Robust Analysis of the Random Model and Weighted Least Squares Regression

BRUCE M. HILL

DEPARTMENT OF STATISTICS
UNIVERSITY OF MICHIGAN
ANN ARBOR, MICHIGAN

1. Introduction

In this article some general approaches to robustness of inference and decision making are formulated. The ideas are not meant to be tied down to any particular statistical ideology and, so far as possible, are intended to be robust to such ideologies as well as towards alternative formulations of the model, likelihood function, prior distribution, and utility function. In general, the spirit of the approach is one of learning to recognize and take advantage of those situations where robustness does exist and to be aware of the alternatives when it does not. The ideas will be conveyed by consideration of an example, which is analyzed at progressively more realistic levels. The example, which is based upon the so-called random model, is of interest in and of itself and is increasingly being recognized as a fundamental data structure in many different areas of statistics. Yet even in its simplest form there is still substantial disagreement as to the interpretation of data arising from such a model. After first analyzing the random model under conventional normality assumptions, we propose a form of analysis which is more robust both to the form of the prior distribution and to departures from normality. Finally, we suggest a random model for weighted regression which leads to new types of estimators for a population regression when the data consists of samples from a number of "blocks."

This article is meant to embody, in microcosm, the full process of statistical inference, beginning with model formulation in its simplest setting,

ISBN 0-12-416550-8

proceeding to obtain various insights under conventional simplifying assumptions, and finally to consider questions of robustness, model reformulation, and applications where full mathematical analysis is not possible. It is hoped that the approach to model building and robustness will be adequately conveyed by the manner in which the examples are treated. In the final section some comments are made which may help to highlight the respects in which the approach differs from others. In addition, interspersed throughout the article, comments are made relating the general philosophy discussed in Section 5 to the specific analyses proposed. Although the problems we are dealing with are, unfortunately, sufficiently complex so that there will inevitably be loose ends, the reader should view the philosophy in Section 5 as an ideal which we are hoping to approximate within the article. Many readers may find it helpful to read Section 5 at this point. Since some aspects of our approach may seem unusual to econometricians, we shall now briefly examine those aspects most likely to cause difficulty.

Section 2 deals entirely with the one-way random model analysis of variance. The analysis of variance is often viewed as a rather trivial special case of regression, so it may seem strange that much emphasis should be placed upon it. However, one of the main purposes of this article is to reveal the complex and subtle issues that arise in the analysis of even simple random models. If we were to turn immediately to a realistic random regression model (such as that in Section 4), then so many other issues would intrude themselves that the first set of issues relating to the general case of random models would be masked. Thus we feel that a much deeper understanding of random model regression will arise if we first carefully explore the relatively simple analysis of variance situation. We hope that the reader will bear with us in this. For those who prefer to think entirely in terms of regression, we remark that much of the analysis of Sections 2 and 3 can be carried over to random model linear regression for which the different groups have the same known slope and the intercepts are random effects. We do not do so because it would only complicate the notation without adding additional insight. Furthermore, in Section 4 we deal with the more general case in which both intercepts and slopes are random effects.

The next source of possible difficulty concerns our use of Bayesian methods of inference and decision making. From our point of view a posterior distribution is meant to approximate the probabilistic knowledge about parameters (or models) that one views as appropriate *a posteriori*. In conjunction with a specific loss function such a posterior distribution yields an optimal Bayes decision rule. It is well known that under modest regularity conditions such decision rules are also admissible. Some methods of Bayesian analysis are, however, easy to misinterpret when looked at from a non-Bayesian point of view. For example, in Section 2 we derive the posterior distribution of an effect μ_i, given the ratio of the between variance to the within variance τ^2. In doing so we do not mean to imply that τ^2 is known. Rather this is simply a convenient intermediate step in the derivation of the unconditional posterior distribution of μ_i which reflects our uncertainty about τ^2. The unconditional analysis is, of course, more difficult and is the stage at which interesting robustness questions arise. In a similar way a Bayesian analysis may derive the posterior distribution for a parameter of relatively little interest as an intermediate step in order to obtain insight and to obtain other results of greater interest. Thus in Section 2 the posterior distribution of the parameter $\gamma = \ln(1 + J\tau^2)$ is of a much simpler form than that of τ^2 or of $\theta(\tau^2) = J\tau^2/(1 + J\tau^2)$ and allows us to determine the important posterior moments of τ^2 and $\theta(\tau^2)$. Generally speaking, those unfamiliar with Bayesian philosophy and methods will have no difficulty if they view a posterior distribution as an ordinary probability distribution appropriate for parameters and conditional upon the data.

Finally, how does the analysis of this article pertain to the evaluation of econometric models? We have attempted to illustrate in the context of random model analysis of variance and regression some of the considerations that we view as essential in the choice of a model. We have suggested a variety of choices to be made: conditioning upon both mean square error between (MSB) and mean square error within (MSW) versus conditioning only upon MSB/MSW as data, *a priori* independence of σ^2 from σ_α^2 versus *a priori* independence of σ^2 from τ^2, use of normal theory versus use of more general distributions, use of $H(x)$ (to be defined in Section 3) versus ordinary least squares theory. In each case we have attempted to make the nature of the choice clear and to offer both mathematical and philosophical guidelines to aid in making such choices. From our point of view realistic problems of statistics are sufficiently complex and subtle that the choice of model and analysis should be based as closely as possible upon such guidelines.

2. One-Way Random Model under Conventional Assumptions

This simple model is increasingly being recognized as a fundamental structure for a great variety of statistical problems ranging from sample surveys to time series. In Section 4 a realistic example of such usage will be presented, but before doing so it will be necessary to examine the model in its simplest form, for even here it is widely misunderstood.

Our starting point will be the conception of the model, which we shall view in a more general sense than is customary in statistics. By the model we shall mean the additive data structure $y_{ij} = \mu_i + \varepsilon_{ij}$. Although we shall not as yet make specific distributional assumptions as to the μ_i and ε_{ij}, we shall think of μ_i as a characteristic of the ith sample category and of ε_{ij} as either the jth measurement error on the ith category, or alternatively as the jth unit error within the ith category. Furthermore, we shall regard the μ_i as though they were a sample from some large finite, or infinite, population, with distribution, say, $H((x-\mu)/\sigma_\alpha)$, i.e., with μ as location parameter and σ_α as scale parameter. We do not assume that the μ_i are necessarily physically sampled from such a population, but merely that our knowledge of them is such that we wish to act as if this were the case. Such an interpretation is especially natural from a subjective Bayesian viewpoint but could also be appropriate from other viewpoints, for example, the pigeonhole model of Cornfield & Tukey (1956). We shall assume that the ε_{ij} have expectation 0, variance σ^2, and distribution $G(\varepsilon/\sigma)$. Furthermore, we assume the ε_{ij} are exchangeable within a given category and, conditional upon σ, are independent for different categories.

The parameters that will be of interest are the individual (realized) μ_i, $\sigma_\alpha^2, \sigma^2$, and $\tau^2 = \sigma_\alpha^2/\sigma^2$. We shall assume $\mu = E[\mu_i | \mu, \sigma_\alpha]$, that $\alpha_i = \mu_i - \mu$, $\sigma_\alpha^2 = \mathrm{Var}(\mu_i | \mu, \sigma_\alpha)$, and, of course, $\sigma^2 = \mathrm{Var}(\varepsilon_{ij} | \sigma)$. At first we shall make conventional normality assumptions, i.e., $H(\cdot)$ and $G(\cdot)$ are both normal distributions. This is done for two reasons. First, in many situations such assumptions are approximately satisfied (perhaps with transformed data). Second, as we shall see in Section 3, the mode of analysis for the normal case provides valuable insights as to the more general case and serves as a takeoff point from which modifications can be made.

We proceed with the conventional assumptions for the one-way unbalanced random model, i.e., $y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$, $i = 1, \ldots, I$, $j = 1, \ldots, J_i$, where $\alpha_i \sim N(0, \sigma_\alpha^2)$, $\varepsilon_{ij} \sim N(0, \sigma^2)$, and $\{\alpha_i, \varepsilon_{ij}\}$ are conditionally independent, given $\sigma_\alpha^2, \sigma^2$. First, consider inference about the μ_i, given $\tau^2 = \sigma_\alpha^2/\sigma^2$. Since $y_{i\cdot} = J_i^{-1} \sum_j y_{ij} \sim N(\mu_i, \sigma^2/J_i)$, given μ_i and σ^2, while $\mu_i \sim N(\mu, \sigma_\alpha^2)$, it follows from elementary Bayesian analysis that, conditional upon μ, σ^2, and σ_α^2, the posterior distribution of the μ_i is

$$\mu_i \sim N(\theta_i y_{i\cdot} + (1-\theta_i)\mu, [1/\sigma_\alpha^2 + J_i/\sigma^2]^{-1}),$$

where $\theta_i = \theta_i(\tau^2) = J_i\tau^2/(1 + J_i\tau^2)$, with the μ_i conditionally independent, *a posteriori*. Since from almost any point of view,

$$\hat{\mu}(\tau^2) = \left[\sum_{i=1}^{I} J_i y_{i\cdot}/(1 + J_i\tau^2)\right] \Bigg/ \left[\sum_{i=1}^{I} J_i/(1 + J_i\tau^2)\right]$$

is the natural estimator of μ, given τ^2, it follows that

$$\hat{\mu}_i(\tau^2) = \theta_i y_{i\cdot} + (1-\theta_i)\hat{\mu}(\tau^2)$$

is an equally natural estimator of μ_i. In particular, it is the Bayes posterior expectation of μ_i, given τ^2, when the prior distribution of μ is diffuse (Hill, 1965, 1977) and is admissible with respect to squared error loss.

Up to this point the analysis has been simple, and hopefully uncontroversial, at least under the assumed model. However, complications crop up as soon as we contemplate inference about τ^2. It has been observed by the author that inference about other parameters in the conventional random model is of an elementary nature, given τ^2, and that all difficulties in the analysis are due solely to the complexity of the information contained in the data concerning τ^2 (Hill, 1965). Let us consider inference about τ^2 in the balanced case $J_i = J$, the unbalanced case having been examined in Hill (1965). The simplest and, as will be shown later, the most robust form of inference about τ^2 is that obtained by conditioning only upon the data MSB/MSW, where

$$\text{MSB} = J\sum_i (y_{i\cdot} - y_{\cdot\cdot})^2/(I-1) \qquad \text{and} \qquad \text{MSW} = \sum_{i,j} (y_{ij} - y_{i\cdot})^2/I(J-1).$$

Since it is the vector $(y_{\cdot\cdot}, \text{MSB}, \text{MSW})$ that forms a sufficient statistic for $(\mu, \sigma_\alpha^2, \sigma^2)$, something is, in principle, lost by conditioning only upon MSB/MSW; this will be examined later. For the present, we shall base our analysis upon the fact that, given τ^2, MSB/MSW $\sim (1 + J\tau^2)F_{m,n}$, where $m = I - 1$, $n = I(J-1)$, and $F_{m,n}$ is a random variable having the F distribution with m and n degrees of freedom. Letting $\gamma = \ln(1 + J\tau^2)$, $\hat{\gamma} = \ln(\text{MSB/MSW})$, the likelihood function for γ based upon the data $\hat{\gamma}$ is $l(\gamma) \propto p_{n,m}(\gamma - \hat{\gamma})$, for $\gamma \geq 0$, where $p_{n,m}(\cdot)$ is the density function of $\ln F_{n,m}$. Although $\gamma \geq 0$, $\hat{\gamma}$ can have any real value. Since $\hat{\gamma}$ serves as a location parameter for $p_{n,m}(\cdot)$, the data merely shift this density to have "center" $\hat{\gamma}$, and then condition $\gamma \geq 0$ is used to truncate the shifted function from below at 0 in order to obtain the likelihood function $l(\gamma)$. In fact the mode of $p_{n,m}(\cdot)$ is 0, so that when $\hat{\gamma} < 0$, $l(\gamma)$ is a portion of the upper tail of $p_{n,m}(\gamma - \hat{\gamma})$ and is monotonically decreasing for $\gamma > 0$; when $\hat{\gamma}$ is 0, $l(\gamma)$ is the portion of $p_{n,m}(\gamma - \hat{\gamma})$ to the right of its mode; and when $\hat{\gamma} \gg 0$, $l(\gamma)$ is approximately $p_{n,m}(\gamma - \hat{\gamma})$ itself since in this instance the truncation cuts off only a small portion of the lower tail. Based upon this likelihood function we see that the nature of inference about γ, and hence τ^2, depends crucially upon the sign of $\hat{\gamma}$, i.e., upon whether or not MSB/MSW ≥ 1. When $\hat{\gamma} > 0$, the maximum of $l(\gamma)$ occurs at $\hat{\gamma}$ and the degree of concentration of $l(\gamma)$ near $\hat{\gamma}$ depends upon the size of m. When $\hat{\gamma} \leq 0$, the maximum of $l(\gamma)$ occurs at 0 and $l(\gamma)$ is monotonically decreasing for $\gamma > 0$. Since $p_{n,m}(x)$ is proportional to

$$e^{nx/2}\Big/\left[1 + \frac{n}{m}e^x\right]^{(m+n)/2},$$

with logarithmic derivative

$$\tfrac{1}{2}mn[(1 - e^x)/(m + ne^x)],$$

the logarithmic derivative can be seen to be negative for $x > 0$ and decreases to the limit $(-m/2)$ as $x \to \infty$. For $\hat{\gamma} < 0$, $l(\gamma)$ declines most sharply from its maximum at 0 when $\hat{\gamma} = -\infty$, and in this case $l(\gamma)$ is proportional to the exponential function $e^{-m\gamma/2}$ for $\gamma > 0$. When $m = I - 1$ is not large, it is clear on the basis of this analysis that the classical conclusion $\tau^2 = 0$ or τ^2 very "small" is not warranted even when the F ratio MSB/MSW is extremely small. On the other hand, when $\hat{\gamma}$ is sufficiently positive, for example, so that only .01 of the lower tail of $p_{n,m}(\gamma - \hat{\gamma})$ is below 0, then, to a good approximation, one may ignore the truncation and simply treat $\gamma - \hat{\gamma}$ as having the distribution of $\ln F_{n,m}$. In this case classical confidence intervals for τ^2 based upon MSB/MSW are in substantial agreement with the above analysis, which may be regarded either as Bayesian with a diffuse prior distribution for τ^2, or as a likelihood analysis ignoring the prior distribution for τ^2, or a fiducial solution. (Unfortunately, however, it is not the fiducial solution obtained by Fisher (1935), who would censor rather than truncate $p_{n,m}(\gamma - \hat{\gamma})$ at 0, lumping the area to the left of 0 at 0. When $\hat{\gamma} \gg 0$, this barely modifies the previous analysis, but, of course, for "very negative" $\hat{\gamma}$, a high fiducial probability may be attached to $\gamma = 0$ or $\tau^2 = 0$, quite contrary to the previous analysis.) When $\hat{\gamma} < 0$, there seems to be no consensus as to the appropriate interpretation from the Neyman–Pearson school; some recommend "confidence" intervals containing negative values although the parameter is nonnegative (Scheffé, 1961, p. 230), others recommend that γ and τ^2 should be taken as 0, and still others recommend rejection of the model. We shall return to the question of model rejection later, merely noting for the present that, when τ^2 is small, negative $\hat{\gamma}$ should arise nearly 50% of the time when the model is "true," so that unless $\hat{\gamma} \ll 0$, there should be little reason to reject the model merely because $\hat{\gamma}$ is negative.

Let us consider more carefully the nature of Bayesian inference about τ^2, which will then lead us to our first confrontation with the robustness problem. Our previous analysis was based upon MSB/MSW alone and led to a simple and seemingly quite natural interpretation of that data. Even in this case one might wish to multiply $l(\gamma)$ by a suitably chosen prior density for γ (as induced by a prior density for τ^2), but, unless there were compelling reasons for such a choice, it would be adequate for many purposes to employ $l(\gamma)$ as an approximate posterior density for anyone whose initial opinions about τ^2 were not overly strong. In this case one might take $\hat{\gamma}$ as the "estimate" of γ when $\hat{\gamma}$ is sufficiently positive and take $2/m$ (the expectation of the limiting exponential distribution) as the "estimate" when $\hat{\gamma}$ is sufficiently negative. Intermediate values of $\hat{\gamma}$ would require numerical integration of $l(\gamma)$, when viewed as an approximate posterior density for γ. Correspondingly, the conditional expectations of $\theta(\tau^2)$ are

$$E[\theta(\tau^2)|\hat{\gamma}] \approx 1 - [n/(n-2)]\text{MSW/MSB} \qquad \text{for} \quad \hat{\gamma} \gg 0$$

and

$$E[\theta(\tau^2)|\hat{\gamma}] \approx 2/(m+2) \qquad \text{for} \quad \hat{\gamma} \ll 0,$$

which may then be used to approximate $E[\mu_i|\text{data}]$.

However, within a Bayesian framework it would be necessary to consider what is lost (and what is gained) by basing inference about τ^2 on MSB/MSW alone, as compared to use of the sufficient statistic ($y_{\cdot\cdot}$, MSB, MSW). So far as $y_{\cdot\cdot}$ is concerned, it seems clear that only very strong prior knowledge of μ and its relationship to σ_α^2 and σ^2 could seriously modify the previous analysis based upon MSB/MSW (Hill, 1965, 1977). The situation with regard to (MSB, MSW) is, however, more complicated. Some insight can be gained by considering a Bayesian analysis with, say, $\phi(\mu, \sigma_\alpha^2, \sigma^2) \propto \phi(\sigma_\alpha^2, \sigma^2)$ as prior density, i.e., with diffuse prior knowledge of μ, which yields the same result as does conditioning upon (MSB, MSW) alone. The posterior density for the variance components is then

$$\phi''(\sigma_\alpha^2, \sigma^2) \propto \phi(\sigma_\alpha^2, \sigma^2)(\sigma^2)^{-n/2} \times \exp\left[-\frac{n\text{MSW}}{2\sigma^2}\right](\sigma^2 + J\sigma_\alpha^2)^{-m/2} \exp\left[-\frac{m\text{MSB}}{2(\sigma^2 + J\sigma_\alpha^2)}\right].$$

These results are somewhat sensitive to the characteristics of the prior density. There are two cases of primary interest. First, there is the case in which τ^2 is *a priori* independent of σ^2, which typically arises when the ε_{ij} are unit errors as opposed to measurement errors. In this case results are very similar to those derived above. Thus if $P(\tau^2, \sigma^2)$ is the initial prior distribution for τ^2 and σ^2 and if $P(\tau^2, \sigma^2) = P_1(\tau^2)P_2(\sigma^2)$, then the posterior density is

$$P''(\tau^2, \sigma^2) \propto P_2(\sigma^2)(\sigma^2)^{-(m+n)/2} \times \exp\left[-\frac{n\text{MSW}}{2\sigma^2}\right] P_1(\tau^2)[1 + J\tau^2]^{-m/2} \exp\left[-\frac{m\text{MSB}}{2\sigma^2(1 + J\tau^2)}\right].$$

Because of the strength of information about σ^2 contained in MSW, the choice of $P_2(\cdot)$ ordinarily has only a minor effect. It is convenient and, when n is large, more or less innocuous to take $P_2(\sigma^2) \propto (\sigma^2)^{-1}$. Then the marginal posterior density for τ^2 is

$$P''(\tau^2) \propto P_1(\tau^2)(1 + J\tau^2)^{n/2} \Bigg/ \left[1 + \frac{n}{m}\frac{\text{MSW}}{\text{MSB}}(1 + J\tau^2)\right]^{(m+n)/2}, \qquad \tau^2 \geq 0,$$

so that, apart from $P_1(\tau^2)$, (MSW/MSB)$(1 + J\tau^2) \sim CF_{n+2,m-2}$, truncated from below at MSW/MSB. This result is in near agreement with the previous analysis based upon MSB/MSW as data as well as the results of Hill (1965, p. 818), where $C = m(n + 2)/n(m - 2)$. Note that, as MSB/MSW goes to 0, $P''(\tau^2)$ tends to $P_1(\tau^2)(1 + J\tau^2)^{-m/2}$ for $\tau^2 \geq 0$. In fact, letting $\delta^2 = \sigma^2/\text{MSW}$, the joint posterior density is

$$P''(\tau^2, \delta^2) \propto P_1(\tau^2)[1 + J\tau^2]^{-m/2}(\delta^2)^{-((m+n)/2)-1} \times \exp\left[-\frac{1}{2\delta^2}\left(n + \frac{m\text{MSB}}{\text{MSW}(1 + J\tau^2)}\right)\right].$$

Thus as MSB/MSW goes to 0, the limiting posterior distribution of (τ^2, δ^2) has density

$$P''(\tau^2, \delta^2) \propto P_1(\tau^2)[1 + J\tau^2]^{-m/2}(\delta^2)^{-((m+n)/2)-1} \exp[-n/2\delta^2],$$

so that τ^2 and δ^2 are independent, and σ^2/MSW is distributed like $n/\chi^2_{(m+n)}$, where $\chi^2_{(j)}$ is a random variable having the chi-square distribution with j degrees of freedom. Of course this limiting distribution should be interpreted only as an approximation, valid when MSB/MSW is sufficiently small. On the whole we may summarize this analysis as providing support for the robustness of our previous inference about τ^2, based only upon MSB/MSW, within the broader model under which τ^2 and σ^2 are *a priori* independent.

The situation is, however, somewhat different when σ_α^2 and σ^2 are *a priori* independent, which is a generally plausible model when σ^2 is the variance of a technical error (for example, of a measuring instrument) rather than that of a unit error. Suppose, for example, that the prior density for $(\sigma_\alpha^2, \sigma^2)$ is $\phi(\sigma_\alpha^2, \sigma^2) \propto \phi_1(\sigma_\alpha^2)\phi_2(\sigma^2)$, where as before, it is ordinarily innocuous to take $\phi_2(\sigma^2) \propto (\sigma^2)^{-1}$. Doing so, the posterior density for (τ^2, δ^2) becomes

$$\phi''(\tau^2, \delta^2) \propto \phi_1(\text{MSW}\delta^2\tau^2)(\delta^2)^{-(m+n)/2} \exp[-n/2\delta^2](1 + J\tau^2)^{-m/2} \times \exp[-m\text{MSB}/2(1 + J\tau^2)\delta^2\text{MSW}].$$

For insight, let $\phi_1(\sigma_\alpha^2)$ have the inverted gamma form, i.e.,

$$\phi_1(\sigma_\alpha^2) \propto (\sigma_\alpha^2)^{-(\gamma_\alpha/2)-1}\exp[-C_\alpha/2\sigma_\alpha^2],$$

where $\gamma_\alpha > 0$ and $C_\alpha > 0$. (As shown in Hill, 1965, p. 811, the case $\gamma_\alpha = C_\alpha = 0$ leads to nonsensical results.) Then as MSB/MSW goes to 0, with MSW held fixed, the posterior density tends to

$$\phi''(\tau^2, \delta^2) \propto (\delta^2)^{-[(m+n+\gamma_\alpha)/2]-1} \exp[-n/2\delta^2](1 + J\tau^2)^{-m/2}(\tau^2)^{-(\gamma_\alpha/2)-1} \times \exp[-C_\alpha/2\text{MSW}\delta^2\tau^2].$$

Hence under this model there is no limiting distribution as MSB/MSW goes to 0 since the limit is different depending upon whether MSB goes to 0, with MSW fixed, or MSW goes to infinity, with MSB fixed. Furthermore, in the latter case the marginal posterior density of τ^2 becomes the improper density $(\tau^2)^{-(\gamma_\alpha/2)-1}(1 + J\tau^2)^{-m/2}$, which must be interpreted as degenerate at 0. Although such degeneracy has been derived here for the special inverted gamma prior distribution of σ_α^2, the result is fairly general, usually holding when σ^2 and σ_α^2 are independent *a priori*. The basic reason for this behavior is that when MSW grows large, the posterior distribution of σ_α^2 tends to the prior distribution of σ_α^2 (Hill, 1965, 1967, 1970, 1975a), while σ^2/MSW has a limiting distribution. Hence typically under these conditions $\tau^2 = \sigma_\alpha^2/\sigma^2$ converges to 0 as MSW goes to infinity. As we saw before, however, when τ^2 and σ^2 are independent *a priori*, then τ^2 has a limiting distribution even as MSB/MSW goes to 0. Correspondingly, in this case it can be shown that σ_α^2/MSW has a limiting distribution as MSW goes to infinity. See Culver (1971) for a general discussion.

The main point to be made is that one may be faced with a nonrobust situation, so that a choice must be carefully made with respect to the appropriate assumptions. This can be illustrated by some numerical examples.

EXAMPLE 1. $I = 3$, $J = 3$, $\gamma_\alpha = 2$, $C_\alpha = 4$, MSB $= 5$, MSW $= 1$. The exact $E(\theta\,|\,\text{data})$ is .84, while the approximation based upon the data MSB/MSW alone is $E(\theta\,|\,\hat{\gamma}) \approx 1 - [n/(n-2)]\text{MSW/MSB} = .70$.

EXAMPLE 2. $I = 20$, $J = 10$, $\gamma_\alpha = 8$, $C_\alpha = 10$, MSB $= 10$, MSW $= 1$. The exact $E(\theta\,|\,\text{data})$ is .91, and the approximation yields .89.

EXAMPLE 3. $I = 5$, $J = 2$, $\gamma_\alpha = 8$, $C_\alpha = 1$, MSB $= 10$, MSW $= 1$. The exact $E(\theta\,|\,\text{data})$ is .08, while the approximation yields .83.

Note that the approximation based upon MSB/MSW alone works well except in Example 3. This stems from the small sample sizes combined with the relatively strong prior input. Thus in Example 3, *a priori* $E(\sigma_\alpha^2) = C_\alpha/(\gamma_\alpha - 2) = \frac{1}{6}$, $\text{Var}(\sigma_\alpha^2) = 2C_\alpha/(\gamma_\alpha - 2)^2(\gamma_\alpha - 4) = \frac{1}{72}$, so there was initially strong opinion that σ_α^2 was "small." Also, the use of $P_2(\sigma^2) \propto (\sigma^2)^{-1}$ is not quite so innocuous here because σ^2 is measured with only five degrees of freedom. It is interesting to compare Example 1 with Example 3. In Example 1 the F ratio is only 5, and the degrees of freedom differ only slightly from those of Example 3, so the huge difference in the two $E(\theta\,|\,\text{data})$ for these examples presumably stems from the much sharper prior opinion about σ_α^2 in Example 3, which overcomes the greater F ratio of 10 and reduces the posterior expectation of θ to the surprisingly small value .08. This dramatically illustrates the nonrobustness of this type of Bayesian analysis based upon the total data and especially its sensitivity to the choice of γ_α and C_α. Finally, in Example 2, where degrees of freedom are ample, we see that life is pleasant in any case.

Now let us relate the above analysis to our general philosophy regarding model building and robustness. We began with conventional normality assumptions for the one-way random model and proceeded to obtain certain estimators and insights as to inference about the parameters. We observed that by conditioning only upon MSB/MSW, we could obtain a "marginal" likelihood function for τ^2 and that this likelihood function allowed a simple and natural form of inference about τ^2. Furthermore, at least when MSB/MSW was sufficiently large and with diffuse prior knowledge of τ^2, Bayesian, fiducial, and Neyman–Pearson approaches led to very nearly the same inference about τ^2. In this context the Bayesian approach merely added a few qualifications as to the appropriateness of such inference, namely, that on the one hand, when MSB/MSW is not large, a truncation from below at zero is crucial with regard to γ, τ^2, or σ_α^2, and on the other hand, particularly when MSB/MSW and m are small, that it may be necessary to assess prior distributions very carefully. Our next observation was that the simple inference based upon MSB/MSW alone was quite robust within the class of prior distributions for which σ^2 is *a priori* independent of τ^2. Thus, provided that one can convince oneself and/or others of the appropriateness of such a class of prior distributions, the inferential problem is quite easy. On the other hand, as both the numerical examples and the limiting analysis as MSW grows large indicate, such an analysis is anything but robust for the class of prior distributions in which σ^2 is *a priori* independent of σ_α^2. Thus if such priors are thought to be relevant, then there is no robust form of inference, and certain hard decisions must be faced.

In this situation there are a few considerations that may be helpful. First, it is sometimes possible by considering potential extreme data (before actually observing it) to decide roughly how one would like to react. Thus as MSW goes to infinity, with τ^2 *a priori* independent of σ^2, σ_α^2/MSW has a limiting distribution, so that under this model a very large MSW gives evidence that σ_α^2 is large; whereas, when σ_α^2 is a *priori* independent of σ^2, a very large MSW provides negligible information about σ_α^2. In some examples the one type of behavior will seem more appropriate, and in other examples, the reverse. Of course another device for choosing between the two types of prior distribution is simply to assess, as best one can, the source of one's prior knowledge about the parameters and choose accordingly. In situations such as this, neither the Bayesian nor any other approach is terribly helpful, but the Bayesian approach at least faces the problem openly, while for other approaches, which ignore the existence of subjective prior knowledge, the distinction between the two types of prior independence is totally irrelevant.

We conclude this section by returning to the question of model rejection. When either $\hat{\gamma}$ is extremely negative, or if several different experiments have led to a preponderance of negative $\hat{\gamma}$, then it is natural to consider alternative models. The model that is most compelling to me for explaining such data is one in which the ε_{ij} within a given category are equally negatively correlated (Hill, 1967, p. 1395, 1970, p. 33). In the extreme case where the negative correlation is $-(J-1)^{-1}$, then $\varepsilon_{i.} = 0$ with probability 1, and MSB $\sim J\sigma_\alpha^2\chi^2(m)/m$, so that no contradictory data can arise. The physical significance of such models is discussed in Hill (1967, p. 1397).

3. Random Model without Normality

Suppose now that normality for the α_i and ε_{ij} is dropped and replaced by the assumption that α_i has distribution $H(\alpha/\sigma_\alpha)$ and ε_{ij} has distribution $G(\varepsilon/\sigma)$, with $E(\alpha_i) = E(\varepsilon_{ij}) = 0$, $\mathrm{Var}(\alpha_i) = \sigma_\alpha^2$, $\mathrm{Var}(\varepsilon_{ij}) = \sigma^2$, and the customary independence assumptions are retained.

Inference about τ^2 can still be based upon MSB/MSW alone, much as before. Thus MSB/MSW can be written as

$$\frac{J\sum_i[(Z_{i.} - Z_{..}) + \tau(X_i - X_.)]^2/(I-1)}{\sum_{i,j}(Z_{ij} - Z_{i.})^2/I(J-1)}$$

where $X_i = \alpha_i/\sigma_\alpha$ and $Z_{ij} = \varepsilon_{ij}/\sigma$, so that the distribution of MSB/MSW depends only upon τ^2. If n is moderately large, then $\sum_{i,j}(Z_{ij} - Z_{i.})^2/n$ should be approximately unity, while MSB, as a sample variance of $f_i = Z_{i.} + \tau X_i$, should approximate $\sigma^2 J\,\mathrm{Var}(f_i) = J(\sigma_\alpha^2 + \sigma^2/J) = \sigma^2 + J\sigma_\alpha^2$. In fact, as a first approximation one might take MSB/MSW $\sim (1 + J\tau^2)F_{m',n'}$, where the degrees of freedom m' and n' could be chosen to yield the right first and second moments for MSB/MSW, depending on the "true" distributions $H(\cdot)$ and $G(\cdot)$. To the extent that such approximations are suitable, which can be investigated both theoretically and by Monte Carlo methods, then inference about τ^2 based upon MSB/MSW will be just as before, with only the degrees of freedom altered. Furthermore, the sensitivity of m' and n' to the choice of $H(\cdot)$ and $G(\cdot)$ can also be investigated by the same methods. When m is moderately large, hopefully inference will be robust for various *a priori* plausible $H(\cdot)$ and $G(\cdot)$.

Turn now to inference about the μ_i, given τ^2, under our more general assumptions. As before, we can argue that given μ, σ^2, and σ_α^2, the density $\sigma_\alpha^{-1}H'((\mu_i - \mu)/\sigma_\alpha)$ can be viewed as a prior distribution for μ_i, while given

σ^2 and μ_i, the function $\Pi_{j=1}^J G'((y_{ij} - \mu_i)/\sigma)$, can be viewed as a conditional likelihood function for μ_i based upon the data $y_{i1}, \ldots, y_{iJ}$. For simplicity suppose first that the ε_{ij} are normally distributed as before, but allow $H(\cdot)$ to be arbitrary. Then conditional upon μ, σ^2, and $\sigma_\alpha^2, H'((\mu_i - \mu)/\sigma_\alpha)$ serves as prior density for μ_i, while $y_{i\cdot} \sim N(\mu_i, \sigma^2/J)$, given μ_i, serves as a measurement for μ_i, which yields the conditional likelihood function $\exp[-(J/2\sigma^2)(y_{i\cdot} - \mu_i)^2]$. Consequently, the conditional posterior density for μ_i is proportional to

$$H'((\mu_i - \mu)/\sigma_\alpha) \exp[-(J/2\sigma^2)(y_{i\cdot} - \mu_i)^2],$$

given μ, σ^2, and σ_α^2.

The behavior of $\hat{\mu}_i \equiv E[\mu_i | \mu, \sigma_\alpha^2, \sigma^2, y_{i1}, \ldots, y_{iJ}]$ will now depend upon the form of $H(\cdot)$. However, although precise analysis is complex, some aspects are intuitively clear. When $H(\cdot)$ is also normal, we saw earlier that $\hat{\mu}_i$ is a weighted average of μ and $y_{i\cdot}$ with weights not depending upon the data. Generally speaking, for "thicker tailed" (than normal) $H(\cdot)$, $\hat{\mu}_i$ will tend to be closer to $y_{i\cdot}$ than in the normal case, while for "thinner tailed" $H(\cdot)$, it will tend to be closer to μ. For any specified function $H(\cdot)$ the exact value can be obtained either analytically or by numerical analysis. For example, if $H(\cdot)$ were double exponential, we would obtain

$$\exp\{-|\mu_i - \mu|/\sigma_\alpha\} \exp[-(J/2\sigma^2)(y_{i\cdot} - \mu_i)^2],$$

while if $H(\cdot)$ were a Cauchy distribution, the corresponding posterior distribution would be of the form

$$(1 + (\mu_i - \mu)^2/\sigma_\alpha^2)^{-1} \exp[-(J/2\sigma^2)(y_{i\cdot} - \mu_i)^2].$$

In the latter case, if μ is sufficiently far out in a tail of $N(y_{i\cdot}, \sigma^2/J)$, then the maximum-likelihood estimate $y_{i\cdot}$ will be approximately $\hat{\mu}_i$. In the former case one anticipates that $\hat{\mu}_i$ will be closer to $y_{i\cdot}$ than when $H(\cdot)$ is normal, but not so close as when $H(\cdot)$ is Cauchy. Through experience it should not be difficult to obtain clear insights as to the effect of altering $H(\cdot)$ upon the posterior distribution of μ_i, given μ, σ_α^2, and σ^2. Note that when $G(\cdot)$ is not normal, an appropriate analysis of this type can be obtained by choosing some function of $y_{i1}, \ldots, y_{iJ}$, say T_i, carrying most if not all of the information about μ_i and employing the distribution of T_i, given μ_i and σ^2, as the likelihood function for μ_i, in place of that for $y_{i\cdot}$, which we used when $G(\cdot)$ was normal. When J is moderately large then of course such a likelihood function will be approximately of normal form, as a consequence of standard theorems on the asymptotic behavior of the likelihood function. However, in this case the curvature of the likelihood function at its maximum will ordinarily depend upon the data, so that no fixed weights independent of the data can be employed, even if $H(\cdot)$ were of normal form. Finally, note that inspection of $y_{ij} - y_{i\cdot}$ and of $y_{i\cdot}$ can lead to intelligent choices as to the appropriate forms of $G(\cdot)$ and $H(\cdot)$, respectively.

Now suppose that we have chosen, on the basis of such considerations, weights θ_i, so that given μ and τ^2,

$$\hat{\mu}_i = \theta_i T_i + (1 - \theta_i)\mu,$$

where T_i is the selected statistic carrying high information about μ_i. Here θ_i may depend upon the data and τ^2. It is now necessary to estimate μ. To illustrate how this might be done, suppose, for example, that $G(\cdot)$ is normal and $H(\cdot)$ is double exponential. Then it would be natural to base inference about μ on $y_{i\cdot} = \mu_i + \varepsilon_{i\cdot}$. Since $H(\cdot)$ has thicker than normal tails, a first approximation (particularly suitable when J is large) is to pretend that the $y_{i\cdot}$ are a sample from $H(\cdot)$ and take $\hat{\mu}$ to be an estimator appropriate for $H(\cdot)$, i.e., $\hat{\mu} = \text{median}\{y_{i\cdot}\}$, for the double exponential. We thus obtain as an approximation in this case an estimator of the form

$$E\{\mu_i | \text{data}, \tau^2\} \approx \theta_i y_{i\cdot} + (1 - \theta_i)\hat{\mu},$$

where we anticipate that $\theta_i > J\tau^2/(1 + J\tau^2)$. Based upon our posterior distribution for τ^2 as derived earlier, we can then obtain an overall estimator for μ_i. Robustness can be examined by altering $H(\cdot)$ and $G(\cdot)$, calculating the appropriate θ_i and $\hat{\mu}$, and noting the magnitude of change in the estimator. If such change is substantial, then of course one must make some choice as to the appropriate $G(\cdot)$ and $H(\cdot)$, guided by inspection of the data.

The approach to nonnormality that we have suggested here is a tentative one in which, following the general philosophy in Section 5, we have attempted to base inference as far as possible on those assumptions in which we have most confidence. Thus inference about τ^2 was based upon a relatively simple approximation to the distribution of MSB/MSW rather than upon all the data. Similarly, by using judgment in the choice of the statistic T_i, which we regard as carrying most of the information about μ_i, we were able to gain some insight about the appropriate modification of the posterior distribution of μ_i when $G(\cdot)$ and $H(\cdot)$ are not of normal form.

4. Random Model Weighted Regression

There is an interesting and important application of the above ideas to random model weighted regressions that arise in sample surveys. Suppose there is a finite population of units and that the least squares linear regression of the dependent variable Y on the independent variable X is of interest. If there are N units in the population with (X_i, Y_i) being the values of the variables for the ith such unit, then we wish to draw inference about the two constants A and B for which $\sigma^2 = \sum_{i=1}^{N} \varepsilon_i^2/N$ is minimal, where $\varepsilon_i = Y_i - (A + BX_i)$, $i = 1, \ldots, N$. Suppose, however, that for convenience or economic reasons the sampling by means of which we hope to estimate A and B takes place within well-defined blocks (some of which may be unsampled) rather than from the population as a whole. For example, we may be interested in the relationship between measures of household income and expenditure in the state of Michigan as a whole, but our sampling may be performed within counties, or perhaps within a county we sample households from within apartment buildings, and so on. The relationship between Y and X for particular counties, or for particular apartment buildings, may or may not be of interest in its own right, but the question that we shall consider concerns how such data can be used to draw inference about the relationship in the population.

To attempt to answer this question, suppose first that the population is physically or conceptually partitioned into M "blocks," with known $S_i \geq 2$ units in the ith block, $\sum_{i=1}^{M} S_i = N$. Let (X_{ij}, Y_{ij}), $i = 1, \ldots, M, j = 1, \ldots, S_i$, label the partitioned variables, and let A_i, B_i, and σ_i^2 be defined for the ith block just as A, B, and σ^2 were for the population. We shall assume that within the population as a whole the least squares regression line is such that knowledge that $X = x$ for a unit carries negligible information about the associated deviation $\varepsilon = Y - (A + Bx)$ from the line. To be specific, we assume, for all x, that $E[Y|X = x, A, B] = A + Bx$, $\mathrm{Var}[Y|X = x, A, B, \sigma^2] = \sigma^2$. We shall make a similar assumption for the regression line within each block. Heuristically, such assumptions insure that deviations from regression lines are viewed as scattered about zero with a variance that does not depend upon the value of the independent variable. The appropriateness of such assumptions depends upon the nature of the blocks and will be discussed critically below. Note that we do not assume independence of deviations.

Let us now explore the relationship between (A, B) and the vector $(\mathbf{A}, \mathbf{B})$, where $(\mathbf{A}, \mathbf{B}) \equiv ((A_1, B_1), \ldots, (A_M, B_M))$ consists of the intercepts and slopes of the individual block regression lines. Sometimes there are deterministic relationships between these quantities. Suppose, for example, that the block averages $X_{i.}$ are all equal, so that they have the common value $X_{..} = (\sum_{i=1}^{M} S_i X_{i.})/N$, and that the block variances $\sum_j (X_{ij} - X_{i.})^2/S_i$ are all equal with common value σ_x^2. Then

$$B = \frac{\sum_{i,j} Y_{ij}(X_{ij} - X_{..})}{\sum_{i,j}(X_{ij} - X_{..})^2} = \frac{\sum_{i,j} Y_{ij}(X_{ij} - X_{i.})}{\sum_{i,j}(X_{ij} - X_{i.})^2}$$

$$= \frac{\sigma_x^2 \sum_i B_i S_i}{\sigma_x^2 \sum_i S_i} = \sum_{i=1}^{M} W_i B_i,$$

where $W_i = S_i/N$, and it then follows easily that $A = \sum_{i=1}^{M} W_i A_i$. Note also that if the slopes B_i are identical, then without any assumption on the independent variable, $A = \sum_{i=1}^{M} W_i A_i$.

However, in the applications that are to be considered it will not ordinarily be appropriate to make such strong assumptions as will yield $B = \sum_{i=1}^{M} W_i B_i$ and $A = \sum_{i=1}^{M} W_i A_i$. Nonetheless both linear and nonlinear approximate Bayes estimators of the relationship between Y and X in the population will be obtained.

Suppose that the data consist of simple random samples without replacement from the blocks, with J_i units from block i, $i = 1, \ldots, M$, and the drawings from different blocks independent. We allow some of the J_i to be 0 and then relabel so that $J_i > 0$ for $i = 1, \ldots, m$ and $J_i = 0$ for $i = m + 1, \ldots, M$. For simplicity assume that $J_i \geq 3$ for $i = 1, \ldots, m$.

Let us evaluate the Bayes posterior expectation of Y for a unit with X having value x. Thus

$$\begin{aligned} H(x) \equiv E[Y|X = x, \text{data}] &= E\{E[Y|X = x, \mathbf{A}, \mathbf{B}, \text{data}]\} \\ &= E\left[\sum_{i=1}^{M} W_i(x)[A_i + B_i x] \,|\, \text{data}\right] \\ &= \sum_{i=1}^{M} W_i(x) E[A_i|\text{data}] + x \sum_{i=1}^{M} W_i(x) E[B_i|\text{data}], \end{aligned}$$

where $W_i(x)$ is the posterior probability that a unit with value $X = x$ belongs to block i. Here, depending on the interpretation of probability, $W_i(x)$ may or may not be regarded as known, but in the absence of other knowledge, is assumed to be evaluated, given the data, as

$$W_i(x) = \frac{P(X = x|\text{block } i) W_i}{\sum_{j=1}^{M} P(X = x|\text{block } j) W_j}$$

where $P(X = x|\text{block } i)$ refers to the distribution of X for block i.

The function $H(x)$ is in general a nonlinear function of x, possessing the usual Bayesian mean square error optimality properties. However, to make explicit use of $H(x)$ requires careful evaluation of the $W_i(x)$, and $H(x)$ may be quite sensitive to their specification. Along the lines of our general philosophy toward robustness, it is natural to look for a simple linear approximation to $H(x)$ that sacrifices some degree of optimality for greater robustness. Consider then the linear function

$$L(x) = \sum_{i=1}^{M} W_i E[A_i|\text{data}] + x \sum_{i=1}^{M} W_i E[B_i|\text{data}].$$

Suppose that there exist at least two values $x_1 \neq x_2$ such that $W_i(x_j) = W_i$, $i = 1, \ldots, M, j = 1, 2$. Hence these values are completely uninformative as to the block to which a unit belongs. In this case $H(x_j) = L(x_j), j = 1, 2$, so that the linear function $L(\cdot)$ intersects the optimal Bayes estimator $H(\cdot)$ in at least two points. This suggests the possibility of using $L(x)$ as an estimator for the population least squares regression line $A + Bx$. Note that when the deterministic relationships

$$A = \sum_{i=1}^{M} W_i A_i, \qquad B = \sum_{i=1}^{M} W_i B_i$$

hold, then in fact $E[A + Bx|\text{data}] = L(x)$. However, the above argument suggests approximating $E[Y|X = x, \text{data}]$ by $L(x)$ under the very weak assumption of the existence of the uninformative values x_1, x_2. In fact, in many applications it will suffice merely to regard $W_i(x_j) \simeq W_i, j = 1, 2$, without even giving careful consideration to the values x_j for which this occurs or to the sharpness of the approximation. We must, of course, rule out cases in which the X_{ij} for the various blocks are known to be nearly disjoint since in this case the value x would virtually identify the block to which the unit belongs. Thus if $X = x$ makes it very likely that the unit is from block i, then

$$H(x) \simeq E[A_i|\text{data}] + xE(B_i|\text{data}),$$

and $L(x)$ would not be appropriate as an estimator.

With the above discussion as motivation, let "data" refer to the sample values $(x_{ij}, y_{ij}), i = 1, \ldots, m, j = 1, \ldots, J_i$, and let us evaluate

$$\tilde{A} \equiv \sum_{i=1}^{M} W_i E[A_i|\text{data}] \qquad \text{and} \qquad \tilde{B} \equiv \sum_{i=1}^{M} W_i E[B_i|\text{data}].$$

There are several ways in which this can be done, depending on the chosen prior distribution for **A** and **B**. The way which corresponds most closely to our previous random model analysis is to assume that the set of As and Bs are conditionally independent samples from normal populations, i.e., $A_i \sim N(\mu_A, \sigma_A^2)$, $B_i \sim N(\mu_B, \sigma_B^2)$, $i = 1, \ldots, M$, with **A** independent of **B**, given μ_A, μ_B, σ_A^2, and σ_B^2. Here we shall illustrate the approximation to $\tilde{B}$. For this purpose we shall take as the informative portion of the data in the ith block the statistics $\hat{A}_i = y_{i\cdot} - \hat{B}_i x_{i\cdot}, \hat{B}_i$ and $\hat{\sigma}_i^2$, where

$$\hat{B}_i = \sum_{j=1}^{J_i} (x_{ij})(y_{ij} - y_{i\cdot}) \Big/ \sum_{j=1}^{J_i} (x_{ij} - x_{i\cdot})^2,$$

and

$$\hat{\sigma}_i^2 = \sum_{j=1}^{J_i} (y_{ij} - \hat{A}_i - \hat{B}_i x_{ij})^2/(J_i - 2), \qquad J_i \geq 3,$$

for $i = 1,2,\ldots,m$. Without further assumptions these are not sufficient statistics for the ith block, but often can be anticipated to contain most of the information concerning A_i, B_i, and σ_i^2. We also anticipate that to a good approximation $\hat{B}_i \sim N(B_i, V_i^2)$, given B_i and V_i^2, where

$$V_i^2 = [1-(J_i-1)/(S_i-1)]\sigma_i^2 \Big/ \sum_{j=1}^{J_i} (x_{ij}-x_{i\cdot})^2$$

and the factor in brackets is the usual finite population correction to allow for sampling without replacement. It follows as in our previous analysis of the random model that

$$E[B_i \,|\, \text{data}, \mu_B, \sigma_B^2/V_i^2] \simeq \theta_i\hat{B}_i + (1-\theta_i)\mu_B,$$

and

$$E[B_i \,|\, \text{data}, \sigma_B^2/V_i^2] \simeq \theta_i\hat{B}_i + (1-\theta_i)E[\mu_B \,|\, \text{data}, \sigma_B^2/V_i^2],$$

where

$$\theta_i = [\sigma_B^2/V_i^2][1+\sigma_B^2/V_i^2]^{-1}$$

and

$$E[\mu_B \,|\, \text{data}, \{\sigma_B^2/V_i^2\}] \simeq \sum_{i=1}^{m} \hat{B}_i/(\sigma_B^2+V_i^2)\left[\sum_{i=1}^{m} 1/(\sigma_B^2+V_i^2)\right]^{-1}.$$

When $J_i \geq 3$, for $i = 1,\ldots,m$, then a crude, but for many purposes adequate, estimate of V_i^2 can be obtained by substituting $\hat{\sigma}_i^2$ for σ_i^2 in the definition of V_i^2, thus leading to an estimate $\hat{\mu}_B$ of $E[\mu_B \,|\, \text{data}, \{\sigma_B^2/V_i^2\}]$. Furthermore, since $\text{Var}[\hat{B}_i \,|\, \mu_B, \sigma_B^2, \sigma_i^2] = \sigma_B^2 + V_i^2$, it follows that an estimate of σ_B^2 can be obtained just as in the usual one-way unbalanced random model, thereby leading to an estimate $\hat{\theta}_i$ of θ_i (Hill, 1965, p. 821). Finally, using such estimates, we evaluate

$$\begin{aligned}\tilde{B} &= \sum_{i=1}^{M} W_i E[B_i \,|\, \text{data}] \simeq \sum_{i=1}^{m} [\hat{\theta}_i\hat{B}_i + (1-\hat{\theta}_i)\hat{\mu}_B]W_i + \sum_{i=m+1}^{M} \hat{\mu}_B W_i \\ &= \hat{\mu}_B + \sum_{i=1}^{m} W_i\hat{\theta}_i(\hat{B}_i - \hat{\mu}_B).\end{aligned}$$

Here our main purpose is not to suggest particular estimators for μ_B and θ_i, since these are rather delicate matters, but rather to show the form which the overall estimator $\tilde{B}$ must take, under our assumptions. In particular it is important to note the quite different roles played by the W_i weights and the θ_i weights (or their estimates). The W_i enter because we are interested in a line for the entire population, while the $\hat{\theta}_i$ are approximations to the optimal weights for estimating B_i in the ith block and, if chosen carefully, will yield admissible estimators that incorporate the prior knowledge that one wishes to bring to bear on the problem.

As mentioned earlier, the above mode of analysis may be modified in certain natural ways.

1. When some J_i are very small, it may be wise to pool some of the blocks, provided there is no indication that such blocks are very disparate. On the other hand, blocks which are viewed as very disparate from the main body of blocks should be analyzed separately.

2 As in Section 3, inspection of the $\hat{B}_i$, and the $\hat{\varepsilon}_{ij} = y_{ij} - \hat{A}_i - \hat{B}_i x_{ij}$, may suggest that normality is inappropriate, in which case the analysis can be modified along the lines of Section 3.

3. One may wish to investigate relationships of the form $B_i = \alpha + \beta S_i +$ error or $\ln \sigma_i^2 = \alpha + \beta \ln S_i +$ error. If the data provide evidence for such relationships, then estimation of such α and β will yield new estimates for $E[B_i | \text{data}]$, of the form $E[B_i | \hat{\alpha}, \hat{\beta}]$, which will then be employed in the evaluation of $\tilde{B}$.

It should be understood that carrying out the original analysis or any of the proposed modifications is a delicate and subtle affair and that no routine analysis is possible. In particular, it should be noted that the assumption that knowledge that $X = x$ carries negligible information about the associated ε is not innocuous. For example, if the pairs (X, Y) in the population were scattered about a convex function but we retained the linear model, then the expectation of errors associated with extremely large or small values x would tend to be positive rather than zero. Of course, in this case one might fit an appropriate convex function of x, but the analysis becomes much more complex and dubious. This suggests some restrictions on the potential applications of the above analysis, namely, to cases where the assumption of negligible information seems appropriate, at least within each block. Thus the blocks must be chosen with judgment.

Finally, let us note that questions as to the appropriateness of $H(x)$ or $L(x)$ are especially subtle since they depend not only upon the choice of criteria for the performance of estimators, but also upon the choice of the function that we use to represent the relationship between Y and X in the population, and hence upon the purposes of the experiment. For example, should one try to estimate the least squares regression line in the population, $A + Bx$, or should one try to estimate the line $\sum_{i=1}^{M} W_i A_i + x \sum_{i=1}^{M} W_i B_i$? In general they are not the same. Tentatively, for forecasting purposes, we would advance the following argument in favor of $L(x)$ over $E[A + Bx | \text{data}]$. Suppose we wish to forecast Y, given that a unit has $X = x$. If the block to which the unit belongs is known, then for squared error loss the Bayes optimal forecast is the posterior expectation of the line for that block at that value x. On the other hand, if the block is unknown, then in the absence of other information it is natural to use $W_i(x)$ for the probability that the unit belongs to block i, and in this case the Bayes optimal forecast is $H(x)$. If the $W_i(x)$ are sufficiently well known, then $H(x)$ is in fact available, and there is no need to consider $L(x)$ or $E[A + Bx|\text{data}]$. Note, for example, that if the X_{ij} are known to belong to disjoint intervals I_i for different blocks, then the $W_i(x)$ become 0 or 1, and $H(x)$ reduces to the appropriate line $E[A_i + B_ix|\text{data}]$ in the interval I_i, $i = 1, \ldots, M$. It is in such a situation of course that $H(x)$ is most preferred over the other candidates. Now suppose that the $W_i(x)$ are not sufficiently well known to be used with confidence for all values x. It may still be the case that there exist certain values x which are viewed as uninformative as to the block to which the unit belongs, so that $W_i(x) = W_i$ for such x. Then at least $H(x) = L(x)$ for such x, so that there is a clear sense in which $L(x)$ approximates $H(x)$, while this is not the case for $E[A + Bx|\text{data}]$. Furthermore, if conditions are such that $A = \sum_{i=1}^{M} W_iA_i$ and $B = \sum_{i=1}^{M} W_iB_i$, then $L(x) = E[A + Bx|\text{data}]$, so use of $L(x)$ amounts precisely to estimation of the population least squares line. When such conditions do not hold, which suggests that the X_{ij} tend to be distributed differently in the different blocks, then, of course, there is the greatest potential benefit to be derived from use of $H(x)$, and $L(x)$ at least approximates $H(x)$ in a clear sense. Thus in the one case we lose nothing by use of $L(x)$ and in the other case we have everything to gain. $H(x)$, itself, would of course be the ideal.

Finally, in the case where there are only two blocks, and the X_{ij} and Y_{ij} are one or zero corresponding to presence or absence of some characteristic, it is easy to see that the considerations we have suggested in regard to use of $H(x)$, the least squares line, and $L(x)$, are analogous to those that arise in Simpson's paradox (Simpson, 1951).

5. On Robustness

Here I shall try to summarize briefly the attitude towards robustness which was illustrated in the analysis of the random model.

1. In any problem there is usually some source of knowledge which has a special status as creating those assumptions about which one is most confident. Such knowledge may concern the model, the likelihood function, the prior distribution, or the utility function. It is then best to make the analysis depend most heavily upon the assumptions about which one is most content. In the initial analysis of the random model, under conventional assumptions, the analysis based on MSB/MSW alone, reflected such a source

of knowledge, and is, in my opinion, ordinarily the most appropriate. An exception would occur if one felt there were compelling reasons for a particular form of prior distribution for the variance components, and such alternative modes of analysis should be kept in mind and investigated. But one should not feel obligated to adopt them merely because they utilize a sufficient statistic or are admissible. Indeed, in the "big world," as opposed to the more customary small world analysis (Hill, 1975a, p. 582, 1977, p. 31), there are no known admissible strategies, and an overly ambitious attempt to achieve such admissibility may produce analyses which are quite absurd. This point of view is more fully explained by Hill (1975b, p. 1169), where it is argued that under realistic assumptions about the global form of a distribution which follows the Zipf–Pareto law in the upper tail, it is best to base inference about the upper tail upon the upper order statistics alone. In this way one obtains an analysis which is robust in the sense that it is not affected by the nuisance parameters that characterize the global distribution. Of course, such an analysis is not based upon a sufficient statistic, nor would it be admissible. The trap of attempting to base inference upon all the data is one which naive Bayesians are especially prone to fall into, but they need not, provided that they keep in mind that a sifting of the totality of available data into its most informative parts for the parameters of interest is necessary in order that any analysis be feasible. Using such subjectively judged informative portion of the data, it is possible to perform a likelihood or Bayesian analysis, which will be relatively insensitive to the nuisance parameters. See also Hill (1969, p. 95) for another example of such robust analysis.

2. The mode of analysis should be such that sensitivity to the various underlying assumptions can be examined. If changes in the assumptions within the range of assumptions that one regards as plausible lead to quite different inferences or decisions, then one must recognize that there is no robustness and that any choice must be made with extreme care. It should be understood that the data can be used to make judicious choices, for example, as to the use of normal theory in the random model.

3. Finally, I should like to suggest that the way to achieve 'true" robustness is to build better models rather than merely to "protect oneself" against small departures from an assumed model. Such a process of building better models was illustrated by the broadening of the usual random model to allow for negatively correlated residuals (Hill, 1967, p. 1395, 1970, p. 33). To worry about small departures from normality in connection with a realistically complex data set is to worry about a mouse when confronted by a tiger. Model building is part of a long-standing scientific tradition, requiring, among other things, creative insight. Even small improvements in an existing model, for example, to cover a wider range of conditions, have enormous impact upon our understanding of real phenomena.

REFERENCES

Cornfield, J., & Tukey, J. W. Average values of mean squares in factorials. *Annals of Mathematical Statistics*, 1956, **27**, 907–949.

Culver, D. H. *A Bayesian analysis of the balanced one-way variance components model.* Ph.D. dissertation, University of Michigan, 1971.

Fisher, R. A. The fiducial argument in statistical inference. *Annals of Eugenics*, 1935, **6**, 391–398.

Hill, B. M. Inference about variance components in the one-way model. *Journal of the American Statistical Association*, 1965, **60**, 806–825.

Hill, B. M. Correlated errors in the random model. *Journal of the American Statistical Association*, 1967, **62**, 1387–1400.

Hill, B. M. Foundations for the theory of least squares. *Journal of the Royal Statistical Society, Series B*, 1969, **31**, 89–97.

Hill, B. M. Some contrasts between Bayesian and classical inference in the analysis of variance and in the testing of models. In D. L. Meyer & R. O. Collier, Jr. (Eds.), *Bayesian statistics.* Itasca, Ill.: Fe E. Peacock Publ., 1970. Pp. 29–36.

Hill, B. M. On coherence, inadmissibility, and inference about many parameters in the theory of least squares. In S. E. Fienberg & A. Zellner (Eds.), *Studies in Bayesian econometrics and statistics* (*in honor of Leonard J. Savage*). Amsterdam: North-Holland Publ., 1975. Pp. 555–584. (a)

Hill, B. M. A simple general approach to inference about the tail of distribution. *Annals of Statistics*, 1975, **3**, 1163–1174. (b)

Hill, B. M. Exact and approximate Bayesian solutions for inference about variance components and multivariate inadmissibility. In A. Aykac & C. Brumat (Eds.), *New developments in the applications of Bayesian methods.* Amsterdam: North-Holland Publ., 1977. Pp. 129–152.

Scheffé, H. *The analysis of variance.* New York: Wiley, 1961.

Simpson, E. The interpretation of interaction in contingency tables. *Journal of the Royal Statistical Society, Series B*, 1951, **13**(2), 238–241.