ITALIAN JOURNAL OF PUBLIC HEALTH

Pinpointing outliers in experimental data: the Hat matrix in Anova for fixed and mixed effects models

Annalisa Orenti[1], Giuseppe Marano[1], Patrizia Boracchi[1], Ettore Marubini[1]

ABSTRACT

The Hat (H) matrix and in particular the elements of its principal diagonal (leverages) have a paramount importance in multiple regression analysis in order to pinpoint possible outliers and/or influential points as components of several regression diagnostics.
This note presents some features of the H matrix and residuals for ANOVA models of experimental designs. For fixed effects models, the values of the elements of H are discussed in completely randomized, randomized complete block and Latin squares designs. The increasing complexity of the design structure leads to different patterns, with increasing values of the corresponding leverages (h_{ii}). For mixed effects models, developments on leverage and residuals for marginal and conditional estimates are illustrated.
The application of H matrix and residuals in fixed effects and mixed effects model is shown in a worked example. It is concluded that for H matrix in mixed models, an important role is played by the values of the variances of the random effects and the error term, and, consequently, by their method of estimation. Marginal and conditional studentized residuals provide different information about the data, and thus should be both used for model checking.

Key words: Outliers; Hat matrix; Anova; Mixed effects model

(1) Department of clinical sciences and community health, University of Milan, Milan, Italy

CORRESPONDING AUTHOR: *Annalisa Orenti, Department of clinical sciences and community health, University of Milan. c/o IRCCS Istituto Nazionale dei Tumori Via Venezian 1, 20133 Milano. e-mail: annalisa.orenti@unimi.it*
DOI: *10.2427/8663*

INTRODUCTION

In regression analysis attention is focused on assessing the role of each observation in determining values of estimators and test statistics (e.g.Weisberg (1)). The careful study of each observation is necessary to pinpoint possible outliers and/or influential points. The Hat (H) matrix and in particular the elements of its principal diagonal (leverages) have a paramount importance in this context as components of several regression diagnostics. The role of such a matrix in linear models underling the analysis of variance (ANOVA) has been less studied. This note tries to fill this gap by presenting some features of H matrix pertinent to both fixed effects and mixed effects models of ANOVA for balanced data layouts.

The scenario presented here will deal with a two-way (factorial) experiment with an equal number of replications arranged in a completely randomized design, in a randomized complete block design and in a Latin square design. It shows the influence of the design structure increasing complexity on the H matrix. Moreover the parallelism between diagnostics in fixed effects models and mixed effects models is developed.

BASIC NOTATIONS AND TERMINOLOGY

Fixed effects model

In matrix notation the standard linear model is:

$y = Xb + e$ [1]

where: **y** (nx1), response vector, **X** (nxp), fixed-effects design matrix, $\boldsymbol{b}$ (px1) vector of fixed parameters to be estimated, **e** (nx1) unknown vector of random errors, which are assumed to be: e~N (0,σ^2I). Letter n specifies the number of observations and p the number of regressors (including the intercept).

In the Ordinary Least Squares (OLS) analysis the parameter estimate vector $\hat{\boldsymbol{b}}$ of $\boldsymbol{b}$ is obtained by minimising, with respect to $\boldsymbol{b}$, the Error Sum of Squares (ESS): ESS = **e**' **e**.

$\hat{\boldsymbol{b}} = \mathbf{(X'Z)^{-1}X'y}$

The predicted values vector ($\hat{y}$) of **y** is:

$\mathbf{y} = \mathbf{X}\hat{\boldsymbol{b}} = \mathbf{X(X'X)^{-1}X'y} = \mathbf{Hy}$ [2], where: $\mathbf{H} = \mathbf{X(X'X)^{-1}X'}$

is the ($n \times n$) projection or leverage matrix (Hat matrix), being h_{ij} its generic term (i,j=1,2,…,n).

From [2] it appears that $\hat{y}_i$ is a linear combination of the observed values y_j, having h_{ij} as coefficients:

$\hat{y}_i = h_{i1}y_1 + h_{i2}y_2 + \ldots + h_{in}y_n$ [3]

"The element h_{ij} of H has a direct interpretation as the amount of leverage or influence exerted on $\hat{y}_i$ by y_j (regardless of the actual value of y_j, since H depends only on X)" (2).

It is easy to see that H is idempotent (**H** = **HH**), symmetrical (**H** = **H'**) and orthogonal (**H'** = $\mathbf{H^{-1}}$). Furthermore it may be shown that: $\sum_{i=1}^{n} h_{ij} = \sum_{j=1}^{n} h_{ij} = 1$ and that: $\sum_{i=1}^{n} h_{ii} = trace(\mathbf{H}) = rank(\mathbf{X}) = p$

It follows that the average leverage ($\bar{h}$) is: $\bar{h} = \frac{\sum_{i=1}^{n} h_{ii}}{n} = \frac{p}{n}$

Moreover it may be shown that:

$\mathrm{var}(\hat{\mathbf{y}}) = \sigma^2\mathbf{H}$ [4]

so that: $\mathrm{var}(\hat{y}_i) = \sigma^2 h_{ii}$.

The error estimate vector ($\hat{\mathbf{e}}$) of **e** is: $\hat{\mathbf{e}} = \mathbf{y} - \hat{\mathbf{y}} = \mathbf{(I-H)y}$ and $\mathrm{var}(\hat{\mathbf{e}}) = \sigma^2\mathbf{(I-H)}$; hence: $\mathrm{var}(\hat{e}_i) = \sigma^2(1-h_{ii})$ [5]

from the idempotency property: $0 \le h_{ii} \le 1$.

Note that the effect of the factor $(1-h_{ii})$ in [5] consists in down-weighting the random error variance σ^2 so that: $\mathrm{var}(\hat{e}_i) \le \mathrm{var}(e_i)$

The estimate $\hat{\sigma}^2$ of σ^2 is obtained from the Residual Sum of Squares (RSS) as:

$$\hat{\sigma}^2 = \frac{RSS}{n-p} \quad [6]$$

where: $RSS = \hat{\mathbf{e}}'\hat{\mathbf{e}}$.

It is worth recalling that *(1-h_{ii})* is a component of diagnostics suitable for pinpointing outlier observations:

$$\text{studentized residual} = \frac{\hat{e}_i}{\hat{\sigma}\sqrt{1-h_{ii}}} \quad [7]$$

Mixed effects model

Mixed effects models are used in experimental designs involving both fixed and random effects, for example, in Randomized Block designs the levels of the blocking factors are often considered as a random sample from a population of levels.

The standard linear mixed model is an extension of [1], namely:

$y = Xb + Zu + e$

where: Z ($n \times g$) random-effects design matrix, U ($g \times 1$) vector of random effects to be estimated. Letter g specifies the number of random effects. The remaining terms have the same meaning as in [1].

A further assumption is that **u** is normally distributed $\mathbf{u} \sim N(\mathbf{0},\mathbf{G})$ and independent of the random error vector **e**.

As a consequence: $\mathrm{var}(\mathbf{y}) = \sigma^2\mathbf{V}$, where:

$$\mathbf{V} = \frac{\mathbf{ZGZ'}}{\sigma^2} + \mathbf{I} = \mathbf{ZDZ'+I}$$

It is important to realize that in mixed effects models the values of σ^2 and the elements of **G** (i.e. the variance components) are needed to derive all the estimates shown below, since the matrix **V** enters in all the expressions defining each estimate.

In the following expressions elements of **V** are assumed to be known, according to the theoretical results on predictions, residuals and

leverage presented here.

For **V** known, the parameter estimate vectors are:

$\hat{\boldsymbol{b}} = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{y}$ [8]

which is the Best Linear Unbiased Estimator for fixed effects and:

$\hat{\mathbf{u}} = \mathbf{D}\mathbf{Z}'\mathbf{V}^{-1}(\mathbf{y}-\mathbf{X}\hat{\boldsymbol{b}})$ [9]

which is the Best Linear Unbiased Predictor for random effects.

From these the vectors of predicted values are derived:

$\hat{\mathbf{y}}_M = \mathbf{X}\hat{\boldsymbol{b}}$ (marginal predictions);

$\hat{\mathbf{y}}_C = \mathbf{X}\hat{\boldsymbol{b}} + \mathbf{Z}\hat{\mathbf{u}}$ (conditional predictions).

The terms marginal and conditional are justified by the fact that $\hat{\mathbf{y}}_M$ and $\hat{\mathbf{y}}_C$ are estimates of the marginal and conditional means of the response vector: $E(\mathbf{y})$ and $E(\mathbf{y} \mid \mathbf{u} = \hat{\mathbf{u}})$ respectively. Accordingly, two kinds of residuals and leverages are defined.

For marginal predictions, different expressions of the **H** matrix were developed in literature (3); here we consider the one which has the advantage of expressing how unusual an observation is in the regression space (3):

$$\mathbf{H}_1 = \mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}$$

The expressions for marginal predictions and residuals according to Martin (4) are:

$$\hat{\mathbf{y}}_M = \mathbf{H}_1\mathbf{y}$$

$$\hat{\mathbf{e}}_M = \mathbf{y} - \hat{\mathbf{y}}_M = (\mathbf{I} - \mathbf{H}_1)\mathbf{y}$$

$$\mathbf{var}(\hat{\mathbf{e}}_M) = \sigma^2(\mathbf{I} - \mathbf{H}_1)\mathbf{V}$$

studentized marginal residual= $\dfrac{\hat{e}_{Mi}}{\hat{\sigma}\cdot v_{ii}\cdot\sqrt{1-h_{1ii}}}$ [10]

where v_{ii} is the i-th element on the principal diagonal of **V**. It can be shown that for most designs, $\hat{\boldsymbol{b}}$ obtained for the mixed effects model [8] converges to $\hat{\boldsymbol{b}}$ obtained for the fixed effects model as the sample size increases (5).

For conditional predictions we refer to the development of Zewotir and Galpin (6). The **H** matrix has the form:

$$\mathbf{H}_2 = \mathbf{I} - \mathbf{V}^{-1} + \mathbf{V}^{-1}\mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1} \quad [11]$$

Accordingly, the expressions for conditional predictions and residuals are:

$$\hat{\mathbf{y}}_C = \mathbf{H}_2\mathbf{y}$$

$$\hat{\mathbf{e}}_C = \mathbf{y} - \hat{\mathbf{y}}_C = (\mathbf{I} - \mathbf{H}_2)\mathbf{y}$$

$$\mathbf{var}(\hat{\mathbf{e}}_C) = \sigma^2(\mathbf{I} - \mathbf{H}_2)$$

studentized conditional residual= $\dfrac{\hat{e}_{Ci}}{\hat{\sigma}\cdot\sqrt{1-h_{2ii}}}$ [12]

It is worth noticing that conditional residuals are obtained by subtracting fixed and random effect estimates from the response vector, thus they may be thought of as an estimate of the random error component **e** . On the other hand, marginal residuals include also the estimates of the random component **u**.

In practice, matrix V is unknown and must be estimated. Among the methods available to estimate σ^2 and the elements of **G** and thus obtaining $\hat{\mathbf{V}}$, the three most frequently implemented in statistical software are mentioned here.

The Method of Moments implies two steps: in the first one the ANOVA table is computed, and the estimates of variance components are obtained by equating the relative mean squares to their expectations; this enables estimating σ^2 and the elements of **G**, and thus obtaining $\hat{\mathbf{V}}$. In the second step $\hat{\boldsymbol{b}}$ and $\hat{\mathbf{u}}$ are estimated according to [8] and [9] after substituting **V** with $\hat{\mathbf{V}}$. This method is appropriate in balanced designs.

Maximum Likelihood (ML) and Restricted Maximum Likelihood (REML) methods are more effective in complex designs; they are routinely implemented in statistical software, and, therefore, are most widely used nowadays. For both methods, the estimates are based on the values maximizing a likelihood function (namely, a restricted likelihood function in REML), under the assumption of normal distribution of residuals and random effects.

In ML the estimates of variance components and fixed effects are obtained jointly. In REML the likelihood is expressed as the product of two terms: the first including only variance components, and the second with both fixed effects and variance components. The estimation procedure starts maximizing the first term; this generates variance components estimates, which are used in the second term to obtain the estimates of the fixed effects (7).

It is known that ML method estimates of the variance components may be biased downward; this is a side-effect of the maximization of the likelihood function, in which the degrees of freedom of the variance components are calculated treating the fixed effects as known values, and are therefore over-estimated. The bias of variance components affects the estimates of the standard errors of the fixed effects, and hence the inherent inference. On the contrary, the REML method recognizes that fixed effects are estimated and gives correct estimates of their standard errors. In

balanced designs REML estimates of variance components overlap those obtained with the Method of Moments.

SCENARIO

3.1 Completely randomized design

The standard presentation of a 2×2 factorial experiment (7) is shown in Figure 1.

FIGURE 1

STANDARD PRESENTATION OF A 2X2 FACTORIAL EXPERIMENT

		A	
		-	+
B	-	(1)	a
	+	b	ab

Capital letters specify the factors (A,B); each of them has two levels (-, +). Factor A and B are fixed, that is the levels explored in the experiment consist of the entire population of possible levels. The four treatment combinations are indicated by small letters. The highest level of each factor is indicated by the presence of the corresponding small letter, whereas the lowest level is indicated by the absence of the corresponding small letter. Conventionally (1) indicates both factors at lowest level.

According to Milliken and Johnson (5) the treatment combinations define the "treatment structure" which, combined via randomization with the "design structure", enables the experimenter to specify the experiment design. When the experimental units are homogeneous the design structure is that of a completely randomized (CR) design.

The model pertinent to a 2×2 factorial experiment arranged in a completely randomized design is:

$$y_{ijr} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + e_{ijr}$$

for i, j=1,2 and r=1,2,..,R

where μ : overall mean; α_i : effect of factor A i-th level; β_j : effect of factor B j-th level; $(\alpha\beta)_{ij}$: interaction effect; with the conditions:

$$\sum_i \alpha_i = \sum_j \beta_j = 0 \text{ and } \sum_{ij}\sum (\alpha\beta)_{ij} = 0 .$$

Furthermore it is assumed that the random error component $e_{ijr} \sim N(0, \sigma^2_{CR})$.

In this presentation we take R (number of replications per cell)=4.

Table 1 reports the X matrix of the experimental design.

We note that:

$$\hat{y}_{ijr} = \hat{\mu} + \hat{\alpha}_i + \hat{\beta}_j + \widehat{(\alpha\beta)}_{ij} = \bar{y}_{ij} = \frac{y_{ij1} + y_{ij2} + y_{ij3} + y_{ij4}}{R}$$

To implement the **H** matrix according to [3], it is convenient to take two preparatory actions; firstly the notation referring to the treatment combinations is simplified by adopting the suffix k (instead of the couple ij) which takes the values 1, 2, 3, 4, to indicate the treatment combinations (1), a, b, ab respectively and secondly two couples of suffixes are introduced for *h: kr* referring to $\hat{y}$ and *k'r'* referring to *y*.

Then, for instance:

$$\hat{y}_{11} = h_{11,11} \cdot y_{11} + h_{11,12} \cdot y_{12} + h_{11,13} \cdot y_{13} + h_{11,14} \cdot y_{14} + 0 \cdot y_{21} + ... + 0 \cdot y_{44} \quad [13]$$

In general:

$$h_{kr,k'r'} = \begin{cases} \frac{1}{R} & \text{for } k = k' \text{ and for every } r \\ 0 & \text{for } k \neq k' \text{ and for every } r \end{cases}$$

TABLE 1

X MATRIX OF 2X2 FACTORIAL EXPERIMENT ARRANGED IN A COMPLETELY RANDOMIZED DESIGN

	μ	α	β	αβ
[1,]	1	-1	-1	1
[2,]	1	-1	-1	1
[3,]	1	-1	-1	1
[4,]	1	-1	-1	1
[5,]	1	1	-1	-1
[6,]	1	1	-1	-1
[7,]	1	1	-1	-1
[8,]	1	1	-1	-1
[9,]	1	-1	1	-1
[10,]	1	-1	1	-1
[11,]	1	-1	1	-1
[12,]	1	-1	1	-1
[13,]	1	1	1	1
[14,]	1	1	1	1
[15,]	1	1	1	1
[16,]	1	1	1	1

Table 2 gives the **H** matrix for the completely randomized 2×2 factorial design. Note that the linear model is saturated since rank (X)=4=number of the design cells. As the number of replications is constant the corresponding **H** matrix is a block diagonal

TABLE 2

H MATRIX OF 2 X 2 FACTORIAL EXPERIMENT ARRANGED IN A COMPLETELY RANDOMIZED DESIGN

	[,1]	[,2]	[,3]	[,4]	[,5]	[,6]	[,7]	[,8]	[,9]	[,10]	[,11]	[,12]	[,13]	[,14]	[,15]	[,16]
[1,]	0,25	0,25	0,25	0,25	0	0	0	0	0	0	0	0	0	0	0	0
[2,]	0,25	0,25	0,25	0,25	0	0	0	0	0	0	0	0	0	0	0	0
[3,]	0,25	0,25	0,25	0,25	0	0	0	0	0	0	0	0	0	0	0	0
[4,]	0,25	0,25	0,25	0,25	0	0	0	0	0	0	0	0	0	0	0	0
[5,]	0	0	0	0	0,25	0,25	0,25	0,25	0	0	0	0	0	0	0	0
[6,]	0	0	0	0	0,25	0,25	0,25	0,25	0	0	0	0	0	0	0	0
[7,]	0	0	0	0	0,25	0,25	0,25	0,25	0	0	0	0	0	0	0	0
[8,]	0	0	0	0	0,25	0,25	0,25	0,25	0	0	0	0	0	0	0	0
[9,]	0	0	0	0	0	0	0	0	0,25	0,25	0,25	0,25	0	0	0	0
[10,]	0	0	0	0	0	0	0	0	0,25	0,25	0,25	0,25	0	0	0	0
[11,]	0	0	0	0	0	0	0	0	0,25	0,25	0,25	0,25	0	0	0	0
[12,]	0	0	0	0	0	0	0	0	0,25	0,25	0,25	0,25	0	0	0	0
[13,]	0	0	0	0	0	0	0	0	0	0	0	0	0,25	0,25	0,25	0,25
[14,]	0	0	0	0	0	0	0	0	0	0	0	0	0,25	0,25	0,25	0,25
[15,]	0	0	0	0	0	0	0	0	0	0	0	0	0,25	0,25	0,25	0,25
[16,]	0	0	0	0	0	0	0	0	0	0	0	0	0,25	0,25	0,25	0,25

one with all blocks of the same size (4×4) and $\bar{h} = \frac{p}{n} = \frac{p}{pR} = \frac{1}{R} = \frac{1}{4} = h_{kr,kr}$ for k, r=1, 2, 3, 4.

So that:

$$\text{var}(\hat{e}_{kr})_{CR} = (1 - h_{kr,kr})\sigma^2_{CR} = (1 - 0.25)\sigma^2_{CR} = 0.75\sigma^2_{CR}$$

[14]

Alternatively, according to [4], the elements of **H** can be obtained in terms of $\text{var}(\hat{y}_{kr})$ and $\text{cov}(\hat{y}_{kr}, \hat{y}_{k^*r^*})$.

Be $\bar{y}_1$,$\bar{y}_2$, $\bar{y}_3$ and $\bar{y}_4$ the averages of the observations in cell (1), a, b and ab respectively. It is peculiar of any saturated model to capture all information regarding the relationship "between" the responses of the design cells, so that $\hat{y}_{kr} = \bar{y}_k$ for r=1, 2, 3, 4. Consider the first block: we know that $\text{var}(\bar{y}_1) = \frac{\sigma^2}{4}$.

On the other hand, from [4], $\text{var}(\hat{y}_{kr}) = \sigma^2 h_{kr,kr}$ thus $h_{kr,kr} = \frac{1}{4} = 0.25$

As regards the terms outside the principal diagonal of the first *4x4* block we observe from [4] that $\text{cov}(\hat{y}_{1r}, \hat{y}_{1r^*}) = \sigma^2 \cdot h_{1r,1r^*}$. However for r=1, 2, 3, 4,

$$\text{cov}(\hat{y}_{1r}, \hat{y}_{1r^*}) = E[(\hat{y}_{1r} - E(\hat{y}_{1r}))(\hat{y}_{1r^*} - E(\hat{y}_{1r^*}))] =$$
$$E[(\bar{y}_1 - E(\bar{y}_1))(\bar{y}_1 - E(\bar{y}_1))] = E[(\bar{y}_1 - E(\bar{y}_1))^2] =$$
$$= \text{var}(\bar{y}_1) = \frac{\sigma^2}{4}$$

Therefore $h_{1r,1r^*} = \frac{1}{4}$ for all r=1, 2, 3, 4. The same considerations apply to the remaining three blocks.

Let's now consider the value of $\boldsymbol{b}$ for two predicted values $\hat{\mathbf{y}}$ belonging to different blocks; for instance:

$$h_{14,24} = \text{cov}(\hat{y}_{14}, \hat{y}_{24}) = E[(\hat{y}_{14} - E(\hat{y}_{14}))(\hat{y}_{24} - E(\hat{y}_{24}))] =$$

$E[(\bar{y}_1 - E(\bar{y}_1))(\bar{y}_2 - E(\bar{y}_2))] = 0$, owing to the independence of the two deviations from the mean as result of the random allocation of experimental units to the different cells of the design.

Following [5]

$$\text{var}(\hat{e}_{kr}) = \sigma^2(1 - h_{kr,kr}) = \sigma^2\left(1 - \frac{1}{R}\right).$$

For $R \to \infty$, $\text{var}(\hat{e}_{kr}) \to \sigma^2$, this is coherent with the observation that for $R \to \infty$ $\hat{\boldsymbol{b}} \to \boldsymbol{b}$, and $\hat{\mathbf{e}} \to \mathbf{e}$.

For R=1, $\text{var}(\hat{e}_{kr}) = 0$, meaning that only one replication of the design does not imply residual variability. In such a case the **H** matrix corresponds to a $p \times p$ identity matrix.

The variance σ^2 is estimated by $\hat{\sigma}^2$ as given in equation [6]. As a consequence:

$$\hat{\text{var}}(\hat{e}_{kr}) = \hat{\sigma}^2\left(1 - \frac{1}{R}\right) = \frac{\hat{\mathbf{e}}'\hat{\mathbf{e}}}{n-p}\left(\frac{R-1}{R}\right) = \frac{\hat{\mathbf{e}}'\hat{\mathbf{e}}}{p(R-1)}\left(\frac{R-1}{R}\right) = \frac{\hat{\mathbf{e}}'\hat{\mathbf{e}}}{pR} = \frac{\hat{\mathbf{e}}'\hat{\mathbf{e}}}{n}$$

Randomized complete blocks design

When the experimental units are heterogeneous the previous design is no longer appropriate; as an alternative a randomized complete blocks (RCB) design could be adopted if the number of experimental units per block is equal to the number of treatments (t) to be investigated. It is assumed that experimental units within each block are homogeneous whereas they differ from block to block. Two further assumptions are that (i) there is no interaction between treatments and blocks and (ii) the blocks effects are fixed.

As in this scenario four treatments are considered, we need four homogeneous units per block and four blocks are necessary to get four replications of the design.

The four treatments are randomly allocated to each of the four units belonging to a given block.

The pertinent model is now:

$$y_{ijl} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \pi_l + e_{ijl}$$

for i, j=1,2 and l=1,2,3,4

where μ, α_i , β_j , $(\alpha\beta)_{ij}$ have already been specified in previous model; π_l is the effect of the l-th block; with the further condition: $\sum_l \pi_l = 0$. Moreover it is assumed that the random error component $e_{ijl} \sim N(0, \sigma^2_{RCB})$.

Table 3 reports the X matrix of this experiment.

Note that:

$$\hat{y}_{ijl} = \hat{\mu} + \hat{\alpha}_i + \hat{\beta}_j + \hat{(\alpha\beta)}_{ij} + \hat{\pi}_l = \bar{y}_{ij.} + \bar{y}_{..l} - \bar{y}_{...}$$

As previously, to implement the **H** matrix the suffix k substitutes the couple ij, so that $\hat{y}_{kl} = \bar{y}_{k.} + \bar{y}_{.l} - \bar{y}_{..}$ Then, according to [3], for example:

$$\hat{y}_{11} = \bar{y}_{1.} + \bar{y}_{.1} - \bar{y}_{..} = \frac{1}{4}y_{11} + \frac{1}{4}y_{12} + \frac{1}{4}y_{13} + \frac{1}{4}y_{14} + \frac{1}{4}y_{11} + \frac{1}{4}y_{21} + \frac{1}{4}y_{31} + \frac{1}{4}y_{41} - \frac{1}{16}\sum_{k,l} y_{kl} =$$
$$= \left(\frac{1}{4}+\frac{1}{4}-\frac{1}{16}\right)y_{11} + \left(\frac{1}{4}-\frac{1}{16}\right)y_{12} + \left(\frac{1}{4}-\frac{1}{16}\right)y_{13} + \left(\frac{1}{4}-\frac{1}{16}\right)y_{14} + \left(\frac{1}{4}-\frac{1}{16}\right)y_{21} - \frac{1}{16}y_{22} - \frac{1}{16}y_{23} - \frac{1}{16}y_{24}$$
$$+\left(\frac{1}{4}-\frac{1}{16}\right)y_{31} - \frac{1}{16}y_{32} - \frac{1}{16}y_{33} - \frac{1}{16}y_{34} + \left(\frac{1}{4}-\frac{1}{16}\right)y_{41} - \frac{1}{16}y_{42} - \frac{1}{16}y_{43} - \frac{1}{16}y_{44}$$

[15]

In general:

$$h_{kl,k'l'} = \begin{cases} \frac{1}{K}+\frac{1}{L}-\frac{1}{KL} & \text{for } k = k' \text{ and } l = l' \\ \frac{1}{L}-\frac{1}{KL} & \text{for } k = k' \text{ and } l \neq l' \\ \frac{1}{K}-\frac{1}{KL} & \text{for } k \neq k' \text{ and } l = l' \\ -\frac{1}{KL} & \text{for } k \neq k' \text{ and } l \neq l' \end{cases}$$

TABLE 3

X MATRIX OF 2X2 FACTORIAL EXPERIMENTARRANGED IN A RANDOMIZED COMPLETE BLOCK DESIGN

	μ	α	β	αβ	π_1	π_2	π_3
[1,]	1	-1	-1	1	1	0	0
[2,]	1	-1	-1	1	0	1	0
[3,]	1	-1	-1	1	0	0	1
[4,]	1	-1	-1	1	-1	-1	-1
[5,]	1	1	-1	-1	1	0	0
[6,]	1	1	-1	-1	0	1	0
[7,]	1	1	-1	-1	0	0	1
[8,]	1	1	-1	-1	-1	-1	-1
[9,]	1	-1	1	-1	1	0	0
[10,]	1	-1	1	-1	0	1	0
[11,]	1	-1	1	-1	0	0	1
[12,]	1	-1	1	-1	-1	-1	-1
[13,]	1	1	1	1	1	0	0
[14,]	1	1	1	1	0	1	0
[15,]	1	1	1	1	0	0	1
[16,]	1	1	1	1	-1	-1	-1

Table 4 reports the pertinent **H** matrix, which is no more a block diagonal matrix; however all the elements of its principal diagonal have the same value: $h_{kl,kl} = \bar{h} = \frac{7}{16} = 0.4375$ for k, l=1, 2, 3, 4.

Coherently:

$$\text{var}(\hat{e}_{kl})_{RCB} = (1 - h_{kl,kl})\sigma^2_{RCB} = (1 - 0.4375)\sigma^2_{RCB} = 0.5625\sigma^2_{RCB}$$

[16]

At a glance to the **X** matrices reported in Table 1 and Table 3, we note that the latter corresponds to the former augmented of the vectors allowing for the block effects. The three vectors π_1 , π_2 , π_3 are orthogonal to the four vectors that span the treatment space (saturated model). Thus the estimates of the treatment effects are expected to remain invariant. On the contrary the blocking effect reduces the residual sum of squares by introducing non null correlation terms between $\hat{y}_{kl}$ pertaining to different treatments. This is evident comparing the **H** matrices in Tables 2 and 4 and it is the result of the difference between the two linear combinations of y , generating the two corresponding $\hat{y}$ with the same suffixes as, for example, it may be seen comparing equations [13] to [15]. The mentioned reduction is "paid" in terms of 12-9=3 d.f. of RSS.

Latin squares design

So far blocking was postulated in one direction, for example forming blocks in terms of inbreeding lines of rats (units). However the experimenter could be interested in blocking in a second direction, irrespective of the first, for example age of rats. In this case the 4 treatment combinations of 4x4 factorial experiment could be arranged in a Latin square (LS) design like the one shown in Figure 2. It is the first 4x4 Latin square given in Table XV by Fisher and Yates (8).

FIGURE 2

ARRANGEMENT OF A 2X2 FACTORIAL DESIGN IN A 4X4 LATIN SQUARE STRUCTURE

(1)	a	b	ab
a	(1)	ab	b
b	ab	a	(1)
ab	b	(1)	a

The *4X4* arrangement of experimental units is blocked in two directions: rows (inbreeding lines) and columns (age classes). To implement a Latin square design, the treatments are randomly allocated to experimental units in the square such that each treatment occurs once and only once in each row and once and only once in each column. This structure assures that the vectors corresponding to the row effects are orthogonal to those corresponding to the column effects and all of them are orthogonal to the vectors specifying the treatment effects.

The standard model for Latin square designs reported in literature (9) is:

$y_{klm} = \mu + \tau_{klm} + \rho_l + \gamma_m + e_{klm}$ for

k, l, m=1, 2, 3, 4

Where: μ is the overall true mean; τ_{klm} is the effect of the k-th treatment combination; ρ_l is the effect of the l-th row; γ_m is the effect of the m-th column. Further conditions: $\sum_l \rho_l = \sum_m \gamma_m = 0$ and $e_{klm} \sim N(0, \sigma^2_{LS})$.

Table 5 reports the **X** matrix of this model.

TABLE 4

H MATRIX OF 2X2 FACTORIAL EXPERIMENT ARRANGED IN A COMPLETE BLOCK DESIGN

	[,1]	[,2]	[,3]	[,4]	[,5]	[,6]	[,7]	[,8]	[,9]	[,10]	[,11]	[,12]	[,13]	[,14]	[,15]	[,16]
[1,]	,4375	,1875	,1875	,1875	,1875	-,0625	-,0625	-,0625	,1875	-,0625	-,0625	-,0625	,1875	-,0625	-,0625	-,0625
[2,]	,1875	,4375	,1875	,1875	-,0625	,1875	-,0625	-,0625	-,0625	,1875	-,0625	-,0625	-,0625	,1875	-,0625	-,0625
[3,]	,1875	,1875	,4375	,1875	-,0625	-,0625	,1875	-,0625	-,0625	-,0625	,1875	-,0625	-,0625	-,0625	,1875	-,0625
[4,]	,1875	,1875	,1875	,4375	-,0625	-,0625	-,0625	,1875	-,0625	-,0625	-,0625	,1875	-,0625	-,0625	-,0625	,1875
[5,]	,1875	-,0625	-,0625	-,0625	,4375	,1875	,1875	,1875	,1875	-,0625	-,0625	-,0625	,1875	-,0625	-,0625	-,0625
[6,]	-,0625	,1875	-,0625	-,0625	,1875	,4375	,1875	,1875	-,0625	,1875	-,0625	-,0625	-,0625	,1875	-,0625	-,0625
[7,]	-,0625	-,0625	,1875	-,0625	,1875	,1875	,4375	,1875	-,0625	-,0625	,1875	-,0625	-,0625	-,0625	,1875	-,0625
[8,]	-,0625	-,0625	-,0625	,1875	,1875	,1875	,1875	,4375	-,0625	-,0625	-,0625	,1875	-,0625	-,0625	-,0625	,1875
[9,]	,1875	-,0625	-,0625	-,0625	,1875	-,0625	-,0625	-,0625	,4375	,1875	,1875	,1875	,1875	-,0625	-,0625	-,0625
[10,]	-,0625	,1875	-,0625	-,0625	-,0625	,1875	-,0625	-,0625	,1875	,4375	,1875	,1875	-,0625	,1875	-,0625	-,0625
[11,]	-,0625	-,0625	,1875	-,0625	-,0625	-,0625	,1875	-,0625	,1875	,1875	,4375	,1875	-,0625	-,0625	,1875	-,0625
[12,]	-,0625	-,0625	-,0625	,1875	-,0625	-,0625	-,0625	,1875	,1875	,1875	,1875	,4375	-,0625	-,0625	-,0625	,1875
[13,]	,1875	-,0625	-,0625	-,0625	,1875	-,0625	-,0625	-,0625	,1875	-,0625	-,0625	-,0625	,4375	,1875	,1875	,1875
[14,]	-,0625	,1875	-,0625	-,0625	-,0625	,1875	-,0625	-,0625	-,0625	,1875	-,0625	-,0625	,1875	,4375	,1875	,1875
[15,]	-,0625	-,0625	,1875	-,0625	-,0625	-,0625	,1875	-,0625	-,0625	-,0625	,1875	-,0625	,1875	,1875	,4375	,1875
[16,]	-,0625	-,0625	-,0625	,1875	-,0625	-,0625	-,0625	,1875	-,0625	-,0625	-,0625	,1875	,1875	,1875	,1875	,4375

TABLE 5

X MATRIX OF 2X2 FACTORIAL EXPERIMENT ARRANGED IN A LATIN SQUARE

		TREATMENTS EFFECTS			ROWS EFFECTS			COLUMNS EFFECTS		
	μ	α	β	αβ	π_1	π_2	π_3	γ_1	γ_2	γ_3
[1,]	1	-1	-1	1	1	0	0	1	0	0
[2,]	1	-1	-1	1	0	1	0	0	1	0
[3,]	1	-1	-1	1	0	0	1	-1	-1	-1
[4,]	1	-1	-1	1	-1	-1	-1	0	0	1
[5,]	1	1	-1	-1	1	0	0	0	1	0
[6,]	1	1	-1	-1	0	1	0	1	0	0
[7,]	1	1	-1	-1	0	0	1	0	0	1
[8,]	1	1	-1	-1	-1	-1	-1	-1	-1	-1
[9,]	1	-1	1	-1	1	0	0	0	0	1
[10,]	1	-1	1	-1	0	1	0	-1	-1	-1
[11,]	1	-1	1	-1	0	0	1	1	0	0
[12,]	1	-1	1	-1	-1	-1	-1	0	1	0
[13,]	1	1	1	1	1	0	0	-1	-1	-1
[14,]	1	1	1	1	0	1	0	0	0	1
[15,]	1	1	1	1	0	0	1	0	1	0
[16,]	1	1	1	1	-1	-1	-1	1	0	0

Note that:

$$\hat{y}_{klm} = \hat{\mu} + \hat{\tau}_k + \hat{\rho}_l + \hat{\gamma}_m = \bar{y}_{k..} + \bar{y}_{.l.} + \bar{y}_{..m} - 2\bar{y}_{...}$$

As previously done the suffix k substitutes the couple ij; however in this case to implement the H matrix according to [3] it is convenient to introduce two triplets of suffixes for *h: klm* referring to $\hat{y}$ and $k'l'm'$ referring to *y*.

Then, for example:

$$\hat{y}_{111} = \bar{y}_{1..} + \bar{y}_{.1.} + \bar{y}_{..1} - 2\bar{y}_{...} = \frac{1}{4}\sum_{l,m} y_{1lm} + \frac{1}{4}\sum_{k,m} y_{k1m} + \frac{1}{4}\sum_{k,l} y_{kl1} - \frac{2}{16}\sum_{k,l,m} y_{klm}$$

In general:

$$h_{klm,k'l'm'} = \begin{cases} \frac{1}{K} + \frac{1}{L} + \frac{1}{M} - \frac{2}{LM} & \text{for } k = k' \text{ and } l = l' \text{ and } m = m' \\ \frac{1}{K} - \frac{2}{LM} & \text{for } k = k', l \neq l', m \neq m' \text{ or } k \neq k', l = l', m \neq m' \text{ or } k \neq k', l \neq l', m = m' \\ -\frac{1}{LM} & \text{for } k \neq k', l \neq l', m \neq m' \end{cases}$$

Table 6 reports the H matrix of this model. As in the previous examples, all the elements of the H matrix principal diagonal have the same value:

$$h_{klm,klm} = \bar{h} = \frac{10}{16} = 0.625 \quad \text{for k, l, m=1, 2, 3, 4.}$$

The variance of $\hat{e}_{klm}$ is now:

$$\text{var}(\hat{e}_{klm})_{LS} = (1 - h_{klm,klm})\sigma^2_{LS} = (1 - 0.625)\sigma^2_{LS} = 0.375\sigma^2_{LS} \quad [17]$$

Increasing the complexity of the design from completely randomized to randomized complete block design or to Latin square design, can be justified by postulating:

$\sigma^2_{RCB} < \sigma^2_{CR}$

or

$\sigma^2_{LS} < \sigma^2_{CR}$

From equations [14] and [16] we get:

$$\text{var}(\hat{e}_i)_{RCB} = 0.5625\sigma^2_{RCB} < \text{var}(\hat{e}_i)_{CR} = 0.75\sigma^2_{CR}$$

Similarly, from equations [14] and [17] we get:

$$\text{var}(\hat{e}_i)_{LS} = 0.375\sigma^2_{LS} < \text{var}(\hat{e}_i)_{CR} = 0.75\sigma^2_{CR} \cdot$$

Finally, observing that h_{ii} increases as the number of directions of orthogonal blocking increases, one can argue that the **H** matrix will correspond to a *nxn* identity matrix if orthogonal blocking will make allowance of the whole random variability. This matrix is the counterpart of the *pxp* identity matrix already mentioned in the subsection Completely randomized design.

A WORKED EXAMPLE

Table 7 reports an *ad hoc* dataset which will be processed in details. It is a *2x2* factorial design in the presence of interaction, with true responses: (1)=5, a=8, b=7, ab=16. The random errors were generated by means of the *rnorm* function of software R: e_{ijl} ~N(0,1). A non homogeneity of experimental units due to inbreeding lines (blocking factor) was postulated. Thus we are dealing with a randomized complete block factorial experiment.

The arrangement of treatments per block adopted in the following example is given in Figure 3.

FIGURE 3

ARRANGEMENT OF 2X2 FACTORIAL DESIGN IN A RANDOMIZED COMPLETE BLOCK STRUCTURE

	UNITS			
	1	2	3	4
Block 1	(1)	a	b	ab
Block 2	b	ab	(1)	a
Block 3	ab	b	a	(1)
Block 4	a	(1)	ab	b

TABLE 6

H MATRIX OF 2X2 FACTORIAL EXPERIMENT ARRANGED IN A LATIN SQUARE

	[,1]	[,2]	[,3]	[,4]	[,5]	[,6]	[,7]	[,8]	[,9]	[,10]	[,11]	[,12]	[,13]	[,14]	[,15]	[,16]
[1,]	0.625	0.125	0.125	0.125	0.125	0.125	-0.125	-0.125	0.125	-0.125	-0.125	0.125	0.125	-0.125	0.125	-0.125
[2,]	0.125	0.625	0.125	0.125	0.125	0.125	-0.125	-0.125	-0.125	0.125	0.125	-0.125	-0.125	0.125	-0.125	0.125
[3,]	0.125	0.125	0.625	0.125	-0.125	-0.125	0.125	0.125	0.125	-0.125	0.125	-0.125	-0.125	0.125	0.125	-0.125
[4,]	0.125	0.125	0.125	0.625	-0.125	-0.125	0.125	0.125	-0.125	0.125	-0.125	0.125	0.125	-0.125	-0.125	0.125
[5,]	0.125	0.125	-0.125	-0.125	0.625	0.125	0.125	0.125	0.125	-0.125	0.125	-0.125	0.125	-0.125	-0.125	0.125
[6,]	0.125	0.125	-0.125	-0.125	0.125	0.625	0.125	0.125	-0.125	0.125	-0.125	0.125	-0.125	0.125	0.125	-0.125
[7,]	-0.125	-0.125	0.125	0.125	0.125	0.125	0.625	0.125	-0.125	0.125	0.125	-0.125	0.125	-0.125	0.125	-0.125
[8,]	-0.125	-0.125	0.125	0.125	0.125	0.125	0.125	0.625	0.125	-0.125	-0.125	0.125	-0.125	0.125	-0.125	0.125
[9,]	0.125	-0.125	0.125	-0.125	0.125	-0.125	-0.125	0.125	0.625	0.125	0.125	0.125	0.125	0.125	-0.125	-0.125
[10,]	-0.125	0.125	-0.125	0.125	-0.125	0.125	0.125	-0.125	0.125	0.625	0.125	0.125	0.125	0.125	-0.125	-0.125
[11,]	-0.125	0.125	0.125	-0.125	0.125	-0.125	0.125	-0.125	0.125	0.125	0.625	0.125	-0.125	-0.125	0.125	0.125
[12,]	0.125	-0.125	-0.125	0.125	-0.125	0.125	-0.125	0.125	0.125	0.125	0.125	0.625	-0.125	-0.125	0.125	0.125
[13,]	0.125	-0.125	-0.125	0.125	0.125	-0.125	0.125	-0.125	0.125	0.125	-0.125	-0.125	0.625	0.125	0.125	0.125
[14,]	-0.125	0.125	0.125	-0.125	-0.125	0.125	-0.125	0.125	0.125	0.125	-0.125	-0.125	0.125	0.625	0.125	0.125
[15,]	0.125	-0.125	0.125	-0.125	-0.125	0.125	0.125	-0.125	-0.125	-0.125	0.125	0.125	0.125	0.125	0.625	0.125
[16,]	-0.125	0.125	-0.125	0.125	0.125	-0.125	-0.125	0.125	-0.125	-0.125	0.125	0.125	0.125	0.125	0.125	0.625

TABLE 7

AD HOC GENERATED DATA SET FOR A RANDOMIZED COMPLETE BLOCK FACTORIAL EXPERIMENT

Y	TREATMENT	BLOCK
3.264	(1)	1
6.202	(1)	2
6.388	(1)	3
4.147	(1)	4
6.302	a	1
8.472	a	2
8.663	a	3
8.563	a	4
6.061	b	1
6.816	b	2
7.397	b	3
7.727	b	4
12.372	ab	1
17.178	ab	2
16.220	ab	3
18.231	ab	4

Fixed effects model

Table 8 reports the ANOVA results. The estimated residual mean square is $\hat{\sigma}^2_{RCB} = 1.201$; together with leverages of **H** matrix reported in Table 4 it enables the computation of the studentized residuals according to [7]. The graph of these residuals against predicted values ($\hat{y}_{kl}$) is drawn in Figure 4. Each point is labelled with a number indicating the block (1, 2, 3, 4) and by a symbol indicating the treatment combination as in Figure 1. The emerging message is the absence of outliers, as all the sixteen points lie within the bands (-2,2).

Table 9 reports the estimates of treatments fixed effects (left side) and of blocks fixed effects (right side). Recalling the first 4 columns of Table 3, it is straightforward to obtain the response estimates for the four cells of the design; for instance for the ab cell: $\hat{y}_{ab} = 9 + 3 + 2.5 + 1.5 = 16$.

Owing to the fact that interaction is significant, the hypothesis concerning the main effects of factors A and B are difficult to interpret. This can be bypassed considering the "simple effects", i.e. the differences (d) at two levels:

- factor A (B present): ab-b
- factor A (B absent): a-(1)

Symmetrically:

- factor B (A present): ab-a
- factor B (A absent): b-(1)

The 95% confidence intervals of the true value of each of these differences are:

$$d \pm t_{9,0.975}\hat{\sigma}_{RCB}\sqrt{\frac{1}{4}+\frac{1}{4}} \quad [18]$$

TABLE 8

ANOVA TABLE OF THE DATA REPORTED IN TABLE 7

SOURCE OF VARIATION	DF	SUM OF SQUARES	MEAN SQUARE	F-VALUE	P-VALUE
A	1	144	144	119.868	<0.0001
B	1	100	100	83.2417	<0.0001
A*B	1	36	36	29.9670	0.0004
BLOCKS	3	21.358	7.119	5.9261	0.0163
RESIDUAL	9	10.812	1.201		

FIGURE 4

GRAPH OF STUDENTIZED RESIDUALS VERSUS PREDICTED VALUES FOR THE DATA REPORTED IN TABLE 7

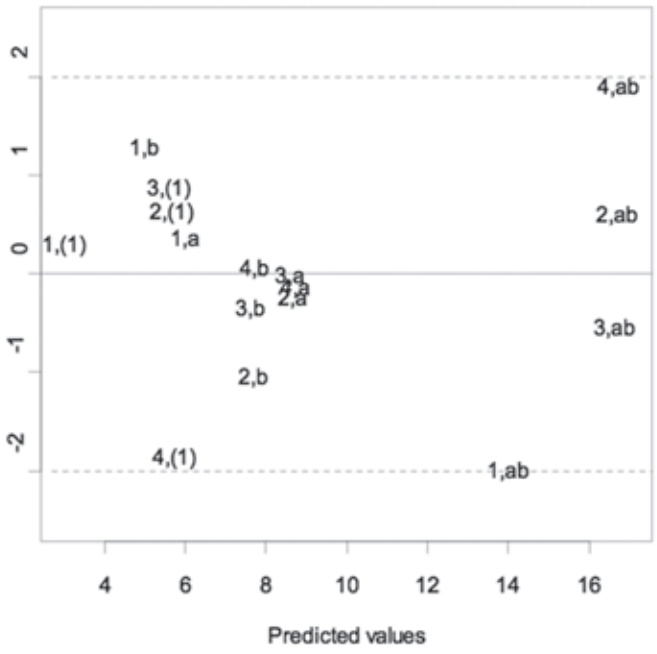

TABLE 9

ESTIMATE OF FIXED EFFECTS FOR THE DATA REPORTED IN TABLE 7

TREATMENTS		BLOCKS ($\tilde{u}_j$)	
$\hat{\mu}$	9.0	1	-2.0
$\hat{\alpha}_2$	3.0	2	0.67
$\hat{\beta}$	2.5	3	0.61
$(\hat{\alpha\beta})_{22}$	1.5	4	0.72

where $t_{9,0.975}$ is the 97.5 centile of the Student t distribution with 9 degrees of freedom.

One could argue that using a fixed effects model for blocks is questionable and alternatively it is more appropriate to consider the blocks as random factors. Consequently the analysis should be carried out in terms of mixed effects model.

Mixed effects model

In this simple example, the mixed effect model includes the fixed component (treatment), the random component (block) with mean 0 and variance $\mathbf{G} = \sigma^2_{\mu}$ and the random error.

Therefore two estimates of variance components are needed: $\hat{\sigma}^2_{\mathbf{u}}$ and $\hat{\sigma}^2_{RCB}$. The REML estimates are: $\hat{\sigma}^2_{\mathbf{u}} = 1.479$ and $\hat{\sigma}^2_{RCB} = 1.201$. The latter is equal to the corresponding estimate from the fixed effects model (Table 8). The same values are obtained according to the Method of Moments from the following expressions: $\hat{\sigma}^2_{RCB}$ =MSResidual, $\hat{\sigma}^2_{\mathbf{u}}$ =(MSBlocks – MSResidual)/4; where 4 is the number of observed block levels and MSBlocks, MSResidual are reported in Table 8.

Table 10 gives the estimates of fixed and random effects of this model. By comparing Table 8 and 10 it is easy to note that the estimates of fixed effects (treatments) are the same, whereas the estimates of random effects (blocks) in the mixed effects model are lower, in absolute value, than the corresponding ones in the fixed effects model. This shows the shrinkage toward zero property of mixed effects models. In the latter the prediction of block effect $\hat{u}_i$ corresponds to the estimated block effect in the fixed effects model $\tilde{u}_i$ multiplied by a coefficient which always lies between 0 and 1, and depends on the variance components (10). In the example shown here:

$$\hat{u}_i = \left(\frac{\hat{\sigma}^2_{\mathbf{u}}}{\hat{\sigma}^2_{\mathbf{u}} + \frac{\hat{\sigma}^2_{RCB}}{n}} \right) \tilde{u}_i$$

TABLE 10

MIXED EFFECTS MODEL ESTIMATES FOR THE DATA REPORTED IN TABLE 7

TREATMENTS		BLOCKS ($\hat{u}$)	
$\hat{\mu}$	9.0	1	-1.663
$\hat{\alpha}_2$	3.0	2	0.557
$\hat{\beta}$	2.5	3	0.507
$(\hat{\alpha\beta})_{22}$	1.5	4	0.599

The scatter plots of the studentized marginal residuals (according to [10]) and the studentized conditional residuals (according to [12]) against predicted values ($\hat{y}_{kl}$) are drawn in Figure 5. This graph is drawn here for marginal residuals, even though we are aware that studentized marginal residuals and marginal predictions are not orthogonal. In panel A the marginal residuals pertaining to block 1 show a clear pattern: they are lower than those pertaining to the other levels, especially point (1, ab) which is under the threshold -2; furthermore, the effect of block 1 (-1.663) contrasts the effects of the other three blocks. This pattern does not appear in panel B, as only marginal residuals include estimates of the random component (blocking factor) as discussed in section Basic Notation and Terminology. However the point (1, ab) is under the threshold here too. The emerging message is that the results produced by block 1 should be thoroughly investigated to assess the validity of the corresponding observations. Nonetheless we can observe that in fixed effects model no residual is outside the range (-2, 2), while the residual plots of the mixed effects model are able to pinpoint the outlier observation (1, ab).

Graph of studentized marginal residuals versus marginal predicted values (Panel A), graph of studentized conditional residuals versus conditional predicted values (Panel B) for the data reported in Table 7

As regards the "simple effects" confidence intervals the formula is the same used in fixed effects model [18]. Consider for instance $d_1 = ab - b$:

$$\mathrm{var}(d_1) = \mathrm{var}(ab) + \mathrm{var}(b) - 2\,\mathrm{cov}(ab, b) =$$

$$\frac{\sigma^2_{RCB} + \sigma^2_{\mathbf{u}}}{4} + \frac{\sigma^2_{RCB} + \sigma^2_{\mathbf{u}}}{4} - 2 \cdot \frac{\sigma^2_{\mathbf{u}}}{4} = \sigma^2_{RCB}\left(\frac{1}{4} + \frac{1}{4}\right)$$

Thus all variances of the simple effects are free of the block variance component $\sigma^2_{\mathbf{u}}$. "This is the manifestation of the randomized complete block design controlling block variation" (11).

Let's come back to the **H** matrices: the matrix of marginal leverages $\hat{\mathbf{H}}_1$ (not reported here) turns out to be practically identical to the hat matrix of the completely randomized fixed effects design (shown in Table 2), in which the block effects are absent. The matrix of conditional leverages $\hat{\mathbf{H}}_2$ (Table 11) turns out to have a pattern similar to that of the hat matrix of the randomized complete block fixed effects design (Table 4), even though the corresponding numerical values are different. In fact $\hat{\mathbf{H}}_2$ depends strongly upon $\hat{\mathbf{V}}$ as shown by [11].

CONCLUDING REMARKS

In fixed effects models the **H** matrix depends only on the design structure, reflected by the **X** matrix. In balanced completely randomized designs the **H** matrix assumes the form of a block diagonal matrix and each $h_{ii} = \frac{1}{R}$, where R is the common number of

FIGURE 5

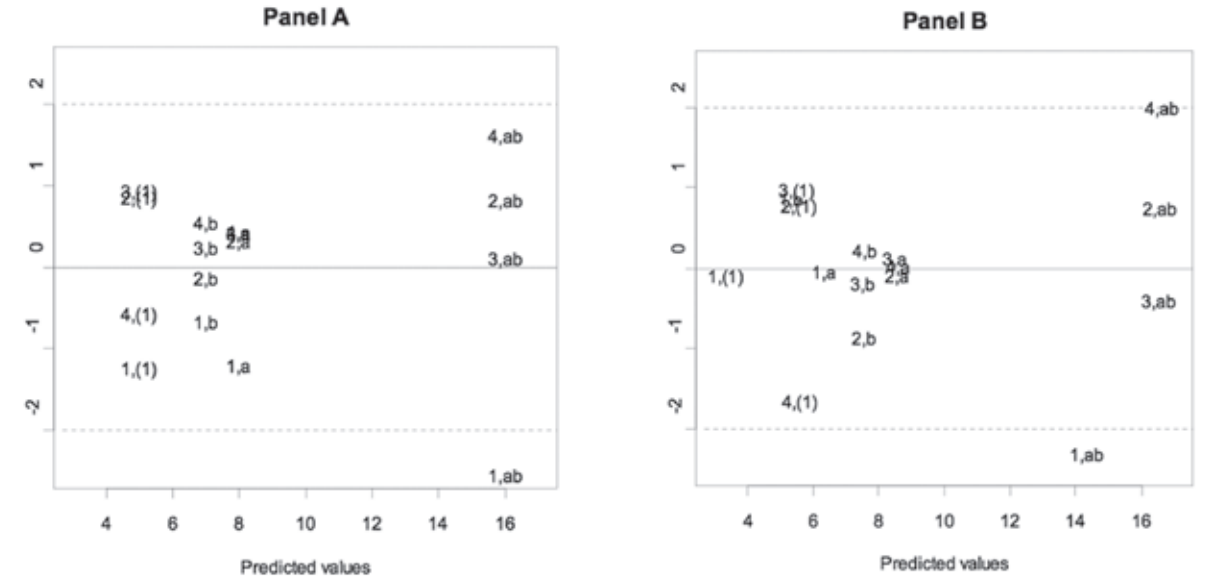

TABLE 11

CONDITIONAL LEVERAGE MATRIX OF THE MIXED EFFECTS MODEL FOR DATA REPORTED IN TABLE 7

	[,1]	[,2]	[,3]	[,4]	[,5]	[,6]	[,7]	[,8]	[,9]	[,10]	[,11]	[,12]	[,13]	[,14]	[,15]	[,16]
[1,]	.4058	.1981	.1981	.1981	.1558	-.0519	-.0519	-.0519	.1558	-.0519	-.0519	-.0519	.1558	-.0519	-.0519	-.0519
[2,]	.1981	.4058	.1981	.1981	-.0519	.1558	-.0519	-.0519	-.0519	.1558	-.0519	-.0519	-.0519	.1558	-.0519	-.0519
[3,]	.1981	.1981	.4058	.1981	-.0519	-.0519	.1558	-.0519	-.0519	-.0519	.1558	-.0519	-.0519	-.0519	.1558	-.0519
[4,]	.1981	.1981	.1981	.4058	-.0519	-.0519	-.0519	.1558	-.0519	-.0519	-.0519	.1558	-.0519	-.0519	-.0519	.1558
[5,]	.1558	-.0519	-.0519	-.0519	.4058	.1981	.1981	.1981	.1558	-.0519	-.0519	-.0519	.1558	-.0519	-.0519	-.0519
[6,]	-.0519	.1558	-.0519	-.0519	.1981	.4058	.1981	.1981	-.0519	.1558	-.0519	-.0519	-.0519	.1558	-.0519	-.0519
[7,]	-.0519	-.0519	.1558	-.0519	.1981	.1981	.4058	.1981	-.0519	-.0519	.1558	-.0519	-.0519	-.0519	.1558	-.0519
[8,]	-.0519	-.0519	-.0519	.1558	.1981	.1981	.1981	.4058	-.0519	-.0519	-.0519	.1558	-.0519	-.0519	-.0519	.1558
[9,]	.1558	-.0519	-.0519	-.0519	.1558	-.0519	-.0519	-.0519	.4058	.1981	.1981	.1981	.1558	-.0519	-.0519	-.0519
[10,]	-.0519	.1558	-.0519	-.0519	-.0519	.1558	-.0519	-.0519	.1981	.4058	.1981	.1981	-.0519	.1558	-.0519	-.0519
[11,]	-.0519	-.0519	.1558	-.0519	-.0519	-.0519	.1558	-.0519	.1981	.1981	.4058	.1981	-.0519	-.0519	.1558	-.0519
[12,]	-.0519	-.0519	-.0519	.1558	-.0519	-.0519	-.0519	.1558	.1981	.1981	.1981	.4058	-.0519	-.0519	-.0519	.1558
[13,]	.1558	-.0519	-.0519	-.0519	.1558	-.0519	-.0519	-.0519	.1558	-.0519	-.0519	-.0519	.4058	.1981	.1981	.1981
[14,]	-.0519	.1558	-.0519	-.0519	-.0519	.1558	-.0519	-.0519	-.0519	.1558	-.0519	-.0519	.1981	.4058	.1981	.1981
[15,]	-.0519	-.0519	.1558	-.0519	-.0519	-.0519	.1558	-.0519	-.0519	-.0519	.1558	-.0519	.1981	.1981	.4058	.1981
[16,]	-.0519	-.0519	-.0519	.1558	-.0519	-.0519	-.0519	.1558	-.0519	-.0519	-.0519	.1558	.1981	.1981	.1981	.4058

replications for the design cells. Moving from a completely randomized to a randomized complete block design or to a Latin squares design, the **H** matrix is not a block diagonal one any longer as no null correlation terms between y_{kl} pertaining to different treatments are introduced. However, in the three mentioned designs the elements of the principal diagonal of the corresponding H matrix are $h_{ii} = \bar{h} = \frac{p}{n}$.

Increasing the complexity of the design structure does increase p and thus increases the corresponding h_{ii} .

As regards mixed effects models, two different **H** matrices are needed: $\mathbf{H}_1$ to compute marginal predictions and residuals, and $\mathbf{H}_2$ to compute conditional predictions and residuals. Furthermore both of them cannot be implemented a *priori* on the knowledge of the design structure, as they depend not only on the **X** matrix, but also on the components of variance matrix **V**, which, in its turn, must be estimated from the data. Furthermore, as different methods of estimating **V** are available, the $\hat{\mathbf{H}}$ matrices depend also on the estimation method chosen. Therefore it is not surprising that studies with the same experimental design give substantially different $\hat{\mathbf{H}}$ matrices.

In fixed effects models the studentized residuals [7] enable pinpointing possible outliers, whereas in mixed effects models marginal [10] and conditional [12] studentized residuals provide different useful information about the data, and thus both should be used for model checking.

References

(1) Weisberg S. Applied linear regression. New York: John Wiley and Sons, 1980

(2) Hoaglin DC, Welsch RE. The Hat Matrix in Regression and ANOVA. The American Statistician; 32(1): 17-22; 1978

(3) Schabenberger O. Mixed model influence diagnostics. SUGI 29 – Statistics and Data Analysis, Paper 189-29. 2004

(4) Martin RJ. Leverage, influence and residuals in regression models when observations are correlated. Communications in Statistics – Theory and Methods 21(6): 1183-1212; 1992

(5) Milliken GA, Johnson DE. Analysis of messy data, Volume I: designed experiments, second edition. New York: Chapman & Hall/CRC, 2009

(6) Zewotir T, Galpin JS. A unified approach on residuals, leverages and outliers in the linear mixed model. Test 16: 58-75; 2007

(7) Corbeil RR, Searle SR. Restricted Maximum Likelihood (REML) Estimation of Variance Components in the Mixed. Technometrics 1976; 18(1): 31-38

(8) Yates F. The design and analysis of factorial experiments. Commonwealth agricultural bureaux, 1958

(9) Fisher RA, Yates F. Statistical tables for biological agricultural and medical research. Edinburgh: Oliver and Boyd, 1963

(10) McCulloch CE, Searle S. Generalized, Linear, and Mixed Models. New York: John Wiley and Sons, 2001

(11) Littell RC, Milliken GA, WW Stroup WW, et al. SAS for mixed models, Second edition. Cary, NY: SAS Institute Inc. 2006

*