

whole is 8.5%. Even though sample mean company R&D intensity (3.7%) is slightly higher than mean federal R&D intensity (3.6%), $\hat{\beta}$ is much closer to $\hat{\beta}_2$ than it is to $\hat{\beta}_1$.

Although the ratio of R&D expenditure to sales is the most commonly used measure of R&D expenditure, an alternative, and perhaps equally valid, measure is the ratio of R&D scientists and engineers employed to total employment. On the second row of Table 2, I present regression coefficients and other statistics for exactly the same sample of firms based on this alternative measure of R&D intensity. The point estimates of $\beta_1$ and $\beta_2$ are very close to their counterparts based on the ratios of R&D expenditure to sales, and the hypothesis $\beta_1 = \beta_2$ is again rejected. But the standard deviation of federal R&D intensity is now only 41% higher than that of company R&D intensity; as Table 1 reveals, this causes a substantial increase in $w$. The correlation coefficient is somewhat lower, $-.09$ compared with .02, tending to reduce $w$ slightly. But $w(1.43, -.0920) = .3179$, which is almost twice as large as (81% higher than) the weight corresponding to the R&D expenditure data. Consequently, the estimate of $\beta$ is 76% higher than the estimate based on the R&D expenditure data.

To summarize, the R&D expenditure- and exployment-intensity data yield virtually identical estimates of the returns to R&D classified by source of funds, but because of differences in the weight $w$, they yield rather different estimates of the returns to total R&D. This example illustrates the need to exercise caution in the interpretation of the coefficient of an aggregate when the coefficients of its components are believed to differ.

[*Received June 1988. Revised July 1989.*]

## REFERENCES

Griliches, Zvi (1957), "Specification Bias in Estimates of Production Functions," *Journal of Farm Economics*, 39, 8–20.

Griliches, Zvi, and Lichtenberg, Frank (1984), "R&D and Productivity Growth at the Industry Level: Is There Still a Relationship?" in *R&D, Patents and Productivity*, ed. Z. Griliches, Chicago: University of Chicago Press, pp. 465–496.

Grunfeld, Yehudi, and Griliches, Zvi (1960), "Is Aggregation Necessarily Bad?" *Review of Economics and Statistics*, 42, February, 1–13.

Madalla, G. S. (1977), *Econometrics*, New York: McGraw-Hill.

Orcutt, G., Watts, H., and Edwards, J. (1968), "Data Aggregation and Information Loss," *American Economic Review*, 58, 773–787.

Theil, Henri (1954), *Linear Aggregation of Economic Relations*, Amsterdam: North-Holland.

# We've Got the Positive Correlation BLUEs

STEVE VERRILL, MICHAEL AXELROD, and MARK DURST*

Intuition suggests that combinations of positively correlated estimates of a quantity have greater variances than combinations of independent estimates of the quantity. In this note we identify circumstances under which best linear unbiased estimators (BLUEs) based on positively correlated measurements are superior to BLUEs based on independent measures.

## 1. INTRODUCTION

Scientific workers often face situations in which they wish to combine multiple unbiased estimates of a single quantity. These estimates may come from different laboratories, experiments, or measuring devices. Under certain circumstances (e.g., the laboratories share samples, researchers, or information), the estimates might be correlated.

The presence of correlation is not necessarily bad, and statisticians know that negative correlation can be exploited to their advantage. If, for example, we wish to combine two measurements, where a positive error in one tends to be associated with a negative error in the other, the best linear combination of these measurements will have a smaller variance than would be possible if the measurements were uncorrelated. In fact, if the correlation is $-1$, the combination will have a zero variance regardless of how imprecise the original measurements were. What is not generally appreciated is that even positive correlation can be useful.

Intuition suggests that positive correlation should increase the variance of a combination because measurements with large positive correlations would seem to be almost a redundancy. Intuition is correct when the measurements have equal precision, but sometimes wrong when the precisions differ. The purpose of this note is to demonstrate that if the variances of the individual measurements are sufficiently different and the positive correlations are sufficiently high, the quality of the combined estimates will *improve*.

## 2. THE BEST LINEAR UNBIASED ESTIMATOR (BLUE)

Let $y_1, \ldots, y_n$ be our estimates, and assume that $E(y_i) = \mu$. Let $\Sigma$ denote the (known) covariance matrix of $\mathbf{y}$. An easy application of the theory of generalized least squares gives us

$$\hat{\mu}_{\text{BLUE}} = \mathbf{1}'\Sigma^{-1}\mathbf{y}/\mathbf{1}'\Sigma^{-1}\mathbf{1} \tag{1}$$

(BLUE is best linear unbiased estimator), which has variance

$$V = 1/\mathbf{1}'\Sigma^{-1}\mathbf{1}. \tag{2}$$

*Steve Verrill is Statistician, Forest Products Laboratory, U.S. Department of Agriculture, Madison, WI 53705. Michael Axelrod is Electrical Engineer, Engineering Research Division, Lawrence Livermore National Laboratory, Livermore, CA 95440. Mark Durst is Statistician, Computing and Mathematics Research Division, Lawrence Livermore National Laboratory, Livermore, CA 94550. This work was performed under the auspices of the U.S. Department of Energy at the Lawrence Livermore National Laboratory under Contract W-7405-Eng-48.

Now consider the special case in which $\text{var}(y_i) = \sigma_i^2$ and $\text{cov}(y_i, y_j) = \rho\sigma_i\sigma_j$. Then, for $\rho = 0$, Equations (1) and (2) become the familiar

$$\hat{\mu}_{\text{BLUE}} = \left(\sum_{i=1}^{n} \frac{y_i}{\sigma_i^2}\right)\left(\sum_{i=1}^{n} \frac{1}{\sigma_i^2}\right)^{-1} \tag{3}$$

and

$$V = \left(\sum_{i=1}^{n} \frac{1}{\sigma_i^2}\right)^{-1}. \tag{4}$$

For $\rho \neq 0$, we have

$$\Sigma = \text{diag}(\sigma_i)[(1 - \rho)I + \rho\mathbf{1}\mathbf{1}']\text{diag}(\sigma_i). \tag{5}$$

Let $d_i$ denote $\sigma_i^{-1}$. From Equation (5) we have

$$\mathbf{1}'\Sigma^{-1}\mathbf{1} = \mathbf{d}'[(1 - \rho)I + \rho\mathbf{1}\mathbf{1}']^{-1}\mathbf{d},$$

where $\mathbf{d}' \equiv (d_1 \cdots d_n)$. Noting that

$$[(1 - \rho)I + \rho\mathbf{1}\mathbf{1}']^{-1} = \frac{I}{1 - \rho} - \frac{\rho\mathbf{1}\mathbf{1}'}{(1 - \rho)(1 + (n - 1)\rho)},$$

we obtain

$$V = \frac{1}{\mathbf{1}'\Sigma^{-1}\mathbf{1}} = \frac{(1 - \rho)(1 + (n - 1)\rho)}{[1 + (n - 1)\rho]\sum_{i=1}^{n} d_i^2 - \rho(\sum_{i=1}^{n} d_i)^2}. \tag{6}$$

Now we can draw several conclusions.

## 3. POSITIVE CORRELATION IS BAD

Suppose that $\sigma_1^2 = \sigma_2^2 = \cdots = \sigma_n^2 = \sigma^2$. Then (6) reduces to

$$V = \sigma^2(1 + (n - 1)\rho)/n. \tag{7}$$

So for $\rho = -1/(n - 1)$, $V = 0$, and for $\rho = 0$, $V = \sigma^2/n$. As $\rho$ increases to 1, $V$ increases to $\sigma^2$, and our $n$ observations have collapsed down to a single observation. This is the source of the notion that positive correlation is "bad."

## 4. POSITIVE CORRELATION IS GOOD

Suppose, however, that $\sigma_i^2 \neq \sigma_j^2$ for some $i, j$. As $\rho$ increases to 1, the numerator in (6) converges to 0, but the denominator converges to

$$n\sum_{i=1}^{n} d_i^2 - \left(\sum_{i=1}^{n} d_i\right)^2 = n\sum_{i=1}^{n} (d_i - \bar{d})^2 > 0 \tag{8}$$

for $\sigma_i \neq \sigma_j$, some $i, j$. (Here, $\bar{d}$ denotes the average of the $d_i$'s.) Thus $V$ *converges to 0* as $\rho$ goes up to 1. See Figure 1 for a plot of $V$ versus $\rho$ in several cases in which $n = 2$.

We can make this result more intuitive by considering the following extreme case. Let $Y_1$ and $Y_2$ be perfectly positively correlated, unbiased estimates of $\mu$, with $\sigma_1^2 \neq \sigma_2^2$. Then $Y_1 = \mu + \sigma_1\varepsilon$ and $Y_2 = \mu + \sigma_2\varepsilon$ (the same $\varepsilon$). Solving for $\mu$ we obtain the unbiased, zero-variance estimator

$$(\sigma_2 Y_1 - \sigma_1 Y_2)/(\sigma_2 - \sigma_1).$$

Finally, having established that positive correlation can yield better results than no correlation, it would be nice to know exactly how large the positive correlation must be for this to hold. We answer this question by setting Expression (6) equal to the variance at $\rho = 0$ and solving for $\rho$. This yields

$$\frac{(1 - \rho)(1 + (n - 1)\rho)}{[1 + (n - 1)\rho]\sum_{i=1}^{n} d_i^2 - \rho(\sum_{i=1}^{n} d_i)^2} = \frac{1}{\sum_{i=1}^{n} d_i^2},$$

or

$$\rho_{\text{break-even}} = \frac{1}{n - 1}\frac{(\sum_{i=1}^{n} d_i)^2 - \sum_{i=1}^{n} d_i^2}{\sum_{i=1}^{n} d_i^2} = 1 - \frac{n}{n - 1}\frac{\sum_{i=1}^{n} (d_i - \bar{d})^2}{\sum_{i=1}^{n} d_i^2}. \tag{9}$$

Thus, as the spread in the $d_i$'s relative to $\sum_{i=1}^{n} d_i^2$ goes up, the amount of positive correlation needed to improve upon no correlation goes down.

## 5. EXAMPLE

Consider the case $n = 2$. Assume that $\sigma_1 < \sigma_2$. By taking the derivative of (6) with respect to $\rho$ and setting it equal to 0, we find that the worst possible variance is achieved at $\rho = \sigma_1/\sigma_2$. Plugging this value back into (6), we see that the maximum value is $\sigma_1^2$. From Equation (9), we have

$$\rho_{\text{break-even}} = 2/(\sigma_1/\sigma_2 + \sigma_2/\sigma_1). \tag{10}$$

In Table 1 and Figure 1 we consider four cases. In all four cases, the $\rho = 0$ variance was set to 1 and we solved for $\sigma_1$, $\sigma_2$, maximum variance $= \sigma_1^2$, and $\rho_{\text{break-even}}$. From this work we see that a 4 to 1 standard deviation ratio leads to BLUE improvements over the $\rho = 0$ case for $\rho \geq .471$. A 10 to 1 ratio reduces this breakeven correlation to .2.

*Remark.* For illustrative purposes, in the foregoing calculations we restricted ourselves to the special case in which the off-diagonal correlations are all equal. More generally, Equations (2) and (4) imply that correlation will be "good" whenever

$$\mathbf{1}'\Sigma\mathbf{1} > \sum_{i=1}^{n} \frac{1}{\sigma_i^2}.$$

As we have seen, this will certainly occur (for example) whenever any pair of measurements has sufficiently different variances and a sufficiently high positive correlation.

*Further Reading.* The interested student can pursue the topics of generalized least squares and BLUEs in Searle (1971), Seber (1977), or Kendall and Stuart (1979). Kleijnen (1974) discussed the use of negative correlation (antithetic variates) as a variance reduction technique in simulation studies. Ku (1969) provided a classic reference that provides an introduction to some of the problems associated with interlaboratory testing.

[*Received March 1988. Revised August 1988.*]

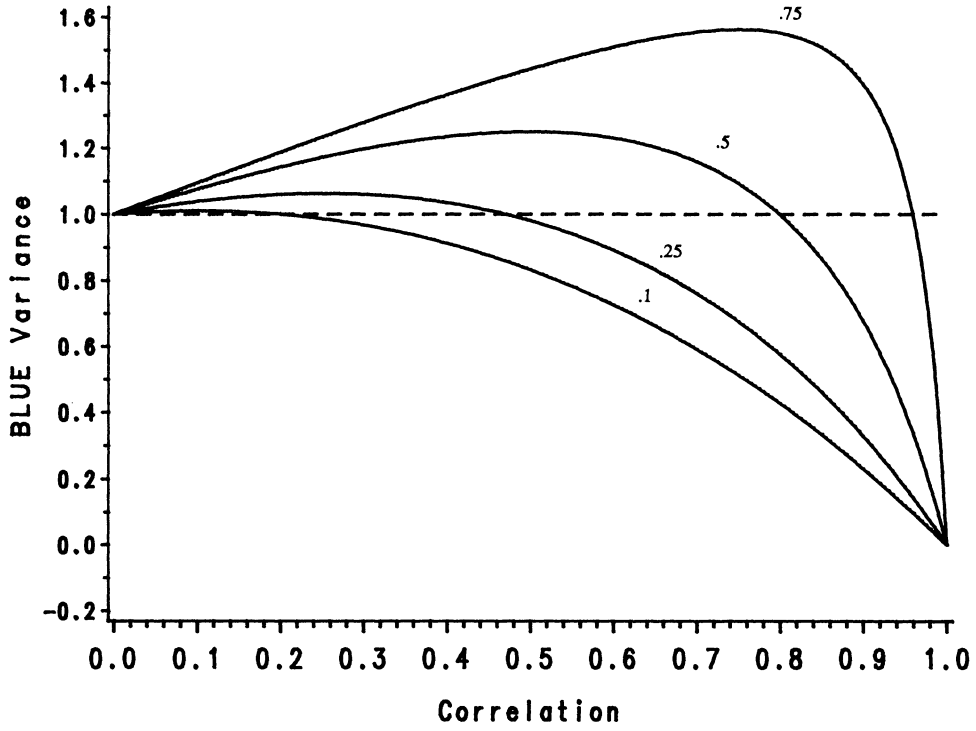


*Figure 1. The numbers indicate the $\sigma_1/\sigma_2$ values associated with the curves. In all four cases, the variance at $\rho = 0$ was 1. Thus the curves cross the dashed line above the break-even $\rho$'s.*

*Table 1. Break-Even Correlation as a Function of the Standard Deviation Ratio*

| $\sigma_1/\sigma_2$ | $\sigma_1$ | $\sigma_2$ | Maximum variance (at $\rho = \sigma_1/\sigma_2$) | Break-even $\rho$ |
|---|---|---|---|---|
| .75 | 1.2500 | 1.6667 | 1.5625 | .9600 |
| .50 | 1.1180 | 2.2361 | 1.2500 | .8000 |
| .25 | 1.0308 | 4.1231 | 1.0625 | .4706 |
| .10 | 1.0050 | 10.050 | 1.0100 | .1980 |

## REFERENCES


Kendall, M., and Stuart, A. (1979), *The Advanced Theory of Statistics* (Vol. 2), New York: Macmillan.

Kleijnen, J. P. C. (1974), *Statistical Techniques in Simulation, Part I*, New York: Marcel Dekker.

Ku, H. (ed.), (1969), *Precision Measurement and Calibration* (Vol. 1; Special Publication 300), Washington, DC: U.S. National Bureau of Standards.

Searle, S. R. (1971), *Linear Models*, New York: John Wiley.

Seber, G. A. F. (1977), *Linear Regression Analysis*, New York: John Wiley.