Comments by John Tukey on Multilevel Models

John Tukey

September 16, 1994

Dear Jan, The purpose of these comments is to be helpful, rather than to criticize. It is hard to take much space in saying that matters are generally correct and carefully expounded. It is easy to take lots of space pointing out missed opportunities. I hope you will not take the spaced used critically as a sign of anything except a desire to be helpful and friendly.

1 Kreft-Aiken-Deleeuw "The Effects of Different Forms of Centering in HLM."

A If this document had been started with 1.0 to 1.5 page executive summary that minced no words, the document's impact would have been much greater. (Presumably the points made would include: Centering at the Grand mean does not alter the answer, but helps speed the computation. Centering at the individual level, is hard to find situations where it is desirable. Centering at the school level, together with a separate regression "among schools" can have a useful role. (need to explain what it is) Choice of algorithms and software needs to be done carefully.)

B Choice of center need not be confined to 0 or the grand mean. What we need the intercept to do for us includes:

- ease computation
- have a useful interpretation for its value
- give us at least information that the regression coefficients do not give us.

Centering at zero does none of these, as the variance contributed to the intercept by crude regression form is usually at least 3 to 10 times that contributed by the (grand mean) intercept. Centering at the grand mean accounts for perhaps 2.5 of those 3. The interpretation within study is clear, but comparing studies with different $\bar{x}$ —* follows. There is something to be said to for centering at the nominal standard values, occasionally near the grand means of most studies. A little information from regression terms is pumped into this standard intercept , but comparability is improved.

C Giving coefficients with z-values (presumably what I would call z^*-values) is an invitation to users to confuse measurement and testing. F-value of 4 or 6 indicates that the

direction of the effect has been tested and well-established. It does not mean, however, that the effect has been well measured–the confidence limits would be 1/2 to 3/2 the point estimate for 4 and 2/3 to 4/3 of 6. Uncertainties by ratios as large as 3, or even 2, is very precise measurement. If we look at the coefficients of the x-expresssions in the table immediately preceding formula (4.1), we find (assuming many degrees of freedom) the following confidence intervals:

- .69 to 1.33
- .69 to 1.33
- .70 to 1.39

Making quite clear the strength of estimated, the looseness of our measurement and the substantial equivalence of the three analysis. (Indeed .7 to 1.33 tells the story.) The variance components are like 1.08 to — and .09 to 25* quite loose indeed, while the covariance components are completely fuzzy.

D The distinction CWC1 and CWC2 is carefully described just above table, and then becomes a will-o-the-wisp, occasionally appearing but omitted in many places where it would clearly matter. (beginning with the two bottom cells in Table 1).

E Is there ever any excuse for not having page numbers?

F An important point that is not emphasized–though perhaps recognized –is that changing the other carriers in a regression changes both the definition and the meaning of a regression coefficient! And that guidance about choices is dominated by what question is being addressed, what answer is need. Improving statistical estimation has to be secondary. Thus, in the 8th line of abstract CGM only "improves parameter estimation" by changing the definition of the estimated intercept for the better, while CWC1 changes the definition of the regression parameters (possibly for the worse). It is the definitions–and the uses of the estimates–of parameters that matter most.

G The last sentence in the abstract, picks up from the top of p.18 (the last text page). I believe this whole argument to be unsound. Whether we end up with β's at two levels or not should not be resolved by theory, but rather by empirical fact. If the two are estimated and found clearly different, then we should work with two. So we ought to begin the analysis with β's at two levels, and then ask "is it reasonable to replace two by one?" (At this point I wonder whether the rule of two applies here (Tukey et al, Foundations of Exploratory ANOVA, 1991) as it presumably should.)

2 Lesser Issues on the Same

1-1-11 (page 1, par 1, line 1) "different realizations of the same random variable." This may be conventional, but seems to me to be too narrow.

1-2 Centering around the grand mean avoids dangers from users thanks but ————————

3-2-3 Last word "grant" ⟶ "grand"

3-3-1up Does this mean by income was reasonably symmetrical?

4-2-3 Formula's ⟶ Formulas

4-1up-1-to-3 Very unclear about units. If we are going to take numbers seriously, they need definitions.

6-2up-2up Since rescaling changes b1 and b2 in the usual sense, presumably what is meant here is re-centering.

7-1-6 last word, varying ⟶ different

7-2up-5up "equivalent models will produce" Is this anything other than the incomplete iteration.

9-1-6 Words missing between "intercept" and "estimate"

10-1up-2up CWC1 ⟶ CWC2

13-last Something is unclear–or sour–given the use of the words "contextual models"

14-1-dn To say that the parameters are the "same" is at best misleading. They may be represented by the same letter, but both values and meanings are different.

14-2-5 Same estimate ⟶ same parameter

15-2-5 "Hence" The conclusion may be true, but the role of orthogonality in proving it is far from clear to me!

16-2-1 Implicated ⟶ implicit, the best ⟶ "The"

16-1up-3up* "is called for" for what reason? Computationally? In principle?

17-2-2up CWC ⟶ CWC1

17-2-last "is not meaningful" Meaningful does not seem to be the proper word. "helpful?"

17-3 Dichotomization can only be considered to be sloppy. If there must be only two, the two end thirds are stronger than the two halves!

18-1-end If, like all predictors, X is measured with error (perhaps because it is a surrogate), then we expect different β's since "error in x" will contribute different fractions to within and between. (Gender is almost rarely a surrogate).

18-5up "theory development" Which comes first "empirical fact" or "theory development"?

Hope these comments have been helpful. Regards, John Tukey.

3 De Leeuw and Kreft, "Questioning Multilevel Models".

Dear Jan, Comments installment 2:

H It is important to emphasize confidence intervals for the quantities that are most responsive to the user's needs. Sooner or later, then, the user needs to be introduced to jacknifing, which can be applied to almost anything. Standard errors for covariances, for example, are better than nothing, but only just barely. Presumably, very few users would want to use them directly, yet if what is really wanted goes entirely in the direction of correlations or in the direction of combinations of regression coefficients, standard errors (equivalently variances of estimate) of disturbances are inadequate, since covariances of estimates will also matter. While this paper is not the place – some place needs to be found for jacknife techniques.

I All x variables are measured with error, either because of overt error in the measurement process or because the x-variable we got our hands on is a surrogate for the ξ-variable we wish we had, and so measures that variable with error. As a consequence there is no apriori reason why β's at two levels should be the same. Quantitative inquiry may lead us to fit a single value for the two, but only because this is justifiable* because of the size of – and degrees of freedom underlying – the error with which the parameter is estimated. (It is wrong to try to estimate every coefficient which we know is there if the estimates are too small. The rule of two (op. Hoaglin, Mosteller, and Tukey, 1991, Fundamentals of Exploratory Analysis of Variance) can usually tell us – rather well – which parameters we do best to keep on estimating.

J If we are to proceed as (I) suggests, we should begin by fitting lots of parameters and then dropping out some and fitting again. The empirical look should precede both the final analysis and consideration of theory.

K The user should understand that it is better to fit what is really wanted, perhaps crudely, perhaps carefully with large errors, than to fit something else that can be estimated with greater precision. It would not be easy to indoctrinate this into many users, but, in the long run, it is essential that this be done. This means more effort on the interpretation of what is estimated by various procedures.

L Diagnostics are more important than most of the fine points discussed in the paper. Again not easy to interpret, again important.

M 100% efficiency is a dangerously misleading target, because actual efficiency can depend greatly on such things as distribution shape and because 100% efficiency under narrow unsupportable assumptions often leads to seriously degraded performance under apparently quite similar circumstances.

N A few % less in efficiency almost never matters. We do not want to distinguish inefficient from the efficient – with ever a slight broadening of assumptions, nothing is efficient! (Asymptotic efficiency might – or might not be – a snare and a delusion.)

O The first approximations to the behavior of an estimating procedure is the pattern of weights in the linearized form of the final iteration. If $A \leq \frac{weightused}{optimalweight} \leq 2A$ (all bearers of weight) Then the weights used must provide $\geq 89\%$ efficiency and ordinarily provide $\geq 97\%$ efficiency. (De minimis non curd lex).

P As noted for the other paper, centercepts are much more interpretable than at zero intercepts. (I believe you got at least part way here.) Using some standard value near the grand mean can make comparisons across studies easier and more effective.

R Models do not analyze data. Models suitably prayed over produce procedures that may be able to analyze data, as you point out (6-2-5). There is no need to worry about the model being true since obviously is not. Isn't the appropriate conclusion that models, far from being matters of truth, are just handles to lead, rather naturally, to procedures which would analyze ideal data rather satisfactorily, but whose performance on real data we are likely to need to study.

S The mistreatment of intercept in the present draft can hardly not make reading the paper almost impossible at the moment! The "intercept" appears at 9-2-2, 9-2-5, 10-1-8 (1 below). Presumably, there will be intercepts in almost every fit suggested by what goes on in this paper. But no single formula has an intercept in it! No reader is warned of this! No excuse is offered for the consequent confusion!

S Again there is a real need for an executive summary.

T The discussion of economists and Klein centering at the top of page 7 is unsound. If we could "have observations on a sufficiently large number of variables", putting all these variables in a linear regression would undoubtably only provide the bath water to accompany the baby out the window. For collinearity would be great enough to describe both x's and y's completely, leaving no residuals (where simple regression defines β). Moreover it is relatively certain that if all these variables describe all behavior it would not do this linearly.

The only reason I can see for dragging in this purple herring is to excuse and admission of correctness for what I routinely call Axiom 1 "People are different!" My own view would be that the only behavior that requires excuse is failure to accept this axiom.

Fortunately this argument seems to have nothing to do with the main thrust of the paper.

At 19-2 the authors call for parsimony and stability. How can they take seriously at 7-1 an indefinite proliferation of predictors.

4 *lesser issues in same*

2-3-5 technique $\longrightarrow$ techniques

3-1-2up Append "or avoiding" at end of line

3-1up-3up A thoughtful document about how to select and code variables would be very valuable.

4-NCES1-2 Here is that non-specific word "independent". Here I presume it means something "treated as a circumstance, not a response".

lines below —* individual ⟶ individuals

4-6th line up on ⟶ of

5-1-top At some other time some discussion is needed.

5-1-9 specific ⟶ specified. I don't believe that we need to specify so narrowly.

ne just below(3) "independent" This is only true if the distributions are specified to be Gaussian ("normal") which is nowhere said to be the case. (For other specifications, "uncorrelated" need not be "independent".)

5-1-2up "or to anything else" Certainly this can only refer to things in the model. And there is nothing else that y's, x's, and σ's in this model. So?? (To quibble further ε_i is not independent of either ε_i^2 or ε_j^3)

6-2-3 "Unlikely to be approximately true" Glad to see the truth recognized!

6-2-4 For "as" read "to generate"

6-2-6 For "it does its" read "the procedures it generates do their"

6-2-7 Insert "the procedures usually generated by" before "the usual* regression".

7-1-3up Insert ", the optimists say", between "should" and "use"

7-1-2up Insert "need to" before "vary"*

7-1-1up Is this "centered" equation likely to be linear?

7-8up I believe it is clearer to say "we do not expect the individual labelled ... to be the same individual in different replications".

8-1-5 Insert "conceivably" between "It" and "makes".

9-2-5 How is "p" here related to "m" on page 8? Or "p" on page 4

9-2-2up Insert "efficiently" before "fitted" (or was something else meant?)

9-(4.1) I don't think this section makes its point.Surely –* for the —*.

11-1-3 "More plausible" not because of the Klein argument but because we have often seen data sets which may fit better. (At least I hope we have!)

11-1up-2up "models" ⟶ "model"

12-1 Append "It is even more important in helping to define better what is to be estimated!"

12-2-4 Restrcited ⟶ restricted

12-1up-1up Delete initial "n".

14-1up-3up Being "not BLUE" is unimportant. Loosing a few % of efficiency is.

5-line below last Insert "(unknown)" before "dispersion"

15-2-7 Insert "asymptotically" begore "fully efficient".

15-2-1up Insert "occassioned" before "negative".

16-NCE* "interest in overall measures" Why? How might the answers be used? How are the answers actually used? These are key questions.

16-(6.1)-2 Note the camel, here "multinormal" has put its nose in the soup again.

line below (83) Append "if we adopt (28)". ((28) isn't part of the model, but is an essential for (34).)

17-2 A few more sentences could change one situation a lot. I believe it to be roughly thus: $\hat{\sigma}_i^2$ is defined in terms of the residuals (why not r_i) of any OLS fit. σ_j^2 is defined in terms of the fit we are studying (why not $\hat{\hat{\sigma}}_i^2$ or something). And so on. (Am I right?)

17-3-1 Why are these useful? How are they useful? Why are no others useful? Even two or three sentences would help!

17-1up-3up deviation ⟶ deviance

17-1up-2up Are the "alternative" definitions more useful or less useful? There is need for a numerical illustration and for what measures are responsive to what detailed questions.

18-(6.2)-end The key question not addressed is – are the differences in efficiency large enough to warrant serious theoretical research.

19-1-3 up ff Has anybody made studies as to what fraction of naive shrinkage actually occurs in such circumstances. Surely a model in which schools differ in level for other reasons as well as for the sampling of students make sense.

19-2-5ff I agree the time has come – and applaud the statement. I would, however, prefer to say "realistic" instead of "critical".

19-2-5 up ff To get the best gains from parsimony one needs to begin by fitting more, and then finding out what to keep.

I suspect the key log in the log jam is that we are supposed to estimate the stochastic part of the "model". This is something I deny vigorously. I believe the approach taken (in similar situations) infinite alternative robustness is the right approach in general, namely to (i) think of different stochastic models as parallel challenges, (ii) consider

procedures to maximize the worst of these performances. If we look at things this way, flexibility in the stochastic model need not imply lack of parsimony in estimation. We just don't estimate these "additional parameters".

20-2-2up "deserve some additional study" I am left wondering whether or not they deserve "immediate use". What do the authors really think?

20-1up Haven't the numerical analysts dealt with this problem adequately? I am not up to date, but I would think that doubling step-length until the estimation goes the other could be used to cure matters.

21-3-6 Insert "or beyond" after "ML/3"

21-3-display If XLISP is here, why not SPLUS? (Shouldn't both go in parentheses?)

22-2-2* By analogy with other fields, one might expect a substantial effect on estimated variability of estimates. Does this happen? If it does dare we not use random coefficients?

22-3-2 "reliably" This word needs a much fuller explanation. Is it something like "OLS with incorrect weights is unbiased, consistent, and probably of higher efficiency than* most* would infer*"?

5 *back to first paper discussed*

3-Table1 Putting the formula numbers for the various models in the boxes would be a great help.

8-top Making the covariance zero from the first line calls for $\bar{x} = 2.84$ and $\hat{\beta} = 2.86$ in true agreement with the first two fits. Making the covariance zero in the 2nd/3rd lines calls for $\bar{x}$ of 3.3 to 3.4. Is this rounding error or should something other than the actual x be used.

Very sincerely yours, John Tukey.