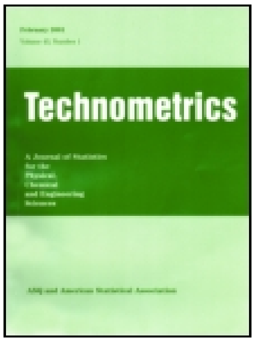

Taylor & Francis
Taylor & Francis Group

Technometrics

ISSN: 0040-1706 (Print) 1537-2723 (Online) Journal homepage: https://www.tandfonline.com/loi/utch20

Bayesian Measures of Explained Variance and Pooling in Multilevel (Hierarchical) Models

Andrew Gelman & Iain Pardoe

To cite this article: Andrew Gelman & Iain Pardoe (2006) Bayesian Measures of Explained Variance and Pooling in Multilevel (Hierarchical) Models, Technometrics, 48:2, 241-251, DOI: 10.1198/004017005000000517

To link to this article: https://doi.org/10.1198/004017005000000517

Bayesian Measures of Explained Variance and Pooling in Multilevel (Hierarchical) Models

Andrew GELMAN

Department of Statistics and Department of Political Science
Columbia University
New York, NY
(*gelman@stat.columbia.edu*)

Iain PARDOE

Charles H. Lundquist College of Business
University of Oregon
Eugene, OR
(*ipardoe@lcbmail.uoregon.edu*)

Explained variance (R^2) is a familiar summary of the fit of a linear regression and has been generalized in various ways to multilevel (hierarchical) models. The multilevel models that we consider in this article are characterized by hierarchical data structures in which individuals are grouped into units (which themselves might be further grouped into larger units), and variables are measured on individuals and each grouping unit. The models are based on regression relationships at different levels, with the first level corresponding to the individual data and subsequent levels corresponding to between-group regressions of individual predictor effects on grouping unit variables. We present an approach to defining R^2 at each level of the multilevel model, rather than attempting to create a single summary measure of fit. Our method is based on comparing variances in a single fitted model rather than with a null model. In simple regression, our measure generalizes the classical adjusted R^2. We also discuss a related variance comparison to summarize the degree to which estimates at each level of the model are pooled together based on the level-specific regression relationship, rather than estimated separately. This pooling factor is related to the concept of shrinkage in simple hierarchical models. We illustrate the methods on a dataset of radon in houses within counties using a series of models ranging from a simple linear regression model to a multilevel varying-intercept, varying-slope model.

KEY WORDS: Adjusted R^2; Bayesian inference; Hierarchical model; Multilevel regression; Partial pooling; Shrinkage.

1. INTRODUCTION

1.1 Explained Variation in Linear Models

Consider a linear regression written as $y_i = (\mathbf{X}\boldsymbol{\beta})_i + \epsilon_i$, $i = 1, \ldots, n$. The fit of the regression can be summarized by the proportion of variance explained,

$$R^2 = 1 - \frac{\mathrm{V}_{i=1}^{n} \epsilon_i}{\mathrm{V}_{i=1}^{n} y_i}, \tag{1}$$

where V represents the finite-sample variance operator, $\mathrm{V}_{i=1}^{n} x_i = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$. (Sample references for classical R^2, also known as the coefficient of determination or squared multiple correlation, include Draper and Smith 1998 and Weisberg 2005.) In a multilevel model (i.e., a hierarchical model with group-level error terms or with regression coefficients $\boldsymbol{\beta}$ that vary by group), the predictors "explain" the data at different levels, and R^2 can be generalized in a variety of ways (for textbook summaries, see Kreft and De Leeuw 1998; Snijders and Bosker 1999; Raudenbush and Bryk 2002; Hox 2002). Xu (2003) reviewed some of these approaches, their connections to information theory, and similar measures for generalized linear models and proportional hazards models. Hodges (1998) discussed connections between hierarchical linear models and classical regression.

The definitions of "explained variance" that we have seen are based on comparisons with a null model, so that

$$R^2 = 1 - \frac{\text{residual variance under the larger model}}{\text{residual variance under the null model}},$$

with various choices of the null model corresponding to predictions at different levels. In this article we propose a slightly different approach, computing (1) at each level of the model and thus coming up with several R^2 values for any particular multilevel model. This approach has the virtue of summarizing the fit at each level and requiring fitting of no additional null models. In defining this summary, our goal is not to dismiss other definitions of R^2, but rather to add another tool to the understanding of multilevel models.

We develop our R^2 measure within a Bayesian framework by evaluating expectations of the variance terms in (1) by averaging over the posterior distribution of the vector of all unknowns in the model. We provide motivation and details in Section 2, and connect to classical definitions in Section 3.

1.2 Pooling in Hierarchical Models

Multilevel models are often understood in terms of "partial pooling," compromising between unpooled and completely pooled estimates. For example, the basic hierarchical model involves data $y_j \sim \mathrm{N}(\alpha_j, \sigma_{\mathbf{y}}^2)$, with population distribution $\alpha_j \sim \mathrm{N}(\mu_{\boldsymbol{\alpha}}, \sigma_{\boldsymbol{\alpha}}^2)$ where $\mu_{\boldsymbol{\alpha}}, \sigma_{\mathbf{y}}$, and $\sigma_{\boldsymbol{\alpha}}$ are known. This can be considered a multilevel model with just the constant predictor at the group level. For each group j, the multilevel estimate of the parameter α_j is

$$\hat{\alpha}_j^{\text{multilevel}} = \omega\mu_{\boldsymbol{\alpha}} + (1-\omega)y_j, \tag{2}$$

where

$$\omega = 1 - \frac{\sigma_{\boldsymbol{\alpha}}^2}{\sigma_{\boldsymbol{\alpha}}^2 + \sigma_{\mathbf{y}}^2} \tag{3}$$

DOI 10.1198/004017005000000517

is a "pooling factor" that represents the degree to which the estimates are pooled together (i.e., based on $\mu_{\boldsymbol{\alpha}}$) rather than estimated separately (based on the raw data y_j). The extreme possibilities, $\omega = 0$ and 1, correspond to no pooling ($\hat{\alpha}_j = y_j$) and complete pooling ($\hat{\alpha}_j = \mu_{\boldsymbol{\alpha}}$). The (posterior) variance of the parameter α_j is

$$\mathrm{var}(\alpha_j) = (1 - \omega)\sigma_{\mathbf{y}}^2. \tag{4}$$

The statistical literature sometimes labels $1 - \omega$ as the "shrinkage factor," a notation we find confusing, because a shrinkage factor of 0 corresponds to complete shrinkage toward the population mean. To avoid ambiguity, we use the term "pooling factor" in this article. The form of expression (3) matches the form of the definition (1) of R^2, a parallelism that we maintain throughout [in the spirit of Gustafson and Clarke (2004), who decomposed components of posterior variance at each level of a Bayesian model].

The concept of pooling is used to help understand multilevel models in two distinct ways: comparing the estimates of different parameters in a group, and summarizing the pooling of the model as a whole. When comparing, it is usual to consider several parameters α_j with a common population (prior) distribution but different data variances; thus $y_j \sim \mathrm{N}(\alpha_j, \sigma_{\mathbf{y}j}^2)$. Then ω_j can be defined as in (3), with $\sigma_{\mathbf{y}j}$ in place of $\sigma_{\mathbf{y}}$. Parameters with more precise data are pooled less toward the population mean, and this can be displayed graphically by a parallel coordinate plot showing the raw estimates y_j pooled toward the posterior means $\hat{\alpha}_j^{\text{multilevel}}$, or a scatterplot of $\hat{\alpha}_j^{\text{multilevel}}$ versus y_j. Pooling of the model as a whole makes use of the fact that the multilevel estimates of the individual parameters α_j, if treated as point estimates, understate the between-group variance (Louis 1984). (See Efron and Morris 1975 and Morris 1983 for discussions of pooling and shrinkage in hierarchical or "empirical Bayes" inference.)

In this article we present a summary measure for the average amount of pooling at each level of a multilevel model. We again develop our pooling measure within a Bayesian framework by evaluating expectations of generalizations of the variance terms in (3) by averaging over the posterior distribution of the vector of all unknowns in the model.

We next introduce an example to motivate the need for summaries such as explained variance and pooling in multilevel models, and then discuss the methods and illustrate their application.

1.3 Example: A Varying-Intercept, Varying-Slope Model for Home Radon Levels

In general, each stage of a multilevel model can have regression predictors and variance components. In this article, we propose summary measures of explained variation and pooling that can be defined and computed at each level of the model. We demonstrate with an example adapted from our own research, a varying-intercept, varying-slope model for levels of radon gas in houses clustered within counties (Price, Nero, and Gelman 1996; Lin, Gelman, Price, and Krantz 1999; Price and Gelman 2005). The model has predictors for both houses and counties, and we introduce it here to show the challenges in defining measures of explained variance and pooling in a multilevel context.

Radon is a carcinogen—a naturally occurring radioactive gas whose decay products are also radioactive—known to cause lung cancer in high concentration, and estimated to cause several thousand lung cancer deaths per year in the United States. The distribution of radon levels in U.S. houses varies greatly, with some houses having dangerously high concentrations. To identify areas with high radon exposures, the Environmental Protection Agency coordinated radon measurements in each of the 50 states.

We illustrate here with an analysis of measured radon in 919 houses in the 85 counties of Minnesota. In performing the analysis, we use a house predictor: whether the measurement was taken in a basement (radon comes from underground and can enter more easily when a house is built into the ground). We also have an important county predictor: a county-level measurement of soil uranium content. We fit the following model:

$$\begin{aligned} y_{ij} &\sim \mathrm{N}(\alpha_j + \beta_j \cdot \mathrm{basement}_{ij}, \sigma_{\mathbf{y}}^2) \\ &\qquad \text{for } i = 1, \ldots, n_j, j = 1, \ldots, J, \\ \begin{pmatrix} \alpha_j \\ \beta_j \end{pmatrix} &\sim \mathrm{N}\left(\begin{pmatrix} \gamma_0 + \gamma_1 u_j \\ \delta_0 + \delta_1 u_j \end{pmatrix}, \begin{pmatrix} \sigma_{\boldsymbol{\alpha}}^2 & \rho\sigma_{\boldsymbol{\alpha}}\sigma_{\boldsymbol{\beta}} \\ \rho\sigma_{\boldsymbol{\alpha}}\sigma_{\boldsymbol{\beta}} & \sigma_{\boldsymbol{\beta}}^2 \end{pmatrix} \right), \\ &\qquad \text{for } j = 1, \ldots, J, \end{aligned} \tag{5}$$

where y_{ij} is the logarithm of the radon measurement in house i in county j, basement$_{ij}$ is the indicator for whether the measurement was in a basement, and u_j is the logarithm of the uranium measurement in county j. The errors in the first line of (5) represent "within-county" variation, which in this case includes measurement error, natural variation in radon levels within a house over time, and variation among houses, beyond what is explained by the basement indicator. The errors in the second line represent variations in radon levels and basement effects between counties, beyond what is explained by the county-level uranium predictor.

Because the sole data-level (i.e., house-level) predictor is dichotomous in this example, it might seem more natural to think in terms of variation in log radon levels for the two categories of home, those with basements and those without. Allowing the intercepts ($\boldsymbol{\alpha}$) and slopes ($\boldsymbol{\beta}$) to vary by county is equivalent to allowing county and basement to be nonadditive factors, and the data indicate that this is indeed the case. Nevertheless, we continue to use the terms "intercepts" and "slopes" here, because our proposed methodology applies more generally to situations with continuous data-level predictors.

We use standard noninformative normal priors for γ_0, γ_1, δ_0, and δ_1 and noninformative uniform priors for $\sigma_{\boldsymbol{\alpha}}$, $\sigma_{\boldsymbol{\beta}}$, and ρ, as recommended by Barnard, McCulloch, and Meng (2000). Section 5 discusses the issue of correlation between the county-level error terms in multilevel models such as this.

The multilevel model allows us to fit a regression to the individual measurements while accounting for systematic unexplained variation among the $J = 85$ counties. Figure 1 shows the data and fitted regression lines within 8 of the 85 counties, and Figure 2 shows the estimated county parameters for all 85 counties and the estimated county-level regression lines, $\gamma_0 + \gamma_1 u$ and $\delta_0 + \delta_1 u$. Various regression diagnostics (includ-

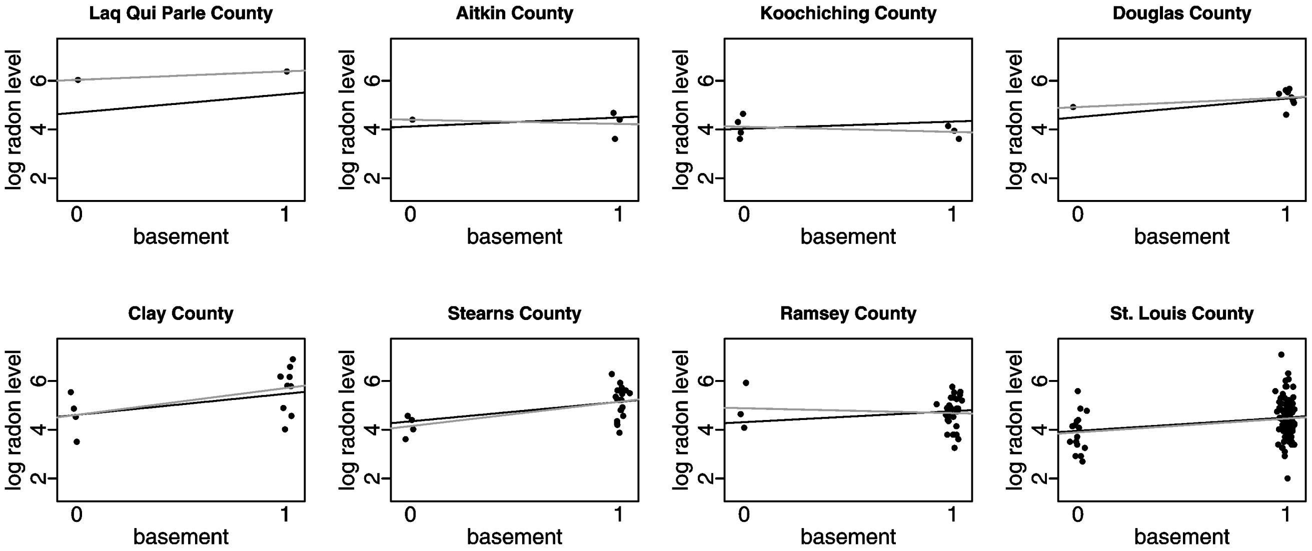

Figure 1. Data (y = log radon, x = jittered basement) and Estimated Regression Lines (black) From the Multilevel Model, $y = \alpha_j + \beta_j \cdot$ basement, for the Radon Example, Displayed for 8 of the 85 Counties j in Minnesota. Both the intercept and the slope vary by county. Because of the pooling of the multilevel model, the fitted lines do not go through the center of the data, a pattern especially noticeable for counties with few observations. For comparison, the gray lines represent the estimated regression lines for models fit to the counties separately.

ing posterior predictive checks; see Gelman, Carlin, Stern, and Rubin 2003, sec. 6.3) suggest that this model fits the data well.

This example illustrates some of the challenges of measuring explained variance and pooling. The model has three levels (data $\mathbf{y}$, intercepts $\boldsymbol{\alpha}$, and slopes $\boldsymbol{\beta}$), with a different variance component at each level ($\sigma_{\mathbf{y}}$, $\sigma_{\boldsymbol{\alpha}}$, and $\sigma_{\boldsymbol{\beta}}$). Here "levels" correspond to the separate variance components rather than to the more usual measurement scales (of which there are two in this case, house and county). Uncertainty in the $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ parameters affects the computation of explained variance for both the data-level model—the simple measure of R^2 from least squares regression will not be appropriate—and the county-level models, because these are second-stage regressions with outcomes that are estimated, not directly observed.

In summarizing the pooling of a batch of parameters in a multilevel model, expression (3) in general cannot be used directly. The difficulty is that it requires knowledge of the unpooled estimates, y_j, in (2). In the varying-intercept, varying-slope radon model, the unpooled estimates are not necessarily available, for example, in a county where all the measured houses have the same basement status.

These difficulties inspire us to define measures of explained variance and pooling that do not depend on fitting alternative models, but rather summarize variances within a single fitted multilevel model.

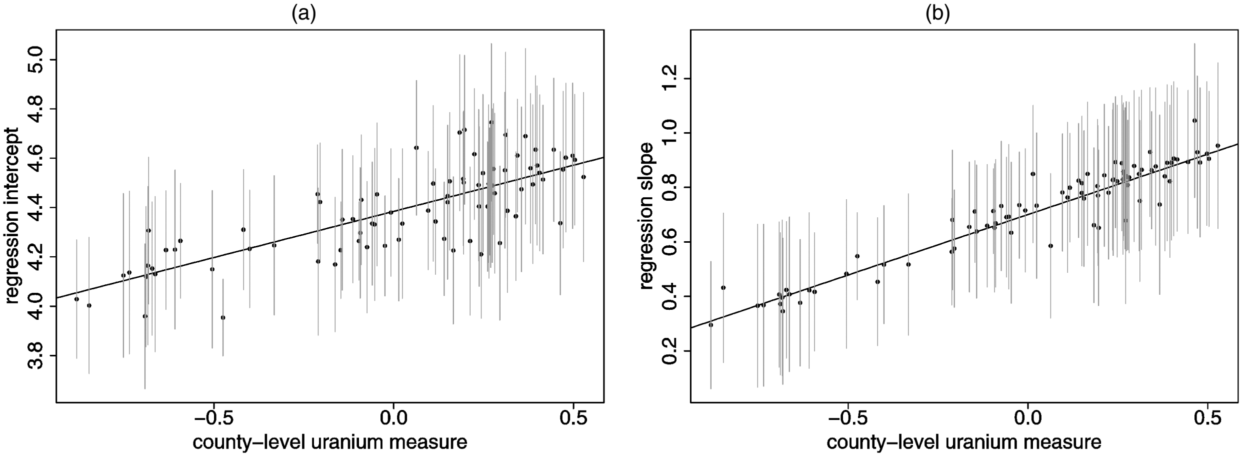

Figure 2. Estimates ± Standard Errors for (a) the County Intercepts α_j, Plotted versus County-Level Uranium Measurement u_j, Along With the Estimated Multilevel Regression Line, $\alpha = \gamma_0 + \gamma_1 u$ and (b) the County Slopes β_j, Plotted versus County-Level Uranium Measurement u_j, Along With the Estimated Multilevel Regression Line, $\beta = \delta_0 + \delta_1 u$. Estimates and standard errors are the posterior medians and standard deviations. For each graph, the county coefficients roughly follow the line, but not exactly; the discrepancies of the coefficients from the line are summarized by the county-level standard deviation parameters σ_α and σ_β.

2. SUMMARIES BASED ON VARIANCE COMPARISONS WITHIN A SINGLE FITTED MODEL

We first define our generalizations of explained variance and pooling factors for each level of a multilevel model, and then (in Sec. 2.4) describe how to compute these summaries using Bayesian posterior simulation draws.

2.1 Notation

We begin by defining a standard notation for a multilevel model with M levels (e.g., $M = 3$ in the radon model of Sec. 1.3). At each level m, we write the model as

$$\theta_k^{(m)} = \mu_k^{(m)} + \epsilon_k^{(m)} \quad \text{for } k = 1, \dots, K^{(m)}, \tag{6}$$

where the $\mu_k^{(m)}$'s are the linear predictors at that level of the model and the errors $\epsilon_k^{(m)}$ come from a distribution with mean 0 and standard deviation $\sigma^{(m)}$. At the lowest (data) level of the model, the $\theta_k^{(m)}$'s correspond to the individual data points (the y_{ij}'s in the radon model). At higher levels of the model, the $\theta_k^{(m)}$'s represent batches of effects or regression coefficients (county intercepts α_j and slopes β_j in the radon model). Because we work with each level of the model separately, we suppress the superscripts (m) for the rest of the article.

The striking similarity of expressions (1) and (3), which define explained variance and pooling, suggests that the two concepts can be understood in a common framework.

2.2 Proportion of Variance Explained at Each Level

For each level (6) of the model, we first consider the variance explained by the linear predictors μ_k. Generalizing from the classical expression (1), we define

$$R^2 = 1 - \frac{\mathrm{E}(\mathbf{V}_{k=1}^K \epsilon_k)}{\mathrm{E}(\mathbf{V}_{k=1}^K \theta_k)}, \tag{7}$$

where "E" represents the posterior mean. In a Bayesian simulation context, these expectations in the numerator and denominator of (7) can be evaluated by averaging over posterior simulation draws, as we discuss in Section 2.4.

R^2 will be close to 0 when the average variance of the errors ϵ_k is approximately equal to the average variance of the θ_k's. R^2 will be close to 1 when the errors ϵ_k are each close to 0 for each posterior sample. Thus R^2 is larger when the μ_k's more closely approximate the θ_k's.

In classical least squares regression, (7) reduces to the usual definition of R^2: the numerator of the ratio becomes the residual variance, and the denominator is simply the variance of the data. Averaging over uncertainty in the regression coefficients leads to a lower value for R^2, as with the classical "adjusted R^2" measure (Wherry 1931). We discuss this connection further in Section 3.1.1. It is possible for our measure (7) to be negative, much like adjusted R^2, if a model predicts so poorly that the estimated error variance is larger than the variance of the data.

2.3 Pooling Factor at Each Level

The next step is to summarize the extent to which the variance of the errors ϵ_k is reduced by pooling across the group-level units. We define the pooling factor as

$$\lambda = 1 - \frac{\mathbf{V}_{k=1}^K \mathrm{E}(\epsilon_k)}{\mathrm{E}(\mathbf{V}_{k=1}^K \epsilon_k)}, \tag{8}$$

where "E" represents the posterior mean. We use the notation λ here to distinguish it from ω in (3) which represents a special case of (8) for the basic hierarchical model; we discuss this connection in Section 3.2.4. The denominator in expression (8) is the numerator in expression (7)—the average variance in the ϵ_k's, that is, the unexplained component of the variance of the θ_k's. The numerator in the ratio term of (8) is the variance among the point estimates (the shrinkage estimators) of the ϵ_k's. If this variance is high (close to the average variance in the ϵ_k's), then λ is close to 0 and there is little pooling. If this variance is low, then the estimated ϵ_k's are pooled closely together, and the pooling factor λ is close to 1.

2.4 Computation Using Posterior Simulations

Multilevel models are increasingly evaluated in a Bayesian framework and computed using posterior simulation, in which inferences for the vector of parameters are summarized by a matrix of simulations (see, e.g., Gilks, Richardson, and Spiegelhalter 1996; Carlin and Louis 2001; Gelman et al. 2003).

We can then evaluate R^2 and λ at each level m of the model using the posterior simulations (not simply the parameter estimates or posterior means), as follows:

1. Evaluate R^2 from (7):
   a. From each simulation draw of the model parameters:
      (1) Compute the vector of θ_k's, predicted values μ_k and the vector of errors, $\epsilon_k = \theta_k - \mu_k$.
      (2) Compute the sample variances, $\mathbf{V}_{k=1}^K \theta_k$ and $\mathbf{V}_{k=1}^K \epsilon_k$.
   b. Average over the simulation draws to estimate $\mathrm{E}(\mathbf{V}_{k=1}^K \theta_k)$ and $\mathrm{E}(\mathbf{V}_{k=1}^K \epsilon_k)$, and then use these to calculate R^2.

2. Evaluate λ from (8) using these same simulation draws in a different way:
   a. For each k, estimate the posterior mean $\mathrm{E}(\epsilon_k)$ of each of the errors ϵ_k as defined in step 1.a(1).
   b. Compute $\mathbf{V}_{k=1}^K \mathrm{E}(\epsilon_k)$—that is, the variance of the K values of $\mathrm{E}(\epsilon_k)$—and then use this, along with $\mathrm{E}(\mathbf{V}_{k=1}^K \epsilon_k)$ from step 1.b, to calculate λ.

We compute R^2 and λ for each level. Figure 3 provides an illustration based on the radon data in Section 1.3. Appendix B gives the computations as implemented in Bugs (Spiegelhalter, Thomas, Best, Gilks, and Lunn 1994, 2003) and R (R Development Core Team 2003).

	R^2 at each level			λ at each level		
Predictors included in the model	y	α	β	y	α	β
Basement (simple linear regression)	.07			0		
County (simple one-way hierarchical model)	.13	0		.04	.54	
Basement + county + uranium	.21	.75		.02	.78	
Basement + county + uranium + basement × county	.22	.22	.34	.04	.81	.90

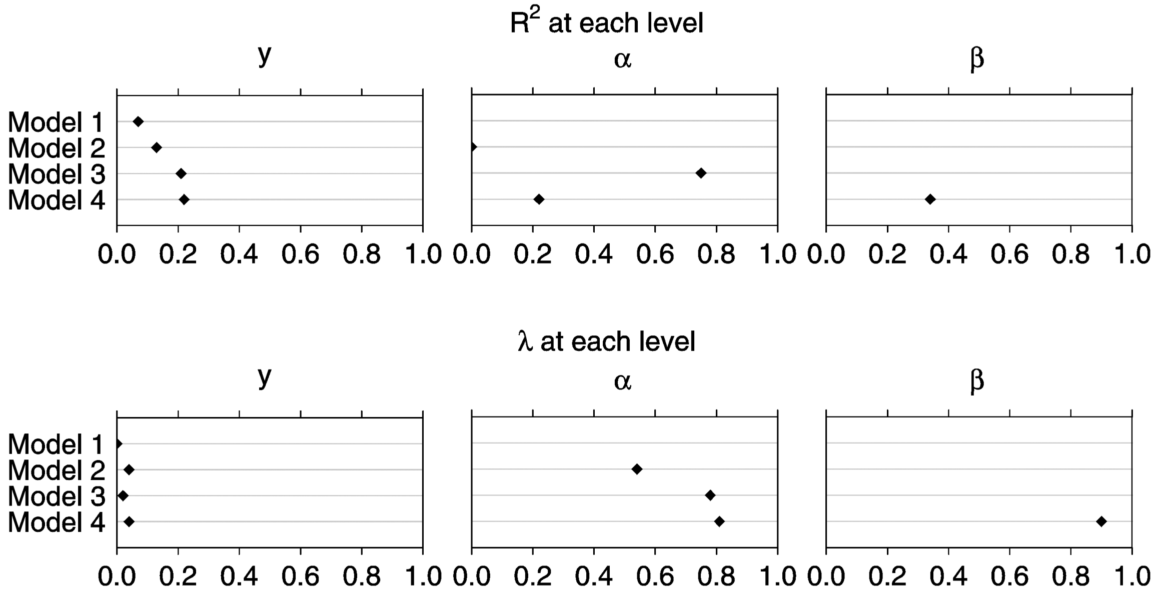

Figure 3. Proportion of Variance Explained and Pooling Factor at the Level of Data y, County-Level Intercepts α, and County-Level Slopes β, for Each of Four Models Fit to the Minnesota Radon Data. Blank entries indicate variance components that are not present in the given model. Results are given in tabular and graphical forms.

2.5 Properties of the Measures of Explained Variance and Pooling

Because R^2 and λ are based on finite-population variances, they are well defined for each level of a multilevel model and work automatically even in the presence of predictors at that level. An alternative approach based on hyperparameters could run into difficulties in such situations, because the hyperparameters may not correspond exactly to the variance comparisons in which we are interested.

For each level of the model, we can interpret our proposed R^2 measure much as we would interpret classical R^2: the proportion of variation in the response (θ) that can be explained by the linear predictor (μ). It therefore suffers the same interpretative drawbacks as classical R^2, such as dependence on the variation in the sample being considered, making comparison between datasets difficult (see Barrett 1974). Also, it does not directly measure predictive accuracy in the manner of model comparison measures such as the Akaike information criterion (AIC) (Akaike 1973) and the deviance information criterion (DIC) (Spiegelhalter, Best, Carlin, and Linde 2002). For example, DIC is a measure of the model as a whole, whereas with our proposed R^2, we are trying to understand how the model fits at each level.

However, for all of its limitations for comparing models and datasets, it seems clear that explained variance will continue to be one of the most commonly used tools in regression. Thus we offer this measure in the spirit of slaking the bottomless thirst that applied researchers have for understanding their regression models. One reason that we consider R^2 as a summary of explained variance *within* a model, rather than in comparison with a null model (see the end of Sec. 1.1) is to avoid the questionable uses of R^2 for model comparison and to focus its use where we believe it is more valid, as a way of understanding the relative importance of predictors and error at each level of a model.

In interpreting the pooling factor λ, .5 would appear to be a clear reference point. A pooling factor $<.5$ suggests a higher degree of within-group information than population-level information. (At the extreme of 0 there is no pooling of estimates toward a common population mean.) Conversely, a pooling factor $>.5$ suggests a higher degree of population-level information than within-group information. (At the extreme of 1, there is complete pooling of estimates toward a common population mean.)

Our proposed pooling factor is also related to the notion of "effective number of parameters" (see Hodges and Sargent 2001; Vaida and Blanchard 2004; Hastie and Tibshirani 1990). For example, in basic hierarchical Bayesian analysis (ignoring posterior uncertainty in the hyperparameters), the effective number of parameters is simply $(1 - \lambda)J$ (thus 0 for complete pooling, J for no pooling).

As a model improves (by adding better predictors and thus improving the μ_k's), we would generally expect both R^2 and

λ to increase for all levels of the model. Increasing R^2 corresponds to more of the variation being explained at that level of the regression model, and a high value of λ implies that the model is pooling the ϵ_k's strongly toward the population mean for that level.

Adding a predictor at one level does not necessarily increase R^2 and λ at other levels of the model, however. In fact, it is possible for an individual-level predictor to improve prediction at the data level but decrease R^2 at the group level (see Kreft and De Leeuw 1998; Gelman and Price 1998; Hox 2002 for discussion and examples of this phenomenon). For the purpose of this article, we merely note that a model can have different explanatory power at different levels.

3. CONNECTIONS WITH CLASSICAL DEFINITIONS

Our general expression for explained variance reduces to classical R^2 for simple linear regression with the least squares estimate for the vector of coefficients. Similarly, for the basic hierarchical model of Section 1.2, our group-level pooling factor is related to the standard definition, conditional on a particular point estimate of the variance components. We present these correspondences here, together with the less frequently encountered pooling factor for the regression model and explained variance for the basic hierarchical model. We illustrate with an applied example in Section 4, and provide further details of the calculations in Appendix A.

3.1 Classical Regression

The classical normal linear regression model can be written as $y_i = (\mathbf{X}\boldsymbol{\beta})_i + \epsilon_i$, $i = 1, \dots, n$, with linear predictors $(\mathbf{X}\boldsymbol{\beta})_i$ and errors ϵ_i that are normal with mean 0 and constant variance σ^2.

3.1.1 Explained Variance and Adjusted R^2. If we plug in the least squares estimate, $\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$, then the proportion of variance explained (7) simply reduces to the classical definition,

$$R^2 = 1 - \frac{\mathrm{E}(\mathrm{V}_{i=1}^{n}\epsilon_i)}{\mathrm{E}(\mathrm{V}_{i=1}^{n}y_i)} = 1 - \frac{\mathbf{y}^T(\mathbf{I}-\mathbf{H})\mathbf{y}}{\mathbf{y}^T\mathbf{I}_c\mathbf{y}},$$

where $\mathbf{I}$ is the $n \times n$ identity matrix, $\mathbf{H} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$, and $\mathbf{I}_c$ is the $n \times n$ matrix with $1 - 1/n$ along the diagonal and $-1/n$ off the diagonal.

In a Bayesian context, to fully evaluate our expression (7) for R^2, one would also average over posterior uncertainty in $\boldsymbol{\beta}$ and σ. Under the standard noninformative prior density that is uniform on $\boldsymbol{\beta}$ and $\log\sigma$ (Gelman et al. 2003, chap. 5), the proportion of variance explained (7) becomes

$$R^2 = 1 - \left(\frac{n-3}{n-p-2}\right)\frac{\mathbf{y}^T(\mathbf{I}-\mathbf{H})\mathbf{y}}{\mathbf{y}^T\mathbf{I}_c\mathbf{y}},$$

where p is the number of columns of $\mathbf{X}$.

This is remarkably similar to the classical adjusted R^2. In fact, if we plug in the classical estimate, $\hat{\sigma}^2 = \mathbf{y}^T(\mathbf{I}-\mathbf{H})\mathbf{y}/(n-p)$, rather than averaging over the marginal posterior distribution for σ^2, then (7) becomes

$$R^2 = 1 - \left(\frac{n-1}{n-p}\right)\frac{\mathbf{y}^T(\mathbf{I}-\mathbf{H})\mathbf{y}}{\mathbf{y}^T\mathbf{I}_c\mathbf{y}},$$

which is exactly classical adjusted R^2. Because $\frac{n-3}{n-p-2} > \frac{n-1}{n-p}$ for $p > 1$, our "Bayesian adjusted R^2" leads to a lower measure of explained variance than the classical adjusted R^2. This makes sense; the classical adjusted R^2 could be considered too high, because it does not account for uncertainty in σ.

3.1.2 Pooling Factor λ. The pooling factor defined in (8) also has a simple form. Evaluating the expectations over the posterior distribution yields

$$\lambda = 1 - \frac{n-p-2}{n-3}.$$

If we plug in the classical estimate, $\hat{\sigma}^2 = \mathbf{y}^T(\mathbf{I}-\mathbf{H})\mathbf{y}/(n-p)$, rather than averaging over the marginal posterior distribution for σ^2, then (8) becomes

$$\lambda = 1 - \frac{n-p}{n-1}.$$

We can see that in the usual setting where the number of regression predictors, p, is small compared with the sample size, n, this pooling factor λ for the regression errors will be close to 0. This makes sense because in this case the classical residuals $(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})_i$ are nearly independent and closely approximate the errors $\epsilon_i = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})_i$. Thus very little shrinkage is needed to estimate these unobserved ϵ_i's.

3.2 One-Way Hierarchical Model

The one-way hierarchical model has the form $y_{ij} \sim \mathrm{N}(\alpha_j, \sigma_{\mathbf{y}}^2)$, $i = 1, \dots, n_j$, $j = 1, \dots, J$, with population distribution $\alpha_j \sim \mathrm{N}(\mu_{\boldsymbol{\alpha}}, \sigma_{\boldsymbol{\alpha}}^2)$, and we can determine the appropriate variance comparisons at each of the model's two levels. Again, this can be considered a multilevel model with just the constant predictor at the group level. For simplicity, we assume that the within-group sample sizes n_j are all equal to a common value n, so that the total sample size is $N = nJ$. The basic hierarchical model of Section 1.2 corresponds to the special case of $n = 1$.

We use the usual noninformative prior density that is uniform on $\mu_{\boldsymbol{\alpha}}$, $\log\sigma_{\mathbf{y}}$, and $\sigma_{\boldsymbol{\alpha}}$ (see, e.g., Gelman et al. 2003, chap. 5). It is not possible to derive closed-form expressions for (7) and (8) averaging over the full posterior distribution. Instead, we present plug-in expressions using the method-of-moments estimators,

$$\begin{aligned} \hat{\sigma}_{\boldsymbol{\alpha}}^2 + \hat{\sigma}_{\mathbf{y}}^2/n &= \frac{\mathbf{y}^T\bar{\mathbf{I}}_c\mathbf{y}}{N}, \\ \hat{\sigma}_{\mathbf{y}}^2 &= \frac{\mathbf{y}^T(\mathbf{I}_c - \bar{\mathbf{I}}_c)\mathbf{y}}{N}, \end{aligned} \qquad (9)$$

where $\mathbf{y} = (y_{11}, \dots, y_{n1}, \dots, y_{1J}, \dots, y_{nJ})^T$ is the N-vector of responses, $\bar{\mathbf{I}}_c$ is the $N \times N$ block-diagonal matrix with $n \times n$ matrices containing elements $1/n - 1/N$ along the diagonal and $n \times n$ matrices containing elements $-1/N$ off the diagonal, and $\mathbf{I}_c$ is the $N \times N$ matrix with $1 - 1/N$ along the diagonal and $-1/N$ off the diagonal. Thus the first estimator in (9) is the sample variance of the J group means [rescaled by $(J-1)/J$], whereas the second estimator is the pooled within-group variance [rescaled by $(n-1)/n$]; we provide further details in Appendix A.

3.2.1 Explained Variance R^2 for the Data-Level Model. Conditional on $\sigma_{\mathbf{y}}$ and $\sigma_{\boldsymbol{\alpha}}$, the proportion of variance explained, (7), at the data level is

$$R^2 = 1 - \frac{\mathbf{y}^T(\mathbf{I}_c - \bar{\mathbf{I}}_c)\mathbf{y} + \omega^2\mathbf{y}^T\bar{\mathbf{I}}_c\mathbf{y} + J(1-\omega)\sigma_{\mathbf{y}}^2}{\mathbf{y}^T\mathbf{I}_c\mathbf{y}}.$$

Plugging in the estimators (9) leads to

$$\begin{aligned} R^2 &= 1 - \left(\frac{n+1}{n}\right)\frac{\mathbf{y}^T(\mathbf{I}_c - \bar{\mathbf{I}}_c)\mathbf{y}}{\mathbf{y}^T\mathbf{I}_c\mathbf{y}} \\ &= 1 - \frac{\hat{\sigma}_{\mathbf{y}}^2/n}{\hat{\sigma}_{\boldsymbol{\alpha}}^2/(n+1) + \hat{\sigma}_{\mathbf{y}}^2/n}. \end{aligned}$$

Subject to finite-sample adjustments, this is approximately equal to the usual value for R^2 in this model, $1 - \sigma_{\mathbf{y}}^2/(\sigma_{\boldsymbol{\alpha}}^2 + \sigma_{\mathbf{y}}^2)$.

3.2.2 Pooling Factor λ for the Data-Level Model. Conditional on $\sigma_{\mathbf{y}}$ and $\sigma_{\boldsymbol{\alpha}}$, the pooling factor, (8), at the data level is

$$\lambda = 1 - \frac{\mathbf{y}^T(\mathbf{I}_c - \bar{\mathbf{I}}_c)\mathbf{y} + \omega^2\mathbf{y}^T\bar{\mathbf{I}}_c\mathbf{y}}{\mathbf{y}^T(\mathbf{I}_c - \bar{\mathbf{I}}_c)\mathbf{y} + \omega^2\mathbf{y}^T\bar{\mathbf{I}}_c\mathbf{y} + J(1-\omega)\sigma_{\mathbf{y}}^2}.$$

Plugging in the estimators (9) leads to

$$\begin{aligned} \lambda &= 1 - \frac{n^2\mathbf{y}^T\bar{\mathbf{I}}_c\mathbf{y} + \mathbf{y}^T(\mathbf{I}_c - \bar{\mathbf{I}}_c)\mathbf{y}}{n(n+1)\mathbf{y}^T\bar{\mathbf{I}}_c\mathbf{y}} \\ &= 1 - \frac{n/(n+1)\hat{\sigma}_{\boldsymbol{\alpha}}^2 + \hat{\sigma}_{\mathbf{y}}^2/n}{\hat{\sigma}_{\boldsymbol{\alpha}}^2 + \hat{\sigma}_{\mathbf{y}}^2/n}. \end{aligned}$$

If the within-group sample sizes n are reasonably large, this data-level pooling factor λ is close to 0, which makes sense because the data-level residuals (estimated errors) are good approximations to the data-level errors (similar to the case of classical regression as discussed in Sec. 3.1.2).

3.2.3 Explained Variance R^2 for the Group-Level Model. At the group level, the one-way hierarchical model has no predictors, and so $R^2 = 0$.

3.2.4 Pooling Factor λ for the Group-Level Model. Conditional on $\sigma_{\mathbf{y}}$ and $\sigma_{\boldsymbol{\alpha}}$, the pooling factor, (8), at the group level is

$$\lambda = 1 - \frac{(1-\omega)\mathbf{y}^T\bar{\mathbf{I}}_c\mathbf{y}}{(1-\omega)\mathbf{y}^T\bar{\mathbf{I}}_c\mathbf{y} + J\sigma_{\mathbf{y}}^2}.$$

Plugging in the estimators in (9) leads to

$$\begin{aligned} \lambda &= 1 - \frac{n\mathbf{y}^T\bar{\mathbf{I}}_c\mathbf{y} - \mathbf{y}^T(\mathbf{I}_c - \bar{\mathbf{I}}_c)\mathbf{y}}{n\mathbf{y}^T\bar{\mathbf{I}}_c\mathbf{y}} \\ &= 1 - \frac{\hat{\sigma}_{\boldsymbol{\alpha}}^2}{\hat{\sigma}_{\boldsymbol{\alpha}}^2 + \hat{\sigma}_{\mathbf{y}}^2/n}. \end{aligned}$$

This expression reduces to (3) by setting n equal to 1 for the basic hierarchical model of Section 1.2.

4. APPLIED EXAMPLE

We apply the methods of Section 2.4 to the home radon level example from Section 1.3. We fit four models:

1. A simple linear regression of log radon level on basement indicators, illustrating the theoretical calculations of Section 3.1
2. A simple one-way hierarchical model of houses within counties, extending the theoretical calculations of Section 3.2 to account for unequal sample sizes and uncertainty in the variance parameters
3. A varying-intercept multilevel model, with basement as an individual-level predictor and log uranium as a county-level predictor
4. The full multilevel varying-intercept, varying-slope model (5), in which the basement effect β is allowed to vary by county.

Figure 3 shows the proportion of explained variance and pooling factor for each level of each model, as computed directly from posterior simulation draws as described in Section 2.4. We discuss the results for each model in turn.

Model 1, Simple Linear Regression. R^2 is very low, suggesting a poorly fitting model, and λ is essentially 0, indicating that the errors are estimated almost independently (which generally holds for a data-level regression model in which there are many more data points than predictors). In comparison, the classical R^2 for this regression, plugging in the least squares estimate for β, is $1 - \mathbf{y}^T(\mathbf{I} - \mathbf{H})\mathbf{y}/(\mathbf{y}^T\mathbf{I}_c\mathbf{y}) = .07$ (see Sec. 3.1.1). The theoretical value for λ for this model is $1 - (n-3)/(n-p-2) = .07$ (see Sec. 3.1.2). These results are all essentially the same, because there is very little uncertainty in β and σ when fitting this simple model, and hence little is changed by moving to fully Bayesian inference.

Model 2, One-Way Hierarchical. At the data level, R^2 shows some improvement over the simple linear regression model but is still quite low. The pooling factor λ remains close to 0. If there were equal sample sizes within each county, then the theoretical value for R^2 for this data-level model, based on plugging in the estimators (9), would come to .13 (see Sec. 3.2.1). Using the posterior simulations accounts for unequal sample sizes and uncertainty in the variance parameters. Similarly, the approximate value for λ for this data level model, plugging in the estimators (9), comes to .05 (see Sec. 3.2.2).

At the county level, $R^2 = 0$, because this model has no county-level predictors. The pooling factor $\lambda = .54$ indicates that the county mean estimates are weighted about equally between the county sample means and the overall population mean. If there were equal sample sizes within each county, then the calculated value for λ for this county-level model, plugging in the estimators (9), comes to .37 (see Sec. 3.2.4). In this case, accounting for unequal sample sizes and uncertainty in the variance parameters leads to a different result.

Model 3, Varying Intercepts. At the data level, R^2 shows further improvement over the one-way hierarchical model but still remains quite low at about 20%. The pooling factor λ remains close to 0.

For the intercept model, $R^2 = .75$ indicates that if the basement effects are restricted to be the same in all counties, then uranium level explains about three-quarters of the variation among counties. The pooling factor implies that the county mean estimates are pooled on average about 80% toward the regression line predicting the county means from their uranium levels.

Model 4, Varying Intercepts and Slopes. At the data level, R^2 is still quite low, indicating that much of the variation in the data remains unexplained by the model (as can be seen in Fig. 1), and λ is still close to 0. For the intercept model, R^2 is 22%, indicating that uranium level explains slightly more than one-fifth of the variation among counties, and λ is about 80%, implying that there is little additional information remaining about each county's intercept. The estimates are pooled on average about 80% toward the regression line [as shown in Fig. 2(a)]. R^2 at the intercept level has decreased from the previous model in which basement effects are restricted to be the same in all counties; allowing the basement effects to vary by county means that there is less variation remaining between counties for uranium level to explain.

For the slope model, R^2 is 34%, implying that the uranium level explains about one-third of the systematic variation in the basement effects across counties. The pooling factor λ is 90%, which tells us that the slopes are estimated mostly from the county-level model, with almost no additional information about the individual counties [as can be seen in Fig. 2(b)].

The fact that much of the information in R^2 and λ is captured in Figures 1 and 2 should not be taken as indicating a flaw of these measures. Just as the correlation is a useful numerical summary of information available in a scatterplot, the explained variance and pooling measures quickly summarize the explanatory power and actions of a multilevel model, without being a substitute for more informative graphical displays.

At each level of each model, R^2 estimates the amount of the variance explained by the predictors, and λ shows the extent to which the errors at that level are pooled by their common prior distribution. The pooling factors for the data are all very close to 0, which is no surprise (see Secs. 3.1.2 and 3.2.2) so in practice it might not be necessary to display them. We include them in this example for completeness.

5. DISCUSSION

We suggest computing our measures for the proportion of variance explained at each level of a multilevel model, (7), and the pooling factor at each level, (8). These can be easily calculated using posterior simulations as detailed in Section 2.4 and illustrated in Appendix B. The measures of R^2 and λ conveniently summarize the fit at each level of the model and the degree to which estimates are pooled toward their population models. Together, they clarify the role of predictors at different levels of a multilevel model. They can be derived from a common framework of comparing variances at each level of the model, which also means that they do not require the fitting of additional null models.

Expressions (7) and (8) are closely related to the usual definitions of adjusted R^2 in simple linear regression and shrinkage in balanced one-way hierarchical models. From this perspective, they unify the data-level concept of R^2 and the group-level concept of pooling or shrinkage, and also generalize these concepts to account for uncertainty in the variance components. Further, as illustrated for the home radon application in Section 4, they provide a useful tool for understanding the behavior of more complex multilevel models.

We define R^2 and λ at each level of a multilevel model, where the error terms at each level can be modeled as independent or correlated. However, even if the error terms are modeled as correlated, the corresponding R^2 and λ measures do not explicitly incorporate information about the correlation. The situation could be considered analogous to multivariate regression (with more than one response variable), where classical R^2 is routinely calculated for each of the response variables (with no adjustment for the modeled correlations among them). Another similar situation occurs in structural equation modeling, where Teel, Bearden, and Sharma (1986) and Bentler and Raykov (2000) proposed measures of explained variance for each related equation, again with no adjustment to account for correlations among the equations. That said, it may be possible to extend our definitions to multivariate versions of explained variance and pooling (although it is far from clear how or why one should do this), but this lies beyond the scope of this article.

The usual care must be taken with intercept terms. For example, when changing the zero point, the interpretation of the intercept changes. Thus with varying-intercept models, the intercepts are of interest only when the zero point has a clear interpretation (as it does in the radon example, where it represents homes without basements).

We have presented our R^2 and λ measures in a Bayesian framework, but they could also be evaluated in a non-Bayesian framework using simulations from distributions representing estimates and measures of uncertainty for the predicted values μ_k and the errors ϵ_k. For example, these might be represented by multivariate normal distributions with a point estimate for the mean and estimated covariance matrix for the variance or, alternatively, by bootstrap simulations.

We have derived connections to classical definitions of explained variance and shrinkage for models with normal error distributions, and also illustrated our methods using a multilevel model with normal error distributions at each level. But (7) and (8) do not depend on any normality assumptions, and in principle, these measures are appropriate variance summaries for models with nonnormal error distributions (see also Goldstein, Browne, and Rasbash 2002; Browne, Subramanian, Jones, and Goldstein 2005). An alternative for generalized linear models could be to develop analogous measures using deviances.

ACKNOWLEDGMENTS

Thank Ronghui Xu, Cajo ter Braak, Bill Browne, Samantha Cook, and the editors and reviewers for helpful comments and the National Science Foundation for financial support.

APPENDIX A: THEORETICAL COMPUTATIONS

A.1 Classical Regression

The classical normal linear regression model can be written as $y_i = (\mathbf{X}\boldsymbol{\beta})_i + \epsilon_i, i = 1, \dots, n$, with linear predictors $(\mathbf{X}\boldsymbol{\beta})_i$

and errors ϵ_i that are normal with mean 0 and constant variance σ^2. If we plug in the least-squares estimate, $\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$, then (7) reduces simply to the classical definition, $R^2 = 1 - \mathbf{y}^T(\mathbf{I}-\mathbf{H})\mathbf{y}/\mathbf{y}^T\mathbf{I}_c\mathbf{y}$.

In a Bayesian context, however, we need to average over posterior uncertainty in $\boldsymbol{\beta}$ and σ. Under the usual noninformative prior density that is uniform on $(\boldsymbol{\beta}, \log\sigma)$, the posterior distribution for $\boldsymbol{\beta}$ (conditional on σ) is $\mathrm{N}((\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}, \sigma^2(\mathbf{X}^T\mathbf{X})^{-1})$. The marginal posterior distribution of σ^2 is then a scaled inverse chi-squared distribution with degrees of freedom $n-p$ and scale factor $\mathbf{y}^T(\mathbf{I}-\mathbf{H})\mathbf{y}/(n-p)$, where $\mathbf{I}$ is the $n\times n$ identity matrix, $\mathbf{H} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$, and p is the number of columns of $\mathbf{X}$. We proceed by first averaging over the posterior distribution for $\boldsymbol{\beta}$ (conditional on σ), so that

$$\begin{aligned}
(n-1)\mathrm{E}\left(\mathop{\mathrm{V}}_{i=1}^{n} \epsilon_i\right)
&= \mathbf{y}_c^T\mathbf{y}_c - 2\mathbf{y}_c^T\mathbf{X}_c\mathrm{E}(\boldsymbol{\beta}) + \mathrm{E}(\boldsymbol{\beta}^T(\mathbf{X}_c^T\mathbf{X}_c)\boldsymbol{\beta})\\
&= \mathbf{y}_c^T\mathbf{y}_c - 2\mathbf{y}_c^T\mathbf{X}_c(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} + \sigma^2\,\mathrm{tr}(\mathbf{X}_c^T\mathbf{X}_c(\mathbf{X}^T\mathbf{X})^{-1})\\
&\quad + \mathbf{y}^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}_c^T\mathbf{X}_c(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}\\
&= \mathbf{y}^T\mathbf{I}_c\mathbf{y} - 2\mathbf{y}^T\mathbf{H}_c\mathbf{y} + \sigma^2\,\mathrm{tr}(\mathbf{H}_c) + \mathbf{y}^T\mathbf{H}_c\mathbf{y}\\
&= \mathbf{y}^T\mathbf{I}_c\mathbf{y} - \mathbf{y}^T\mathbf{H}_c\mathbf{y} + (p-1)\sigma^2\\
&= \mathbf{y}^T(\mathbf{I}-\mathbf{H})\mathbf{y} + (p-1)\sigma^2
\end{aligned}$$

and

$$\begin{aligned}
(n-1)\mathrm{E}\left(\mathop{\mathrm{V}}_{i=1}^{n} y_i\right) &= (n-1)\mathop{\mathrm{V}}_{i=1}^{n} y_i\\
&= \mathbf{y}^T\mathbf{I}_c\mathbf{y},
\end{aligned}$$

where $\mathbf{I}_c$ is the $n\times n$ matrix with $1-1/n$ along the diagonal and $-1/n$ off the diagonal, $\mathbf{H}_c = \mathbf{I}_c\mathbf{H}$, $\mathbf{X}_c = \mathbf{I}_c\mathbf{X}$, and $\mathbf{y}_c = \mathbf{I}_c\mathbf{y}$. Conditional on σ, the proportion of variance explained, (7), is then

$$R^2 = 1 - \frac{\mathbf{y}^T(\mathbf{I}-\mathbf{H})\mathbf{y} + (p-1)\sigma^2}{\mathbf{y}^T\mathbf{I}_c\mathbf{y}}.$$

Because the marginal posterior expected value for σ^2 is $\mathbf{y}^T(\mathbf{I}-\mathbf{H})\mathbf{y}/(n-p-2)$, the proportion of variance explained, (7), fully averaging over posterior uncertainty in $\boldsymbol{\beta}$ and σ, is

$$R^2 = 1 - \left(\frac{n-3}{n-p-2}\right)\frac{\mathbf{y}^T(\mathbf{I}-\mathbf{H})\mathbf{y}}{\mathbf{y}^T\mathbf{I}_c\mathbf{y}}.$$

Similarly, conditional on σ,

$$(n-1)\mathop{\mathrm{V}}_{i=1}^{n}\mathrm{E}(\epsilon_i) = \mathbf{y}^T(\mathbf{I}-\mathbf{H})\mathbf{y},$$

and the pooling factor, (8), is then

$$\lambda = 1 - \frac{\mathbf{y}^T(\mathbf{I}-\mathbf{H})\mathbf{y}}{\mathbf{y}^T(\mathbf{I}-\mathbf{H})\mathbf{y} + (p-1)\sigma^2}.$$

Averaging over the marginal posterior distribution of σ, this becomes

$$\lambda = 1 - \frac{n-p-2}{n-3}.$$

A.2 One-Way Hierarchical Model

The one-way hierarchical model has the form $y_{ij} \sim \mathrm{N}(\alpha_j, \sigma_{\mathbf{y}}^2)$, $i = 1,\dots,n_j$, $j = 1,\dots,J$, with population distribution $\alpha_j \sim \mathrm{N}(\mu_{\boldsymbol{\alpha}}, \sigma_{\boldsymbol{\alpha}}^2)$. For simplicity, we assume that the within-group sample sizes n_j are all equal to a common value n. Under the usual noninformative prior density that is uniform on $(\mu_{\boldsymbol{\alpha}}, \log\sigma_{\mathbf{y}}, \sigma_{\boldsymbol{\alpha}})$, the posterior distribution for α_j (conditional on $\sigma_{\mathbf{y}}$ and $\sigma_{\boldsymbol{\alpha}}$) is $\mathrm{N}(\omega\mu_{\boldsymbol{\alpha}} + (1-\omega)\bar{y}_{.j}, (1-\omega)\sigma_{\mathbf{y}}^2/n)$, where $\omega = 1 - \sigma_{\boldsymbol{\alpha}}^2/(\sigma_{\boldsymbol{\alpha}}^2 + \sigma_{\mathbf{y}}^2/n)$ and $\bar{y}_{.j}$ is the sample mean within group j.

In what follows, it helps to set up matrix notation for this setting. Let $\mathbf{y} = (y_{11},\dots,y_{n1},\dots,y_{1J},\dots,y_{nJ})^T$ be the N-vector of responses, where $N = nJ$. Then if $\mathbf{I}_c$ is the $N\times N$ matrix with $1-1/N$ along the diagonal and $-1/N$ off the diagonal, the mean-centered vector of responses can be written as $\mathbf{y}_c = \mathbf{I}_c\mathbf{y}$. Similarly, let $\boldsymbol{\alpha} = (\alpha_1,\dots,\alpha_1,\dots,\alpha_J,\dots,\alpha_J)^T$ be an N-vector of J stacked sets of population group means, each set containing n replicates, and let $\bar{\mathbf{y}} = (\bar{y}_{.1},\dots,\bar{y}_{.1},\dots,\bar{y}_{.J},\dots,\bar{y}_{.J})^T$ be a similar N-vector of stacked sets of sample group means. Then if $\bar{\mathbf{I}}_c$ is the $N\times N$ block-diagonal matrix with $n\times n$ matrices containing elements $1/n - 1/N$ along the diagonal and $n\times n$ matrices containing elements $-1/N$ off the diagonal, the mean-centered vector of population means can be written as $\boldsymbol{\alpha}_c = \bar{\mathbf{I}}_c\boldsymbol{\alpha}$, and the mean-centered vector of sample means can be written as $\bar{\mathbf{y}}_c = \bar{\mathbf{I}}_c\mathbf{y}$. Finally, let $\bar{\mathbf{I}}$ be the $N\times N$ block-diagonal matrix with $n\times n$ matrices containing elements (1) along the diagonal and $n\times n$ matrices containing elements (0) off the diagonal, so that the posterior distribution of $\boldsymbol{\alpha}_c$ (conditional on $\sigma_{\mathbf{y}}$ and $\sigma_{\boldsymbol{\alpha}}$) can be written as $\mathrm{N}((1-\omega)\bar{\mathbf{y}}_c, \bar{\mathbf{I}}(1-\omega)\sigma_{\mathbf{y}}^2/n)$.

We proceed by averaging over the posterior distribution for $\boldsymbol{\alpha}$ (conditional on $\sigma_{\mathbf{y}}$ and $\sigma_{\boldsymbol{\alpha}}$) to find conditional expressions for (7) and (8) at each level of the model. In this case, further averaging over the marginal posterior distributions of $\sigma_{\mathbf{y}}$ and $\sigma_{\boldsymbol{\alpha}}$ does not result in closed-form solutions. As an alternative, we plug in particular point estimates for the variance components to find unconditional expressions.

At the data level, conditional on $\sigma_{\mathbf{y}}$ and $\sigma_{\boldsymbol{\alpha}}$,

$$\begin{aligned}
(N-1)\mathrm{E}\left(\mathop{\mathrm{V}}_{i,j}(y_{ij}-\alpha_j)\right)
&= \mathbf{y}_c^T\mathbf{y}_c - 2\mathbf{y}_c^T\mathrm{E}(\boldsymbol{\alpha}_c) + \mathrm{E}(\boldsymbol{\alpha}_c^T\boldsymbol{\alpha}_c)\\
&= \mathbf{y}_c^T\mathbf{y}_c - 2\mathbf{y}_c^T(1-\omega)\bar{\mathbf{y}}_c\\
&\quad + \mathrm{tr}(\bar{\mathbf{I}})(1-\omega)\sigma_{\mathbf{y}}^2/n + (1-\omega)^2\bar{\mathbf{y}}_c^T\bar{\mathbf{y}}_c\\
&= \mathbf{y}^T\mathbf{I}_c\mathbf{y} + (\omega^2-1)\mathbf{y}^T\bar{\mathbf{I}}_c\mathbf{y} + J(1-\omega)\sigma_{\mathbf{y}}^2\\
&= \mathbf{y}^T(\mathbf{I}_c - \bar{\mathbf{I}}_c)\mathbf{y} + \omega^2\mathbf{y}^T\bar{\mathbf{I}}_c\mathbf{y} + J(1-\omega)\sigma_{\mathbf{y}}^2
\end{aligned}$$

and

$$\begin{aligned}
(N-1)\mathrm{E}\left(\mathop{\mathrm{V}}_{i,j} y_{ij}\right) &= (N-1)\mathop{\mathrm{V}}_{i,j} y_{ij}\\
&= \mathbf{y}^T\mathbf{I}_c\mathbf{y}.
\end{aligned}$$

So, conditional on $\sigma_{\mathbf{y}}$ and $\sigma_{\boldsymbol{\alpha}}$, the proportion of variance explained, (7), at the data level is

$$R^2 = 1 - \frac{\mathbf{y}^T(\mathbf{I}_c - \bar{\mathbf{I}}_c)\mathbf{y} + \omega^2\mathbf{y}^T\bar{\mathbf{I}}_c\mathbf{y} + J(1-\omega)\sigma_{\mathbf{y}}^2}{\mathbf{y}^T\mathbf{I}_c\mathbf{y}}.$$

Similarly,

$$(N-1)\mathop{\mathrm{V}}_{i,j}\mathrm{E}(y_{ij}-\alpha_j)=\mathbf{y}^T(\mathbf{I}_c-\bar{\mathbf{I}}_c)\mathbf{y}+\omega^2\mathbf{y}^T\bar{\mathbf{I}}_c\mathbf{y},$$

and, conditional on $\sigma_{\mathbf{y}}$ and $\sigma_{\boldsymbol{\alpha}}$, the pooling factor, (8), at the data level is

$$\lambda=1-\frac{\mathbf{y}^T(\mathbf{I}_c-\bar{\mathbf{I}}_c)\mathbf{y}+\omega^2\mathbf{y}^T\bar{\mathbf{I}}_c\mathbf{y}}{\mathbf{y}^T(\mathbf{I}_c-\bar{\mathbf{I}}_c)\mathbf{y}+\omega^2\mathbf{y}^T\bar{\mathbf{I}}_c\mathbf{y}+J(1-\omega)\sigma_{\mathbf{y}}^2}.$$

At the group level, conditional on $\sigma_{\mathbf{y}}$ and $\sigma_{\boldsymbol{\alpha}}$,

$$\begin{aligned}n(J-1)\mathrm{E}\left(\mathop{\mathrm{V}}_{j=1}^{J}(\alpha_j-\mu_{\boldsymbol{\alpha}})\right)&=n(J-1)\mathrm{E}\left(\mathop{\mathrm{V}}_{j=1}^{J}\alpha_j\right)\\&=(1-\omega)^2\mathbf{y}^T\bar{\mathbf{I}}_c\mathbf{y}+J(1-\omega)\sigma_{\mathbf{y}}^2\end{aligned}$$

and

$$n(J-1)\mathop{\mathrm{V}}_{j=1}^{J}\mathrm{E}(\alpha_j-\mu_{\boldsymbol{\alpha}})=(1-\omega)^2\mathbf{y}^T\bar{\mathbf{I}}_c\mathbf{y}.$$

So the proportion of variance explained, (7), is 0, whereas the pooling factor, (8), is

$$\lambda=1-\frac{(1-\omega)\mathbf{y}^T\bar{\mathbf{I}}_c\mathbf{y}}{(1-\omega)\mathbf{y}^T\bar{\mathbf{I}}_c\mathbf{y}+J\sigma_{\mathbf{y}}^2}.$$

To find unconditional expressions, we plug in the following point estimates for the variance components:

$$\hat{\sigma}_{\boldsymbol{\alpha}}^2+\hat{\sigma}_{\mathbf{y}}^2/n=\frac{\mathbf{y}^T\bar{\mathbf{I}}_c\mathbf{y}}{N}$$

and

$$\hat{\sigma}_{\mathbf{y}}^2=\frac{\mathbf{y}^T(\mathbf{I}_c-\bar{\mathbf{I}}_c)\mathbf{y}}{N}.$$

These estimators are just rescalings of the sample variance of the J group means and the pooled within-group variance,

$$\frac{\mathbf{y}^T\bar{\mathbf{I}}_c\mathbf{y}}{N}=\left(\frac{J-1}{J}\right)\mathop{\mathrm{V}}_{j=1}^{J}\bar{y}_{.j}$$

and

$$\frac{\mathbf{y}^T(\mathbf{I}_c-\bar{\mathbf{I}}_c)\mathbf{y}}{N}=\left(\frac{n-1}{n}\right)\mathop{\mathrm{M}}_{j=1}^{J}\mathop{\mathrm{V}}_{i=1}^{n}y_{ij}.$$

The plug-in estimate of (7) at the data level is then

$$\begin{aligned}R^2&=1-\left(\frac{n+1}{n}\right)\frac{\mathbf{y}^T(\mathbf{I}_c-\bar{\mathbf{I}}_c)\mathbf{y}}{\mathbf{y}^T\mathbf{I}_c\mathbf{y}}\\&=1-\frac{\hat{\sigma}_{\mathbf{y}}^2/n}{\hat{\sigma}_{\boldsymbol{\alpha}}^2/(n+1)+\hat{\sigma}_{\mathbf{y}}^2/n},\end{aligned}$$

whereas the plug-in estimate of (8) at the data level is

$$\begin{aligned}\lambda&=1-\frac{n^2\mathbf{y}^T\bar{\mathbf{I}}_c\mathbf{y}+\mathbf{y}^T(\mathbf{I}_c-\bar{\mathbf{I}}_c)\mathbf{y}}{n(n+1)\mathbf{y}^T\bar{\mathbf{I}}_c\mathbf{y}}\\&=1-\frac{n/(n+1)\hat{\sigma}_{\boldsymbol{\alpha}}^2+\hat{\sigma}_{\mathbf{y}}^2/n}{\hat{\sigma}_{\boldsymbol{\alpha}}^2+\hat{\sigma}_{\mathbf{y}}^2/n}.\end{aligned}$$

Finally, the plug-in estimate of (8) at the group level is

$$\begin{aligned}\lambda&=1-\frac{n\mathbf{y}^T\bar{\mathbf{I}}_c\mathbf{y}-\mathbf{y}^T(\mathbf{I}_c-\bar{\mathbf{I}}_c)\mathbf{y}}{n\mathbf{y}^T\bar{\mathbf{I}}_c\mathbf{y}}\\&=1-\frac{\hat{\sigma}_{\boldsymbol{\alpha}}^2}{\hat{\sigma}_{\boldsymbol{\alpha}}^2+\hat{\sigma}_{\mathbf{y}}^2/n},\end{aligned}$$

which is equivalent to the standard definition of the group-level pooling factor for the basic hierarchical model of Section 1.2 in which $n=1$.

APPENDIX B: IMPLEMENTATION IN BUGS AND R

A key feature of the methods described here is their easy implementation in a simulation-based computing environment. We illustrate this by programming the R^2 and λ computations for the radon model in the popular Bayesian package Bugs (Spiegelhalter et al. 1994, 2003), as called from the general statistical computing software R (R Development Core Team 2003; Gelman 2003).

From R, we first set up the model and run it in Bugs:

```
radon.data <- list ("N", "x", "y",
                    "J", "county", "u")
radon.inits <- function () {
  list (g=cbind(rnorm(J), rnorm(J)),
        sigma.y=runif(1), gamma.0=rnorm(1),
        gamma.1=rnorm(1), sigma.a=runif(1),
        delta.0=rnorm(1), delta.1=rnorm(1),
        sigma.b=runif(1), rho=runif(1,-1,1))}
radon.parameters <- c ("y.hat", "a", "b",
                       "e.y", "e.a", "e.b",
                       "gamma.0", "gamma.1",
                       "delta.0", "delta.1",
                       "sigma.y", "sigma.a",
                       "sigma.b", "rho")
radon.r2 <- bugs (radon.data, radon.inits,
                  radon.parameters,
                  "rsquared.bug",
                  n.chains=3, n.iter=20000)
```

We then use the resulting simulation draws to compute R^2 and λ for each of the three levels of the model:

```
attach.bugs (radon.r2)

# data level summaries
rsquared.y <- 1 - mean (apply (e.y, 1, var))
                  / var (y)
lambda.y <- 1 - var (apply (e.y, 2, mean))
                / mean (apply (e.y, 1, var))

# summaries for the intercept model
rsquared.a <- 1 - mean (apply (e.a, 1, var))
                  / mean (apply (a, 1, var))
lambda.a <- 1 - var (apply (e.a, 2, mean))
                / mean (apply (e.a, 1, var))

# summaries for the slope model
```

```
rsquared.b <- 1 - mean (apply (e.b, 1, var))
                  / mean (apply (b, 1, var))
lambda.b <- 1 - var (apply (e.b, 2, mean))
                / mean (apply (e.b, 1, var))

print (round (c (rsquared.y, rsquared.a,
                 rsquared.b), 2))
# 0.22 0.22 0.34
print (round (c (lambda.y, lambda.a,
                 lambda.b), 2))
# 0.04 0.81 0.90
```

Finally, we show the Bugs code for the three-level model (as saved in the file `rsquared.bug`):

```
model {
  for (i in 1:N){
    y[i]   dnorm (y.hat[i], tau.y)
    y.hat[i] <- a[county[i]]
              + b[county[i]]*x[i]
    e.y[i] <- y[i] - y.hat[i]
  }
  tau.y <- pow(sigma.y, -2)
  sigma.y ~ dunif (0, 1000)

  for (j in 1:J){
    a[j] <- g[j,1]
    b[j] <- g[j,2]
    g[j,1:2] ~ dmnorm (g.hat[j,1:2],
                       tau.g[1:2,1:2])
    g.hat[j,1] <- gamma.0 + gamma.1*u[j]
    e.a[j] <- a[j] - g.hat[j,1]
    g.hat[j,2] <- delta.0 + delta.1*u[j]
    e.b[j] <- b[j] - g.hat[j,2]
  }
  gamma.0 ~ dnorm (0, .0001)
  gamma.1 ~ dnorm (0, .0001)
  delta.0 ~ dnorm (0, .0001)
  delta.1 ~ dnorm (0, .0001)

  tau.g[1:2,1:2] <- inverse(sigma.g[,])
  sigma.g[1,1] <- pow(sigma.a, 2)
  sigma.a ~ dunif (0, 100)
  sigma.g[2,2] <- pow(sigma.b, 2)
  sigma.b ~ dunif (0, 100)
  sigma.g[1,2] <- rho*sigma.a*sigma.b
  sigma.g[2,1] <- sigma.g[1,2]
  rho ~ dunif (-1, 1)
}
```

[Received July 2004. Revised April 2005.]

REFERENCES

Akaike, H. (1973), "Information Theory and an Extension of the Maximum Likelihood Principle," in *2nd International Symposium on Information Theory*, eds. B. Petrov and F. Csáki, Budapest: Akadémiai Kiadó, pp. 267–281.

Barnard, J., McCulloch, R., and Meng, X. L. (2000), "Modeling Covariance Matrices in Terms of Standard Deviations and Correlations, With Application to Shrinkage," *Statistica Sinica*, 1, 1281–1311.

Barrett, J. P. (1974), "The Coefficient of Determination—Some Limitations," *The American Statistician*, 28, 19–20.

Bentler, P. M., and Raykov, T. (2000), "On Measures of Explained Variance in Nonrecursive Structural Equation Models," *Journal of Applied Psychology*, 85, 125–131.

Browne, W. J., Subramanian, S. V., Jones, K., and Goldstein, H. (2005), "Variance Partitioning in Multilevel Logistic Models That Exhibit Over-Dispersion," *Journal of the Royal Statistical Society*, Ser. B, to appear.

Carlin, B. P., and Louis, T. A. (2001), *Bayes and Empirical Bayes Methods for Data Analysis* (2nd ed.), London: CRC Press.

Draper, N. R., and Smith, H. (1998), *Applied Regression Analysis* (3rd ed.), New York: Wiley.

Efron, B., and Morris, C. (1975), "Data Analysis Using Stein's Estimator and Its Generalizations," *Journal of the American Statistical Association*, 70, 311–319.

Gelman, A. (2003), "Bugs.R: Functions for Calling Bugs From R," available at *www.stat.columbia.edu/∼gelman/bugsR/*.

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (2003), *Bayesian Data Analysis* (2nd ed.), London: CRC Press.

Gelman, A., and Price, P. N. (1998), Discussion of "Some Algebra and Geometry for Hierarchical Models, Applied to Diagnostics," by J. S. Hodges, *Journal of the Royal Statistical Society*, Ser. B, 60, 497–536.

Gilks, W. R., Richardson, S., and Spiegelhalter, D. (eds.) (1996), *Practical Markov Chain Monte Carlo*, New York: Chapman & Hall.

Goldstein, H., Browne, W. J., and Rasbash, J. (2002), "Partitioning Variation in Multilevel Models," *Understanding Statistics*, 1, 223–232.

Gustafson, P., and Clarke, B. (2004), "Decomposing Posterior Variance," *Journal of Statistical Planning and Inference*, 119, 311–327.

Hastie, T., and Tibshirani, R. (1990), *Generalized Additive Models*, London: CRC Press.

Hodges, J. S. (1998), "Some Algebra and Geometry for Hierarchical Models, Applied to Diagnostics" (with discussion), *Journal of the Royal Statistical Society*, Ser. B, 60, 497–536.

Hodges, J. S., and Sargent, D. J. (2001), "Counting Degrees of Freedom in Hierarchical and Other Richly-Parameterised Models," *Biometrika*, 88, 367–379.

Hox, J. (2002), *Multilevel Analysis: Techniques and Applications*, Mahwah, NJ: Lawrence Erlbaum.

Kreft, I., and De Leeuw, J. (1998), *Introducing Multilevel Modeling*, London: Sage.

Lin, C. Y., Gelman, A., Price, P. N., and Krantz, D. H. (1999), "Analysis of Local Decisions Using Hierarchical Modeling, Applied to Home Radon Measurement and Remediation" (with discussion and rejoinder), *Statistical Science*, 14, 305–337.

Louis, T. A. (1984), "Estimating a Population of Parameter Values Using Bayes and Empirical Bayes Methods," *Journal of the American Statistical Association*, 78, 393–398.

Morris, C. (1983), "Parametric Empirical Bayes Inference: Theory and Applications" (with discussion), *Journal of the American Statistical Association*, 78, 47–65.

Price, P. N., and Gelman, A. (2005), "Should You Measure the Radon Concentration in Your Home?" in *Statistics: A Guide to the Unknown* (4th ed.), eds. R. Peck et al., N. Scituate, MA: Duxbury Press, pp. 149–170.

Price, P. N., Nero, A. V., and Gelman, A. (1996), "Bayesian Prediction of Mean Indoor Radon Concentrations for Minnesota Counties," *Health Physics*, 71, 922–936.

R Development Core Team (2003), *R: A Language and Environment for Statistical Computing*, Vienna: R Foundation for Statistical Computing, available at *www.r-project.org*.

Raudenbush, S. W., and Bryk, A. S. (2002), *Hierarchical Linear Models* (2nd ed.), Thousand Oaks, CA: Sage.

Snijders, T. A. B., and Bosker, R. J. (1999), *Multilevel Analysis*, London: Sage.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., and Linde, A. V. D. (2002), "Bayesian Measures of Model Complexity and Fit" (with discussion), *Journal of the Royal Statistical Society*, Ser. B, 64, 583–639.

Spiegelhalter, D. J., Thomas, A., Best, N. G., Gilks, W. R., and Lunn, D. (1994, 2003), "BUGS: Bayesian Inference Using Gibbs Sampling," MRC Biostatistics Unit, available at *www.mrc-bsu.cam.ac.uk/bugs/*.

Teel, J. E., Bearden, W. O., and Sharma, S. (1986), "Interpreting LISREL Estimates of Explained Variance in Nonrecursive Structural Equation Models," *Journal of Marketing Research*, 23, 164–168.

Vaida, F., and Blanchard, S. (2004), "Conditional Akaike Information for Mixed-Effects Models," technical report, Harvard School of Public Health, Dept. of Biostatistics.

Weisberg, S. (2005), *Applied Linear Regression* (3rd ed.), New York: Wiley.

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

Xu, R. (2003), "Measuring Explained Variation in Linear Mixed-Effects Models," *Statistics in Medicine*, 22, 3527–3541.