(largest contemplated) linear model Ω by Jaeckel's method, say using (3.15). Having obtained the resulting $\mathbf{e}(\hat{\beta}_{\rm JW})$ residuals $\mathbf{e}(\hat{\beta}_{\rm JW})$ and an estimate of $\hat{\sigma}_{\rm R}$, form pseudo values for the data as follows,

$$Y_i^* = [X\hat{\beta}_{\rm JW}]_i + \hat{\sigma}_{\rm R}\left[R_i(\hat{\beta}_{\rm JW}) - \frac{(N+1)}{2}\right], \qquad i = 1, \cdots, n.$$

The heuristics of Bickel (1976) can I believe be rigorized in this case also to conclude that if we now act as if the Y_i^* were the data, apply ordinary least squares methods in fitting subhypotheses and then calculate the usual F statistics we are asymptotically right in the sense that the asymptotic null distribution and power functions of these statistics agree with the χ^2 approximations to the corresponding Hettmansperger-McKean statistics. We expect more. For instance, application of Tukey's method of multiple comparisons to the pseudo observations should have the same efficiency (say in terms of length of the intervals) with respect to the method applied to the original observations as the Wilcoxon test has to the t test.

Of course the asymptotic χ^2 approximations here too will be inadequate as Draper points out. However, one might hope that the same empirical observations made by Draper continue to hold, viz., using the classical degrees of freedom for F works adequately.

Let me add a caution. As Draper points out what is done here guarantees robustness only against heavy tails. In particular, sensitivity to high leverage points among the $[X\beta]_i$ is not affected. Nor is sensitivity to heteroscedasticity, dependence, transformation of the Y scale, etc. Perhaps the pseudo values could be used as a first step in procedures where the second step fitting method would address these departures and of course, one would then iterate.

It is worth noting that the scope of the methods discussed by Draper has recently been enlarged by Tsiatis (1986) to handle the case of right censoring of the Y_i. It's not clear what happens to the pseudo value-based procedures in this context.

Finally, it is worth remembering that the scope of purely rank-based procedures is much greater than what is suggested by the Kruskal-Wallis, Friedman-Tukey tests. In particular, ranks not rank of residual procedures are appropriate when one considers transformation models of the form

$$h(Y_i) = [X\beta]_i + e_i \quad i = 1, \cdots, n$$

where the e_i are assumed to come from some parametric family but h is an unknown monotone transformation. See Doksum (1987) and Bickel (1986) for example.

I congratulate David Draper on this clear insightful presentation.

ADDITIONAL REFERENCES

Bickel, P. J. (1976). Another look at robustness: A review of reviews and some new developments (with discussion). *Scand. J. Statist.* **3** 145–168.

Bickel, P. J. (1986). Efficient testing in a class of transformation model. Papers on semiparametric models at the ISI Centenary Session, C.W.I., Amsterdam.

Doksum, K. (1987). An extension of partial likelihood methods for proportional hazard models to general transformation models. *Ann. Statist.* **15** 325–345.

Tsiatis, A. (1986). Estimating regression parameters using linear rank statistics for censored data. Technical Report, School of Public Health, Harvard Univ.

Comment

R. Douglas Martin

Dr. Draper has provided a very nice exposition and review of two rank-based robust methods for fixed effects ANOVA problems. In so doing, he concentrates on (i) the formal structure of the methods and (ii) robust inference based on rank-based analogues of the classical test statististics, where robust inference is taken to mean robustness of validity and efficiency. Given the author's commitment to focus on the R-estimate approach, I would only wish that he had given some emphasis to examples, and in so doing revealed the exploratory data-analytic use of the methods. As far as the focus on rank-based methods goes, I have a pragmatically motivated reservation based on a concern I share with Draper, namely, robust methods are not widely available in the major statistical packages.

As Draper points out, R-estimates comprise just one of three major classes of robust estimates, with L-estimates and M-estimates being the other two, and

R. Douglas Martin is Professor of Statistics, Department of Statistics, University of Washington, Seattle, Washington 98195.

these classes hardly exhaust the range of proposals found in the literature. For example, other approaches include robust Pitman-type estimates and robust minimum distance estimates. Although the wide range of possibilities reflected in these various approaches has made the robustness area a rich one for research, this has not been altogether helpful for the potential user of such methods, who may be confused by the wide range of possibilities. Researchers have their own individual preferences, and we have no "general committee" to recommend one class over another. Furthermore, it does not seem to be a very convincing sales pitch to tell the applied statistician to "pick one at random, because that is better than using no robust procedure." All in all, the actual use of robust procedures has probably been hindered considerably by the existence of far too many robust statistics.

Thus, if we really want to see robust methods widely used by practitioners, we must somehow focus on as few types of estimates as possible from a statistical performance point of view. Only by such a focus will we see a satisfactory emergence of robust methods in many statistical software packages, and their inclusion as a standard part of the statistics curriculum offering (much as courses on nonparametric tests are standard fare). It is with this motivation that I make the ensuing remarks.

I agree with the author that L-estimates have yet to prove themselves with regard to providing a unified approach for a wide range of statistical models. Although M and R methods are considerably more unified than L-estimates, I do not find R-estimates as appealing as M-estimates in this regard. Furthermore, because R methods only have "simple, closed form expressions" for a few special models, this argument does not provide a very convincing case for preferring this class.

The main justification for having a strong preference for one class over the other should be based on dominating robustness properties. In the event that no class clearly dominates the other, then I believe we should emphasize to applied statisticians that class which has the best combination of intuitive appeal, transparency and generality. From this point of view the M-estimate class seems preferable. Its intuitive appeal and transparency features for general linear models include (a) close connection with maximum likelihood procedures, (b) a weighted least squares interpretation and (c) a constant $E\Psi^2(\varepsilon)/E^2\Psi'(\varepsilon)$ in the asymptotic variance-covariance formula, which by now is becoming relatively familiar and which is easily estimated from the data in a natural way (perhaps with some small sample size correction). By way of contrast with (c), the constant $\int f^2$ for the Wilcoxon scores does not have a ring of general familiarity, and it is rather more complicated to estimate from the data.

The M-estimate class has considerable generalizability for a wide range of statistical models, both for ANOVA setups and for other problems (e.g., time series, see for example Martin and Yohai, 1985). Within the ANOVA realm, M-estimates provide natural estimates in factorial experiments (Carroll, 1980), for random and fixed effects models, and also for variance components in a REML-like manner (Fellner, 1986). Although Draper mentions that the rank-based methods may be extended to mixed models, I wonder really how natural the methods are, even for estimating random effects, let alone variance components?

One more distracting feature of R-estimates. If one wishes to see widespread use of R methods, then the point which Draper makes of clearly distinguishing rank-based methods from rank tests will have to be made time and time again, as far too many users will assume that the former have the same nonparametric properties as the latter.

Now the preceding ceteris paribus comments are moot if in fact R methods clearly dominate M methods (and others) from a robustness point of view. Evidence to date suggests that this will not be the case, as Draper and other workers might well agree. However, a complete and thorough analysis remains to be done.

Along these lines one would prefer to see a range of robustness concepts brought to bear on the problem, as well as more extensive Monte Carlo studies of robustness of level and power. In particular, use should be made of two basic tools for robustness: the *influence curve* for estimation and testing, and the *breakdown point* for estimation and testing (see Hampel, 1974; Ylvisaker, 1977; Huber, 1981; Lambert, 1981; Donoho and Huber, 1983; Field and Ronchetti, 1985; Hampel, Ronchetti, Rousseuw and Stahel, 1986, and references therein).

Perhaps a few words concerning the breakdown point may indicate its potential utility in selecting statistics for ANOVA situations. The breakdown point is a "global" measure of robustness which has considerable transparency as a concept. It is easy to explain the concept to any scientist or engineer, and illustrate it by comparing the sample mean and median, which have breakdown points of 0 and 0.5, respectively. It is particularly important to have high breakdown point procedures when using robust methods for exploratory purposes. Focusing on the problem of estimating location, for purposes of illustration, one finds the Hodges-Lehmann R-estimate has a breakdown point (BP) of only 0.29, whereas the median (which is both an R- and M-estimate) has the desirable high breakdown point (HBP) of 0.5.

In estimating contrasts for ANOVA situations, the BP will be much smaller. Consider the simplest case of testing for equality of location parameters in a balanced two-sample situation with $n/2$ observations in each sample. Any two-sample test statistic constructed as the difference of location estimates for each of the samples will break down as soon as either of the two location estimates breaks down. The best we can hope for is to achieve the HBP of 0.5 *for each of the samples*. But this results in a BP of 0.25 relative to the entire sample size of n. The situation will quite clearly be worse for the higher way ANOVA situations of major interest. The basic problem in ANOVA is that typically there are only a small number of observations per parameter, and thus a small number of wild points can spoil the various contrasts of interest. Thus, there will be a premium on maintaining the highest possible BP in ANOVA situations. From this point of view, the BP of 0.5 for each parameter obtained with median-based methods (e.g., median polish where applicable) seems preferable for example to the Hodges-Lehmann estimate BP of 0.29.

Returning to the M-estimates versus R-estimates issue for a moment, it is important to note that the latter have the attractive property that they do not require estimation of a nuisance scale parameter. On the other hand, M-estimates based on the popular psi functions, e.g., Huber's favorite or Tukey's bisquare, do require an auxiliary scale estimate. For the small sample sizes per parameter that occur in many ANOVA situations, this may result in loss of robustness in level and power relative to R methods. This issue needs to be studied in detail.

One other robustness consideration is worth noting. For those situations where one really needs to estimate an effect rather than a true contrast, it will be important to control the maximum bias due to asymmetric contamination. In this regard it may be useful to consider *min-max bias robust* estimates. For example, Huber (1964) showed that among all translation equivariant estimates of location, the median solves the problem for ε contamination models. Hence, in that case the HBP and min-max bias properties coincide, but this is not always the case. Recently Yohai, Zamar and I have been working on min-max bias robust estimates of scale (Martin and Zamar, 1987) and regression (Martin, Yohai and Zamar, 1987). In the solutions found to date, the breakdown point BP = BP(ε) will be relatively close to 0.5 for all but very small ε in an ε-contaminated model, and furthermore BP(ε) $\rightarrow 0.5$ as $\varepsilon \rightarrow 0.5$. Perhaps the min-max bias robust approach will be of some utility in ANOVA problems.

ADDITIONAL REFERENCES

CARROLL, R. J. (1980). Robust methods for factorial experiments with outliers. *Appl. Statist.* **29** 246–251.

DONOHO, D. L. and HUBER, P. J. (1983). The notion of breakdown point. In *A Festschrift for Erich L. Lehmann* (P. J. Bickel, K. A. Doksum and J. L. Hodges, Jr., eds.) 157–184. Wadsworth, Belmont, Calif.

FELLNER, W. H. (1986). Robust estimation of variance components. *Technometrics* **28** 51–60.

FIELD, C. A. and RONCHETTI, E. (1985). A tail area influence function and its application to testing. *Sequential Anal.* **4** 19–41.

HAMPEL, F. R. (1974). The influence curve and its role in robust estimation. *J. Amer. Statist. Assoc.* **69** 383–393.

HAMPEL, F. R., RONCHETTI, E. M., ROUSSEEUW, P. J. and STAHEL, W. A. (1986). *Robust Statistics: The Approach Based on Influence Functions.* Wiley, New York.

HUBER, P. J. (1964). Robust estimation of a location parameter. *Ann. Math. Statist.* **35** 73–101.

LAMBERT, D. (1981). Influence functions for testing. *J. Amer. Statist. Assoc.* **76** 649–657.

MARTIN, R. D. and YOHAI, V. J. (1985). Robustness in time series and estimating ARMA models. In *Handbook of Statistics* **5** (E. J. Hannan, P. R. Krishnaiah and M. M. Rao, eds.) 119–155. North-Holland, Amsterdam.

MARTIN, R. D., YOHAI, V. J. and ZAMAR, R. H. (1987). Min-max bias robust regression. *Ann. Statist.* To appear.

MARTIN, R. D. and ZAMAR, R. H. (1987). Min-max bias robust estimates of scale. Technical Report 72, Dept. Statistics, Univ. Washington.

YLVISAKER, D. (1977). Test resistance. *J. Amer. Statist. Assoc.* **72** 551–556.

Rejoinder

David Draper

Let me open the rejoinder by thanking the discussants for their insightful and kind comments. They have not given me much to disagree with, but (since it is at least as much the job of the rejoinderer to be contentious as it is for the discussants themselves) I'll see what I can do. I begin with some remarks about the influence of robustness work on actual practice to date; continue with some comments on the relevance of expert systems research to the comparison of modeling strategies; and finally devote most of my