Intrinsic negative variance components: case of non-normal functions of random variables

Ronian Siew

2307-939 Homer Street, Vancouver, British Columbia, V6B2W6, Canada
(ronian@inopticalsolutions.com)

Abstract

Studies from https://doi.org/10.6084/m9.figshare.3840753 and http://doi.org/10.6084/m9.figshare.3978087 are here extended to include the possibility that stochastic physical parameters (i.e., parameters that are functions of random variables) may not be normally distributed in general. It is shown that such parameters would possess at least two intrinsic negative variance components that are always negative for regression models that apply a second-degree Taylor series. Plausible implications of their impact on the random effects ANOVA model are discussed.

Keywords: Variance components analysis; Regression analysis; Mathematical models; Analysis of variance (ANOVA)

Earlier studies [1, 2] had shown that, when a second-degree multivariable Taylor series is used as a regression model for a stochastic physical parameter with two factors, at least one intrinsic negative variance component could exist for the variance of that stochastic physical parameter, and a specific practical example was provided in reference 1 where this would occur. In reference 2, equations were modified to include the possibility that the probability distribution of functions of random variables may be non-normal, but consequences to this modification were not discussed. In fact, reference 2 stated that it would not affect the main results of the study. However, as we shall now see, the consideration of non-normal stochastic functions would actually result in more than a single intrinsic negative variance component.

Following from equation 2 in reference 2, a stochastic physical parameter that is a function of two random factors may be expressed as

$$P_{ij} = P(\bar{x},\bar{y}) + a(x_i - \bar{x}) + b(y_j - \bar{y}) + c(x_i - \bar{x})(y_j - \bar{y}) + d(x_i - \bar{x})^2 + e(y_j - \bar{y})^2. \quad (1)$$

The mean $\bar{P}$ of P_{ij} is

$$\bar{P} = \frac{1}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n} P_{ij}. \quad (2)$$

Assuming that x and y are independent normal random variables, substituting Eq. (1) into (2) yields

$$\bar{P} = P(\bar{x},\bar{y}) + \frac{1}{m}\sum_{i=1}^{m} d(x_i - \bar{x})^2 + \frac{1}{n}\sum_{j=1}^{n} e(y_j - \bar{y})^2$$

$$= P(\bar{x},\bar{y}) + d(\delta_x)^2 + e(\delta_y)^2, \quad (3)$$

Where $(\delta_x)^2$ and $(\delta_y)^2$ are given by $\frac{1}{m}\sum_{i=1}^{m}(x_i - \bar{x})^2$ and $\frac{1}{n}\sum_{j=1}^{n}(y_j - \bar{y})^2$ respectively. The variance of P_{ij} is

$$\sigma^2 = \frac{1}{mn-1}\sum_{i=1}^{m}\sum_{j=1}^{n}\left(P_{ij} - \bar{P}\right)^2. \quad (4)$$

Substituting Eqs. (1) and (3) into (4) and performing the square in (4) yields

$$\sigma^2 = \frac{1}{mn-1}\sum_{i=1}^{m}\sum_{j=1}^{n}\big[a^2(x_i-\bar{x})^2 + b^2(y_j-\bar{y})^2$$

$$+c^2(x_i-\bar{x})^2(y_j-\bar{y})^2 + d^2(x_i-\bar{x})^4 + e^2(y_j-\bar{y})^4$$

$$-2d^2(\delta_x)^2(x_i-\bar{x})^2 - 2de(\delta_y)^2(x_i-\bar{x})^2$$

$$-2e^2(\delta_y)^2(y_j-\bar{y})^2 - 2de(\delta_x)^2(y_j-\bar{y})^2$$

$$+2de(x_i-\bar{x})^2(y_j-\bar{y})^2 + d^2(\delta_x)^4 + e^2(\delta_y)^4\big]. \quad (5)$$

Now, note that there are intrinsic negative variance components in Eq. (5), some of them depend on the sign of the coefficients d and e, while the terms $-2d^2(\delta_x)^2(x_i-\bar{x})^2$ and $-2e^2(\delta_y)^2(y_j-\bar{y})^2$ do not. Hence, under all present model assumptions, there are at least two variance components that are always negative.

Up until now, although we have not yet considered the inclusion of residuals to the regression model (or equivalently, we have not included an "error" term), it still is rather interesting at this point to also note the implications of Eq. (5) to the two-factor random effects analysis of variance (ANOVA) model, which is really just a first-degree Taylor series of two variables with the addition of the second-degree cross-term or so-called "interaction" term. The ANOVA mathematical model (without repeated level measurements) may be expressed as

$$P_{ij} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ij}, \quad (6)$$

where ε_{ij} is the error term, and $\mu = \bar{P}$, the mean [3]. Assuming that all variable terms in Eq. (6) are independent normal random variables, substituting Eq. (6) into (4) and performing the square in (4) yields

$$\sigma^2 = \frac{1}{mn-1}\sum_{i=1}^{m}\sum_{j=1}^{n}\big[{\alpha_i}^2 + {\beta_j}^2 + {(\alpha\beta)_{ij}}^2 + {\varepsilon_{ij}}^2\big]. \quad (7)$$

Now, if there were no nonlinear terms (i.e., the quadratic terms) in the regression model of Eq. (1), then $d = 0$, and $e = 0$ in Eq. (5). Under such a condition, a comparison of Eq. (7) with Eq. (5) motivates associating ${\alpha_i}^2$ with $a^2(x_i-\bar{x})^2$, ${\beta_j}^2$ with $b^2(y_j-\bar{y})^2$, ${(\alpha\beta)_{ij}}^2$ with $c^2(x_i-\bar{x})^2(y_j-\bar{y})^2$, and ${\varepsilon_{ij}}^2$ would be a residual term. But the presence of nonlinearity in the model (i.e., if $d \neq 0$ and $e \neq 0$) and having a non-normal stochastic physical parameter would result in the intrinsic negative variance components in Eq. (5). Hence, some of the second-powered positive terms in Eq. (5) would be reduced by negative terms. In fact, by making the associations mentioned above for the variance components in Eq. (7) with the components in Eq. (5), we may write

$${\alpha_i}^2 = (x_i-\bar{x})^2[a^2 - 2d^2(\delta_x)^2 - 2de(\delta_y)^2], \quad (8)$$

$${\beta_j}^2 = (y_j-\bar{y})^2[b^2 - 2e^2(\delta_y)^2 - 2de(\delta_x)^2], \quad (9)$$

$${(\alpha\beta)_{ij}}^2 = (x_i-\bar{x})^2(y_j-\bar{y})^2[c^2 + 2de]. \quad (10)$$

Eqs. (8) – (10) may be considered one example of the regression approach to ANOVA [4]. These equations imply that, depending on the signs of the coefficients d and e, there would be conditions that could lead to natural negative variance components in an ANOVA model. Since, as was discussed in the conclusions section of reference 1, a regression model that uses a second-degree Taylor polynomial would apply to situations where variations are significantly large relative to the mean, Eqs. (8) – (10) imply that negative variance component estimates may at times be a consequence of high variability in the data being studied, and one may therefore require an ANOVA mathematical model to include higher order terms.

On more fundamental grounds, perhaps it should also be noted that variance components in the ANOVA mathematical model is actually the result of an attempt at quantifying variability in terms of a simple sum of constituent parts. In fact, this is an over-simplification. If a phenomenon is believed to be influenced by several factors, then it stands to reason that this phenomenon would be a function of several variables. If this function exists, then it is not necessarily a simple sum of those

variables, each multiplied by coefficients of proportionality. If, instead, physics shows us that a physical phenomenon or process is described by a complex multivariable function, then the only way to partition its behavior into a sum of constituent pars is to perform a Taylor series expansion of this multivariable function, which is an approximation to that function. But that is precisely what the ANOVA mathematical model does (and also regression). That is, the ANOVA mathematical model is a linear sum of terms, and as such, it is an approximate description of the behavior of a phenomenon or process. This would be true whenever ANOVA is used to partition variability into a sum of constituent variance components, such as in gauge and measurement systems analysis (MSA).

As a final remark, we note again that, if the ANOVA mathematical model is fundamentally a first-degree Taylor series plus a second-degree cross-term (i.e., interaction), then the inclusion of higher degree terms from the Taylor series should help to converge the ANOVA model towards a more accurate description of the phenomenon or process it is modelling. If the series converges (within the radius of convergence), then all variance components (including both positive and negative components) would sum up to a finite positive total variance. Consequently, we may speculate that if any intrinsic higher order negative variance component terms have been left out of a model, then one or more of the estimated variance components might need to become negative in order to compensate for the absence of those higher order negative terms. Hence, truncation of higher order terms in an ANOVA mathematical model could, in theory, lead to the appearance of one or more negative variance component estimates for the first-degree terms (e.g., the $\alpha_i{}^2$ or $\beta_j{}^2$ terms), or even the second-degree "interaction" cross-term $(\alpha\beta)_{ij}{}^2$. We might go further to speculate that this is a consequence of a sort of "conservation of variance" principle. That is, total variance is a finite "conserved" positive quantity, just as energy in physics is a conserved quantity (in loose analogy). However, note that total variance could theoretically be negative if its standard deviation is a complex number (i.e., if the standard deviation is given by a number A + iB where i is the square root of negative one, and A may be zero, leaving only the imaginary part iB). In this case, the total variance is still a finite conserved quantity in the sense that it is a finite negative quantity. Still, perhaps in most cases, variation of physical parameters would be described by positive variances. An example, which is perhaps most fundamental, is the general form of Heisenberg's uncertainty principle in quantum mechanics [5], which is written as the product of the squares of the position and momentum standard deviations (i.e., their variances).

References

[1] R. Siew, "Do 'intrinsic' negative variance components exist for population variances of stochastic physical parameters?" Figshare (2016). DOI: https://doi.org/10.6084/m9.figshare.3840753.

[2] R. Siew, "Intrinsic negative variance components: errata," Figshare (2016), DOI: https://dx.doi.org/10.6084/m9.figshare.3978087.

[3] H. Sahai and M. I Ageel, *The Analysis of Variance: Fixed, Random, and Mixed Models*, (Birkhauser, Boston, 2000), p. 180.

[4] R. E. Walpole and R. H. Myers, *Probability and Statistics for Engineers and Scientists*, 3rd Ed., (Macmillan, New York, 1985), pp. 460 – 463.

[5] D. J. Griffiths, *Introduction to Quantum Mechanics*, (Prentice Hall, New Jersey, 1995), p. 109.