Paper PO-013

MIXED_DX: A SAS® Macro for Two-Level Linear Model Diagnostics

Bethany A. Bell, Grant B. Morgan, Jason A. Schoeneberger
University of South Carolina
John M. Ferron, Jeffrey D. Kromrey
University of South Florida

ABSTRACT

Multilevel modeling has become a common analytic technique across a variety of disciplines including medicine and the social and behavioral sciences. However, because many researchers who use multilevel modeling in their research do not report if the data were screened for potential violations of distributional assumptions and outliers, it is unclear if these diagnostic procedures are being conducted. Thus, in an effort to make the process of checking the assumptions for two-level linear models easier for the applied researcher, this paper provides a SAS macro for conducting two-level linear model diagnostics, including examinations of residual normality, linearity, homogeneity of variance, and influential outliers. By utilizing data from PROC MIXED ODS tables, the macro produces box-and-whisker plots, summary tables of statistical output, histograms, and plots of residuals by predicted values. Macro outputs are produced for both level-1 and level-2 data. This paper provides the macro programming language, as well as results from an executed example of the macro.

INTRODUCTION

As multilevel modeling techniques (also called hierarchical linear modeling or mixed modeling) continue to be used with increasing frequency across a variety of disciplines, it is important for researchers to conduct appropriate model diagnostics when using these statistical methods. Like most parametric statistical procedures, there are certain distributional assumptions underlying the validity of the Type I error control when conducting multilevel analyses. Thus, just as researchers were trained to evaluate the underlying assumptions for multiple regression and ANOVA, they, too, need to assess and report on the assumptions associated with multilevel models. More specifically, although most of the assumptions associated with two-level linear models are similar to OLS model assumptions (i.e., residual normality, linearity, and homogeneity of variance), few articles in which multilevel modeling techniques are used contain this information, despite recommendations to be transparent when reporting multilevel modeling research processes and findings (Ferron, Hogarty, Dedrick, Hess, Niles, & Kromrey, 2008).

One likely reason for the omission of this information could be related to the cumbersome nature of conducting model diagnostics of two-level linear models. First, it is not easy to locate all of the data necessary to conduct diagnostics for both level-1 and level-2 residuals. For example, the level-1 and level-2 residuals are stored in two different ODS tables, which not only have to be requested separately, but the commands necessary to request each ODS table occurs at two different places within the PROC MIXED code (i.e., the ODS table for level-1 residuals is an option that can be added to the model statement whereas the ODS table for level-2 residuals is requested through an ODS OUTPUT command). Second, the ODS table that contains the level-2 residuals does not contain predicted values for level-2 units. Thus, the data must be manipulated to generate these values. Third, although PROC MIXED does allow the INFLUENCE option on the model statement, the oodles of data produced by this option can be overwhelming, especially to a novice PROC MIXED user. Thus, in an effort to reduce the burden of conducting these important, yet unwieldy, analyses this paper provides a SAS macro to conduct diagnostic evaluations of two-level linear models, including examinations of residual normality, linearity, homogeneity of variance, and influential outliers. Diagnostics are conducted for both level-1 and level-2 data and the macro includes the data screening techniques recommended by authors such as Hox (2002), Longford (2001), and Raudenbush and Bryk (2002).

STATISTICAL ASSUMPTIONS AND THEIR EVALUATION

Raudenbush and Bryk (2002) correctly assert that the validity of inferences based on models depend on the degree to which assumptions are upheld about the structural and random parts of the model. They go on to suggest that “skillful analysts pay close attention to the assumptions required by their models” (p. 253). Unfortunately, based on reporting practices, it seems that the assumptions frequently go unexamined or under-examined by analysts conducting multilevel modeling (Ferron et al., 2008). However, it is not clear if this common omission is a function of a lack of understanding or the effort and time needed to assess model adequacy. This work seeks to address the latter issue by facilitating the generation of information necessary to assess to the tenability of the model assumptions.

Under Ordinary Least Squares (OLS) regression, the error terms are assumed to be normally distributed, independent, and homoscedastic. These same assumptions apply to linear multilevel models. In particular, the assumption of normally distributed errors must be made for both level-1 and level-2 variables, with violations adversely affecting level-2 estimated standard errors and inferential statistics (Raudenbush & Bryk, 2002). The effects of the violation at level-1 may include distorted random effect coefficients and variance-covariance components. To examine the distribution of the errors, data analysts should consider the following options. First, one might plot the standardized residuals against their normal scores to observe how closely the plot follows a diagonal line (Hox, 2002). Second, to assess the distribution of the overall residuals at levels one and two, one might also use histograms or box-and-whisker plots. Similarly, box-and-whisker plots of the level-1 residuals for each level-2 unit and of the level-2 residuals for each level-2 effect will help assess deviations from a normal distribution and identify extreme values. Third, to assess assumptions of normality, linearity, and homogeneity of variance simultaneously, analysts should examine the plot of predicted values against the residuals. This should be done for level-1 residuals, as well as the residuals for each level-2 effect. Scatterplots that contain roughly equivalent frequencies of points above and below their mean value, with no particular structure, provide evidence that the assumptions have not been violated (Hox, 2002).

To supplement the visual level-1 assumption diagnostics, summary statistics (i.e., skewness, kurtosis, variance and standard deviation, and a statistical test for normality such as the Shapiro-Wilk or Kolmogorov-Smifnov test) for level-1 residuals within each level-2 unit should also be reviewed for information regarding level-1 normality and variance homogeneity. In addition, Levene's test can be used to assess homogeneity of variance of the level-1 residuals across each level-2 unit. Next, when evaluating the tenability of assumptions for the level-2 residuals, multivariate summary statistics such as skewness, kurtosis, and Mahalanobis distances should be generated and reviewed. Multivariate normality can also be evaluated visually through a histogram of Mahalanobis distances. And, as with the level-1 residuals, Levene's test for assessing homogeneity of variance of the level-2 residuals can also be conducted for each level-2 effect.

The MIXED_DX macro provides analysts with a comprehensive approach to assessing the degree to which one or more of the linearity, normality, or homogeneity of variance assumptions has been violated. The output includes both visual (e.g., box-and-whisker plots, histograms, scatter plots) and statistical information for both level-1 and level-2 residuals. A unique feature of our macro is the calculation of level-2 predicted values for each level-2 effect. Unlike the level-1 ODS output table in SAS (generated through the outp option) that contains both residuals and predicted values for each level-1 observation, the level-2 ODS output table in SAS (SolutionR) does not contain predicted values. By calculating predicted values for each level-2 effect, our macro allows analysts to create and examine residual by predicted scatterplots for level-2 effects (something that cannot be done using the default SAS output alone). Furthermore, the MIXED_DX macro generates output for all level-2 units, with an option for a user-defined minimum of cases per level-2 unit to be included in these analyses.

INFLUENCE

Nearly all statistical analyses involve post-hoc diagnostic investigations to determine how well models fit the data. One of the more well-established diagnostics for ordinary least-squares linear regression is the detection of influential observations in determining parameter estimates. Previous work by Cook (1977), Belsley, Kuh, and Welsch (1980), and Andrews & Pregibon (1978) provided useful statistics for determining how much influence a particular observation had over estimated parameters or the performance of the model. With the development of more complex multilevel models, the detection of influential observations becomes more complex as well. For example, mixed model analyses generate more results than may be of potential interest to the researcher. A researcher may be interested in predicting a particular outcome variable, in which case fixed effect results are of primary interest, whereas another researcher might be interested in explaining the variance in a particular outcome, thus variance component results become the primary concern. Further, with multilevel models, the concept of influence is often applied to higher-level units (i.e., group or cluster influence) vs. with OLS diagnostics where influence diagnostics are concerned with the influence of individual observations. This interest in the influence of higher-level clusters or groups calls for set-deletion analyses where an entire cluster or group is removed to determine its influence vs. the removal of individual observations as is done in OLS diagnostic algorithms.

Littell, Milliken, Stroup, and Wolfinger (2006) recommend a "drill-down" approach to mixed model influence diagnostics, beginning with a global assessment of the influence on the overall model, followed by a more detailed exploration of the case-sets should they be warranted. The likelihood distance (Cook & Weisberg, 1982) provides an assessment of a unit's influence on the overall model. A group or unit's influence on parameter estimates can be determined using Cook's *D* or the multivariate DFFITS statistics. The larger the value for these statistics, the greater the influence a unit has on parameter estimates (i.e., the change in the parameter estimate is large relative to the variability of the estimate; Schabenberger, 2004). Researchers can also examine a unit's effect of the precision of an estimate through the covariance trace (COVTRACE) and covariance ratio (COVRATIO) statistics. For these statistics, benchmark values used to determine a unit's level of influence are zero and one, respectively (Schabenberger, 2004). Each of these influence diagnostics can be generated for both fixed effects and covariance parameters, however calculation for the latter requires an iterative process due to the potential impact of observations on covariance matrices. The INFLUENCE option within PROC MIXED allows the researcher to utilize non-iterative or iterative diagnostics, and provides the option to control the number of iterations when re-calculating estimates and covariance matrices (Littell et. al., 2006).

In addition to examining a unit's influence on the change in parameter estimates and the change in the precision of estimates, influence on fitted and predicted values can also be examined through the Predicted Residual Sum of Squares (PRESS) statistic. The PRESS provides a comparison of the predicted marginal mean and the observed mean when the predicted value is calculated without the deleted observation in question (Schabenberger, 2004). Again, larger values suggest more influential units. Other measures of overall influence provided through the INFLUENCE option in PROC MIXED include RMSE and Restricted Likelihood Distance (RLD). RMSE values represent the model RMSE with a specific level-2 unit deleted. On the other hand, RLD functions more as an 'indicator' statistic in that it does not convey what part of the model is likely to change given the removal of a specific level-2 unit. Instead, it suggests that the overall influence of a particular level-2 unit stands out comparatively to other level-2 units (Littell et al., 2006). To determine what model components are influenced by a given level-2 unit, analysts need to examine the individual influence statistics such as MDFFITS or COVRATIO.

To facilitate detection of influential observations in a two-level linear model, the MIXED_DX macro creates a ranked summary table of the influence statistics automatically created in the SAS ODS influence table (Influence). More specifically, for each statistic provided in the summary table, the macro employs the RANK procedure to determine the percentile rank of each statistic's value for each level-2 unit. MIXED_DX allows the user to choose a threshold percentile rank of interest for the detection of influential observations via the 'PR = ' argument to the macro. The default, 'PR = 90', will flag any level-2 unit with a percentile rank greater than 90 for each statistic included in the ODS output. To further facilitate inspection of the summary table, particularly for data sets with a large number of level-2 units, we summed the number of flagged statistics and sorted the table in descending order, thus, placing level-2 units most frequently identified as influential at the top of the summary table. In addition, the MIXED_DX influence output is calculated at the level-2 unit (controlled by the 'EFFECT =' option in Influence command in PROC MIXED) and based on five iterations (controlled by the 'ITER =' option in the Influence command in PROC MIXED).

MACRO MIXED_DX

The MIXED_DX macro inputs consist of ODS table names generated from the MIXED procedure including (a) ModelInfo,, (b) Dimsensions, (c) SolutionF, (d) SolutionR, (e) Level1, and (f) Influence, as well as two user specified arguments, MIN_NJ and PR. MIN_NJ defines the smallest level-2 unit to be included in the within-unit box-and-whisker plots and normality assessments (allowing the exclusion of very small units from these analyses), whereas PR defines the threshold percentile rank of influence diagnostics that trigger the flagging of level-2 units. If the MIN_NJ subanalysis is not wanted, the user will need to override the default value of five by specifying some value for MIN_NJ such as MIN_NJ = 1.

The MIXED_DX macro produces numeric (i.e., output located in the output window) and visual (i.e., output located in the graph window) output for both level-1 and level-2 data.

Level-1 numeric output includes: (a) normality summary statistics table for the overall level-1 residual and for each level-2 unit, (b) normality summary statistics table for the overall level-1 residual and for each level-2 unit where n_j > min_nj, (c) PROC UNIVARIATE plot output for the variance of all level-1 residuals, (d) PROC UNIVARIATE plot output for the variance of level-1 residuals where n_j > min_nj, and (e) PROC GLM output using Levene's Test for homogeneity of variance of level-1 residuals.

Level-1 visual output includes: (a) side-by-side box-and-whisker plots for the overall level-1 residual and for each level-2 unit, (b) side-by-side box-and-whisker plots for the overall level-1 residual and for each level-2 unit where n_j > min_nj, (c) histogram of the variance of all level-1 residuals, (d) histogram of the variance of all level-1 residuals where n_j > min_nj, and (e) plot of level-1 residuals by predicted values.

Level-2 numeric output includes: (a) PROC UNIVARIATE plot output of level-2 residuals for each level-2 effect, (b) multivariate normality and outlier summary table, (c) Mahalanobis distance values for level-2 units, (d) PROC UNIVARIATE plot output output of Mahalanobis distance values for level-2 units, and (e) table of ranked influence diagnostics.

Level-2 visual output includes: (a) plot of level-2 residuals by predicted values for each level-2 effect, (b) histogram of level-2 residuals for each level-2 effect, and (c) histogram of Mahalanobis distance values.

EXAMPLE OF MACRO MIXED_DX

Below is an example of PROC MIXED code used in conjunction with the macro MIXED_DX, followed by select examples of the numeric and visual output.

```
ods exclude influence SolutionR;
title;
proc mixed data = temp covtest noclprint NAMELEN=32;
class schoolid;
model mathach = SIZE FEMALE SES MEANSES SES*MEANSES SES*SIZE FEMALE*SIZE/
solution outp=L1Resid influence(effect=schoolid iter=5);
random  intercept FEMALE SES / sub=schoolid solution type=un;
ods output SolutionR=L2Resid SolutionF=Fixed ModelInfo=ModelStuff
Dimensions=DatStuff Influence=influence; run;
%mixed_dx (ModelI = modelstuff, Dims = datstuff, solnF = Fixed, solnR =
L2Resid, Level1 = L1Resid, Influence = influence, min_nj = 5, pr = 90);
run;
```

Figure 1 contains one ‘window’ of the level-1 residual box-and-whisker plot output for the overall level-1 residual (the first plot in the series) and for each level-2 unit. These plots can be used to evaluate the normality and heterogeneity of variance assumption for the overall level-1 residual, as well as for the level-1 residuals within each level-2 unit. Figure 2 contains similar output as Figure 1, however, instead of including box-and-whisker plots for all level-2 units, the output in Figure 2 displays the overall level-1 residual and the residuals for level-2 units in which $n_j \geq 5$, where the sample size of five was specified through the MIN_NJ = 5 option in the call to the macro. As with Figure 1, the box-and-whisker plots included in Figure 2 allow an analyst to evaluate the tenability of the normality and heterogeneity of variance assumptions for level-1 residuals. However, instead of examining level-1 residuals within each level-2 unit, this part of the MIXED_DX output s limited to level-2 units with a user specified minimum number of observations. This approach allows users to focus on level-2 units with enough observations that normality and equal variances might be plausible (i.e., when level-2 units do not have many observations in them, normally distributed errors and equal variances with other, more dense level-2 units is highly improbable given the tenets of Central Limit Theorem).

The histogram displaying the distribution of the variance of level-1 residuals for all level-2 units shown in Figure 3 provides users with summary information regarding the magnitude of level-1 residual variances across level-2 units. Although this part of the output is not directly related to evaluating the heterogeneity of variance assumption for level-1 residuals, it offers supplemental information regarding the nature of the variability in level-1 residuals. Figure 4 displays the plot of level-1 residuals by predicted values for all observations included in the PROC MIXED analysis. Similarly, Figures 5 and 6 contain plots of level-2 residuals by predicted values for the intercept and the random effect for the level-1 variable ‘female’ (i.e., level-2 effect = female), respectively. Each scatterplot provides users a way to evaluate simultaneously the normality, linearity, and homoscedasticity of level-1 (Figure 4) and level-2 residuals (Figures 5 and 6).

Figure 7 contains a histogram displaying the distribution of Mahalanobis distance values for each level-2 unit which allows researchers to easily examine the range of values to determine if potentially problematic outliers are evident. Moreover, when evaluating data for possible outliers or violations of statistical assumptions, nicely summarized distributions such as the histogram in Figure 7 are particularly helpful and efficient when there are oodles of level-2 units.

Figures 8, 9, 10, and 11 contain different elements of the numerical output generated by MIXED_DX. Specifically, Figure 8 is one ‘window’ of the normality summary table for level-1 residuals, Figure 9 contains the PROC GLM out for Levene’s test of homogeneity of variance of level-1 residuals, Figure 10 includes the level-2 multivariate normality and outlier assessment table, and Figure 11 includes one ‘window’ of the ranked influence diagnostic summary table for level-2 units. The statistical output provided in the normality summary table for level-1 residuals (Figure 8) can be used to supplement the visual evaluation of the normality and homogeneity of variance assumption of level-1 residuals. Similarly, the Levene’s test of homogeneity of variance output provided in Figure 9 can also be used to supplement any visual assessment of equal variances of level-1 residuals across level-2 units. The statistical output in Figure 10 allows users a method of evaluating the multivariate normality of level-2 residuals. Also, results from the inferential examination of the largest Mahalanobis distance can be used to supplement the visual evaluation of potentially problematic outliers. By providing a nice summary of the influential nature of each level-2 unit, the ranked influence diagnostics output in Figure 11 provides users a way to easily examine which, if any, level-2 units might be influencing the magnitude or precision of fixed and/or random effects in their model.

CONCLUSIONS

The macro MIXED_DX was created to facilitate the evaluation of statistical assumptions of linear two-level multilevel models, including the examination of residual normality, linearity, and homogeneity of variance, as well as influential outliers. By utilizing data provided in various ODS tables in PROC MIXED, MIXED_DX makes it much easier for the applied data analyst to evaluate assumptions for both level-1 and level-2 residuals, through both numeric and visual assessments such as statistical summary tables, box-and-whisker plots, histograms, and plots of residuals by predicted values. Furthermore, MIXED_DX is easily modifiable. For example, users can add to the macro to invoke procedures that are not currently included, such as examining the distribution of level-2 sample sizes. Conversely, users can also easily comment out any unwanted elements of the macro. Thus, given the ease in calling the macro when estimating two-level linear models via PROC MIXED, we hope that more researchers will not only evaluate the distributional assumptions and outliers, but also report said evaluation in their published research.

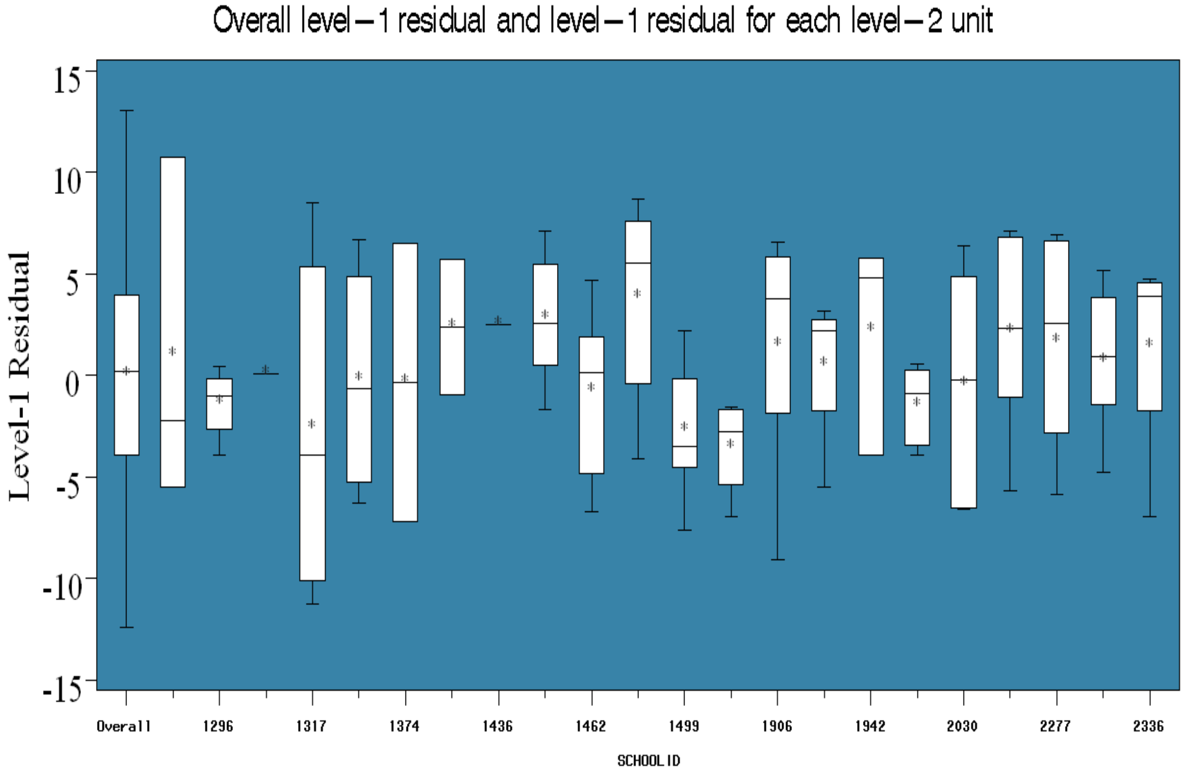

Figure 1. Distribution of level-1 residuals: Overall residual and for each level-2 unit

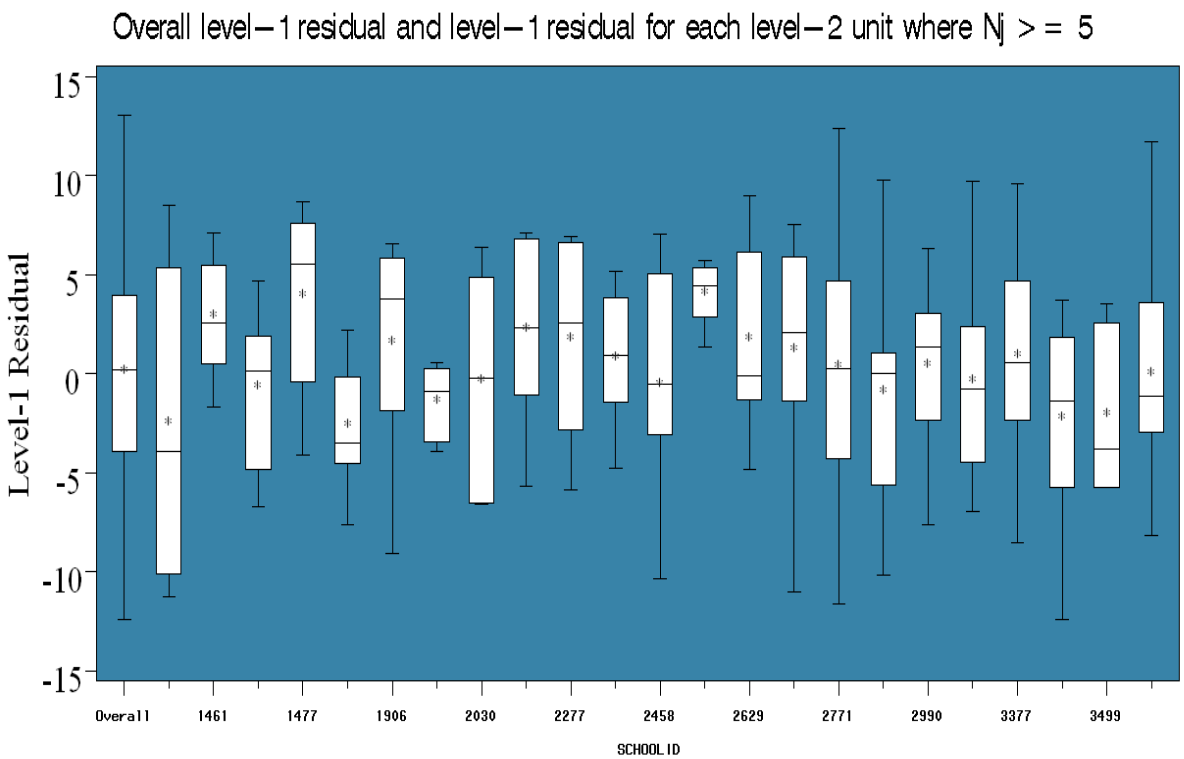

Figure 2. Distribution of level-1 residuals: Overall residual and for each level-2 unit where $\boldsymbol{n_j} \geq 5$

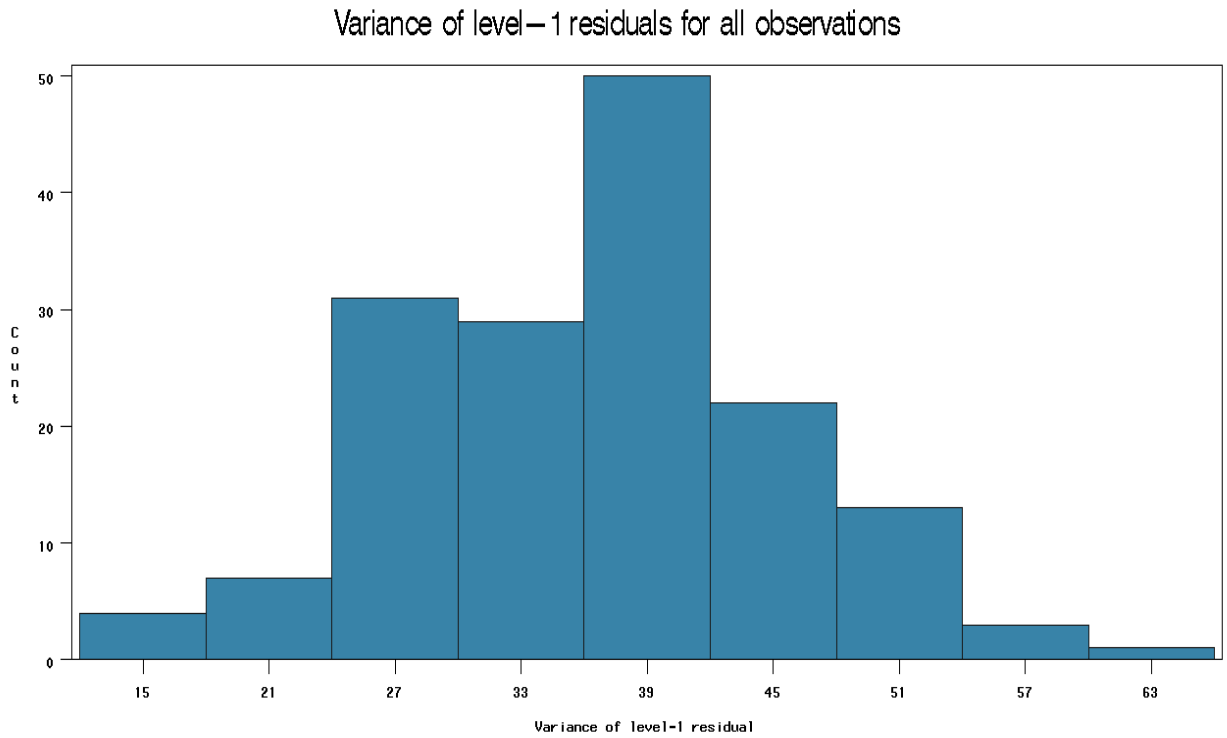

Figure 3. Distribution of the variance of level-1 residuals for all observations

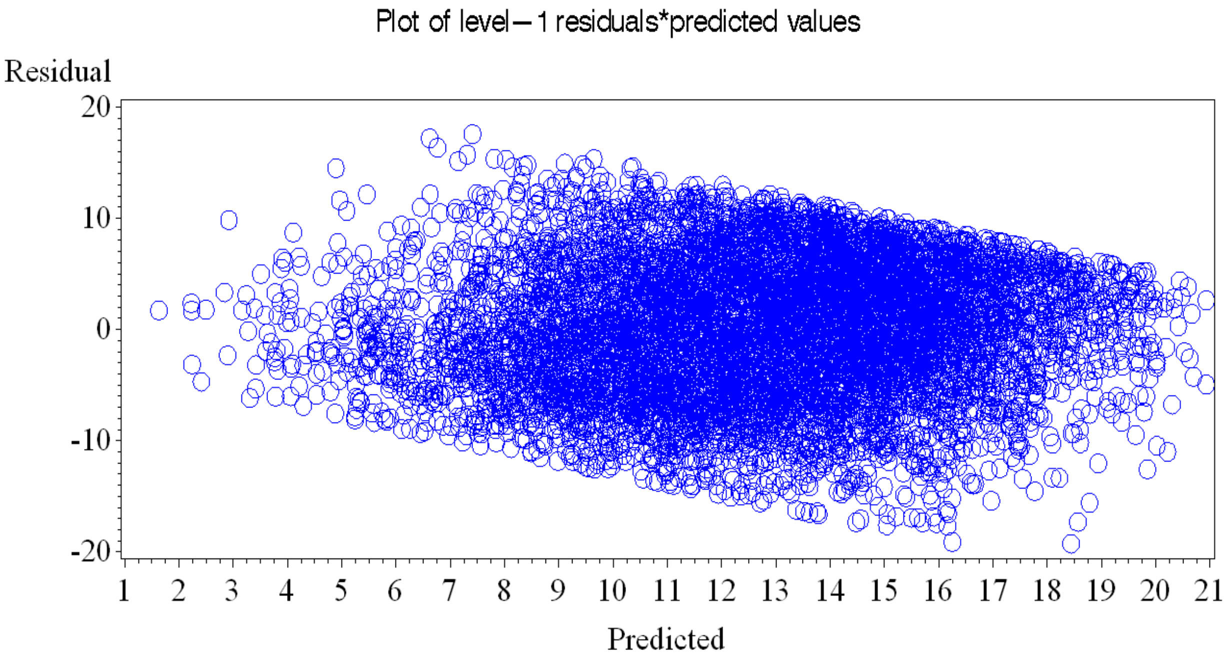

Figure 4. Level-1 residual*predicted value for all observations

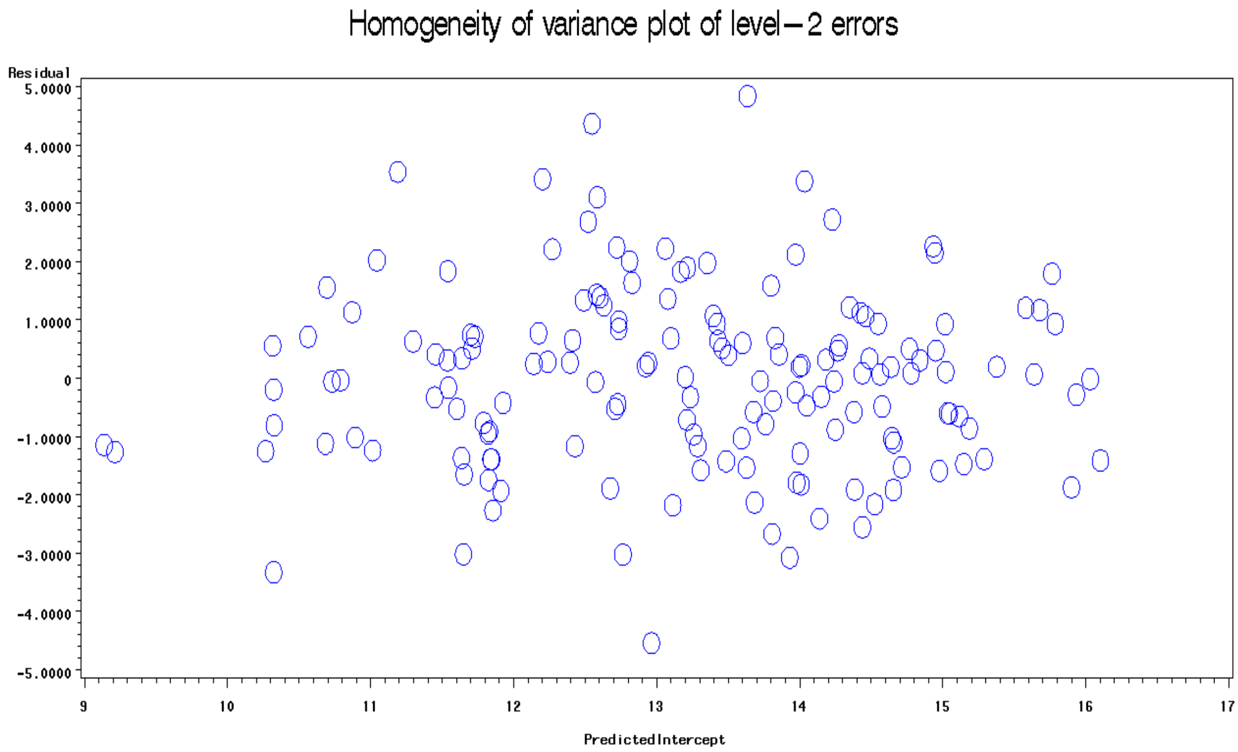

Figure 5. Level-2 residual*predicted intercept value for all level-2 units

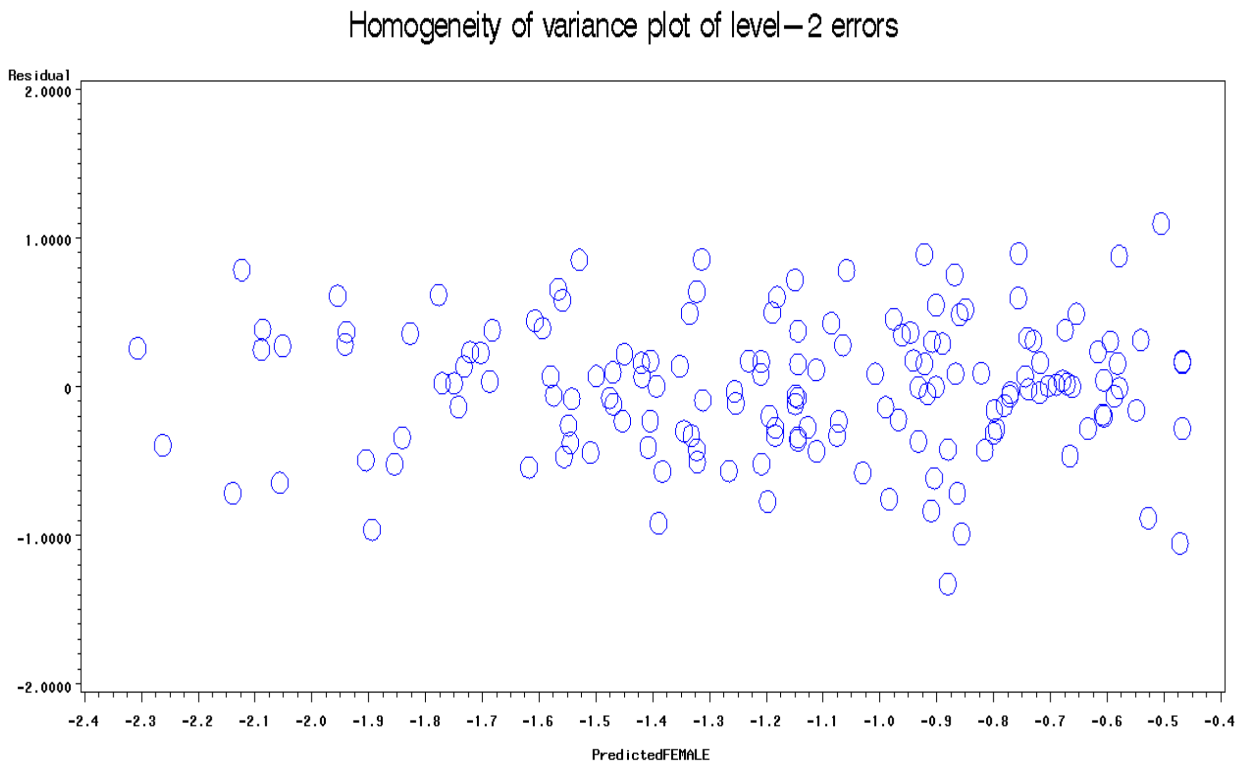

Figure 6. Level-2 residual*predicted variance component for level-1 variable ‘female’ for all level-2 units

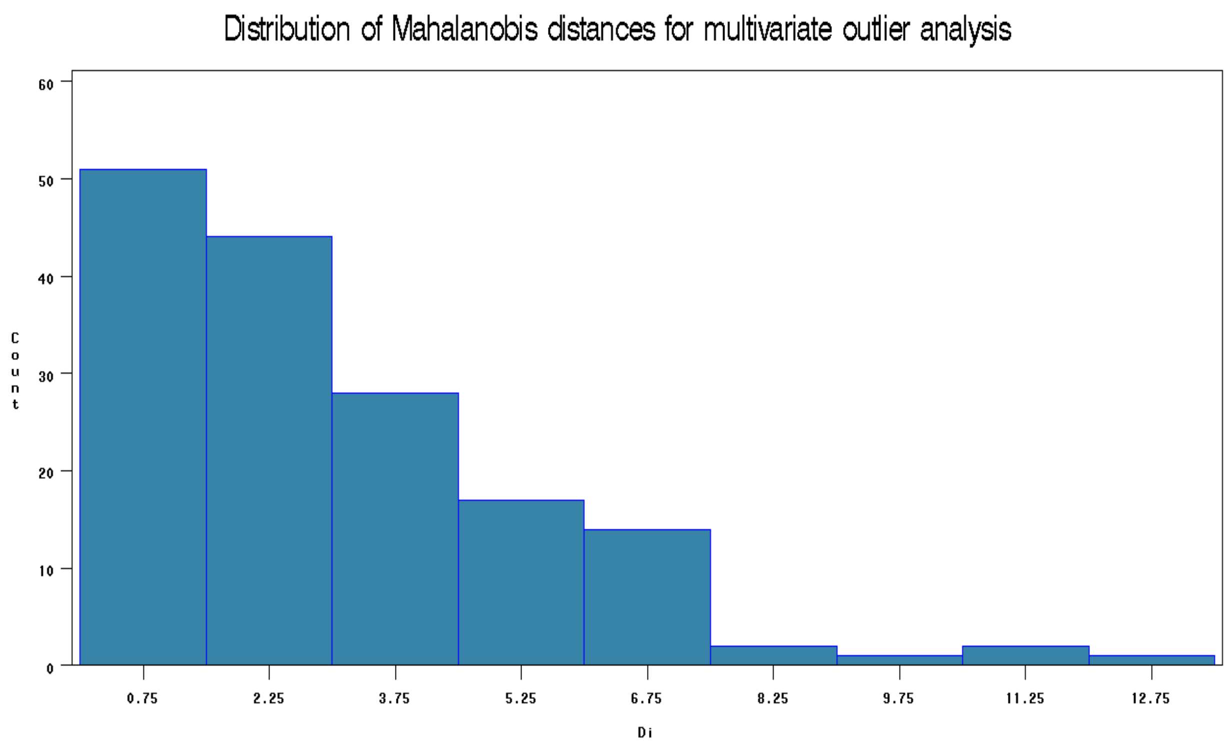

Figure 7. Distribution of Mahalanobis distance values for each level-2 unit

Normality summary statistics for the overall residual and each level-2 unit

Level-2 Unit	Skewness	Kurtosis	Shapiro-Wilk	p-value	Variance	Std Dev	N
Overall	-0.1464	-0.5220	0.0321	0.0100	35.6292	5.9690	7185
1224	0.2481	-1.1449	0.9375	0.0143	54.4537	7.3793	47
1288	-0.4288	-0.6287	0.9475	0.2206	46.0958	6.7894	25
1296	0.6568	1.3472	0.9562	0.0711	28.5501	5.3432	48
1308	-0.3487	-0.5644	0.9483	0.3423	38.2406	6.1839	20
1317	0.1423	-0.8975	0.9604	0.1052	29.4795	5.4295	48
1358	-0.4556	0.2699	0.9683	0.4938	25.4989	5.0496	30
1374	-0.3089	-0.7325	0.9599	0.3471	61.3113	7.8302	28
1433	-0.4066	-0.7341	0.9572	0.1883	13.8575	3.7226	35
1436	-0.7585	0.1641	0.9476	0.0450	17.8752	4.2279	44
1461	-1.4369	2.4512	0.8717	0.0011	34.2199	5.8498	33
1462	0.5766	-0.4901	0.9467	0.0139	41.0545	6.4074	57
1477	-0.4332	-0.7634	0.9522	0.0171	50.5179	7.1076	62
1499	0.6850	0.0731	0.9523	0.0337	32.1495	5.6701	53
1637	0.7102	-0.0295	0.9374	0.1050	45.9021	6.7751	27
1906	-0.5982	-0.1637	0.9573	0.0561	40.8894	6.3945	53
1909	-0.2498	-1.3750	0.9193	0.0333	34.4495	5.8694	28
1942	-0.9756	0.7643	0.8972	0.0084	30.7441	5.5447	29
1946	0.0692	-1.1729	0.9503	0.0843	39.2686	6.2665	39
2030	-0.0596	-0.7441	0.9765	0.4578	38.4415	6.2001	47
2208	-0.2361	-1.1965	0.9445	0.0087	33.6639	5.8021	60
2277	0.5186	-0.3404	0.9664	0.0916	33.4067	5.7799	61
2305	-0.5484	-0.0718	0.9647	0.0541	26.2418	5.1227	67
2336	-0.5344	-0.2465	0.9653	0.1742	32.5628	5.7064	47

Figure 8. Partial output from MIXED_DX normality summary table for level-1 residuals: Overall and for each level-2 unit

```
                   Levenes test of homogeneity of variance of level-1 residuals

                                     The GLM Procedure

Dependent Variable: Absolute_resid   (Absolute value of level-1 residual)

                                              Sum of
       Source                      DF        Squares     Mean Square    F Value    Pr > F

       Model                      159     3269.56971        20.56333       1.85    <.0001

       Error                     7025    77957.87122        11.09721

       Corrected Total           7184    81227.44092

                   R-Square     Coeff Var      Root MSE    Absolute_resid Mean

                   0.040252      67.55124      3.331247               4.931437

       Source                      DF       Type I SS     Mean Square    F Value    Pr > F

       SCHOOLID                   159     3269.569707       20.563331       1.85    <.0001


       Source                      DF     Type III SS     Mean Square    F Value    Pr > F

       SCHOOLID                   159     3269.569707       20.563331       1.85    <.0001
```

Figure 9. MIXED_DX output for Levene's homogeneity of variance test of level-1 residuals

```
Multivariate normality and outlier assessment
---------------------------------------------
Multivariate Skewness:
         b1p =              0.45811311
         Chi-Square =       12.5624477
         df =                       10
         p-value =          0.24917459

Multivariate Kurtosis:
         b2p =              14.7863678
         Z (Upper) =        -0.2466812
         p-value (Upper) =  0.59742254
         Z (Lower) =        -0.6775666
         p-value (Lower) =  0.24902327

Largest Mahalanobis Distance:
         Level 2 Unit =          3427
         D-Squared =         13.135085
         F =                4.71470232
         df numerator =              3
         df denominator =          156
         p-value            0.00353276
```

Figure 10. MIXED_DX output for level-2 multivariate normality and outlier assessment

Ranked influence diagnostics

Obs	SCHOOLID	flags	Nobs	Iter	PRESS	MDFFITS	CovRatio	CovTrace	CookDCP	MDFFITSCP	CovRatioCP	CovTraceCP	RMSE	RLD
1	3427	8	49	2	2406.76	0.01810	0.8530	0.1512	0.53129	0.60558	0.7554	0.2546	6.05025	0.7456
2	8367	8	14	2	1244.06	0.02893	0.8735	0.1305	0.25961	0.30125	0.7396	0.2786	6.03687	0.4877
3	2277	7	61	2	2273.07	0.04285	0.9057	0.0948	0.25883	0.28665	0.7984	0.2094	6.04184	0.6579
4	2305	7	67	2	2192.31	0.02412	0.9872	0.0110	0.12724	0.12935	0.9405	0.0570	6.04602	0.3455
5	6074	7	56	2	2990.11	0.02145	0.9742	0.0239	0.12442	0.13011	0.9222	0.0756	6.03473	0.3118
6	2655	6	52	2	2623.47	0.04934	1.0959	0.0982	0.13220	0.14074	0.9957	0.0044	6.03880	0.5516
7	6990	6	53	2	2177.66	0.02582	0.9600	0.0364	0.18515	0.20090	0.9699	0.0234	6.04228	0.3824
8	3533	5	48	2	2500.44	0.01383	0.9214	0.0779	0.29336	0.32961	0.8320	0.1706	6.03095	0.4105
9	3716	5	41	2	2114.35	0.03330	1.0337	0.0361	0.12585	0.13249	0.9420	0.0555	6.03160	0.4154
10	8628	5	61	2	3382.65	0.01206	0.9408	0.0570	0.21589	0.22709	0.9976	0.0072	6.03865	0.3254
11	2526	4	57	2	1873.14	0.01222	0.9795	0.0192	0.09784	0.10099	0.9773	0.0187	6.04531	0.1929
12	2639	4	42	2	1574.98	0.03415	1.2044	0.1955	0.03084	0.03170	1.0291	0.0310	6.03686	0.3174
13	7345	4	56	2	2558.10	0.00704	0.9630	0.0353	0.15047	0.15451	1.0219	0.0252	6.03364	0.2177
14	7688	4	54	2	2021.59	0.00813	0.9567	0.0428	0.12721	0.13235	0.9245	0.0734	6.04651	0.2013
15	8193	4	43	2	1573.01	0.01828	0.9898	0.0095	0.09397	0.09611	0.9133	0.0874	6.04648	0.2618
16	8627	4	53	2	2543.96	0.01852	1.1416	0.1364	0.03256	0.03279	1.0667	0.0658	6.03346	0.1939

Figure 11. Partial output from MIXED_DX ranked influence diagnostics summary table for each level-2 unit

REFERENCES

Andrews, D.F. & Pregibon, D. (1978). Finding the outliers that matter. *Journal of the Royal Statistical Society, Series B, 40,* 85-93.

Belsley, D.A., Kuh, E., & Welsch, R.E. (1980). *Regression diagnostics: Identifying influential data and sources of collinearity.* New York: Wiley.

Cook, R.D. (1977). Detection of influential observations in linear regression. *Technometrics, 19,* 15-18.

Cook, R.D. & Weisberg, S. (1982). *Residuals and influence in regression.* New York: Chapman and Hall.

Ferron, J.M., Hogarty, K.Y., Dedrick, R.F., Hess, M.R., Niles, J.D., & Kromrey, J.D. (2008). Reporting results from multilevel analyses. In A.A. O'Connell & D. B. McCoach (Eds.), *Multilevel modeling of educational data* (pp. 391-426). Charlotte, NC: Information Age Publishing, Inc.

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

Littell, R. C., Milliken, G. A., Stroup, W. W., & Wolfinger, R., D. (2006). *SAS for mixed models* (2nd ed.). Cary, NC: SAS Institute Inc.

Longford, N. (2001). Simulation-based diagnostics in random-coefficient models. *Journal of the Royal Statistical Society Series A-Statistics in Society, 164,* 259-273.

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

Schabenberger, O. (2004). *Mixed model influence diagnostics*. *Proceedings of the 29th Annual SAS Users Group International Conference*.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Bethany A. Bell
University of South Carolina
College of Education
Wardlaw 133
Columbia, SC 29208
Work Phone: 803-777-2387
E-mail:babell@sc.edu
Web: http://www.ed.sc.edu/bell/

MACRO MIXED_DX

```
%macro MIXED_DX (ModelI = modelstuff, Dims = datstuff, solnF = Fixed, solnR =
L2Resid, Level1 = L1Resid, Influence = influence, min_nj = 5, pr = 90);
/*
* +------------------------------------------------------------------------------------------------------+
Important information about the types of models that are supported by
macro MIXED_DX:
* +------------------------------------------------------------------------------------------------------+;
a. MIXED_DX has been tested for use in SAS 9.1.3 and 9.2
b. MIXED_DX utilizes the following SAS components: Base SAS, SAS/STAT,
SAS/GRAPH ,and SAS/IML
c. MIXED_DX works for 2-level linear multilevel models estimated in PROC MIXED
d. MIXED_DX works correctly for models that have converged and have positive
G-matrices
e. MIXED_DX is designed to read model effect names up to 32 characters in length. If a
user has a n effect that is longer than 32 characters, the 'NAMELEN= ' option in
PROC MIXED must be updated accordingly.
f. If a user's model contains any interactions, they must be created on the model
statement (e.g., model y=x1 x2 w1 w2 x1*w1)
g. Users cannot use the class statement to create dummy variables in PROC MIXED.
That is, any dummy variables used in the model statement need to be created prior
to the MIXED procedure (e.g., in a data step).
h. Users need to specify their desired value for MIN_NJ, which defines the smallest
level-2 unit to be included in the within-unit box-and-whisker plots and normality
assessments. The default is set to 5.Even if this subanalysis is not wanted, the
user must override the default value of 5 and specify some value for MIN_NJ such
as min_nj = 1.
i. Users need to specify their desired value for PR, which defines the threshold
percentile rank of influence diagnostics that trigger the flagging of level-2 units.
The default is set to 90.

As you use the macro, please feel free to send feedback, comments, and/or suggestions
to, Bethany A. Bell, at babell@sc.edu.
The complete MIXED_DX SAS program file is available for download from
http://www.ed.sc.edu/bell/
*/
* +------------------------------------------------------------------------------------------------------+
  Define the variable &SUBJ whose value is the clustering variable name from PROC
MIXED
* +------------------------------------------------------------------------------------------------------+;
%global subj;
data getsubj; set &ModelI; if Descr = 'Subject Effect'; call symput('SUBJ',Value);
* +------------------------------------------------------------------------------------------------------+
  Obtain level-1 residual means, variances, and sample sizes for each unit
* +------------------------------------------------------------------------------------------------------+;
proc sort data = &Level1; by &subj;
proc means noprint data = &Level1; by &subj;
        var Resid;
        output out = L1Summary mean = MN_Resid Var = Var_Resid N = Nj;
data L1Summary; set L1summary; label Var_Resid = 'Variance of level-1 residual';
* +------------------------------------------------------------------------------------------------------+
   Obtain level-1 residual means, variances, and sample sizes for each unit
   where Nj >= &min_nj
* +------------------------------------------------------------------------------------------------------+;
data L1Summary2; set L1Summary; if Nj GE &min_nj;
* +------------------------------------------------------------------------------------------------------+
   Organizing the data to generate box-plots for the overall level-1 residual and for
   level-2 units where Nj >= &min_nj
* +------------------------------------------------------------------------------------------------------+;
**Applying format to subject variable to denote Overall Residual;
proc format; value subjlbl    0 = 'Overall'; value $ subjlbl '0' = 'Overall';
data L1ResidAll; set &Level1; &subj = 0;
data box; set &Level1 L1ResidAll;
        &subj=trim(left(&subj));
        format &subj subjlbl.;
proc sort data = box; by &subj;
proc sort data = L1summary; by &subj;
data box2; merge box L1summary; by &subj;
        if &subj = 0 or Nj ge &min_nj;
        format &subj subjlbl.;
proc boxplot data=box;
        plot resid*&subj/vaxis = axis1 boxstyle = schematic cframe = steel
        cboxes = black
        cboxfill = white idsymbol = circle idcolor = bbl boxwidth = 2;
        symbol v = * color = bbl f='times new roman' h=1.5;
        axis1 label=(f='times new roman' h=2.5 a=90 r=0 'Level-1 Residual')
        value=(f='times new roman' h=2.5);
        title 'Overall level-1 residual and level-1 residual for each level-2 unit';
proc boxplot data=box2;
        plot resid*&subj/vaxis = axis1 boxstyle = schematic cframe = steel
        cboxes = black
        cboxfill = white idsymbol = circle idcolor = bbl boxwidth = 2;
        symbol v = * color = bbl f='times new roman' h=1.5;
        axis1 label=(f='times new roman' h=2.5 a=90 r=0 'Level-1 Residual')
        value=(f='times new roman' h=2.5);
        title "Overall level-1 residual and level-1 residual for each level-2 unit
        where Nj >= &min_nj";
* +------------------------------------------------------------------------------------------------------+
  Creating a normality summary statistics table for the overall level-1 residual and
  each level-2 unit
* +------------------------------------------------------------------------------------------------------+;
proc univariate data = box normal noprint; var resid; by &subj;
output  out=table   n=nT  skewness=skew  kurtosis=ku  normal=SW  var=varT  std=stdT
probn=pvalue;
data normal_table;
        set table;
        file print header=h notitles;
 put @1 &subj @25 skew 9.4 @35 ku 9.4 @45 SW 10.4 @60 pvalue 6.4 @70 varT 9.4
@80 stdT 9.4 @90 nT 5.0;
return;
H: put @1 'Normality summary statistics for the overall residual and each level-2 unit'//
```

```
@1 'Level-2 Unit' @26 'Skewness' @36 'Kurtosis' @47 'Shapiro-Wilk' @60 'p-value'
@71 'Variance' @82 'Std Dev' @93 'N'/;
* +--------------------------------------------------------------------------------------------------------+
   Creating a normality summary statistics table for the overall level-1 residual and
   each level-2 unit where Nj >= &min_nj
* +--------------------------------------------------------------------------------------------------------+;
proc univariate data = box2 normal noprint; var resid; by &subj;
output out=table2  n=nT skewness=skew kurtosis=ku normal=SW var=varT std=stdT
probn=pvalue;
data normal_table2;
        set table2;
        file print header=h notitles;
put @1 &subj  @25 skew 9.4 @35 ku 9.4 @45 SW 10.4 @60 pvalue 6.4 @70 varT 9.4
@80 stdT 9.4 @90 nT 5.0;
return;
H: put @1 "Normality summary statistics for the overall residual and each level-2 unit
 where Nj >= &min_nj"//
@1 'Level-2 Unit' @26 'Skewness' @36 'Kurtosis' @47 'Shapiro-Wilk' @60 'p-value'
@71 'Variance' @82 'Std Dev' @93 'N'/;
* +--------------------------------------------------------------------------------------------------------+
   Histograms for overall level-1 residual and for each level-2 unit
* +--------------------------------------------------------------------------------------------------------+;
proc univariate data = L1summary normal plot; var var_resid; id &subj;
        histogram var_resid /vscale=count cfill=steel;
title 'Variance of level-1 residuals for all observations';
*+--------------------------------------------------------------------------------------------------------+
Histograms for overall level-1 residual and for level-2 units where Nj >= &min_nj
* +--------------------------------------------------------------------------------------------------------+;
proc univariate data = L1summary2 normal plot; var var_resid; id &subj;
        histogram var_resid /vscale=count cfill=steel;
title "Variance of level-1 residuals for each level-2 unit where Nj >= &min_nj";
* +--------------------------------------------------------------------------------------------------------+
   Obtaining the absolute value of level-1 residuals and then running a one-way
   ANOVA to test for equal variances, which is analogous to Levenes Test
* +--------------------------------------------------------------------------------------------------------+;
data Levene; set &Level1;
        Absolute_resid = abs(resid);
        label Absolute_resid = '(Absolute value of level-1 residual)';
        title 'Levenes test of homogeneity of variance of level-1 residuals';
proc glm data=levene; class &subj; model Absolute_resid=&subj;
* +--------------------------------------------------------------------------------------------------------+
   Assessment of level-1 residual normality and equal variance through the scatterplot
* +--------------------------------------------------------------------------------------------------------+;
symbol1 font =, color=blue interpol=none height=2.5 v=circle;
proc gplot data = &Level1;
        plot resid*pred /vaxis = axis1 haxis =axis2;
        title1 'Plot of level-1 residuals*predicted values';
        axis1 label=(f='times new roman' h=2.5 'Residual')
                value=(f='times new roman' h=2.5);
                 axis2 label=(f='times new roman' h=2.5 'Predicted')
                 value=(f='times new roman' h=2.5);
```

```
* +--------------------------------------------------------------------------------------------------------+
   Define the variable &dat whose value is the dataset name from PROC MIXED
* +--------------------------------------------------------------------------------------------------------+;
%global dat;
data getdat; set &ModelI; if Descr = 'Data Set'; call symput('DAT',Value);
run;

* +--------------------------------------------------------------------------------------------------------+
Define the variable &n2 whose value is the number of level-2 units from PROC MIXED
* +--------------------------------------------------------------------------------------------------------+;
%global n2;
data getn2; set &Dims; if Descr = 'Subjects'; call symput('n2',Value);
run;
* +--------------------------------------------------------------------------------------------------------+
   Make a little data set containing names of the main effects used in PROC MIXED
* +--------------------------------------------------------------------------------------------------------+;
data preds; set &solnF; keep effect; if effect='Intercept' then delete;
        if index(effect, '*')>0 then delete;
* +--------------------------------------------------------------------------------------------------------+
   Count the number of main effects used in PROC MIXED
* +--------------------------------------------------------------------------------------------------------+;
data _null_; set preds end=lastrec;
        if lastrec then do;
                call symput('totalvars',_n_);
        end;
run;
* +--------------------------------------------------------------------------------------------------------+
   Make a macro variable to calculate predicted values (pred1, pred2, etc.) for each
   main effect in PROC MIXED
* +--------------------------------------------------------------------------------------------------------+;
%do nn=1 %to &totalvars;
        %global pred&nn;
        data _null_; set preds; if &nn=_n_; call symput("pred&nn",effect); run;
%end;
* +--------------------------------------------------------------------------------------------------------+
   Make a mini-macro to generate the list of main effects in PROC MIXED
* +--------------------------------------------------------------------------------------------------------+;
%macro create_predlist;
        %do k=1 %to &totalvars;
                &&pred&k
        %end;
%mend create_predlist;
* +--------------------------------------------------------------------------------------------------------+
   Compute means and variances of each main effect variable for each level-2 unit
* +--------------------------------------------------------------------------------------------------------+;
proc sort data = &dat; by &subj;
proc means data = &dat nway noprint; class &subj;
        var %create_predlist;
        output out=m mean=;
        output out=s std=;
```

```
* +---------------------------------------------------------------------------------------------------+
   Compute the sum of the within unit stds for each main effect variable
* +---------------------------------------------------------------------------------------------------+;
proc means data=s nway noprint;
            var %create_predlist;
            output out=ss sum=;
* +---------------------------------------------------------------------------------------------------+
   Make data set containing names of the level-2 variables
            (those where within unit std = 0)
* +---------------------------------------------------------------------------------------------------+;
proc transpose data=ss out=sss;
data sss2; set sss; if COL1=0 and _NAME_ NE '_TYPE_';
data sss3; set sss2; rename _NAME_ = Effect;
data sss4; set sss3; lev2=1; keep effect lev2;
* +---------------------------------------------------------------------------------------------------+
    Count the number of level-1 predictors with random slopes
* +---------------------------------------------------------------------------------------------------+;
data _null_; set &SolnR end=lastrec;
            if lastrec then do;
                        call symput('totalrand',_n_);
                        call symput('numlev1',_n_/&n2);
            end;
run;

* +---------------------------------------------------------------------------------------------------+
  Make data set containing names of the level-1 variables with random slopes
* +---------------------------------------------------------------------------------------------------+;
data lev11; set &solnR (FIRSTOBS=1 OBS=&numlev1);
data lev12; set lev11; keep effect;
data lev13; set lev12;
            effecttxt=cats("'",effect,"'");
            lev1r=1;
            inter1=cats("*",effect); inter2=cats(effect,"*");
            lenint1=length(inter1); lenint2=length(inter2); lenint3=lenint1+1;
* +---------------------------------------------------------------------------------------------------+
  Make a macro variable (rand1, rand2, etc.) for each random effect in PROC MIXED
* +---------------------------------------------------------------------------------------------------+;
%do mm=1 %to &numlev1;
            %global rand&mm;
            %global int1&mm;
            %global int2&mm;
            %global len1&mm;
            %global len2&mm;
            %global len3&mm;
            data _null_;
                        set lev13;
                        if &mm=_n_;
                        call symput("rand&mm",Effecttxt);
                        call symput("int1&mm",inter1);
                        call symput("int2&mm",inter2);
                        call symput("len1&mm",lenint1);
                        call symput("len2&mm",lenint2);
                        call symput("len3&mm",lenint3);
                        run;
%end;
proc sort data = m; by &subj;
* +---------------------------------------------------------------------------------------------------+
 For each random effect: Make data set containing level-2 variables, level-2 errors, and
predicted values
* +---------------------------------------------------------------------------------------------------+;
%do ii=1 %to &numlev1;
* +---------------------------------------------------------------------------------------------------+
  For each random effect: Make a prediction equation, compute predicted values and
plot the level-2 residuals with the predicted values
* +---------------------------------------------------------------------------------------------------+;
            %if &ii=1 %then %do;
                        data eq1; set &solnF;
                        data eq2; set sss4;
                        proc sort data=eq1; by effect;
                        proc sort data=eq2; by effect;
                        data eq4; merge eq1 eq2; by effect; if Effect=&&rand&ii or lev2=1;
                        data eq5; set eq4;
                                    term=cats("+",estimate,"*",effect);
                                    if effect='Intercept' then term=cats("1*",estimate);
data _null_; set eq5 end=lastrec; if lastrec then do;
            call symput('totalterms',_n_);
                        end;
                        run;
            %do z=1 %to &totalterms;
            %global term&z;
data _null_; set eq5; if &z=_n_; call symput("term&z",term); run;
            %end;
%macro create_eqs;
            %do w=1 %to &totalterms;
             &&term&w
            %end;
%mend create_eqs;
data r1; set &solnR; if Effect=&&rand&ii;
proc sort data=r1; by &subj;
data lev2dat; merge r1 m; by &subj; rename estimate = Residual;
data lev2datp; set lev2dat; Predicted=%create_eqs;
label predicted=Predicted&&rand&ii;
*+---------------------------------------------------------------------------------------------------+
Assessment of level-2 residual normality and equal variance through scatterplots
   for each level-2 effect
* +---------------------------------------------------------------------------------------------------+;
* +---------------------------------------------------------------------------------------------------+
   If no variance in the predicted values, do not produce a scatter plot
* +---------------------------------------------------------------------------------------------------+;
proc means noprint data = lev2datp;
            var predicted;
            output out = check_var var = predicted_variance;
data back; if _n_ = 1 then set check_var; set lev2datp; retain predicted_variance;
            if predicted_variance = 0 then do;
```

```
                    predicted = .;   residual = .;
          end;
proc gplot data=back;
          plot residual*predicted;
          title "Homogeneity of variance plot of level-2 errors";
%end;
%if &ii>1 %then %do;
data eq1; set &solnF; keep effect estimate;
data eq2; set sss4;
data eq3; set lev13; keep effect lev1r;
proc sort data=eq1; by effect;
proc sort data=eq2; by effect;
proc sort data=eq3; by effect;
data eq3; merge eq1 eq2 eq3; by effect;
data eq4; set eq3;
          inter=cats(&&rand&ii,"*");
          inter2=cats("*",&&rand&ii);
          length effect2 $&&len1&ii;
          effect2=effect;
          length effect5 $&&len2&ii;
          length effect6 $&&len2&ii;
          effect3=reverse(effect);
          effect4=cats(effect3);
          effect5=effect4;
          effect6=reverse(effect5);
data eq5; set eq4; if effect=&&rand&ii or inter=effect2 or inter2=effect6;
data eq6; set eq5;
if index(effect, '*')>0 then do;
          if inter=effect2 then do;
                    effect7=substr(effect, &&len3&ii);
          end;
          if inter2=effect6 then do;
                    effect8=cats(reverse(effect));
                    effect9=substr(effect8, &&len3&ii);
                    effect7=cats(reverse(effect9));
          end;
end;
term=cats("+",estimate,"*",effect7);
if effect=&&rand&ii then term=cats("1*",estimate);
run;
data _null_; set eq6 end=lastrec;
          if lastrec then do;
          call sympût('totalterms',_n_);
          end;
run;
%do z=1 %to &totalterms;
          %global term&z;
          data _null_; set eq6; if &z=_n_; call symput("term&z",term); run;
%end;
%macro create_eqs;
 %do w=1 %to &totalterms;
          &&term&w
 %end;
%mend create_eqs;
          data r1; set &solnR; if Effect=&&rand&ii;
proc sort data=r1; by &subj;
data lev2dat; merge r1 m; by &subj;
          rename estimate = Residual;
data lev2datp; set lev2dat; Predicted=%create_eqs;
          label predicted=Predicted&&rand&ii;
* +--------------------------------------------------------------------------------------------+
  If no variance in the predicted values, do not produce a scatter plot
* +--------------------------------------------------------------------------------------------+;
proc means noprint data = lev2datp;
          var predicted;
          output out = check_var var = predicted_variance;
data back; if _n_ = 1 then set check_var; set lev2datp; retain predicted_variance;
          if predicted_variance = 0 then do;
                    predicted = .;  residual = .;
          end;
proc gplot data=back;
          plot residual*predicted;
          title "Homogeneity of variance plot of level-2 errors";
%end;
%end;
* +--------------------------------------------------------------------------------------------+
   Assessment of univariate and multivariate normality for level-2 residuals
* +--------------------------------------------------------------------------------------------+;
proc sort data = &solnR; by effect &subj;
proc univariate data = &solnR normal plot; by effect;
          var estimate; id &subj;
          histogram estimate /vscale=count cfill=steel;
          title 'Distribution of level-2 residuals';
proc means noprint data = &solnR; by effect; var estimate;
          output out = geteffects mean = dummyvar;
title ;
data geteffects; set geteffects; keep effect;
proc iml;
          use geteffects;
          read all var {effect} into effects;
          use &solnR;
          read all var {estimate} into allresids;
          read all var {&SUBJ} into allsubj;
          N_obs = NROW(allresids)/NROW(effects);
          y1 = J(N_obs,NROW(effects),0);
          startrow = 1; endrow = N_obs;
          do col = 1 to NROW(effects);
                    y1[,col] = allresids[startrow:endrow,1];
                    startrow = endrow+1;
                    endrow = endrow + N_Obs;
          end;
          subjlist = allsubj[1:N_Obs,1];
          n1=nrow(y1); p1=ncol(y1);
          ybar1 = 1/n1 * y1` * J(N1,1);
```

```
y1barm = repeat(ybar1,1,n1);
s1 = 1/(n1-1)*y1`*(i(n1)-1/n1*j(n1))*y1;
d1 = (y1`-y1barm)`*inv(s1)*(y1`-y1barm);
D1i2 = vecdiag(d1);
sighat1 = 1/n1*y1`*(i(n1)-1/n1*j(n1))*y1;
g1 = (y1`-y1barm)`*inv(sighat1)*(y1`-y1barm);
b1p = SUM(g1##3)/n1##2;
b2p = TRACE(g1##2)/n1;
chiskew = (((p1+1)*(n1+1)*(n1+3))/(6*((n1+1)*(p1+1)-6)))*b1p;
dfskew = (1/6)*p1*(p1+1)*(p1+2);
pskew = 1 - probchi(chiskew,dfskew);
zupper = (b2p - p1*(p1+2))/sqrt(8*p1*(p1+2)/n1);
zlower = (b2p - (p1*(p1+2)*(n1+p1+1))/n1)/sqrt(8*p1*(p1+2)/(n1-1));
pupper = 1 - probnorm(zupper);
plower = probnorm(zlower);
d1i2max=d1i2[1,1];
d1maxsubj = subjlist[1,1];
do i = 1 to n1;
        if d1i2[i,1]>=d1i2max then do;
                d1i2max=d1i2[i,1];
                d1maxsubj = subjlist[i,1];
        end;
end;
Fout = ((n1-p1-1)/p1)*((1/(1-(n1*d1i2max/(n1-1)**2)))-1);
df1=p1;
df2=n1-p1-1;
pout= 1 - probf(Fout,df1,df2);
file print;
        put     @1 'Multivariate normality and outlier assessment' /
                @1 '----------------------------------------------' /
                @1 'Multivariate Skewness:' /
                @10 'b1p =' @30 b1p best10. /
                @10 'Chi-Square =' @30 chiskew best10. /
                @10 'df = ' @30 dfskew best10. /
                @10 'p-value =' @30 pskew best10. //
                @1 'Multivariate Kurtosis: ' /
                @10 'b2p =' @30 b2p best10. /
                @10 'Z (Upper) = ' @30 Zupper best10. /
                @10 'p-value (Upper) = ' @30 pupper best10. /
                @10 'Z (Lower) = ' @30 Zlower best10. /
                @10 'p-value (Lower) = ' @30 plower best10. //
                @1 'Largest Mahalanobis Distance:' /
                @10 'Level 2 Unit =' @30 d1maxsubj /
                @10 'D-Squared = ' @30 d1i2max best10. /
                @10 'F = ' @30 fout best10. /
                @10 'df numerator = ' @30 df1 best10. /
                @10 'df denominator =' @30 df2 best10. /
                @10 'p-value' @30 pout best10.;
create v1 from D1i2 (|COLNAME={'Di'}|);
append from D1i2;
create v2 from subjlist (|COLNAME={'Level_2_Unit'}|);
append from subjlist;
```

```
        data v3; merge v1 v2;
        proc sort data = v3; by descending Di;
        proc print noobs; var Level_2_Unit Di;
                title 'Mahalanobis distance values for level-2 units';
proc univariate data = v3 normal plot;  var Di; id Level_2_Unit;
        histogram Di /vscale=count cfill=steel;
        title 'Distribution of Mahalanobis distances for multivariate outlier analysis';
* +-----------------------------------------------------------------------------------------------+
  Code to create a ranked summary table of the oodles of influence diagnostics
        created in SAS
* +-----------------------------------------------------------------------------------------------+;
proc rank data=&influence groups=100 out=rinfluence;
        var press CookD MDFFITS COVRATIO COVTRACE CookDCP MDFFITSCP
        COVRATIOCP COVTRACECP  RMSE RLD;
        ranks rpress rCookD rMDFFITS rCOVRATIO rCOVTRACE rCookDCP
        rMDFFITSCP rCOVRATIOCP
        rCOVTRACECP rRMSE rRLD;
**create flags to count 'influential' markers for sorting;
data rinfluence; set rinfluence;
        if (rpress gt &pr) then press_flag = 1;
        if (rCookD gt &pr) then CookD_flag = 1;
        if (rMDFFITS gt &pr) then MDFFITS_flag = 1;
        if (rCOVRATIO lt (100-&pr)) then COVRATIO_flag = 1;
        if (rCOVTRACE gt &pr) then COVTRACE_flag = 1;
        if (rCookDCP gt &pr) then CookDCP_flag = 1;
        if (rMDFFITSCP gt &pr) then MDFFITSCP_flag = 1;
        if (rCOVRATIOCP gt &pr) then COVRATIOCP_flag = 1;
        if (rCOVTRACECP gt &pr) then COVTRACECP_flag = 1;
        if (rRMSE gt &pr) then RMSE_flag = 1;
        if (rRLD gt &pr) then RLD_flag = 1;
        array flag(1:11) press_flag CookD_flag MDFFITS_flag COVRATIO_flag
COVTRACE_flag CookDCP_flag MDFFITSCP_flag COVRATIOCP_flag
COVTRACECP_flag RMSE_flag RLD_flag;
        flags=0;
        do i=1 to 11;
                if flag(i)=1 then flags=flags+1;
        end;
proc sort data = rinfluence; by descending flags;
proc print data = rinfluence;
        var  &subj flags nobs iter cookd mdffits covratio covtrace CookDCP
MDFFITSCP COVRATIOCP COVTRACECP PRESS RMSE RLD;
        title 'Ranked influence diagnostics';
        run;
%mend MIXED_DX;
* +-----------------------------------------------------------------------------------------------+
Example of how users need to set up their PROC MIXED code for use with the
MIXED_DX macro. Specifically, the MIXED_DX macro is set up to read output
generated from 2-level models using the specific options and output data files as
specified below. Users only need to update the code where indicated by
'USER……..'.
* +-----------------------------------------------------------------------------------------------+;
```

```
ods exclude SolutionR influence;
title;
proc mixed data = USER ADD DATA NAME covtest noclprint NAMELEN=32;
class USER ADD CLUSTER VARIABLE NAME;
model USER SPECIFY MODEL HERE/ solution outp=L1Resid
        influence(effect=USER ADD CLUSTER VARIABLE NAME iter=5);
random  USER SPECIFY MODEL HERE / sub=USER ADD CLUSTER VARIABLE
NAME solution  type=USER SPECIFY VARIANCE STRUCTURE HERE;
ods   output   SolutionR=L2Resid   SolutionF=Fixed   ModelInfo=ModelStuff
Dimensions=DatStuff Influence=influence;
run;
%mixed_dx (ModelI = modelstuff, Dims = datstuff, solnF = Fixed, solnR = L2Resid,
Level1 = L1Resid, Influence = influence,  min_nj = 5, pr = 90);
run;
```