SORT 32 (1) January-June 2008, 77-92

An alternative analysis of variance

Nicholas T. Longford*

SNTL, Reading, United Kingdom, and Universitat Pompeu Fabra, Barcelona,

Abstract

The one-way analysis of variance is a staple of elementary statistics courses. The hypothesis test of homogeneity of the means encourages the use of the selected-model based estimators which are usually assessed without any regard for the uncertainty about the outcome of the test. We expose the weaknesses of such estimators when the uncertainty is taken into account, as it should be, and propose synthetic estimators as an alternative.

MSC: 62J10

Keywords: Mean squared error, model selection, shrinkage estimation, synthetic estimation.

1 Introduction

For most students of statistics, analysis of variance (ANOVA) is their first encounter with model-based inference. In the standard one-way setting of K groups with n_k observations each, drawn independently and at random from the respective normal distributions $\mathcal{N}(\mu_k, \sigma^2)$, $k = 1, \ldots, K$, the two contending models are

A, no constraints on the expectations μ_k, $k = 1, \ldots, K$;
B, homogeneity; $\mu_1 = \mu_2 = \ldots = \mu_K$.

* N. T. Longford is Director, SNTL Statistics Research and Consulting, Reading, UK, and Academic Visitor, Departament d'Economia i Empresa, Universitat Pompeu Fabra, Barcelona, Spain. Email: NTL@SNTL.co.uk.

He acknowledges support by the Grants SEC2003–04476, SAB2004–0190 and SEJ2006–13537 from the Spanish Ministry of Science and Technology. Professor J. Wywiał pointed out an error in an earlier version of this paper.

Received: September 2007

Accepted: November 2007

The centrepiece of a typical analysis is the test of the hypothesis that model B is valid. The possible outcomes of the test are that we have evidence against model B, that is, for rejecting the hypothesis that B is valid, or that we have failed to find such evidence. These outcomes are often interpreted as:

a, model B is not appropriate;
b, model B might be appropriate.

Thus, rejection of B is equated to ruling B out, discarding the possibility of the error of the first kind. Logical consistency is corroded further when after failing to reject B we proceed to adopt B, and act as if B were valid. Any model selection, such as one based on the likelihood ratio or an information criterion, entails similar inconsistency if the imperfection of the model selection process is ignored.

In this paper, we explore the properties of estimators based on selected models, relate their distributions to mixtures and propose, as an alternative, synthetic estimators which linearly combine the contending estimators. We focus on the standard setting of ANOVA, using elementary tools, and demonstrate that the deeply ingrained two-stage strategy of identifying a suitable model and basing all subsequent inferences on this model is not conducive to efficient estimation.

The next section introduces the setting and describes the model-selection based estimator of the expectation of a group. The following section defines synthetic estimators which combine the alternative (constituent) estimators, and identifies their principal difficulty – estimation of the weights to be accorded to the constituent estimators. This problem is addressed in Section 4. The properties of several versions of these estimators are explored by simulations in Section 5. Section 6 incorporates prior information, within the frequentist framework, about the deviation $\mu_1 - \mu$ of the expectation of the target group from the overall expectation. Section 7 applies synthesis to estimating the within-group variance σ^2. The concluding section discusses the full potential of synthetic estimators and some of its implications.

2 Estimating μ_1

There are two obvious candidates for estimating the expectation μ_1: the sample mean $\hat{\mu}_1 = \bar{y}_1$ of the n_1 observations in group 1 and $\hat{\mu}$, the sample mean of all the $n = n_1 + \cdots + n_K$ observations $\mathbf{y} = (y_{11}, y_{21}, \ldots, y_{n_K K})^\top$. We refer to them as *single-model based* estimators. Their respective distributions are

$$\hat{\mu}_1 \sim \mathcal{N}\left(\mu_1, \frac{\sigma^2}{n_1}\right)$$

$$\hat{\mu} \;\sim\; \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right).$$

We regard the mean squared error (MSE) as the arbiter of the quality of an estimator; smaller MSE (greater efficiency) is preferred. The estimators $\hat{\mu}_1$ and $\hat{\mu}$ have respective MSEs σ^2/n_1 and $\sigma^2/n + (\mu_1 - \mu)^2$. This suggests that when we entertain only $\hat{\mu}_1$ and $\hat{\mu}$ as possible estimators of μ_1 we should use $\hat{\mu}_1$ when we believe that

$$(\mu_1 - \mu)^2 \;>\; \sigma^2 g_1\,, \tag{1}$$

where $g_1 = 1/n_1 - 1/n$, and use $\hat{\mu}$ otherwise. We define g_k, $k = 2, \ldots, K$, similarly to g_1. Setting aside for the moment the fact that $\mu_1 - \mu$ is not known, the criterion based on (1) implies that our choice between models A and B should not be made by a hypothesis test in a blanket fashion for all subsequent inferences that are based on $\mathbf{y}$. For example, the expectation for a group with a small sample size n_1 may be estimated more efficiently by $\hat{\mu}$, whereas the expectation of another group, with a greater sample size n_2, may be estimated more efficiently by $\hat{\mu}_2$. This calls into question the presumption that a valid model is the ideal basis for efficient estimation. If the expectations μ_k were in a sufficiently narrow range, without all of them coinciding, $\hat{\mu}$ would be more efficient than $\hat{\mu}_1$ for estimating μ_1, even though it would be based on the invalid model B.

Let $\mathcal{I}$ be the indicator of the event that model B is selected by the established ANOVA or a similar procedure. In the conventional approach, the analysis concludes with the *selected-model based* estimator

$$\hat{\mu}_1^\dagger = (1 - \mathcal{I})\,\hat{\mu}_1 + \mathcal{I}\hat{\mu}\,, \tag{2}$$

stating that it is unbiased and has the sampling variance

$$\mathrm{var}_\dagger\left(\hat{\mu}_1^\dagger\right) \;=\; (1 - \mathcal{I})\,\mathrm{var}(\hat{\mu}_1) + \mathcal{I}\,\mathrm{var}(\hat{\mu})\,. \tag{3}$$

(Note that this 'variance' is not a constant, but a random variable with a scaled Bernoulli distribution.) That is, if A is selected, then we report $\hat{\mu}_1$ and associate it with sampling variance σ^2/n_1, and if B is selected, then we report $\hat{\mu}$ and associate it with variance σ^2/n, claiming in both cases that the estimator is unbiased. Even if σ^2 is known or its efficient estimator is substituted in (3), this conclusion is flawed, and this is easy to show by simulations. The key to such a demonstration is to apply the model selection, between the data-generation and estimation steps, in each replication. Figure 1 displays the empirical (simulated) distributions of the estimator $\hat{\mu}_1^\dagger$ and the estimator of $\sqrt{\mathrm{var}_\dagger(\hat{\mu}_1^\dagger)}$ obtained by substituting the conventional estimates for the variances on the right-hand side of (3). The diagram is based on the setting with $K = 5$ groups, each with $n_k = 7$ observations, $\mu_1 = 1$, $\mu_2 = \ldots = \mu_5 = -\frac{1}{4}$, so that $\mu = 0$, and $\sigma^2 = 1$. The size of the F-test for selecting between A and B is set to $\alpha = 0.05$.

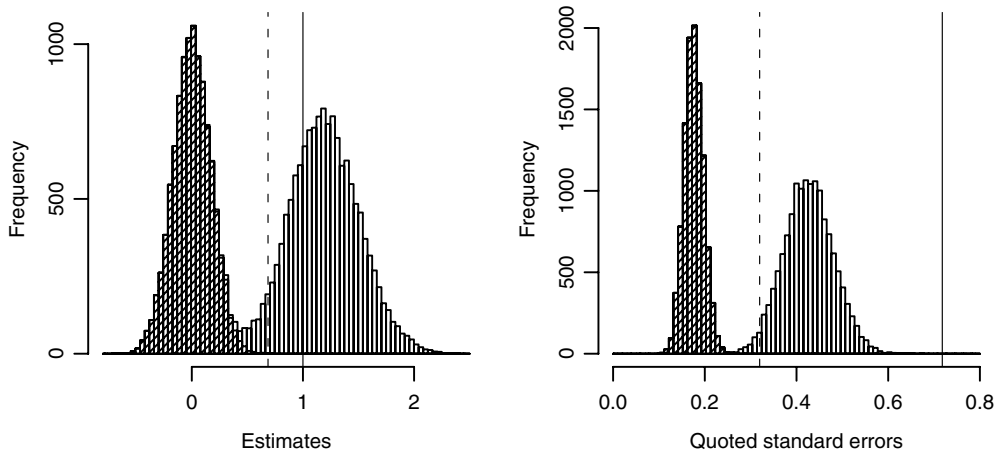

Figure 1: *The empirical distribution of* $\hat{\mu}_1^\dagger$ *and of the estimator of its standard error,* $\sqrt{\widehat{var}_\dagger(\hat{\mu}_1^\dagger)}$*, for the setting with* $\mu_1 = 1$*,* $\mu_2 = \ldots = \mu_5 = -\frac{1}{4}$ *and* $\sigma^2 = 1$*. The solid vertical lines indicate the respective targets,* $\mu_1 = 1$ *and* $MSE(\hat{\mu}_1^\dagger; \mu_1) = 0.718$*, and the vertical dashes the empirical expectations of the estimators. The shaded parts correspond to the outcomes when the null-hypothesis was not rejected. Based on 25 000 replications.*

The shaded part of the histogram of $\hat{\mu}_1^\dagger$ in the left-hand panel represents the replications in which the null hypothesis was not rejected and $\hat{\mu}$ was evaluated. The vertical lines mark the expectation (dashes, $\mathrm{E}(\hat{\mu}_1^\dagger) = 0.686$) and the target (solid line, $\mu_1 = 1$). The estimator $\hat{\mu}_1^\dagger$ is biased and distinctly not normally distributed; in fact, its density is bimodal, with a small value at its expectation. The conditional distributions $(\hat{\mu}_1 \,|\, \mathcal{I} = 1)$ and $(\hat{\mu} \,|\, \mathcal{I} = 0)$ have a narrow overlap; 0.45, a value near the point where the two conditional densities intersect, is the 0.46-percentile of $(\hat{\mu} \,|\, \mathcal{I} = 0)$ and the 99.23-percentile of $(\hat{\mu} \,|\, \mathcal{I} = 1)$.

The empirical distribution of the estimator of the standard error of $\hat{\mu}_1^\dagger$ based on (3) is drawn in the right-hand panel. Every one of the 25 000 generated estimates falls short of the empirical root-MSE, which is equal to 0.718; the empirical mean of these estimates is 0.320. Clearly, any reference to a χ^2 distribution for the estimated sampling variance would be grossly erroneous. In summary, the model selection is a highly non-ignorable process.

Some of the findings from this example can be confirmed analytically and generalised. For instance, unless the appropriate model is selected with certainty or $\mu_1 = \mu$, estimator $\hat{\mu}_1^\dagger$ is biased:

$$\begin{aligned} \mathrm{E}\left(\hat{\mu}_1^\dagger\right) &= p_\mathrm{A}\,\mathrm{E}\left(\hat{\mu}_1 \,|\, \mathcal{I} = 0\right) + p_\mathrm{B}\,\mathrm{E}\left(\hat{\mu} \,|\, \mathcal{I} = 1\right) \\ &= \mu_1 + p_\mathrm{B}\left\{\mathrm{E}\left(\hat{\mu} \,|\, \mathcal{I} = 1\right) - \mathrm{E}\left(\hat{\mu}_1 \,|\, \mathcal{I} = 1\right)\right\}, \end{aligned}$$

where $p_A = P(\mathcal{I} = 0)$ and $p_B = 1 - p_A$. Note that $E(\hat{\mu} \mid \mathcal{I} = 0) \neq \mu_1$ and $E(\hat{\mu}_1 \mid \mathcal{I} = 1) \neq \mu$, because the sample means $\hat{\mu}_1$ and $\hat{\mu}$ are correlated with $\mathcal{I}$.

The variance formula (3) remains incorrect even when $\mathcal{I}$ is replaced by its expectation p_B:

$$\begin{aligned} \mathrm{MSE}\left(\hat{\mu}_1^{\dagger}; \mu_1\right) &= p_A \mathrm{MSE}(\hat{\mu}_1 \mid \mathcal{I} = 0) + p_B \mathrm{MSE}(\hat{\mu} \mid \mathcal{I} = 1) \\ &= p_A \mathrm{var}(\hat{\mu}_1 \mid \mathcal{I} = 0) + p_B \mathrm{var}(\hat{\mu} \mid \mathcal{I} = 1) \\ &\quad + p_A \{E(\hat{\mu}_1 \mid \mathcal{I} = 0) - \mu_1\}^2 + p_B \{E(\hat{\mu} \mid \mathcal{I} = 1) - \mu_1\}^2 . \end{aligned}$$

The last two terms are positive, unless $p_A p_B = 0$ or $\mu_1 = \mu$. Their total differs from the squared bias of (3) in estimating the MSE, because $\mathrm{var}(\hat{\mu}_1 \mid \mathcal{I} = 0) \neq \mathrm{var}(\hat{\mu}_1)$ and $\mathrm{var}(\hat{\mu} \mid \mathcal{I} = 1) \neq \mathrm{var}(\hat{\mu})$. The conditioning in these two variances is essential; $\mathcal{I}$ is independent of neither $\hat{\mu}_1$ nor $\hat{\mu}$.

3 Synthetic estimation

As an alternative to the selected-model based estimator $\hat{\mu}_1^{\dagger}$, we consider the convex combination

$$\tilde{\mu}_1 = (1 - b_1)\hat{\mu}_1 + b_1\hat{\mu} ,$$

and set the coefficient b_1 (or, in a more rigorous notation, b_{μ_1}) so as to minimise $M(b_1) = \mathrm{MSE}(\tilde{\mu}_1; \mu_1)$. After substituting $\mathrm{cov}(\hat{\mu}_1, \hat{\mu}) = \sigma^2/n$ and rearranging the terms, we obtain

$$\begin{aligned} M(b_1) &= (1 - b_1)^2 \frac{\sigma^2}{n_1} + b_1^2 \frac{\sigma^2}{n} + 2b_1(1 - b_1)\frac{\sigma^2}{n} + b_1^2(\mu_1 - \mu)^2 \\ &= b_1^2 \left\{ g_1 \sigma^2 + (\mu_1 - \mu)^2 \right\} - 2b_1 g_1 \sigma^2 + \frac{\sigma^2}{n_1} . \end{aligned}$$

This function attains its minimum at

$$b_1^* = \frac{g_1}{g_1 + \dfrac{(\mu_1 - \mu)^2}{\sigma^2}} .$$

If b_1^* were established with precision, the *ideal* synthetic estimator

$$\tilde{\mu}_1(b_1^*) = (1 - b_1^*)\hat{\mu}_1 + b_1^*\hat{\mu} \tag{4}$$

would be more efficient than both $\hat{\mu}_1$ and $\hat{\mu}$ because these constituent estimators are equal to $\tilde{\mu}_1(0)$ and $\tilde{\mu}_1(1)$, whereas $b_1^* \in (0, 1)$. The MSE of $\tilde{\mu}_1(b_1^*)$ is $\sigma^2(1/n_1 - b_1^* g_1)$. For the setting of Figure 1, this corresponds to root-MSE 0.362.

In practice, b_1^* has to be estimated, so the estimator $\tilde{\mu}_1(\hat{b}_1^*)$ is less efficient than $\tilde{\mu}_1(b_1^*)$, and may be less efficient than $\hat{\mu}_1$ or $\hat{\mu}$. In model selection we incur a similar loss due to uncertainty about the validity of model B. However, even if model B could be ruled out, $\hat{\mu}$ may be more efficient than $\hat{\mu}_1$, because the squared bias of $\hat{\mu}$ may be smaller than the variance reduction $g_1\sigma^2$; see (1). With errors of both kinds in model selection, $\hat{\mu}_1^\dagger$ is likely to have MSE greater than $\min\{\sigma^2/n_1, \sigma^2/n + (\mu_1 - \mu)^2\}$. The key questions therefore are how much efficiency is lost and how much are the respective MSEs underestimated in the two approaches.

For selected-model based estimation, we can, in principle, choose any selection process (binary statistic) $\mathcal{I}$. Similarly, for synthetic estimation, we can choose any estimator of the coefficient b_1^*. Except for Hjort and Claeskens (2003), statistical literature offers little rationale for choosing a different model-selection process $\mathcal{I}$ for one target (μ_1) than for another (μ_2 or σ^2). In contrast, in synthetic estimation, the ideal coefficients b_k^* for the distinct expectations μ_k coincide only when $n_1 = \ldots = n_K$.

4 Estimating b_1^*

To address the problem without any distractions, we assume first that σ^2 is known. The obvious way of estimating b_1^* is naively, by

$$\hat{b}_1 = \frac{g_1}{g_1 + \dfrac{(\hat{\mu}_1 - \hat{\mu})^2}{\sigma^2}}\,. \tag{5}$$

As $\mathrm{E}(\hat{\mu}_1 - \hat{\mu})^2 = (\mu_1 - \mu)^2 + g_1\sigma^2$, $1/\hat{b}_1$ overestimates $1/b_1^*$. This does not imply that $\hat{b}_1$ underestimates b_1^*, but cases when $\mathrm{E}(\hat{b}_1) < b_1^*$ arise only in some esoteric situations. If we adjust $(\hat{\mu}_1 - \hat{\mu})^2$ for its bias in estimating $(\mu_1 - \mu)^2$, we obtain the estimator $\hat{b}_1 = g_1\sigma^2/(\hat{\mu}_1 - \hat{\mu})^2$. Its values can exceed unity, in which case $\hat{\mu}_1$ would be associated with negative 'weight' $1 - \hat{b}_1$. This estimator of b_1^* is therefore not suitable. In any case, unbiased estimation of $1/b_1^*$ does not lead to unbiased estimation of b_1^*.

When there are several groups and we have no external information about the relative sizes of the expectations μ_k or deviations $\mu_k - \mu$, we might estimate $(\mu_1 - \mu)^2$ by an (approximately) unbiased estimator of the group-level variance

$$\sigma_{\mathrm{B}}^2 \;=\; \frac{1}{K}\sum_{k=1}^{K}(\mu_k - \mu)^2\,.$$

Such an estimator would be biased for $(\mu_1 - \mu)^2$, but its MSE may be much smaller, because it draws on the information in all the groups. The variance σ_{B}^2 can be estimated without bias by moment matching. We define a statistic quadratic in the outcomes $\mathbf{y}$, such as

$$S_{\mathrm{B}} = \frac{1}{K}\sum_{k=1}^{K}(\hat{\mu}_k - \hat{\mu})^2\,,$$

and match it with its expectation, $\mathrm{E}(S_{\mathrm{B}}) = \bar{g}\,\sigma^2 + \sigma_{\mathrm{B}}^2$, where $\bar{g} = \frac{1}{K}(g_1 + \cdots + g_K)$. This yields the estimator

$$\hat{\sigma}_{\mathrm{B}}^2 = S_{\mathrm{B}} - \bar{g}\,\hat{\sigma}^2\,,$$

where $\hat{\sigma}^2$ is an unbiased estimator of σ^2. The estimator $\hat{\sigma}_{\mathrm{B}}^2$ can be interpreted as adjusting S_{B} for its bias in estimating σ_{B}^2. It can attain negative values, and these should be truncated at zero, even though the resulting estimator is biased. The pooled within-group variance estimator is the obvious choice for $\hat{\sigma}^2$.

Yet another way of estimating b_1^* is to consider the consequences of its under- and overestimation. By underestimating b_1^* we tend to err on the side of attaching greater weight to $\hat{\mu}_1$, reducing the bias but increasing the variance of $\tilde{\mu}_1$, which nevertheless has the upper bound σ^2/n_1. In contrast, by overestimating b_1^* we tend to err on the side of attaching greater weight to $\hat{\mu}$, the bias of which does not have an *a priori* bound. This suggests that underestimating b_1^*, and hence overestimating $(\mu_1 - \mu)^2$ or σ_{B}^2, is preferable. This has the added advantage that the greater denominator in $\hat{b}_1$ brings about greater stability. Instead of overestimating the denominator of b_1^*, we may simply use the estimator $\tilde{\mu}_1(r_1\hat{b}_1)$ with an *a priori* set constant factor $r_1 < 1$.

The synthetic estimator can be described as a *shrinkage* estimator, pulling the unbiased estimator $\hat{\mu}_1$ toward the more stable but biased estimator $\hat{\mu}$. The estimator $\tilde{\mu}_1(r_1\hat{b}_1)$ thus involves reduced shrinkage. When b_1^* is estimated using $\hat{\sigma}_{\mathrm{B}}^2$, $\tilde{\mu}_1$ has the same form as empirical Bayes estimators (Efron and Morris, 1972), which borrow strength across the groups. The only essential difference arises due to the different meaning of the between-group variance. It relates to a finite set of groups in our fixed-effects and to an infinite population of groups in the random-effects setting. Through the involvement of $\hat{\sigma}_{\mathrm{B}}^2$ in $\hat{b}_1$, the synthetic estimator $\tilde{\mu}_1(\hat{b}_1)$ can borrow strength across the groups, or exploit their similarity, even without assuming a superpopulation of groups.

5 Empirical assessment

An analytical expression can be derived for neither $\mathrm{MSE}(\hat{\mu}_1^{\dagger};\mu_1)$ nor $\mathrm{MSE}\{\tilde{\mu}_1(\hat{b}_1);\mu_1\}$, and so these quantities can only be estimated by simulations. However, setting up such

simulations is easy, not much more difficult than programming a single replication of the data-generating, model selection and estimation processes.

We lose no generality by reducing our attention to parameter values $\mu = 0$, $\sigma^2 = 1$ and, owing to symmetry, to positive deviations $\Delta = \mu_1 - \mu$. As the benchmark, we use the ANOVA estimator $\hat{\mu}_1^\dagger$ based on the F-test with the conventional size $\alpha = 0.05$. Instead of the outcomes y_{ik} it suffices to generate the sample means $\hat{\mu}_k$ as random draws from $\mathcal{N}(\mu_k, \sigma^2/n_1)$ and the within-group corrected sums of squares $s_k^2 = \sum_i (y_{ik} - \hat{\mu}_k)^2$ independently from suitably scaled χ^2 distributions:

$$(n_k - 1)\frac{s_k^2}{\sigma^2} \sim \chi^2_{n_k - 1}\,.$$

For a fixed set of sample sizes $n_1, \ldots, n_K$, we execute 5 000 replications of the ANOVA and synthetic estimators for each value of Δ on a grid of 76 equidistant points in the closed interval $[0, 3]$. For each estimator, we evaluate its empirical bias and MSE. For ANOVA estimators, the probability of rejecting model B can be obtained as the tail of the appropriate non-central F-distribution.

Figure 2 displays the root-MSEs of several estimators of μ_1, as functions of the (absolute) deviation Δ. It exposes the gross inefficiency of the selected-model based estimator $\hat{\mu}_1^\dagger$ for a wide range of deviations Δ. The synthetic estimator $\tilde{\mu}_1(\hat{b}_1)$ with naively estimated b_1^* is more efficient than the estimator with $\hat{b}_1 = g_1/(g_1 + \hat{\sigma}_{\mathrm{B}}^2)$, except for very small values of Δ. In our setting, $\sigma_{\mathrm{B}}^2 = \frac{1}{4}$ is much smaller than $(\mu_1 - \mu)^2 = 1$, so $\hat{\sigma}_{\mathrm{B}}^2$ is

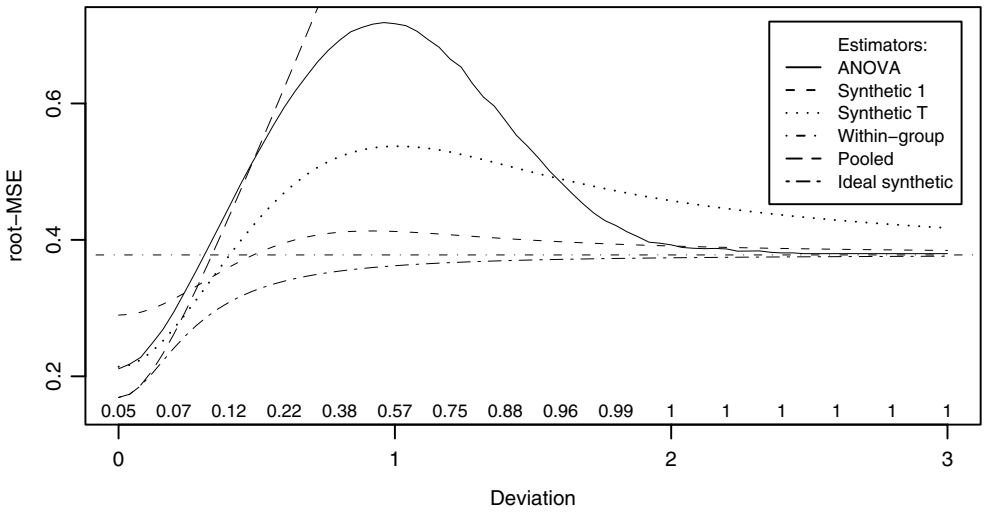

Figure 2: *The root-MSEs of the estimators of μ_1 : $\hat{\mu}_1^\dagger$ given by (2) (solid line), $\tilde{\mu}_1(\hat{b}_1)$ with $\hat{b}_1$ given by (5) (short dashes), $\tilde{\mu}_1(\hat{b}_1)$ with $\hat{b}_1 = g_1/(g_1 + \hat{\sigma}_{\mathrm{B}}^2)$ (dots), $\hat{\mu}_1$ (dots and short dashes), $\hat{\mu}$ (long dashes) and the ideal synthetic estimator $\tilde{\mu}_1(b_1^*)$ given by (4) (dots and long dashes). The figures at the bottom of the diagram are the probabilities of rejecting model B by the standard F-test.*

a poor estimator of $(\mu_1 - \mu)^2$. For large values of Δ, $\hat{\mu}_1^\dagger$ is marginally more efficient than $\tilde{\mu}_1(\hat{b}_1)$, and both of them approach the efficiency of $\hat{\mu}_1$, without surpassing it. In fact, for $\Delta > 2.5$, the null hypothesis is rejected and $\hat{\mu}_1^\dagger = \hat{\mu}_1$ in all 5 000 replications.

The power of the F-test, that is, the probabilities $\beta(\Delta)$ of rejecting model B, are given at the bottom of the diagram for $\Delta = 0, 0.2, \ldots, 3$. The model-selection estimator $\hat{\mu}_1^\dagger$ is less efficient than $\hat{\mu}_1$ when $\Delta = 1.4$ and $\beta(1.4) = 0.88$ (root-MSE 0.57) and when $\Delta = 0.4$ and $\beta(0.4) = 0.12$ (root-MSE 0.45). Thus, the probability of the 'correct' model choice is a poor proxy for the efficiency of $\hat{\mu}_1^\dagger$.

Figure 3 contains the plot of the root-MSEs of the synthetic estimators of μ_1 with reduced shrinkage, $\tilde{\mu}_1(r\hat{b}_1)$, with the naive estimator $\hat{b}_1$ of b_1^* and $0.6 \le r \le 1.0$. It shows that by reducing the shrinkage we improve estimation around $\Delta = 1$, where the MSE is largest, at the expense of losing some efficiency for small values of Δ, where the MSE is smallest. The root-MSE of the ideal synthetic estimator $\tilde{\mu}(b_1^*)$ is smaller than for any synthetic estimator with estimated b_1^*.

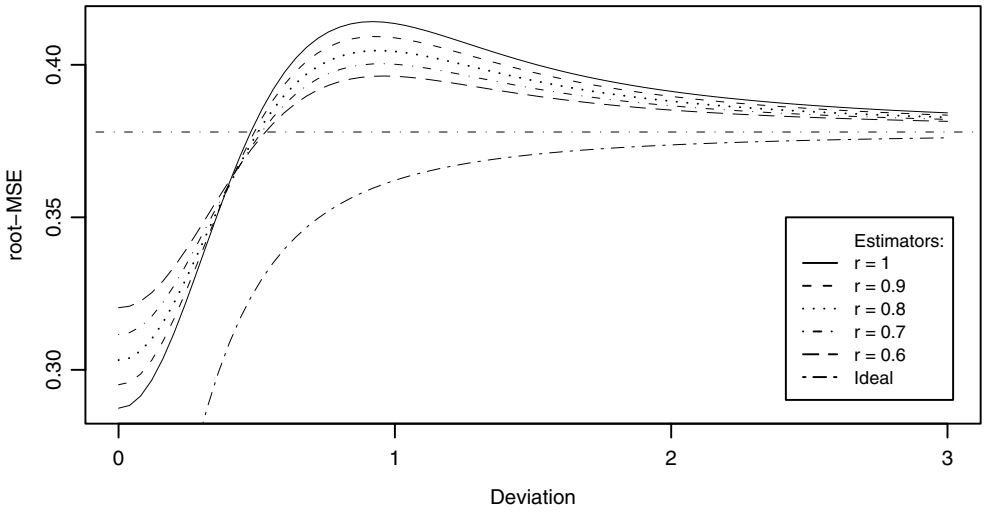

Figure 3: *The root-MSEs of the estimators of $\tilde{\mu}_1(r\hat{b}_1)$ with naively estimated b_1^* and $r = 0.6$, 0.7, 0.8, 0.9 and 1.0.*

We conclude this section with Figure 4 which presents the empirical distributions of $\tilde{\mu}_1(\hat{b}_1)$ using the naive estimator $\hat{b}_1$ given by (5), and of the naive estimator of its root-MSE, $\hat{\sigma}^2(1/n_1 - g_1\hat{b}_1)$. The same setting for the simulation and the same layout for the diagram are used as in Figure 1. The distributions in Figure 4 are unimodal and do not deviate substantially from normality and a scaled χ^2 distribution with many degrees of freedom. The estimator $\tilde{\mu}_1$ has a small bias but, more importantly, its root-MSE is much smaller than for $\hat{\mu}_1^\dagger$. Because the uncertainty about b_1^* is ignored, the root-MSE of $\tilde{\mu}_1$ is underestimated, by 0.063, but much less blatantly than it is for $\hat{\mu}_1^\dagger$. Admittedly, this

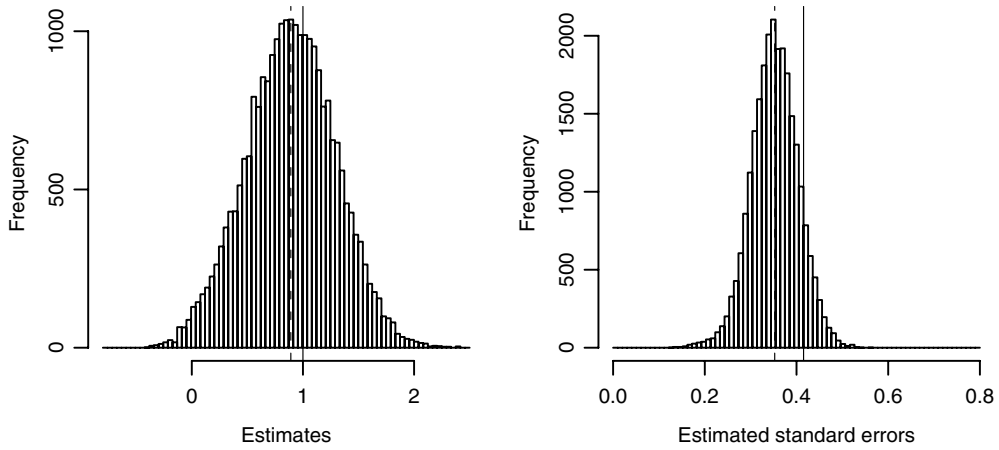

Figure 4: *The empirical distributions of $\tilde{\mu}_1(\hat{b}_1)$ and of the naive estimator of its standard error for the same setting as in Figure 1. Based on 25 000 replications.*

comparison of $\tilde{\mu}_1$ and $\hat{\mu}_1^{\dagger}$ is somewhat unfair, because the setting is least favourable for the latter. However, for $\mu_1 = 1$, $\tilde{\mu}_1$ also has (nearly) the highest root-MSE, see Figure 2.

6 Prior information about Δ

An unsatisfactory property of all of the estimators explored so far is that they fail to outperform $\hat{\mu}_1$ uniformly. In this section we make amends on this count, although for that we require an upper bound on the absolute deviations $|\Delta|$. Suppose we are confident that $|\Delta|$ is smaller than an *a priori* set value Δ_*. We apply the synthetic estimator that is optimal when $\Delta = \Delta_*$, and assess its properties when in fact $|\Delta| \leq \Delta_*$. Let $B_\Delta = g_1/(g_1 + \Delta^2/\sigma^2)$ and $B_* = B_{\Delta_*}$. The MSE of the estimator $\tilde{\mu}_1(B_*)$ is

$$B_*^2 \left(g_1 \sigma^2 + \Delta^2\right) - 2B_* g_1 \sigma^2 + \frac{\sigma^2}{n_1}.$$

For fixed Δ_*, this is an increasing (linear) function of Δ^2, so it attains its maximum in $[0, \Delta_*^2]$ for $\Delta^2 = \Delta_*^2$. At this point, $\tilde{\mu}_1(B_*)$ coincides with the ideal synthetic estimator, and so it is efficient. Therefore, among the synthetic estimators it is the *minimax* estimator; any other synthetic estimator has a greater maximum MSE within $\Delta \in (-\Delta_*, \Delta_*)$. Figure 5 provides an illustration using our earlier setting with $K = 5$ groups, $n_1 = \ldots = n_K = 7$, and $\sigma^2 = 1$, and with Δ_* set to 1.5. The estimator $\tilde{\mu}_1(B_{1.5})$ is uniformly more efficient than $\hat{\mu}_1$ for $\mu_1 \in (\mu - \Delta_*, \mu + \Delta_*)$, and is less efficient than $\hat{\mu}$ only when $|\Delta|$ is very small.

If we are justified to set Δ_* lower, say, to 0.75, then the estimator $\tilde{\mu}_1(B_{0.75})$ is uniformly more efficient than $\tilde{\mu}_1(B_{1.5})$ while $|\Delta| < 0.75$. It is more efficient for slightly greater

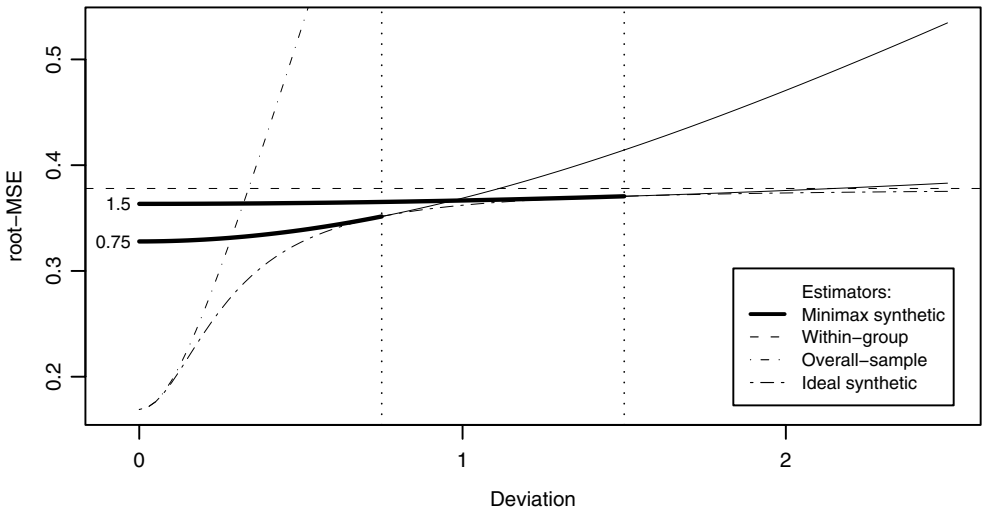

Figure 5: *The root-MSEs of the minimax synthetic estimators. The same setting is used as in Figures 2 and 3. The root-MSEs of the minimax synthetic estimators are drawn by thick solid lines in the ranges of deviations Δ assumed to be plausible and by thin solid lines outside them. The upper bound Δ_* is indicated at the left-hand margin and by the vertical dots.*

absolute deviations $|\Delta|$, but its MSE increases more steeply than for $\Delta_* = 1.5$. Using a value of Δ_* that is too small has more severe consequences than making a conservative choice of a greater value of Δ_*.

7 Estimating σ^2

Synthetic estimation can be applied also to σ^2, by combining the two candidate estimators

$$\hat{\sigma}^2_{\mathrm{A}} = \frac{1}{n-K} \sum_{k=1}^{K} \sum_{i=1}^{n_k} (y_{ik} - \hat{\mu}_k)^2$$

$$\hat{\sigma}^2_{\mathrm{B}} = \frac{1}{n-1} \sum_{k=1}^{K} \sum_{i=1}^{n_k} (y_{ik} - \hat{\mu})^2 \ ,$$

based on the respective models A and B. These two estimators are connected by the orthogonal decomposition

$$(n-1)\,\hat{\sigma}^2_{\mathrm{B}} = (n-K)\,\hat{\sigma}^2_{\mathrm{A}} + \sum_{k=1}^{K} n_k\,(\hat{\mu}_k - \hat{\mu})^2 \ . \qquad (6)$$

Their distributions are scaled χ^2,

$$(n-K)\,\frac{\hat{\sigma}_{\mathrm{A}}^2}{\sigma^2} \sim \chi^2_{n-K,0}$$

$$(n-1)\,\frac{\hat{\sigma}_{\mathrm{B}}^2}{\sigma^2} \sim \chi^2_{n-1,\gamma}\,,$$

where the subscripts indicate the number of degrees of freedom and the noncentrality, and $\gamma = \sum_k n_k(\mu_k - \mu)^2/\sigma^2$. The orthogonal decomposition in (6) implies that

$$\mathrm{cov}\left(\hat{\sigma}_{\mathrm{A}}^2, \hat{\sigma}_{\mathrm{B}}^2\right) = \frac{n-K}{n-1}\,\mathrm{var}\left(\hat{\sigma}_{\mathrm{A}}^2\right) = \frac{2\sigma^4}{n-1}\,.$$

Further,

$$\mathrm{var}\left(\hat{\sigma}_{\mathrm{A}}^2\right) = \frac{2\sigma^4}{n-K}$$

$$\mathrm{var}\left(\hat{\sigma}_{\mathrm{B}}^2\right) = \frac{2\sigma^4}{n-1} + \frac{4\gamma\sigma^4}{(n-1)^2}\,.$$

The convex combination

$$\tilde{\sigma}^2 = (1-b_{\mathrm{W}})\,\hat{\sigma}_{\mathrm{A}}^2 + b_{\mathrm{W}}\,\hat{\sigma}_{\mathrm{B}}^2$$

attains its minimum MSE in estimating σ^2 for

$$\begin{aligned} b_{\mathrm{W}}^* &= \frac{\mathrm{var}\left(\hat{\sigma}_{\mathrm{A}}^2\right) - \mathrm{cov}\left(\hat{\sigma}_{\mathrm{A}}^2, \hat{\sigma}_{\mathrm{B}}^2\right)}{\mathrm{var}\left(\hat{\sigma}_{\mathrm{A}}^2\right) + \mathrm{var}\left(\hat{\sigma}_{\mathrm{B}}^2\right) - 2\mathrm{cov}\left(\hat{\sigma}_{\mathrm{A}}^2, \hat{\sigma}_{\mathrm{B}}^2\right) + \left\{\mathrm{E}\left(\hat{\sigma}_{\mathrm{A}}^2\right) - \mathrm{E}\left(\hat{\sigma}_{\mathrm{B}}^2\right)\right\}^2} \\ &= \frac{\dfrac{2\sigma^4}{n-K} - \dfrac{2\sigma^4}{n-1}}{\dfrac{2\sigma^4}{n-K} - \dfrac{2\sigma^4}{n-1} + \dfrac{4\gamma\sigma^4}{(n-1)^2} + \dfrac{\gamma^2\sigma^4}{(n-1)^2}} \\ &= \frac{2q}{2q+4\gamma+\gamma^2}\,, \end{aligned} \tag{7}$$

where $q = (K-1)(n-1)/(n-K)$.

In a balanced design, with $n_k \equiv n/K$, and many groups K, so that $n \gg n_k$,

$$b_{\mathrm{W}}^{*} \doteq \frac{2K}{2K + 4n\dfrac{\sigma_{\mathrm{B}}^2}{\sigma^2} + n^2\dfrac{\sigma_{\mathrm{B}}^4}{\sigma^4}}$$

$$< \frac{1}{1 + 2n_k\dfrac{\sigma_{\mathrm{B}}^2}{\sigma^2}},$$

after approximating q by K and γ by $n\sigma_{\mathrm{B}}^2/\sigma^2$. This is smaller than the coefficient b_1^* for estimating μ_1, when Δ^2 in its denominator is replaced by σ_{B}^2. This comparison of b_1^* and b_{W}^* can be motivated as follows. For estimating μ_1, we seek to exploit the $n - n_1$ observations from outside group 1, which could increase the effective number of observations from n_1 to n. In many settings this amounts to a several-fold increase. In contrast, for estimating σ^2, we seek to engage $K - 1$ degrees of freedom in addition to $n - K$. Usually this amounts to only a modest increase. For example, in the setting studied earlier, $n_1 = 7$, $n = 35$ and $K = 5$, synthetic estimation of μ_1 seeks to exploit the information in 28 observations in addition to the seven in group 1, whereas estimation of σ^2 draws on only four degrees of freedom in addition to $n - K = 30$. We should be disposed much less favourably toward submodel B for estimating σ^2 than for μ_1, because synthesis has a very modest potential for gain in precision given the threat of bias associated with the contentious $K - 1$ degrees of freedom. This reinforces our earlier conclusion that selecting the same model for estimating several targets associated with a dataset need not be optimal for all of them. Combining the single-model based estimators using the same set of weights for several targets, as is done by Bayesian model averaging (Kass and Raftery, 1995; Hoeting *et al.*, 1999) may also be suboptimal – estimators have to be averaged with target-specific weights.

8 Discussion and conclusions

We introduced synthetic estimators for the ANOVA setting and showed that their weaknesses (low efficiency for some parameter values) are not as pronounced as for the selected-model based estimators. The principle applied, of combining alternative estimators instead of selecting one of them, is applicable much more generally; see Longford (2003) for its application to ordinary regression and Longford (2007) to estimation of the MSE in a setting similar to ANOVA.

A synthetic estimator $\tilde{\theta}$ which has constituent estimators $\hat{\theta}_{\mathrm{A}}$ and $\hat{\theta}_{\mathrm{B}}$, based on respective models A and B, can be paired with a model-selection based estimator $\hat{\theta}$ which selects between the models A and B and the respective single-model based

estimators $\hat{\theta}_A$ and $\hat{\theta}_B$. The synthetic estimator $\tilde{\theta}$ has a greater potential than the model-selection based estimator $\hat{\theta}$, because its ideal version $\tilde{\mu}(b^*)$ is more efficient than the ideal version of $\hat{\theta}$, equal to the single-model based estimator with the smaller MSE. The synthetic estimator can be extended to more than two constituent estimators; see Longford (2005, Chap. 11) for details.

Our conclusion calls for a revision of the dictum that identification of a (parsimonious) valid model is an essential prerequisite for efficient estimation. We have shown that estimators based on *some* invalid models, namely submodels of valid models, are efficient for *some* targets. The presented perspective calls into question the effectiveness of all model-selection procedures, including those based on information criteria, which seek to maximise the probability of the appropriate choice without assessing the consequences of an inappropriate model choice, however small its probability may be.

Diagnostic procedures can be considered similarly. Since their application may lead to a revision of the model or of the dataset, the estimator involved is a mixture of the (numerous) estimators that are specific to the possible outcomes of the diagnostic process. This should not be regarded as an unqualified discouragement from applying diagnostic procedures. However, when the outcome of the procedure is subject to uncertainty, the estimator that incorporates the application of a procedure has a distribution different from its version that skips the application. If we are certain that a particular diagnostic procedure is unnecessary, the estimator is more efficient than if we were not certain and applied it, even if the actual outcome of the procedure is negative, failing to find any contradiction with the model assumptions. Applying a comprehensive battery of diagnostic procedures is not a good practice because it inflates the chances of an inappropriate revision of the model or dataset following a false positive finding. Instead, procedures should be carefully selected to respond to the analyst's uncertainties as assessed prior to data inspection.

Model selection and diagnostic procedures do not come for free in inference, and we do not act with scientific integrity when we quote the (conditional) properties of estimators given their selection, ignoring the uncertainty associated with the application of such procedures. The lack of integrity is exacerbated when we do not inform about the details of the procedures applied, or when they are applied informally and afford no simple description. Addressing these problems is essential for integrity in the conduct of statistical inference.

Unbiased estimation of MSEs of synthetic estimators is an open problem, just as it is for selected-model based estimators and estimators that are applied following diagnostic and data-cleaning procedures. Exploration by simulations is the only solution available at present even for the simplest settings. Methods based on expansions, which are valid only asymptotically, cannot resolve the issues we have highlighted in this paper because model uncertainty and the trade-off between sampling variance and bias are essentially small-sample issues.

References

Efron, B., and Morris, C. N. (1972). Limiting the risk of Bayes and empirical Bayes estimators – Part II: The empirical Bayes case. *Journal of the American Statistical Association*, 67, 130-139.

Hjort, N. L., and Claeskens, G. (2003). Frequentist model average estimators, *Journal of the American Statistical Association*. 98, 879-899.

Hoeting, J. A., Madigan, D., Raftery, A. E., and Volinsky C. T. (1999). Bayesian model averaging: a tutorial. *Statistical Science*, 14, 382-401.

Kass, R. E., and Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, 90, 773-795.

Longford, N. T. (2003). An alternative to model selection in ordinary regression. *Computing and Statistics*, 13, 67-80.

Longford, N. T. (2005). *Missing Data and Small-Area Estimation. Modern Analytical Equipment for the Survey Statistician.* New York, NY: Springer-Verlag.

Longford, N. T. (2007). On standard errors of model-based small-area estimators. *Survey Methodology*, 33, 69-79.