EDA for HLM

Jake Bowers
Katherine W. Drake
Department of Political Science
Center for Political Studies
University of Michigan *

February 24, 2005

*Email for corresponding author: jwbowers@umich.edu

EDA for HLM

> Somewhere along the line in the teaching of statistics in the social sciences, the importance of good judgment got lost amid the minutiae of null hypothesis testing. It is all right, indeed essential, to argue flexibly and in detail for a particular case when you use statistics. Data analysis should not be pointlessly formal. It should make an interesting claim; it should tell a story that an informed audience will care about, and it should do so by intelligent interpretation of appropriate evidence from empirical measurements or observations.
>
> (Abelson, 1995, page 2)

> With neither prior mathematical theory nor intensive prior investigation of the data, throwing half a dozen or more exogenous variables into a regression, probit, or novel maximum-likelihood estimator is pointless. No one knows how they are interrelated, and the high-dimensional parameter space will generate a shimmering pseudo-fit like a bright coat of paint on a boat's rotting hull.
>
> (Achen, 1999, page 26)

> The greatest value of a picture is when it forces us to notice what we never expected to see.
>
> (Tukey, 1977, page vi)

Abstract

It is well know that the most important first step in executing a regression analysis involves producing and interpreting appropriate crosstabulations and plots. When it comes to executing a multilevel regression analysis, however, what is the crosstab or bivariate scatterplot analogue? In this article we propose that within-macro-unit regression coefficients provide the raw material necessary to justify some of the crucial decisions required by a multilevel model. We present several ways to display and assess the patterns within these coefficients. If analysts do such exploratory data analysis *before* estimating a multilevel model, they will be able to (1) ground modeling decisions firmly in characteristics of the data they are analyzing, and (2) have the opportunity to apply their scientific judgement to the details of their data to learn something new about politics.

Any given strategy for summarizing a data set requires many decisions which analysts must justify. Although the number of judgment calls involved in a simple bivariate linear regression is large (See, e.g. Berk, 2004, page 79), the number required for graceful and appropriate use of hierarchical linear regression models (HLMs) is even larger. This short article is about what to do *before* estimating a hierarchical model but *after* research design and data collection.[1] We do not focus here on *diagnostics*, which provide hints to the data analyst about whether and how a set of assumptions that drives a particular modeling strategy may have gone right or wrong.[2] Instead, exploratory data analysis (EDA) is what we do before we estimate a model. The procedures we advocate do not involve any formal hypothesis testing, and thus are not subject to the problems of multiple testing and incorrect coverage probabilities of confidence intervals that can arise when model checking shades into data mining.

Although exhortations to "look at the data" are common in introductory statistics classes, and most analysts have a sense of how to do this in the context of simple linear regression models, it is often not clear how to look at data that are organized in a complex manner. In this short article we demonstrate a few ways to engage in this prior investigation in order to help analysts develop good judgment about complicated data structures in which units (like individual humans or votes) are nested (at least partially) within other units (like political institutions or individual legislators). These data structures are interesting to political scientists for many substantive reasons (e.g. how do institutions influence voters' behavior), but they raise problems for conventional data analyses using, say, OLS or logit models. Growing availability of multilevel datasets and increased awareness of the problems involved in ignoring the structure of these datasets have led many analysts to use multilevel models (represented either with a likelihood function or a posterior density function). We do not offer these techniques as the only possible ways to look at complex data — rather they represent what we've invented and found in various places throughout the literature on these models. We hope that data analysts think of these techniques as starting points or exemplars for their own EDA — as ways to kick off the intensive examination of data before

[1] When we say "hierarchical linear model" we mean a multilevel, mixed effects, random coefficients, or even random effects model. And, what we have to say here applies whether the analyst produces quantities of interest by maximizing a likelihood function or by sampling from a posterior distribution.

[2] We strongly recommend Gelman (2004, 2003) as guides to model checking after estimating a multilevel model.

model specification and estimation.

Since each modeling decision in a multilevel model (or any model) has important implications for the results, it is crucial to have a reasoned basis for them. One reasoned basis is strong prior beliefs. Such beliefs can be "strong" because they rest on tightly argued logical foundations or because they arise from a massive quantity of past findings. Another such reasoned basis is the structure of the data at hand. In this article we do not address the use of prior mathematical theory in model specification and estimation, although understanding the implications of such theory for complicated multilevel phenomenon will be an important task for future scholars (Achen, 2002). Instead, we will present a few ways to produce a reasoned basis for the decisions that lead to a given model specification — without actually running many such models and then evaluating them with diagnostics.[3]

So far, we've portrayed our purpose as helping data analysts avoid embarrassing mistakes in model specification and estimation. However, there is another reason to do EDA that is much more important, from our perspective, than establishing a reasoned basis for the many decisions necessary for any given model. The reason is that EDA can allow analysts to discover new (and unexpected) things about the world. We do not claim that EDA is somehow "model free". It clearly is not, and has not been so since its earliest formal description by Tukey. The very idea that EDA is useful in helping analysts "discover something new" in their data presumes that there are certain patterns that would be expected and other patterns that would be surprising. This is one reason why we are writing this article about EDA in the context of a particular set of models. For example, Gelman (2004, 2003) shows that having an explicit model enables EDA to be yet more powerful. In the end, we think that political scientists would like to learn new things about the political world. The confirmatory approach helps us change our beliefs about what we have learned (e.g. alerting us when there is too little evidence to make what we thought we had learned plausible). The

[3] Running such diagnostics is another fine and complementary approach, although it has some problems with how a frequentist would interpret the p-values produced from such procedures.

exploratory approach, however, comes first, and "finding and revealing the clues" (Tukey, 1977, page 21) is how we guess we might learn something worth testing. That is,

> Exploratory data analysis can never be the whole story, but nothing else can serve as the foundation stone — as the first step. (Tukey, 1977, 3)

Thus, we advocate EDA before estimation because (1) it helps people justify the decisions that are driving the modeling (thus, avoiding hanging an article on a Normality or constant-effects assumption that is untenable in the data), and (2) it enhances the ability of the area expert to learn from the costly data she has collected. It is a shame to spend blood, sweat, tears (and NSF funds) to collect data only to summarize it with a single number "controlling for" many other things — regardless of the fanciness of the technique being used. While we definitely think that confirmation and discovery go hand in hand, we are concerned that discovery has gotten short shrift among political scientists. Much too often we see arguments about important political problems that hang entirely on an $\alpha = .05$ significance test on a single coefficient in a model with an unjustified and implausible probability story. We are concerned that as the modeling machinery becomes more complex, people will spend more time figuring out how to run HLM or WinBUGs and less time learning about their data — and thus, less time learning new things about the world.

The "Standard" Multilevel Model

When most people say "multilevel model" or "hierarchical linear model" they nearly always refer to a particular setup. Here, we explain this archetypal model so that we can know what to expect, and therefore what might be surprising.[4]

Like any statistical model, a multilevel model involves two main components: (1) a structural model which species the functional form of the relationship among the variables, and (2) a stochastic model which uses probability distributions to encode information about how the

[4] This standard model is the one that is hard coded into the lme() command in R and Splus (Pinheiro and Bates, 2000), into the HLM program (Raudenbush and Bryk, 2002), and into most other common multilevel modeling code. Other common programs for estimating this standard model include gllamm for Stata `http://www.gllamm.org` and proc mixed for SAS `http://www.sas.com`.

values of the variables were produced in the world. In the vast majority of uses of multilevel models, the structural model is linear and involves interaction terms, such that for a simple model with two levels, one level-1 explanatory variable, X, and one level-2 explanatory variable, Z,

$$Y_{ij} = \beta_{0j} + \beta_{1j} X_{ij} + e_{ij} \tag{1}$$

$$\beta_{0j} = \gamma_{00} + \gamma_{01} Z_{1j} + u_{0j} \tag{2}$$

$$\beta_{1j} = \gamma_{10} + \gamma_{11} Z_{1j} + u_{1j} \tag{3}$$

Combining the previous three equations, we have:

$$Y_{ij} = \gamma_{00} + \gamma_{01} Z_{1j} + u_{0j} + (X_{ij})(\gamma_{10} + \gamma_{11} Z_{1j} + u_{1j}) + e_{ij} \tag{4}$$

$$= \gamma_{00} + \gamma_{01} Z_{1j} + \gamma_{10} X_{ij} + \gamma_{11} Z_{1j} X_{ij} + (u_{0j} + u_{1j} X_{ij} + e_{ij}), \tag{5}$$

where $j = 1 \ldots J$ for the number of level-2 units and $i = 1 \ldots n_j$ for the number of level-1 units within a given level-2 unit. Equation 5 shows that the influence of X on Y is assumed to be linear across the whole range of X, and that the slope of this line is expected to vary at a constant rate across the range of Z.

One can also write this model combining the structural decisions with the stochastic ones such that:

$$\mathbf{y} | \mathbf{X}, \boldsymbol{\beta}, \boldsymbol{\Sigma}_y \sim N(\mathbf{X}\boldsymbol{\beta}, \boldsymbol{\Sigma}_y) \tag{6}$$

$$\boldsymbol{\beta} | \mathbf{Z}, \boldsymbol{\gamma}, \boldsymbol{\Sigma}_\beta \sim N(\mathbf{Z}\boldsymbol{\gamma}, \boldsymbol{\Sigma}_\beta) \tag{7}$$

This probability model for an HLM is equally standard. Nearly all analysts decide that the values of Y ought to be seen as arising from a process governed by the Normal Distribution. The extra variation in the intercept and slope parameters β_{0j} and β_{1j} is also nearly always understood as arising from a multivariate Normal distribution.[5] Nearly always, $\mathbf{\Sigma}_y$ is assumed to contain σ^2 for the variance of y_{ij}, $\rho\sigma^2$ for the covariance of ys within the same level-2 unit, and 0 otherwise.

The major benefit of this model is that it allows the analyst to specify directly the structure of her dataset in her statistical model, and to estimate relationships taking this into account.[6] However, this structure is assumed to arise from the mixture of Normal probability distributions and a model that is linear in the parameters (i.e. $E(\mathbf{y}|\mathbf{X},\mathbf{Z},\boldsymbol{\beta},\boldsymbol{\gamma}) = \mathbf{X}\boldsymbol{\beta}+\mathbf{Z}\boldsymbol{\gamma}$). This structure implies that

$$\mathbf{y}|\mathbf{X},\boldsymbol{\beta},\mathbf{\Sigma}_y \sim N(\mathbf{X}\mathbf{Z}\boldsymbol{\gamma},\mathbf{\Sigma}_y + \mathbf{X}\mathbf{\Sigma}_\beta\mathbf{X}^T) \tag{8}$$

$$\boldsymbol{\beta}|\mathbf{y},\mathbf{Z},\boldsymbol{\gamma},\mathbf{\Sigma}_\beta \sim N((\mathbf{X}^T\mathbf{\Sigma}_y^{-1}\mathbf{X})^{-1}(\mathbf{X}^T\mathbf{\Sigma}_y^{-1}\mathbf{y} + \mathbf{\Sigma}_\beta^{-1}\mathbf{Z}\boldsymbol{\gamma}),(\mathbf{X}^T\mathbf{\Sigma}_y^{-1}\mathbf{X})^{-1}). \tag{9}$$

That is, that the coefficients in $\boldsymbol{\beta}$ are an average of the within-unit regressions weighted by the variation within those regressions (See, Chapter 10 Gill, 2002, for more explanation of this model in the Bayesian context). Thus, an important benefit of these models occurs when an analyst has many level-2 units with very little information in each unit such that it is possible to estimate coefficients with reasonable standard errors. This benefit only possible, however, if they analyst is prepared to commit to pooling the data in this particular way.

[5]In the Bayesian context, the parameters in $\boldsymbol{\gamma}$ are often given their own probability distributions (called "hyper-priors") which themselves are governed by parameters fixed by the data analyst (See Chapter 15 of Gelman et al., 2004, for more on the Bayesian perspective on these kinds of models) .

[6]For much more discussion of the benefits, weaknesses, details, and implementation of multilevel models see: (Steenbergen and Jones, 2002; Snijders and Bosker, 1999; Kreft and Leeuw, 1998; Longford, 1993; Goldstein, 1999; Pinheiro and Bates, 2000; Raudenbush and Bryk, 2002; Singer and Willett, 2003; McCulloch and Searle, 2001).

In general, there are four ways to estimate the coefficients in (5). The first way is to ignore the multilevel or hierarchical structure of the data, and to estimate this model using OLS. The approach is not appropriate mainly because it involves ignoring the fact that the error is not e_{ij} but $(u_{0j} + u_{1j}X_{ij} + e_{ij})$, which produces heteroskedasticity and serial correlation. More importantly, it is not reasonable to assume that the units inside of one level-2 unit are *exchangeable* with the units in another level-2 unit. Roughly, the fact that the level-1 units are not exchangeable means that it doesn't make sense to treat them all as if they arose from a common probability distribution. This means that one cannot write down a likelihood function as a simple product of identical distributions, and it also implies that the responses y_{ij} cannot be seen as arising independent of the level-2 unit within which they are nested. The very fact that an analyst wants to estimate coefficients for Z suggests that he does not believe the level-1 units are exchangeable. Exchangeability is a weaker property than independence or identical distribution (iid), but it is a precondition for using probability distributions to pool information from disparate observations.[7] If the level-1 units are not plausibly exchangeable then they are not independent, the degrees of freedom available for hypothesis testing will be too large, and the hypothesis tests on the coefficients will be too liberal. One strategy that can help analysts out of these many problems is to include dummy variables for each level-2 unit in their equation; this dramatically increases the number of coefficients estimated. Fixed effects allow for the intercept to be different for each level-2 unit, but for the slopes to be fixed across units.[8]

The second approach is to collapse the data such that the mean of y within each level-2 unit $(\bar{y}_{.j})$ is regressed on the mean of x within each level-2 unit $(\bar{x}_{.j})$ and z. This approach implies that the analyst believes that the level-1 units are identical within level-2 units, and thus the mean provides as much information about y and x as the individual observations. If this

[7]See Gill (2002, Chapter 10) and Gelman et al. (2004) for accessible discussions of exchangeability in Bayesian data analysis.

[8]See Wooldridge (2002, Chapter 10), Davidson and MacKinnon (1993, Chapter 9.10), and Mundlak (1978) for more on fixed effects models.

decision is not correct, then the analyst has needlessly thrown away a lot of information — and, more importantly, no longer has a model of an individual level process. In the end, analysts cannot interpret the coefficients of this aggregate model as if they directly reflected something about causal relationships among attributes of level-1 units without risking the ecological fallacy (Achen and Shively, 1995).

The third approach is to estimate the set of β_{0j} and β_{1j} separately for each level-2 unit. Each within unit regression may satisfy the exchangeability (and iid) assumptions required for such models. This approach has the benefit of allowing the analyst to inspect the set of regression coefficients for heterogeneity (which is what we advocate here as EDA), but it is not often a reasonable strategy for estimating the parameters in (5). The approach implies that the analyst believes that the level-2 units have nothing in common, and it tends to produce a lot of regression estimates that are hard to combine in a principled way. Furthermore, each within unit regression is based on very little information, and thus may produce imprecise estimates.

The multilevel model approach attempts to combine the benefits of the second two approaches: the between units regression and the within units regression. By specifying a probability distribution for the coefficients (i.e. assuming that the coefficients themselves are exchangeable) the analyst can overcome the problem of the non-exchangeability of y (that is $y|\beta$ can be exchangeable and conditionally independent even if y is not, as long as β is exchangeable). If it makes sense to assume that the coefficients are drawn from a common multivariate distribution, then lack of information inside of one level-2 unit can be compensated for by information from the pooled sample overall.

From the standpoint of scientific inference, the within-unit regression (or the fixed effects regression) produces results that only allow the analyst to generalize to those particular units, while the multilevel model — by thinking about the level-2 units as a sample from a population or as having values produced by a general stochastic process — allows for generalization to other level-2 units subject to that same process. But, in order to get the multilevel model off the ground, both the probability distribution of $y|\beta$ and the probability distribution of β must be specified. In the standard multilevel framework, both of these probability distributions are assumed to be Normal. Thus, one of the key tasks of EDA for HLM is to assess whether this specification is reasonable. Abelson (1995) reminds us of the implications of these different approaches by linking the different models to the scientific task of generalizing beyond the boundaries of a specific analysis or experiment when he says:

> In sum, it is fundamental to specify the boundaries of generalization of one's claims. At one extreme, (the fixed effects model), these boundaries are usually painfully narrow. At the other extreme (the random effects model), the price of trying to broaden the boundaries is either to make them somewhat vague — or worse, to lose warrant to claim any generality at all. This is simply the way the research life is. One does not deserve a general result by wishing it. ... *There is no free hunch.*" (page 142, emphasis in original).

Summing up the Assumptions and Decisions

Once a researcher decides that a multilevel model is an appropriate way to summarize some aspects of her data — and to attempt to cast doubt on some null hypothesis using such a summary — she faces a number of decisions.

First are decisions about exactly how the variables relate to each other. Do straight lines do a good job of summarizing the relationship between X, Z, and Y?

Second are decisions about the ways in which the values of the variables were produced. Is the Normal distribution useful for thinking about $p(y|X, Z, \beta, \gamma)$? Is it useful for thinking about $p(\beta|Z, \gamma)$? Are these assumptions reasonable? Or desirable? Even if the slopes and intercepts "look like" they could have been generated by such a process, is it desirable to summarize them in this way? Is it useful to think of the values of β as all emerging from one single distribution? (That is, if the values in $\mathbf{y}$ are only exchangeable conditional on $\boldsymbol{\beta}$, are the values of $\boldsymbol{\beta}$ exchangeable?)

In what follows we confront such questions as we engage in some exploratory data analysis of a multilevel dataset.

An Application: Education and Political Participation

One of the most persistent and important findings in the political participation literature to date is that individuals who have more formal education are more likely to get involved in politics than those who have less. The robustness of this relationship has spurred scholars to question the mechanism by which it arises. That is, discovering that education is a strong predictor of political participation has raised the question about what it is, exactly, that education does to facilitate political participation. The most recent answers to this question provide two mechanisms: education influences political participation via provision of "civic skills" and "civic status". These theoretical mechanisms have received empirical support from findings that individuals who have money, time, or organizational abilities — that is, people who have the "civic skills" to participate in politics — are those more likely to do so. And, individuals who know a lot of other people, particularly politically active people — that is, people who have "civic status" — are also more likely to participate in politics than those who have less extensive and politically involved social networks.[9] Rosenstone and Hansen (1993) summarize the distinction between the operation of skills and status succinctly:

> ...When political participation requires that knowledge and cognitive skills be brought to bear, people with more education are more likely to participate than people with less education. Participation, that is, requires resources that are appropriate to the task.
>
> On the other hand, education also indicates both the likelihood that people will be contacted by political leaders and the likelihood that they will respond. Educated people travel in social circles that make them targets of both direct and indirect mobilization. Politicians and interest groups try to activate people they know personally and professionally. (Rosenstone and Hansen, 1993, page 76)

[9]The most recent and extensive articulation and defense of the "civic skills" and "civic status" points of view are provided by Verba, Schlozman and Brady (1995) and Nie, Junn and Stehlik-Barry (1996). Verba, Schlozman and Brady (1995) coined the term "civic skills." I use the term "civic status" to describe the findings and argument of Nie, Junn and Stehlik-Barry (1996). Huckfeldt (1979) presents an in-depth analysis of how status within social networks can influence political participation.

Our question is whether there are places where the relationship between education and political participation changes depending on the social context — and in particular on the educational context. In places where few people have college degrees (like West Virginia, which has 15% with BA+ education), those who do have BAs ought to find that their educational status places them into more politically relevant social network positions — and thus they ought to be more advantaged by their education than people who are just one college educated person among many (like those living in Massachusetts, with 33% of the population college educated).[10] If education most strongly predicts political involvement in places with few highly educated people — like West Virginia — and only weakly does so in places with many highly educated people — like Massachusetts—, then we would think that education is mainly acting to allocate politically relevant status to people. And, if an additional year of education provides the same boost to participation in all places, regardless of the educational inequality in the context, then we might think that education is mainly about providing individuals the skills necessary to overcome the costs of political participation. Of course, the first pattern of results might also suggest that education is providing skills that are politically relevant in only some places — that somehow political involvement in West Virginia is qualitatively different in terms of skills or status required than in Massachusetts. And, the second pattern might also indicate that politically relevant social status is structured in the same way in all of the places that we examine. However, either set of findings would both contribute to the existing literature and, more importantly, suggest new avenues for in-depth research into how institutions and behavior interact to produce politics.

The idea that individual political behavior is crucially constrained and shaped by social, economic, and political context is not a new idea. For example, Almond and Verba (1963) contend that while psychological accounts of political activity are useful, these explanations

[10] This is the argument of Nie, Junn and Stehlik-Barry (1996), only they are more concerned about changes in the national educational context over time, rather than differences among places in educational context in one moment in time.

assume that the effects will travel across different contexts. This assumption can be a problem, from their perspective, especially since political psychology "is not made exchangeable in terms of political process and performance" across places (Almond and Verba, 1963, page 33). In fact, they found that differing political structures created different patterns of political attitudes. And, of course, the literature linking vote turnout to registration laws is an excellent example of how fruitful such cross-level theories can be (Wolfinger and Rosenstone, 1980; Squire, Wolfinger and Glass, 1987; Highton, 2000, See for example).

Could we shed some light on the mechanism linking education to participation by examining how the link changes depending on the social context in which an individual lives? This is the question we pursue here. The analyses that follow are not meant to be an authoritative investigation of how social context changes the influence of education on political participation. However, they are an example of how one might do some of the intensive investigation of data that we think ought to precede all applications of multilevel models.

We pursue this hunch and represent educational context with data from the 2000 Census on the percentage of the population in a state that has completed a college degree and individual education and political participation with data from the 2000 National Election Study (NES). If we were to represent these ideas as they usually are, we would write the following structural model:

$$\begin{aligned}\text{Participation}_{ij} =& \gamma_{00} + \gamma_{01}\%\text{ College Educated}_j + \gamma_{10}\text{Education}_{ij} \\ & + \gamma_{11}\%\text{ College Educated}_j \cdot \text{Education}_{ij} + \\ & (u_{0j} + u_{1j}\text{Education}_{ij} + e_{ij})\end{aligned} \tag{10}$$

We also proceed as if we believed the standard two-level probability model written in equations (6) and (7).[11]

The NES provided 11 questions about the non-voting political involvement of respondents. We summed the "yes" responses to these questions to create a variable containing the total number of acts engaged in by a respondent in the past year (min=0, max=10, median=1).[12] The NES respondents also reported their educational status, which produced a variable ranging from 0 years of formal education to 17 or more years (min=0, max=17, median=14). Figure 1 shows the positive overall relationship that we would have expected given the weight of past literature. A regression and a non-parametric smoother both summarize this generally positive relationship.

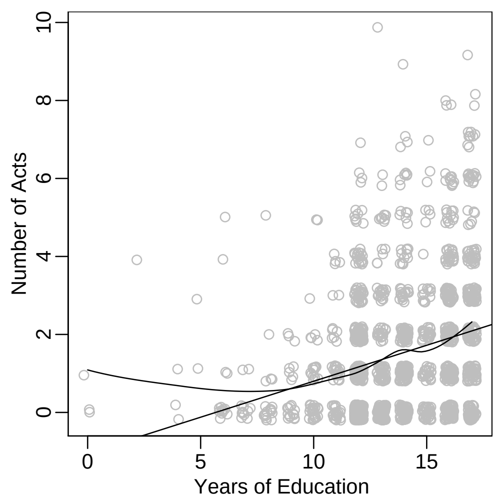

Figure 1: Education and Non-Voting Political Participation in the NES 2000

Note: Points are jittered to show density. The straight line is from OLS. The curved line is from a local non-parametric fit with a nearest neighbor fraction of .5.

[11]We are aware of several additional complications and modifications that might be made to this model due to the clustering within states that arises from the general sampling design of the NES (Stoker and Bowers, 2002) and specifics of the 2000 NES (Bowers and Ensley, 2003). We set these concerns aside for this paper since here the data play an illustrative role in the service of our discussion of methods.

[12]The types of participation summed here are: Did R try to influence vote of others, Did R display button/sticker/sign, Did R go to meetings/rallies etc., Did R do any other campaign work, Did R contribute to candidate, Did R give money to party, Did R give to group for/against candidate, Worked on community issue in last year, Contacted public official to express in last year, Attend community meeting about issue in last year, Taken part in Protest or march in last year.

A Multilevel Dataset

For those readers who are new to these kind of models, we first show what a multilevel dataset looks like. Table 1 shows 10 respondents from two states. The respondent ID and state are in the first two columns, education and participation for each individual are in the next columns. A characteristic of most multilevel datasets is that the macro-level variables are included in the same file with the micro-level variables by repeating the same value for each micro-level unit. Here, we see that the percent college educated is the same for each respondent from the same state.

	State	Years of Education	Participation Acts	% of State Pop $\geq$ BA
1	Alabama	16	0	0.19
2	Alabama	7	0	0.19
3	Alabama	17	1	0.19
5	Alabama	12	1	0.19
6	Alabama	16	1	0.19
1050	New York	12	1	0.27
1051	New York	12	0	0.27
1052	New York	15	2	0.27
1053	New York	14	1	0.27
1054	New York	17	3	0.27
1055	New York	15	0	0.27

Table 1: A Multilevel Dataset: Persons within States

Batches of Lines as Data

The general principle behind our approach to EDA in the context of this proposed multilevel model of education and political participation is to use the within state regressions as data. Since decisions about the probability models (including whether Normality is reasonable and what units can be seen as exchangeable with each other) involve specifying models for the within state regression coefficients, we propose using within state regressions to allow analysts to take a look at the entities that they are deciding about. This means that we estimated a separate regression of participation on education for each of the 46 states that had more than 1 respondent with valid data for education and our participation scale — leaving us with 1539 respondents across 46 states.

We collect the coefficients from the within state regressions into a new dataset, and add the state-level variables. We think that working with within-state regressions as data helps sharpen intuitions about what a given multilevel model is doing. As noted above with the two probability models, $\boldsymbol{\beta}$ depends on some second-level variables, $\mathbf{Z}$, with some effect, $\boldsymbol{\gamma}$. This new data we created has our $\boldsymbol{\beta}$, which we can think of as our dependent variable in the second level, as well as our Z (percent college educated in a given state), which we can think of as our independent variable.

Table 2 shows the first 5 rows from this dataset. The last column contains a variable that we created to categorize the states by the percent of their population aged 25-65 with a college degree.[13] We see that all of the 5 states in this table have positive relationships between education and participation — but that these relationships vary a great deal. In the end, a table is very limited in its ability help us understand the output from 46 regressions. For this reason, the rest of our data analysis will use plots.

	State	$\hat{\beta}_{0j}$	$\hat{\beta}_{1j}$	$\mathrm{SE}_{\beta_{0j}}$	$\mathrm{SE}_{\beta_{1j}}$	% of State Pop $\geq$ BA	State Education Groups
1	Alabama	0.33	0.08	0.86	0.07	0.19	Low
2	Arizona	1.33	0.02	3.58	0.24	0.24	Middle
3	Arkansas	−1.53	0.27	1.80	0.15	0.17	Low
4	California	−1.41	0.19	0.72	0.05	0.27	High
5	Colorado	−5.61	0.51	2.35	0.16	0.33	High

Table 2: A Dataset of Within-State Coefficients

The first question we have is whether there is any appreciable variation in the within state regression coefficients that is worth attempting to explain with a state level variable. Pinheiro and Bates (2000) have developed a plot for just this purpose. For each state, they suggest plotting the point estimate with confidence intervals as line segments on either side. We present such a figure with 95% confidence intervals in figure 2. This figure is immediately useful in showing which states have very little information available for the within state

[13] "Low" refers to the lowest quarter of the states (those with between 15% and 21% of their populations with BAs), "Mid" refers to the middle 50 percent of the states (with between 21% and 27% BAs in their population), and "High" refers to the highest quarter of states (27% to 39%).

regressions (New Mexico has wide confidence intervals and Maine, Nebraska, and Delaware have none at all). This display is also useful for showing variation in the point estimates. In this case, the state level variation in slopes and intercepts is obscured by the confidence interval and point estimate for New Mexico.[14] If this plot did not show appreciable variation across states, then we would guess that the state is not the right contextual unit within which to seek variation in the influence of education on political participation. Looking at this plot excluding New Mexico, however, does suggest a fair amount of variation in the coefficients at the state level, so we will proceed to the next step of our exploratory data analysis.

Once we have a sense that some state level variables can account for some of the variation in the individual level relationships, we can begin to look at our full model. Figure 3 is a first glance at this model, showing the within state regressions and the relevant bivariate scatterplots for each state.[15] The panels are plotted in increasing order of the state level education context from bottom to top, and left to right. The state with the least proportion college educated is West Virginia (with 15% of the population reporting a college degree or higher to the Census in 2000), and the "state" with the most college educated inhabitants is the District of Columbia (with 39%). Each panel shows the scatterplot of the individual level data for that state as gray dots. The black straight lines are the OLS fits — restricted to only plot within the range of the education of the individuals within that state. Thus, the OLS fit for Louisiana runs from 0 to 17+ years of education while the fit for Wisconsin is only plotted from 11 years of education and up. The jagged dark gray lines connect the means of the participation variable at each value of the education variable — this is the

[14] The huge confidence interval for New Mexico can be diminished (and this plot improved) by mean deviating individual level education — using either the grand mean, or the within state mean. This makes the intercept a more meaningful quantity since few states included any respondents with 0 years of education. Demeaning education, or centering it at a meaningful value, like 12 years, also makes the correlation between the slopes and the intercepts much less severe, and enables the standard multilevel likelihood maximization algorithms to work more effectively. We do not work with centered data here, in part to avoid a long discussion about centering data in the multilevel context, and also to emphasize our goal of allowing the analyst to look a displays directly without mental re-scaling of the quantities of theoretical interest.

[15] We have made our lives a bit easier in this article by only focusing on a bivariate relationship at level-1. For multivariate situations we recommend partial regression plots — or the relevant bivariate scatterplot with a predicted line fit to it holding constant the other variables in the equation.

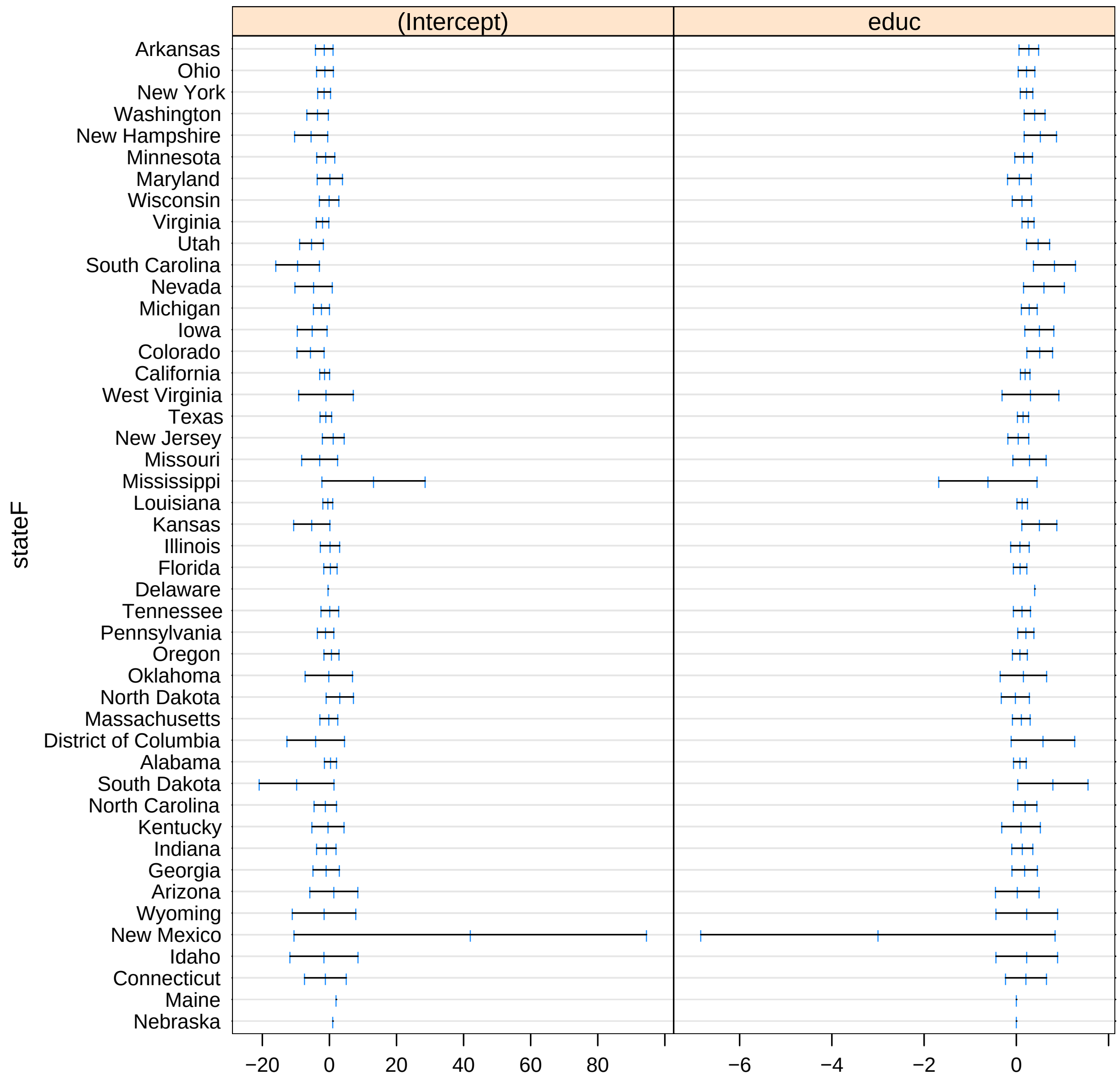

Figure 2: Assessing Between State Variation of Within State Regressions

quantity that is being smoothed with the regression line. We include this line of means to inform our decision about how education ought to relate to participation at the individual level. This plot does not show any clear departures from linearity that cannot be explained by data scarcity.

Does education appear to affect participation more in states where there are few college

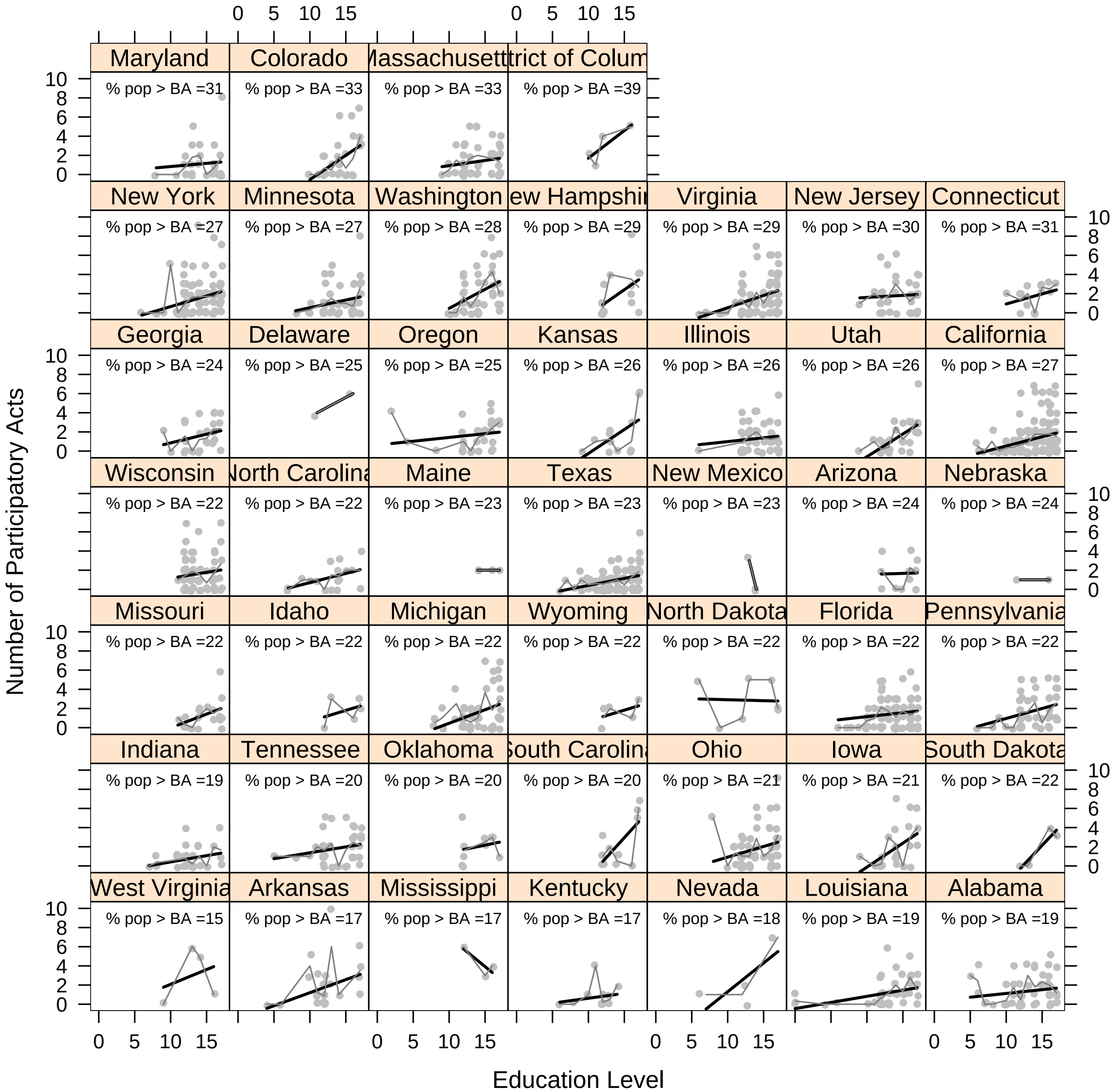

Figure 3: Participation versus Education within States

Notes: The straight lines in each panel are OLS fits. The broken lines connect the mean level of participation at each value of education. Points are slightly jittered to show density.

educated inhabitants? The slopes of the OLS fits on the bottom two rows of this plot do not appear systematically steeper than those in the top few rows. However, it appears that the slopes vary even across states more than a few observations.[16]

[16]We recognize that we have not taken other confounding factors into account. As noted at the beginning of the article, these are just preliminary looks at the data and relationships between our variables of interest.

Although we always begin with a plot like this in our own exploratory data analysis, we have discovered that it would be nice to have a more concise depiction of the relationships in our data. One way to deal with the problem of having too many plots is to plot the within state lines for each state in the groups we created for percent college educated.

Figure 4 displays such a plot for this model and this dataset. Each plot collects the regression lines for states with a given level of college educated population — low, middle and high. The thick black lines in each panel have the mean intercept and mean slope of all of the lines in a given panel — weighted by the number of observations used in each within state regression.[17] If the educational context of states changes the way that an individual's education influences her political participation, then we ought to see systematic differences in the slopes of the lines (either the thick ones or the thin ones) between the panels of this figure. This kind of pattern is not evident here — even the weighted mean lines appear to have nearly the same slope. This picture makes us rethink our initial ideas about state level educational context as structuring or conditioning the way that individual education drives individual political participation. Of course, it is possible that a presidential election year might drive participation equally across the country — but it is also possible that primaries and efforts to maximize electoral votes would cause different states to receive much more or much less mobilizing efforts on the parts of the national parties. In any case, this plot casts doubt on our model in (10) in so far as the influence of state level education is concerned.

As a final assessment of the relationship between the βs and the state level variable, we regress our coefficients (both the intercept and the slope of education) on our state level variable, using the standard errors of the coefficients as weights in our regression. Figure 5 plots the slopes against percent college educated in the left panel, and plots the intercepts against percent college educated in the right panel using gray dots for the point estimates and gray lines for rough 95% confidence intervals. The black straight lines are the weighted

[17] One could also do this kind of plot grouping local, non-parametric lines according to levels of a macro-level variable if the analyst were concerned to justify decisions about linearity across the whole range of both X and Z.

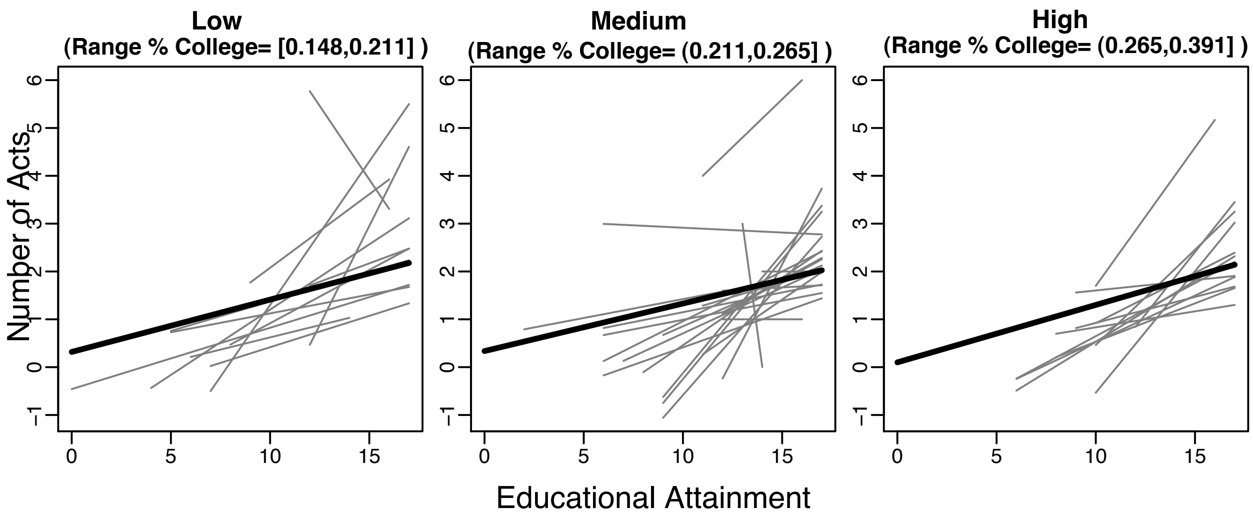

Figure 4: Within State Regressions Grouped by Levels of State-Level Educational Context

OLS fit of the coefficients on the state level variable. The curved black lines are the weighted local non-parametric fit of the coefficients on the state level variable. A thin gray line is drawn at 0 to aid in visual assessments of the flatness of the slopes. States whose residuals were in the top decile are labeled.

If the plot labeled "Slopes" shows a relationship between the coefficients and the percent college educated in a state, then we are correct in assuming that the educational context of a state, measured by the percent of college educated, changes the effect of education on the individual. In this case, however, it is hard to believe that the slope in the left panel is anything but flat ($\hat{\gamma}_{11} = .56$). The non-parametric fit does not show any appreciable non-linearities in this relationship.

The plot labeled "Intercepts" suggests that states with more people who are college educated also tend to be states that tend to have slightly lower amounts of political participation among the less educated members of their populations ($\hat{\gamma}_{01} = -17.15$). Removing Mississippi and New Mexico from the regression flattens the slope to $\hat{\gamma}_{01} = -7.1$. In the end, this slope is still too flat taking into account the scale and range of the state level education variable to give us much confidence in the model we specified in (10).

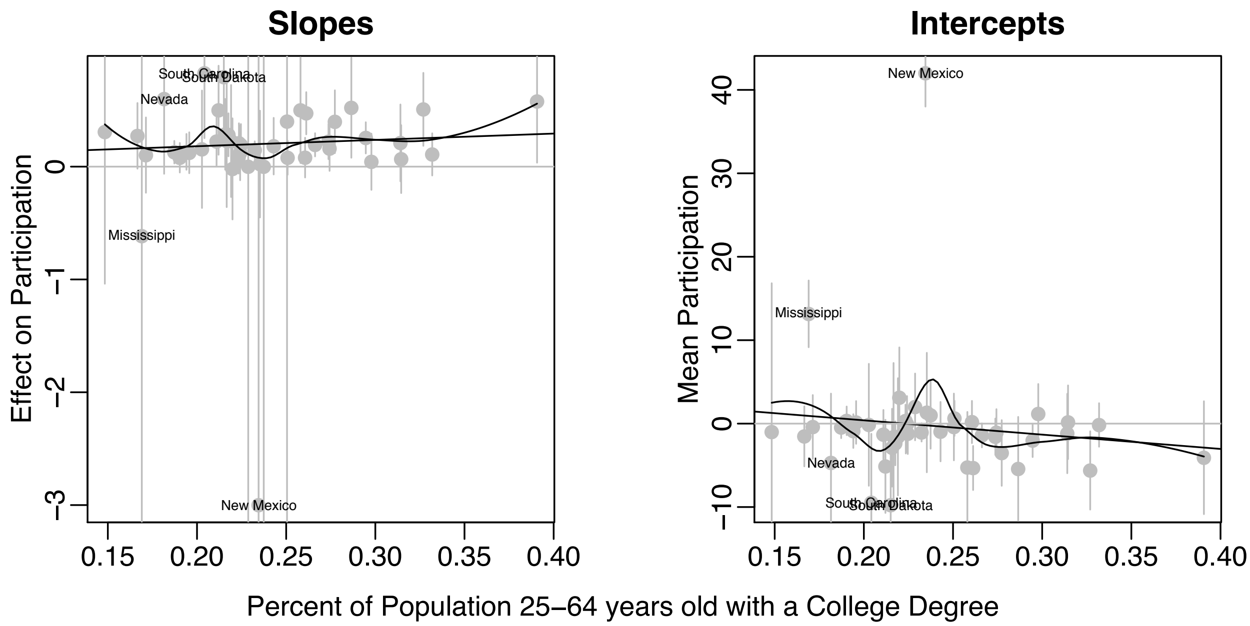

Figure 5: Modeling the Within State Coefficients as a Function of The Percent College Educated in Each State

Notes: States with the largest absolute residuals (in the top 10%) are labeled. For both plots the outlying states are the same: they are Mississippi, Nevada, New Mexico, South Carolina, and South Dakota. The straight lines are OLS fits weighted by the standard errors of the coefficients plotted in gray (with ± 2 SE confidence intervals plotted as vertical line segments). The curvy black line represents a locally quadratic non-parametric fit to the data with a nearest neighbor fraction of .5 also weighted by the standard errors of the coefficients.

This graph confirms our earlier suspicions that have been building since Figure 3 that percent of college educated does not have much to do with the effect of education on an individual within the state. These plots are especially useful, though, as they help us to separate out the intercept and the slope of education. Earlier, we were looking at the effect of state context on both of these with the same graph. Graphing the regressions of the intercept and the slope on our second level variable separately allows us to make intelligent arguments about what exactly is being affected by our second level variables. It is possible that our slope is constant across different contexts but that our intercept may vary, or vice versa.

These plots, like the others, also highlight two states whose slopes diverge from the mass of the others. These two states (Mississippi and New Mexico) have negative slopes. A few other states (Maine, Arizona, Nebraska) have flat slopes for this relationship. It is probable that these unexpected estimates are due to the small sample sizes within those states as seen in Figure 3. Imagine, however, that sample sizes within each of these states were quite large. In that case, plots like these would have shown us something unexpected — places where the well established relationship between education and political participation did not hold! Thus, together, Figures 3, 4, and 5 have provided us with information that:

1. there is state level heterogeneity in this relationship such that thinking about the individual level units as exchangeable without conditioning on state (or perhaps town?) might be a mistake,
2. the structural relationships between education and political participation is reasonably linear within states,
3. the proportion of a state having a college degree does not seem to influence the relationship between education and political participation much, if at all — and a lack of a linear relationship is not hiding a strong non-linear relationship, and
4. in our imaginary world of lots of data per state, we would have discovered something unexpected — which might lead us to change both our estimation and theoretical tactics.

Assessing the Probability Model

The standard model presented above in (6) requires that the values of the political participation variable be produced by a normal distribution (conditional on education). A series of simple histograms of participation by the quartiles of education, shown in Figure 6, shows that a likelihood based on a single normal distribution is probably not plausible. After looking at these plots, we would be tempted to use a Poisson distribution as our model of counts of acts of political participation.[18]

[18]Once we begin to consider distributions other than the Normal for our likelihood, the algorithms for estimating such hierarchical generalized linear models (HGLMs) or generalized linear mixed models (GLMMs)

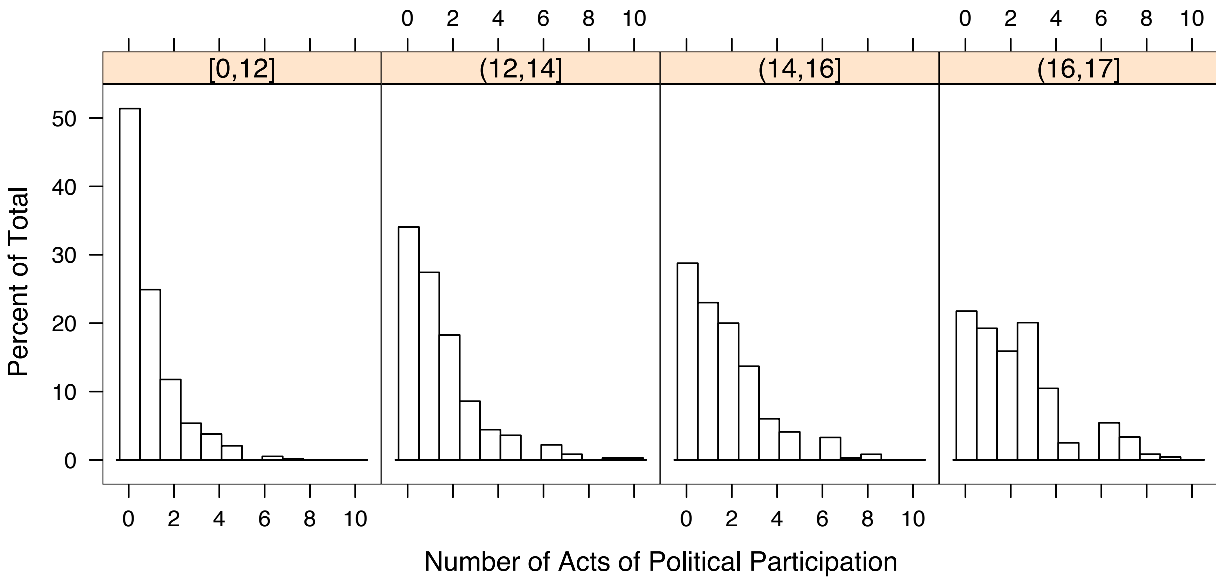

Figure 6: The Distribution of Participation Conditional on Education

The second decision about probability distributions has to do with the βs. In equation (7) we followed standard practice and specified that they were to be jointly Normally distributed. Is this decision plausible? As a rough first glance we produced plots of the quantiles of a standard normal distribution against the within state intercepts and slopes (ordered by their observed quantile). Such quantile-quantile plots (or qq-plots), shown here in Figure 7, are a common device for assessing the match between observed data and a univariate probability distribution. Since our assumption is about a multivariate distribution, these plots are not a direct assessment of our assumption, but they are a rough and quick (and easy to interpret) look nonetheless.

In each case, the middle of the distribution of our coefficients coincides rather well with what would be expected were they drawn from a normal distribution. And in each case, the tails of the observed distributions are wider. Inspection of density plots of these coefficients

become much slower and have difficulties with convergence and accuracy; even with a Normal mixture (or prior) distribution, closed-form solutions to the likelihood maximization problem are less likely to be found, and the software must use approximations to the integrals that arise. In such situations, the Bayesian-inspired algorithms in software like MCMCpack (Martin and Quinn, 2005) and WinBUGS can take just as long to run (with their own concerns about convergence), but have the potential to do a better job with the numerical integration than the likelihood based approximations.

(not shown here) suggests that these heavy tails are mainly due to those few states which, depending on one's perspective, were (1) providing noisy results from too few within state observations or (2) were suggesting unexpected patterns of education and political participation. One might decide to interpret these plots as suggesting that perhaps a t-distribution or some other wide-tailed distribution would be a better way to specify the prior or mixture distribution for this multilevel model. Another possibility would be to make this probability distribution more flexible (perhaps using a Beta distribution), in order to capture the division of states into those where education seems to matter greatly and those where it does not.[19]

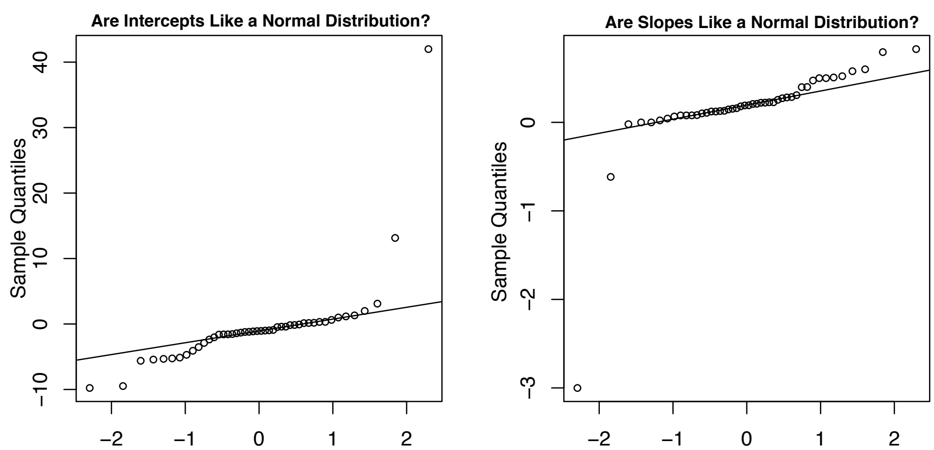

Figure 7: Normal Q-Q Plots of the Within State Regression Coefficients

The standard specification has the within state coefficients generated according to a single multivariate normal probability distribution governed by $\mathbf{Z}\boldsymbol{\gamma}$ for the means, and with the variance-covariance matrix $\boldsymbol{\Sigma}_{\beta}$. This matrix allows for correlations between the coefficients, such that, for example, states with very high levels of participation among people with zero education could also tend, systematically, to be those states where the influence of education on participation is high.

EDA makes the most sense when the data are compared to an explicit model (Gelman,

[19]Our discussion here presumes that if these assessments cast doubt on the Normal-normal model established in most multilevel modeling code, that the analyst will use Markov Chain Monte-Carlo techniques in order to estimate her quantities of interest.

2004). In this case, what would the model look like? In our case, the multivariate normal distribution we care about is bivariate, and so may be thought of as a hill with the slopes defining one axis, the intercepts defining the other axis, and the probability density at each (slope,intercept) point determining the height of the hill at that point. Although we could graph this hill in 3D, we find it more useful to use a contour plot to represent such a density. A contour plot positions the viewer right on top of the hill, looking down. Figure 8 shows two different bivariate normal distributions generated using the means and variances of the within-state regression coefficients that we've already estimated. In the left panel, we are looking down on a perfectly round hill that summarizes the scatter of gray points underneath (the points are 1000 draws from the distribution drawn with the contour lines). If our actual coefficients were to look like this, then we would feel confident in our decision in (7) and would also know that $\mathbf{\Sigma}_\beta$ in that equation has 0s on the off-diagonal elements. The right panel shows the density for a bivariate normal with the same means, variances, *and covariance* as our within-state regression coefficients (i.e. the covariance between the slopes and the intercepts). If we were to see a picture like this, we would again feel confident in using a multivariate normal distribution to summarize the variation in the within-state slopes and intercepts over the states. This picture also tells us that we should not fix the off-diagnoal elements of $\mathbf{\Sigma}_\beta$ to be zero, and that states where individual level education strongly predicts political participation also tend to be those states where the participation level among those with zero education is estimated to be extremely low.

Now that we have made clear what we expect to see based on the standard model, we are ready to look at what the data actually look like. Figure 9 plots the joint distributions of our slopes and intercepts — excluding Mississippi and New Mexico. The left panel (labeled "Theoretical Density") shows the scatterplot overlaid with what we might have expected to see if a bivariate normal distribution were the true density. The right panel (labeled "Estimated Density") is the same scatterplot, only this time overlaid with a non-parametric estimate of the density function. While the univariate plots suggested two long tailed distributions, this

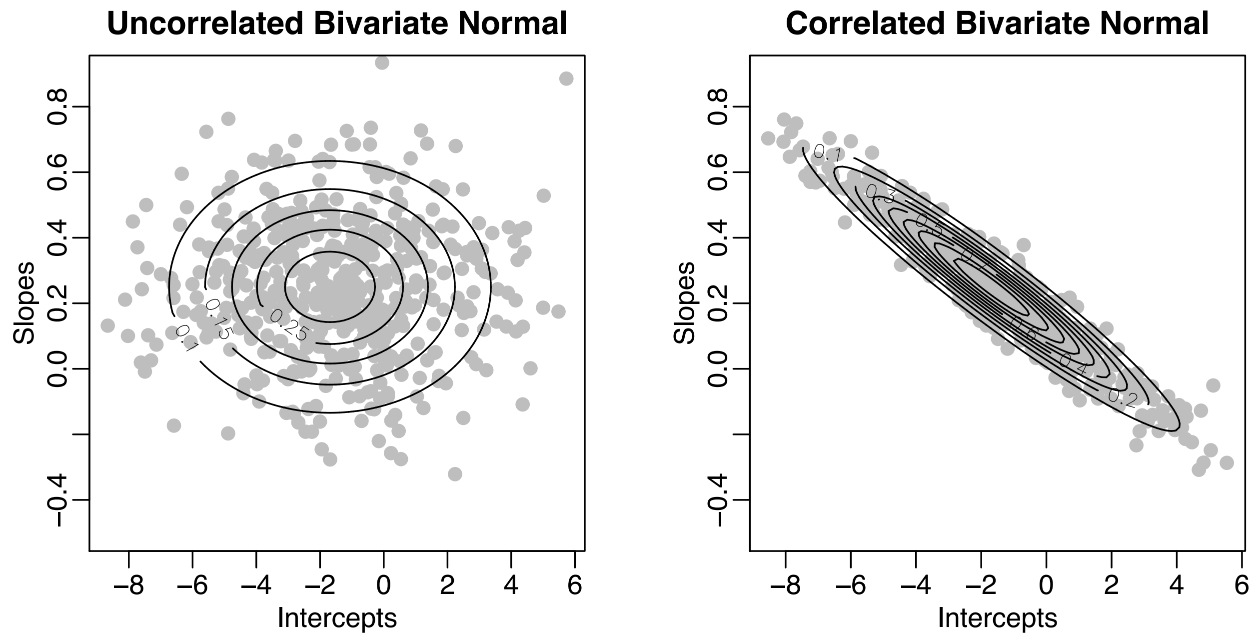

Figure 8: Hypothetical Bivariate Normal Distributions

Note: Each plot contains 500 draws from a bivariate normal distribution with mean equal to the mean of the within-state regression coefficients (excluding New Mexico and Mississippi) and variances equal to the variance of these coefficients. The "Correlated Bivariate Normal" plot uses the observed covariance between the within-state slopes and within-state intercepts to generate the density.

plot suggests that the two long tails coincide, such that it might not be reasonable to see these βs as exchangeable without conditioning on which cluster of states they arise from — the small "mound" in the middle of the plot or the large "hill" on the bottom and right of the plot.

Conclusion

Have we learned anything new about political participation from this exercise? Putting aside the pared down nature of our application, we can say that we have identified a few states where our model does not seem to hold (possibly just due to a paucity of data). It does not appear that state level percent college educated has much to do with the bivariate relationship between an individual's education and her political participation. If this result appears in more complete models, and in other datasets, at other geographic levels of analysis, it will fruitfully complicate our understanding about how education allocates political relevant positions in social networks in our country.

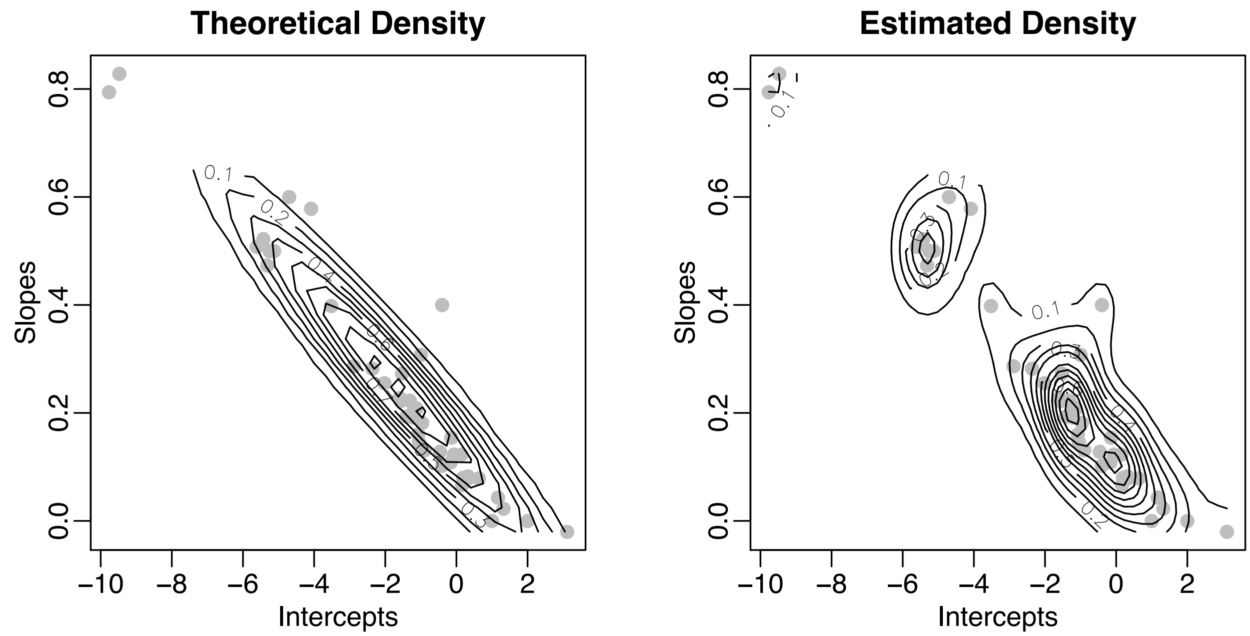

Figure 9: Bivariate Density Plot of the Within State Regression Coefficients

Notes: The contour lines in the right panel show the bivariate Normal density calculated at the points of the observed data — using the means and covariance matrix of the slopes and intercepts as the parameters. The contour lines in the left panel represent a non-parametric Normal kernel density with bandwidth 2 for the Intercepts and .25 for the Slopes.

Notice that we have learned something from our data without a single hypothesis test or distributional assumption. Our use of linear regression here was mainly as a data smoother or as data about which we had modeling decisions to make. We began this paper with a quote from Abelson (1995) about what good data analysis should do. We think that scholars can do a lot of good multilevel data analysis in Abelson's sense by making plots like those we've produced here.

Were we able to justify the decisions implied by the standard multilevel model? Not very well. If we were to estimate a multilevel model of this phenomenon, after engaging in this kind of exploratory data analysis, we would be tempted to use a Poisson distribution for our likelihood and perhaps a mixture of two Normal distributions for our prior. And, we would check our model's predictions against our observed data following the suggestions in (Gelman, 2003, 2004).

Plots of within-unit fits are a common feature in the textbooks on multilevel modeling that we listed on page 7; they are quite often useful in motivating what it is that such models actually do (i.e. produce a coefficient estimate that is in some way an average of the within unit coefficients weighted by the amount of information contributed by each unit). However, such teaching devices are not commonly found in the working data analyst's tool kit, which is strange given that at least two of those textbooks explicitly include chapters on visualizing and describing multilevel data (Pinheiro and Bates, 2000; Singer and Willett, 2003). We think that looking at the within state regression coefficients as data in their own right is an important aspect of EDA and belongs in both textbooks and journal articles.

In general, we think that taking a close look at the micro-level relationships within macro-units enables analysts to demystify what can often seem like the black box of multilevel models. In addition, inspecting such data displays calls on the substantive knowledge and judgment of the scholar. And, good judgment about modeling decisions is exactly what exploratory data analysis is about. If in pursuit of such good judgment unexpected patterns emerge, so much the better.

Of course, the plots and ideas we've presented here are not meant to be a particular set of techniques to be applied everywhere. They are meant to stimulate scholars to develop their own data displays, and we've cribbed many of them from the scholars we've cited. The basic idea is that multilevel models are partly about understanding patterns in $\mathbf{y}$ as a function of $\mathbf{X}$, but also about understanding patterns in $\boldsymbol{\beta}$. Scholars are used to exploring the relationships within their datasets using crosstabulations and bivariate scatterplots, but are perhaps not as used to exploring the relationships between their coefficients. In this article we've tried to suggest and present a few different ways that data analysts can make their varying coefficients easier to handle.

> A basic problem about any body of data is to make it more easily and effectively handleable by minds —our minds, her mind, his mind. To this general end:
>
> - anything that makes a simpler description possible makes the description more easily handleable.

- anything that looks below the previously describe surface makes the description more effective.

(Tukey, 1977, page v)

References

Abelson, Robert. 1995. *Statistics as Principled Argument*. New York: Lawrence Erlbaum.

Achen, Christopher H. 1999. "Warren Miller and the Future of Political Data Analysis." *Political Analysis* 8:142–146.

Achen, Christopher H. 2002. "Toward A New Political Methodology: Microfoundations and ART." *Annual Review of Political Science* 5:423–450.

Achen, Christopher H. and W. Phillips Shively. 1995. *Cross-Level Inference*. Chicago: University of Chicago Press.

Almond, Gabriel Abraham and Sidney Verba. 1963. *The Civic Culture; Political Attitudes and Democracy in Five Nations*. Princeton, N.J.: Princeton University Press.

Berk, Richard. 2004. *Regression Analysis: A Constructive Critique*. Sage.

Bowers, Jake and Michael Ensley. 2003. Issues in Analyzing Data from the Dual-Mode 2000 American National Election Study. Technical Report No. 64. National Election Studies.

Davidson, Russell and James G. MacKinnon. 1993. *Estimation and inference in econometrics*. New York: Oxford University Press.

Gelman, Andrew. 2003. "A Bayesian formulation of exploratory data analysis and goodness-of-sit testing." *International Statistical Review* 71:369–382.

Gelman, Andrew. 2004. "Exploratory Data Analysis for Complex Models (with Discussion by Andreas Buja and Rejoinder)." *Journal of Computational and Graphical Statistics* 13:755–787.

Gelman, Andrew, John B. Carlin, Hal S. Stern and Donald B. Rubin. 2004. *Bayesian Data Analysis*. 2nd ed. ed. Chapman and Hall/CRC.

Gill, Jeff. 2002. *Bayesian Methods: A Social and Behavioral Sciences Approach.* Chapman and Hall/CRC.

Goldstein, H. 1999. *Multilevel Statistical Models.* London: Edward Arnold. Internet Edition, available at http://www.arnoldpublishers.com/support/goldstein.htm.

Highton, Ben. 2000. "Residential Mobility, Community Mobility, and Electoral Participation." *Political Behavior* 22:109–120.

Huckfeldt, R. Robert. 1979. "Political Participation and the Neighborhood Social Context." *American Journal of Political Science* 23:579–592.

Kreft, I. and J. De Leeuw. 1998. *Introducing Multilevel Modeling.* London, UK: Sage.

Longford, N. T. 1993. *Random Coefficient Models.* Oxford: Clarendon Press.

Martin, Andrew D. and Kevin M. Quinn. 2005. "Markov Chain Monte Carlo Package (MCMCpack)." `http://mcmcpack.wustl.edu/doku.php`.

McCulloch, Charles E. and Shayle R. Searle. 2001. *Generalized, Linear, and Mixed Models.* New York: John Wiley and Sons.

Mundlak, Yair. 1978. "On the Pooling of Time Series and Cross Section Data." *Econometrica* 46:69–85.

Nie, Norman, Jane Junn and Kenneth Stehlik-Barry. 1996. *Education and Democratic Citizenship in America.* Chicago: University of Chicago Press.

Pinheiro, José C. and Douglas M. Bates. 2000. *Mixed-Effects Models in S and S-PLUS.* Springer-Verlag.

Raudenbush, Stephen W. and Anthony S. Bryk. 2002. *Hierarchical Linear Models: Applications and Data Analysis Methods.* 2nd ed. Thousand Oaks, CA: Sage Publications.

Rosenstone, Steven and John Mark Hansen. 1993. *Mobilization, Participation and Democracy in America*. MacMillan Publishing.

Singer, Judith D. and John B. Willett. 2003. *Applied Longitudinal Data Analysis: Modeling Change and Event Occurrence*. Oxford University Press.

Snijders, T. and R. Bosker. 1999. *Multilevel Modeling: An Introduction to Basic and Advanced Multilevel Modeling*. London, UK: Sage.

Squire, Peverill, Raymond E. Wolfinger and David P. Glass. 1987. "Residential Mobility and Voter Turnout." *The American Political Science Review* 81:45–66.

Steenbergen, Marco R. and Bradford S. Jones. 2002. "Modeling Multilevel Data Structures." *American Journal of Political Science* 46:218–237.

Stoker, Laura and Jake Bowers. 2002. "Designing Multi-Level Studies: Sampling Voters and Electoral Contexts." *Electoral Studies* 21:235–267.

Tukey, John W. 1977. *Exploratory Data Analysis*. Addison-Wesley.

Verba, Sidney, Kay Lehman Schlozman and Henry Brady. 1995. *Voice and Equality: Civic Voluntarism in American Politics*. Cambridge: Harvard University Press.

Wolfinger, Raymond E. and Steven J. Rosenstone. 1980. *Who Votes?* New Haven: Yale University Press.

Wooldridge, Jeffrey M. 2002. *Econometric analysis of cross section and panel data*. Cambridge, Mass.: MIT Press.