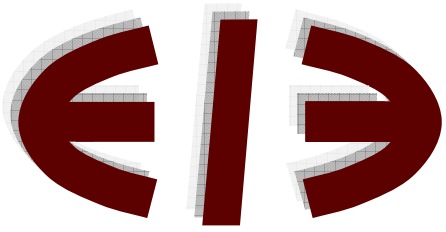
EERI
Economics and Econometrics Research Institute

Bayesian measures of explained variance and pooling in multilevel (hierarchical) models

Andrew Gelman and Iain Pardoe

EERI Research Paper Series No 4/2004



EERI
Economics and Econometrics Research Institute
Avenue de Beaulieu
1160 Brussels
Belgium

Tel: +322 299 3523
Fax: +322 299 3523
www.eeri.eu

Bayesian measures of explained variance and pooling in multilevel (hierarchical) models*

Andrew Gelman† Iain Pardoe‡

April 13, 2004

Abstract

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 we consider in this paper are characterized by hierarchical data structures in which individuals are grouped into units (which themselves might be further grouped into larger units), and there are variables 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 comparing to 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.

Keywords: adjusted R-squared, Bayesian inference, hierarchical model, multilevel regression, partial pooling, shrinkage

1 Introduction

1.1 Explained variation in linear models

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

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

where V represents the finite-sample variance operator, $\mathop{\mathrm{V}}_{i=1}^{n} x_i = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$. In a multilevel model (that is, a hierarchical model with group-level error terms or with regression coefficients β that vary by group), the predictors "explain" the data at different levels, and R^2 can be generalized in a variety of ways

*We thank Ronghui Xu, Cajo ter Braak, and Bill Browne for helpful comments and the National Science Foundation for financial support.

†Department of Statistics and Department of Political Science, Columbia University, New York, USA, gelman@stat.columbia.edu, http://www.stat.columbia.edu/∼gelman/

‡Charles H. Lundquist College of Business, University of Oregon, Eugene, USA, ipardoe@lcbmail.uoregon.edu

(for textbook summaries, see Kreft and De Leeuw, 1998, Snijders and Bosker, 1999, Raudenbush and Bryk, 2002, and Hox, 2002). Xu (2003) reviews some of these approaches, their connections to information theory, and similar measures for generalized linear models and proportional hazards models. Hodges (1998) discusses 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 paper we shall 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 no additional null models to be fit. 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.

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 $y_j \sim \text{N}(\alpha_j, \sigma_y^2)$, with population distribution $\alpha_j \sim \text{N}(\mu_\alpha, \sigma_\alpha^2)$ and hyperparameters $\mu_\alpha, \sigma_y, \sigma_\alpha$ known. For each group j, the multilevel estimate of the parameter α_j is

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

where

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

is a "pooling factor" that represents the degree to which the estimates are pooled together (that is, based on μ_α) 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 ($\hat{\alpha}_j = \mu_\alpha$), respectively. The (posterior) variance of the parameter α_j is

$$\text{var}(\alpha_j) = (1-\omega)\sigma_y^2. \tag{4}$$

The statistical literature sometimes labels $1-\omega$ as the "shrinkage" factor, a notation we find confusing since a shrinkage factor of zero corresponds to complete shrinkage towards the population mean. To avoid ambiguity, we use the "pooling factor" terminology in this paper. The form of expression (3) matches the form of the definition (1) of R^2, a parallelism we shall continue throughout.

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, $y_j \sim \text{N}(\alpha_j, \sigma_{y\,j}^2)$. Then ω_j can be defined as in (3), with $\sigma_{y\,j}$ in place of σ_y. Parameters with more precise data are pooled less towards 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 $\hat{\alpha}_j^{\text{multilevel}}$ vs. 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 paper we present a summary measure, λ, for the average amount of pooling at each level of a multilevel model. We shall introduce an example to motivate the need for such summaries, and then discuss the method and illustrate its 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 paper, 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. The model has predictors for both houses and counties, and we introduce it here in order to show the challenges in defining R^2 and λ 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. In order to identify the 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,

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

where y_{ij} is the logarithm of the radon measurement in house i in county j, $\texttt{basement}_{ij}$ 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 and third lines represent variations in radon levels and basement effects between counties, beyond what is explained by the county-

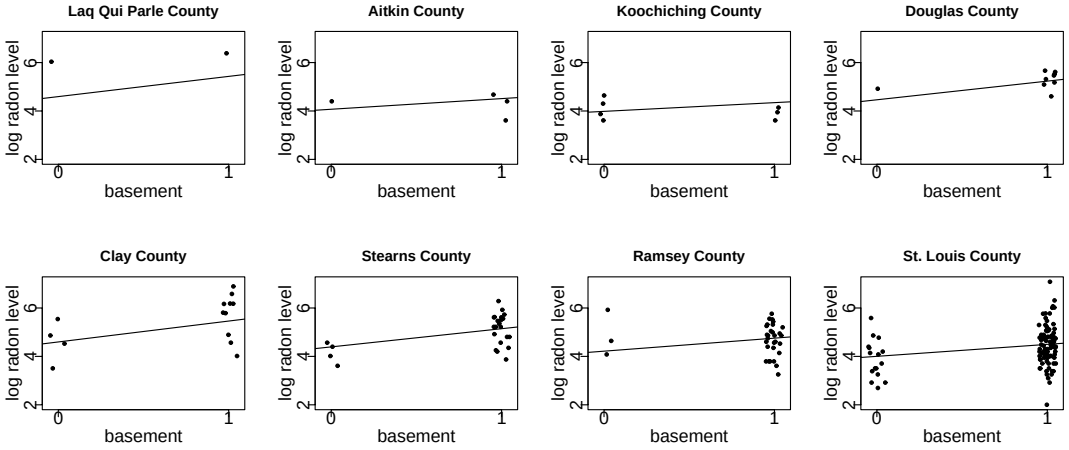


Figure 1: Jittered data and estimated regression lines from the multilevel model, $y = \alpha_j + \beta_j \cdot \text{basement}$, for radon data, displayed for 8 of the 85 counties j in Minnesota. Both the intercept and the slope vary by county. Because of the pooling of the multilevel model, the fitted lines do not go through the center of the data, a pattern especially noticeable for counties with few observations.

level uranium predictor. The between-county errors, α_j and β_j, are modeled as independent—see Section 5 for discussion of this point.

The hierarchical 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 counties, and Figure 2 shows the estimated county parameters and the county-level regression lines.

This example illustrates some of the challenges of measuring explained variance and pooling. The model has three levels, with a different variance component at each level. 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 α and β parameters affects the computation of explained variance for the data-level model—the simple measure of R^2 from least-squares regression will not be appropriate—and also for the county-level models, since 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) cannot in general 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.

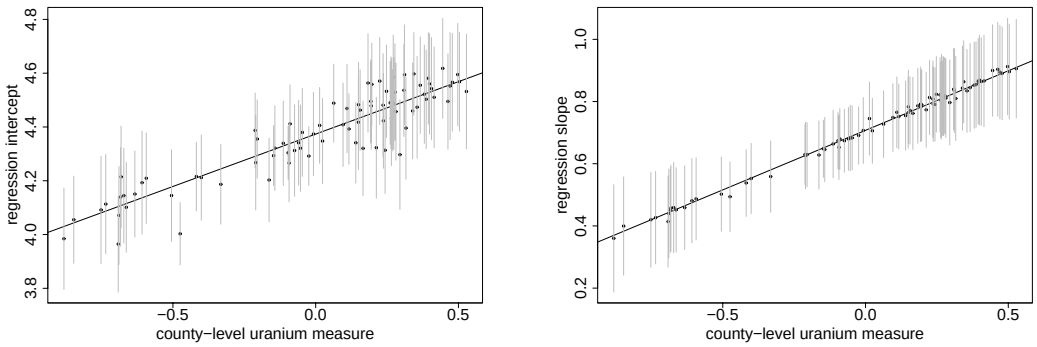


Figure 2: (a) Estimates ± standard errors for the county intercepts α_j, plotted vs. county-level uranium measurement u_j, along with the estimated multilevel regression line, $\alpha = \gamma_0 + \gamma_1 u$. (b) Estimates ± standard errors for the county slopes β_j, plotted vs. county-level uranium measurement u_j, along with the estimated multilevel regression line, $\beta = \delta_0 + \delta_1 u$. 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 hierarchical standard deviation parameters $\sigma_\alpha, \sigma_\beta$.

2 Summaries based on variance comparisons within a single fitted model

We define our generalizations of R^2 and pooling factors for each level of a multilevel model and then in Section 2.5 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. (For example, $M = 3$ in the radon model of Section 1.3.) At each level m, we write the model as,

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

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 zero 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 shall suppress the superscripts (m) for the rest of the paper.

The striking similarity of expressions (1) and (3), which define R^2 and λ, respectively, suggests that the two concepts can be understood in a common framework. We consider each to represent the fraction of variance explained, first by the linear predictor μ and then by the hierarchical model for ϵ.

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

$$R^2 = 1 - \frac{\mathrm{E}\left(\mathop{\mathrm{V}}_{k=1}^{K} \epsilon_k\right)}{\mathrm{E}\left(\mathop{\mathrm{V}}_{k=1}^{K} \theta_k\right)}. \tag{7}$$

In a Bayesian simulation context, the expectations in the numerator and denominator of (7) can be evaluated by averaging over posterior simulation draws, as we discuss in Section 2.5.

R^2 will be close to 0 when the average residual error variance is approximately equal to the average variance of the θ_k's. R^2 will be close to 1 when the residual errors ϵ_k are each close to zero 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, on average, the residual 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 residuals ϵ_k is reduced by the pooling of the hierarchical model. We define the pooling factor as

$$\lambda = 1 - \frac{\mathop{\mathrm{V}}_{k=1}^{K} \mathrm{E}(\epsilon_k)}{\mathrm{E}\left(\mathop{\mathrm{V}}_{k=1}^{K} \epsilon_k\right)}. \tag{8}$$

The denominator in this expression 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 will be 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 will be close to 1.

In Section 3.2.4, we discuss connections between the pooling factor (8) and the pooling factor ω defined in (3) for the basic hierarchical model.

2.4 Properties of the measures of explained variance and pooling

Since R^2 and λ are based on finite-population variances, they are well-defined for each level of a multilevel model, and automatically work even in the presence of predictors at that level. An alternative approach

based on hyperparameters could run into difficulties in such situations since the hyperparameters may not correspond exactly to the variance comparisons we are interested in.

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 towards 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, and Hox, 2002, for discussion and examples of this phenomenon). For the purpose of this paper, we merely note that a model can have different explanatory power at different levels.

2.5 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 et al., 1996, Carlin and Louis, 2001, and 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:
      i. Compute the vector of θ_k's, predicted values μ_k and the vector of residuals, $\epsilon_k = \theta_k - \mu_k$.
      ii. Compute the sample variances, $\mathop{\mathrm{V}}_{k=1}^{K} \theta_k$ and $\mathop{\mathrm{V}}_{k=1}^{K} \epsilon_k$.
   (b) Average over the simulation draws to estimate $\mathrm{E}\left(\mathop{\mathrm{V}}_{k=1}^{K} \theta_k\right)$ and $\mathrm{E}\left(\mathop{\mathrm{V}}_{k=1}^{K} \epsilon_k\right)$, 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)i above.
   (b) Compute $\mathop{\mathrm{V}}_{k=1}^{K} \mathrm{E}(\epsilon_k)$—that is, the variance of the K values of $\mathrm{E}(\epsilon_k)$—and then use this, along with $\mathrm{E}(\mathop{\mathrm{V}}_{k=1}^{K} \epsilon_k)$ from step 1(b) to calculate λ.

We compute R^2 and λ for each level; see Figure 3 for an illustration based on the radon data in Section 1.3. Appendix B shows the computations as implemented in Bugs (Spiegelhalter et al., 1994, 2003) and R (R Development Core Team, 2003).

3 Connections to 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 $y_i = (X\beta)_i + \epsilon_i, i = 1, \ldots, n$, with linear predictors $(X\beta)_i$ and errors ϵ_i that are normal with zero mean and constant variance σ^2.

3.1.1 Explained variance and adjusted R^2

If we plug in the least-squares estimate, $\hat{\beta} = (X^T X)^{-1} X^T y$, then the proportion of variance explained (7) simply reduces to the classical definition,

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

where I is the $n \times n$ identity matrix, $H = X(X^T X)^{-1} X^T$, and 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 β and σ. Under the standard noninformative prior density that is uniform on $(\beta, \log \sigma)$, the proportion of variance explained (7) becomes,

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

where p is the number of columns of X.

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

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

which is exactly classical adjusted R^2. Since $\frac{n-3}{n-p-2} > \frac{n-1}{n-p}$ 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, since the classical adjusted R^2 could be considered too high since 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, $\hat{\sigma}^2 = y^T(I–H)\,y/(n–p)$, 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 to the sample size, n, this pooling factor λ for the regression errors will be close to zero. This makes sense because, in this case, the classical residuals $(y - X\hat{\beta})_i$ are nearly independent, and they closely approximate the errors $\epsilon_i = (y - X\beta)_i$. 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, $y_{ij} \sim \mathrm{N}(\alpha_j, \sigma_y^2),\ i = 1, \ldots, n_j,\ j = 1, \ldots, J$, with population distribution $\alpha_j \sim \mathrm{N}(\mu_\alpha, \sigma_\alpha^2)$, and we can determine the appropriate variance comparisons at each of the two levels of the model. 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 $(\mu_\alpha, \log\sigma_y, \sigma_\alpha)$. 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,

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

where $y = (y_{11}, \ldots, y_{n1}, \ldots, y_{1J}, \ldots, y_{nJ})^T$ is the N-vector of responses, $\bar{I}_c$ 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 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 $(J–1)/J$), while 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 σ_y and σ_α, the proportion of variance explained, (7), at the data level is

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

Plugging in the estimators (9) leads to

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

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

3.2.2 Pooling factor λ for the data-level model

Conditional on σ_y and σ_α, the pooling factor, (8), at the data level is

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

Plugging in the estimators (9) leads to

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

If the within-group sample sizes n are reasonably large, this data-level pooling factor λ is close to zero, which makes sense because the data-level residuals are good approximations to the data-level errors (similar to the case of classical regression as discussed in Section 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 σ_y and σ_α, the pooling factor, (8), at the group level is

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

Plugging in the estimators in (9) leads to

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

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 2.5 to the example from Section 1.3 of home radon levels. 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 hierarchical model, with basement as an individual-level predictor and log uranium as a county-level predictor.

4. The full 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 2.5. We discuss the results for each model in turn.

1. Simple linear regression:

   - R^2 is very low, suggesting a poorly fitting model, and λ is essentially zero, 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). By comparison, the classical R^2 for this regression, plugging in the least-squares estimate for β, is $1 - y^T(I-H)\,y/y^T I_c\,y = 0.07$ (see Section 3.1.1). The theoretical value for λ for this model, is $1 - (n-3)/(n-p-2) = 0.07$ (see Section 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, hence little is changed by moving to fully-Bayesian inference.

2. One-way hierarchical model:

   - 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 zero. If there were equal sample sizes within each county, the theoretical value for R^2 for this data level model, based on plugging in the estimators (9), comes to 0.13 (see Section 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 0.05 (see Section 3.2.2).

Predictors included in the model	R^2 at each level:			λ at each level:		
	y	α	β	y	α	β
Basement (simple linear regression)	0.07			0.00		
County (simple one-way hierarchical model)	0.12	0		0.04	0.54	
Basement + county + uranium	0.21	0.73		0.03	0.77	
Basement + county + uranium + basement × county	0.21	0.53	0.83	0.03	0.81	0.97

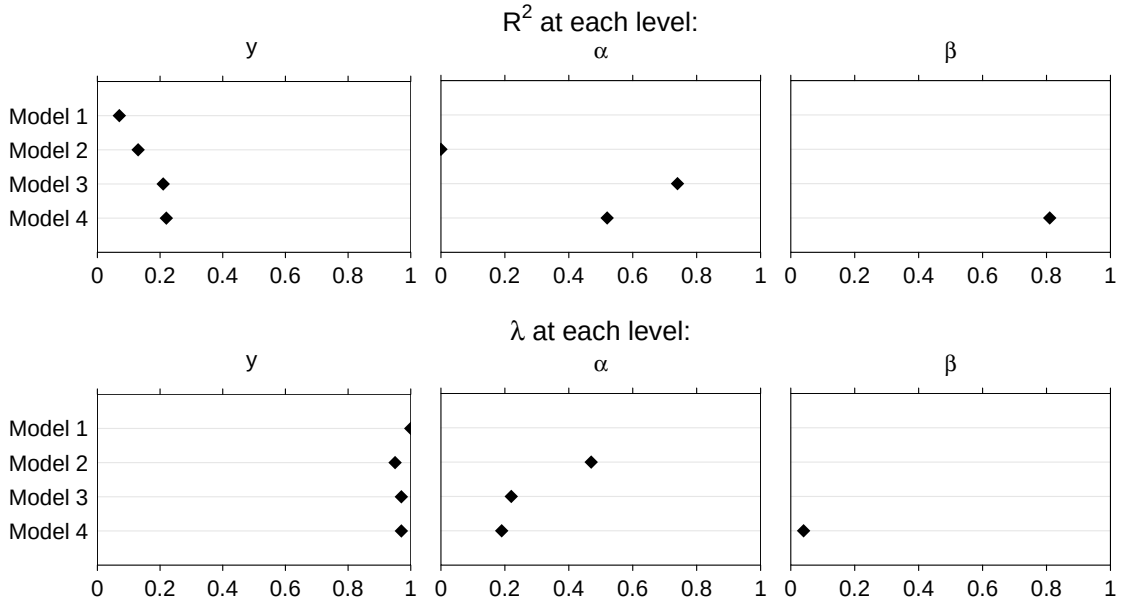


Figure 3: Proportion of variance explained and pooling factor at the level of data y, county-level intercepts α, and county-level slopes β, for each of four models fit to the Minnesota radon data. Blank entries indicate variance components that are not present in the given model. Results shown in tabular and graphical forms.

- At the county level, $R^2 = 0$ because this model has no county-level predictors. The pooling factor $\lambda = 0.54$ 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, the calculated value for λ for this county level model, plugging in the estimators (9), comes to 0.37 (see Section 3.2.4. In this case, accounting for unequal sample sizes and uncertainty in the variance parameters leads to a very different result.

3. Varying-intercept model:

- At the data level, R^2 shows further improvement over the one-way hierarchical model, but still remains quite low. The pooling factor λ remains close to zero.
- For the intercept model, $R^2 = 0.73$ indicates that if the basement effects are restricted to be the same in all counties, 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.

4. Varying-intercept, varying-slope model:

   - 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 Figure 1), and λ is still close to zero.
   - For the intercept model, R^2 is close to 50%—indicating that uranium level explains about half 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 is apparent in Figure 2a). 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 over 80%, implying that the uranium level explains much of the systematic variation in the basement effects across counties. The pooling factor λ is almost all the way to 1, which tells us that the slopes are almost entirely estimated from the county-level model, with almost no additional information about the individual counties (as can be seen in Figure 2b).

   The fact that much of the information in R^2 and λ is captured in Figures 1 and 2 should not be taken as 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.

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 2.5 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 towards 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 are modeled as independent. However, models such as the full varying-intercept, varying-slope model used in the home radon

application can be generalized to allow for correlated intercepts and slopes. The assumption of uncorrelated intercepts and slopes is often reasonable when there are useful predictors available for each grouping unit (as is the case for the home radon application). Nevertheless, it would be useful to extend R^2 and λ for use in situations where such an assumption was not reasonable.

We have presented our R^2 and λ measures in a Bayesian framework. However, 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 residuals 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. However, (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 et al., 2002, and Browne et al., 2003). An alternative for generalized linear models could be to develop analogous measures using deviances.

A Theoretical computations

A.1 Classical regression

The classical normal linear regression model can be written as $y_i = (X\beta)_i + \epsilon_i, i = 1, \ldots, n$, with linear predictors $(X\beta)_i$ and errors ϵ_i that are normal with zero mean and constant variance σ^2. If we plug in the least-squares estimate, $\hat{\beta} = (X^T X)^{-1} X^T y$, then (7) simply reduces to the classical definition, $R^2 = 1 - y^T(I-H)\,y / y^T I_c\, y$.

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

$$\begin{aligned}
(n-1)\,\mathrm{E}\left(\mathop{\mathrm{V}}_{i=1}^{n} \epsilon_i\right) &= y_c^T y_c - 2 y_c^T X_c \mathrm{E}(\beta) + \mathrm{E}(\beta^T (X_c^T X_c)\beta) \\
&= y_c^T y_c - 2 y_c^T X_c (X^T X)^{-1} X^T y + \sigma^2 \mathrm{tr}(X_c^T X_c (X^T X)^{-1}) + \\
&\qquad y^T X (X^T X)^{-1} X_c^T X_c (X^T X)^{-1} X^T y \\
&= y^T I_c\, y - 2 y^T H_c\, y + \sigma^2 \mathrm{tr}(H_c) + y^T H_c\, y \\
&= y^T I_c\, y - y^T H_c\, y + (p-1)\sigma^2 \\
&= y^T (I-H)\, y + (p-1)\sigma^2, \\
(n-1)\,\mathrm{E}\left(\mathop{\mathrm{V}}_{i=1}^{n} y_i\right) &= (n-1) \mathop{\mathrm{V}}_{i=1}^{n} y_i \\
&= y^T I_c\, y,
\end{aligned}$$

where I_c is the $n \times n$ matrix with $1 - 1/n$ along the diagonal and $-1/n$ off the diagonal, $H_c = I_c H$, $X_c = I_c X$, and $y_c = I_c\, y$. Conditional on σ, the proportion of variance explained, (7), is then

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

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

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

Similarly, conditional on σ,

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

and the pooling factor, (8), is then

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

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, $y_{ij} \sim \mathrm{N}(\alpha_j, \sigma_y^2),\ i = 1, \ldots, n_j,\ j = 1, \ldots, J$, with population distribution $\alpha_j \sim \mathrm{N}(\mu_\alpha, \sigma_\alpha^2)$. 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 $(\mu_\alpha, \log \sigma_y, \sigma_\alpha)$, the posterior distribution for α_j (conditional on σ_y and σ_α) is $\mathrm{N}(\omega\mu_\alpha + (1-\omega)\bar{y}_{.j},\ (1-\omega)\sigma_y^2/n)$, where $\omega = 1 - \sigma_\alpha^2/(\sigma_\alpha^2 + \sigma_y^2/n)$ 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 $y = (y_{11}, \ldots, y_{n1}, \ldots, y_{1J}, \ldots, y_{nJ})^T$ be the N-vector of responses, where $N = nJ$. Then if 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 $y_c = I_c\, y$. Similarly, let $\alpha = (\alpha_1, \ldots, \alpha_1, \ldots, \alpha_J, \ldots, \alpha_J,)^T$ be an N-vector of J stacked sets of population group means, each set containing n replicates, and let $\bar{y} = (\bar{y}_{.1}, \ldots, \bar{y}_{.1}, \ldots, \bar{y}_{.J}, \ldots, \bar{y}_{.J})^T$ be a similar N-vector of stacked sets of sample group means. Then if $\bar{I}_c$ 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 $\alpha_c = \bar{I}_c\, \alpha$, and the mean-centered vector of sample means can be written $\bar{y}_c = \bar{I}_c\, y$. Finally, let $\bar{I}$ 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 α_c (conditional on σ_y and σ_α) can be written $\mathrm{N}((1–\omega)\bar{y}_c,\ \bar{I}\,(1–\omega)\sigma_y^2/n)$.

We proceed in two stages. First, we average over the posterior distribution for α (conditional on σ_y and σ_α) 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 σ_y and σ_α does not result in closed-form solutions. As an alternative, we then plug-in particular point estimates for the variance components to find unconditional expressions.

At the data level, conditional on σ_y and σ_α,

$$\begin{aligned}
(N-1)\,\mathrm{E}\left(\underset{\mathrm{i,j}}{\mathrm{V}}\,(y_{ij} - \alpha_j)\right) &= y_c^T y_c - 2y_c^T \mathrm{E}(\alpha_c) + \mathrm{E}(\alpha_c^T \alpha_c) \\
&= y_c^T y_c - 2y_c^T (1-\omega)\bar{y}_c + \mathrm{tr}(\bar{I})\,(1-\omega)\sigma_y^2/n + (1-\omega)^2 \bar{y}_c^T \bar{y}_c \\
&= y^T I_c\, y + (\omega^2 - 1)\, y^T \bar{I}_c\, y + J(1-\omega)\sigma_y^2 \\
&= y^T (I_c - \bar{I}_c)\, y + \omega^2\, y^T \bar{I}_c\, y + J(1-\omega)\sigma_y^2, \\
(N-1)\,\mathrm{E}\left(\underset{\mathrm{i,j}}{\mathrm{V}}\, y_{ij}\right) &= (N-1)\ \underset{\mathrm{i,j}}{\mathrm{V}}\, y_{ij} \\
&= y^T I_c\, y.
\end{aligned}$$

So, conditional on σ_y and σ_α, the proportion of variance explained, (7), at the data level is

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

Similarly,

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

and conditional on σ_y and σ_α, the pooling factor, (8), at the data level is

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

At the group level, conditional on σ_y and σ_α,

$$\begin{aligned} n(J-1)\,\mathrm{E}\left(\underset{\mathrm{j=1}}{\overset{\mathrm{J}}{\mathrm{V}}}(\alpha_j-\mu_\alpha)\right) &= n(J-1)\,\mathrm{E}\left(\underset{\mathrm{j=1}}{\overset{\mathrm{J}}{\mathrm{V}}}\alpha_j\right) \\ &= (1-\omega)^2\,y^T\bar{I}_c\,y+J(1-\omega)\,\sigma_y^2, \\ n(J-1)\underset{\mathrm{j=1}}{\overset{\mathrm{J}}{\mathrm{V}}}\mathrm{E}(\alpha_j-\mu_\alpha) &= (1-\omega)^2\,y^T\bar{I}_c\,y. \end{aligned}$$

So, the proportion of variance explained, (7), is zero, while the pooling factor, (8), is

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

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

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

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

$$\begin{aligned} \frac{y^T\bar{I}_c\,y}{N} &= \left(\frac{J-1}{J}\right)\underset{\mathrm{j=1}}{\overset{\mathrm{J}}{\mathrm{V}}}\bar{y}_{\cdot j}, \\ \frac{y^T(I_c-\bar{I}_c)\,y}{N} &= \left(\frac{n-1}{n}\right)\underset{\mathrm{j=1}}{\overset{\mathrm{J}}{\mathrm{M}}}\underset{\mathrm{i=1}}{\overset{n}{\mathrm{V}}}y_{ij}. \end{aligned}$$

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

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

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

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

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

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

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

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 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 (a.raw=rnorm(J), b.raw=rnorm(J), sigma.y=runif(1),
        gamma.0.raw=rnorm(1), gamma.1.raw=rnorm(1), sigma.a.raw=runif(1), xi.a=rnorm(1),
        delta.0.raw=rnorm(1), delta.1.raw=rnorm(1), sigma.b.raw=runif(1), xi.b=rnorm(1))}
radon.parameters <- c ("a", "b", "sigma.y", "y.hat", "e.y",
                       "gamma.0", "gamma.1", "sigma.a", "a.hat", "e.a",
                       "delta.0", "delta.1", "sigma.b", "b.hat", "e.b")
radon.r2 <- bugs (radon.data, radon.inits, radon.parameters, "radon.r2.bug",
                  n.chains=3, n.iter=10000, n.thin=10)
```

It is then simple to 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))
# 0.21 0.53 0.83
print (round (c (lambda.y, lambda.a, lambda.b), 2))
# 0.03 0.81 0.97
```

Finally, we show the Bugs code for the three-level model. The implementation shown below looks somewhat complicated because we have used parameter expansion (Liu, Rubin, and Wu, 1998) to increase the speed of convergence of the hierarchical model. The regression parameters a and b are defined in terms of "raw" parameters $a^{\text{raw}}, b^{\text{raw}}$ and multiplicative factors ξ. The parameter-expansion formulation is not needed for computing R^2 and λ but in practice is an important tool for speeding computations in hierarchical models (see Gelman et al, 2003, Sections 11.9 and 15.4). The Gibbs sampler works faster by separately updating the raw parameters and the ξ's.

```
#   File radon.r2.bug with Bugs code for radon model with varying intercept and slope
#   Redundant multiplicative parameterization (Liu, Rubin, and Wu, 1998) used to improve
#   speed of convergence:  xi.a, xi.b, and the "raw" parameters are intermediate quantities.
```

```
  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] <- xi.a*a.raw[j]
      a.raw[j] ~ dnorm (a.raw.hat[j], tau.a.raw)
      a.raw.hat[j] <- gamma.0.raw + gamma.1.raw*u[j]
      a.hat[j] <- xi.a*a.raw.hat[j]
      e.a[j] <- a[j] - a.hat[j]

      b[j] <- xi.b*b.raw[j]
      b.raw[j] ~ dnorm (b.raw.hat[j], tau.b.raw)
      b.raw.hat[j] <- delta.0.raw + delta.1.raw*u[j]
      b.hat[j] <- xi.b*b.raw.hat[j]
      e.b[j] <- b[j] - b.hat[j]
    }
    xi.a ~ dnorm (0, .0001)
    gamma.0.raw ~ dnorm (0, .0001)
    gamma.1.raw ~ dnorm (0, .0001)
    gamma.0 <- xi.a*gamma.0.raw
    gamma.1 <- xi.a*gamma.1.raw
    tau.a.raw <- pow(sigma.a.raw, -2)
    sigma.a.raw ~ dunif (0, 1000)
    sigma.a <- abs(xi.a)*sigma.a.raw

    xi.b ~ dnorm (0, .0001)
    delta.0.raw ~ dnorm (0, .0001)
    delta.1.raw ~ dnorm (0, .0001)
    delta.0 <- xi.b*delta.0.raw
    delta.1 <- xi.b*delta.1.raw
    tau.b.raw <- pow(sigma.b.raw, -2)
    sigma.b.raw ~ dunif (0, 1000)
    sigma.b <- abs(xi.b)*sigma.b.raw
}
```

References

Afshartous, D., and De Leeuw, J. (2002). Decomposition of prediction error in multilevel models. Technical report, Department of Statistics, University of California, Los Angeles.

Box, G. E. P., and Tiao, G. C. (1973). *Bayesian Inference in Statistical Analysis*. New York: Wiley Classics.

Browne, W. J., Subramanian, S. V., Jones, K., and Goldstein, H. (2003). Variance partitioning in multilevel logistic models that exhibit over-dispersion. Technical report, School of Mathematical Sciences, University of Nottingham.

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

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. `www.stat.columbia.edu/~gelman/bugsR/`

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (2003). *Bayesian Data Analysis*, second edition. 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 B*.

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

Goldstein, H., Browne, W. J., and Rasbash, J. (2002). Partitioning variation in multilevel models. *Understanding Statistics* **1**, 223–232.

Hodges, J. S. (1998). Some algebra and geometry for hierarchical models, applied to diagnostics (with discussion). *Journal of the Royal Statistical Society B* **60**, 497–536.

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

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

Liu, C., Rubin, D. B., and Wu, Y. N. (1998). Parameter expansion to accelerate EM: the PX-EM algorithm. *Biometrika* **85**, 755–770.

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.

R Development Core Team (2003). R: a language and environment for statistical computing. Vienna: R Foundation for Statistical Computing. `www.r-project.org`

Raudenbush, S. W., and Bryk, A. S. (2002). *Hierarchical Linear Models*, second edition. Thousand Oaks, Calif.: Sage.

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

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, Cambridge, England. `www.mrc-bsu.cam.ac.uk/bugs/`

Wherry, R. J. (1931). A new formula for predicting the shrinkage of the coefficient of multiple correlation. *Annals of Mathematical Statistics* **2**, 440–457.

Xu, R. (2003). Measuring explained variation in linear mixed effects models. *Statistics in Medicine* **22**, 3527–3541.