Statistics & Experimental Design with R

Barbara Kitchenham
Keele University

Keele University

Analysis of Variance

Multiple groups with Normally distributed data

Experimental Design

- LIST
  - Factors you may be able to control
- BLOCK
  - Factors under your control
    - Some factors could be used to restrict scope of experiment
    - E.G. Restrict to Post graduate students
- MEASURE
  - Factors that cant be controlled
  - Possible co-variates
- RANDOMLY
  - Assign units to treatments within blocks

ANOVA

- Basic Terminology
  - ANOVA stands for Analysis of Variance
  - Consider the problem of deciding whether testing method A is better method B
    - You recruit 20 testers (subjects/participants)
    - Randomly assign 10 to standard method (called a control)
    - Randomly assign 10 to the new method
    - Give them a testing problem & measure outcome (e.g. number of defects detected)
    - The two treatments together are referred to as a **factor** with two **levels**
  - Number of defects is called "**dependent variable**"
  - Method is called the "**independent variable**"
    - Takes on two values A or B
  - When you have equal number of participants in each treatment condition
    - **Balanced** design
    - Otherwise **unbalanced**
  - This is called a **one-way between -groups ANOVA**

Basic Experimental Designs

- One-way ANOVA means participants classified in one dimension i.e. treatment
  - There can be many treatments
  - Treatments can be independent
    - E.g. Testing methods A, B, C, etc.
  - Treatment may be related
    - Based on the extent of a treatment
    - E.g. Extent of training one day, two days, or 5 days

More Complex Designs

- Consider a testing experiment comparing three methods
  - Want to assess how well the methods work with programs of different complexity
  - Assume three methods and three levels of complexity: easy, average, hard
- This experiment has two factors
  - Testing method and complexity
  - For each testing method we want to investigate each complexity condition
- Also interested in the effect of complexity level on the outcome of each method
  - Which is called the **interaction** between the factors
- For a balanced design we would need the number of participants to be a multiple 9
  - product of number of conditions in each factor
- This design is called a **3 by 3 Factorial experiment**

Within-subject Designs

Keele University

- Alternatively suppose we have three testing methods and testing problems all of average complexity
- If each participant tried out each method
  - 20 participants result in 60 observations
  - 20 for each testing method
  - In this case we can treat the individual participants as a blocking factor
    - Analysing the data to remove the effect of difference among participants
    - Hopefully reducing the variance used for our tests
- This give us a *within-subjects* design

Basic On-way ANOVA Model

- Fixed effects model

$$x_{ij} = A + E_j + e_{ij}$$

- x_{ij} is i-th member of group j
- *A* is an overall average effect common to all observations
- E_j is a "fixed" or constant difference from *A* due to the jth population common to all members of j
- e_{ij} is a random error $\sim N(0,\sigma^2)$
- H0 is all E_j are zero and population mean = *A*

Model parameters

$$\bar{x}_{.j} = \frac{1}{n_j}\sum_{i=1}^{n_j} x_{ij} = \frac{1}{n_j}\left(\sum_{i=1}^{n_j}(A + E_j + e_{ij})\right) \quad \bar{x}_{.j} = A + E_j + \bar{e}_{.j}$$

$$\bar{x}_{..} = \frac{1}{N}\sum_{j=1}^{k} n_j \bar{x}_{.j} = \frac{1}{N}\left(\sum_{j=1}^{k} n_j \left(A + E_j + \bar{e}_{.j}\right)\right)$$

$$\bar{x}_{..} = A + \frac{\sum_{j=1}^{k} n_j E_j}{N} + \bar{e}_{..} = A + \bar{e}_{..}$$

Assuming $\frac{\sum_{j=1}^{k} n_j E_j}{N} = 0$

$$x_{ij} - \bar{x}_{.j} = e_{ij} - \bar{e}_{.j}$$

Independent of E_j

$$x_{.j} - \bar{x}_{..} = E_j + \bar{e}_{.j} - \bar{e}_{..}$$

Partitioning Sums of Squares

$$SS = \sum_{j=1}^{k}\sum_{i=1}^{n_j}(x_{ij} - \bar{x}_{..})^2 = \sum_{j=1}^{k}\sum_{i=1}^{n_j}((x_{ij} - \bar{x}_{.j}) + (\bar{x}_{.j} - \bar{x}_{..}))^2$$

$$= \sum_{j=1}^{k}\sum_{i=1}^{n_j}(x_{ij} - \bar{x}_{.j})^2 + \sum_{j=1}^{n_j} n_j(\bar{x}_{.j} - \bar{x}_{..})^2$$

SSW: $$\sum_{j=1}^{k}\sum_{i=1}^{n_j}(x_{ij} - \bar{x}_{.j})^2 = \sum_{j=1}^{k}\sum_{i=1}^{n_j}(e_{ij} - \bar{e}_{.j})^2 = \sigma^2\sum_{j=1}^{k}(n_j - 1) = \sigma^2(N-k)$$

SSB: $$\sum_{j=1}^{k}(\bar{x}_{.j} - \bar{x}_{..})^2 = \sum_{j=1}^{k}(E_j + \bar{e}_{.j} - \bar{e}_{..})^2 = \sigma^2(k-1) + \sum_{j=1}^{k} n_j E_j^2$$

Rational for F test

- Distribution of ratio of two chi-squared variables is known and called F distribution
- So distribution of ratio of two sample variances (i.e. s_1^2/s_2^2) follows the F distribution
- If distribution of measured values is Normal in each group and H0 true
  - Ratio of [SBG/(k-1)]/[SWG/(N-k)]
  - F with degrees of freedom k-1 and N-k respectively

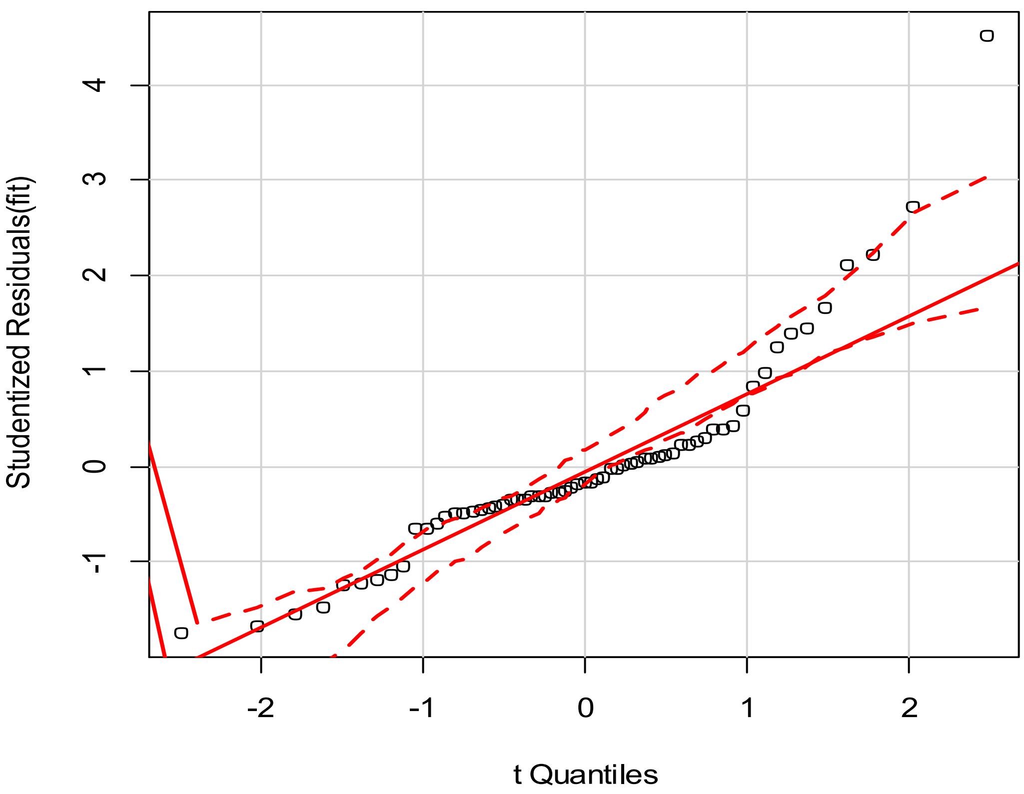
One-Way ANOVA Table

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F-ratio
Between Groups	SSB	$\nu=k-1$	$MSB=SSB/\nu$	MSB/MSW
Within Groups	SSW	$\nu=N-k$	$MSW=SSW/\nu$	
Total	SS			

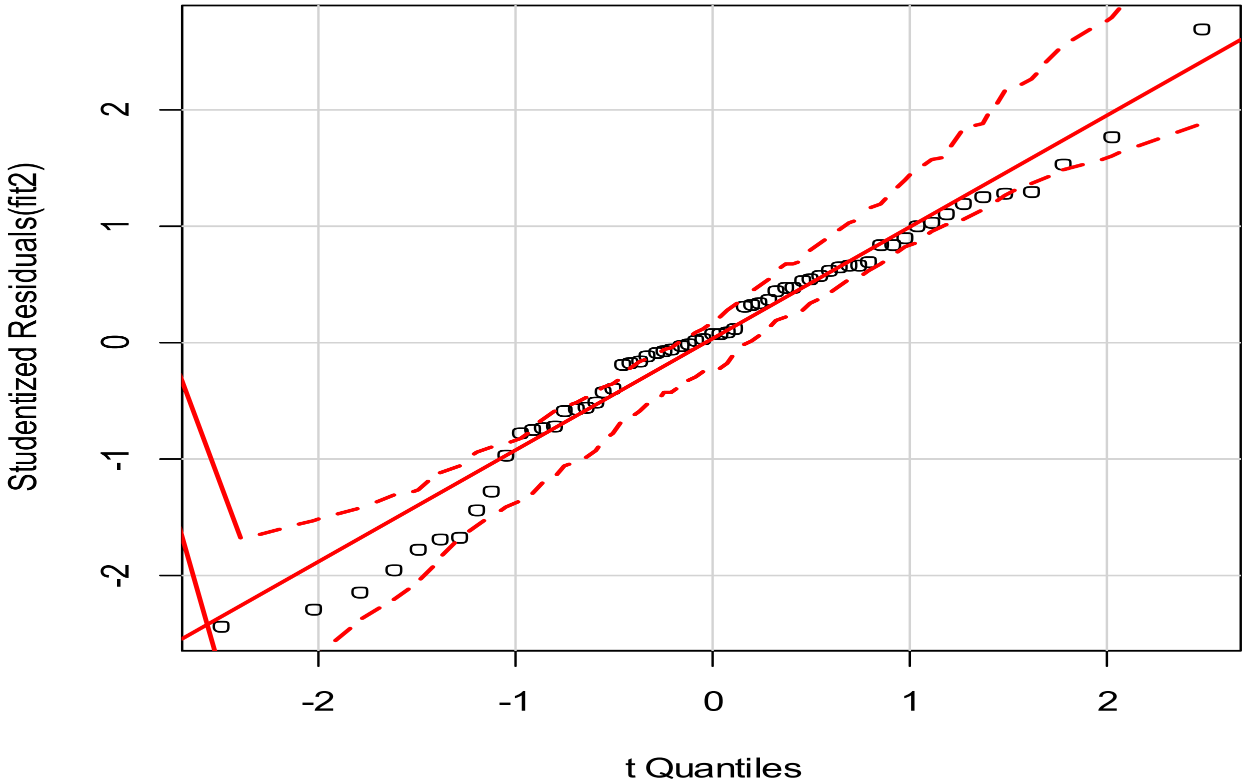
ANOVA for COCOMO Productivity with Mode as main factor

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F-ratio
Between Groups	1.197	2	0.598	13.33 *** (p=1.62e-05)
Within Groups	2.693	60	0.0499	
Total	3.89	62	0.0627	

QQPlot of Productivity data analysis



QQPlot of ANOVA based on Log(Productivity)



Keele University

Standard ANOVA designs

- Blocked designs
  - Blocking is used for controllable nuisance parameters
  - Simplest design is randomised blocks design
    - Has treatment factor (T) with k-levels
    - Blocking Factor B
    - Each Block has an observation for each treatment
  - E.g. Block are student grades
    - Match k-tuples of students based on grade
    - Randomly assign one subject per block to each of k treatments
  - Interaction between blocks & treatments ignored

ANOVA Design for Randomised Blocks

Blocks	Treatments		
	T1	T2	T3
B1	S1	S2	S3
B2	S4	S5	S6
B3	S7	S8	S9

Source	SS	df	MS	F
Treatments	SS Between Treatments	k-1	MST= SST/ df(T)	MMST/ ME
Blocks	SS Between Blocks	j-1	MSB= SSB/ df(B)	
Error	SS Within Treatments and Blocks	(k-1) × (j-1)	ME= SSE/ df(E)	

Latin-Square

- Two-way Blocking
  - Example would be
    - Participants each try a set of different treatments
      - Individual participants are one block
      - Order that participants are assigned to each treatment is other block

Subjects	Order		
	First	Second	Third
S1	T1	T2	T3
S2	T2	T3	T1
S3	T3	T1	T2

Factorial Design

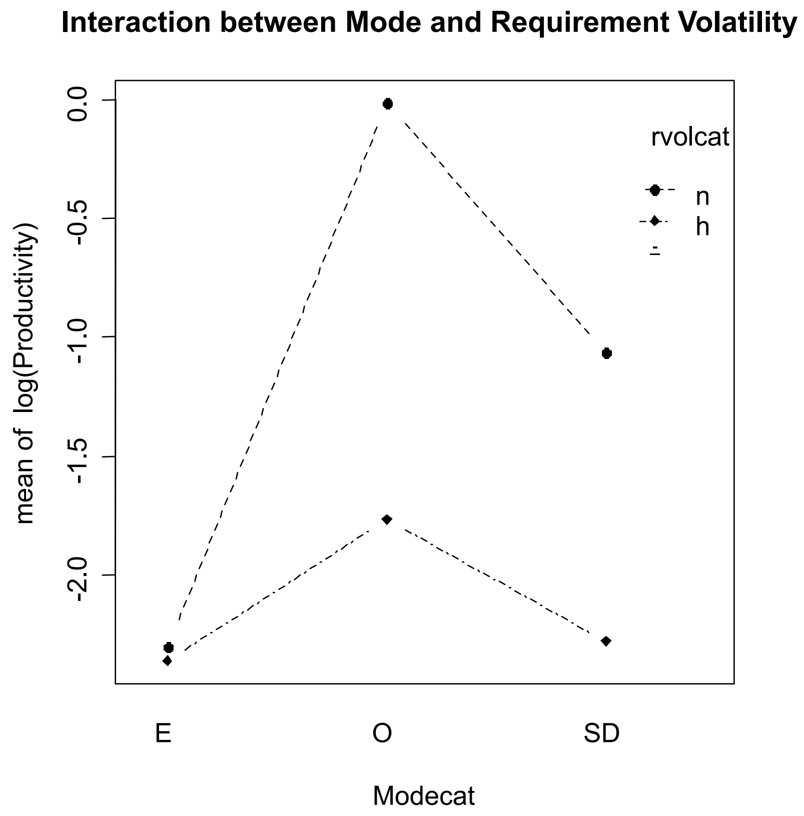
Factor B	Factor A		
	Level 1	Level 2	Level 3
Level 1	P1,P2,P3	P4,P5,P6	P7,P8,P9
Level 2	P10,P11,P12	P13,P14,P15	P16,P17,P19

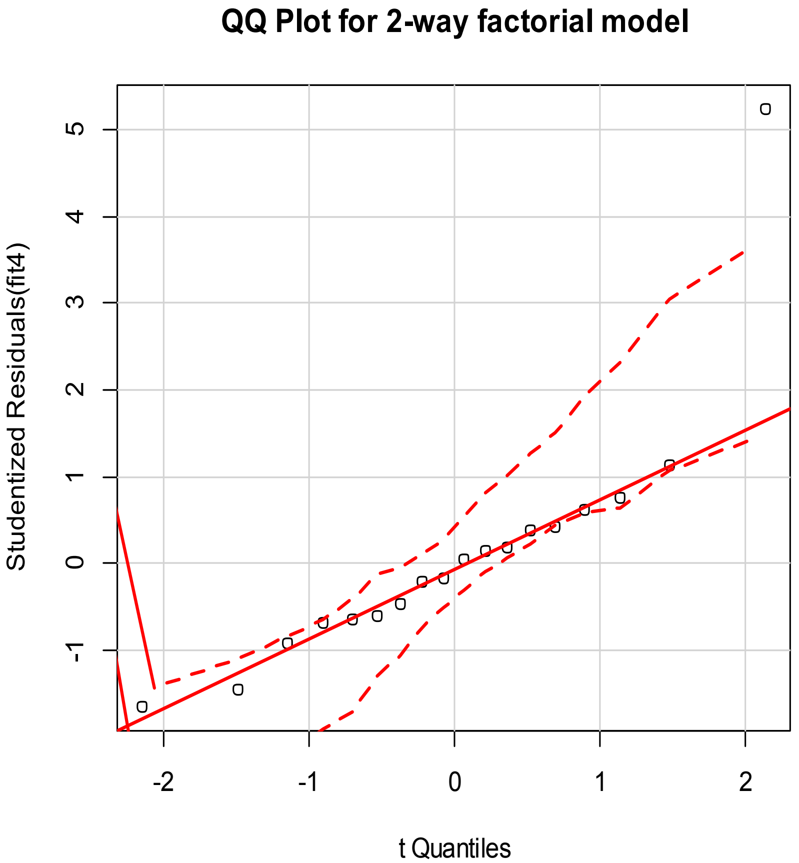
Source	SS	df	MS	F
Factor A	SS Between Factor A levels	k-1	MSA= SSA/df(A)	MSA/MSE
Factor B	SS Factor B levels	j-1	MSB= SSB/df(B)	MSB/MSE
Interaction	SS Due to Interaction between A and B	(k-1) × (j-1)	MSAB= SSAB/df(AB)	MSAB/MSE
Error	SS Within cells	k×j × (n-1)	MSE= SSE/df(E)	

Factor Analysis Example

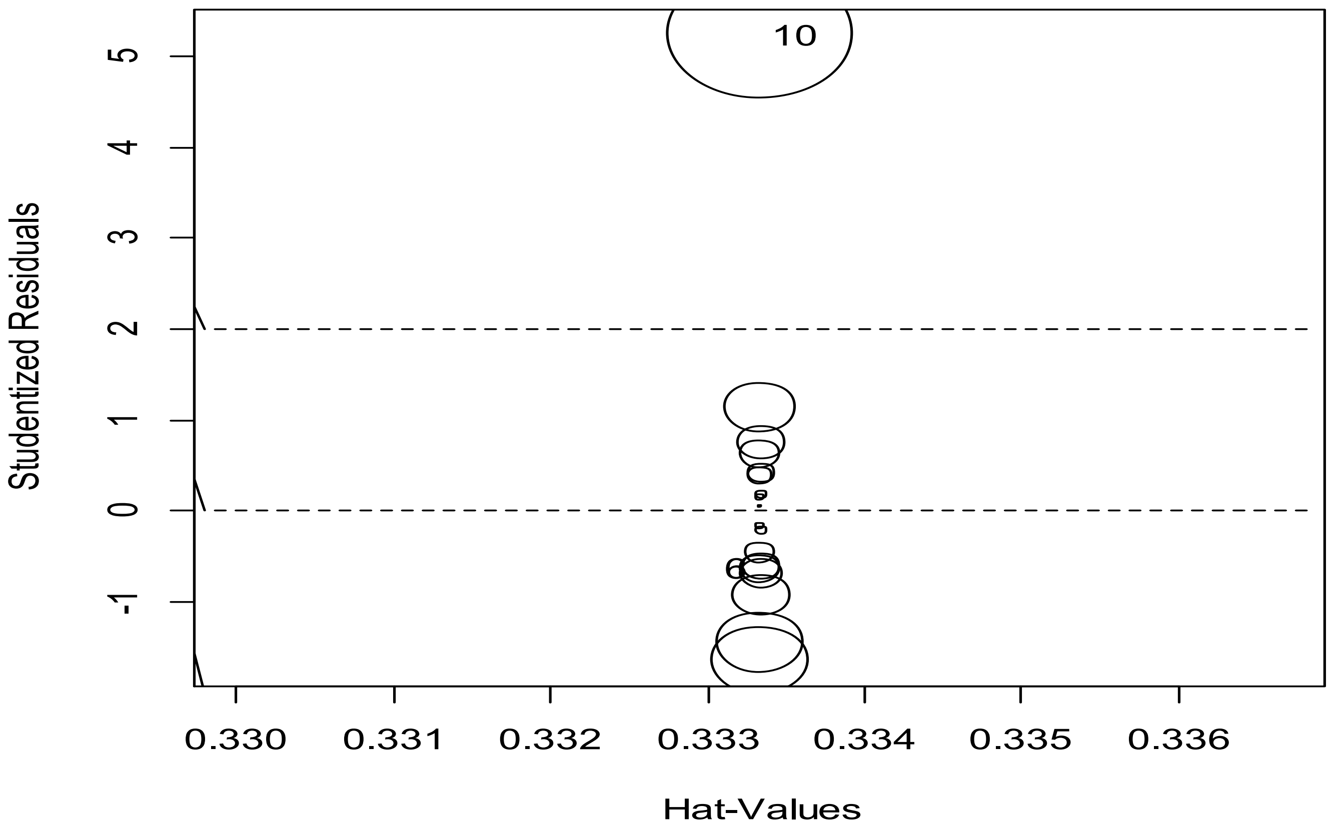
- Use a subset of the COCOMO data base
- Select 6 projects from each Mode category
- Such that 3 project in each Mode category
  - Have high requirements volatility
  - Have normal requirements volatility
- One factor with 3 levels and one factor with two levels
  - Balanced 2*3 Factor Analysis

Log(Productivity) Analysis

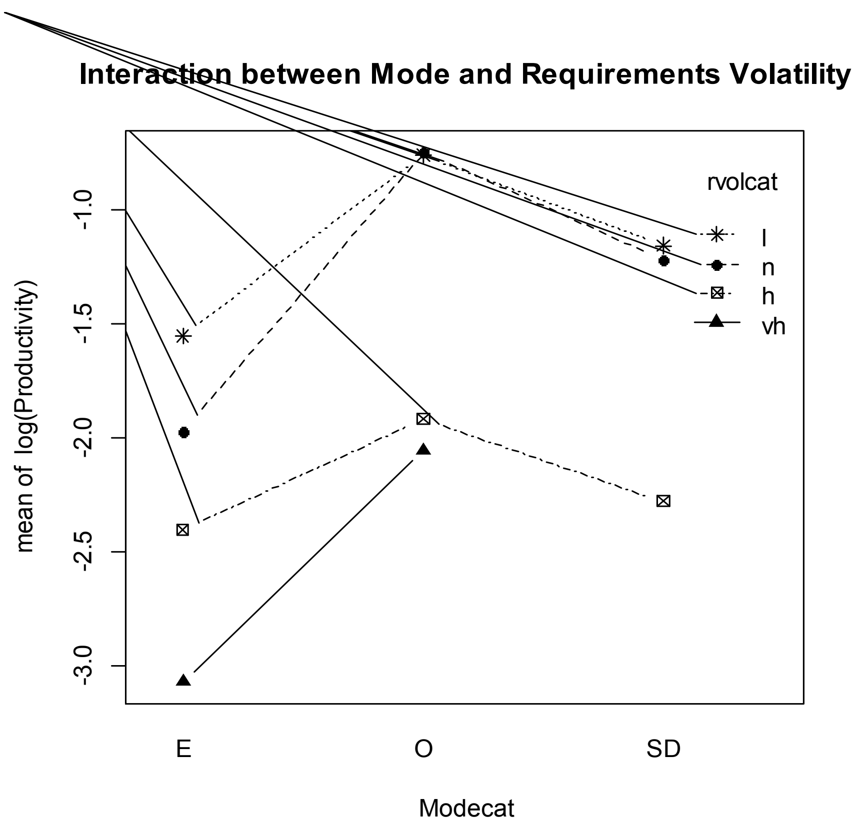


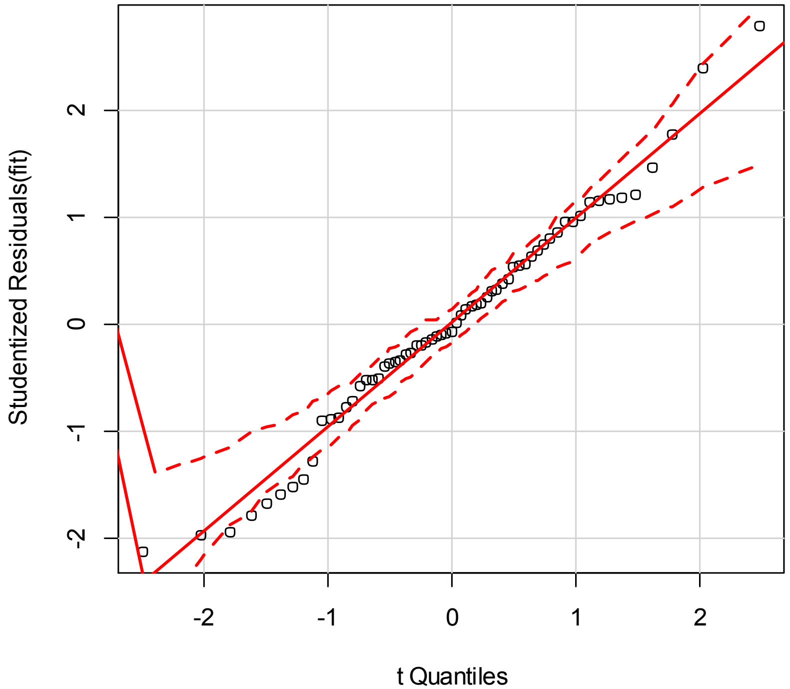


Influence Plot for Log(Productivity)



Full COCOMO Dataset





AOV Order dependency

- For full data set factors are not balanced
- Analysis differs depending on which factor entered first

Mean Log(Productivity) with number of project in each in parenthesis				
Mode	Requirements Volatility			
	L	N	H	VH
E	-1.5554 (1)	-1.9730 (11)	-2.404 (11)	-3.0700 (5)
O	-0.7644 (2)	-0.7511 (15)	-1.9205 (4)	-2.0554 (2)
SD	-1.1595 (2)	-1.2211 (7)	-2.2785 (3)	NA (0)

Term	Fitting First	Requirements Volatility	Mode	Residuals
MS	Mode	4.2***	10.318 **	0.395
MS	Req Vol	7.496 ***	5.373 ***	0.395
df		3	2	57

Random Effects and Mixed Effects

- Random effects model (n observations in each group)

  $$x_{ij} = \mu + \alpha_j + e_{ij}$$

  - where $\alpha_j \sim N(0, \sigma_a^2)$
- Compared with fixed effects
  - α_j are random variables not fixed quantities to be estimated
  - Null hypothesis $\alpha_j = 0$ is the same
  - Under H1, expected value of MSBG= $n\sigma_a^2 + \sigma^2$
  - Differences between models if H0 is false
- Often used to assess different ways of measuring something
  - So main purpose of analysis is to estimate σ_a^2
  - Rarely used in SE except for meta-analysis
- Mixed effects model includes some fixed and some random factors
  - In such models, the F tests may differ from the equivalent fixed effects model
- Mixed and Random effects not handled in basic R configuration

Different types of model

- Is the productivity of different platforms different?
  - Obtain productivity measures from projects produced on the different platforms
  - Fixed effects
- Are two methods of measuring function points equivalent
  - Find 20 FP counters and 10 projects
    - Assign 2 counters to each project
    - Let each counter use both methods on their assigned project
    - Mixed effects
      - Project effect - fixed
      - Method – fixed
      - Person effect - random
      - With-in person error term
      - Between method error term
  - Important to use the correct tests
    - Between method error term must be used to compare methods

Impact of Model type on 2-way Factorial

Mean Squares	Fixed Effects	Random Effects	Mixed Model: A fixed, B Random
A	$\sigma^2 + nbk_A^2$	$\sigma^2 + n\sigma_{AB}^2 + nb\sigma_A^2$	$\sigma^2 + n\sigma_{AB}^2 + nbk_A^2$
B	$\sigma^2 + nak_B^2$	$\sigma^2 + n\sigma_{AB}^2 + na\sigma_B^2$	$\sigma^2 + na\sigma_B^2$
AB	$\sigma^2 + nk_{AB}^2$	$\sigma^2 + n\sigma_{AB}^2$	$\sigma^2 + n\sigma_{AB}^2$
Error	σ^2	σ^2	σ^2

SE Example

- Test Case Prioritization
- Design:
  - 18 techniques
    - 16 different test case prioritisation techniques
    - 2 control techniques
    - Ran experiments in groups of 4 techniques
  - 8 C programs
    - Generated 29 different versions with a random number of non-interfering faults
    - From available set of regression tests for program
      - Extracted 50 different test sets per program version for each method
  - Each experiment could generate
    - $4 \times 8 \times 29 \times 50 = 46400$ observations
    - Although not all combinations possible

Example of ANOVA table

Source	SS	df	MS	F
Program	3472054	7	49615.6	1358
Techn	97408	3	32469.2	88.9
Program*Techn	182322	21	8682.0	23.77
Error	9490507	259086	365.22	

- Is this analysis valid?

Model

- Each observation is based on
  - Program - Fixed
  - Treatment - Fixed
  - Interaction between Treatment and Program
  - Within each program the version used
    - Random effect
  - Within each version test case used for each method
    - Random effect

$$y_{ijkl} = p_i + T_j + (pT)_{ij} + v_{(ij)k} + \epsilon_{(ijk)l}$$

ANOVA Problems

- F-test requires the ratio two chi-squared variables
  - Variance of a Normal variable is chi-squared
  - Also assume the variances are equal for each group
- Affects of non normality and heteroscedastcity
  - Worse if sample sizes differ
- F test is not robust for heavy-tailed or skewed distributions

MANOVA

- Analysis of variance generalised to multiple outcome variables
- Consider analysing Duration, KDSI & Effort (after log transformation) within Mode
- Need to setup a data matrix containing only y variables
- Then use manova(y~Modecat)
  - Need library(MASS)

MANOVA Results

Modecat	Log(Effort)	Log(Dur)	Log(AKDSI)
E	5.8093	2.9453	3.48624
SD	4.7885	2.5510	3.3134
O	3.6552	2.4936	2.5862

- F=8.27 with 6 and 118 degrees of freedom
- p=1.744e-07
- R command summary.aov(fit)
  - Shows ANOVA for each variable separately
  - Only Effort significant at $p<0.05$
- Require
  - Multivariate Normality
  - Homogeneity of variance-covariance matrices

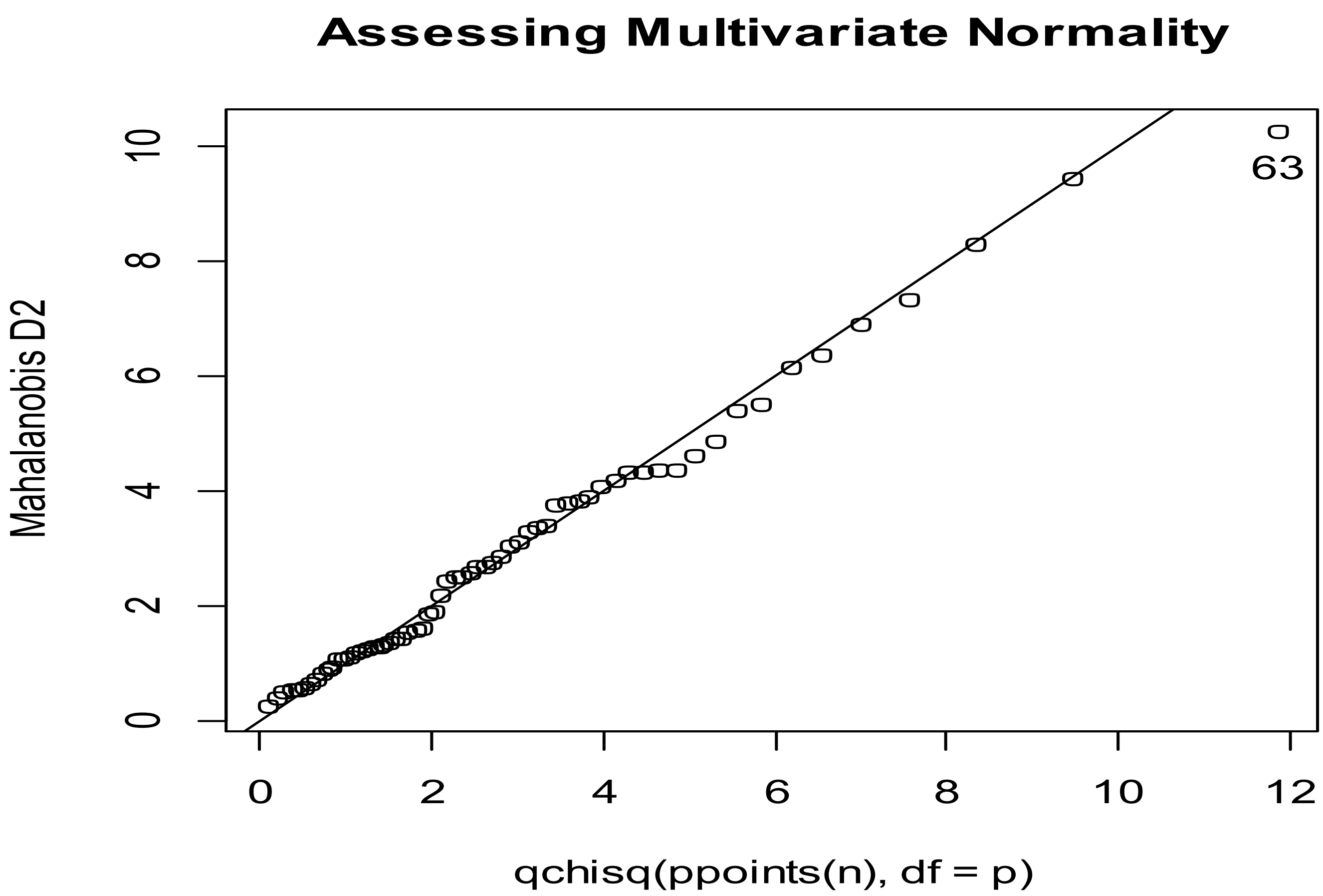
Mahalanobis Distance

- With p×1 multivariate random vector **x** with
  - mean $\bar{X}$
  - variance-covariance matrix **S**
- Mahalobis d² is distance between **x** and squared $\bar{X}$
  - Chi-squared with p degrees of freedom
- Check normality by a qqplot of chi-squared

$$d^2 = \left[1 + \left(x - \bar{X}\right)' S^{-1} \left(x - \bar{X}\right)\right]$$

- Points should be close to lines with slope 1 and intercept 0

qqplot of d^2



Assessing Multivariate Normality
Mahalanobis D2
qchisq(ppoints(n), df = p)
63
0
2
4
6
8
10
12

Robust two-way analyses

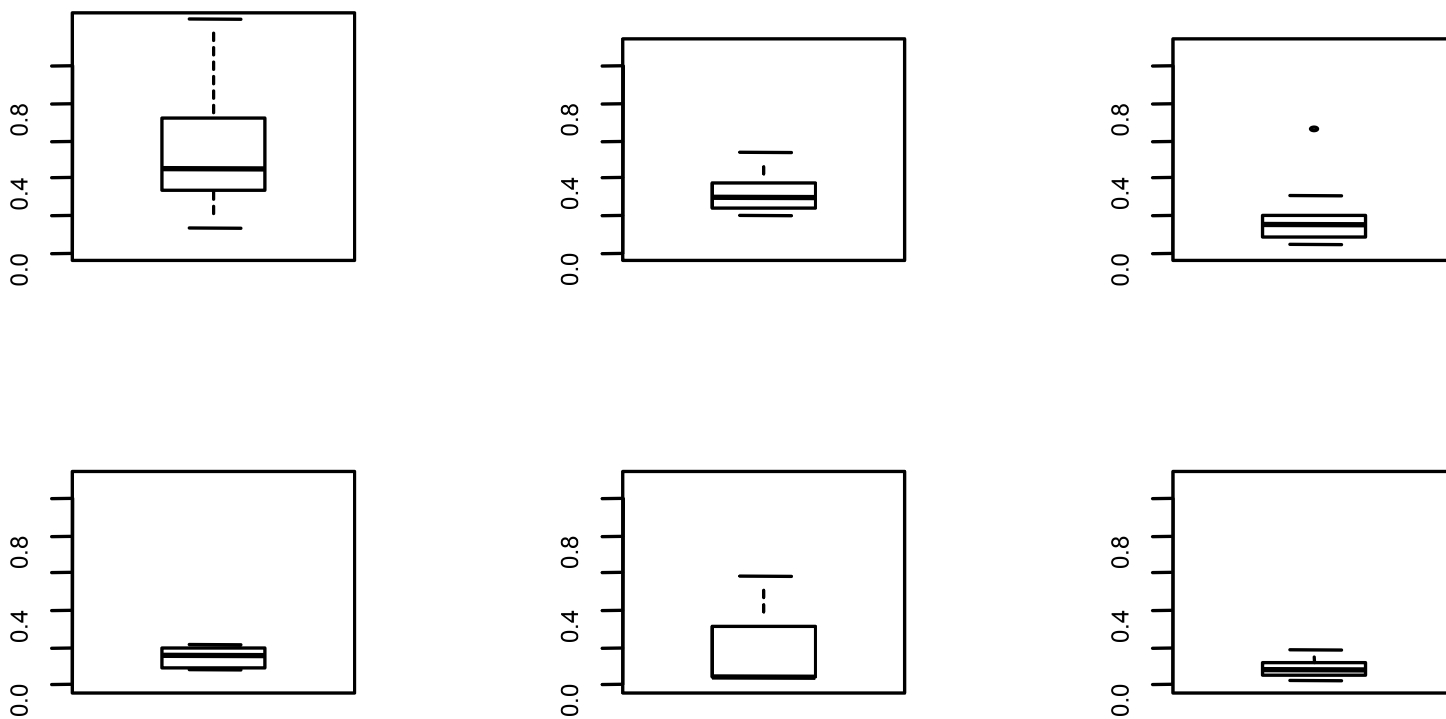
- Trimmed means can be used in a two-way factorial design
- Can cope with lack of balance
  - Same results irrespective of order
- Needs a reasonably large number of units in each cell
  - Command is t2way(J,K,w,tr=p)
  - W is a list with J×K entries
  - Might need to use p=.1 rather than .2 if small numbers of observations per cell
- Recoded rvol categories so
  - Normal & Low counted as one category
  - High and Very high together counted as one category

Constructing List Variable

- w[[1]] contains the values for factor A level 1 and factor B level 1
- w[[2]] … w[[J]] contain the values for factor A level 1 and factor B levels 2 to J
- w[[J+1]] …w[[2J]] contains values for factor A level 2 and factor B levels 1…J
- w[[K(J-1) +1]]…w[[KJ]] contains values for factor A level K and factor B levels 1 to J

Productivity per Cell

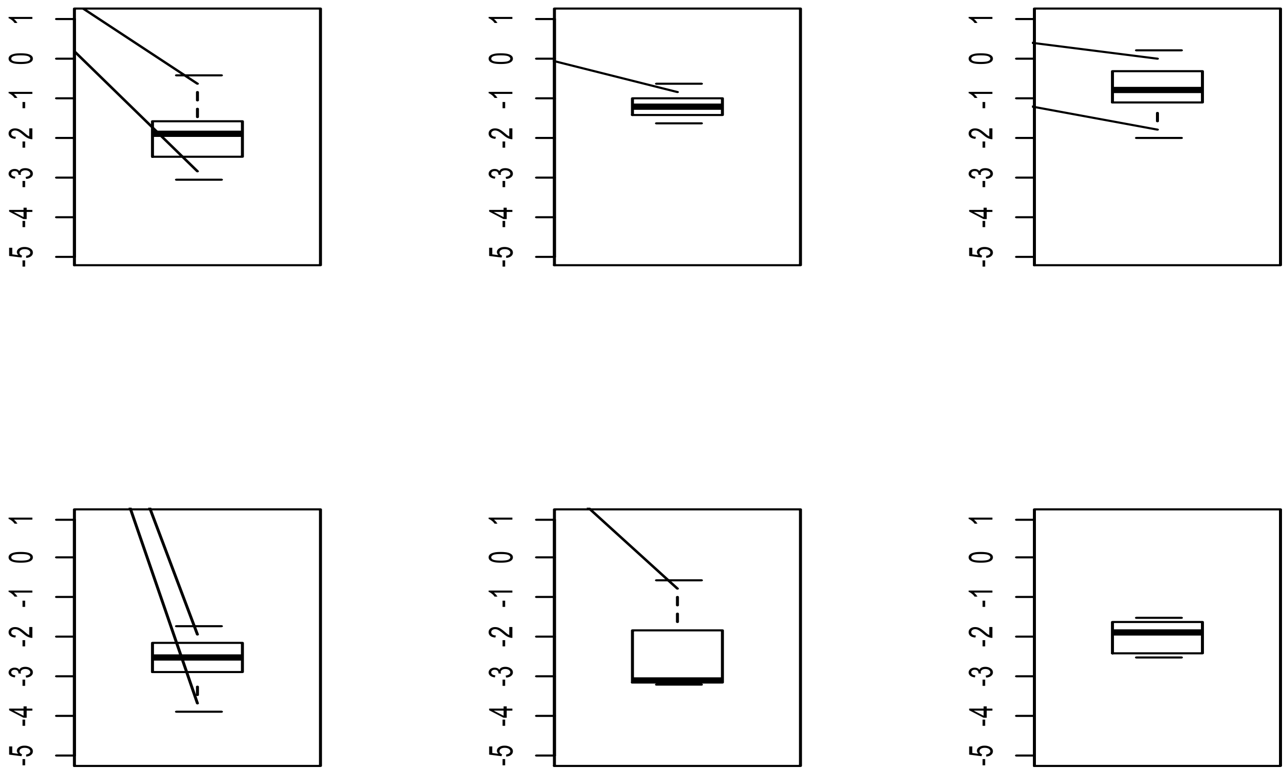
Rvolcat	Mode		
	Organic	Semi-detached	Embedded
N or L	0.5378 (17)	0.3137 (9)	0.1871 (12)
H or VH	0.1507 (6)	0.2 231(3)	0.0866 (16)



Trimmed means results

- Effect due to Requirement Volatility significant ($p=0.05$)
- Effect due to Mode significant ($p=0.001$)
- Interaction significant ($p=0.014$)
- Different results if log(Productivity)
  - Mode ($p=0.002$), Rvol($p=0.031$), Interaction ($p=0.27$)
- Similar results if log(Productivity) & trim=0
  - Mode ($p=0.002$), Rvol ($p=0.029$), Interaction ($p=0.383$)

Log(Productivity)



Non-Parametric Analysis

- Akritas, Arnold & Brunner method
  - Works for unbalanced Factorial design
    - Same results irrespective of order
  - Function: bdm2way(J,K,x)
  - J=number of levels in Factor A
  - K= number of levels in factor B
- Based on w as a list variable (same as for trimmed means)
- Reports the relative effect size

COCOMO Example

- Productivity for factors
  - Requirements volatility (two levels)
  - Mode category E,SD,O
- Requirements volatility effects (p=0.059)
- Mode effects (p=0.205)
- Interaction effects (p=0.624)

Relative effect size	Mode		
Requirements Volatility	Embedded	Semi-Detached	Organic
Normal	0.4140	0.6693	0.7988
High	0.2202	0.3360	0.3995

Additional facilities

- Trimmed means
  - Available for three-way designs
  - Randomised effects
  - Linear contrasts for complex designs
  - MANOVA
  - Not all techniques available in standard R configuration
- With a good transformation available
  - Can transform data and use tr=0
    - For facilities not available in standard R

Conclusions

- ANOVA can easily get too complex to understand
  - Always choose the simplest design possible
  - Preferably one that is fully specified in a statistical text book
  - Main problems are mixed designs with multiple levels and error terms
- ANOVA is reliant on normal distributions but
  - Possible to use trimmed means for Robust analyses
    - However, may be better to transform data
  - Non-parametric methods for designs as complex as two-way factorial designs available in WRS library
    - Allow for unbalanced designs
- ANCOVA covered by regression analysis
- MANOVA facilities available
  - Standard R facilities
  - Trimmed means