Exegeses on Linear Models

Paper presented to the S-PLUS User's Conference
Washington, DC, 8-9th October, 1998

W N Venables
The University of Adelaide

May 13, 2000

Contents

1 Preliminaries and introduction **1**

2 A view of regression models **2**

2.1 Extending the model 3

2.2 Additive models 4

3 Some simple examples **5**

3.1 The Janka hardness data 5

3.2 An additive model that seems to work 9

4 Variance heterogeneity **9**

5 The SAS factor **11**

5.1 'Type III' sums of squares 12

5.2 An example: The rat genotype data 14

6 Enhancements **17**

7 Mixed effects, multistratum models and variance components **18**

7.1 An example: the petroleum data of Nilon H Prater 20

A Data sets used **22**

1 Preliminaries and introduction

An *exegesis* is a marginal note or footnote expanding on the text, particularly in ecclesiastical contexts such as in canon law. Some non-ecclesiastical exegeses are quite famous, such as the one that says "I have found a truly remarkable proof of this assertion but there is not sufficient space to write it down here". Others deserve to be, such as the pencil annotation I once discovered in a *very* old book on the English monarchy in my college library. Beside the paragraph describing how Charles II departed this world and was buried it continued the final sentence with "…and whose epitaph might have been 'Here lies a profligate, licentious, perfidious, lascivious old scoundrel!'"

There is an element of whimsy in the title, of course, with an oblique suggestion that perhaps people take linear models a little too seriously these days. On the other hand, as Brian Ripley and I say in our joint book, linear models form the core of modern applied statistical practice, so they are indeed rather important. In this talk I would like to present a view of linear models that is in many ways the folklore of the subject, certainly for some of the more experienced (probably meaning simply 'older') practitioners but unfortunately is not presented often enough or clearly enough in the current crop of textbooks or, more particularly, in the current slew of wordy software manuals.

In so doing I hope to make some side comments—exegeses—that I hope provoke people into a minor, largely personal re-examination of linear models and, more importantly for us here at this conference, of the software used to support the praxis of linear models. If they prove controversial this perhaps will serve to get your minds into gear and yourselves engaged in fruitful, or perhaps heated dialogue, as befits one of these rare gatherings where so many of us who know each other only at the ends of an email link finally get to see one another face to face - and slug it out.

2 A view of regression models

Let Y be a *quantitative* response variable, x_1, x_2, ..., x_p potential *quantitative* explanatory variables and Z a standard normal variate. A model that says almost nothing useful about the situation, but which nevertheless cannot seriously be challenged is that

$$Y = f(\boldsymbol{x}, Z) \tag{1}$$

where $f(\cdot)$ is some unknown, hopefully tame function.

Using a standard normal to inject error into the dependency is not without serious loss of generality since any continuous distribution can be obtained by transformation from it, that is

$$E = F^{-1}\{\Phi(Z)\}$$

is a random variable with distribution function $F(\cdot)$, assuming continuity.

Assuming a steady-as-she-goes context, we might try to replace (1) by a local approximation we recognize will only apply in the neighbourhood of some point $\boldsymbol{x}_0$ in design space. There are many ways we might do this, but surely the first and most obvious one, at least in this slowly varying universe we are assuming, is a first order Taylor series:

$$Y \approx f(\boldsymbol{x}_0, 0) + \sum_{i=1}^{p} f^{(i)}(\boldsymbol{x}_0, 0)(x_i - x_{i0}) + f^{(p+i)}(\boldsymbol{x}_0, 0)Z$$

or, in more conventional notation

$$Y \approx \beta_0 + \sum_{i=1}^{p} \beta_i(x_i - x_{i0}) + \sigma Z \tag{2}$$

It does seem to be the case that most multiple regression models are really only intended to be local linear approximations to something more complex, and hence will have a rather closely proscribed domain of validity.

There is a tendency to simplify (2) even further and bundle all constant parts with respect to covariates into the intercept term and write it as

$$Y = \beta_0^\star + \sum_{i=1}^{p} \beta_i x_i + \sigma Z$$

which may well be formally the same model, but I argue that this is not a practice to be recommended. for several reasons. The first of these is that if the covariates x_i are all, say, positive but in any case reasonably far from 0, the constant term is the value of the mean at a point $\boldsymbol{x} = \mathbf{0}$ likely to be far outside the domain in which the model is intended to apply, and hence likely to mislead naive clients as we shall shortly see in an example.

There may also be good numerical reasons for anchoring the $x-$variables at an origin somewhere near the centre of the section of design space covered by observations, but I do not wish to push this point too far.

2.1 Extending the model

Following the same simple Taylor series heuristic, if we wished to improve the linear approximation, and hopefully extend the practical domain of validity, we might next consider two terms in the Taylor series expansion. This gives an approximating model of the form

$$\begin{aligned} Y &= \beta_0 + \sum_{i=1}^{p} \beta_i (x_i - x_{i0}) + \sum_{i=1}^{p} \sum_{j=i}^{p} \beta_{ij} (x_i - x_{i0})(x_j - x_{j0}) + \\ &\quad \left\{ \sigma + \sum_{i=1}^{p} \gamma_i (x_i - x_{i0}) \right\} Z + \delta Z^2 \end{aligned} \tag{3}$$

Of course there is no guarantee that the heuristic remains particularly cogent and at some point practical considerations and experience with the context have to take over, but it is worth considering what it is this simple idea is telling us to look for in extending the linear model, namely

- curvature in the main effects (quadratic terms in one variable),
- linear $\times$ linear interactions (cross product terms in two variables),
- variance heterogeneity (terms in $(x_i - x_{i0})Z$) and
- skewness (the term in Z^2),

which are all very commonly encountered problems in real life linear modelling, often ignored but sometimes modelled in a more structured and context dependent way.

2.2 Additive models

The so-called *additive models* of (Hastie and Tibshirani, 1990) extend the first order model in a different way, namely to a form

$$Y = \beta_0 + \sum_{i=1}^{p} g_i(x_i) + \sigma Z$$

where $g_i(x)$ may be a term consuming several degrees-of-freedom, or even a smoothing spline or locally weighted regression function that has to be estimated by non-standard methods, but typically (though not always) it involves only one covariate.

My first reaction when I saw this was that it makes the strong assumption of no interactions between variables, but relatively weak assumptions about the form of the main effects. I can remember expressing this objection to Trevor Hastie at a conference in 1993. This was clearly not the first time this question had been raised since he was well and truly ready for me and had many reasons why such a model was useful in practice. The reasons included that such a modelling strategy isolated the effect of each variable and allowed the user to examine them separately to see which were important and which were not.

Privately I remained not completely convinced, I must say. As a consultant I am very familiar with the kind of user who simply cannot think in a way that admits interactions but who just wants to know "which of my variables are important, which are not and what are the important ones doing". (This is very closely akin to the experimental design philosophy that varies one variable, only, at a time—the sort of thing that factorial design principles should have put out of business for good but, sadly, has not, yet.) But GAMs were receiving such accolades at the time I did rather feel like the little boy trying to suggest that the king had no clothes.

More recently, however, I was at an experimental design lecture in Oxford where the lecturer expressed precisely the same reservation about additive models as I had done some years before, which made me feel rather more confident about the king's state of dress. (The speaker, by the way, was Sir David Cox.)

To put the contrary case, there are indeed practical situations where the possibility of important interactions can safely be downgraded *à priori* and additive models can prove a very useful investigative or diagnostic device. Indeed I use them often this way myself and encourage my students to do the same. But I do seriously suggest that the question of interactions must in some way—however informally or by appeal to context—be investigated and laid to rest first, since rather simple interactive models can sometimes be approximated fairly well by complicated main-effect models, but these models do not hold up well in prediction and can be misleading if used for interpretation.

3 Some simple examples

3.1 The Janka hardness data

This is a favourite data set of mine with an Australian flavour to it, indeed an Australian timber flavour, which was the (now politically rather unfashionable) business my late father worked in all his life, and me off and on for most of my adolescence.

The dependent variable is the Janka hardness of timber samples, a structurally important quantity known to be closely related to density, which is the only recorded determining variable. The data comes from an old but still fascinating textbook, ((Williams, 1959)), by E J Williams, who worked in the CSIRO Division of Forest Products before joining Melbourne University. The full data set is given in Table 1 on page 23.

Williams simply used the data as an example where a quadratic regression was necessary rather than simple linear regression. It was before the days when plots were commonly shown in textbooks but fortunately complete data sets were. Had Williams done the plot I think he would have noticed some interesting features that these days would attract much greater attention.

A plot of the data is shown in Figure 1. Some of the general features I have suggested we should anticipate happening with approximating linear models are almost apparent even from this rather simple plot. The curvature in the main effect is more-or-less clear (which is all that Williams was concerned with), but so is the variance heterogeneity, in this case an increase in variance as the density increases.

The quadratic regression can easily be shown to be an adequate degree of polynomial regression but a plot of residuals versus fitted values makes the variance heterogeneity and skewness starkly apparent, as shown in Figure 2

There is a plausible argument that much hangs on the value leading to the largest residual, but even omitting this point the variance heterogeneity can be detected by standard devices. Moreover some standard checks for outlying residuals fail to rule it out of court even as it stands.

Box and Cox transformations are interesting with this data set, as it turns out you need something like a square root transformation to make the quadratic regression linear, but something close to a log transformation to make the variance stable and remove the skewness credibly.

With this cue John Nelder once suggested to me using a generalized linear model with a square root link to make the linear predictor linear in density rather than quadratic, and a variance function proportional to the square of the mean:

$$\mu = (\beta_0 + \beta_1 x)^2, \qquad \text{Var}[Y] \propto \mu^2$$

which implies that the original observations have a gamma distribution. This model works remarkably well, giving deviance residuals uniform beyond suspicion and a fairly

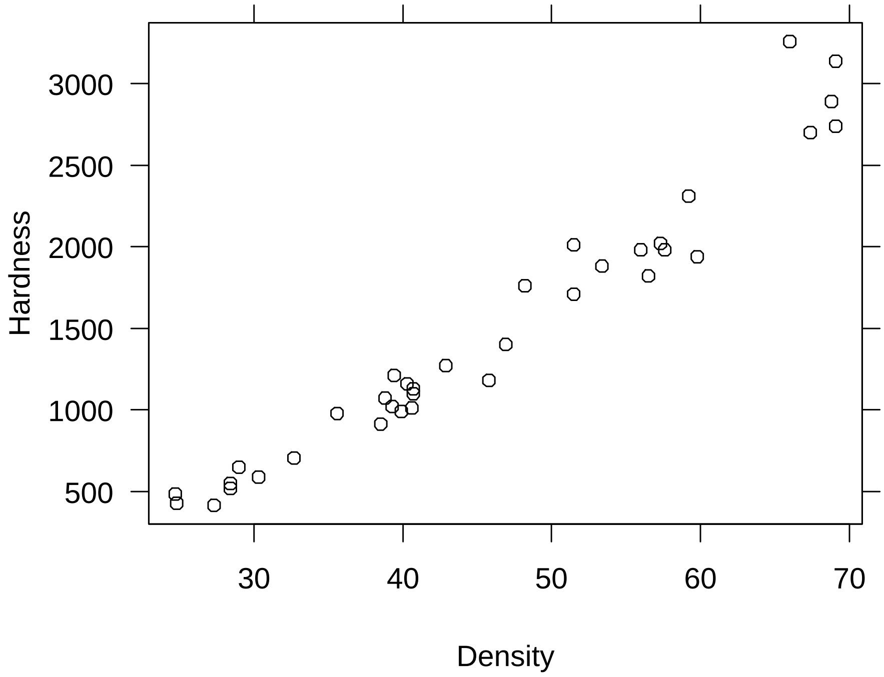

Figure 1: The Janka hardness of timber data. Source: Williams (1959)

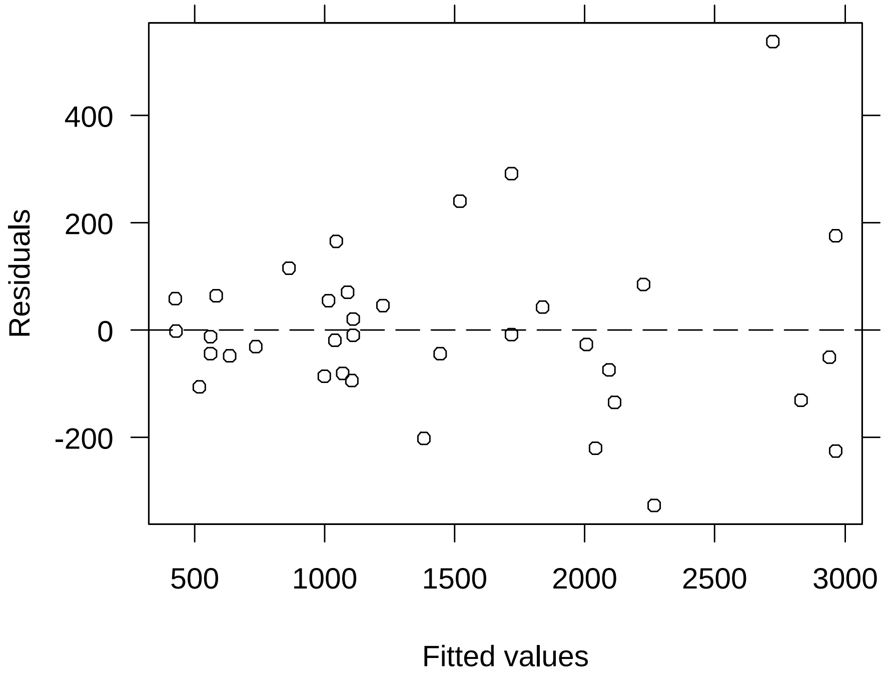

Figure 2: Residuals versus fitted values after a quadratic regression of Hardness on Density in the Janka data.

simple expression for hardness in terms of density. Two features should be acknowledged, though

- The predictions from the gamma model are virtually identical to those obtained by Williams using an ordinary quadratic regression. Where the more sophisticated model does offer some advantages, though, is in assessing the errors of prediction. The less dense timbers have a much tighter error bound with the gamma model than with the normal, and the heavier timbers a wider one, as seems to be appropriate.
- The real loss of degrees of freedom for estimation is probably slightly larger than it appears, considering the amount of data data snooping that went into the modelling proposal, but that could be checked by some of the standard means, such as cross-validation, bootstrap validation or using a confirmatory sample if available.

One final important point will be made with this simple example. Look at the differences between fitting the polynomial models, here taken up to degree 3 for illustration, in a zero centred and a median centred form.

```
> jank.1 <- lm(Hardness ~ Density + Density^2, Janka)

> round(summary(jank.1)$coefficients, 2)
                Value Std. Error t value Pr(>|t|)
 (Intercept) -118.01     334.97   -0.35     0.73
     Density    9.43      14.94    0.63     0.53
I(Density^2)    0.51       0.16    3.25     0.00

> jank.1a <- update(jank.1, . ~ . + Density^3)
> round(summary(jank.1a)$coefficients, 2)
                Value Std. Error t value Pr(>|t|)
 (Intercept) -641.44    1235.65   -0.52     0.61
     Density   46.86      86.30    0.54     0.59
I(Density^2)   -0.33       1.91   -0.17     0.86
I(Density^3)    0.01       0.01    0.44     0.66
```

Figure 3: Fitting quadratic and cubic models for the Janka data in zero centred form

Notice how the coefficients change wildly in going from the second to the third degree curve. This is sharply in contrast to the case where density is measured from its median value.

Although this is not even an approximately orthogonal model specification, the change in coefficients is small, with sensible signs intact.

```
> Janka$density <- Janka$Density - median(Janka$Density)

> jank.2 <- lm(Hardness ~ density + density^2, Janka)
> round(summary(jank.2)$coefficients, 2)
               Value Std. Error t value Pr(>|t|)
 (Intercept) 1165.82      36.85   31.63        0
     density   51.99       2.63   19.74        0
I(density^2)    0.51       0.16    3.25        0

> jank.2a <- update(jank.2, . ~ . + density^3)
> round(summary(jank.2a)$coefficients, 2)
               Value Std. Error t value Pr(>|t|)
 (Intercept) 1174.02      41.70   28.15     0.00
     density   50.41       4.47   11.27     0.00
I(density^2)    0.42       0.26    1.58     0.13
I(density^3)    0.01       0.01    0.44     0.66
```

Figure 4: Fitting quadratic and cubic models for the Janka data in median centred form

More importantly in some respects, with the zero centred form the intercept and linear terms are *not significant!* For some users this is a powerful reason to remove them and recover the degrees of freedom for error. Stepwise regression programs also find it difficult to avoid removing such variables from the model out of sequence, which is just one of the reasons that they can be so dangerous.

The proponents of removing the intercept and slope terms from the zero centred form argue that regressions through the origin are often quite sensible models and 'zero density' timbers in this context could be expected to have 'zero hardness', although just why the curve should also be flat at the origin is somewhat more problematical.

I would argue very strongly, though, that these removals are not justified in this case, (and with the median centred form the question does not even arise). This is only ever intended to be a *local* model with limited scope for extrapolation, so arguing what the behaviour of the curve should be for densities near zero is evidently to argue about what happens at a point well beyond the intended scope of the model.

More fundamentally, though, one could argue that if there is a natural group of transformations under which the problem is invariant, the model building process should also be invariant under the same group of transformations. In the present case it is reasonable to require that the model building process be invariant with respect to changes of location and scale in the density. This immediately requires that as long as there is a term in the polynomial model of degree k, no term of degree less than k should be considered for exclusion.

Under these conditions we say that the intercept and linear terms are *marginal* to the quadratic term, a concept guaranteed to generate controversy like no other in this area.

Marginality is also at the crux of another thorny issue in linear models, namely that of Type III sums of squares, to which I shall return below.

3.2 An additive model that seems to work

To show my even-handedness on the issue, I present below a textbook example where additive models do seem to offer a useful perspective on the data, but I have concealed my attempts to resolve the question of interactions beforehand, which to my satisfaction led me to assume that in this case they could safely be ignored.

The data comes from (Draper and Smith, 1981), but for convenience is reproduced here in Table 4 on page 24. The response is the annual average wheat yield for the state of Iowa from 1930 through 1962 and the potential covariates are the state average rainfalls and temperatures, each for 4 months of the growing season and the year itself as a surrogate for incremental varietal improvements. The only variables that seem to have any potential explanatory capacity are `Rain0` (pre-season precipitation), `Rain2` (mid-season precipitation), `Temp4` (temperature at harvest time) and `Year`. Fitting a smoothing spline term in each of these four variables gives an interesting and speculatively interpretable picture, as shown in 5. The apparent halt to varietal improvements during and immediately after the second world war is about the only thing pretty convincingly established, and credible, but not something easy to pick up any other way that I know of.

4 Variance heterogeneity

It is manifestly *not* the case, of course, that all linear models are only approximate and local in scope. The egregious exceptions that come to mind are the linear models that arise from designed experiments such as variety trials in agriculture or clinical trials in medicine. We do seem to have a peculiar attitude towards variance heterogeneity that shows itself most clearly in what we provide for about the simplest such case, namely the two sample $t-$test.

The S-PLUS function, `t.test`, following a long and venerable tradition, allows users to test equality of means either (a) assuming equality of underlying variances or (b) not making such an assumption (which will always be false in practice, of course). Option (b), (which thankfully is *not* the default in S-PLUS, though it is very often the default

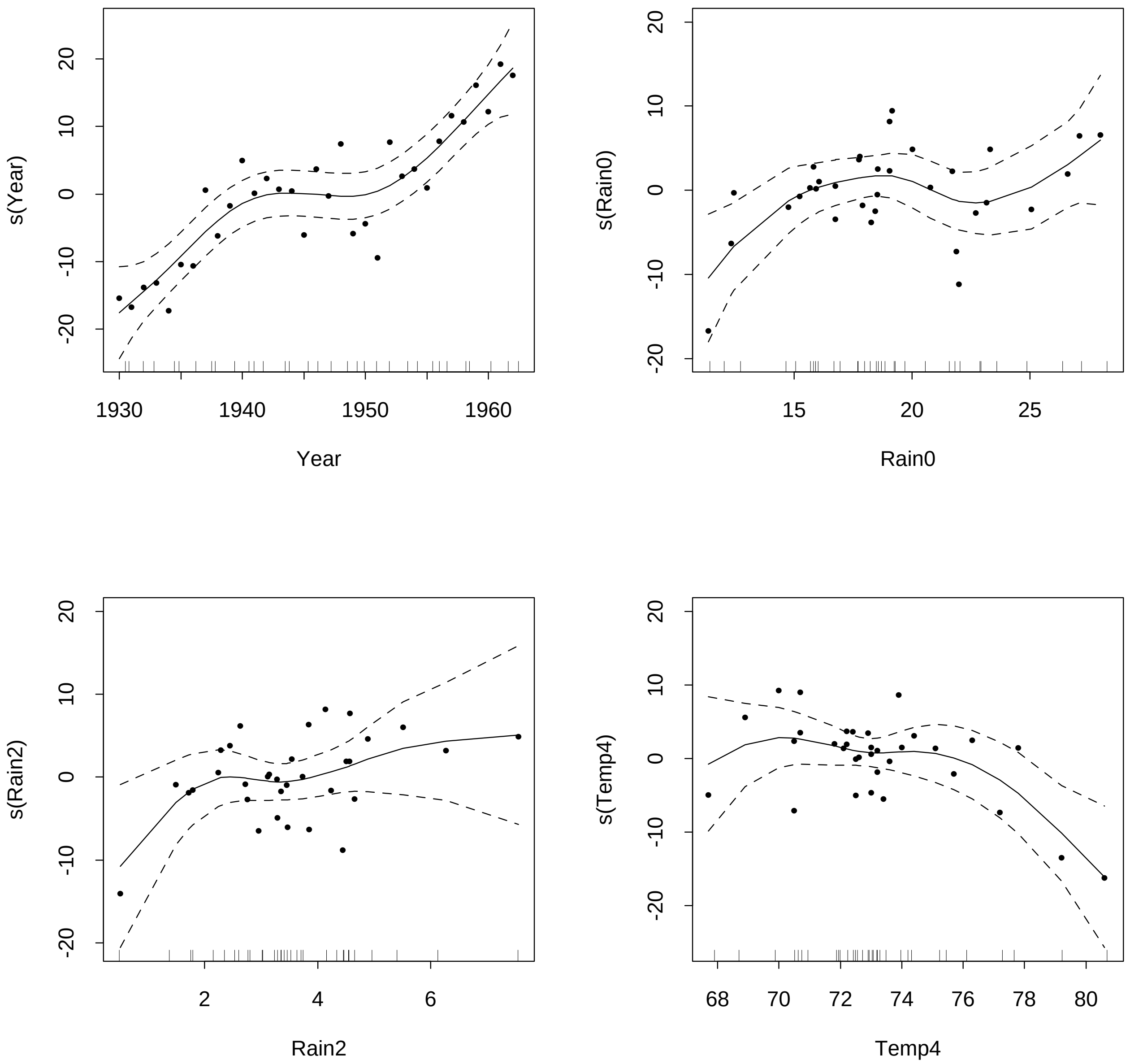

Figure 5: Additive model components for the Iowa wheat yield data.

elsewhere) gives naive users a cosy feeling of protection that perhaps their test makes sense even if the variances happen to come out wildly different.

What's the matter with that?

For one thing, not so naive users wonder why such a fuss is made of equal variance in the two sample $t-$test but elsewhere it is a non-problem, it seems. In the three sample case the only software provided is the one-way analysis of variance where equality of variance is, it seems, silently dismissed as a non-issue.

The main problem I see with it, though, and I see it often, is that it leads users away from the important considerations of the problem. Variance heterogeneity is an important problem in all linear models and it can be the *vital* issue. If two samples show large and established differences of variance, this is usually much *more* important than any difference of means. If one treatment gives a higher yield than another, this is often not of much interest if it also has a much higher error variance causing it to be unreliable.

In a sense, the onus is on the user to investigate variance heterogeneity along with investigating mean differences. This is often done by diagnostic methods such as we did in the Janka data. Providing things like the modified $t-$test that 'can be used even if the assumption of equal variances is not met' is only encouraging naive users to miss the whole point. To be candid I am not sure if this problem can really be fixed by changing the software—in the long run users are (often) adults and must be trusted not to do silly things—but as an educator and a user I feel obliged to voice my concern at what I see happening.

5 The SAS factor

Many more S-PLUS users than I had expected would be the case are in fact refugees from the SAS regime. Even some of the authors of the voluminous SAS documentation are now well-known names in the S-PLUS world and some, though very few it seems, continue to play a fairly prominent role in both camps.

I have involuntarily used SPSS many years ago but I have never seriously used SAS. To prepare this talk I thought I would have to look closely at the SAS documentation, at least, to see how other software systems treat linear models these days, and since SAS seems to be to statistical computing what Microsoft is to personal computing, it had to be the logical next place to look. The only documentation I had available ((SAS Institute Inc., 1990)) dated from 1990, but I think for my purposes that would be current enough.

It must be said that although it is often very waffly and occasionally rather patronising the documentation for SAS is at least comprehensive and very often statistically astute (if I might be permitted a similarly patronising rejoinder!) My first and strongest impression, though, is that SAS attempts to provide all functionality for all occasions—programming is such a delicate and exacting job there must be no scope for ordinary users to do it for themselves.

The goal of providing all necessary statistical and data analysis functionality in the one awkward and foolishly consistent framework is uncannily like Microsoft and, (one hopes and prays), ultimately futile, but in the meantime what happens is that the program *defines* the subject rather than the subject dictating what the program should do. SAS, it seems, has become the gold standard, the output of SAS programs the ultimate point of reference for correct and appropriate statistical calculations and the SAS terminology is rapidly taking over as standard terminology. This is very Microsoft-like indeed and very worrying for anyone who cares about the profession.

Nowhere, it seems to me, is this SAS *coup d'état* more evident than in the way Analysis of Variance concepts are handled.

There is no essential distinction between linear models we happen to call *Regression* and those we call *Analysis of Variance* models, but with the latter the feature that makes them somewhat special is that the natural parametrisation leads to design matrices not of full column rank. This in turn gives rise to the important notion of *estimability* of linear functions of the parameters.

In a sense this notion is easy. The means of the observations themselves are the only well-defined functions of the parameters; any function is estimable if and only if it can be written as a linear function of the observation means, that is the coefficient vector must be in the *row space* of the design matrix.

Many programs overcome this minor obstacle by omitting parameters in such a way that those remaining are all estimable. S-PLUS does this when it uses the `contr.treatment` contrast matrices, but for other contrast matrices the original redundant parameters are replaced by a smaller set of more complicated linear functions of them which are, usually, estimable.

The clear and unequivocal message is, though, that a linear function of the parameters is either estimable or it is not; there are no shades of grey, no half way houses and no subtle distinctions. With this in mind, the title of Chapter 9 of the SAS/STAT manual is worrying: *The Four Types of Estimable Functions*. Closer inspection reveals that this notion is really a kind of transferred epithet, and there are really four types of hypotheses being considered, another equally curious distinction being made without a difference.

One sentence bears quoting: "The four types of hypotheses available in GLM may not always be sufficient for a statistician to perform all desired hypothesis tests, but they should suffice for the vast majority of analyses." In other words, the distinctions being made are limitations on the program rather than differences of principle in the subject, but then, it seems, the program really is defining the subject.

5.1 'Type III' sums of squares

I was profoundly disappointed when I saw that S-PLUS 4.5 now provides "Type III" sums of squares as a routine option for the summary method for `aov` objects. I note that it is not yet available for multistratum models, although this has all the hallmarks of an oversight (that is, a bug) rather than common sense seeing the light of day. When the decision was being taken of whether to include this feature, "because the FDA requires it" a few of my colleagues and I were consulted and our reply was unhesitatingly a clear and unequivocal "No", but it seems the FDA and SAS speak louder and we were clearly outvoted.

So what is the problem with Type III sums of squares that is worth making such a fuss about?

It's difficult to resist the same sort of words we use to explain the phenomenon to second year students. In fact I will not resist.

In a two-factor experiment the factors A and B are said to have an *interaction* if the change in the response mean resulting from a change in levels for factor A, say, *depends* on which level of factor B is applied.

If any change in level for factor A always produces the same change in the response mean regardless of the level of factor B, the same must be true for factor B as well, and the two factors are said not to have an interaction, or to act *additively*.

If factors act additively their *main effects* are well defined and testing whether their main effects are zero or not is a sensible and useful thing to do.

If there is an interaction between factors A and B, it is difficult to see why the main effects for either factor can be of any interest, since to know what the effect of changing an $A-$level on the response will be depends on which $B-$level is in force.

More pointedly, when there is an interaction in force, a routine test of the main effects can be shown to be testing an hypothesis that depends on the design of the experiment rather than on the parameters alone. To overcome this manifestly arbitrary aspect of the testing procedure, Type III sums of squares arise from testing an hypothesis—equally arbitrary—connected with the parameters that does not vary with the design.

To hark back to a previous idea, testing main effects in the presence of an interaction is a violation of the *marginality* principle. This is not a totally rigid principle, but in all common practical situations the sensible thing is to respect it. Just as there are situations where it does make sense to fit a regression line through the origin, though, or to constrain a fitted quadratic curve to be flat at some specified point, there are some very special occasions where some clearly defined estimable function of the parameters that would qualify as a definition of main effect to be tested, even when there is an interaction in place, but like the regression through the origin case, such instances are extremely rare and special.

Just as providing a switch in the two-sample $t-$test to cover unequal variances encourages users to think that the annoying problem of variance inhomogeneity is nailed down and can be ignored, providing a deceptive "Type III" sum of squares option in he Analysis of Variance summary encourages users to think that the annoying problem of interactions can be ignored and the main effect question—the one that everyone understands, even though in the presence of interaction it makes little sense—can be safely settled without worrying about pesky interactions, at least to the FDA's satisfaction, and what more could anyone ever want?

There is even a class of user now days who sees the significance stars (which fortunately S-PLUS does not yet provide, for who knows how long?) rather like the gold stars my grandson sometimes gets on his homework. Three solid gold stars on the main effects will do very nicely, thank you, and if there are a few little stars here and there on the interactions, so much the better!

5.2 An example: The rat genotype data

The non-problem that Type III sums of squares tries to solve only arises because it is so simple to do silly things with orthogonal designs. In that case main effect sums of squares are uniquely defined and order independent. Where there is any failure of orthogonality, though, it becomes clear that in testing hypotheses, as with everything else in statistics, it is your responsibility to know clearly what you mean and that the software is faithfully enacting your intentions.

Henri Scheffé in his classic text on *The Analysis of Variance* (Scheffé, 1959), gives an example of a 4×4 double classification with unequal cell frequencies, though not wildly so, that will do to illustrate some of these points. For convenience the data set is reproduced here in Table 2 on page 23.

Litters of rats are separated from their natural mother and given to another female to raise. The two factors are the mother's genotype and the litter's genotype and the response is the average weight gain per member of the litter. (There is an acknowledged component of variance ignored in this, but as Scheffé says, it is likely to be very small.) The natural models and the relations between them form a lattice that immediately makes clear what any sum of squares means. This is shown in Figure 6.

Any path from the top of the lattice to the bottom gives sums of squares for testing each model *within* the one immediately above it. Thus testing for *Litters ignoring Mothers* and for *Litters eliminating Mothers* are both "main effect" sums of squares for Litters, but correspond to different tests of hypotheses, one assuming Mother's genotype does not have an effect, the other allowing for the possibility that it may. For an orthogonal experiment, confusingly in a way, these sums of squares are numerically identical, even if the hypotheses they test are, conceptually at least, different.

The difference here is only small, as we can easily check in Figure 7. Interactions appear ignorable, so main effect tests start to make sense, and it becomes clear that the mother's genotype exerts a sizeable influence on average weight gain and litter's genotype does not. This has to be checked graphically, of course, and when that step is taken it becomes clear that there is at least one large outlier and much hangs on whether that observation is correct or not.

When I use this example with my students, though, the first thing I ask them to do is to fit the full model and use `drop1` to suggest the next step. Brilliantly, `drop1` fingers the interaction term, and the interaction term *only*:

```
> drop1(gen.1)
Single term deletions

Model:
Wt ~ Litter * Mother
               Df Sum of Sq       RSS  F Value    Pr(F)
```

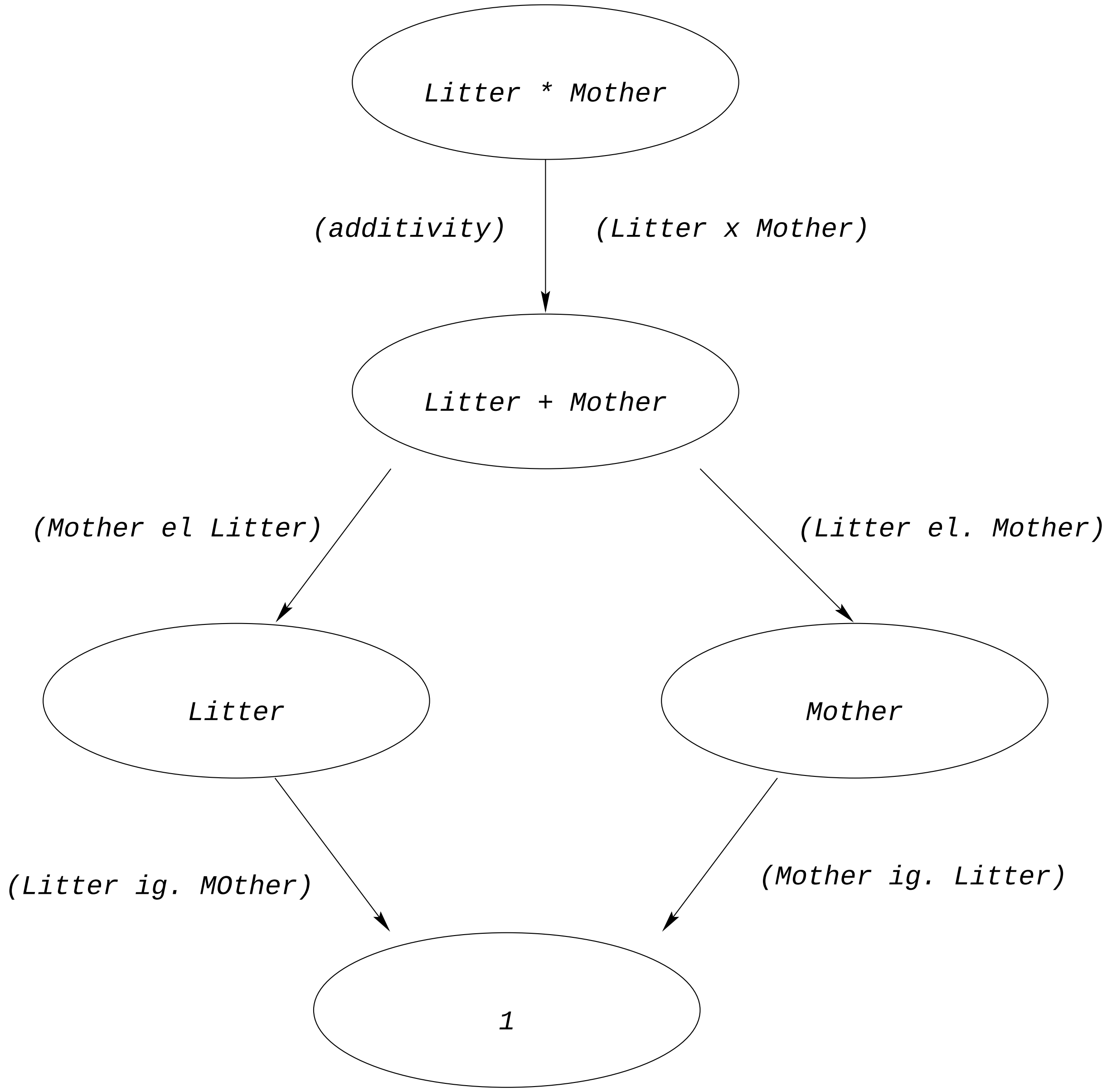

Figure 6: Lattice of models for the rat genotype foster feeding data

```
      <none>              2440.816
Litter:Mother  9  824.0725 3264.889 1.688108 0.120053
```

To my delight I see that marginality constraints between factor terms are by default honoured and students are not led down the logically slippery 'Type III sums of squares' path. We discuss why it is that no main effects are shown, and it makes a useful tutorial point.

The irony is, of course, that Type III sums of squares were available all along if only people understood what they really were and how to get them. If the call to `drop1` contains any

```
> names(genotype)
[1] "Litter" "Mother" "Wt"
> gen.1 <- aov(Wt ~ Litter*Mother, genotype)
> anova(gen.1)
Analysis of Variance Table

Response: Wt

Terms added sequentially (first to last)
              Df Sum of Sq Mean Sq F Value   Pr(F)
Litter         3      60.2   20.05  0.3697 0.77522
Mother         3     775.1  258.36  4.7632 0.00574
Litter:Mother  9     824.1   91.56  1.6881 0.12005
Residuals     45    2440.8   54.24
> gen.2 <- aov(Wt ~ Mother*Litter, genotype)
> summary(gen.2)
              Df Sum of Sq Mean Sq F Value   Pr(F)
Mother         3     771.6  257.20  4.7419 0.00587
Litter         3      63.6   21.21  0.3911 0.76000
Mother:Litter  9     824.1   91.56  1.6881 0.12005
Residuals     45    2440.8   54.24
>
```

Figure 7: Two paths through the lattice for the genotype data example

formula as the second argument, the sections of the model matrix corresponding to all non-intercept terms are omitted *seriatim* from the model, giving some sort of test for a main effect:

```
> drop1(gen.1, .~.)
Single term deletions

Model:
Wt ~ Litter * Mother
              Df Sum of Sq      RSS  F Value     Pr(F)
       <none>              2440.816
       Litter  3   27.6559 2468.472 0.169959 0.9161176
```

```
       Mother  3  671.7376 3112.554 4.128153 0.0114165
Litter:Mother  9  824.0725 3264.889 1.688108 0.1200530

> drop1(gen.2, .~.)
Single term deletions

Model:
Wt ~ Mother * Litter
              Df Sum of Sq      RSS  F Value     Pr(F)
       <none>              2440.816
       Mother  3  671.7376 3112.554 4.128153 0.0114165
       Litter  3   27.6559 2468.472 0.169959 0.9161176
Mother:Litter  9  824.0725 3264.889 1.688108 0.1200530
```

They are indeed order independent, so what are they?

Provided you have used a contrast matrix with zero-sum columns they will be unique, and they are none other than the notorious 'Type III sums of squares'. If you use, say, `contr.treatment` contrasts, though, so that the columns do not have sum zero, you get nonsense. This sensitivity to something that should in this context be arbitrary ought to be enough to alert anyone to the fact that something silly is being done.

From the looks of things, MathSoft has put a lot of work into re-programming the computation via the tedious SAS formula for doing so. Even then it is not complete, as it does not work for multistratum experiments.

Well don't look at me like that—you didn't really expect me to tell them how to do it the easy way, did you?

6 Enhancements

For all the solemn wringing of hands and knitting of brows going on above it must still be said that S-PLUS offers the best environment and suite of tools for actually doing linear modelling on the sorts of consulting jobs that arise in practice. The area covered by just the four fitting functions `lm`, `aov`, `glm` and `nls` is handled in SAS by an unbelievable array of `PROC`s each with some special features and other special limitations. Even the notion of "General linear model" in SAS simply means a linear model that is allowed to have *both* factors *and* quantitative explanatory variables. Just how general can you get?

Nevertheless there are some features of the linear modelling software in S-PLUS that I would like to see enhanced and even simplified. I offer them here as suggestions for future work.

- A mechanism for *declaring* marginality between non-factor terms, such as between powers of a single variable, or linear terms and their products is urgently needed.

This would then guide functions like `drop1`, `step` and our own function `stepAIC` in deciding which terms are legitimate candidates for removal at any particular stage. This would greatly enhance the value of these otherwise quite risky tools.

- The `%in%` is redundant and serves no very useful purpose. It could very easily be made a synonym for the `:` operator as it is in R, in fact, but no one (other than Ross Ihaka, who did it and me, who wormed it out of him) has realised it as yet!) and silently dropped in some future release.

- For teaching purposes it would be useful to have a switch that required users to include the intercept term in formulae if it is needed. This would definitely help more students than it would hinder. In other words it should be possible to override the automatic intercept term.

- The `:` operator is the primary one in terms of which the other combination operators could be defined. For example

  – `a*b*c` should be understood to mean `(1+a):(1+b):(1+c)` expanded algebraically and nonzero constant multipliers removed.

  – `a/b` should be understood to mean `a:(1+b)` or `a + a:b`, again with suitable simplifications.

  – `(a+b+c)^2` should be able to generate a general degree 2 regression model. As it stands, powers of terms are replaced by linear terms, which is sensible if the term is a factor, but unhelpful if the term is a quantitative predictor.

7 Mixed effects, multistratum models and variance components

There is little reason to believe that the current explosion of interest in what are called ‘random effects models’ or ‘mixed effects models’ is a passing phase. The proof of that is that it is not a new subject at all, but in one form or another has been around for a very long time, but with the different traditions in which it has independently arisen having little knowledge of what is going on over the fence.

A mixed effects linear model is one where some of the coefficients are regarded as themselves random variables, and interest focuses on the properties of the distribution giving rise to them, in particular its variance, known for some reason as a *variance component*.

In some traditions there is interest in ‘estimating’ the unobserved instances of the random variables themselves, but rather than call them estimates the fashion is to give them a different name such as BLUPs, posterior modes, and several others. I favour a different name from ‘estimate’ as well, but my preference is to call them ‘residuals’ since they do have exactly the same status as ordinary residuals from a simple fixed effects regression model, which are possibly the simplest special case. Summing the squares of these residuals and dividing by (a) the number of them or (b) the degrees of freedom to estimate the variance component is the choice that has become known as (a) maximum likelihood or (B) REML, although it is often presented in much more guarded and qualified terms.

There are two way of looking at a mixed effects model, at least. In one way the focus is on the unobserved random variables and the fixed regression coefficients, and we set about estimating both.

In the second way the random effects model are taken as a kind of paradigm which *might* be applicable, but the real difference with simple mixed effects models is that the observations are now acknowledged to be dependent, but with a very highly structured variance matrix, the parameters of which are often of interest in themselves.

Consider the example of a feeding experiment again. We have a number k of chicken coops each taking m chickens giving us $n = km$ bird weight gains at the end of the experiment. Each coop can only take one diet, so all information comparing diets comes from comparisons of coop total weight gains, but the birds within the coop might well be of different genotypes, say, so information on some fixed effects is available within coops.

Given that there is competition between birds for the food on offer, and only a total amount of food, there is every reason to anticipate that correlation between individual animal weights, the so-called intra-class correlation will be *negative* and so the variance component estimate on the boundary, but really uninformative. In this case the usual paradigm of an additive component for "coops" is not appropriate, and the variance component question rather silly, but the usual multistratum analysis is quite sensible and valid.

The feature of random effects linear models that makes them easy to handle within the framework of linear models is the fact that, even though the variance matrix is not scalar, *the eigenvectors* are nevertheless *known*. This means it is possible to find orthonormal spanning sets for the stable subspaces as columns of matrices, say $Z_1, Z_2, \ldots, Z_k$ such that $Z = [Z_1\, Z_2\, \cdots\, Z_k]$ is an orthogonal matrix and the components of $Z_i^T Y$ are independent and have equal variance. More completely, $Z^T Y$ has a diagonal variance matrix with k distinct diagonal entries that we might well call the canonical variance components.

The sets of linear functions themselves are called the *strata* of the experiment, but in reality are none other than the known stable subspaces of the variance matrix.

The simple case occurs when it is possible to re-parameterize the regression coefficient vector so that each $Z_i^T Y$ has a mean depending on a different subset of the coefficient parameters. In this case the likelihood completely factorizes and each set of linear functions, $Z_i^T Y$, has all information on the coefficient parameters on which its distribution depends. There is still a problem if the canonical variance components are functions of fewer than k underlying parameters, but this is fairly uncommon in practice.

In the more difficult case it is not possible to partition the coefficient parameters in this way, and information on some linear functions of the coefficients has to come from two or more strata. Finding best linear unbiased estimates of the coefficient vector under these circumstances used to be a problem known as *the recovery of interblock information*. The maximum likelihood estimates are, of course, weighted estimates of the estimates from the separate strata, and whether you use maximum likelihood or REML estimation for the canonical variances at that point starts to make a difference.

The S-PLUS function `aov` with the *Error* special function within the formula to declare the strata provides a stratum-by-stratum analysis, no more. If the design is orthogonal this is usually all that is needed. The function `lme` at present provides a method of handling some simple cases of the more general problem (certainly covering the bulk of practical cases, particularly with the next planned release) but certainly not all. A general development effort in this area, and more importantly almost, its extension to generalized linear models and non-linear models is urgently needed, as much for commercial reasons as for anything else.

7.1 An example: the petroleum data of Nilon H Prater

This is a now famous example that if N H Prater were still around would no doubt amuse him greatly. The data set comes from a magazine article (Prater, 1956). Prater collected information on the gasoline yield from crude oils at various stages, known as "end points" of the refining process. Each crude oil also has three measurements made on it, namely the specific gravity and two different vapour pressures.

By sorting the data it becomes clear that although there are 32 observations there are really only 10 different crudes involved. There is no hint of this in Prater's article, though it is now taken as well established, just as if he had, in fact. For convenience the data set is given here in Table 3 on page23.

A plot of the data shows that within each crude the regression of yield on end point is aggressively linear, (with perhaps one small exception) and almost unbelievably parallel, although there are large and presumably important differences between the intercepts.

Fitting separate regressions for the 10 samples consumes 20 regression parameters and 1 variance, and for just 32 observations this is rather too many.

A mixed effects model with random intercepts and slopes has only 2 fixed effects parameters, but 4 variance parameters; (3 variances and one covariance). The degree of parametrisation is close to acceptable, though using the BLUPs of the random terms gives something of the flexibility of the 21 parameter model above. In essence the model still adapts to some variation in parameters between crudes, but some of the more extreme cases have estimates noticeably contracted towards the centre.

Consider now models with fixed slope but variable intercepts.

```
> fm <- aov(Y ~ EP + No, petrol)
> gm <- aov(Y ~ EP + Error(No), petrol)
```

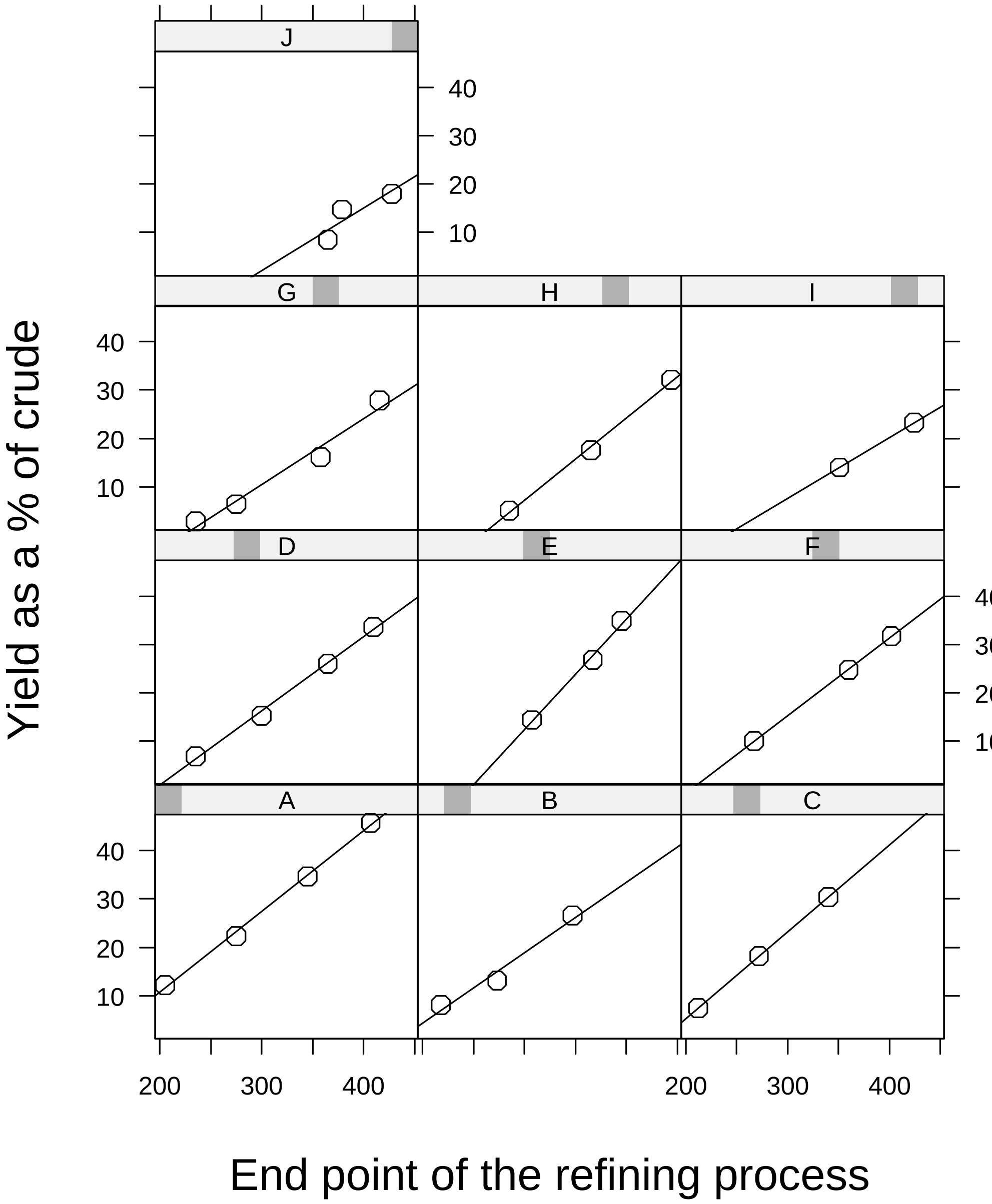

Figure 8: The refinery data of Nilon H Prater. Yield versus end point of the refining process for 10 crude oil samples.

```
> hm <- lme(Y ~ EP, random = ~1, cluster = ~No, data = petrol)

> coef(fm)
 (Intercept)       EP      No1       No2     No3       No4
     -33.708 0.15873 -4.1374 -0.20999 -1.762 -0.98419
      No5    No6      No7      No8      No9
 -0.83673 -1.369 -1.2734 -1.3524 -1.6188
```

```
> coef(gm[[2]])
       EP
 -0.014912

> coef(gm[[3]])
      EP
 0.15873

> coef(hm)
  (Intercept)      EP
A     -19.974 0.15757
B     -28.200 0.15757
C     -24.756 0.15757
D     -31.205 0.15757
E     -30.822 0.15757
F     -31.903 0.15757
G     -37.255 0.15757
H     -39.129 0.15757
I     -42.234 0.15757
J     -47.585 0.15757
```

The first fit above has a model with 10 intercepts and one slope gives an estimate of $\widehat{\beta} = 0.15873$.

The second model recognizes that (normalized) crude totals and within crude contrasts will have different variances. The estimate $\widehat{\beta}$ from within crude contrasts (for technical reasons stratum 3 rather than stratum 2) is the same as that for the previous case, though its estimate of error may be slightly different. In addition there is an estimate of the regression coefficient from within block totals (stratum 2) but with a nonsensical value since the amount of information available from that stratum is practically negligible.

Finally the random effects model gives a very similar value, $\widehat{\beta} = 0.15757$ since it combines the information from within and between crudes.

I contend that a recognition that what's really going on in linear mixed effects models is a variance matrix with known stable subspaces is the key observation that turns it into a problem where some of the standard methods of univariate fixed effects linear models are possibly applicable.

A Data sets used

Density	24.7	24.8	27.3	28.4	28.4	29.0	30.3	32.7	35.6	38.5	38.8	39.3
Hardness	484	427	413	517	549	648	587	704	979	914	1070	1020
Density	39.4	39.9	40.3	40.6	40.7	40.7	42.9	45.8	46.9	48.2	51.5	51.5
Hardness	1210	989	1160	1010	1100	1130	1270	1180	1400	1760	1710	2010
Density	53.4	56.0	56.5	57.3	57.6	59.2	59.8	66.0	67.4	68.8	69.1	69.1
Hardness	1880	1980	1820	2020	1980	2310	1940	3260	2700	2890	2740	3140

Table 1: The Janka hardness data of Australian hardwoods.

Mother: A				Mother: B				Mother: I				Mother: J			
A	B	I	J	A	B	I	J	A	B	I	J	A	B	I	J
61.5	60.3	37.0	59.0	55.0	50.8	56.3	59.5	52.5	56.5	39.7	45.2	42.0	51.3	50.0	44.8
68.2	51.7	36.3	57.4	42.0	64.7	69.8	52.8	61.8	59.0	46.0	57.0	54.0	40.5	43.8	51.5
64.0	49.3	68.0	54.0	60.2	61.7	67.0	56.0	49.5	47.2	61.3	61.4	61.0		54.5	53.0
65.0	48.0		47.0		64.0			52.7	53.0	55.3		48.2			42.0
59.7					62.0					55.7		39.6			54.0

Table 2: The rat genotype data. Litters of rats from one of four genotypes, A, B, I or J, were fostered by mothers of one of the same four genotypes. The response is the litter mean growth rate.

No	SG	VP	V10	(EP,	Y)	(EP,	Y)	(EP,	Y)	(EP,	Y)
A	50.8	8.6	190	(205,	12.2)	(275,	22.3)	(345,	34.7)	(407,	45.7)
B	40.8	3.5	210	(218,	8.0)	(273,	13.1)	(347,	26.6)		
C	40.0	6.1	217	(212,	7.4)	(272,	18.2)	(340,	30.4)		
D	38.4	6.1	220	(235,	6.9)	(300,	15.2)	(365,	26.0)	(410,	33.6)
E	40.3	4.8	231	(307,	14.4)	(367,	26.8)	(395,	34.9)		
F	32.2	5.2	236	(267,	10.0)	(360,	24.8)	(402,	31.7)		
G	41.3	1.8	267	(235,	2.8)	(275,	6.4)	(358,	16.1)	(416,	27.8)
H	38.1	1.2	274	(285,	5.0)	(365,	17.6)	(444,	32.1)		
I	32.2	2.4	284	(351,	14.0)	(424,	23.2)				
J	31.8	0.2	316	(365,	8.5)	(379,	14.7)	(428,	18.0)		

Table 3: The petrol refining data of Nilon H Prater. Ten crude oils with petroleum yield measured at different end points of the refining process.

Year	Rain0	Temp1	Rain1	Temp2	Rain2	Temp3	Rain3	Temp4	Yield
1930	17.75	60.2	5.83	69.0	1.49	77.9	2.42	74.4	34.0
1931	14.76	57.5	3.83	75.0	2.72	77.2	3.30	72.6	32.9
1932	27.99	62.3	5.17	72.0	3.12	75.8	7.10	72.2	43.0
1933	16.76	60.5	1.64	77.8	3.45	76.4	3.01	70.5	40.0
1934	11.36	69.5	3.49	77.2	3.85	79.7	2.84	73.4	23.0
1935	22.71	55.0	7.00	65.9	3.35	79.4	2.42	73.6	38.4
1936	17.91	66.2	2.85	70.1	0.51	83.4	3.48	79.2	20.0
1937	23.31	61.8	3.80	69.0	2.63	75.9	3.99	77.8	44.6
1938	18.53	59.5	4.67	69.2	4.24	76.5	3.82	75.7	46.3
1939	18.56	66.4	5.32	71.4	3.15	76.2	4.72	70.7	52.2
1940	12.45	58.4	3.56	71.3	4.57	76.7	6.44	70.7	52.3
1941	16.05	66.0	6.20	70.0	2.24	75.1	1.94	75.1	51.0
1942	27.10	59.3	5.93	69.7	4.89	74.3	3.17	72.2	59.9
1943	19.05	57.5	6.16	71.6	4.56	75.4	5.07	74.0	54.7
1944	20.79	64.6	5.88	71.7	3.73	72.6	5.88	71.8	52.0
1945	21.88	55.1	4.70	64.1	2.96	72.1	3.43	72.5	43.5
1946	20.02	56.5	6.41	69.8	2.45	73.8	3.56	68.9	56.7
1947	23.17	55.6	10.39	66.3	1.72	72.8	1.49	80.6	30.5
1948	19.15	59.2	3.42	68.6	4.14	75.0	2.54	73.9	60.5
1949	18.28	63.5	5.51	72.4	3.47	76.2	2.34	73.0	46.1
1950	18.45	59.8	5.70	68.4	4.65	69.7	2.39	67.7	48.2
1951	22.00	62.2	6.11	65.2	4.45	72.1	6.21	70.5	43.1
1952	19.05	59.6	5.40	74.2	3.84	74.7	4.78	70.0	62.2
1953	15.67	60.0	5.31	73.2	3.28	74.6	2.33	73.2	52.9
1954	15.92	55.6	6.36	72.9	1.79	77.4	7.10	72.1	53.9
1955	16.75	63.6	3.07	67.2	3.29	79.8	1.79	77.2	48.4
1956	12.34	62.4	2.56	74.7	4.51	72.7	4.42	73.0	52.8
1957	15.82	59.0	4.84	68.9	3.54	77.9	3.76	72.9	62.1
1958	15.24	62.5	3.80	66.4	7.55	70.5	2.55	73.0	66.0
1959	21.72	62.8	4.11	71.5	2.29	72.3	4.92	76.3	64.2
1960	25.08	59.7	4.43	67.4	2.76	72.6	5.36	73.2	63.2
1961	17.79	57.4	3.36	69.4	5.51	72.6	3.04	72.4	75.4
1962	26.61	66.6	3.12	69.1	6.27	71.6	4.31	72.5	76.0

Table 4: The Iowa wheat yield data. The yield (in bushels/acre) for the state of Iowa, with average monthly temperatures and rainfalls as covariates. The year is a surrogate for variety improvements.

References

Draper, N. R. and Smith, H. (1981). *Applied Regression Analysis*, second edn, Wiley, New York.

Hastie, T. J. and Tibshirani, R. J. (1990). *Generalized Additive Models*, Chapman & Hall, London/New York.

SAS Institute Inc. (1990). *SAS/STAT User's Guide, Version 6*, Vol. 1, fourth edition edn, SAS Instutute, Inc., Cary, NC.

Prater, N. H. (1956). Estimate gasoline yields from crudes, *Petroleum Refiner* **35**: 236–238.

Scheffé, H. (1959). *The Analysis of Variance*, Wiley, New York.

Williams, E. J. (1959). *Regression Analysis*, Wiley, New York.