Upper-Level Expectation in Random Coefficient Logit Model

Lee, SeongWoo* · Ryu,sungHo*

다중 로짓 모형에서의 상위차원의 예측치 통계에 관한 연구

이성우 · 류성호

* 서울대학교 농업생명과학대학 농경제사회학부

초 록

본 연구는 다음의 두 가지 목적이 있다. 첫째, 각종 실증분석에 있어서의 다중모형의 효율성에 대한 소개와, 둘째, 다중모형의 분석에 있어서 상위단계의 예측되는 가치를 측정하기 위한 새로운 통계를 소개하는 데 있다. 다중모형의 이론적 틀은 광범위하게 사용되는 기존의 1단계 모형의 통계적 문제점(이분산 등)을 보완하고, 현실을 더욱 실체적으로 파악한다는 측면에서 앞으로 지역분석의 중추적 틀로서 자리매김하리라 예상되고 있다. 본 연구는 이러한 다중모형의 효율성을 가상 자료가 아닌 실제 자료를 이용하여 검증하였으며, 특히 기존에 제시되지 않은 다중로짓모형에서의 상위수준의 잔차 또는 예측치를 계산하는 통계량을 제시하였다. 이 새로운 통계량은 실증분석에 있어서의 관찰치와의 상관관계와 그 분산의 분석에 있어서 잘 행위하고 있는 것으로 나타났다.

I. Introduction

A persistent theme in planning involves understanding the influence of the community, or more generally, the social context on individual behavior. Although theoretical interest in the problem has been long standing, the statistical tools to address them properly have been found in years. The statistical complications stem from the hierarchical or nested structure of the requisite data, *i.e.*, individual observations are organized into larger units or clusters, which, in turn, may be grouped into still larger units. Since observations from the same group tend to be more alike than observations from different groups, the classical assumption of independence between observations is likely to be violated.

Hierarchical structure of data is particularly clear in many social and geographical analyses. Examples of such systems include housing units, administrative units such as, provinces, cities, counties, etc. The recognition of scientific concerns with multi-dimensional orientation and the conception of hierarchically organized data imply that we should take that dimension into account when we analyze data. Despite the prevalence of these concerns, however, past studies have often failed to address them adequately in the data analysis.

This paper considers aspects of the specification and estimation of multi-level logit model particularly focused on the hypothesis that the hyper-level residual is zero, which has not yet been cultivated in the literature. We start with an overview of the special problems encountered when analyzing grouped data in Section Ⅱ. In section Ⅲ, we introduce the random coefficient linear model and its extension to binary response data, along with the specification of a new statistic for the calculation of hyper-level expectation in the multi-level logit model. After explaining data and variables in Section Ⅳ, Section Ⅴ and Ⅵ summarize interpretations and conclusions.

II. Theoretical Concerns

Three most important concepts--cross-level inference, spatial heterogeneity, and spatial dependency--disregarded in the existing literature are discussed in this section.

Cross-level inferences are interactions between explanatory variables defined at different levels of the hierarchy (Hox and Kreft, 1994). When variables from different levels are analyzed at one single level, it becomes an important problem to identify the proper level to which all variables must be aggregated or disaggregated for statistical analysis. There is the possibility of committing a fallacy in analyzing data at one level and making inferences to another level when the researcher interprets results. This fallacy is best represented by the well-known 'ecological fallacy' (Robinson, 1950) and 'atomistic fallacy' (Alker, 1969).

Spatial heterogeneity A model takes on its operational form when it is applied to any specific real world context. If the realizations of a model in widely different contexts are identical, then the model is independent of geographical situation. In contrast, a model is contextually dependent if its realization varies in different operational geographies. For example, housing is, in general, characterized by surmounting geographic contrasts, making knowledge of spatial differences essential for understanding housing dynamics. Thus the individual behavior of residential choice shows quite different responses depending upon particular local housing market conditions (Timmermans and Noortwijk, 1995).

Spatial dependency In general, observations within a group that are close in space are expected to be more similar than observations in distant groups (Anselin, 1988;1992). In general, groups are rarely formed at random but rather on the basis of some homogeneity (Blalock, 1984). Ignoring the values of group similarity (intraclass correlations) leads to Type I errors that are much larger than the nominal significance level (Hox and Kreft, 1994). This is particularly so considering the discrete spatial distribution of local housing markets. It can be easily anticipated that individuals who live in the same geographic area are more likely to be alike, in some way, than people in the other geographic areas. Moreover, with hierarchically structured data sets such as houses nested in areas, the characteristics of the dependency is expected as "the norm" (Jones and Bullen., 1994).

Multi-level modeling addresses precisely these concerns. Instead of reducing the world to one fixed equation, it recognizes that there are different relationships for different places or contexts. The next section draws a methodological procedure encapsulating these theoretical concerns.

III. Methodological Concerns

III-1. Random Coefficient Linear Model

Provided we have two levels of observation, the household (micro) level and the MSA (macro) level. We hypothesize that the micro values of the response variable in some way depend on each MSA and that the effects of the micro determinants may vary systematically as a function of idiosyncratic MSA characteristics. Without individual subscript for the convenience, suppose there are n_j -element household level dependent variable vector y_j , regressor matrix X_j defined by m groups (j=1 to m) of MSAs and p household level regressors (s=1 to p) with the total number of observations $N=\sum_{j=1}^{j=m} n_j$. Define a household level equation identically for each MSA:

$$y_j = X_j \beta_j + \varepsilon_j \qquad (1)$$

where β is a p×1 vector of unknown regression parameter,j=1,⋯,J macro level units and MSAs are free to have different numbers of individual observations.

Assuming ε_j are independently distributed as $N(0_j, \Sigma_j)$. If we assume $\Sigma_j = \sigma_j^2 I$ that is, independent and constant-variance observations, then equation (1) is a standard linear model. Because equation (1) poses no unusual estimation or computation problems, the fixed effects regression model has been used frequently in multilevel situation (Kallan, 1993; Lee et al., 1995).

A more realistic model can be explored by letting each intercept and slope vary in the MSA level, termed a random coefficient model. Assuming β_j is a random sample from a multivariate normal, $\beta_j \sim N_p(\beta, \Xi)$ uncorrelated with ε_j, this is equivalent to the random coefficient model

$$y_j = X_j\beta + Z_j\gamma_j + \varepsilon_j \qquad (2)$$

where the matrix Z_j are stacked by selection of certain interests of variables (columns of X_j - *e.g.*, tenure and housing vintage in model 7) and $\gamma_j = \beta_j - \beta$ is the vector of deviations of the regression coefficients β_j from the their expectation β. In this case, the matrix Z contains the intercept(=1) as its first column and its variance is presented by σ_γ^2 . We also denote σ_ε^2 corresponding to household level intercept variance term.

Let $var(\gamma) = \Xi$, and $var(\varepsilon) = \sigma^2 I$ and $cov(\gamma, \varepsilon) = 0$, so that $E(y) = X\beta$ and the variance of y has the following structure:

$$\Sigma_j \begin{bmatrix} X_1 \Xi X_1' + \sigma^2 I & 0 & 0 & 0 \\ 0 & X_2 \Xi X_2' + \sigma^2 I & 0 & 0 \\ 0 & 0 & \cdots & 0 \\ 0 & 0 & 0 & X_j \Xi X_j' + \sigma^2 I \end{bmatrix} \qquad (3)$$

Among the various forms of the covariance structure of Ξ (Jennrich et al., 1986; Littell et al., 1996), we adopt a banded main diagonal covariance structure where needed. This covariance structure is relevant in our study in that we are interested in the variation of tenure and housing age which are specified as categorical variables in the empirical models. Jennrich and Schluchter (1986) proved that the covariance structure work quite well in their simulation study.

Two special cases from the model (2) are worthy of attention and correspond to the analysis of our study. A random effects analysis of variance model (ANOVA) empirically first drawn by Moellering and Tobler (1972), is obtained by setting to zero all of the coefficients of X_j and Z_j except both levels of intercepts:

$$y_j = \beta_1 + \gamma_j + \varepsilon_j \qquad (4)$$

where β_1 is a constant term indicating the grand mean of y as is shown in model 1, 2, and 8 in this paper . This model includes no capacity for explaining variability in y_j at either the household or MSA level, but it does include two sources of random variability in y_j .

Another important case where only Z_j is a vector of ones, so called random intercept model (Bryk et al., 1992) has the following form:

$$. \ y_j = X_j\beta + \gamma_j + \varepsilon_j \qquad (5)$$

This model depicts a picture as a series of parallel lines with the same fixed slope but varying intercepts ($\beta_1 + \gamma_j$). Because of the computational efficiency, many applied researchers adopt this model (Ward and Dale, 1992; Duncan et al., 1993; O'Campo et al., 1995).

III-2. Random Coefficient Logit Model

We now consider the case where y is a vector of binary outcome that corresponds to our interest. We follow the general procedure as in Longford (1993) and Wolfinger and O'Connell (1993). In this study, we assume j=1,2,$\cdots$,100 as a MSA level random sample with household level units i=1,2,$\cdots$, n_j. We use $\eta = logit(\gamma)$ as the canonical link where γ is the MSA-level components of the random vector. Thus probability of outcome

$$\gamma_{ij} = Prob(y_{ij} = 1) = p$$
$$\gamma_{ij} = Prob(y_{ij} = 0) = 1 - p$$

is related to the linear predictor by the logit link

$$\eta = log\left(\frac{p}{1-p}\right) = X_{ij}\beta + Z_{ij}\gamma_j \qquad (6)$$

assuming that γ_j has a multivariate normal density with

$E(\gamma)=0$ and $var(\gamma)=\Xi$ as in (3). We further assume that the households within MSAs are conditionally independent, given the random vector γ_j , then the unrestricted log-likelihood related to y is

$$L(\beta, \Xi \mid y)=\sum_j log \int \cdots \int P_j(\gamma_j)\Phi(\gamma_j)d\gamma_j \qquad (7)$$

where $\Phi(\gamma)$ is the density of the multivariate standard normal distribution and $P_j(\gamma_j)$ is the conditional likelihood for MSA j. Because restricted maximum likelihood (MLR) is 'statistically sound' (Dempster et al., 1981) and less biased than unrestricted maximum likelihood (MLU) estimates (Wong and Mason, 1985) as in (7), we adopt MLR throughout the analyses. MLR can be specified as

$$L_{MLR}=L_{MLU}+\left[-\frac{1}{2} log\{det(X' \ \Sigma^{-1}X)\}\right] \qquad (8)$$

where L_{MLU} is defined by (7) and Σ is the variance of y as in (3). Whole estimation procedures are carried out by a %GLIMMIX macro function recently developed by SAS (Littell et al., 1996). However, Equation (8) is numerically intractable, and so we adopts restricted pseudo-likelihood (REPL) as its approximation (for details, see Wolfinger and O'Connell, 1993). Among the several options in SAS for numerical integration, we adopt Newton-Raphson algorithm that is believed to be very rapid for well-identified models (Longford, 1993).

III-3. Computation of Hyper-Level Expectation in Random Coefficient Logit Model

Unlike the multilevel linear model as in (2), the multilevel logit model connected to canonical binomial link as in Equation (12) is immune to household level assumption of $E(\varepsilon)=0$, however, we need to test the assumption of MSA level error distribution $E(\gamma)=0$. The statistic for this test is given by (Goldstein, 1987) and the form

$$\hat{\gamma}_{j=}[\{(\ \textstyle\sum_j \ \hat{q}_{ij})/ n_j\}\times\{ n_j \ \ \hat{\sigma}_\gamma{}^2\}]\div(n_j \sigma_\gamma^2+ \ \hat{\sigma}_\varepsilon{}^2), \qquad (9)$$

Table 1. Variables and Descriptive Statistics

Variable	Description	Total HHs	
		Mean or %	S.D
Dep. Var.	Moved during the last 15 months period	20	
	Same house during the last 15 months period	80	
Ind. Vars.			
RACE	Other races(1)	24	
	NH White(0)	76	
TENURE	Renter(1)	37	
	Owner(0)	63	
HINC	Deviation from mean household income divided by1000	42	38
HINCSQ	Square of HINC	3231	8594
AGE	Deviation from mean age	48	17
AGESQ	Square of AGE	2621	1807
HHSIZE	Deviation from mean household size	3	1.5
HSIZESQ	Square of HHSIZE	9	11.5
FAMTYPE	Non-married, Divorced, Separated, others (1)	45	
	Married couple (0)	55	
PREVMIG	If place of birth is different from the residence of 5 years ago (1)	23	
	If place of birth is the same as the residence of 5 years ago (0)	77	
VINT_80	Housing built in the 1980s	19	
VINT_70	Housing built in the 1970s	20	
	Housing built in the 1960s (omitted reference group)	44	
VINT_P60	Housing built in the pre-1960	17	

where $\hat{q}_{ij}$ is composite residuals calculated by substracting model estimates of $\hat{y}_{ij}$ from observed y_{ij} , n_j is MSA level units, $\hat{\sigma}_{\gamma}^2$ is an estimated variance at the MSA level, and $\hat{\sigma}_{\varepsilon}^2$ is an estimated variance at the household level. Equation (9) is made for multilevel linear model, so a new statistic which fits for multi-level logit model must be adopted. Once the individual characteristics of the households have been taken into account, the MSA-level residuals can be seen as estimates of the remaining differences between the MSAs. In case of Equation (6) of binary logit model, the composite residual becomes

$$\hat{q}_{ij} = [\{ y_{ij} - exp(X\,\hat{\beta}_{ij} - Z\,\hat{\gamma}_j)\} \div \{1 + (y_{ij} - exp(X\,\hat{\beta}_{ij} - Z\,\hat{\gamma}_j))\}] \qquad (10)$$

IV. Data and Variables

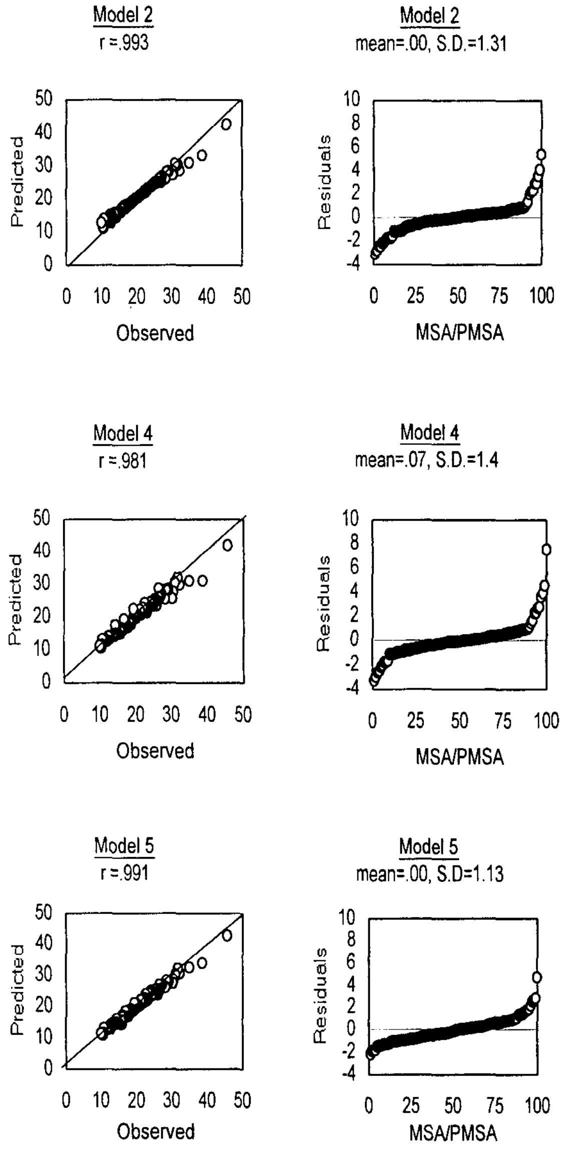
Data used in the analysis are drawn from the 1990 U.S. decennial census of population and housing, specifically the Public Use Microdata Sample (PUMS) file A, which is a 5% sample of all households in the U.S. Because of the

Table 2. Summary of Two-Level Random Coefficient Models

			Model 1	Model 2	Model 3	Model 4	Model 5
			Coeff.	Coeff.	Coeff.	Coeff.	Coeff.
FIXED	Household						
		INTERCEPT	-1.3773***	-1.3774***	-3.0407***	-3.0415***	-3.0599***
		RACE			0.0290	0.0301	0.0343
		TENURE			1.3314***	1.3323***	1.3801***
		HINC			-0.0023***	-0.0023***	-0.0023***
		HINCSQ			0.0000***	0.0000***	0.0000***
		AGE			-0.0500***	-0.0500***	-0.0500***
		AGESQ			0.0008***	0.0008***	0.0008***
		HHSIZE			-0.0664***	-0.0664***	-0.0637***
		HSIZESQ			0.0070*	0.0070**	0.0065*
		FAMTYPE			0.1356***	0.1361***	0.1361***
		PREVMIG			1.0501***	1.0499***	1.0555***
		VINT_80			0.6764***	0.6754***	0.6798***
		VINT_70			0.1404***	0.1395***	0.1326***
		VINT_P60			-0.1525***	-0.1509***	-0.1325***
RANDOM	Household						
		INTERCEPT	1.0000	0.9938***	1.0000	0.9111***	0.8693***
	MSA						
		INTERCEPT	0.1260***	0.1259***	0.0433***	0.0460***	0.0156*
		TENURE					0.0755***
		VINT_80					0.0750***
		VINT_70					0.0069
		VINT)P60					0.0100
	Deviance		43,054	43,054	31,942	31,934	31,701

Model 1: two-level null model (intercept only). assuming binomial level 1 variance.
Model 2: as Model 1, but unconstrained level-1 variance.
Model 3: as Model 1 , but includes individual-level explanatory variables.
Model 4: as Model 3, but unconstrained level-1 variance.
Model 5: as Model 4, but with two random parts in which the relationship with tenure and housing vintage is random between MSAs.
*** Asymptotic t-test or Z-test significant at the 0.01 level.
** Asymptotic t-test or Z-test significant at the 0.05 level.
* Asymptotic t-test or Z-test significant at the 0.10 level.

Figure 1. Model Diagnostics for Equation (10)



great size of this sample, multivariate models are estimated with a sub-sample of the PUMS-A file, amounting to a 1-in-1000 sample. We have extracted data for the 100 largest metropolitan areas (MSA or PMSA). This number is large enough for us to adequately evaluate the metropolitan-level effects.

A series of household-level and metropolitan-level variables (income, age, household size) and their square terms were were selected for this study. The three continuous variables deviated around their mean for the sample. This 'centering' schema has several advantages particularly in the multi-level modeling (see Bryk and Raudenbush 1992; Kreft et al., 1995). We provide detailed descriptions for the variables of this study in Table 1.

V. Results

The results of this operation are shown in Figure 1. Combining owners and renters, the overall expected mobility rates in each area can also be generated. Figure 1-a plots the expected mobility rate versus the observed for each metropolitan area, as computed from Models 2, 4 and 5. The correspondence between expected and actual mobility rates is very close (r=0.99). We note, however, that high mobility areas tend to be under-predicted while low mobility areas are over-predicted. The right hand side of Figure 1-b shows that MSA-level residual distribution called 'shrunken residual' (Goldstein, 1995) is quite well-behaved in our models satisfying the assumption of $E(\gamma)=0$. As expected, Model 3 equipped with the shrunken estimates $\hat{\gamma}_j$ generates a much better fit, as shown by the mean residual of 0 and the lower standard deviation.

VI. Conclusion

Even though most regional data contain information on group membership, this information is often poorly used. Group variables are often entered indiscriminately into the analysis along with individual variables, or group information is simply ignored. Our example highlights features of analysis using multi-level model, and how analysis by more traditional regression models can lead to biased estimates of uncertainty and different conclusions. As demonstrated, multi-level model provides a powerful tool for analysis of grouped data where the number of individuals within groups varies.

A hierarchical logistic regression model is proposed for studying data with group structure and a binary response variable. The group structure is defined by the presence of micro observations embedded within groups (macro observations), and the specification is at both of these levels. While hyper-level expectation term or computation of macro-level residual in multi-level analysis has been ignored, the present study provides a new statistic that handles the caveat.

The statistic has been behaved particularly well. It is our hope that, by using the statistic, researchers will find more realistic results.

References

Alker H S, 1969, "A typology of ecological fallacies" pp. 69-86 in Dogan M, Rokkan S (eds.). *Quantitative Econological analysis* (MIT Press, Mass.)

Anselin L, 1992, "Space and applied econometrics" *Regional Science and Urban Economics* 22 307-316

Anselin L, 1988, "Model validation in spatial econometrics: a review and evaluation of alternative approaches" *International Regional Science Review* 11 279-316

Blalock H, 1984, contextual-effects models: theoretical and methodological issues" *Annual Review of Sociology* **10** 353-372

Bryk A S, Raudenbush S W, 1992, *Hierarchical Linear Models* (Sage, Newbury Park)

Dempster A P, Rubin D B, Tsutakawa R K, 1981 "Estimation in covariance components models" *Journal of the American Statistical Association* 76 341-353

Duncan C, Jones K, Moon G, 1993, "Do places matter? a multi-level analysis of regional variations in health-related behaviour in Britain" *Social Science and Medicine* 37 725-733

Goldstein H, 1995, *Multilevel Statistical Models* (John Wiley & Sons Inc., New York)

Goldstein H, 1987, *Multilevel Models in Educational and Social Research* (Griffin, London)

Hox J J, Creft I G G, 1994, "Multilevel analysis methods" *Sociological Methods & Research* 22 283-299

Jennrich R I, Schluchter M D, 1986, "Unbalanced repeated-measures models with structured covariance matrices" *Biometrics* 42 805-820

Jones K, 1991, *Multi-Level Models for Geographical Research* (CATMOG 54)

Jones K, Bullen N, 1994, "Contextual models of urban house prices: a comparison of fixed- and random-coefficient models developed by expansion" *Economic Geography* 70 252-272

Kallan J E, 1993 "A multilevel analysis of elderly migration" *Social Science Quarterly* 74 403-419

Kreft I G G, DeLeeuw J, Aiken L S, 1995, "The effect of different forms of centering in hierarchical linear models" *Multivariate Behavioral Research* 30 1-21

Lee B A, Oropesa R S, Kanan J W, 1995, "Neighborhood context and residential mobility" *Demography* 31 249-270

Littel R C, Milliken G A, Stroup W W, Wolfinger R D, 1996, *SAS Systems for Mixed Models* (SAS Institute Inc., N.C)

Longford, N T, 1993, *Random Coefficient Models* (Oxford, London)

Moellering H, Tobler W, 1972 "Geographical variances" *Geographical Analysis* 4 34-50

O'Campo P, Gielen A C, Faden R R, Xue X, Kass N, Wang M C, 1995, "Violence by male partners against women during the childbearing year: a contextual analysis" *American Journal of Public Health* 85 1092-1097

Robinson W S, 1950, "Ecological correlations and the behavior of individuals" *American Sociological Review* 15 351-357

Timmermans H, Noortwijk L V, 1995, "Context dependencies in housing choice behavior" *Environment and Planning A* 27 181-192

Ward C, Dale A, 1992, "Geographical variation in female labour force participation:an application of multilevel modelling" *Regional Studies* 26 243-255

Wolfinger R, O'Connell M, 1993, "Generalized linear models: a pseudo likelihood approach" *Journal of Statistical Computation and Simulation* 48 233-243

Wong G W, Mason W M, 1985, "The hierarchical logistic regression model for multilevel analysis" *Journal of American Statistical Association* 80 513-524