# On Formulations of Models for Factor Effects

Presented at LinStat2014

Linköping University

Aug. 24-28, 2014

by Lynn Roy LaMotte

LSUHSC School of Public Health, New Orleans, USA

The setting

A fixed model for the observation vector $\boldsymbol{Y}$

$$\mathrm{E}(Y) = X\beta$$

that includes possible effects of factors A, B, ..., perhaps with other terms.

There is interest in testing factor effects, such as A main effects, AB interaction effects, etc.

Balanced models that include only factor effects:

classical ANOVA sums of squares

standard hypotheses

Unbalanced, or covariates:

Confusion and controversy

Statistical computing packages provide test statistics

based on sums of squares (Type II, III)

not on hypotheses.

Which sum of squares? On what basis?

Macnaughton 1998

Langsrud 2003

McCullagh 2000

Nelder 1977

Yates 1934

Hector, von Felten, Schmid 2010 J. Animal Ecology

Hocking (Methods and Applications of Linear Models, Third Edition, Wiley 2013)

Formulates models and hypotheses for factor effects in two-factor models

some specific development for empty cells

no general development for testing estimable part

How did we get here?

ANOVA sums of squares, Fisher and others, balanced settings.

Unbalanced: orthogonality of SSs lost, confounding.

Yates 1934. Tests marginal means definition of A effects ($\bar{\eta}_{i\cdot}$ all equal) in model that includes A, B, and AB effects.

1960s: Type III SSs extend Yates's construction to general linear models so long as all dfs are estimable.

How are factor effects defined in terms of the cell means?

How can the estimable part of a factor effect be identified and tested?

Can models be formulated such that each factor effect can be tested by omitting its columns from the model?

Do Type III tests test factor effects?

## Two Factors

| Levels of |  | FLC | Responses | Cell Means |
|---|---|---|---|---|
| A | B | $\ell = (\ell_1, \ell_2)$ | $s = 1, \ldots, n_\ell$ | $\eta_\ell = \mathrm{E}(Y_{\ell,s})$ |
| 1 | 1 | $(1, 1)$ | $\{y_{(1,1),1}, \ldots, y_{(1,1),n_{(1,1)}}\}$ | $\eta_{(1,1)}$ |
| 1 | 2 | $(1, 2)$ | $\{y_{(1,2),1}, \ldots, y_{(1,2),n_{(1,2)}}\}$ | $\eta_{(1,2)}$ |
| 1 | 3 | $(1, 3)$ | $\{y_{(1,3),1}, \ldots, y_{(1,3),n_{(1,3)}}\}$ | $\eta_{(1,3)}$ |
| 2 | 1 | $(2, 1)$ | $\{y_{(2,1),1}, \ldots, y_{(2,1),n_{(2,1)}}\}$ | $\eta_{(2,1)}$ |
| 2 | 2 | $(2, 2)$ | $\{y_{(2,2),1}, \ldots, y_{(2,2),n_{(2,2)}}\}$ | $\eta_{(2,2)}$ |
| 2 | 3 | $(2, 3)$ | $\{y_{(2,3),1}, \ldots, y_{(2,3),n_{(2,3)}}\}$ | $\eta_{(2,3)}$ |

## Factor effects defined

Two factors, A and B at *a* and *b* levels, resp.

Cell means $\boldsymbol{\eta} = (\eta_{\ell_1\ell_2})$, $\ell_2$ fastest, an *ab*-vector

$$U_m = (1/m)\mathbf{1}_m\mathbf{1}_m', \qquad S_m = \mathrm{I}_m - U_m$$

A effects: $(S_a \otimes U_b)\boldsymbol{\eta} = H_{10}\boldsymbol{\eta} = (\bar{\eta}_{\ell_1\cdot} - \bar{\eta}_{\cdot\cdot})$

B effects: $(U_a \otimes S_b)\boldsymbol{\eta} = H_{01}\boldsymbol{\eta} = (\bar{\eta}_{\cdot\ell_2} - \bar{\eta}_{\cdot\cdot})$

AB effects: $(S_a \otimes S_b)\boldsymbol{\eta} = H_{11}\boldsymbol{\eta} = (\eta_{\ell_1\ell_2} - \bar{\eta}_{\ell_1\cdot} - \bar{\eta}_{\cdot\ell_2} + \bar{\eta}_{\cdot\cdot})$

$\mathbb{K}$: $n_{..}$ rows, $ab$ columns.

Subject $s$ has $\ell_s$ FLC.

$\mathbb{K}$ has 1 in $s$-th row, $\ell_s$-th column, 0s elsewhere.

If all $n_\ell$s are $n$, $\mathbb{K}$ is a row-permutation of $\mathrm{I}_{ab} \otimes \mathbf{1}_n$.

$\boldsymbol{\mu}$: the $n_{\cdot\cdot}$ mean vector

The regression model

$$\boldsymbol{\mu} = \mathbb{K}\boldsymbol{\eta}$$

The observation model

$$y_{\ell_1\ell_2 s} = \mu_0 + \alpha_{\ell_1} + \beta_{\ell_2} + \gamma_{\ell_1\ell_2} + \epsilon_{\ell_1\ell_2 s}$$

The dummy-variable model for $\boldsymbol{\eta}$ in terms of $\boldsymbol{\theta} = (\mu_0, \boldsymbol{\alpha}', \boldsymbol{\beta}', \boldsymbol{\gamma}')'$:

$$\begin{aligned}
\boldsymbol{\eta} &= (\overbrace{\mathbf{1}_a \otimes \mathbf{1}_b}^{\text{1: 00}}, \overbrace{\mathrm{I}_a \otimes \mathbf{1}_b}^{\text{A: 10}}, \overbrace{\mathbf{1}_a \otimes \mathrm{I}_b}^{\text{B: 01}}, \overbrace{\mathrm{I}_a \otimes \mathrm{I}_b}^{\text{AB: 11}})\boldsymbol{\theta} \\
&= (E_{00}, E_{10}, E_{01}, E_{11})\boldsymbol{\theta} = E\boldsymbol{\theta}.
\end{aligned}$$

The DV model for $\boldsymbol{\mu}$: $\boldsymbol{\mu} = \mathbb{K}E\boldsymbol{\theta} = X_E\boldsymbol{\theta}$.

Objective: Test $\mathrm{H}_0 : H_{10}\boldsymbol{\eta} = \mathbf{0}$,

that there are no A main effects.

In terms of $\boldsymbol{\theta}$,

$$\mathrm{H}_0 : (S_a \otimes \mathbf{1}_b)\boldsymbol{\alpha} + (S_a \otimes U_b)\gamma = \mathbf{0}.$$

This doesn't zero out any term (A, B, AB),

so we can't get the restricted model by dropping terms

from the full model.

What is the *correct* sum of squares?

A digression on testing linear hypotheses.

Model $\boldsymbol{\mu} = X\boldsymbol{\beta}$

Test $\mathrm{H}_0 : G'\boldsymbol{\beta} = \mathbf{0}$.

Restricted-Model – Full Model (RMFM) difference in SSE

tests

the *estimable part* of $G'\boldsymbol{\beta}$.

Some justification.

Restricted model:

$$\{X\beta : G'\beta = \mathbf{0}\} = \mathrm{sp}(XN),$$

where $\mathrm{sp}(N) = \mathrm{sp}(G)^{\perp}$.

Numerator SS is

$$\boldsymbol{y}'(\mathbf{P}_X - \mathbf{P}_{XN})\boldsymbol{y}.$$

*Equivalent estimable conditions* are $(\mathbf{P}_X - \mathbf{P}_{XN})X\beta = \mathbf{0}$:

$$\{X\beta : G'\beta = \mathbf{0}\} = \{X\beta : (\mathbf{P}_X - \mathbf{P}_{XN})X\beta = \mathbf{0}\}.$$

And

$$\mathrm{sp}(X'(\mathbf{P}_X - \mathbf{P}_{XN})) = \mathrm{sp}(X') \cap \mathrm{sp}(G),$$

the space of estimable functions among $G'\beta$.

Such tests are not provided directly in statistical packages.

SSEs for deleted-variables models can be had in any package.

$X = (X_1, X_2)$, restricted model $\mathrm{sp}(X_1)$: SSE after omitting the columns in $X_2$.

A formulation of Factor Effects models in terms of ...

factor effects.

Illustration: three factors, A, B, and C.

Factor effects defined as $H_j\eta$.

(1): $H_{000} = U_a \otimes U_b \otimes U_c$

A: $H_{100} = S_a \otimes U_b \otimes U_c$

B: $H_{010} = U_a \otimes S_b \otimes U_c$

AB: $H_{110} = S_a \otimes S_b \otimes U_c$

etc.

A model for $\eta$ that includes (1), A, B, AC

$$\boldsymbol{\eta} = (H_{000} + H_{100} + H_{010} + H_{101})\boldsymbol{\tau} = H_{\mathcal{J}}\boldsymbol{\tau}.$$

Including $H_{\boldsymbol{j}}$ permits $\boldsymbol{j}$ effects, excluding $H_{\boldsymbol{j}}$ prohibits $\boldsymbol{j}$ effects.

Thus, in this model, $H_{001}\boldsymbol{\eta} = \mathbf{0}$ – the C marginal means are 0 although the AC marginal means might be not 0.

The model for the mean vector is

$$\boldsymbol{\mu} = \mathbb{K}\boldsymbol{\eta} = (\mathbb{K}H_{\mathcal{J}})\boldsymbol{\tau}$$

Additional terms for covariates $\boldsymbol{x}$ and factor-covariate interaction effects

$$D_{\boldsymbol{x}} \mathbb{K} H_{\mathcal{L}} \boldsymbol{\theta}$$

Any DV model can be expressed with $H_j$ matrices.

Every model built with $H_j$s is a factor-effects (FE) model.

Some FE models cannot be expressed with DVs.

The restricted model for testing a set of effects (one or more) is specified by leaving out the corresponding $H$ matrices from the sum.

The model becomes

$$\mathrm{sp}(\mathbb{K}(H_{\mathcal{J}} - H_*)),$$

where $H_*$ is the sum of the $H$ matrices for the effects to be tested.

The same model can be specified by concatenating the $H$ matrices instead of summing them.

Then to test a set $* = \{\boldsymbol{j}_{*1}, \ldots\}$ of effects,
omit the corresponding sets of columns from
$\mathbb{K}(H_{\boldsymbol{j}_1}, \ldots, H_{\boldsymbol{j}_t})$.

Other ways to define columns for factor effects.

Let $M_r$ be a matrix such that $\mathrm{sp}(\mathbf{1}_r, M_r) = \Re^r$.

For two factors, let

$$
\begin{aligned}
B_{00} &= \mathbf{1}_a \otimes \mathbf{1}_b \\
B_{10} &= M_a \otimes \mathbf{1}_b \\
B_{01} &= \mathbf{1}_a \otimes M_b \\
B_{11} &= M_a \otimes M_b.
\end{aligned}
$$

The model (1), A, B is then

$$\boldsymbol{\eta} = (B_{00}, B_{10}, B_{01})\boldsymbol{\theta}$$

because $\mathrm{sp}(B_{00}, B_{10}, B_{01}) = \mathrm{sp}(H_{00} + H_{10} + H_{01})$.

Dummy variable models can be built with $B_j$ matrices.

## Alternative definitions of $M_a$, $a = 4$

| $\ell$ | $M_a = \mathrm{I}_a$ | | | | $M_a = aS_a$ | | | | "Effect" | | | Orth. Poly. | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 1 | 0 | 0 | 0 | 3 | -1 | -1 | -1 | 1 | 0 | 0 | -3 | 1 | -1 |
| 2 | 0 | 1 | 0 | 0 | -1 | 3 | -1 | -1 | 0 | 1 | 0 | -1 | -1 | 3 |
| 3 | 0 | 0 | 1 | 0 | -1 | -1 | 3 | -1 | 0 | 0 | 1 | 1 | -1 | -3 |
| 4 | 0 | 0 | 0 | 1 | -1 | -1 | -1 | 3 | -1 | -1 | -1 | 3 | 1 | 1 |

Keep any $a - 1$ columns for full rank.

Not all models formed from such $B$ matrices are FE models.

Example: “Reference coding” omits one column of $\mathrm{I}_a$ in the DV formulation. Omitting $B_{100}$ from the saturated model gives a model that is not a FE model.

A model that includes all “contained” effects is a FE model.

If columns of $M_r$ are *contrasts*

then every model built with $B$ matrices is a FE model

because $\mathrm{sp}(M_r) = \mathrm{sp}(S_r)$,

which implies (after some argument) that $\mathrm{sp}(\boldsymbol{B_j}) = \mathrm{sp}(\boldsymbol{H_j})$.

(drum roll)

Conclusion

In models in which factor effects are included with $B$ matrices built from $M_r$ matrices whose columns are contrasts,

the restricted model for the $\boldsymbol{j}$ effect is obtained by deleting the corresponding columns from the $X$ matrix.

The RMFM SS so obtained is the correct numerator SS for testing the estimable part of the effect in question.