`apsl2lme`: A Model-Selection Diagnostic Tool for Hierarchical Linear Models

Kim Nimon
University of Texas at Tyler

A primary goal in regression is to choose the simplest model that provides the best fit to the observed data (Thompson, 2006; West, Welch, & Galecki, 2014). In ordinary least squares regression, this may be a simple process of examining the relationship between the number of predictors and the resulting Multiple R^2 or Adjusted R^2 (Thompson, 2006). However, in multilevel modeling, the model selection process is more complicated. Not only must researchers consider which variables to include in the model, they must also determine whether level-1 variables should be modeled as random effects (West et al., 2104). The purpose of this paper is to present a software-based model-selection diagnostic tool that supports two-level models with a single grouping factor.

Multilevel modeling (MLM) supports data that are measured in clusters or at multiple levels of a hierarchy. A typical MLM dataset includes some level-1 units (e.g., students or measurement occasions) nested inside level-2 units (e.g., schools or years). Although three-, four- and even five-level structures are plausible, multilevel models are typically easier to interpret when they are limited to two levels.

Because hierarchical linear modeling (HLM) is a generalization of ordinary least squares (OLS) (Newman, Newman, & Salzman, 2010, p. 1), one might expect that researchers apply similar techniques to analyze clustered data as are available for non-clustered data. However, gaps in knowledge, tools, and analytic strategies prohibit such application. For example, in OLS regression, much literature and software exists to help researchers report the appropriate effect size for an OLS regression model (e.g., Nimon, Oswald, & Roberts, 2013; Yin & Fan, 2001) and interpret results in the face of multicollinearity including commonality and dominance analysis (e.g., Courville & Thompson, 2001; Nimon, Lewis, Kane & Hayes, 2008; Nimon & Oswald, 2013). However, effect sizes for multilevel models are still in development (LaHuis, Hartman, Hakoyama, & Clark, 2014; Luo & Azen, 2013) and few studies have examined the consequences of multicollinearity on the context of multilevel models (Shieh & Fouladi, 2003). Further, current techniques that consider "predictor importance" in multilevel models (i.e., dominance analysis, Pratt's index) are only applicable to random intercept regression models (Liu, Zumbo, Wu, 2014; Luo & Azen). Similarly, although software is available to help researchers identify parsimonious OLS regression models by considering all-possible-subsets (e.g., Lumley, 2017), it appears that such software has yet to be made available to support multilevel models. The purpose of this paper is to make such software available and to demonstrate its use.

Thompson (2006) noted "in predictive applications, the researcher may seek a parsimonious (smaller) set of predictors that may still yield an acceptable R^2" (p. 277). In OLS, this is a straightforward process of computing the R2 for every combination of predictors. For example, researchers can plot the resulting R^2s from an all-possible-subsets regression by the number of predictors to inform the researcher's judgment as to the number of predictors to retain. Such analyses is aided by the fact that R^2 increases each time a predictor is added to a model. However, the property of monotonicity for R^2 analogues in multilevel models is limited to models with random intercepts as the "problems raised by a random slope model are still not yet solved" (Liu et al., 2014, p. 7).

Researchers wanting to compare multi-level models from an all-possible-subsets analysis may therefore consider the deviance (-2 log-likelihood) as measure of model fit as a lower deviance always implies better fit. The disadvantage of the deviance is that a model fit to the same data with more parameters will always have better fit (i.e., smaller deviance) (Luke, 2004). While smaller deviance is good in that researchers generally want to maximize the fit of the model to the data, researchers also often desire parsimonious models – those that are able to explain the data with as few parameters as possible. Therefore, researchers may also consider the Akaike Information Criterion (AIC. Akaike, 1973) and/or Bayesian Information Criterion (BIC, Schwarz, 1978) when making model comparisons. Both AIC and BIC are based on the deviance where smaller is better but incorporate different penalties for the number of model parameters estimated. The AIC penalizes the fit of a model by adding twice the number of parameters being estimated

to the deviance while the BIC applies a greater penalty by adding the product of the number of parameters estimated by the natural logarithm of the number of observations in the model (West, Welch, & Galecki, 2014). However, by their nature, AIC and BIC are not monotonic.

apsl2lme and plot.apsl2lme

In order to facilitate data analysis and accessibility, the statistical package R was used. R is a free statistical programming language and environment for the Unix, Windows and Mac families of operating systems (R Core Team, 2019). Two new functions, `apsl2lme` and `plot.apsl2lme` were written in R to perform the multilevel equivalent of an all-possible -subsets analysis.

apsl2lme

The function `apsl2lme` receives output from the `lme` function and conducts a series of multilevel analyses using the variables identified in the `lme` output. For more information on the use of the `lme` function, readers are encouraged to consult Crawley (2013). The `apsl2lme` function identifies all possible combinations of fixed effects and all possible combinations of random effects for each fixed effect permutation. Note that the `apsl2lme` function considers all possible random effects even if they are not identified in the original function. In this way, the function serves as a tool for model building that supports the "top-down strategy" and "step-up strategy" presented by West et al. (2014). For each model derived, the `apsl2lme` function calls the `lme` function and captures the resulting model summary statistics. However, before executing these permutations, the function eliminates any missing data so that meaningful comparisons can be made between fit indices. The `apsl2lme` function also summarizes model metrics and fit indices and captures model formulas that do not converge. All models are identified by a unique identifier (i.e., model ID) for subsequent analysis.

plot.apsl2lme

The function `plot.apsl2lme` receives output from the `apsl2lme` object and plots model IDs, depicting differences in fit indices and variance components by degrees of freedom. Three plots are produced depicting relationships between fit indices (specifically -2LL, AIC and BIC) and degrees of freedom. Two plots are produced depicting relationships between variance components and degrees of freedom. The plots only include the variance components for the intercept ($\sigma^2_{u_{0j}}$) and the residual variance ($\sigma^2_{e_{ij}}$) since additional variance components for random effects only represent a subset of all models tested.

Calling Sequence

As depicted in the Appendix, the typical usage sequence involves: (a) loading the `nlme` library (Pinheiro, Bates, DebRoy, Sarkar, & R Core Team, 2018), `graphics` library (R Core Team, 2018), and the `foreign` library (R Core Team, 2017), (b) calling the `lme` function with a well-specified model, (c) calling the `apsl2lme` function with the output of the `lme` function, and (d) calling the `plot.apsl2lme` function with the output of the `apsl2lme` function.

To illustrate the usage, we replicated the analysis reported in Kreft and de Leeuw (1998) where a two-level model was built to analyze math achievement scores from a subset of scores from the NELS-88 dataset (available at https://stats.idre.ucla.edu/wp-content/uploads/2016/02/imm23-1.sav). Identified as Model 6 (Equation 1), school-level variables 'public' and student-level variables 'homework' and 'white' were used to model variance in math achievement scores. The student-level variables were also modeled as random effects. The complete model tested was then:

$$math_{ij} = \gamma_{00} + (\gamma_{10} + u_{1j})homework_{ij} + (\gamma_{20} + u_{2j})white_{ij} + \gamma_{10}public_j + u_{0j} + e_{ij} \qquad (1)$$

After loading the necessary libraries and software and reading in the dataset (see Appendix, lines 3 – 12), Equation 1 was specified using the function `lme` (see Appendix, lines 14 – 15). The results of the call to `lme` was saved to an object named `lme.out` which is presented in Figure 1.

```
(lme.out<-lme(data=Dataset,math~homework+white+public,random=
~white+homework| schid,method='ML'))

Linear mixed-effects model fit by maximum likelihood
 Data: dt
       AIC      BIC    logLik
  3640.854 3687.625 -1809.427

Random effects:
 Formula: ~homework + white | schid
 Structure: General positive-definite, Log-Cholesky parametrization
            StdDev   Corr
(Intercept) 8.025397 (Intr) homwrk
homework    3.960801 -0.849
white       4.902999 -0.513  0.142
Residual    7.152161

Fixed effects: list(formfix)
               Value Std.Error  DF   t-value p-value
(Intercept) 48.17410  2.271634 494 21.206802  0.0000
homework     1.94666  0.880458 494  2.210968  0.0275
white        2.67606  1.507105 494  1.775630  0.0764
public      -4.93191  1.582288  21 -3.116947  0.0052
 Correlation:
         (Intr) homwrk white
homework -0.663
white    -0.537  0.083
public   -0.478  0.011  0.006

Standardized Within-Group Residuals:
       Min         Q1        Med         Q3        Max
-2.3131498 -0.6867736 -0.0041486  0.7068647  2.7363296

Number of Observations: 519
Number of Groups: 23

            numDF denDF  F-value p-value
(Intercept)     1   494 4518.088  <.0001
homework        1   494    4.424  0.0359
white           1   494    3.225  0.0732
public          1    21    9.715  0.0052
```

Figure 1. Original model (Model 18).

Results

Processing the results of the object returned from `lme`, the `apsl2lme` function derives all possible models and returns an object with the following components:

- `ModelComparison` - This component presents the following information for each model analyzed: fixed effects formula, random effects formula, degrees of freedom, deviance, AIC, BIC, residual error variance, variance of intercept, and variance of each level 1 variable. For models where variance in the slope of level 1 variables are not modeled, the table indicates NA.

Table 1. Example Model Comparison Output from `apsl2lme`

#	Fixed Effects	Random Effects	*df*	-2LL	AIC	BIC	Var (Res)	Var (Intr)	Var (homework)	Var (white)
1	math ~ 1	~1 | schid	3	3800.78	3806.78	3819.53	81.24	24.85	NA	NA
2	math ~ homework	~1 | schid	4	3730.49	3738.49	3755.50	71.14	20.23	NA	NA
3	math ~ homework	~homework | schid	6	3639.04	3651.04	3676.55	53.30	59.28	16.79	NA
4	math ~ white	~1 | schid	4	3785.20	3793.20	3810.21	79.61	18.67	NA	NA
5	math ~ white	~white | schid	6	3784.80	3796.80	3822.31	79.01	21.18	NA	5.36
6	math ~ public	~1 | schid	4	3793.67	3801.67	3818.68	81.22	17.20	NA	NA
7	math ~ homework + white	~1 | schid	5	3717.60	3727.60	3748.86	70.18	14.68	NA	NA
8	math ~ homework + white	~homework | schid	7	3627.86	3641.86	3671.62	52.57	55.14	16.33	NA
9	math ~ homework + white	~white | schid	7	3716.23	3730.23	3759.99	69.22	18.22	NA	10.97
10	math ~ homework + white	~homework + white | schid	10	3625.32	3645.32	3687.84	51.54	61.90	16.38	9.38
11	math ~ homework + public	~1 | schid	5	3725.66	3735.66	3756.92	71.13	15.77	NA	NA
12	math ~ homework + public	~homework | schid	7	3634.84	3648.84	3678.60	53.34	56.25	16.37	NA
13	math ~ white + public	~1 | schid	5	3776.99	3786.99	3808.25	79.60	11.88	NA	NA
14	math ~ white + public	~white | schid	7	3775.07	3789.07	3818.84	78.58	21.09	NA	13.20
15	math ~ homework + white + public	~1 | schid	6	3711.95	3723.95	3749.46	70.18	10.68	NA	NA
16	math ~ homework + white + public	~homework | schid	8	3623.25	3639.25	3673.27	52.64	52.28	15.84	NA
17	math ~ homework + white + public	~white | schid	8	3707.88	3723.88	3757.90	68.67	20.86	NA	31.66
18	math ~ homework + white + public	~homework + white | schid	11	3618.85	3640.85	3687.63	51.15	64.41	15.69	24.04

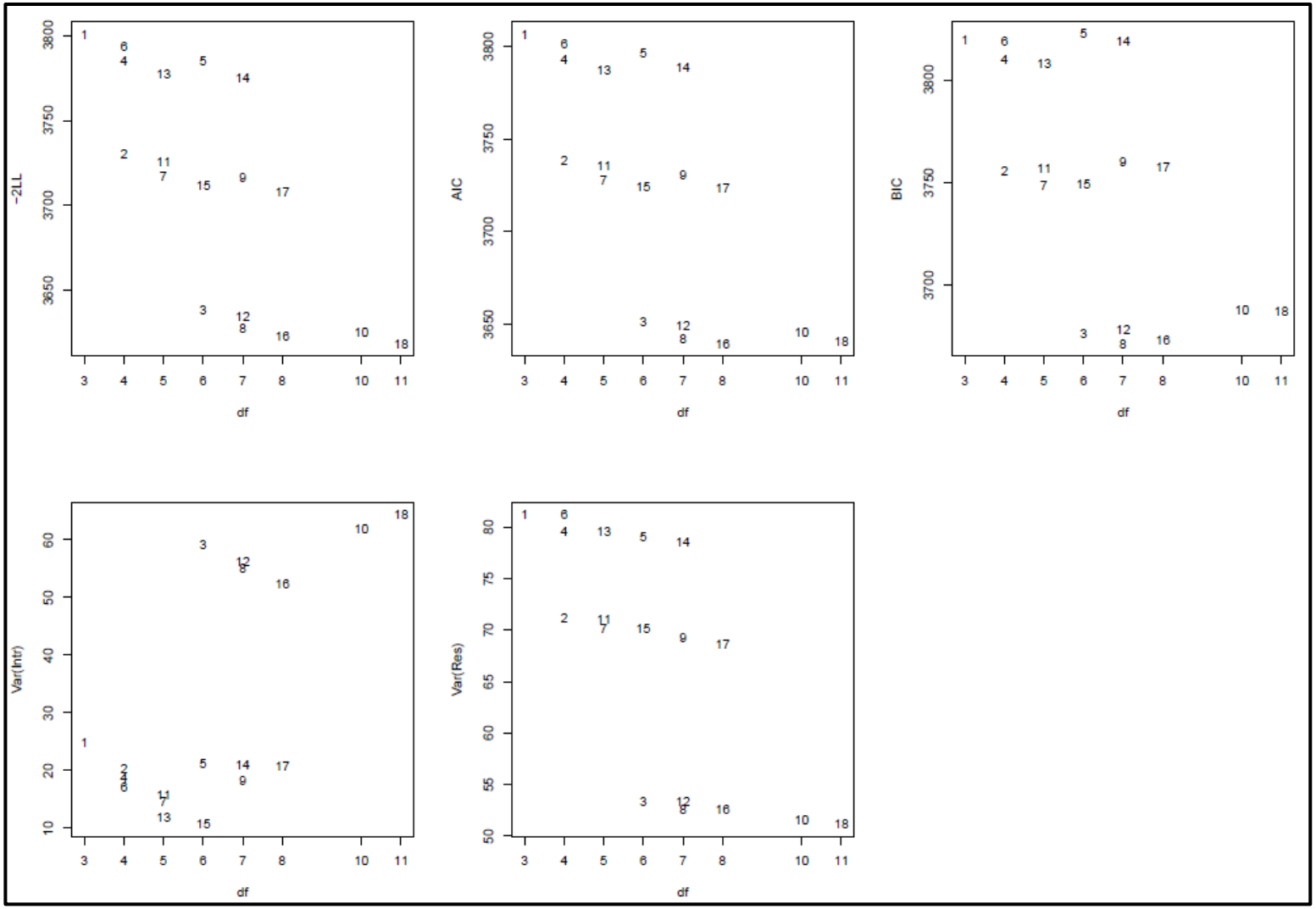

Figure 2. Example model ID comparisons of relationships between degrees of freedom (df) and -2LL (upper left), AIC (upper middle), BIC (upper right), level 2 intercept variance (lower left), and level 1 residual variance (lower middle).

- `ModelSummary` - This component contains the summary output from the `lme` command and fixed effect *F* test results for each model identified in the `ModelComparison`.
- `InvalidModels` - This component contains the fixed effects and random effects formula for models that did not converge.

After running the command to conduct the all-possible-subsets analysis (see Appendix, line 17), and examining the `ModelComparison` object (see Appendix, line 18), one can see that `apsl2lme` ran 18 separate models as depicted in Table 1 (see Appendix, line 19 to save `ModelComparison` as a comma separated file). The results help identify other competing models to the original model (Model 18) which is the best fitting model according to deviance. For example, by sorting the `ModelComparison` object by the different fit indices, one can see that Model 18 is the second best fitting model according to AIC and the fifth according to BIC (see Table 2 and Appendix, lines 21 - 30). One also sees that the residual variance is highest in Model 1 and lowest in Model 18 but that the variance of the intercept is smallest in Model 15.

Processing the results of the object returned from `apsl2lme, plot.apsl2lme` produces a graphical presentation of model IDs (see Appendix, line 32). As depicted in Figure 2, five graphs are produced. The upper right hand graph plots models IDs by deviance (-2LL) and degrees of freedom (*df*). The upper middle graph plots model IDs by AIC and *df*. The upper right graph plots model IDs by BIC and *df*. The lower left hand graph plots model IDs by intercept variance [Var(Int)] and *df*. The lower middle graph plots model IDs by residual variance [Var(Res)] and *df*. These plots allow researchers to examine model fit and variance components by degrees of freedom. For example, one can see that given 8 degrees of freedom, model 16 outperforms model 17 according to AIC and BIC.

Discussion

A primary goal in regression is to choose the simplest model that provides the best fit to the observed data (cf., Thompson, 2006; West et al., 2014). In OLS regression, this is the somewhat simple process of examining the relationship between the number of predictors in the model and the resulting Multiple R^2 or Adjusted R^2 (Thompson, 2006). However, in multilevel modeling, the model selection process is more complicated. Not only must researchers consider which variables to include in the model, they must also determine whether level 1 variables should be modeled as random effects (West et al.). Although such decisions should be guided by theory, there may be times when researchers are forced to consider more technical approaches in determining the components of their models (Kreft & de Leeuw, 1998). Oftentimes, researchers compute the statistical significance of a *t*-statistic to help decide whether an effect should be kept in a model. However, such techniques are fraught with difficulties in the face of correlated variables (Kreft & de Leeuw, 1998; Shieh & Fouladi, 2003). Kreft and de Leeuw (1998) noted that in the absence of theory, researchers should use model fit as a criterion as it is a more reliable measure than individual parameter estimates.

As no single information criterion stands apart as the best criterion to use when selecting multilevel models (Gurka, 2006), the `apsl2lme` function provides three measures of model fit: (a) deviance, (b) AIC, and (c) BIC. For ease of interpretation, all models tested in the `apsl2lme` function employ the maximum likelihood (ML) method of estimation. For all three measures of model fit, smaller values are considered better fit.

In practice, researchers can use the `apsl2lme` function to identify differences in model fit based on all possible combinations of parameters. In the event that not all possible combinations of parameters are of interest, researchers can compare a subset of models that are most theoretically appropriate to their applications (e.g., subsets based on degrees of freedom or random effect formula). The benefit of the `apsl2lme` function is that all possible models are available for the researcher to investigate. No longer does a researcher have to worry if a decision to exclude a variable early in the model-building stage is still appropriate when other variables are deleted or entered into the model.

As previously noted, the model selection process should be guided by theory. However, the technical approach of selecting models based on model fit might also be used to inform or confirm theory. In Figure 1, one sees that models 3, 8, 10, 12, and 16 are comparable to the original model (18) analyzed by Kreft and de Leeuw (1998) and are among the best fitting models. One can also see that models 4, 5, 6, 13, and 14 are not that much better than the null model with no predictors. These data seem to suggest that home-

Table 2. Model # Sorted by Global Fit Indices and Variance Components

Order	-2LL	AIC	BIC	Var(Res)	Var(Intr)
1	18	16	8	18	18
2	16	18	16	10	10
3	10	8	3	8	3
4	8	10	12	16	12
5	12	12	18	3	8
6	3	3	10	12	16
7	17	17	7	17	1
8	15	15	15	9	5
9	9	7	2	7	14
10	7	9	11	15	17
11	11	11	17	11	2
12	2	2	9	2	4
13	14	13	13	14	9
14	13	14	4	5	6
15	5	4	6	13	11
16	4	5	14	4	7
17	6	6	1	6	13
18	1	1	5	1	15

Note. Indices and component sorted from lowest to highest with the exception of Var(Intr).

work is a critical component in explaining variance in math achievement and that its effect on math achievement differs from school to school regardless of what other variables are included in the model. The data (models 5, 9, 14, and 17) also suggest that the effect of being white on math achievement does not vary by school regardless of what other variables are included in the model. Although answering questions such as these may be of interest to researchers, it seems unlikely that such an exhaustive model comparison strategy would be possible without the support of software that consider the combinations of all potential fixed and random effects.

Conclusions and Future Development

The `apsl2lme` and `plotlaps.l2lme` functions presented in this paper provide researchers a straight forward approach to comparing models derived from multilevel data in the context of two levels with one grouping factor. The R functions are currently available at:

https://www.dropbox.com/s/qvb66pnpn3vew6f/apsl2lme.r?dl=0.

It is the intention of the authors to continue development on this function and to publish the functions on The Comprehensive R Archive Network (CRAN).

Further improvements could include updating the software to accommodate three level data structures, supporting output from `lmer` (Bates, Maechler, & Bolker, & Walker, 2015), and migrating the software so that it could be utilized with other statistical software packages (e.g., SPSS, SAS). Further extensions to the software could also include computing the Extension Information Criterion (EIC, Ishiguro, Sakamoto, & Kitagawa, 1997).

It is recommended that researchers employing multilevel modeling techniques consider analyzing competing models to determine the parsimonious solutions to answer their questions. When the problems raised by models with random slopes are solved (cf. Liu et al., 2014), researchers may also want to couple with results of the all-possible-subsets analyses described in this paper with "predictor importance" metrics (e.g., commonality analysis coefficients, dominance analysis coefficients, Pratt's index) that will likely emerge once R^2 analogues in multilevel models are developed that have the property of monotonicity.

References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In B. N. Petrov & F. Csaki (Eds.) *2nd International Symposium on Information Theory* (pp. 267-281). Budapest: Akademia Kiado.

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lm4. *Journal of Statistical Software, 67*(1), 1-48. doi:10.18637/jss.v067.i01.

Courville, T., & Thompson, B. (2001). Use of structure coefficients in published multiple regression articles: Beta is not enough. *Educational and Psychological Measurement, 61*, 229-248. doi:10.1177/0013164401612006

Crawley, M. J. (2013). *The R book* (2nd ed.). West Sussex, England: Wiley.

Gurka, M. J. (2006). Selecting the best linear mixed model under REML. *American Statistician, 60*(1), 19-26. doi:10.1198/000313006X90396

Ishiguro, M., Sakamoto, Y., & Kitagawa, G. (1997) Bootstrapping log likelihood and EIC, an extension of AIC. *Annals of the Institute of Statistical Mathematics*, *49*, 411-434. doi: doi.org/10.1023/A:1003158526504

Kreft, I. G. G., & de Leeuw, J. (1998). *Introducing multilevel modeling*. Thousand Oaks, CA: Sage. doi: 10.4135/9781849209366

LaHuis, D. M., Hartman, M. J., Hakoyama, S., & Clark, P. C. (2014). Explained variance measures for multilevel models. *Organizational Research Methods, 17*, 433-451. doi: 10.1177/1094428114541701

Liu, Y., Zumbo, B. D., & Wu, A. D. (2014). Relative importance of predictors in multilevel modeling. *Journal of Modern Applied Statistical Methods, 13*(1), 2-22. doi: 10.22237/jmasm/1398916860

Luke, D. A. (2004). *Multilevel modeling* (Vol. 143). Thousand Oaks, CA: Sage. doi: 10.4135/9781412985147

Lumley, T., using Fortran code by A. Miller (2017). *leaps: regression subset selection*. R package version 3.0, Retrieved from https://CRAN.R-project.org/package=leaps

Luo, W., & Azen, R. (2013). Determining predictor importance in hierarchical linear models using dominance analysis. *Journal of Educational and Behavioral Statistics*, *38*(1), 3-31. doi: 10.3102/1076998612458319

Newman, D., Newman, I., & Salzman, J. (2010). Comparing OLS and HLM models and the questions they answer: Potential concerns for type VI errors. *Multiple Linear Regression Viewpoints*, *36*(1), 1-8.

Nimon, K., Lewis, M., Kane, R., & Haynes, R. M. (2008). An R package to compute commonality coefficients in the multiple regression case: An introduction to the package and a practical example. *Behavior Research Methods, 40*, 457-466. doi:10.3758/BRM.40.2.457

Nimon, K., & Oswald, F. L. (2013). Understanding the results of multiple linear regression: Beyond standardized regression coefficients. *Organizational Research Methods, 16*, 650-674. doi:10.1177/1094428113493929

Nimon, K., Oswald, F., & Roberts, J. K. (2013). *yhat: Interpreting regression effect*. R package version 2.0-0, Retrieved from http://CRAN.R-project.org/package=yhat

Pinheiro, J., Bates, D., DebRoy, S., & Sarkar, D., & R Core Team (2018). *nlme: Linear and nonlinear mixed effects models*. R package version 3.1-131.1, Retrieved from https://CRAN.R-project.org/package=nlme.

R Core Team (2017). R: A language and environment for statistical computing [Computer software]. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from http://www.R-project.org.

R Core Team (2018). R: A language and environment for statistical computing [Computer software]. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from http://www.R-project.org.

R Core Team (2019). R: A language and environment for statistical computing [Computer software]. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from http://www.R-project.org.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics, 6*, 461-464. doi: 10.1214/aos/1176344136

Shieh, Y., & Fouladi, R. (2003). The effect of multicollinearity on multilevel modeling parameter estimates and standard errors. *Educational and Psychological Measurement, 63*, 951-985. doi:10.1177/0013164403258402

Thompson, B. (2006). *Foundations of behavioral statistics: An insight-based approach*. New York: Guilford Press.

West, B. T., Welch, K., B., & Galecki, A. T. (2014). *Linear mixed models: A practical guide using statistical software* (2nd ed.). Boca Raton, FL: Chapman & Hall.

Yin, P., & Fan, X. (2001). Estimating R^2 shrinkage in multiple regression: A comparison of different analytical methods. *Journal of Experimental Education, 69*, 203-225. doi:10.1080/00220970109600656.

Send correspondence to: Kim Nimon
University of Texas at Tyler
Email: kim.nimon@gmail.com

APPENDIX

R Code for Illustrative Analyses

```
library (nlme)
library (graphics)
library (foreign,pos=4)
source("apsl2lme.r")

Dataset <-
read.spss("https://stats.idre.ucla.edu/wp-content/uploads/2016/02/imm23-
1.sav",
   use.value.labels=FALSE, to.lower.case=TRUE,max.value.labels=Inf,
   to.data.frame=TRUE)
colnames(Dataset) <- tolower(colnames(Dataset))

(lme.out<- lme(data=Dataset, math~homework+white+public,
     random = ~white+homework| schid,method='ML',na.action=na.omit))

apsOut<-apsl2lme(lme.out)
(MC<-apsOut$ModelComparison)
write.csv(MC,"aout.csv")

(mcdev<-MC[order(MC$"-2LL"),])
(mcaic<-MC[order(MC$AIC),])
(mcbic<-MC[order(MC$BIC),])
(mcres<-MC[order(MC$"Var(Res)"),])
(mcint<-MC[order(MC$"Var(Intr)",decreasing=T),])
fit<-cbind(rownames(mcdev),rownames(mcaic),rownames(mcbic),rownames(mcres),
     rownames(mcint))
colnames(fit)<-colnames(MC)[4:8]
fit
write.csv(fit,"fit.csv")

plot.apsl2lme(MC)
```