Statistical Science
2019, Vol. 34, No. 4, 584–590
https://doi.org/10.1214/19-STS746
Main article: https://doi.org/10.1214/18-STS693, https://doi.org/10.1214/18-STS694
© Institute of Mathematical Statistics, 2019

Comment: Models Are Approximations!

Anthony C. Davison, Erwan Koch and Jonathan Koh

Abstract. This discussion focuses on areas of disagreement with the papers, particularly the target of inference and the case for using the robust 'sandwich' variance estimator in the presence of moderate mis-specification. We also suggest that existing procedures may be appreciably more powerful for detecting mis-specification than the authors' RAV statistic, and comment on the use of the pairs bootstrap in balanced situations.

Key words and phrases: Bootstrap, designed experiment, infinitesimal jackknife, model mis-specification, regression diagnostics, sandwich variance estimator.

1. INTRODUCTION

The authors of these papers have thought hard about fundamental issues in modelling, and although we disagree with their main conclusions, their work repays study. As the basic issues are posed most simply in the first article, we shall largely confine our discussion to it, referring to the articles as Paper 1 and Paper 2, or jointly as the Papers. Like the authors, we avoid technical issues, assuming the existence of all necessary derivatives, moments and matrix inverses.

One dictionary definition of 'model' is 'a simplified description, especially a mathematical one, of a system or process'. This clearly implies that the model is not the reality, and in any application it is essential to ensure that the inevitable simplification is not so radical that the model becomes useless for its intended purpose. Although the key issue in choosing a model is the goal of modelling, this varies so much that universal prescriptions are dangerous. Two classical goals are to summarise and to predict, but if it is to be useful for either, a statistical model should not traduce key aspects of the data. This entails careful data exploration in advance of modelling, and, after fitting models, the use of diagnostics of their fit, often graphical procedures, typically supplemented with test statistics when the visual evidence is equivocal. If a divergence between the model and data is found, then its likely impact on the conclusions needs to be assessed, and the benefit of dealing with it weighed against the cost of doing so. A crucial aspect is the range of validity of the model, which depends on its relationship to available theory, relevant previous experience and so forth—both summary and prediction will be more secure when inference is broadly based.

It is useful to separate model specification into primary and secondary aspects. Primary aspects relate to the major questions to be answered, whereas secondary aspects involve assumptions needed to answer the major questions but not crucial in themselves. In many cases, variation in a mean response due to changes in an explanatory variable is primary, whereas assumptions about the response variance or the error distribution are secondary: they affect uncertainty assessment for primary aspects, but are not in themselves usually central to the questions of interest—though, as mentioned above, there are no universal nostrums; covariances are key in Panaretos, Kraus and Maddocks (2010), for example.

The authors argue that if the data are generated by a nonlinear model and a linear model is fitted, then 'regressors are not ancillary, hence can't be treated as fixed'. Since one is never certain that a correct model has been fitted, this implies that regressors should always be treated as random. If correct, this astonishingly broad conclusion would run counter to at least a

Anthony C. Davison is Professor, École Polytechnique Fédérale de Lausanne, EPFL-FSB-MATH-STAT, Station 8, 1015 Lausanne, Switzerland (e-mail: Anthony.Davison@epfl.ch). Erwan Koch is Instructor, École Polytechnique Fédérale de Lausanne, EPFL-FSB-MATH-STAT, Station 8, 1015 Lausanne, Switzerland (e-mail: Erwan.Koch@epfl.ch). Jonathan Koh is PhD candidate, École Polytechnique Fédérale de Lausanne, EPFL-FSB-MATH-STAT, Station 8, 1015 Lausanne, Switzerland (e-mail: Jonathan.Koh@epfl.ch).

century of statistical practice. We regard it as incorrect, however, for several reasons, the first of which relates to the goal of analysis. In our view, the appropriate estimand in most settings is not the best population linear approximation touted in the Papers, $\beta(P)$, but the 'X-conditional parameter'

$$\begin{aligned}\beta(X) &= \operatorname*{argmin}_{\beta\in\mathbb{R}^{p+1}} \mathrm{E}\{(Y - X\beta)^2 \mid X\} \\ &= (X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}\mathrm{E}(Y \mid X),\end{aligned}$$

which depends only on the conditional distribution of the response given the explanatory variable X, and not on the distribution of X. Dependence of the target of inference $\beta(X)$ on the explanatory variables X might seem odd if we think in classical terms of a perfectly specified model, but on deeper reflection it makes sense, because the fitted model can only provide an approximation to the true mean response in the design region $\mathcal{X}$: the validity of any model is uncheckable outside $\mathcal{X}$ based on the data alone. For this reason, textbooks stress that one should avoid extrapolating a fitted regression, and, if extrapolation is essential, then its results should be treated with extreme caution.

A related basic issue is that $\beta(P)$ is inestimable from any finite dataset unless the true regression relationship is linear, whereas $\beta(X)$ can be estimated unbiasedly. This suggests that $\beta(X)$, not $\beta(P)$, should be the target in regression analysis—in which case X must be treated as fixed, because it determines the best approximation for the values of the explanatory variables *actually available to the investigator*. As mentioned above, the appropriate model depends on the goal of the analysis: if the objective is to provide an approximation useful in the region $\mathcal{X}$ in which values of X are known, then $\beta(X)$ is relevant, and not $\beta(P)$. Of course, as more data become available X changes, and with it $\mathcal{X}$ and the target of estimation, $\beta(X)$; this is natural, because a larger sample should allow a better approximation to a broader reality, even though $\beta(P)$ itself remains inestimable.

2. DIAGNOSTICS

The authors 'emphasize that diagnostics should be part of every regression analysis'. When diagnostic procedures are used correctly, large divergences between an assumed model and the underlying data should be detected and the fit improved so that the only remaining divergences between the fitted model and the data are on the borderline of detectability, or are irrelevant to the main goal of the analysis. Thus divergences between the assumed and true models that can easily be detected are not of interest.

This train of thought implies that in the canonical decomposition in equation (5) of Paper 1 it is realistic to suppose that $\eta(X) = n^{-1/2}g(X)$ and that heteroscedasticity is of the general form $\mathrm{var}(y \mid X) = \sigma^2\exp\{n^{-1/2}h(X)\}$, where $g(X)$ and $h(X)$ are of order one in probability, $O_{\mathrm{p}}(1)$. In large samples, such departures are not certain to be detected and, therefore, may perturb the residuals from the linear fit, whereas situations such as those shown in Figure 1 of Paper 1 should be detected using standard techniques and thus can be mitigated by fitting a more complex model. Since the Papers argue that the sandwich standard error is needed to offset the effects of mis-specification, it seems worthwhile to see how it performs under mis-specification on the borderline of detectability.

To assess this, we performed an experiment with $X_1, \dots, X_n \overset{\text{iid}}{\sim} U(0, 10)$, conditional on which

$$\begin{aligned}Y_j &= \beta_0 + \beta_1 X_j + n^{-1/2}\gamma X_j^{1.7} \\ &\quad + \exp(n^{-1/2}\delta X_j)\varepsilon_j, \quad j = 1, \dots, n,\end{aligned} \tag{1}$$

with $\delta \geq 0$ and $\varepsilon_1, \dots, \varepsilon_n \overset{\text{iid}}{\sim} \mathcal{N}(0, 1)$ independent of $\gamma \overset{\text{iid}}{\sim} \mathcal{N}(\mu_\gamma, \sigma_\gamma)$. As n increases this model has nonlinearity and heteroscedasticity on the border of detectability for fixed values of μ_γ, σ_γ and δ. Residuals for example datasets from this model, shown in the right-hand panels of Figure 1, suggest mis-specification of a homoscedastic linear model without this being blatantly obvious, precisely the setting in which using a 'robustified' standard error may seem worthwhile. Here, a cautious data analyst, wary of data-dredging, would likely decide not to elaborate the model. Is it worthwhile to follow the advice in the Papers and use the sandwich standard error, in hope of guarding against the effects of mis-specification?

For each of 20,000 simulated datasets of sizes $n = 50$, 100 and 200, we computed the classical and the sandwich standard errors, S_{c} and S_{s}, for the ordinary least squares estimates of β_0 and β_1. The corresponding 'true' values, S_{t}, were estimated by computing the standard deviations of the estimates based on the simulations. Figure 2 shows boxplots of the ratios $S_{\mathrm{c}}/S_{\mathrm{t}}$ and $S_{\mathrm{s}}/S_{\mathrm{t}}$ for four different configurations: no mis-specification, nonlinearity, heteroscedasticity and both nonlinearity and heteroscedasticity. In all boxplots, the means of both the classical and sandwich standard errors converge to the 'true' value. Although the sandwich standard error is on average slightly closer to the 'true' value for small n, its variability is so large com-

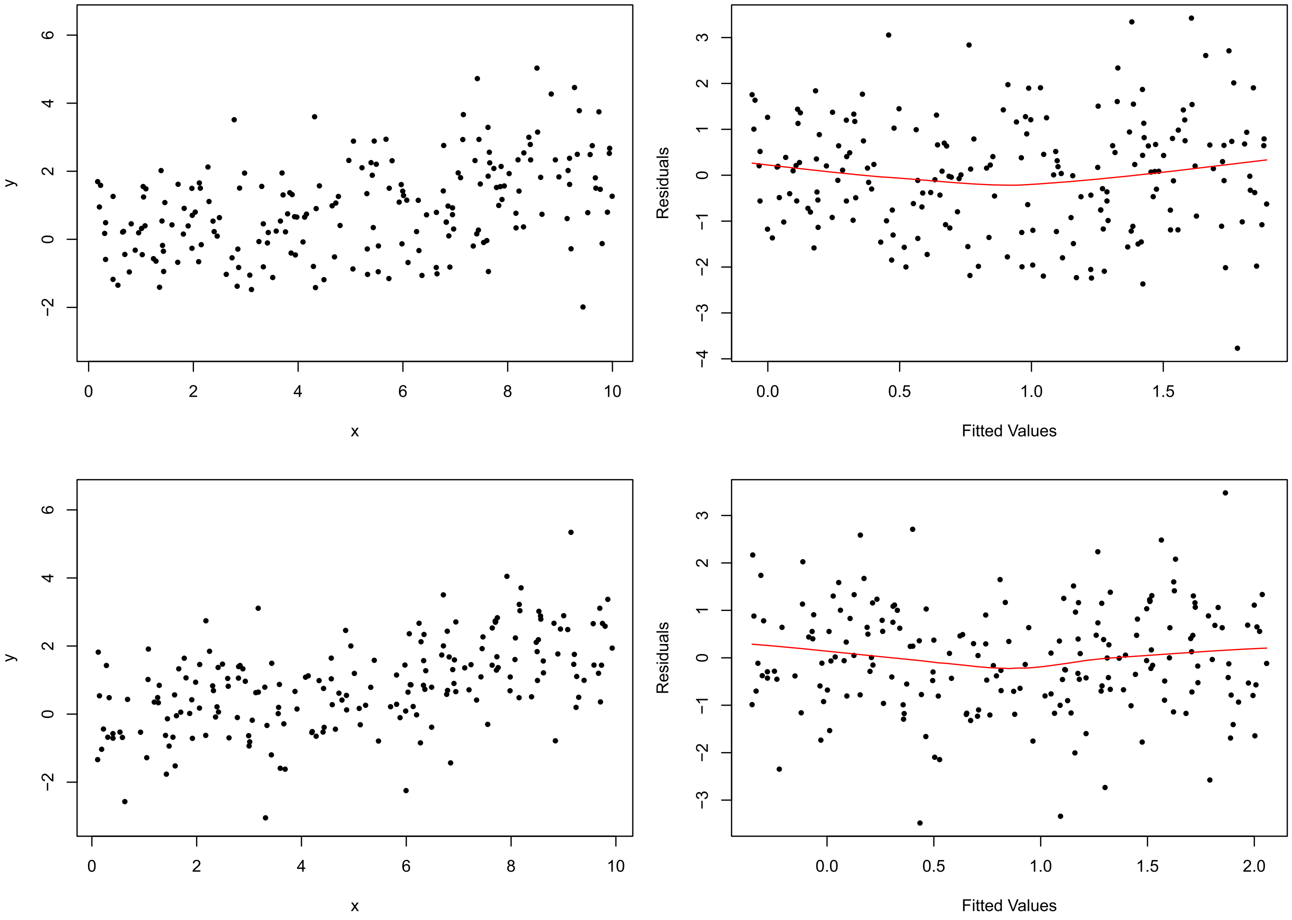


FIG. 1. *Two mis-specified datasets of size* $n = 200$ *drawn from model* (1) (*left*) *and plot of residuals against fitted values for a straight-line regression fit* (*right*), *with loess lines* (*red*). *Top panels*, $(\delta, \mu_\gamma, \sigma_\gamma) = (0.6, 0.7, 0.1)$, *and bottom panels*, $(\delta, \mu_\gamma, \sigma_\gamma) = (0.2, 0.7, 0.1)$.

pared to the classical standard error that, in all cases, the latter has lower relative root mean square error. It is straightforward to check that the ratio $S_\mathrm{s}/S_\mathrm{t}$ equals $1 + O_p(n^{-1})$ in such cases, as is suggested by the boxplots; the term of order n^{-1} depends on the mis-specification but disappears in large samples.

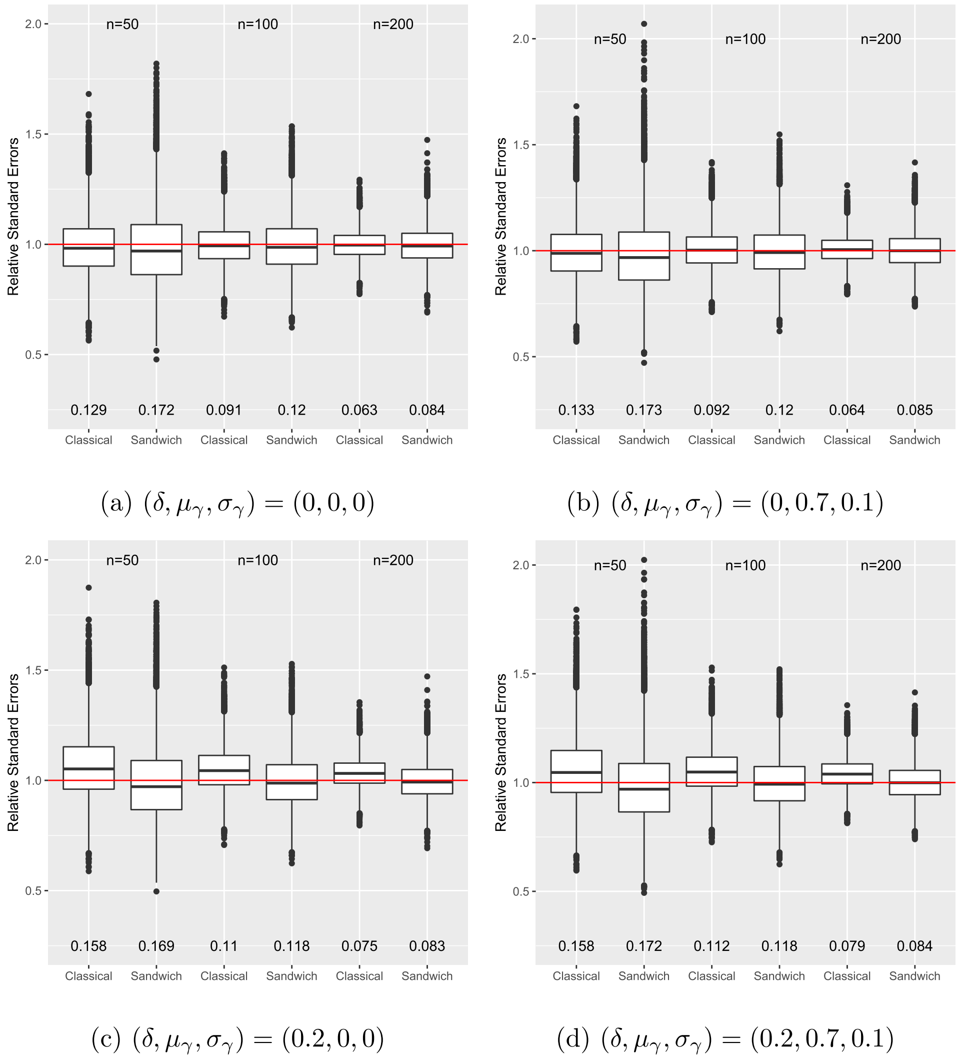
These very limited simulations suggest that the classical standard error is preferable to its sandwich counterpart when mis-specification is not obvious. Further investigation is of course warranted, but if the results here are representative then it appears that the sandwich standard error should be avoided unless one has decided to fit a visibly incorrect model.

3. NEW OR OLD DIAGNOSTICS?

The RAV test is intended to provide a generic test for covariates whose standard errors may be incorrect. However, the efficiency loss due to using the sandwich standard error can be 50% under ideal conditions (Hinkley and Wang, 1991), and smaller but still appreciable under mild mis-specification (Figures 2 and 3), leading one to question the power of a RAV-based test compared to existing tests for mis-specification. To assess this for the simple model (1), we compared RAV with the Cook and Weisberg (1983) test for variance heteroscedasticity and Tukey's 'one degree of freedom for nonadditivity' (Tukey, 1949), respectively. We took the power 1.7 in (1) so that the Tukey test is not on its home ground. There is a huge literature on regression diagnostics, references to which can be found in any relevant textbook, and we made no effort to seek optimal diagnostics, merely using two well-established tools that should be known to every rookie data analyst.

Table 1 shows that RAV has much lower power than these standard tests. The same is true even in cases of both nonlinearity and heteroscedasticity, so one might question the value of RAV if it can detect only the most blatant problems, such as in the sketches in Paper 1.

If the RAV is used regardless, then its construction as a ratio of variances suggests that in finite samples a chi-squared or F approximation to its null distribu-



(a) $(\delta, \mu_\gamma, \sigma_\gamma) = (0, 0, 0)$

(b) $(\delta, \mu_\gamma, \sigma_\gamma) = (0, 0.7, 0.1)$

(c) $(\delta, \mu_\gamma, \sigma_\gamma) = (0.2, 0, 0)$

(d) $(\delta, \mu_\gamma, \sigma_\gamma) = (0.2, 0.7, 0.1)$

FIG. 2. *Boxplots of the relative classical and sandwich standard errors* $S_{\mathrm{c}}/S_{\mathrm{t}}$ *and* $S_{\mathrm{s}}/S_{\mathrm{t}}$ *for the least squares estimator of intercept,* $\hat\beta_0$. *Results are for* 20,000 *simulated datsets with sample sizes* $n = 50$, 100 *and* 200 *in four configurations*: *no mis-specification* (*top left*); *nonlinearity* (*top right*); *heteroscedasticity* (*bottom left*); *nonlinearity and heteroscedasticity* (*bottom right*). *In all cases the mis-specifications are on the border of detectability. The relative root mean square errors* (*RMSE*) *of the standard errors are shown below the boxplots.*

tion would be vastly better than appealing to its limiting normality, and this is borne out by the plots in Appendix F of Paper 1.

4. PAIRS BOOTSTRAP

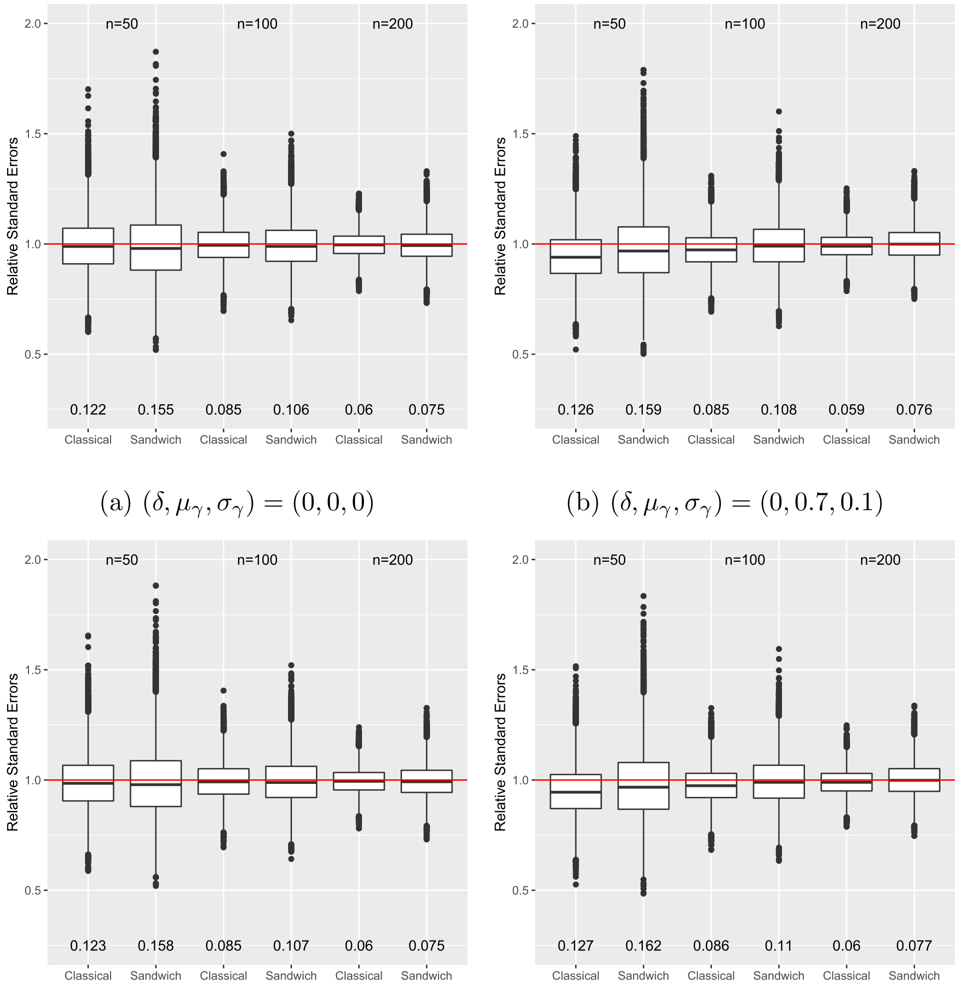
Paper 1 states that ‘Many results are … not new …’ This applies in particular to the relation between the sandwich variance and the pairs bootstrap, an account of which is given in Davison and Hinkley (1997), Sections 2.7, 6.2, 6.3. For the convenience of the reader, we summarise the key elements below.

The argument rests on the infinitesimal jackknife expansion of a statistical functional $t(F)$, where F represents a distribution. When F represents the distribution of $z = (x, y)$, with scalar y and $p \times 1$ vector x, the coefficient corresponding to a linear regression of y on x can be written as

$$t(F) = \left\{\int xx^{\mathrm{T}} F(\mathrm{d}x, \mathrm{d}y)\right\}^{-1} \int xyF(\mathrm{d}x, \mathrm{d}y); \tag{2}$$

in the notation of the Papers, $t(F) = \beta(P)$. The infinitesimal jackknife stems from the functional Taylor series expansion (Fernholz, 1983)

$$\begin{aligned} t(G) = t(F) &+ \int L_t(z; F)G(\mathrm{d}z) \\ &+ \frac{1}{2}\iint Q_t(z_1, z_2; F)G(\mathrm{d}z_1)G(\mathrm{d}z_2) \\ &+ \cdots, \end{aligned} \tag{3}$$



(a) $(\delta, \mu_\gamma, \sigma_\gamma) = (0, 0, 0)$

(b) $(\delta, \mu_\gamma, \sigma_\gamma) = (0, 0.7, 0.1)$

(c) $(\delta, \mu_\gamma, \sigma_\gamma) = (0.2, 0, 0)$

(d) $(\delta, \mu_\gamma, \sigma_\gamma) = (0.2, 0.7, 0.1)$

FIG. 3. *As for Figure 2, but for the slope estimator $\hat{\beta}_1$.*

where the first derivative of t at F, the influence function (Hampel, 1974, Hampel et al., 1986), is defined by

$$L_t(z; F) = \left.\frac{\partial t\{(1-\varepsilon)F + \varepsilon H_z\}}{\partial \varepsilon}\right|_{\varepsilon=0},$$

with H_z the step function jumping from zero to one at z. The second derivative is

$$\begin{aligned} &Q_t(z_1, z_2; F) \\ &\quad = \left.\frac{\partial^2 t\{(1-\varepsilon_1-\varepsilon_2)F + \varepsilon_1 H_{z_1} + \varepsilon_2 H_{z_2}\}}{\partial \varepsilon_1 \partial \varepsilon_2}\right|_{\varepsilon_1=\varepsilon_2=0}, \end{aligned}$$

TABLE 1
Empirical power (%) for tests at the 5% nominal level of the Cook and Weisberg (1983) and RAV tests with $(\delta, \mu_\gamma, \sigma_\gamma) = (0.6, 0, 0)$, and for the Tukey (1949) and RAV tests with $(\delta, \mu_\gamma, \sigma_\gamma) = (0, 0.7, 0.1)$, for 1000 samples of sizes 50, 100 and 200. 10,000 permutations were used for RAV

Sample size	50	100	200		50	100	200
Cook & Weisberg	61	64	67	Tukey	44	46	46
RAV	7.3	7.0	6.1	RAV	7.7	5.7	5.6

with higher-order derivatives defined similarly. It can be checked that

$$\mathrm{E}\{L_t(Z;F)\} = \int L_t(z;F)F(\mathrm{d}z) = 0,$$

$$\mathrm{E}\{Q_t(Z,Z';F)\} = \int Q_t(z,z';F)F(\mathrm{d}z)F(\mathrm{d}z') = 0,$$

where $Z, Z' \overset{\text{iid}}{\sim} F$, and that for the least squares parameter (2) (Hinkley, 1977),

$$L_t(z;F) = \left\{\int x'x'^{\mathrm{T}}F(\mathrm{d}z')\right\}^{-1} x\{y - x^{\mathrm{T}}t(F)\}.$$

For statistical applications, G is replaced by the empirical distribution function $\hat{F}$ placing masses $1/n$ on the elements of a random sample $Z_1, \ldots, Z_n$ from F. Then the first-order approximation

$$t(\hat{F}) \doteq t(F) + \int L_t(z;F)\hat{F}(\mathrm{d}z)$$
$$= t(F) + n^{-1}\sum_{j=1}^{n} L_t(Z_j;F)$$

yields

$$\mathrm{E}\{t(\hat{F})\} \doteq t(F),$$
$$\mathrm{var}\{t(\hat{F})\} \doteq n^{-1}\,\mathrm{var}\{L_t(Z;F)\}$$
$$= n^{-1}\int L_t^2(z;F)F(\mathrm{d}z).$$

Higher-order approximations appear on including further terms from (3).

In the bootstrap setting, G is replaced by the empirical distribution function of a bootstrap sample, $\hat{F}^*$, and F is replaced by $\hat{F}$, giving a first-order approximation to the bootstrap statistic of the form

$$t(\hat{F}^*) \doteq t(\hat{F}) + \int L_t(z;\hat{F})\hat{F}^*(\mathrm{d}z)$$
$$= t(\hat{F}) + n^{-1}\sum_{j=1}^{n} f_j^* L_t(z_j;\hat{F}),$$

where f_j^* is the number of times that z_j appears in the bootstrap sample, and the joint distribution of $(f_1^*, \ldots, f_n^*)$ is multinomial with denominator n and mean vector $(1, \ldots, 1)$. It is easy to check that the bootstrap variance of $t(\hat{F}^*)$, conditional on the original sample $z_1, \ldots, z_n$, is approximately

$$\text{(4)} \quad \mathrm{var}^*\{t(\hat{F}^*)\} = n^{-2}\sum_{j=1}^{n} L_t^2(z_j;\hat{F}) + O_p^*(n^{-3/2}),$$

with an obvious modification for the m out of n bootstrap. The error term relates to the bootstrap approximation. The right-hand side of (4) typically underestimates the bootstrap variance (Efron and Stein, 1981).

In the case of the least squares parameter (2),

$$L_t\{(x_j,y_j);\hat{F}\} = n(X^{\mathrm{T}}X)^{-1}x_j(y_j - x_j^{\mathrm{T}}\hat{\beta}),$$

where $\beta = (X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}y$ is the ordinary least squares estimate, so, with $\hat{D} = \mathrm{diag}\{(y_j - x_j^{\mathrm{T}}\hat{\beta})^2 : j = 1, \ldots, n\}$ (Hinkley, 1977, Fox, Hinkley and Larntz, 1980),

$$\text{(5)} \quad \begin{aligned} \mathrm{var}^*\{t(\hat{F}^*)\} &= (X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}\hat{D}X(X^{\mathrm{T}}X)^{-1} \\ &\quad + O_p^*(n^{-3/2}); \end{aligned}$$

this is the sandwich variance estimate for the least squares estimator. The downward bias of $\hat{D}$ can be reduced by using modified residuals, but in any case the familiar trade-off between efficiency and robustness rears its ugly head, as mentioned above and in Hinkley and Wang (1991).

Nowhere above is it supposed that the data stem from a linear model, homoscedastic or not, but simply that the $z_j = (x_j, y_j)$ $(j = 1, \ldots, n)$ have been sampled independently at random from F. If $t(F)$ is to be useful, however, it should be roughly linear in X, and, as discussed above, sensible data analysis will ensure that this is the case by eliminating obvious discrepancies between the fitted model and the data.

5. DESIGNED EXPERIMENTS

In a designed experiment, the design matrix X is chosen in advance of observing the responses y, but the fitted model will rarely be perfect. Yet X was chosen deliberately: does it make sense to regard it as random because the model might be inadequate? In this setting, the design is an experimental ancillary (Kalbfleisch, 1975), and so is correctly treated as fixed.

Bootstrapping pairs in designed settings may lead to certain parameters being inestimable in most resamples. As an example, consider Darwin's data on self- and cross-fertilised *Zea mays* plants (Fisher, 1935, Table 1). When a model with 15 parameters for the matched pairs and a fertilisation effect is fitted, all 16 parameters are estimable only in around 9% of resamples, and the 'model-trusting' and 'model-robust' standard errors for the fertilisation effect, both equal to 1.22, substantially underestimate the corresponding bootstrap standard error, 1.66. Related discussion is given in Example 6.5 of Davison and Hinkley (1997). This example is extreme, but the repetition of rows in the resampled design matrix can greatly change its eigenvalues in many other situations, both designed and observational: bootstrapping is not a panacea. In particular, subsetting of bootstrap output may be necessary to ensure that the conclusions based on the resamples are relevant to the data actually observed.

ACKNOWLEDGEMENTS

We thank the editor for the invitation to contribute to this discussion, and the Swiss National Science Foundation for financial support.

REFERENCES

COOK, R. D. and WEISBERG, S. (1983). Diagnostics for heteroscedasticity in regression. *Biometrika* **70** 1–10. MR0742970

DAVISON, A. C. and HINKLEY, D. V. (1997). *Bootstrap Methods and Their Application*. *Cambridge Series in Statistical and Probabilistic Mathematics* **1**. Cambridge Univ. Press, Cambridge. MR1478673

EFRON, B. and STEIN, C. (1981). The jackknife estimate of variance. *Ann. Statist.* **9** 586–596. MR0615434

FERNHOLZ, L. T. (1983). *Von Mises Calculus for Statistical Functionals*. *Lecture Notes in Statistics* **19**. Springer, New York. MR0713611

FISHER, R. A. (1935). *The Design of Experiments*. Oliver and Boyd, Edinburgh.

FOX, T., HINKLEY, D. and LARNTZ, K. (1980). Jackknifing in nonlinear regression. *Technometrics* **22** 29–33. MR0559682

HAMPEL, F. R. (1974). The influence curve and its role in robust estimation. *J. Amer. Statist. Assoc.* **69** 383–393. MR0362657

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. MR0829458

HINKLEY, D. V. (1977). Jacknifing in unbalanced situations. *Technometrics* **19** 285–292. MR0458734

HINKLEY, D. V. and WANG, S. (1991). Efficiency of robust standard errors for regression coefficients. *Comm. Statist. Theory Methods* **20** 1–11. MR1114631

KALBFLEISCH, J. D. (1975). Sufficiency and conditionality. *Biometrika* **62** 251–268. MR0386075

PANARETOS, V. M., KRAUS, D. and MADDOCKS, J. H. (2010). Second-order comparison of Gaussian random functions and the geometry of DNA minicircles. *J. Amer. Statist. Assoc.* **105** 670–682. MR2724851

TUKEY, J. W. (1949). One degree of freedom for non-additivity. *Biometrics* **5** 232–242.