PSYCHOMETRIKA—VOL. 74, NO. 1, 97–105
MARCH 2009
DOI: 10.1007/S11336-008-9080-1

A NOTE ON COMPARING THE ESTIMATES OF MODELS FOR CLUSTER-CORRELATED OR LONGITUDINAL DATA WITH BINARY OR ORDINAL OUTCOMES

DANIEL J. BAUER

UNIVERSITY OF NORTH CAROLINA AT CHAPEL HILL

When using linear models for cluster-correlated or longitudinal data, a common modeling practice is to begin by fitting a relatively simple model and then to increase the model complexity in steps. New predictors might be added to the model, or a more complex covariance structure might be specified for the observations. When fitting models for binary or ordered-categorical outcomes, however, comparisons between such models are impeded by the implicit rescaling of the model estimates that takes place with the inclusion of new predictors and/or random effects. This paper presents an approach for putting the estimates on a common scale to facilitate relative comparisons between models fit to binary or ordinal outcomes. The approach is developed for both population-average and unit-specific models.

Key words: categorical data, mixed model, multilevel model, ordinal, binary.

Clustered and longitudinal data are quite common in psychology, social science, and health research. A common feature of both data structures is dependence among the observations within units. Two general approaches for modeling dependent data are to fit a *population-average model* that permits the model residuals to correlate within units, for instance, using Generalized Estimating Equations (GEE) (Liang & Zeger, 1986), or to fit a *unit-specific model* with random effects, as in the Generalized Linear Mixed Model (GLMM) (McCulloch & Searle, 2001). These models are also referred to as marginal and conditional models, respectively.

For linear models, a strategy often employed in practice is to begin an analysis by first fitting a relatively simple model to the data. Complexity is then added to the model in steps. When fitting linear mixed models, for instance, it is common for an analysis to begin with the partitioning of variance into within- and between-group components via a so-called "null" or "empty" model (absent fixed predictors). In subsequent steps, predictors at the individual and/or group levels are added to the model, and the proportional reduction in the variance component estimates may be calculated as an index of effect size (e.g., 20% of between-group variance explained as a function of a set of group-level predictors; Raudenbush & Bryk, 2002, but see Snijders & Bosker, 1994, for warnings about this practice). An investigator may also wish to include certain key predictors in the model initially and then observe how the effects of these predictors change with the entrance of other covariates to the model, for instance when assessing mediation.

For linear (continuous) outcomes, this kind of model building is straightforward because the coefficient and variance estimates can be compared directly from one fitted model to the next. In contrast, direct comparisons between models are often not possible when the outcome is binary or ordered-categorical. The problem is that the addition of a new predictor or random effect to the model implicitly rescales the coefficients of the prior predictors as well as the variances and covariances of any random effects. As such, some changes in the model coefficients represent differences in scale as opposed to substantively meaningful changes brought about by the inclusion

Requests for reprints should be sent to Daniel J. Bauer, Department of Psychology, University of North Carolina, Chapel Hill, NC 27599-3270, USA. E-mail: dbauer@email.unc.edu

of the additional effects. This rescaling makes it difficult to compare the results of sequentially fit models for binary or ordered-categorical outcomes.

A solution to this problem was developed by Winship and Mare (1983, 1984) for probit and logit models fit to nonclustered (independent) data. The current paper extends this work to take account of the more complex variance structures of models for cluster-correlated or longitudinal data.

1. Scaling with Binary or Ordinal Outcomes

Let us begin by considering a population-average model for an ordinal outcome. Consider an ordered categorical response y_{ij} for the ith observation within cluster j with possible observed values $c = 1, \ldots, C$. A model for the cumulative probabilities of the category responses may be written as

$$g\big[P(y_{ij} \le c)\big] = \upsilon^c - \mathbf{x}'_{ij}\boldsymbol{\beta}, \quad c = 1, \ldots, C-1 \tag{1}$$

where g is the link function, for instance, logit, probit, or complementary log–log, and $\mathbf{x}_{ij}$ is a vector of covariates. Included in $\mathbf{x}_{ij}$ is the regression constant, producing an intercept within the vector of fixed effects $\boldsymbol{\beta}$. To identify this intercept, we set $\upsilon^1 \equiv 0$. When y is binary, this term thus drops out and the usual models for binary outcomes are obtained. Other parameterizations of this model are also possible.

Additionally, some way must be found to accommodate the dependence within clusters. With the GEE approach, this is typically done through the specification of a working correlation matrix for the Pearson residuals. For instance, a compound symmetric working correlation matrix might be specified for hierarchically clustered data, or an autoregressive or unstructured working correlation matrix might be specified for longitudinal data.

The unit-specific modeling approach, i.e., GLMM, differs from the GEE approach in the method used to accommodate dependence due to clustering. Specifically, rather than posit a working correlation matrix for the observed residuals, dependence is modeled through the introduction of random effects. Equation (1) is thus modified to be

$$g\big[P(y_{ij} \le c)\big] = \upsilon^c - \big(\mathbf{x}'_{ij}\boldsymbol{\beta} + \mathbf{z}'_{ij}\mathbf{u}_j\big) \tag{2}$$

where $\mathbf{u}_j$ is the vector of random effects for the covariates $\mathbf{z}_{ij}$. It is assumed that $\mathbf{u}_j \sim N(\mathbf{0}, \boldsymbol{\Sigma}_{\mathbf{u}})$. Further, the observed responses y_{ij} are typically assumed to be independent, conditional on the covariates and random effects.

One way to motivate the above models is by considering the observed variable y to be a coarse representation of an underlying continuous latent variable y^*. This is a particularly natural way to think about ordinal outcome variables (McCullagh & Nelder, 1989, pp. 151–155), and is also applicable for many binary outcomes. Pursuing this idea, we can specify the relationship between y^* and y through a threshold model

$$y_{ij} = c \quad \text{if } \upsilon^{c-1} < y^*_{ij} \le \upsilon^c. \tag{3}$$

The thresholds are assumed to be strictly increasing and the first and last thresholds are defined as $\upsilon^0 \equiv -\infty$ and $\upsilon^C \equiv \infty$, respectively.

Because y^* is continuous, it (unlike y) can be expressed as a linear function of the covariates. Corresponding to the population-average model, we can therefore write

$$y^*_{ij} = \mathbf{x}'_{ij}\boldsymbol{\beta} + r_{ij} \tag{4}$$

where r is the error of prediction. Alternatively, the unit-specific model would be written

$$y_{ij}^{*} = \mathbf{x}_{ij}'\boldsymbol{\beta} + \mathbf{z}_{ij}'\mathbf{u}_j + r_{ij}. \tag{5}$$

In either case, y^* is never actually observed, so the scale of r is arbitrary. Choosing the distribution of r to be the standard logistic distribution, with variance $\pi^2/3$ is equivalent to defining g in (1) or (2) to be the logit link function. Alternatively, choosing the distribution of r to be standard normal, with variance 1, is equivalent to defining g to be the probit link function. Assuming instead an extreme value (Gumbel) distribution for $-r$, with mean equal to the Euler constant and variance $\pi^2/6$, gives rise to the complementary log–log model.

The idea of a underlying response variable is popular within psychometrics, econometrics, and many social sciences, but it is by no means necessary for formulating the models in (1) and (2). Nor will it always be appropriate to assume an underlying response variable, particularly for some binary outcomes in the health sciences (i.e., presence or absence of disease). Nevertheless, this notion can help to clarify the implicit rescaling that takes place as new predictors (or random effects) are entered into the model (Snijders & Bosker, 1999, pp. 225–226). Before expanding on this point, however, let us first consider how this implicit rescaling complicates the assessment of sequentially fit models.

The linear model provides a useful comparison. If the dependent variable y is continuous, then the residual variance, σ_r^2, can be estimated directly from the data. Further, as new predictors enter the model, this residual variance shrinks as more of the variance in y is accounted for by the predictors. The total variance of y remains constant, setting a constant scale for the model estimates. Comparing between fitted models is therefore straightforward. In contrast, such comparisons are not simple when the outcome is binary or ordered-categorical. The difficulty is that the metric of y^* is unknown and so σ_r^2 cannot be estimated from the data. Instead, it is fixed by the choice of distribution (e.g., $\pi^2/3$ for the logistic distribution). Since the variance of r is constant, the marginal variance of y^* must increase in proportion to the variance accounted for by the new covariates (or random effects) included in the model. This change in the metric of y^* then rescales the estimates, complicating the comparison of estimates between models (Fielding, 2004; Snijders & Bosker, 1999, pp. 225–226).[1]

To illustrate this difference in the behavior of the two models, let us consider a simulated data example. Using SAS 9.1, data was generated for 1,000 clusters randomly varying in size from 1 to 10 observations (with equal probability). In the first step, values for the underlying latent variable y^* were generated from a linear mixed model:

$$y_{ij}^{*} = \beta_0 + \beta_1 x_{1ij} + \beta_2 x_{2ij} + \beta_3 x_{3j} + u_j + r_{ij}. \tag{6}$$

In the second step, the binary variable y was generated from y^* as follows

$$y_{ij} = 1 \quad \text{if } y_{ij}^{*} > 0; \qquad \text{else } y = 0. \tag{7}$$

In the population, $\beta_0 = 0$, $\beta_1 = \beta_2 = \beta_3 = 1$, u_i and r_{ij} are independent and normally distributed with means of zero and variances of $\sigma_u^2 = 0.03$ and $\sigma_r^2 = 0.15$, respectively. The covariates x_1 and x_2 are independent and normally distributed with means of 0 and variances of 0.4. The intraclass correlation coefficients (ICCs) for x_1 and x_2 are both 0.25 in the population. The cluster-level covariate x_3 is normally distributed with a mean of zero and a variance of 0.02 and was simulated independently of x_1 and x_2. The values of the population model were chosen for ease of

[1] Note that the difference in behavior between linear and generalized linear models can also be understand through consideration of geometric properties of the link function, as shown by Neuhaus and Jewell (1993), without appeal to the notion of an underlying latent variable.

interpretation: the predictors are uncorrelated, their effects are of equal magnitude, the marginal distribution of y_{ij}^* is a standard normal, and the marginal proportion for the "event" is 0.50.

Of interest here are three models. The initial model includes x_1 as the only explanatory variable, the second model adds x_2, and the third model also includes x_3. If the outcome variable were continuous, the comparison of these three models would be straightforward: given the lack of correlation between the predictors, the effect of x_1 should be unchanged through the inclusion of x_2, and the effects of x_1 and x_2 should be unaffected by including x_3; adding x_2 to the model should decrease both the between-cluster and within-cluster residual variances (in this case, by 0.1 and 0.3, respectively, to reflect the between-cluster and within-cluster variances of x_2); including the cluster-level covariate x_3 in the model should further decrease the between-cluster residual variance (in this case by 0.02), but leave unchanged the within-cluster residual variance. Indeed, just this pattern of results is obtained if linear mixed models are fit to the simulated data for the underlying latent response, as can be seen in the first section of Table 1. Identical results are obtained by fitting marginal models using GEE with a compound symmetric working corre-

TABLE 1.
Estimates for models fit to simulated data.

Model/Effects	Random effects models[a]			Marginal models[b]		
	Model 1	Model 2	Model 3	Model 1	Model 2	Model 3
Linear						
β_0	0.02	0.00	−0.01	0.02	0.00	−0.01
β_1	1.00	1.00	1.00	1.00	1.00	1.00
β_2		1.02	1.01		1.02	1.01
β_3			1.05			1.05
σ_u^2	0.15	0.05	0.03			
σ_r^2	0.46	0.15	0.15	0.61	0.20	0.18
Probit						
β_0	0.03	0.00	−0.02	0.03	0.00	−0.01
β_1	1.51	2.65	2.65	1.30	2.25	2.37
β_2		2.62	2.60		2.24	2.34
β_3			2.59			2.33
σ_u^2	0.36	0.38	0.24			
σ_r^2	1.00	1.00	1.00	1.00	1.00	1.00
Rescaled probit[c]						
β_0	0.02	0.00	−0.01	0.02	0.00	−0.01
β_1	1.01	1.02	1.02	1.01	1.01	1.02
β_2		1.00	1.00		1.01	1.00
β_3			1.00			1.00
σ_u^2	0.16	0.06	0.04			
σ_r^2	0.44	0.15	0.15	0.60	0.20	0.18

Note. With the exception of the linear models, values of σ_r^2 were not estimated and rather were determined either through the use of the probit link (fixed at 1) or through rescaling the results to transform the marginal variance of y^* to a fixed value.

[a]Random effects models were fit using maximum likelihood in SAS PROC MIXED (linear) or NLMIXED (probit). For probit models, maximum likelihood was conducted using numerical integration with 20 quadrature points.

[b]Marginal models were fit using generalized estimating equations in SAS PROC GENMOD with a compound symmetric (exchangeable) working correlation matrix for the Pearson residuals.

[c]Results reported for "Rescaled Probit" were obtained by rescaling the estimates from the probit models so that the implied marginal variance of y^* would be 1.

lation matrix, with the only exception that the residual variance is pooled rather than partitioned between levels.

For the binary y variable, however, the results of fitting the same three models, but using a probit formulation, are quite different. As can be seen in the second section of Table 1, regardless of the modeling approach, the effect of x_1 is considerably larger in magnitude in the second model than the first, despite the fact that the two predictors are uncorrelated. Likewise, the inclusion of x_3 produced an apparent increase in the effects of x_1 and x_2 when comparing between marginal models. This increase did not, however, occur for the random effects models. Last, a seemingly paradoxical pattern of effects emerges in the random-effects models: the inclusion of x_2 in Model 2 appears to slightly increase the cluster-level residual variance relative to Model 1, despite the fact that x_2 varies both within and between clusters and has a significant fixed effect. These "odd" results are entirely a consequence of the fact that the estimates of the two models are on different scales, making direct comparisons between the two sets of estimates difficult.[2]

2. Commensurate Scaling

Winship and Mare (1983, 1984) noted that one could rescale the results of a series of standard logit or probit models so that the variance of y^* would be held constant, enabling effect estimates to be compared between models. Their results are also directly applicable to population-average models for binary and ordinal responses.[3] Specifically, if (4) holds, then the marginal variance of y^* can be expressed as

$$V(y^*) = \boldsymbol{\beta}' \boldsymbol{\Sigma}_{\mathbf{x}} \boldsymbol{\beta} + \sigma_r^2 \tag{8}$$

where $\boldsymbol{\Sigma}_{\mathbf{x}}$ is the covariance matrix of $\mathbf{x}$, and σ_r^2 is determined by the choice of distribution for r. To transform this model so that the variance of the underlying response variable is equal to a specific value, say a, note that

$$\left(\boldsymbol{\beta}' \boldsymbol{\Sigma}_{\mathbf{x}} \boldsymbol{\beta} + \sigma_r^2\right)^{-1} V(y^*) = 1. \tag{9}$$

Therefore,

$$a\left(\boldsymbol{\beta}' \boldsymbol{\Sigma}_{\mathbf{x}} \boldsymbol{\beta} + \sigma_r^2\right)^{-1} V(y^*) = a \tag{10}$$

and the appropriate scaling factor is then defined as

$$s_1 = a^{1/2}\left(\boldsymbol{\beta}' \boldsymbol{\Sigma}_{\mathbf{x}} \boldsymbol{\beta} + \sigma_r^2\right)^{-1/2}. \tag{11}$$

The model in (4) can now be rescaled by multiplying each term by the scaling factor so that

$$s_1 y_{ij}^* = \mathbf{x}_{ij}'(s_1 \boldsymbol{\beta}) + s_1 r_{ij} \tag{12}$$

and the transformed underlying latent response variable $s_1 y_{ij}^*$ has a marginal variance equal to a. The rescaled coefficients are contained in the vector $s_1 \boldsymbol{\beta}$ and the rescaled residual variance is $s_1^2 \sigma_r^2$. If a set of models are fit in sequence, the results of each model can be rescaled

[2] One can also think of the results of the first model as biased due to the omission of the covariate x_2. This bias arises in many generalized linear models even when the omitted covariates are uncorrelated with the present predictors, as noted by Neuhaus and Jewell (1993).

[3] In previous work, Fielding (2004) proposed a different approach for rescaling the results of mixed-effects models for ordinal outcomes that involves computating conditional mean scores for each category of the outcome variable. This approach is, however, somewhat ad hoc and unlikely to work well for binary outcomes or ordinal outcomes with few categories. Accordingly, our focus here is on extending the Winship and Mare (1983, 1984) approach.

using a common a value by inserting sample estimates into (11). The coefficient estimates can then be compared directly from one model to the next just as they are for linear models.

Unit-specific models, such as the GLMM, introduce the further complication of random effects. The Winship and Mare (1983, 1984) approach remains applicable, but requires modification to account for the more complex variance structure (see Snijders & Bosker, 1999, pp. 225–226). With random effects in the model, e.g., (5), the variance of y^* is

$$V(y^*) = \boldsymbol{\beta}'\boldsymbol{\Sigma}_{\mathbf{x}}\boldsymbol{\beta} + \boldsymbol{\mu}_{\mathbf{z}}'\boldsymbol{\Sigma}_{\mathbf{u}}\boldsymbol{\mu}_{\mathbf{z}} + \mathrm{VEC}(\boldsymbol{\Sigma}_{\mathbf{z}})'\mathrm{VEC}(\boldsymbol{\Sigma}_{\mathbf{u}}) + \sigma_r^2 \tag{13}$$

where $\boldsymbol{\mu}_{\mathbf{z}}$ is the vector of means for $\mathbf{z}$, $\boldsymbol{\Sigma}_{\mathbf{z}}$ is the covariance matrix of $\mathbf{z}$, and the VEC operator places the elements of a $q \times q$ matrix into a $q^2 \times 1$ vector. The scaling factor for the GLMM is then

$$s_2 = a^{1/2}\left(\boldsymbol{\beta}'\boldsymbol{\Sigma}_{\mathbf{x}}\boldsymbol{\beta} + \boldsymbol{\mu}_{\mathbf{z}}'\boldsymbol{\Sigma}_{\mathbf{u}}\boldsymbol{\mu}_{\mathbf{z}} + \mathrm{VEC}(\boldsymbol{\Sigma}_{\mathbf{z}})'\mathrm{VEC}(\boldsymbol{\Sigma}_{\mathbf{u}}) + \sigma_r^2\right)^{-1/2}. \tag{14}$$

This is a new result of the current paper, although it is related to other expressions found in the literature (see Appendix for an informal derivation and references to related work).

Using this result, the rescaled GLMM is

$$s_2(y_{ij}^*) = \mathbf{x}_{ij}'(s_2\boldsymbol{\beta}) + \mathbf{z}_{ij}'(s_2\mathbf{u}_j) + s_2 r_{ij}. \tag{15}$$

Thus, the rescaled fixed-effect regression coefficients are obtained as $s_2\boldsymbol{\beta}$, the rescaled variance-covariance matrix of the random effects is $s_2^2\boldsymbol{\Sigma}_{\mathbf{u}}$, and the rescaled residual variance is $s_2^2\sigma_r^2$. In the course of fitting successive models, the sample estimates for the required terms can be inserted into (14) to hold the sample variance of the transformed underlying response variable at a constant a value. This rescaling permits the comparison of fixed effects as well as variance components as the model is made more complex.

Another difference between the population-average and unit-specific modeling approaches concerns the inclusion of cluster-level covariates. In the GLMM, the variance associated with omitted cluster-level covariates is absorbed by the random effects. Including observed cluster-level covariates reduces only the variances of the random effects, having no net impact on the marginal variance of y^*. Thus, direct model comparisons can be made for GLMMs that differ only in the inclusion of cluster-level covariates, even without rescaling the estimates as described above. The same is not true for population-average models. Because these models include no random effects, the inclusion of cluster-level covariates will impact the marginal variance of y^*. The results will then need to be put on a common scale if one wishes to compare the model coefficients.

To demonstrate these points, let us return to our simulated data example. To make valid comparisons between the three probit models, the estimates must be put on a common scale. In the third section of Table 1, the results of each model have been rescaled so that the model-implied marginal variance of y^* is 1. Not coincidentally, this is the scale of the underlying latent response variable to which we previously fit linear mixed models. Considering first the results obtained from the random effects models, it can be seen that the rescaled results match those reported for the linear models quite closely. Whereas the original probit estimates suggested an increase in the effect of x_1 in the presence of x_2, the rescaled estimates indicate that in fact the effect of x_1 is unchanged by controlling for x_2. Further, the paradoxical increase in random intercept variance associated with the addition of x_2 to the model is no longer observed—the rescaled results indicate that in fact both the cluster-level and individual-level residual variances decrease in magnitude when x_2 is included in the model. In contrast, it was previously noted that the inclusion of the cluster-level covariate x_3 had no effect on the raw coefficients obtained for the other covariates. The reason is that x_3 drew variance from the random intercept, and thus had

no net impact on the marginal variance of y^*. Since the marginal variance of y^* was constant between Models 2 and 3, these models could, in fact, be compared even without rescaling.

Turning next to the marginal model results, we again see from the rescaled results that the effect of x_1 is actually unchanged by entering x_2 into the model. Unlike the GLMM, the inclusion of x_3 in Model 3 also produced an apparent increase in the effects of x_1 and x_2. Because there is no between-cluster residual in Model 2 to account for omitted cluster-level covariates, the inclusion of x_3 changes the marginal variance of y^* and, therefore, also the scale of the estimated coefficients. Returning the coefficients to a common scale shows that the effects of x_1 and x_2 are unaffected by the addition of x_3 to the model.

3. Conclusion

The primary goal of this paper was to draw attention to the importance of scaling when comparing the results of sequential models fit to cluster-correlated or longitudinal data with binary or ordinal outcomes. The approach recommended here for comparing sequentially fit models is to rescale the model results so that the estimates will be commensurate between the fitted models. In some instances, such as the comparison of GLMMs that differ only in the inclusion of cluster-level covariates, this rescaling is unnecessary, but in many other cases comparing the magnitude of raw estimates will be misleading. This scaling issue is most easily appreciated by considering the observed binary or ordinal variable to be coarse categorization of an underlying continuous variable. Equating the model-implied variance for the underlying response variable places the estimates from any two models on the same scale. The relative difference between the coefficients can then be computed and interpreted straightforwardly. If desired, formal tests of the relative difference in the coefficients can also be obtained by bootstrapping the rescaled estimates, for instance to provide a test of mediation (e.g., Shrout & Bolger, 2002). In cases where the notion of an underlying continuous response variable is not sensible, the proposed rescaling approach may have less appeal.

Note that the rescaled estimates are intended to be used primarily for making relative comparisons between models. The absolute values of the rescaled estimates (i.e., in the third section of Table 1) are meaningful only in the metric of the underlying latent variable y^*. For instance, setting $a = 1$, each fixed effect would indicate standard deviations change in y^* per unit change in x. Often, one wishes to interpret the coefficients of the model more directly in terms of the observed y, for instance by computing odds or hazard ratios from exponentiated estimates when using the logit or complementary log–log links, respectively. The rescaled estimates can not be similarly exponentiated and interpreted. It is not necessary, however, to focus attention only on the rescaled estimates. One may interpret exponentiated raw coefficients when communicating the substantive significance of the estimates obtained from the final model, yet also use rescaled estimates to compute proportions of variance explained, or to evaluate evidence of mediation from the sequence of preceding models.

Appendix

What follows is an informal proof of (13). Equation (5) expresses the underlying response variable as a linear function of the fixed covariates and random effects:

$$y_{ij}^* = \mathbf{x}_{ij}'\boldsymbol{\beta} + \mathbf{z}_{ij}'\mathbf{u}_j + r_{ij}. \tag{16}$$

The variance of y^* is then

$$V\left(y_{ij}^*\right) = V\left(\mathbf{x}_{ij}'\boldsymbol{\beta} + \mathbf{z}_{ij}'\mathbf{u}_j + r_{ij}\right). \tag{17}$$

From the assumption of independence of r from $\mathbf{x}$ and $\mathbf{u}$, it follows that

$$V\left(y_{ij}^{*}\right) = V\left(\mathbf{x}_{ij}'\boldsymbol{\beta} + \mathbf{z}_{ij}'\mathbf{u}_j\right) + V(r_{ij}). \tag{18}$$

In addition, the covariance of $\mathbf{x}_{ij}'\boldsymbol{\beta}$ and $\mathbf{z}_{ij}'\mathbf{u}_j$ must be zero. This can be seen by choosing any two terms βx and zu and expressing their covariance as

$$\begin{aligned}\mathrm{COV}(\beta x, zu) &= E(\beta xzu) - E(\beta x)E(zu)\\ &= \beta E(xzu) - \beta E(x)E(zu).\end{aligned} \tag{19}$$

Since u is assumed to be independent of both x and z, this can be rewritten as

$$\mathrm{COV}(\beta x, zu) = \beta E(xz)E(u) - \beta E(x)E(z)E(u). \tag{20}$$

Since it is also assumed that $E(u) = 0$, it follows that

$$\mathrm{COV}(x\beta, zu) = E(u)\left[\beta E(xz) - \beta E(x)E(z)\right] = 0.$$

Given the independence of $\mathbf{x}_{ij}'\boldsymbol{\beta}$ and $\mathbf{z}_{ij}'\mathbf{u}_j$, we may simplify (18) to

$$V\left(y_{ij}^{*}\right) = V\left(\mathbf{x}_{ij}'\boldsymbol{\beta}\right) + V\left(\mathbf{z}_{ij}'\mathbf{u}_j\right) + V\left(r_{ij}\right). \tag{21}$$

Of these terms, $\mathbf{x}'\boldsymbol{\beta}$ is simply a weighted linear combination of random variables and it is well known that its variance is, therefore,

$$V\left(\mathbf{x}_{ij}'\boldsymbol{\beta}\right) = \boldsymbol{\beta}'\boldsymbol{\Sigma}_{\mathbf{x}}\boldsymbol{\beta}. \tag{22}$$

More difficult is the variance of $\mathbf{z}'\mathbf{u}$ since both terms are random variables.

Applying results in Goodman (1960), the variance of any given term zu can be shown to be

$$V(zu) = E(z)^2V(u) + E(u)^2V(z) + V(u)V(z) \tag{23}$$

$$= E(z)^2V(u) + V(u)V(z) \tag{24}$$

$$= \mu_z^2\sigma_u^2 + \sigma_z^2\sigma_u^2 \tag{25}$$

where the simplification in the second step arises due to the customary assumption that $E(u)$ is zero. The covariance between any two terms z_1u_1 and z_2u_2 can also be written as

$$\mathrm{COV}(z_1u_1, z_2u_2) = E(z_1u_1z_2u_2) - E(z_1u_1)E(z_2u_2). \tag{26}$$

Given the independence of u_1 and u_2 from z_1 and z_2, (26) may be simplified to

$$\mathrm{COV}(z_1u_1, z_2u_2) = E(z_1z_2)E(u_1u_2) - E(z_1)E(u_1)E(z_2)E(u_2). \tag{27}$$

Given the assumption that $E(u_1) = E(u_2) = 0$, the second term in (27) is zero and the first term may be expanded as

$$\mathrm{COV}(z_1u_1, z_2u_2) = E(z_1z_2)E(u_1u_2) \tag{28}$$

$$= \left[E(z_1)E(z_2) + \mathrm{COV}(z_1, z_2)\right]\left[\mathrm{COV}(u_1, u_2)\right] \tag{29}$$

$$= \mu_{z_1}\mu_{z_2}\sigma_{u_1u_2} + \sigma_{z_1z_2}\sigma_{u_1u_2}. \tag{30}$$

Generalizing the notation from (25) and (30) into matrix form produces the expression

$$V\left(\mathbf{z}'_{ij}\mathbf{u}_j\right) = \boldsymbol{\mu}'_{\mathbf{z}}\boldsymbol{\Sigma}_{\mathbf{u}}\boldsymbol{\mu}_{\mathbf{z}} + \mathrm{VEC}(\boldsymbol{\Sigma}_{\mathbf{z}})'\mathrm{VEC}(\boldsymbol{\Sigma}_{\mathbf{u}}) \tag{31}$$

where VEC is an operator that places the $q \times q$ elements of a matrix into a $q^2 \times 1$ vector. As a special case, a random intercept would be defined by including $z_0 = 1$ in $\mathbf{z}$, which will have a mean of one, variance of zero, and covariance of zero with all other z's.

The final term in (21), $V(r_{ij})$, is defined by the choice of conditional distribution for the underlying response. Using the notation $V(r_{ij}) = \sigma_r^2$, the logistic distribution sets $\sigma_r^2 = \pi^2/3$, the standard normal distribution sets $\sigma_r^2 = 1$, and the extreme value distribution sets $\sigma_r^2 = \pi^2/6$. Substituting this result along with (22) and (31) into (21) provides the expression

$$V\left(y^*_{ij}\right) = \boldsymbol{\beta}'\boldsymbol{\Sigma}_{\mathbf{x}}\boldsymbol{\beta} + \boldsymbol{\mu}'_{\mathbf{z}}\boldsymbol{\Sigma}_{\mathbf{u}}\boldsymbol{\mu}_{\mathbf{z}} + \mathrm{VEC}(\boldsymbol{\Sigma}_{\mathbf{z}})'\mathrm{VEC}(\boldsymbol{\Sigma}_{\mathbf{u}}) + \sigma_r^2 \tag{32}$$

as shown in (13).

Although derived here independently, it is important to note that the expression in (32) is consistent with formulas presented elsewhere for calculating the variance of a continuous outcome in linear mixed-effects models. First, setting $y^* = y$ to indicate that the outcome is continuous (i.e., using an identity link), subtraction of the first term from (32) yields the residual variance of y conditional on $\mathbf{x}$, equivalent to (21) of Snijders and Bosker (1994). Second, to calculate the marginal variance of y, (7.10) of Snijders and Bosker (1999, p. 108) reinserts the first term in (32), however, it also introduces an incorrect partition of the second term. Also of note, in Snijders and Bosker (1994, 1999, p. 108), the third term of (32) is written as $\mathrm{trace}(\boldsymbol{\Sigma}_{\mathbf{u}}\boldsymbol{\Sigma}_{\mathbf{z}})$ rather than $\mathrm{VEC}(\boldsymbol{\Sigma}_{\mathbf{z}})'\mathrm{VEC}(\boldsymbol{\Sigma}_{\mathbf{u}})$. The two expressions are, however, equivalent. Finally, Snijders and Bosker (1999, pp. 225–226) present a formula for the implied variance of y^* in multilevel logit and probit models with random intercepts for the purpose of calculating the underlying variance explained by the model. They do not extend this formula to permit the inclusion of additional random effects.

References

Fielding, A. (2004). Scaling for residual variance components of ordered category responses in generalised linear mixed multilevel models. *Quality & Quantity*, *38*, 425–433.

Goodman, L.A. (1960). On the exact variance of products. *Journal of the American Statistical Association*, *55*, 708–713.

Liang, K.-Y., & Zeger, S.L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, *73*, 13–22.

McCullagh, P., & Nelder, J.A. (1989). *Generalized linear models* (2nd ed.). Boca Raton: Chapman & Hall.

McCulloch, C.E., & Searle, S.R. (2001). *Generalized, linear, and mixed models*. New York: Wiley.

Neuhaus, J.M., & Jewell, N.P. (1993). A geometric approach to assess bias due to omitted covariates in generalized linear models. *Biometrika*, *80*, 807–815.

Raudenbush, S.W., & Bryk, A.S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Newbury Park: Sage.

Shrout, P.E., & Bolger, N. (2002). Mediation in experimental and nonexperimental studies: new procedures and recommendations. *Psychological Methods*, *7*, 422–445.

Snijders, T.A.B., & Bosker, R.J. (1994). Modeled variance in two-level models. *Sociological Methods & Research*, *22*, 342–363.

Snijders, T.A.B., & Bosker, R.J. (1999). *Multilevel analysis*. Thousand Oaks: Sage.

Winship, C., & Mare, R.D. (1983). Structural equations and path analysis for discrete data. *American Journal of Sociology*, *89*, 54–110.

Winship, C., & Mare, R.D. (1984). Regression models with ordinal variables. *American Sociological Review*, *49*, 512–525.

Manuscript Received: 24 OCT 2007
Final Version Received: 3 JUN 2008
Published Online Date: 9 SEP 2008