Design of Engineering Experiments
Part 3 – Analysis of Variance (2)

- Comparing more that two factor levels…**the analysis of variance**
  - ANOVA decomposition of total variability
  - Statistical testing & analysis
  - Checking assumptions, model validity
  - Post-ANOVA testing of means
  - Box-Cox method
- **Sample size** determination

Model Adequacy Checking in the ANOVA
Text reference, Section 3-4, pg. 75

- **Checking assumptions** is important
- Normality
- Constant variance
- Independence
- Have we fit the right model?
- Later we will talk about what to do if some of these assumptions are **violated**

Least Squares Estimation of the Model Parameters

(text reference, section 3-9.1, page 107-108)

$$\text{Overall mean}: \hat{\mu} = \bar{y}_{..} \quad \text{(grand average of the obs.)}$$

$$\text{Treatment effect}: \hat{\tau}_i = \bar{y}_{i.} - \bar{y}_{..} \text{ (diff. bw. trt. avg. and the grand avg.)}$$

$$\text{Constraint}: \sum_{i=1}^{a} \hat{\tau}_i = 0$$

$$\text{Fitted Value}: \hat{y}_{ij} = \hat{\mu} + \hat{\tau}_i = \bar{y}_{i.}$$

$$\text{Residual Value}: \ e_{ij} = y_{ij} - \hat{y}_{ij} \ \text{ (estimator of } \varepsilon_{\text{ij}})$$

(text reference, section 3-9.1, page 107-108)

$$L = \sum_{i=1}^{a} \sum_{j=1}^{n} \varepsilon_{ij}^{2} = \sum_{i=1}^{a} \sum_{j=1}^{n} (y_{ij} - \mu - \tau_i)^2$$

$$\left.\frac{\partial L}{\partial \mu}\right|_{\hat{\mu}, \hat{\tau}_i} = 0; \qquad \left.\frac{\partial L}{\partial \hat{\tau}_i}\right|_{\hat{\mu}, \hat{\tau}_i} = 0 \qquad i = 1, 2, \ldots, a$$

$$-2 \sum_{i=1}^{a} \sum_{j=1}^{n} (y_{ij} - \hat{\mu} - \hat{\tau}_i) = 0$$

$$-2 \sum_{j=1}^{n} (y_{ij} - \hat{\mu} - \hat{\tau}_i) = 0 \qquad i = 1, 2, \ldots, a$$

$$\begin{cases} an\,\hat{\mu} + n\hat{\tau}_1 + n\hat{\tau}_2 + \cdots + n\hat{\tau}_a = an\,\bar{y}_{..} \\ n\hat{\mu} + n\hat{\tau}_1 = n\bar{y}_{1.} \\ n\hat{\mu} + n\hat{\tau}_2 = n\bar{y}_{2.} \\ \vdots \\ n\hat{\mu} + n\hat{\tau}_a = n\bar{y}_{a.} \end{cases}$$

$a + 1$ unknowns, but only a equations. Therefore, no unique solutions.

Add constraint : $\sum_{i=1}^{a} \hat{\tau}_i = 0$

Least Squares Estimation of the Model Parameters

(text reference, section 3-9.1, page 108)

$$\text{Constraint}: \sum_{i=1}^{a} \hat{\tau}_i = 0$$

- There are an infinite number of possible constraints that could be used to solve the equation.

However, it really doesn't matter

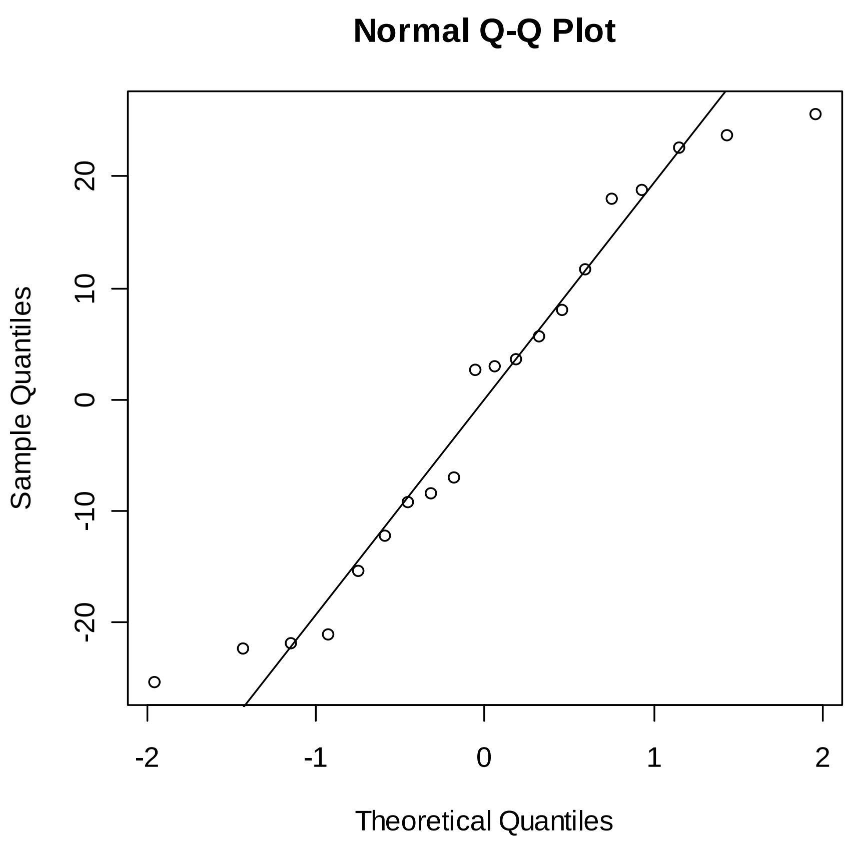
- We are usually interested in differences among the treatment effects $\tau_i - \tau_j$ which can be uniquely estimated.

Model Adequacy Checking in the ANOVA

- Examination of **residuals** (see text, Sec. 3-4, pg. 75)

$$e_{ij} = y_{ij} - \hat{y}_{ij}$$
$$= y_{ij} - \bar{y}_{i.}$$

- R package generates the residuals
- **Residual plots** are very useful
- **Normal probability plot** of residuals



- Focus on the central values of the plot than on the extremes
- Sometime the defect is caused by outlier
- F-test is robust to normality assumption

Other Important Residual Plots

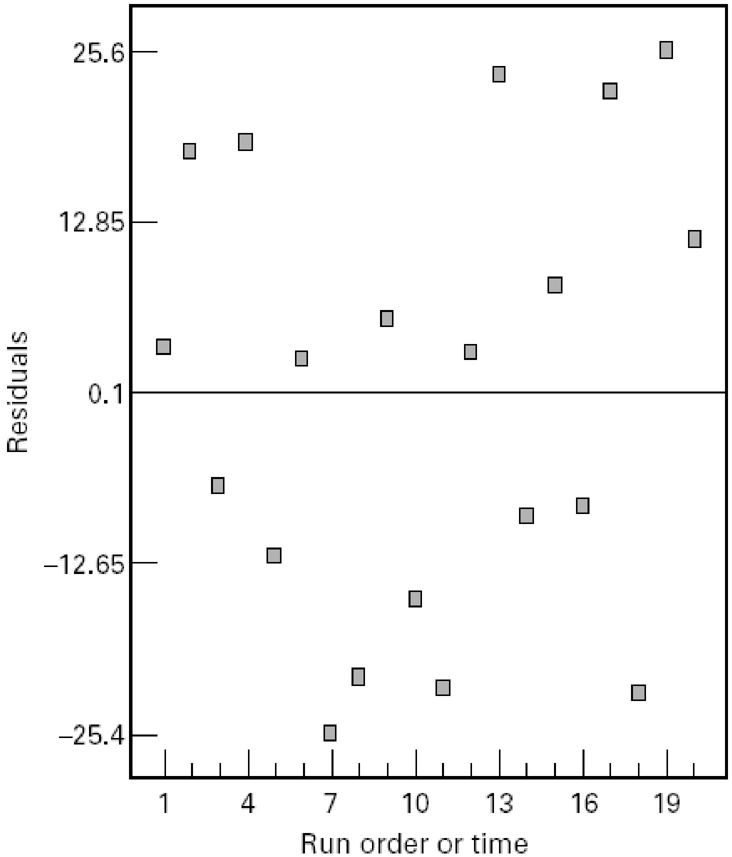


Figure 3-5 Plot of residuals versus run order or time.

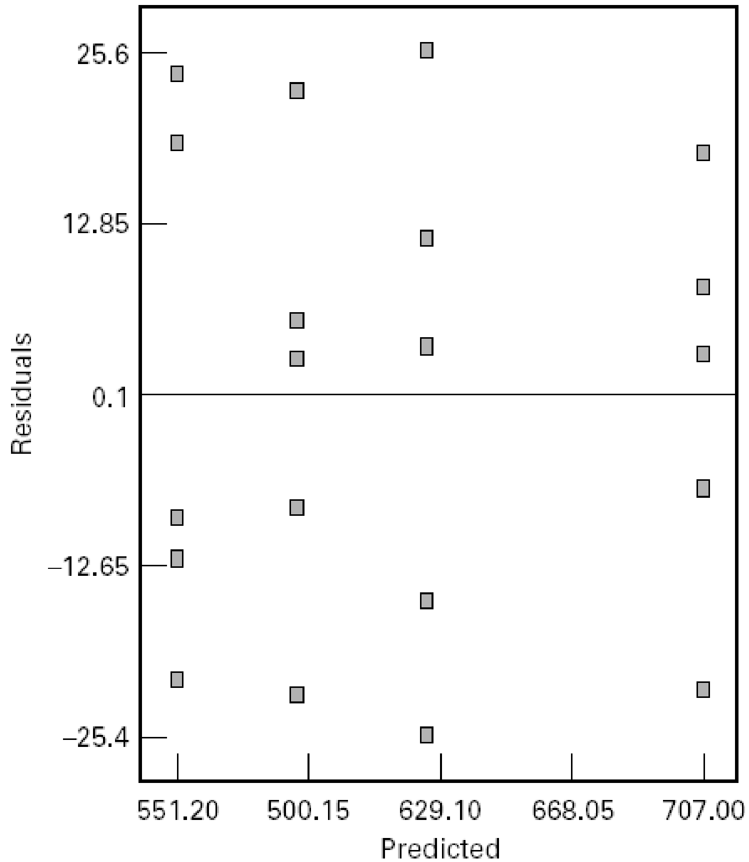


Figure 3-6 Plot of residuals versus fitted values.

A tendency to have runs of positive and negative residuals indicates positive correlation.

- Unusual structure indicates nonconstant variance
- Equal sample sizes whenever possible

R Model Adequacy Checking

```
attach(plasma)
# 1-way ANOVA
plasma.aov = aov(EtchRate~Power)

# Model Adequacy Checking: 1-way ANOVA
>resid=unlist(plasma.aov$resid)
>fitted=unlist(plasma.aov$fitted)

#QQ plot of residuals for normality assumption
>qqnorm(resid)
>qqline(resid)

#plot of residuals versus fitted values for equal variance assumption
>plot(fitted, resid)
>abline(h=0)

#plot of residuals versus observation order for independence assumption
>plot(order, resid)
>abline(h=0)
```

Bartlett's Tests for Equivality of variances of *a* normal distributions

warning: sensitive to normality assumption

- Assumption: all *a* samples from **normal** distributions
- Statistical hypotheses: $H_0: \sigma_1^2 = \sigma_2^2 \cdots = \sigma_a^2$

  H_1: above not true for at least one σ_i^2
- Test Statistic:

$$\chi_0^2 = 2.3026 \frac{q}{c} \overset{\text{Under } H_0}{\sim} \chi^2(a-1)$$

$$\text{where } q = (N-a)\log_{10} S_p^2 - \sum_{i=1}^{a} (n_i - 1)\log_{10} S_i^2$$

$$c = 1 + \frac{1}{3(a-1)}\left(\sum_{i=1}^{a} (n_i - 1)^{-1} - (N-a)^{-1}\right)$$

$$S_p^2 = \frac{\sum_{i=1}^{a} (n_i - 1) S_i^2}{N-a} \text{ and } S_i^2 \text{ is the sample variance of the } i\text{th population}$$

- Reject region or P-value: Reject if $\chi_0^2 > \chi_{\alpha,a-1}^2$ or if p-value<α.
- Conclusion

Example 3-4: Etch Rate Data

$$\chi_0^2 = 2.3026\frac{q}{c} = 2.3026\frac{(0.17)}{(1.10)} < \chi_{0.05,3}^2 = 7.81$$

$$\text{where } \mathrm{q} = (\mathrm{N-a})\log_{10} S_p^2 - \sum_{i=1}^{a}(n_i - 1)\log_{10} S_i^2$$

$$= (20-4)\log_{10}(333.7) - 4(\log_{10} 400.7 + \log_{10} 280.3 + \log_{10} 431.3 + \log_{10} 232.5)$$

$$= 0.17$$

$$c = 1 + \frac{1}{3(a-1)}\left(\sum_{i=1}^{a}(n_i - 1)^{-1} - (N-a)^{-1}\right)$$

$$= 1 + \frac{1}{3(3)}\left(\frac{4}{4} - \frac{1}{16}\right) = 1.10$$

$$S_p^2 = \frac{\sum_{i=1}^{a}(n_i - 1)S_i^2}{N-a}$$

$$= \frac{4(400.7) + 4(280.3) + 4(421.3) + 4(232.5)}{16}$$

$$= 333.7$$

R code and output

- #Bartlett's test for equal variances

```
>bartlett.test(EtchRate~Power)
```

Bartlett test of homogeneity of variances

data: EtchRate by Power

Bartlett's K-squared = 0.4335, df = 3, p-value = 0.9332

Modified Levene Tests for Equivality of variances of *a* normal distributions

note: robust to violation of normality assumption

1. Calculate the absolute deviation of the observations y_{ij} in each treatment from the treatment median, say $\tilde{y}_i$. Denote these deviations by

$$d_{ij} = | y_{ij} - \tilde{y}_i | \begin{cases} i = 1,2,\ldots,a \\ j = 1,2,\ldots,n_i \end{cases}$$

2. Use the usual ANOVA F statistic for testing the equality of means applied to the absolute deviation

R code and output

```
#Modified Levene test for equal variances
#obtain absolute deviation
>EtchRate.median=rep(tapply(EtchRate, Power, median),each=5)
>d<-abs(EtchRate-EtchRate.median)
#F-test on d
>Levene.aov<-aov(d~Power)
>summary(Levene.aov)
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Power	3	82.00	27.33	0.1959	0.8977
Residuals	16	2232.80	139.55		