Multilevel Models with Spatial Data

Richard A. Berk, Jan de Leeuw, Ming Zheng, Yan Xiong, with
Craig Schuman, Domingo Ochavillo and Cindy Lin

Department of Statistics, UCLA

May 14, 2022

1 Introduction

Multilevel statistical models are a relatively recent development. The basic idea is that one's data are nested, and the goal is to characterize how the different levels are related. The classic application is in educational research.

One might want to study how the socio-economic background of students is related to the scores on some standardized test. But students are assembled into classes. One might then be interested in how the relationship between a student's background and test score differs depending on the teacher-student ratio of the class. Perhaps students from higher income households are better able to capitalize on a favorable teacher-student ratio.

It would seem that one could explore this with conventional regression analysis by including the relevant interaction term. But this would only take you part of the way. The clustering of students within classes could well mean that the regression disturbances were not independent. Then, conventional estimates of the standard errors would probably be too small. Falsely narrow confidence intervals and falsely powerful hypothesis tests would follow. A Multilevel model for linear regression would solve such problems in principle. There are a number of excellent textbook treatments available including Kreft and De Leeuw (1998) and Bryk and Raudenbush (2002).

Nested data are common in ecological research, but can be somewhat more complicated to analyze. First, errors of the multilevel model can be related not just as a function of clustering but as because of spatial proximity

more generally. The errors for observations that are closer together in space may be more alike than the errors for observations that are farther apart. A failure to properly account for such spatial autocorrelation will lead, as before, to incorrect estimates of the standard errors. Second, one has allow for the errors (or equivalently the conditional distribution of the response variable) to have a non-normal but still specified distribution, such as the binomial or the Poisson. Such flexibility is found in the generalized linear model. For example, a binary outcome such as polluted or not might suggest a binomial distribution as operationalized in logistic regression.

It this paper (and especially in the lengthy appendix), we present a generalization of multilevel regression models that allows for spatially dependent errors and for the class of link functions and conditional error distributions found with the generalized linear model. An application will also be presented.

2 Some Basics: A Very Simple Illustration

Suppose one is interested in the diversity of fish species living around coral reefs and in particular the impact of preserves. Do fishing preserves increase species diversity? Data come from a set of transects from two different kinds of reefs: one protected from all fishing and one unprotected. Along each transect, data are collected from four locations 5 meters apart. Transects are taken here as the observational units at the lowest level while the reef is taken to be the observational unit at the highest level. For this simple illustration, we will ignore locations within transects. The response variable is a measure of species diversity. The two predictors are the percentage of the ocean floor that is sandy and whether the reef is protected or not. The former is taken to operate at the first (or "micro") level while the later is taken to operate at the second (or "macro") level.

At the"micro" or "first" level let

$$diversity_{ij} = \beta_{0j} + \beta_{1j} sandy_{ij} + \epsilon_{ij}, \tag{1}$$

where parameters are subscripted as 0 and 1, i refers to the transect, j to the reef, and ϵ_{it} is the usual regression error term. A key point is that the errors are assumed to be independent. The subscript j attached to the slope and intercept implying that each reef can have its own equation.

The researcher would like to systematically characterize the similarities and differences in these equations across the two reefs. One strategy would be to employ a fixed effects model with an indicator variables for reef. In principle, one could allow for different intercepts and different slopes.

But this may be unsatisfactory for at least three reasons. First, allowance would have to be made for the possibility of different error variances in the different schools. Second, the number of parameters to be estimated could be very large. Third, the role of the whether the reef is protected or not has not been considered. In response, one might specify the following relationships (as but one of several possibilities) for the *parameters* in the first-level model. One might write:

$$\beta_{0j} = \gamma_{00} + \gamma_{01} protected_j + \delta_{0j} \tag{2}$$

and,

$$\beta_{1j} = \gamma_{10} + \gamma_{11} protected_j + \delta_{1j}, \tag{3}$$

where reef-level errors δ_{0j} and δ_{1j} are uncorrelated with each other and the transect-level errors.

The intercept is now a random variable with γ_{00} the mean diversity for unprotected transects, γ_{01} the change in average diversity in the protected transects, and δ_{0j} the source of the uncertainty. The slope is now also a random variable with γ_{01} the slope in unprotected transects, γ_{11} the change in the slope in protected transects, and δ_{1j} the source of uncertainty. Whether the reef is protected is allowed to affect both average species diversity and how the composition of the ocean floor translates into species diversity. Substituting equations **??**, **??** into equation **??**,

$$\begin{aligned} diversity_{it} = (\gamma_{00} + \gamma_{01} protected_j + \delta_{0j}) + \\ (\gamma_{10} + \gamma_{11} protected_j + \delta_{1j}) sandy_{ij} + \epsilon_{ij}, \end{aligned} \tag{4}$$

which simplifies to

$$\begin{aligned} diversity_{ij} = \gamma_{00} + \gamma_{01} protected_j + \gamma_{10} sandy_{ij} + \\ \gamma_{11}(protected_j \times sandy_{ij}) + (\delta_{1j} sandy_{ij} + \delta_{0j} + \epsilon_{ij}). \end{aligned} \tag{5}$$

Equation **??** is a conventional linear regression with a main effect for the composition of the ocean floor and for whether the reef is protected, and an interaction effect between composition of the ocean floor and whether the reef is protected. As such, there is nothing mysterious. However, the errors no longer have constant variance because of the product of *sandy* and

the reef-level errors associated with the first-level slope parameter. Applying least squares to equation **??** will produce unbiased estimates of the two regression coefficients and intercept, but because of the usual constant variance assumption, get the estimates of the standard errors wrong. However, there are a wide variety of consistent estimators that do a much better job with the standard errors.

3 Some Generalizations

The multilevel framework can be generalized beyond this simple example in a number of ways. Perhaps most obviously, one can include a number of predictors at each observational level. The risk is high levels of multicollinearity. When multilevel models are constructed, it is rare for reserchers to do the substitution illustrated in equation **??**. It may then not be apparent how many interaction terms are being added to the model and how much instability can be generated as a result.[1]

A second elaboration of the bacic model is to employ a mix of random and fixed effects for different coefficients. A major hurdle is to make the case that random effects are sensible. It is one thing to assert as a technical matter that the model's intercepts, for instance, behave as if drawn independently and at random from a particular distribution, but quite another to make a convincing case that nature happens to operate in so convenient a manner. One must argue from subject-matter knowledge that such an account makes good sense. Typically this will be very difficult to do.

A third elaboration is to employ response variables within the framework of the generalized linear model. Common examples are binary response variables and count response variables. While moving beyond the normal linear regression model to the binomial or the Poisson may be relatively small step conceptually, for multilevel models it complicates enormously the estimation procedures. One consequence is that there may be more than one solution to the estimation problem so that the algorithm may converge to a local, not global result. It is vital, therefore, that the output be carefully examined to

[1]It is often very useful to do those substitutions and estimate the resulting model with ordinary least squares. Recall that the estimates of the regression coefficients are in principle unbiased. Then, one has easy access to all of the usual regression diagnostics, including those for multicollinearity. Careful examination of those diagnostics can be very enlightening and rather humbling.

determine if it makes scientific sense. It can also be helpful to estimate the model's parameters several times using different start values. If all of the results look about the same, all is probably well. If not, one would need to decide which results to accept and at the very least, report that there seem to be several sets of plausible parameter estimates.

Finally, there is the matter of how to represent dependence among the disturbances. Here we will focus on spatial autocorrelation, although the same basic issues are raised by temporal autocorrelation. Note that there are two sets of disturbances, one at the level of the transect and one at the level of the reef. If in both cases, of the disturbances are not independent (i.e., if they are spatially correlated), and one fails to take that formally into account, the estimated regression coefficients can be unbiased, but the standard errors will be wrong. To fix this problem, spatial dependence needs to be part of the model.

There are two options. The first is capture as much of the spatial dependence as possible in the predictors. This can be done with variables that would be included on scientific grounds alone, or variables that are functions of location per se. For example, if there are spatial coordinates, functions of these can be included as predictors at both levels. Ideally, the functions will have some instructive scientific interpretation. And there is the added benefit that special software may not be needed (e.g., for a fixed effects model, the usual software for the generalized linear model will suffice). Of course, high multicollinearity can still be an important complication.

The other option is to build the spatial dependence into the variance-covariance matrix of the disturbances. Most commonly this would be done at first level (e.g., at the level of the transect) but in principle, it can be done at either level or even both levels. Within each cell in the matrix is some decreasing function of Euclidian distance; observations farther apart are less alike. An exponential function is one popular approach. The reciprocal of the distance is another. But, perhaps the key point is that the dependence is being treated as a nuisance. There is no desire to extract a scientific story about the role of distance per se.

Whether a researcher decides to capture potential dependence among the disturbances in the structural part of the model or in through one or more disturbance covariance matrices, there are never an guarantees that the dependence is eliminated. There is rarely any theory or past research providing a convincing formulation for how the disturbances are related over and above the effects of the predictors. In the end, the decision about how

best to proceed may legitimately be a matter of convenience. In particular, the choice could depend on the available software and how it performs with different models. And it is probably true in general that representing the spatial dependence in the structural part of the model will lead to more reliable computation.

In the model we estimate below, we address the spatial autocorrelation through the variance-covariance matrix of the disturbances at the first level. We focus on the first level because the the first level units (e.g., sites) can be quite close to one another. The second level units (e.g., reefs), in contrast are often too far apart for spatial autocorrelation to be an issue.

The disturbance variance-covariance matrix of the level one units is $n_j \times n_j$, where n_j is the total number of first level units within second level j. That is, we allow for a different disturbance variance-covariance matrix for each second level unit. We use the reciprocal of the Euclidian distance between these units as the representation of the role of distance. More formally,

$$\epsilon_j = \theta W_j \epsilon_j + \zeta_j, \tag{6}$$

where ζ_j is a conventional regression error term assumed to behave as if drawn independently and at random from some specified distribution. We standardize W so that all of the rows and columns add to 1.0. Thus, θ can range from -1 to +1, and can be thought of as a single "autoregressive" parameter. Positive values imply that the spatial autocorrelation is positive (the usual case) while negative values imply that the spatial autocorrelation is negative. Values near 0 imply that spatial autocorrelation was either never an important problem or that the spatial autocorrelation has already been absorbed by the predictors in the model. Equation **??** denotes that each disturbance at the first level is a) a linear combination of all other disturbances within that level two unit, weighted by the reciprocal of the distance between them and b) a new, conventional disturbance.[2]

4 Empirical Example

We have data from coral reefs along Olango Island in the Phillipines. There are 33 sites with 4 transects in each. There are 14 sites in areas that are protected; fishing is prohibited. There are 19 sites that are in unprotected

[2]The weight assigned in W to a disturbance with itself is 0.

areas; fishing is allowed (and is common). And the fishing can include such very distructive practices such as poisoning fish. The data we analyze is an aggregate or average over 4 equally spaced observations along each transect. Thus, transects are our lowest level, and the second level is sites in the multilevel spatial model.

To keep the exposition simple for now, we use the same formulation shown in equations **??** through **??**. The only difference is that the specific response is the number of different fish species. Whether a reef is protected is coded 1 if the reef is protected and 0 otherwise.[3]

Predictor	Coefficient	Std. Error
Protected (γ_{01})	5.58	3.75
% Sandy Bottom (γ_{10})	-0.18	0.04
Protect X % Sandy (γ_{11})	0.05	0.08
Constant (γ_{00})	27.6	2.10
θ (AR parameter)	.44	–

Table 1: Model for Species Counts Estimated by Augmentation Algorithm (N=132)

Focusing first on the regression coefficients From Table **??**, one can see that if a reef is unprotected there are on average nearly 28 distinct fish species at a site. At these unprotected sites, for each addition percent of the bottom that is sandy, the number of species drops by .18; for every additional 10%, the number of species drops by 1.8. In the protected sites, the number of fish species is greater by 5.58. Finally, in the protected sites, the the negative impact of a sandy bottom on the number of species is a bit less pronounced. The regression coefficient of -.18 is now -.13. For every 10% increase in sandy bottom, the number of species is reduced by 1.3.

The autogregressive parameter is .44, which is of moderate size. There is some meaningful spatial autocorrelation in the residuals. When this is taken into account, we see that the percent of the bottom that is sandy is easily twice the standard error. The impact of a protecting a reef about 1.5 times its standard error, statistically significant at the .10 level for a one-tailed test. The coefficient for the interaction effect is less than its standard error. One

[3]Also, where as before we talked about reefs and transects, now we talk about sites and transects. Earlier, there seemed no point in getting into details of the data collection.

should treat any formal tests with great caution, in part because the data were not collected by random sampling, and there is no compelling model-based sampling alternative. But if one chooses to take formal tests seriously, the interaction effect is certainly be discarded.[4]

Predictor	Coefficient	Std. Error
Protected (γ_{01})	3.14	3.39
% Sandy Bottom (γ_{10})	-0.12	0.04
Protect X Sandy (γ_{11})	0.09	0.08
Constant (γ_{00})	25.52	1.54
θ (AR parameter)	–	–

Table 2: Model for Species Counts Estimated by Ordinary least Squares (N=132)

Table **??** shows the results when ordinary least squares is applied. From a descriptive point of view, the results are about the same. But even though the standard errors are (incorrectly) smaller in the OLS case,[5] the coefficient for protecting the reefs is now smaller than its standard error. Hence the case for protecting the reefs is not nearly a strong.

5 Conclusions

Multilevel models can be constructed for ecological data so that in many cases the model corresponds better than conventional regression models to how the data were generated. Multilevel models can also be more elegant and technically interesting. However, there is a price. The added complexity can lead to a number estimation complications with the result that computer output cannot be taken at face value. This is particularly true if formal statistical inference is an important part of the enterprise. An important recommendation, therefore, is that conventional least squares regression should be applied along with multilevel models. If the results are not dramatically different,

[4] A normal-normal plot of the residuals suggested that we were not terribly far off base assuming normal errors.

[5] The standard errors that look the same are actually smaller in the next decimal place, which is not report in the table.

the multilevel formulation may be preferred. If large differences are found, the multilevel results must be treated with great caution.

There are a number of decisions the researcher has to make as multilevel models are built. When are random effects preferred to fixed effects? If random effects make sense, which particular formulations are most appropriate. When there is concern about spatial dependence, how should that be addressed, in the structural part of the model or in the variance-covariance matrix of the errors? These concerns and others are beyond what one would normally address in a regress analysis. With greater flexibility comes more model specification decisions. And commonly, these decisions will have to be made with very little guidance from subject-matter theory. There is the real risk that arbitrary decisions will be made with the the results that findings will be arbitrary as well.

6 References

Bryk, A.S., and Raudenbush, S.W. (2002) *Hierarchical Linear Models: Applications and Data Analysis Methods*, second edition. Newbury Park, CA: Sage Publications.

Kreft, I. and De Leeuw, J. (1998) *Introduction to Multilevel Modeling.* Newbery Park, CA: Sage Publications.

Ord, J.K. (1975) "Estimate Methods for Models for Spatial Interaction." *Journal of the American Statistical Association* 70: 120-126.