Journal of Modern Mathematics and Statistics 5 (4): 96-101, 2011
ISSN: 1994-5388

Exploring Response Variable Distributional Consequences Following Alterations in the Level of a Multilevel Model

Imande Michael Tyolumun
Department of Mathematics and Computer Science, Benue State University,
P.M.B. 102119, Makurdi, Nigeria

Abstract: Distributional assumptions made in respect of the variables and parameters of single or multi-level models often include those relating to the response or explanatory variables. It is anticipated that erroneous and misleading conclusions could be drawn on the distribution of the response variable if differing multi-level models predict the response variable values. This study explores, using educational data sets for illustrative analysis, distributional consequences that could be ascribed to the response variable of a multi-level model if the level of a model varies. It is shown how response variable population distribution parameter estimates can vary with varying levels in the K-level Model and also how inferences on confidence interval limits of the response variable can also be misleading in an inappropriate K-level Multi-level Modeling Framework.

Key words: Multi-level data, multi-level model, distribution, distribution parameters, response, population, Nigeria

INTRODUCTION

In fixed or mixed effects single level models such as Linear Models, Generalized Linear Models, Incomplete Block Designs and Split-plot Designs, assumptions or specifications are often given in respect of the distributions of the associated response variables. These response variable distributional assumptions or specifications often constitute the basis for the application of appropriate parameter estimation techniques or valid statistical tests of significance.

There is an extensive wealth of literature discussing the import and implications that emanate from response variable distributional assumptions or specifications in single level fixed effects models (Johnston and DiNardo, 1997; McCullagh and Nelder, 1989; Johnston, 1984; Weisberg, 2005; Kariya and Kurata, 2004; Seber and Lee, 2003; Palta, 2003). In more complex mixed models such as multi-level models that fit data known to emanate from clearly recognizable hierarchically clustered populations, it is known that inferential consequences on fixed and random parameters can be radically altered on account of the level describing the model (Goldstein, 2003). This study examines response variable distributional consequences that could result in a multi-level model following alterations in the level of the model. Two differently conceptualized 4-level Models are explored and it is shown that higher level multi-level models adjudged to be superior also gave response variable distribution parameter estimates that were different from estimates of same response variable obtainable from seeming inferior lower level models. This goes to show the extent to which inferences on the significance of the response variable distribution parameters could be altered on miss-specifying the level of a multi-level model.

MATERIALS AND METHODS

Data structure: The illustrative data employed to enable exploration was drawn from an educational environment. There were basically two data sets herein named Datasets 1 and 2. The data are derived from 50 randomly selected secondary schools in Benue State of Nigeria. Dataset 1 constituted a 4-level data structure (and hence a 4-level Model conceptualization) in which there were 9,999 level 1 units (here students), 450 level 2 units (here subjects or subject groups), 150 level 3 units (here classes) and 50 level 4 units (here schools). The clustering was such that for any original sample n_j ($20 \le n_j \le 30$) of the students from each school j, the n_j was replicated into 9 clusters giving rise to 9 n_j level 1 units for school j (j = 1, 2, ..., 50). In other words, the same n_j students in school j were mirrored in 9 clusters or groups and in particular for each school j, researchers had 9 n_j level 1 units nested in 9 level 2 units that were further nested in 3 level 3 units. Dataset 2 also constituted a 4-level data structure but here there were 6,666 level 1 units (students), 300 level 2 units (subjects or subject groups),

150 level 3 units (classes) and 50 level 4 units (schools); in this dataset, the seeming confounding characteristics in Dataset 1 were reduced by removing the level 2 unit or cluster relating to Common Entrance (CE) and variables based on it.

Description of variables:

- $Navgstem_{ij}$ is the student's final STM score; a level 1 response variable
- $Ncescore_{ij}$ is the student's entrance score; a level 1 predictor variable
- $Normscore_{ij}$ is the JSS1 school STM score student's subject score per class; a level 1 predictor variable
- $Navg1stem_l$ is the final school STM score; a level 4 predictor variable
- $Navgce_l$ is the school common entrance score; a level 4 predictor variable
- $Navg2stem_l$ is the JSSCE school STM score ; a level 4 predictor variable
- $Navg3stem_j$ is the final school STM score; a level 4 predictor variable
- $Navgsub_j$ is score per subject; a level 2 predictor variable
- $Navgincls_k$ is score in class; a level 3 predictor variable
- $Schstatus_l$ is the school status (i.e., whether school is owned as private or public); a categorical predictor variable
- $Schsystem_l$ is the school system; it is a categorical level 4 predictor variable with the categorized into Boardsytem, Daysystem or Bothsystem
- $Schgender_l$ is school gender; it is categorical level 4 predictor variable with school gender categorized into Boys school (Boysch), Girls school (Girlsch) or Mixed (Mixedsch)
- $Nrsqindex_l$ is the school staff quality index (an indication of academic staff quality or strength in any particular school. This is estimated by dividing the total number of qualified academic staff by the entire estimated student population in the school; it is level 4 predictor variable
- $PSStatus_l$ is an indication of electric power supply status in a school; it is a categorical level 4 predictor variable with power supply categorized into school generator, PHCN both or none
- $Labav_l$ is an indication of the availability of science laboratories in a school; it is a categorical level 4 predictor variable lab available categorized into no science lab, one science lab or two or more science labs
- CONS is a constant used for dummy variables and usually carries a value of one; it is a level 1 predictor

The K-level Model: This may be expressed in the compact form:

$$Y = X\gamma+ZU+Z^{(1)}e \quad (1)$$

Where, Y is a column vector of true unobservable responses each assumed continuous:

$$Z = [Z^{(k)}, Z^{(k-1)}, ..., Z^{(2)}]$$

and:

$$U' = [u^{(k)}, u^{(k-1)}, ..., u^{(2)}]$$

The $Z^{(k)}$'s are block diagonal matrices having diagonal elements as $Z_j^{(k)}$ ($j = 1, 2, ..., m_k$) while $u^{(k)}$, X and γ are column matrices with elements, respectively, $u_j^{(k)}$, X_j ($j = 1, 2, ..., m_k$) and γ_{h0} ($h = 0, 1, ..., p$). Researchers assume that $Z^{(1)}e$ and U are normally distributed with zero mean and researchers, symbolically, write:

$$Z^{(1)} e = r \sim N (O, \sigma^2 \ddot{I}^*) \quad (2)$$

and:

$$U \sim N (O, T^*) \quad (3)$$

Where, $\ddot{I}^*$ and T^* are appropriate block diagonal matrices comprising, respectively, the blocks of unit matrices and blocks of variance-covariance matrices of the residual vectors associated with the K-level Model (that is the residual contributions from the levels 2, 3, ..., k in the K-level Model).

Researchers infer from Eq. 1-3 that Y is normally distributed with E (Y) = $X\gamma$ and variance-covariance matrix:

$$V_k = V = E\left[\tilde{E}\tilde{E}'\right] = \sum_1 (V_{k(l)})$$

where:

$$\tilde{E} = ZU + Z^{(1)}e$$

The notation V_k here referring to the covariance (or variance-covariance) matrix associated with the response vector for the K-level Model and $V_{k(l)}$ ($l = 1, 2, ..., k$), respectively, denote the contributions to the covariance matrix of the response vector from levels k, k-1, ..., 1 in a K-level Model.

The level 1 residuals are assumed to be independent across level 1 units. Similarly, levels 2, 3, ..., k residuals are assumed to be independent across levels 2, 3, ..., k units, respectively. It should be noted also that V_k is a block diagonal matrix with block diagonal elements $V_{k(l)}$ ($l = 1, 2, ..., k$) and each of these elements is also block diagonal comprising blocks in their composition.

Models investigated

Dataset 1 case: The 4-level Model considered is:

$$\begin{aligned}\text{Navgstem}_{ijkl} &= \beta_{01} + \beta_{1j}(\text{Normscore} - \text{m(Subject)})_{ijkl} + \beta_{21}(\text{Ncescore} - \text{m(Subject)})_{jkl} + \beta_3(\text{Navg3stem} - \text{gm})_l + \\ &\beta_4\text{Daysystm}_l + \beta_5\text{Bothsystm}_l + \beta_6\text{Girlsch}_l + \beta_7\text{Mixedsch}_l + \beta_8(\text{Normscore} - \\ &\text{m(Subject)}.(\text{Navgsub} - \text{m(Class)}_{ijkl} + \beta_9(\text{Normscore} - \text{m(Subject)}.\text{Schstatus\_1}_{ijkl} + \\ &\beta_{10}(\text{Normscore} - \text{m(Subject)}.\text{Psstatus\_2}_{ijkl} + \beta_{11}(\text{Normscore} - \text{m(Subject)}.\text{Psstatus\_3}_{ijkl} + \\ &\beta_{12}(\text{Normscore} - \text{m(Subject)}.\text{Psstatus\_4}_{ijkl} + \beta_{13}(\text{Nrsqindex} - \text{gm})_l + e_{ijkl}\end{aligned}$$

$$\begin{aligned}&\beta_{01} = \beta_0 + f_{01} \\ &\beta_{1j} = \beta_1 + u_{1jkl} \\ &\beta_{21} = \beta_2 + f_{21}\end{aligned} \quad , \quad \begin{aligned}&\begin{bmatrix} f_{01} \\ f_{21}\end{bmatrix} \sim N(0, \Omega_f) \;\; \Omega_f = \begin{bmatrix} \sigma^2_{f0} & \\ \sigma_{f02} & \sigma^2_{f2}\end{bmatrix} \\ &u_{0jkl} \sim N(0, \sigma^2_{u0}) \\ &e_{ijkl} \sim N(0, \sigma^2_e)\end{aligned} \tag{4}$$

The levels 3, 2 and single level models are respectively given by Eq. 5-7 below:

$$\begin{aligned}\text{Navgstem}_{ijk} &= \beta_{0k} + \beta_{1j}(\text{Normscore} - \text{m(Subject)})_{ijkl} + \beta_{2k}(\text{Ncescore} - \text{m(Subject)})_{jkl} + \beta_3(\text{Navgincls} - \text{gm})_k + \\ &\beta_4(\text{Normscore} - \text{m(Subject)}.\ (\text{Navgsub} - \text{m(Class)}_{ijkl} + e_{ijkl}\end{aligned}$$

$$\begin{aligned}&\beta_{0k} = \beta_0 + v_{0k} \\ &\beta_{1j} = \beta_1 + u_{1jk} \\ &\beta_{2k} = \beta_2 + v_{2k}\end{aligned} \quad , \quad \begin{aligned}&\begin{bmatrix} v_{0k} \\ v_{2k}\end{bmatrix} \sim N(0, \Omega_v) \;\; \Omega_v = \begin{bmatrix} \sigma^2_{v0} & \\ \sigma_{v02} & \sigma^2_{v2}\end{bmatrix} \\ &u_{0jk} \sim N(0, \sigma^2_{u0}) \\ &e_{ijk} \sim N(0, \sigma^2_e)\end{aligned} \tag{5}$$

$$\begin{aligned}\text{Navgstem}_{ij} &= \beta_{0j} + \beta_{1j}(\text{Normscore} - \text{m(Subject)})_{ijkl} + \beta_{2j}(\text{Ncescore} - \text{m(Subject)})_{jkl} + \\ &\beta_3(\text{Navgsub} - \text{gm})_j + \ e_{ijkl}\end{aligned}$$

$$\begin{aligned}&\beta_{0j} = \beta_0 + u_{0j} \ , \beta_{1j} = \beta_1 + u_{1j} \ , \beta_{2j} = \beta_2 + u_{2j} \\ &\begin{bmatrix} u_{0j} \\ u_{1j} \\ u_{2j}\end{bmatrix} \sim N(0, \Omega_f) \;\; \Omega_f = \begin{bmatrix} \sigma^2_{u0} & & \\ \sigma_{u01} & \sigma^2_{u1} & \\ \sigma_{u02} & \sigma_{u12} & \sigma^2_{u2}\end{bmatrix}, \quad e_{ij} \sim N(0, \sigma^2_e)\end{aligned} \tag{6}$$

$$\begin{aligned}\text{Navgstem}_i &= \beta_0 + \beta_1(\text{Normscore} - \text{m(Subject)})_i + \beta_2(\text{Ncescore} - \text{m(Subject)})_i + \beta_3(\text{Navg3stem} - \text{gm})_i + \\ &\beta_4\text{Daysystm}_i + \beta_5\text{Bothsystm}_i + \beta_6\text{Girlsch}_i + \beta_7\text{Mixedsch}_i + \beta_8(\text{Normscore} - \\ &\text{m(Subject)}.(\text{Navgsub} - \text{m(Class)}_i + \beta_9(\text{Normscore} - \text{m(Subject)}.\text{Schstatus\_1}_i + \beta_{10}(\text{Normscore} - \\ &\text{m(Subject)}.\text{Psstatus\_2}_i + \beta_{11}(\text{Normscore} - \text{m(Subject)}.\text{Psstatus\_3}_i + \beta_{12}(\text{Normscore} - \\ &\text{m(Subject)}.\text{Psstatus\_4}_i + \beta_{13}(\text{Nrsqindex} - \text{gm})_i + e_i\end{aligned}$$

$$e_i \sim N(0, \ \sigma^2_e) \tag{7}$$

Dataset 2 case: The 4-level Model considered is:

$$\begin{aligned}\text{Navgstem}_{ijkl} &= \beta_{01} + \beta_{1k}(\text{Normscore} - \text{m(Subject)})_{ijkl} + \beta_2(\text{Nrsqindex} - \text{gm})_l + \beta_3(\text{Navg3stem} - \text{gm})_l + \beta_4\text{Daysystm}_l + \\ &\beta_5\text{Bothsystm}_l + \beta_6\text{Schstatus\_1}_l + e_{ijkl}\end{aligned}$$

$$\begin{aligned}&\beta_{01} = \beta_0 + f_{01} \quad , \ \beta_{1k} = \beta_1 + v_{1kl} \\ &f_{01} \sim N(0, \ \sigma^2_{f0}) \quad , \ v_{0kl} \sim N(0, \ \sigma^2_{v0}) \\ &e_{ijkl} \sim N(0, \ \sigma^2_e)\end{aligned} \tag{8}$$

The levels 3, 2 and single level models are respectively given by Eq. 9-11 below:

$$\begin{aligned}\text{Navgstem}_{ijk} &= \beta_{0k} + \beta_{1k}(\text{Normscore} - \text{m(Subject)})_{ijk} + \\ &\beta_2(\text{Navgincls} - \text{gm})_k + \beta_3(\text{Normscore} - \\ &\text{m(Subject)}).(\text{Navgsub} - \text{m(Class)}_{ijk} + e_{ijk}\end{aligned}$$

$$\begin{aligned}\beta_{0k} &= \beta_0 + v_{0k} \\ \beta_{1k} &= \beta_1 + v_{1k}\end{aligned}$$

$$\begin{bmatrix} v_{0k} \\ v_{1k} \end{bmatrix} \sim N(0, \Omega_v) \quad \Omega_v = \begin{bmatrix} \sigma^2_{v0} & \\ \sigma_{v01} & \sigma^2_{v1} \end{bmatrix} \qquad (9)$$

$$e_{ijk} \sim N(0, \sigma^2_e)$$

$$\begin{aligned}\text{Navgstem}_{ij} &= \beta_{0j} + \beta_{1j}(\text{Normscore} - \text{m(Subject)})_{ij} + \\ &\beta_2(\text{Navgsub} - \text{gm})_j + e_{ij} \\ \beta_{0j} &= \beta_0 + u_{0j} \\ \beta_{1j} &= \beta_1 + u_{1j} \\ e_{ij} &\sim N(0, \sigma^2_e)\end{aligned} \qquad (10)$$

$$\begin{aligned}\text{Navgstem}_i &= \beta_0 + \beta_1(\text{Normscore} - \text{m(Subject)})_i + \\ &\beta_2(\text{Nrsqindex} - \text{gm})_i + \beta_3(\text{Navg3stem} - \text{gm})_i + \\ &\beta_4\text{Daysystm}_i + \beta_5\text{Bothsystm}_i + \beta_6\text{Schstatus\_1}_i + e_i\end{aligned}$$

$$e_i \sim N(0, \sigma^2_e) \qquad (11)$$

It should be noted that for each of the Datasets there are of course, several alternative K-level Model formulations for each k (k = 1-4) besides the ones reflected above but the adopted ones here were affirmed (using the MLWiN software) to have the least deviance and generally retained the highest number of statistically relevant predictor variables.

In this study, three variants of single level model formulations are considered from each of the two data sets to enable some insight into differing conclusions that may result in respect of the response variable distribution. In the first variant (this being the only variant which equations are explicitly depicted in this study), all variables used in the k-level full structure are recast as single level variables. In the second variant, the 2-level is recast in a single level framework while the third variant simply removes from the 2-level Model all variables originally defined as belonging to 2 or higher levels.

RESULTS AND DISCUSSION

To explore the distributional consequences of the response variable as model levels vary, the K-level Models in each of Datasets 1 and 2 are estimated via Iterative Generalized Least Squares (IGLS) technique implemented in MLWiN package 2.20 and response variable values from each of the fitted models generated. Datasets 1 and 2 shall generate 9,999 and 6,666 response values, respectively for each K-level Model. The fitted model deviance, population mean, standard deviation and variance of the response variable are then estimated for each K-level Model. From Table 1 and 2, the Asterisked (*) Model designations denote the estimation results associated with the second and third variants of the Single Level Model formulations alluded to in the preceding section.

Table 1 and 2 relating to Dataset 1 indicated that higher level models returned lower model deviances in general (an indication of better model fit with increasing model level) and where as the response variable exhibited common mean response value (-0.000036), the response variable variances (and hence standard deviations) differed with differing model levels; tending towards 1 as

Table 1: Response variable mean and variance estimates for 9,999 response values

		Response variable estimate					
Models	Levels	Mean	Variance	SD	Minimum	Maximum	Model deviance
2.4	4	-0.000036	0.4936	0.7026	-3.8692	3.3022	21940
2.5	3	-0.000036	0.4757	0.6897	-3.5929	3.0749	22761
2.6	2	-0.000036	0.4298	0.6556	-3.4350	2.9268	23856
2.7	1	-0.000036	0.3915	0.6257	-2.6789	2.0753	23404
2.7	1	-0.000036	0.1215	0.3485	-1.2748	1.1617	27077
2.7	1	-0.000036	0.0391	0.1977	-0.8018	0.9904	27974

Table 2: Response variable mean and variance values based on 9,999 collected

	Raw values			
Mean	Variance	SD	Minimum	Maximum
-3.6E-05	0.9997	0.9998	3.95	3.31

Table 3: Response variable mean and variance estimates for 6,666 response values

		Response variable estimate					
Models	Levels	Mean	Variance	SD	Minimum	Maximum	Model deviance
2.8	4	-0.00012	0.4803	0.6930	-3.3956	3.4906	14785
2.9	3	-0.00012	0.4667	0.6832	-3.3588	3.3938	15173
2.10	2	-0.00012	0.4425	0.6652	-3.2900	2.8113	15644
2.11	1	-0.00012	0.3733	0.6110	-2.2999	2.1704	15803
2.11	1	-0.00012	0.1814	0.4259	-1.7555	1.6744	17583
2.11	1	-0.00012	0.0766	0.2768	-1.2541	1.6379	18386

Table 4: Response variable mean and variance values based on 6,666 collected

	Raw values			
Mean	Variance	SD	Minimum	Maximum
-0.00012	1.0002	1.0001	-3.82	3.25

the model level increased. Thus as model level increases the response variable had estimated distribution parameters (namely mean and standard deviation) getting closer to those of the response values based on raw collected data which has mean -0.000036 and standard deviation 0.9998. The interval limits of response variable values also differed with differing model levels as the model levels increased the interval limits increased from (-0.8018, 0.9904) to (-3.8692, 3.3022). The interval limits of response variable values from the raw collected data are (3.95, 3.31). That is increasing the model level results in predicting response variable values that are within a range that is very close to that obtainable in the raw collected data.

Table 3 and 4 shows patterns that are similar to what obtained in the preceding tabular values relating to Dataset 1; models with higher levels yielded lower model deviances and also the mean and variance distributional parameters of the response variable better approximated the actual raw collected data values as the model level increased. The interval limits of the response variable values (as predicted by the models) increased with increasing model level; getting closer to the raw collected data response variable values interval limit of (-3.82, 3.25). Thus as in the case of the tabular values of Dataset 1, increasing the model level results in predicting response variable values that are within a range much closer to that obtainable in the raw collected data.

CONCLUSION

It is well documented in the multi-level modeling literature that misleading inferences could be made on the relevance or otherwise of predictor variables in formulated linear models (and indeed on the adequacy or otherwise of the entire model) when such models are erroneously cast as Single-level Models or in an inappropriate K-level Model framework (Goldstein, 2003; Bryk and Randenbush, 1992, 2002; Hox, 1995). That model adequacy can be called to question when a model is fit in an inappropriate K-level framework makes it plausible to examine possible inferential differences that could result on the response variable of a multi-level model if the model level is altered. In the illustrative analysis employed, it is seen in the hierarchically structured educational data explored that judging from model deviance values and increasing level of the K-level Model improved model adequacy. The normal distribution assumption in respect of the response variable (here called Navgstem) appears to have been more closely met by the response variable values generated by the 3 and 4-level Models than what obtained in the lower level models. Single-level Model formulations particularly generated response variable values that largely formed only a sub-range of the large range of true response variable values and the estimated distribution population parameter variances, especially differed significantly from what obtained in the 4-level Model or raw collected data cases. Confidence interval inference investigations on the response variable could also yield misleading answers (especially in respect of the response variable standard deviation or variance) if the values of the response variable are generated or simulated from an in appropriate K-level Model.

It is opined that if studies on the variability of the response variable in a Hierarchical Linear Model are consequential then care must be taken to ensure that the multi-level model is cast in an appropriate K-level Model framework.

REFERENCES

Bryk, A.S. and S.W. Raudenbush, 1992. Hierarchical linear models: Advanced Quantitative techniques in the social sciences. 2nd Edn., Sage Publications, Newbury Park, ISBN: 9780803946279, Pages: 265.

Bryk, A.S. and S.W. Randenbush, 2002. Hierarchical linear models. 2nd Edn., Sage Publications, Newbury park, California, ISBN: 9780761919049, Pages: 512.

Goldstein, H., 2003. Multilevel Statistical Models. 3rd Edn., Edward Arnold, New York, London,.

Hox, J.J., 1995. Applied Multilevel Analysis. TT-Publikaties, Amsterdam, ISBN: 90-801073-2-8, Pages: 126.

Johnston, J. and J. DiNardo, 1997. Econometric Methods. 4th Edn., McGraw Hill, New York.

Johnston, J., 1984. Econometric Methods. 3rd Edn., McGraw-Hill, New York.

Kariya, T. and H. Kurata, 2004. Generalized Least Squares. Wiley, New York.

McCullagh, P. and J.A. Nelder, 1989. Generalized Linear Models. 2nd Edn., Chapmann and Hall, London, ISBN: 0-412-31760-5, Pages: 536.

Palta, M., 2003. Quantitative Methods in Population Health: Extensions of Ordinary Regression. Wiley and Sons, Hoboken, New Jersey.

Seber, G.A.F. and A.J. Lee, 2003. Linear Regression Analysis. 2nd Edn., Wiley-Interscience, New York, ISBN: 9780471415404, Pages: 557.

Weisberg, S., 2005. Applied Linear Regression. 3rd Edn., John Wiley and Sons, USA.