
> *hist*(*resid*)
> *boxplot*(*resid*)

Also the scatterplot of residuals against the Nscores (ideal values). If the points fall on a line the assumption of normality holds.



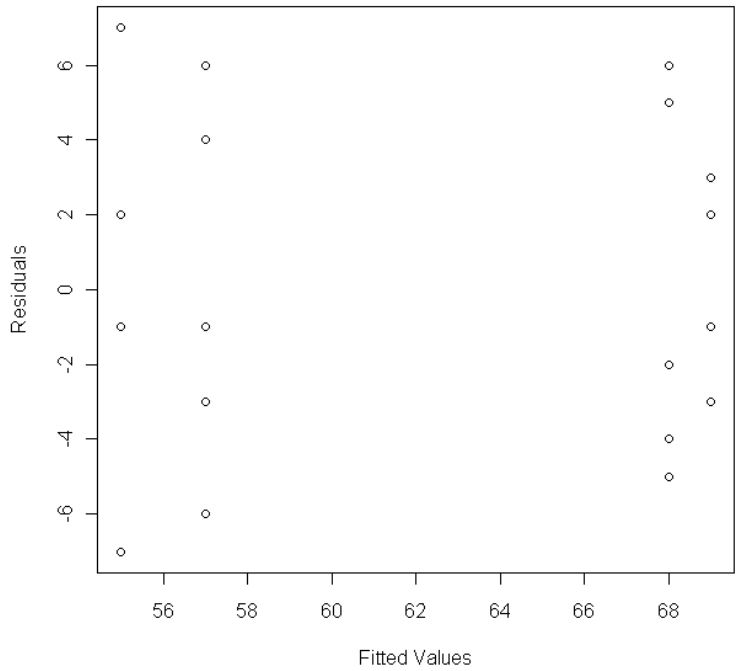
> *qqnorm*(*resid*)
> *qqline*(*resid*)

The plot of the residuals against the order of the data displays no particular pattern, hence we can assume the data are random.



$>$ *plot*(*resid*)

Use the plot of the residuals against the fitted values (predicted values) to check for non-constant variance.



$> \ pred < -predict(model)$

$> \ pred < -plot(pred, resid)$

Statistical test for constant variance: Bartlett test of homogeneity.

$> \ bartlett.test(y \sim group, det)$

Bartlett test of homogeneity of variances

data: y by group

Bartlett's K-squared = 2.1899, df = 3, p-value = 0.5339

The Bartlett test gives a p-value of 0.5339 which is greater than 0.05, hence we accept the null hypothesis that the variances are equal.