Letters to the Editor

Grömping, U. (2007), "Estimators of Relative Importance in Linear Regression Based on Variance Decomposition," *The American Statistician*, 61, 139–147: Comment by Menard and Response.

Two Simple Estimators of Relative Importance in Linear Regression Based on Variance Decomposition

Grömping (2007) considers two alternative estimators of relative importance in linear regression, one attributed to Lindeman, Merenda, and Gold (1980) and further explicated by Kruskal (1987), and another attributed to Feldman (2005) which is more computationally intensive but has more desirable properties. Dismissed without detailed consideration is an approach attributed to Hoffman (1960), the product of the standardized regression coefficient and the marginal correlation, because this product sometimes produces a negative estimate for the component of the explained variance attributed to a predictor. Here, two alternative estimators are proposed, one a slight modification of the approach attributed to Hoffman (1960) that produces estimates for components of explained variance attributable to the respective predictors based on the standardized coefficient and the zero-order correlation, and a second based only on the standardized regression coefficient. Both of the estimators suggested here have the advantages of being computationally simple and being adaptable to use with both the sample R^2 and the adjusted R^2 used to estimate the population explained variance, as well as to other reasonable measures of explained variance or explained variation.

In ordinary least squares multiple regression analysis (OLS), an outcome or dependent variable Y is modeled as a linear function of a set of predictors $X_1, X_2, \ldots, X_K$ using the formula $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_K X_K$, where the estimates for the β coefficients, represented by $b_0, b_1, b_2, \ldots, b_K$ are calculated using a closed-form solution to minimize the sum of the squared errors in Y. Here, to simplify the notation, all regression coefficients are assumed to be standardized (unstandardized regression coefficients are irrelevant to the present discussion). In some instances, when the relationship between Y and the predictors is nonlinear and the same for Y and each of the predictors, a nonlinear transformation of Y is used to linearize the equation so the right side of the equation remains the same, and for present purposes, it will be assumed that any such transformation has been applied to the dependent variable if appropriate. The strength of the relationship of the predictors to the dependent variable is measured by R^2, the variance in the dependent variable explained by the predictors. Various formulas can be used to calculate R^2 (Kvålseth 1985), but we ignore distinctions among these different formulas for this discussion on the assumption that the researcher has selected an appropriate R^2 statistic for the problem at hand. Most typically, R^2 is calculated as $R^2 = 1 - (\text{SSE}/\text{SST})$ where SSE is the error sum of squares and SST is the total sum of squares. When R^2 is calculated on sample data, it tends to be biased upwards and the bias is greater for smaller numbers of cases N and larger numbers of predictors K; so an adjusted R^2, $R^2_{\text{adj}} = 1 - (1 - R^2)(N - 1)/(N - K - 1)$, may be used in place of the sample R^2 as an estimator for the population explained variance (see, e.g., Tabachnick and Fidell 2007, pp. 153–154, who attribute this measure to Wherry 1931).

As reviewed in some detail by Grömping (2007), there is often a desire to rank order or to quantify the relative influence of different predictors in a linear regression model in a way that (a) decomposes the explained variance into shares allocated to the predictors, with the sum of the shares being equal to the explained variance; (b) with all shares being nonnegative; (c) with the share allocated to X_k being zero if the standardized regression coefficient β_k (or in practice its estimate, b_k) is equal to zero; and (d) with the share allocated to X_k being nonzero if β_k (or in practice its estimate, b_k) is not equal to zero. Grömping examined two estimators of relative importance, one attributed to Lindeman, Merenda, and Gold (1980) and Kruskal (1987), hereafter the LMGK estimator; and another attributed to Feldman (2005). The LMGK estimator was based on the average increment to R^2 for X_k. To calculate the average increment to R^2 for X_k, each variable is entered into the model, one at a time, and the increment to R^2 when X_k is added to the model is averaged over all possible orderings of the entry of X_k into the model. This average increment to R^2 approach has the disadvantage that it may violate criterion (c), termed the exclusion criterion by Grömping, since if a predictor is correlated at all with the dependent variable, if it is entered into the prediction equation first, it will necessarily have a nonzero average increment to R^2 [a point more clearly evident in the discussion of this approach by Kruskal (1987)] and hence a nonzero estimate of its contribution to the explained variance. In practice, of course, one can test whether the exclusion principle is actually violated by the LMGK estimator for any specific analysis. As indicated by Lindeman et al. (1980, p. 20; see also Menard 2004 and Grömping 2007) there are $K!$ possible orderings of the predictors, imposing a nontrivial computational burden for a large number of predictors; and as noted by Menard (2004, p. 221), the number of calculations for each predictor increases exponentially, with 2^{K-1} orderings that are unique with respect to the ordering of each predictor X_k relative to the variables that precede it in entry into the equation. As described by Grömping (2007), the proportional marginal variance decomposition approach of Feldman (2005) has larger variation and higher computational burden, but avoids violation of the exclusion criterion.

Dismissed by Grömping as unsuitable was a suggested estimator of relative importance attributed to Hoffman (1960) equal to the product of the standardized regression coefficient and the zero-order correlation. The standardized coefficient x correlation product is $\text{SCCP}_k = b_k r_k$, where b_k is the standardized regression coefficient and r_k is the zero-order correlation between the predictor and the dependent variable. This estimator was not considered in detail because it is possible with correlated predictors to obtain a negative value (and hence a negative estimated contribution to the explained variance) using $b_k r_k$. With some very minor modification, however, $b_k r_k$ can produce estimates of relative importance that (a) sum to R^2, (b) are all nonnegative, (c) are zero when $b_k = 0$, and (d) are nonzero when $b_k \neq 0$ (if also $r_k \neq 0$).

First, note that $\Sigma_k b_k r_k = R^2$ (Tatsuoka 1971, pp. 31–34). This is true even when some of the $b_k r_k < 0$. The reason why $b_k r_k$ would be less than zero for some $b_k r_k$ is that the regression coefficient b_k is sometimes opposite in sign to the correlation coefficient r_k. In practice, this may arise in the presence of a suppressor effect involving a near-zero correlation opposite in sign to the regression coefficient obtained when the suppressor variable is included in the equation. The production of a negative estimated component of the explained variance is also consistent with an unbiased estimate of a parameter whose true value is zero and which, in repeated sampling, may have both negative and positive values unless otherwise constrained to be nonnegative, and when it does occur, the standardized regression coefficient is typically small, close to zero (and the unstandardized regression coefficient is typically not statistically significant for small to moderate samples, but may be statistically significant for large samples). As an alternative to SCCP, we can consider the absolute value of $b_k r_k$, $\text{ASCCP}_k = |b_k r_k|$. This necessarily produces nonnegative estimates of the contribution of a predictor to R^2, and when all $b_k r_k$ are nonnegative, $R^2 = \Sigma_k b_k r_k = \Sigma_k |b_k r_k|$; but if at least one $b_k r_k$ is negative, $R^2 < \Sigma_k |b_k r_k|$.

The resolution to this problem is to norm each $|b_k r_k|$ by the ratio of R^2 to $\Sigma_k |b_k r_k|$, producing a normed standardized coefficient x correlation product $\text{NSCCP}_k = |b_k r_k| R^2 / \Sigma_k |b_k r_k|$. If all $b_k r_k$ are nonnegative to begin with, $R^2 = \Sigma_k |b_k r_k| = \Sigma_k b_k r_k$, so each $b_k r_k$ is multiplied by one, hence unchanged. If at least one $b_k r_k$ is less than zero, then because $R^2 = \Sigma_k b_k r_k$ even when some $b_k r_k < 0$, $R^2 < \Sigma_k |b_k r_k|$, and hence $R^2 / \Sigma_k |b_k r_k| < 1$. The resulting products $|b_k r_k| R^2 / \Sigma_k |b_k r_k|$ will sum to R^2, that is, $\Sigma_k [|b_k r_k| R^2 / \Sigma_k |b_k r_k|] = R^2 \Sigma_k |b_k r_k| / \Sigma_k |b_k r_k| = R^2$, satisfying the first criterion for an estimator of relative importance, summation to R^2. Because all of the elements of $|b_k r_k| R^2 / \Sigma_k |b_k r_k|$ are nonzero (one squared term and two in absolute values), the estimates of the contributions of each predictor to the explained variance are all nonnegative, satisfying the second criterion. Because $|b_k r_k| R^2 / \Sigma_k |b_k r_k| = 0$ whenever b_k is equal to zero the third criterion, the exclusion criterion, is also satisfied. Finally, unless b_k or r_k is equal to zero, $b_k r_k \neq 0$. At issue here is the extent to which the fact that $b_k r_k = 0$ when $r_k = 0$ even if $b_k \neq 0$ is a matter of concern (an issue with both NSCCP and ASCCP, and also the original SCCP proposed by Hoffman 1960). If *no* $b_k r_k = 0$, the issue does not really arise, and the inclusion criterion is satisfied. Second, if $b_k r_k = 0$ *only* because $r_k = 0$ (and $|b_k| > 0$) to whatever degree of precision is used in the calculation, this can be detected and one can decide whether, in the context of the specific

©American Statisticial Association DOI: 10.1198/000313007X223739

analysis, this is sufficient to justify the additional computational burden imposed by other estimators such as the LMGK or Feldman estimators. Computational effort for NSCCP is minimal, and the computational burden increases linearly rather than exponentially as additional variables are added to the model, so if other considerations do not dictate otherwise, NSSCP seems more practical for general use than the other estimators discussed to this point.

Why would we want to base our ranking or numerical estimate of relative importance on the marginal variance, zero-order correlation, or anything other than or in addition to the coefficient estimates in the model itself? It seems clear that for the Hoffman (1960) SCCP estimator (and its variants) in particular that the inclusion of the zero-order correlation in computing the relative importance estimator was not motivated by any perceived conceptual need to include the zero-order correlation as a component of relative importance, but was instead based on the empirical relationship $\Sigma_k b_k r_k = R^2$. The appropriateness of including the marginal variance, as in the Feldman estimator, may be questioned on similar grounds. On a purely conceptual level, it seems intuitively sensible to consider using only the (here standardized) coefficients produced by the model to assess relative importance, and in fact, when not trying to *numerically* decompose the explained variance into shares attributable to the individual predictors, the common practice *is* to use the standardized coefficient as a criterion for the relative importance of a predictor in the model. For uncorrelated predictors, $R^2 = \Sigma_k (b_k)^2$; but when the predictors are correlated, this relationship no longer holds. Otherwise, however, the squared standardized coefficient $\text{SSC}_k = (b_k)^2$ does (rather obviously) satisfy the criteria for an estimator of relative importance: it is always nonnegative, it is always zero when b_k is zero, and always nonzero when b_k is nonzero (this last possibly requiring a degree of numerical precision—places after the decimal point—higher for $(b_k)^2$ than for b_k).

An alternative to $(b_k)^2$, still based only on the standardized coefficients in the model, would be to use the normed squared standardized coefficient $\text{NSSC}_k = (b_k)^2 R^2 / \Sigma_k (b_k)^2$. It is readily apparent that the construction of this estimator parallels the construction of the NSCCP measure described earlier; it simply involves taking the formula for the NSCCP, substituting b_k for r_k, and dropping the absolute value signs (since, with only squared terms in the equation, it is no longer necessary to take absolute values). As indicated by a reviewer of an earlier version of this letter, one could in principle construct any number of estimators of this form, $M_k R^2 / \Sigma_k M_k$, where M_k is some expression, most likely a function of one or more parameter estimates, specific to each predictor X_k; and hence for NSCCP, $M_k = |b_k r_k|$ and for NSSC, $M_k = (b_k)^2$. Although it is true, as suggested by the reviewer, that one can construct any number of arbitrary estimators using this approach, not all estimators are arbitrary. The specification of M_k matters for this family of estimators, and both NSCCP and NSSC are based on quantities with some conceptual or empirical appeal themselves as estimators of relative importance.

The normed squared standardized coefficient estimator $\text{NSSC}_k = (b_k)^2 R^2 / \Sigma_k (b_k)^2$ ignores marginal correlations and preserves the ordering in terms of magnitude of the standardized regression coefficients, effectively rendering it redundant with the standardized coefficient *for ranking alone*; but NSSC also provides us with an estimator of the *numerical value* of the contribution of each predictor to the explained variance, which satisfies the rule that the shares attributed to each predictor sum to the explained variance, something the standardized coefficient by itself (except when all predictors are uncorrelated) does not do. If only the *ordering* of the predictors, and not the quantitative calculation of their contribution to the explained variance, is of interest, then one can simply use the standardized coefficients (squared or in absolute values: $(b_k)^2$ or $|b_k|$) themselves; it is only if one is interested in a quantitative estimate of the contribution to the explained variance attributable to each predictor that NSSC is of interest. NSSC thus does provide an estimator which (1) provides estimates for shares of explained variance that sum to R^2, (2) ensures that the share of the explained variance attributed to each predictor is always nonnegative, (3) satisfies the exclusion principle, since $(b_k)^2 R^2 / \Sigma_k (b_k)^2 = 0$ whenever $b_k = 0$, and (4) satisfies the inclusion principle, since $(b_k)^2 R^2 / \Sigma_k (b_k)^2 \neq 0$ whenever $b_k \neq 0$. As with NSCCP, computational effort is minimal.

Additionally, if R^2 is being calculated from sample data, and an estimator of relative importance in the population (taking into account the bias of R^2 in estimating the population explained variance) is desired, it is straightforward to substitute R^2_{adj} into the equations for either NSCCP or NSSC. NSCCP and NSSC can thus be used to partition either the calculated sample explained variance (R^2) or the estimated population explained variance (R^2_{adj}) with equal facility. In essence, $b_k r_k / \Sigma_k |b_k r_k|$ or $(b_k)^2 / \Sigma_k (b_k)^2$ is the unitless core measure of relative importance, and multiplication by R^2 or R^2_{adj} expresses it in terms of the proportion or percentage of variance explained. For that matter, one can in practice take $b_k r_k / \Sigma_k |b_k r_k|$ or $(b_k)^2 / \Sigma_k (b_k)^2$ as a general unscaled measure for relative importance, and multiply it by *any* reasonable measure of explained variance or explained variation [e.g, the likelihood ratio R^2 analog in logistic regression, R^2_L, or an index of predictive efficiency such as τ_p; see Menard (2000) for a discussion of explained variation measures in logistic regression analysis], and the measure of explained variation can be decomposed into components attributable to the respective predictors. To justify the use of NSCCP for measures of explained variation which are not variance-based (e.g., a proportional reduction in the log-likelihood), however, one must first decide whether it makes sense to use a variance-based estimator (since $b_k r_k$ is a component of explained *variance*) in conjunction with such measures of explained *variation*. NSSC may be easier to justify extending to logistic regression analysis, since standardized logistic regression coefficients may or may not be based on variance in the dependent variable (Menard 2004). The ability to extend the use of $b_k r_k / \Sigma_k |b_k r_k|$ and $(b_k)^2 / \Sigma_k (b_k)^2$ to measures of explained variance or variation other than R^2, particularly R^2_{adj}, is here regarded as a virtue of these measures, one shared by the LMGK estimator, as indicated by Menard (2004), but a characteristic which is not readily apparent for the Feldman (2005) estimator as described in Grömping (2007).

The Feldman (2005) estimator of relative importance fulfills all four of the criteria listed by Grömping (2007) for an estimator of relative importance, but has a high computational burden, and its applicability beyond the OLS R^2 is not readily apparent. The LMGK estimator of relative importance satisfies three of the four criteria for an estimator of relative importance, but may not satisfy the exclusion criterion, can be extended beyond the OLS R^2 and has a lower but still substantial computational burden. The issue of computational burden is not trivial; many statistical applications involve models with large numbers of predictors, and unless an estimator of relative importance is able to accommodate such models, it is unlikely to receive widespread use. This is not a problem for two other estimators of relative importance, NSCCP and NSSC.

The normed standardized coefficient x correlation estimator $\text{NSCCP}_k = |b_k r_k| R^2 / \Sigma_k |b_k r_k|$ preserves the ordering in terms of magnitude (ignoring sign) of the products in Hoffman's (1960) $\text{SCCP}_k = b_k r_k$. Relative to SCCP, when there is at least one negative value of $b_k r_k$, $|b_k r_k| R^2 / \Sigma_k |b_k r_k|$ will result in estimators of relative importance that are smaller in absolute value (because they were, arguably, inflated by the presence of a negative estimate of the contribution to the explained variance). Most typically, negative values of $b_k r_k$ will represent near-zero effects, in which case the negative value of $b_k r_k$ may represent either a suppressor effect involving a reversal of sign between r_k and b_k, or may simply represent the fact that variance *estimates*, unless constrained a priori to be positive, may take on negative values when the true value of the contribution of the predictor to the explained variance is equal or near to zero in a random sample. The use of $|b_k r_k| R^2 / \Sigma_k |b_k r_k|$ effectively constitutes such an a priori constraint on the estimate of the contribution to the explained variance. Given the fact that $\text{NSCCP}_k = |b_k r_k| R^2 / \Sigma_k |b_k r_k|$ satisfies the criteria of (a) assigning shares of the explained variance to each predictor such that the sum of the shares is equal to R^2 (or, if preferred, R^2_{adj} or some other measure of explained variation), (b) producing nonnegative estimates of the contribution to R^2, (c) being zero when $b_k = 0$, and (d) being nonzero when $b_k \neq 0$ (if also $r_k \neq 0$; and one can decide first whether $b_k r_k = 0$ only because $r_k = 0$ and whether, if so, this requires the use of an alternative estimator); coupled with (e) the computational simplicity of $|b_k r_k| R^2 / \Sigma_k |b_k r_k|$ compared to the alternatives described by Grömping (2007), the use NSCCP deserves more serious consideration as an estimator of relative importance, at least the same consideration given to the LMGK estimator (which violates the exclusion principle) and the Feldman estimator [which imposes the greatest computational burden of the measures considered here and by Grömping (2007)].

Part of that consideration, however, needs to be whether, conceptually, one really wants or needs to introduce the zero-order correlation (or, in the case of the Feldman estimator, the marginal variance) as a component in assessing the relative importance of a predictor. When one is simply ranking predictors (and not attempting to provide a numerical estimate of their contribution to explained variance), particularly when those predictors are measured on different scales, one routinely uses the standardized regression coefficient, without reference to marginal variances or correlations. More consistent with this approach than either the LMGK, Feldman, or NSCCP estimators would be the use of $\text{NSSC}_k = (b_k)^2 R^2 \Sigma_k (b_k)^2$, which is based *only* on the standardized coefficient, but which satisfies *all* of the criteria suggested by Grömping (2007) for an estimator of relative importance, *and*, along with NSCCP, imposes the lowest computational burden of any of the estimators. Furthermore, like NSCCP, NSSC is easily extended to the decomposition of R^2_{adj}, and even more than NSCCP, NSSC lends itself to models beyond ordinary least squares multiple regression, in particular logistic regression. For these reasons, NSSC seems the most promising choice for an estimator of relative importance in linear regression and related models for the relationship between a single outcome variable and multiple predictors, with NSCCP possibly a viable alternative if one believes there is a sound conceptual reason for incorporating the zero-order correlation into an estimator of the relative importance of the predictors.

Scott Menard
Sam Houston State University

REFERENCES

Feldman, B. (2005), "Relative Importance and Value," unpublished manuscript (Version 1.1, March 19, 2005). Available online at *http://www.prismanalytics.com/docs/RelativeImportance050319.pdf*.

Grömping, U. (2007), "Estimators of Relative Importance in Linear Regression Based on Variance Decomposition," *The American Statistician*, 61, 139–147.

Hoffman, P. J. (1960), "The Paramorphic Representation of Clinical Judgment," *Psychological Bulletin*, 57, 116–131.

Kruskal, W. (1987), "Relative Importance by Averaging Over Orderings," *The American Statistician*, 41, 6–10.

Kvålseth, T. O. (1985), "Cautionary Note About R^2," *The American Statistician*, 39, 279–285.

Lindeman, R. H., Merenda, P. F., and Gold, R. Z. (1980), *Introduction to Bivariate and Multivariate Analysis*, Glenview, IL: Scott, Foresman.

Menard, S. (2000), "Coefficients of Determination for Multiple Logistic Regression Analysis," *The American Statistician*, 54, 17–24.

——— (2004), "Six Approaches to Calculating Standardized Logistic Regression Coefficients," *The American Statistician*, 58, 218–223.

Tabachnick, B. G., and Fidell, L. S. (2007), *Using Multivariate Statistics* (4th ed.), Boston: Allyn and Bacon.

Tatsuoka, M.M. (1971), *Multivariate Analysis: Techniques for Educational and Psychological Research*, New York: Wiley.

Wherry, R. J., Sr. (1931), "A New Formula for Predicting the Shrinkage of the Coefficient of Multiple Correlation," *Annals of Mathematical Statistics*, 2, 440–457.

Response

I thank Dr. Menard for sharing his thoughts and welcome the opportunity for a discussion. First, I would like to reiterate an important aspect of my article that appears to have been overlooked by Menard. Although I have listed the exclusion criterion (a variable with $\beta_k = 0$ should be allocated a zero share) under desirability criteria because of Feldman's (2005) emphasis of this criterion, exclusion is by no means desirable for all relative importance questions that might be of interest. While it tends to be a reasonable criterion for prediction-related relative importance, it does not usually make sense if interest is in explanation-related relative importance (see also the discussion of Figure 1 in Section 2.2 of my article, pp. 140–141). I therefore disagree with Menard's general statement that Feldman's (2005) PMVD has more desirable properties than LMG (LMG = LMGK in Menard's notation).

In the following, I will discuss SCCP, NSCCP, and NSSC, that is, those of Menard's proposals that provide contributions with sum R^2. SCCP, the original proposal by Hoffman (1960), will be discussed first, and I will argue for preferring LMG and PMVD to SCCP and all other proposals currently used in the market. Then, I will comment on the suitability of modifying SCCP and the choice of modification made for NSCCP. Next, I will cite Johnson and Lebreton (2004) for a reason why I do not consider NSSC (or SSC) an appropriate single choice for a relative importance metric. Finally, I will discuss and illustrate the computational burden of calculating LMG and PMVD that appears to be Menard's main reason for proposing alternative methods. As in my article, I will argue in terms of estimands rather than estimates, as this contributes to clear focus on the meaning of metrics.

Like LMG and PMVD, SCCP *naturally* sums to R^2. The decision against including SCCP in my article was driven equally by space constraints as by SCCP's unfortunate property of sometimes producing negative shares of R^2 for some variables. LMG and PMVD deliver a *natural* decomposition of R^2 into non-negative shares, that is, they neither need normalization nor forced non-negativity. Thus, it is particularly plausible that they have an inherent rationale that is adequate for decomposition—even if that rationale may have not been fully understood so far. On the other hand, if a metric (like NSCCP or NSSC) has to be artificially normalized in order to deliver a decomposition of R^2, it is unlikely to have any nonarbitrary rationale as far as decomposition is concerned. Likewise, if a metric (like SCCP) delivers shares that sum to R^2 but can sometimes become negative, it does not suggest itself for a natural decomposition of R^2. Such a metric is likely to possess fundamental properties that render it unsuitable for the *decomposition* purpose even in specific situations where no negative shares have occurred. I have found in simulations (not published) that SCCP shares are even more variable than PMVD shares (which were shown in my article to be substantially more variable than LMG shares). This is a further disadvantage of SCCP. Hence, LMG and PMVD are the only methods currently used in the market that are natural choices for decomposing R^2. SCCP might have a natural purpose other than decomposition in relation to R^2, even though such purpose has not been discovered so far. A possible trigger for further thoughts into this direction may come from game theory: as Feldman (2005) pointed out, not only LMG and PMVD have game-theoretic justifications (see also my original article), but also SCCP (called covariance decomposition by Feldman), which is a sample application of the Aumann-Shapley pricing theory (see Feldman for details).

NSCCP is introduced by Menard as a remedy against the negativity of certain SCCP$_k$: non-negativity is achieved by using $|r_k b_k|$ and normalizing to the correct sum again. Apart from general reservations against artificial normalization, there is also a flaw in the justification of NSCCP. Menard—restricting his considerations to empirical quantities and application experience—argues that negative contributions SCCP$_k$ will typically be small and should be considered as unbiased estimates for an actual zero. This is inappropriate: SCCP$_k = r_k b_k$ for the kth regressor is, of course, a consistent estimator for the product of the true underlying zero-order correlation ρ_k with the true underlying standardized coefficient β_k (using Menard's notation). These quantities can and do have opposite signs in various scenarios. Theoretical examples can easily be constructed (see, e.g., Johnson and Lebreton 2004) and practical examples can also be found [see, e.g., the example data on Fertility in Switzerland used by Grömping (2006) for which the variable Agriculture has a positive zero-order correlation (p value < 0.05) and a negative coefficient (p value < 0.05), leading to an SCCP-value of -0.11]. Thus, SCCP$_k$ estimates a distinctly negative estimand at times. Setting aside general reservations against normalization and even restricting considerations to situations for which the estimand of SCCP$_k$ is zero due to either $\rho_k = 0$ or $\beta_k = 0$ or both, taking the absolute value of negative SCCP$_k$ does not seem to be an appropriate choice. It would appear to be more adequate to replace negative values by zero; but of course, that is not my recommendation either because the estimand may very well be negative and is presumably not suitable for the decomposition problem in general.

Menard brings up the conceptual question whether it is useful to include the zero-order correlation in addition to the standardized coefficient into the assessment of relative importance, that is, whether there is a reason for moving away from solely using the squared standardized coefficient or NSSC. Here, Johnson and Lebreton's (2004) relatively vague definition of relative importance as "the proportionate contribution each predictor makes to R^2, considering both its direct effect (i.e., its correlation with the criterion) and its effect when combined with the other variables in the regression equation" gives an answer. Following this definition, the squared standardized coefficient and NSSC are inappropriate, since they do not account for the direct effect of the predictor at all. Likewise, a quantification based solely on the zero-order correlation must be considered inappropriate. SCCP would be considered appropriate according to the Johnson and Lebreton definition, if no further criteria like non-negativity or inclusion were requested (note that Johnson and Lebreton heavily criticize SCCP). In summary, the squared standardized coefficient and NSSC do not adequately cover all aspects of relative importance and should therefore not be used as the single metric for measuring relative importance.

The computational burden of LMG and the even higher computational burden of PMVD remain a problem, of course. For cases with many variables, one may have to live with the fact that it is not practically feasible to compute LMG and PMVD, at least when treating each variable individually. However, with computers becoming more powerful it will be possible to tackle larger problems. Currently, on a Windows PC (Core2Duo E6600 with 2.4GHz) using R-package `relaimpo` (Grömping 2006) for calculating point estimates without bootstrapping for confidence intervals, six minutes CPU time are sufficient for calculating LMG contributions for 20 variables or PMVD contributions for 16 variables. PMVD would need about 1,650 minutes for 20 variables on this computer. Note that these relatively favorable results cannot be expected to carry over to logistic regression, for instance, since the R-package `relaimpo` exploits the fact that a linear regression can be calculated solely based on the empirical covariance matrix of the response and all regressors. If one is willing to combine variables into prespecified groups of related variables and to decompose R^2 just between such groups, the computational burden is driven by the number of groups instead of the number of variables, so that more individual variables can be accommodated. It is also possible to "adjust out" some baseline variables and to only decompose the remaining R^2 among the remaining variables, if appropriate. Thus, it is possible to handle many problems of realistic size on a modern computer. If, nevertheless, the application is too large for employing LMG or PMVD, I recommend looking at several of the simple metrics discussed by Darlington (1968) in comparison—among them SSC and zero-order correlation—preferably without artificially transforming them into a decomposition of R^2. Whenever it is feasible to compute LMG and PMVD, I would certainly make these metrics my first choice, with a preference for LMG in case of explanation-related relative importance questions.

Ulrike Grömping
TFH Berlin—University of Applied Sciences

REFERENCES

Darlington, R. B. (1968), "Multiple Regression in Psychological Research and Practice," *Psychological Bulletin*, 69, 161–182.

Feldman, B. (2005), "Relative Importance and Value," manuscript (Version 1.1, March 19 2005). Available online at *http://www.prismanalytics.com/docs/RelativeImportance050319.pdf*.

Grömping, U. (2006), "Relative Importance for Linear Regression in R: The Package relaimpo," *Journal of Statistical Software*, 17, 1. Available online at *http://www.jstatsoft.org/counter.php?id=167&url=v17/i01/v17i01.pdf&ct=1*.

Hoffman, P. J. (1960), "The Paramorphic Representation of Clinical Judgment," *Psychological Bulletin*, 57, 116–131.

Johnson, J. W., and Lebreton, J. M. (2004), "History and Use of Relative Importance Indices in Organizational Research," *Organizational Research Methods*, 7, 238–257.

Rosalsky, A. (2007), "A Simple and Probabilistic Proof of the Binomial Theorem," *The American Statistician*, 61, 161–162: Comment by Mukhopadhyay and Response

The title of Professor Rosalsky's recent article and the heading of its Section 2 were both specific and catchy. But, as I read it, I began wondering about what it was that Professor Rosalsky had truly proven.

I argue that in some sense he did not prove the *binomial theorem* as promised. I have also given some additional remarks. I hope that readers of *TAS* would benefit if they read Rosalsky (2007) and this letter side by side.

For completeness, let me restate the *binomial theorem*: For fixed but arbitrary real numbers a, b and a positive integer n, the following holds:

$$(a+b)^n = \Sigma_{j=0}^{n} \binom{n}{j} a^j b^{n-j}. \tag{1}$$

We interpret 0^j as 0 whatever be the integer $j > 0$ and identify 0^0 with 1.

I believe that in Section 2, Rosalsky (2007) *assumes* the following: There exists positive integers $C(n, j)$, $j = 0, 1, \ldots, n$ such that

$$(a+b)^n = \Sigma_{j=0}^{n} C(n, j) a^j b^{n-j}, \tag{2}$$

with fixed but arbitrary real numbers a, b and a positive integer n. Relying upon (2) or *assuming* (2), Professor Rosalsky had set out to identify $C(n, j)$ with $\binom{n}{j}$.

"By straightforward induction argument, for all $n \geq 1$," Rosalsky (2007) *claims* that (2) holds. So, one may object to my assertion that Rosalsky (2007) *assumes* (2). Let me explain. When I use induction to conclude (2), as a byproduct, I also see exactly what these $C(n, j)$'s are supposed to be. Thus, I feel uneasy about *claiming* (2) via "straightforward induction argument" and still pretending to wonder about what these $C(n, j)$'s may be. This may become a big issue and hence I rework briefly an earlier explanation given in response to a referee: Clearly, Equation (2) holds when $n = 1$ with $C(1, j) = \binom{1}{j}$, $j = 0, 1$. When $n = 2$, I note:

$$(a+b)^2 = (a+b)a + (a+b)b = a^2 + 2ab + b^2 = \Sigma_{j=0}^{2} \binom{2}{j} a^j b^{2-j}$$
$$\Rightarrow C(2, j) = \binom{2}{j}, \quad j = 0, 1, 2. \tag{3}$$

Next, assume that $(a+b)^k = \Sigma_{j=0}^{k} \binom{k}{j} a^j b^{k-j}$ for fixed k. Then, since I have $(a+b)^{k+1} = (a+b)^k a + (a+b)^k b$, by simply combining similar terms, I express $(a+b)^{k+1}$ as follows:

$$\Sigma_{j=0}^{k} \binom{k}{j} a^{j+1} b^{k-j} + \Sigma_{j=0}^{k} \binom{k}{j} a^j b^{k+1-j} = \Sigma_{j=0}^{k+1} C(k+1, j) a^j b^{(k+1)-j}$$
$$\text{with} \quad C(k+1, j) = \binom{k}{j} + \binom{k}{j-1}, \tag{4}$$

which proves (2). But, $C(k+1, j)$ from (4) obviously simplifies to:

$$\frac{k!}{j!(k-j)!} + \frac{k!}{(j-1)!(k-j+1)!} = \frac{k!}{j!(k-j+1)!}[(k-j+1)+j]$$
$$= \frac{k!(k+1)}{j!(k+1-j)!} = \binom{k+1}{j}. \tag{5}$$

Note that this simple process of induction identifies $C(n, j)$'s both explicitly and right away in one line under Equation (4)! Of course, I may pretend I do not see (5) and then move along to identify the $C(n, j)$'s, but why? What is the urgency for a full-page of explanation in *TAS* to identify the $C(n, j)$'s when the same thing can be accomplished trivially in one single line *assuming* (or *under*) Equation (4)? One only needs to hop to (5) from (4). How difficult or nonelegant is that?

Perhaps Professor Rosalsky has some other method of induction in mind to come up with (2) where $C(n, j)$'s happen to naturally hang out there unknown *not involving* a, b. And if he does, would a reader use that other possible induction or the one that is shown here? I do not know about others, but personally I would take a simpler proof any day.

The bottom line is this: The statements made in (1) and (2) are equivalent. In all fairness, instead of proving the *binomial theorem*, it is my understanding that Rosalsky (2007) has proved that (2) ⇒ (1) only. But, again is that not trivial in view of a one-line justification of (5) from (4)? A major purpose of this letter is to caution others, if I may: Be careful about overly flashy claims or jargons in publications since these may just mislead readers by masking one's message.

The next set of comments addresses Rosalsky's (2007) Equation (6). I will take the liberty to rewrite as follows: Let d_j, $j = 0, 1, \ldots, n$, be fixed real numbers and n be a fixed positive integer. Then, I have:

$$\Sigma_{j=0}^{n} d_j x^j = 0 \quad \text{for all} \quad x, 0 < x < \infty \Rightarrow d_j = 0, \quad j = 0, 1, \ldots, n. \tag{6}$$

Rosalsky (2007) appealed to "zero polynomials" to arrive at the conclusion laid out in (6). But, from my own experience, I have observed that a large part of the undergraduate population has not heard about what it is that these "zero polynomials" are supposed to do. To them, the conclusion stated in (6) may not be nearly as obvious. Hence, I share two other methods to resolve this matter which I have discussed in my classes with varying degrees of success depending on the level of sophistication of my students.

Resolution 1: This is aimed at students with some knowledge of elementary calculus. A polynomial $f(x) = \Sigma_{j=0}^{n} d_j x^j$, $0 < x < \infty$, is right-continuous at 0, and hence $\lim_{x \downarrow 0} f(x) = d_0$, but this must also coincide with 0 since $f(x)$ is identically 0 in any arbitrary right-hand neighborhood of 0. Hence, I claim that $d_0 = 0$. Thus, I have:

$$f(x) = \Sigma_{j=1}^{n} d_j x^j = x(\Sigma_{j=1}^{n} d_j x^{j-1}) = 0 \quad \text{for all} \quad x, 0 < x < \infty$$
$$\Rightarrow \Sigma_{j=1}^{n} d_j x^{j-1} = 0 \quad \text{for all} \quad x, 0 < x < \infty.$$

I can use my previous argument again to claim that $d_1 = 0$. This way one can successively verify that $d_j = 0$, $j = 2, 3, \ldots, n$, the required conclusion stated in (6).

Resolution 2: This is aimed at students with some knowledge of elementary linear algebra. It is straightforward to claim that the set of all polynomials of degree $k \leq n$ in x is a vector space with n fixed. Clearly, the set of vectors $\mathcal{S} = \{1, x, x^2, \ldots x^n\}$ can generate any polynomial of degree $k \leq n$ in x. On the other hand, suppose that a vector is left out from this set $\mathcal{S}$, for example, x^2. Then, the polynomial such as x^2 or $x + 3x^2$, for example, can no longer be generated by the remaining vectors in $\mathcal{S} - \{x^2\}$. In other words, the set of vectors $\mathcal{S}$ is a minimal generator for the vector space of all polynomials in x of degree $k \leq n$. Hence, the vectors in $\mathcal{S}$ ought to be linearly independent. This will lead to the conclusion laid out in (6).

Nitis Mukhopadhyay
University of Connecticut

Response

I thank Professor Nitis Mukhopadhyay for having taken an interest in my article and for the comments and insight he provided. Professor Mukhopadhyay took issue with various aspects of the proof of the binomial theorem that I presented and I am pleased to respond to his concerns.

Professor Mukhopadhyay's letter presented the standard well-known proof of the binomial theorem and felt that this standard proof is far better than the one I presented. I respect his viewpoint, but I stand by the validity and usefulness of my proof. There were not any "overly flashy claims or jargons" in my article which "may mislead readers" as suggested by Professor Mukhopadhyay.

In my article, I indicated that relation (2) of Rosalsky (*TAS*, 2007) follows by a straightforward induction argument without offering details. The details parallel those of the beginning of the standard proof of the binomial theorem: It is clear that $C(1,0) = C(1,1) = 1$ and by induction, one readily obtains

$$\begin{aligned} C(n+1,j) &= C(n,j-1) + C(n,j), \quad 1 \le j \le n \\ C(n+1,0) &= C(n,0), \quad C(n+1,n+1) = C(n,n) \end{aligned}$$

thereby yielding relation (2) of Rosalsky (*TAS*, 2007).

Now to complete the proof of the binomial theorem, the exact values of the $\{C(n,j) : 0 \le j \le n\}$ need to be explicitly identified. The standard proof of the binomial theorem is somewhat technical and detail oriented in that one has to know *in advance* exactly what to look for in the $C(n,j)$, namely combinatorial coefficients. Then the completion of the argument indeed proceeds very quickly as Professor Mukhopadhyay made abundantly clear. However, in my proof, by appealing to the binomial distribution, the exact values of the $\{C(n,j) : 0 \le j \le n\}$ jump right out at the *end* of the proof without requiring any algebraic manipulation at all involving the $\{C(n,j) : 0 \le j \le n\}$ or the combinatorial coefficients in the binomial distribution. This is a key feature (and I believe an advantage) of my proof which, in contradistinction to the standard proof, can thus be used by the instructor who would like her or his students to actually discover the binomial theorem for themselves in the process of deriving the exact form for the expansion of positive integer powers of sums of two terms. I readily acknowledge in having been remiss in my failing to emphasize this matter in my article.

I agree with Professor Mukhopadhyay that readers of *TAS* should take a look at my article and his letter side by side. The reader can then decide for herself or himself as to which proof is preferable for classroom presentation.

Professor Mukhopadhyay also took issue concerning my assertion that relation (6) of Rosalsky (*TAS*, 2007) guarantees that all the coefficients are zero. This is immediate because the left-hand side of (6) is the zero polynomial. Apparently Professor Mukhopadhyay felt that some additional detail should be included which I gladly offer: If not all the coefficients are zero, then there is a largest $j = j_0 \in [0,n]$ for which that coefficient is nonzero; letting $x \to \infty$ leads to an immediate contradiction (even if $j_0 = 0$) since the right-hand side of relation (6) of Rosalsky (*TAS*, 2007) is identically zero. Of course, the instructor may find it to be helpful to some students to provide the above few lines of additional detail to tie up the argument, but including such details in the article itself should not be necessary for virtually all readers of a journal of the caliber of *TAS*.

Professor Mukhopadhyay so kindly offered two interesting and alternative ways of tying up the argument, but I feel that they are not any simpler or better than what I have offered.

Let me conclude by saying that I have never liked the standard proof of the binomial theorem ever since I first saw it during my high school days in the mid 1960s. I say this because the standard proof requires no real understanding of $\binom{n}{j}$ apart from the formula

$$\binom{n}{j} = \frac{n!}{(n-j)!j!}, \quad 0 \le j \le n.$$

The proof in my article does not have this (negative) feature. Although the binomial theorem is definitely *not* a deep mathematical result, it is extremely important. For example, it is a key tool used in the proof that

$$\sum_{n=0}^{\infty} \frac{1}{n!} = \lim_{n\to\infty} \left(1 + \frac{1}{n}\right)^n.$$

Any new proof of it should be most welcome and will almost surely have both its proponents and its critics. I am most grateful to the Editorial Board of *TAS* for letting me share my perspective and also for letting Professor Mukhopadhyay share his.

Andrew Rosalsky
University of Florida