The Sociological Methodologist

Newsletter of the Methodology Section, American Sociological Association
Summer 2007

Chair: Thomas A. DiPrete *Chair-Elect:* Ross M. Stolzenberg *Secretary/Treasurer:* Guang Guo
Council: Kenneth C. Land, Pamela Paxton, Aimee Dechter, Trond Petersen, Ross L. Matsueda, and Douglas Heckathorn
Newsletter Editor: Lawrence E. Raffalovich *Assistant Editor:* David P. Armstrong

Contents

From the Chair

by Thomas A. DiPrete
tad61@columbia.edu

The methodology section is sponsoring three sessions at the 2007 ASA meetings in New York. I am very pleased that Charles Tilly has agreed to be our featured speaker this year. He will give the O.D. Duncan Honorary Lecture on the topic of "Describing, Measuring, and Explaining Struggle." Chuck is primarily known for his substantive and theoretical research, but he has had a keen interest in methodology throughout his career and has made many important methodological contributions to historical sociology and in particular the study of contentious politics. The Duncan Lecture will take place on Tuesday, August 14th at 12:30pm. In addition, the section will sponsor two additional sessions. The first is a special session on Latent Growth Curve Models and has been organized by Pam Paxton, and the second session has been organized as an open session by Guang Guo. Both sessions will take place on the 14th of August as will the section's business meeting, which will start at 9:30, immediately after the meeting of the methodology section council.

A second important news item concerns the awarding of section awards at the methodology section reception, which will take place at the Hilton on Monday evening, August 13th, at 6:30. This year, the awards committee chose Stanley Lieberson and Arthur Stinchcombe as the co-winners of the Lazarsfeld Award. These awards honor two lifetimes of contributions to sociological methodology, which have given *Making It Count*, "Small N's and Big Conclusions," "Einstein, Renoir, and Greeley: Some Thoughts about Evidence in Sociology," *Theoretical Methods in Social History*, *The Logic of Social Research*, and many other insightful discussions and commentary on nearly every major methodological orientation that can be found in our discipline. In addition, we will honor Mitch Duneier as the second winner of the Goodman Award for his many contributions to the methodology of urban ethnography as represented in *Sidewalk, Sim's Table,* and many other publications. Finally, the section is pleased to announce that Paul von Hippel has been awarded this year's Clifford Clogg award for his paper, "Regression With Missing Y's: An Improved Strategy for Analyzing Multiply Imputed Data," which is forthcoming in the 2007 volume of *Sociological Methodology.* Please check the ASA program schedule for the locations of the business meeting, our two invited sessions, the Duncan lecture by Chuck Tilly (all of which take place on Tuesday), and the methodology section reception, which takes place on Monday evening. I look forward to seeing a large attendance by section members and other fellow sociologists at our events in New York.

This is my last year as chair of the methodology section. Immediately after the ASA meetings in August, Rafe Stolzenberg, who currently is chair-elect, will take over as chair, and Tim Liao will be the new chair-elect. I want to thank all the section members who voted in the recent election, and I wish both Rafe and Tim well in their leadership of the section during the next couple of years.

Computation of Fixed Effects Models

by David J. Armor
darmor@gmu.edu

This research note concerns a computational problem discovered in the course of trying to replicate a fairly well-known paper on black peer effects by Hanushek, Kain, and Rivkin (HKR). Although this paper has not yet been published, it has been circulated in various versions between 2002 and 2006. It has been widely cited because of its potentially important policy implications for school segregation and desegregation.

Applying a sophisticated fixed effects model to longitudinal Texas state testing data, the HKR papers report a strong negative effect of black peers on black achievement. The effect is considerably stronger than that found by most prior studies of black peer effects, and because of its more rigorous statistical analysis, the study has been used to defend continuing school desegregation plans such as those in Seattle and Jefferson County, Kentucky, just reviewed by the Supreme Court. Fortunately, the achievement issue did not play a role in the decision just issued by the Supreme Court.

The problem is that the paper relied on incorrect computational procedures (using Stata software) to estimate two or more large fixed effect sets, and therefore the coefficients in all versions of the paper circulated so far are incorrect. The purpose of this research note is to describe what went wrong in the computational procedures so that others might avoid the same problem when doing fixed effect models using Stata or perhaps other packages.[1] Secondarily, the note discusses some of the rationale for choosing fixed effect models (versus random effects) which leads to the computational issues.

A Note on Fixed Effect Models

Some of the discussions about fixed vs. random effect models with panel (longitudinal) data are unclear regarding the reasons for choosing one or the other approach for particular variables. The variables in question include the unit of analysis (such as subjects when persons are measured over time), time (e.g., years), and natural groupings of the units of analysis (e.g., schools in education studies or firms in economic studies). In particular, some sources (particularly manuals for statistical software) suggest or imply, incorrectly, that a variable should be treated as a random effect if its values are a (random) sample from some population.

This advice leads some analysts to chose random effects for units of analysis (subjects) and fixed effects for the time variable; grouping variables are also sometimes treated as random effects if there are large numbers of them (e.g., schools) or fixed if small numbers (e.g., states). Neither sampling nor size considerations are the proper bases for deciding on fixed vs. random effects.

Rather, the decision should be based on whether or not the variable in question is correlated with other predictors (independent variables) in the model, whether measured or unmeasured. If one believes that the variable is not correlated with other predictors of the main dependent variable, then random effects is the preferred model. But if the alternative assumption is true, which is more typical in social science applications, then fixed effects is the more appropriate choice.

Generally, when one is not certain about the two assumptions, fixed effects is the more conservative choice. Another way of expressing the advantage of fixed effects models, particularly when used for

[1] I have not used SPSS or SAS for panel data, so do not know whether the problem can arise using those packages.

subjects, is that it removes the effect of any time-invariant characteristic such as gender, race, or relatively stable socioeconomic characteristics (such as parent education), some of which may be unmeasured in the data at hand. A similar advantage accrues to using fixed effects for grouping variables such as schools or firms, so that the effect of any time-invariant characteristic of the group entity is also removed. A relatively clear exposition of fixed vs. random effects models can be found in Johnston and DiNardo (1997).

The HKR Model

The HKR papers model academic achievement test scores A_{isg} where i indexes subjects, s school, and g grade level. Achievement is modeled separately for black and white students. Two types of models are estimated, a gain model and a lag model as described by the following two equations:

(1) $A_{isg} - A_{is(g-1)} = X_{isg}B + S_{sg}C + P_{sg}D + u_{isg}$ (Gain model)

(2) $A_{isg} = \theta A_{is(g-1)} + X_{isg}B^* + S_{sg} C^* + P_{sg} D^* + u_{isg}$ (Lag model)

Achievement is measured as percent of math or reading questions answered correctly, so gain scores are meaningful. The coefficients B for a vector of student characteristics X, the coefficients C for school characteristics S, and the coefficients D for peer effects P are assumed to be the same from one grade or year to the next (e.g., parent SES exerts the same effect on annual achievement gains in grade 3 as in grade 4, etc.) in this model. The Texas analysis includes three cohorts of students, and their equations show a superscript c for cohort which is omitted here for ease of notation. [2] All cohorts are combined in the analysis, but as made clear below dummies are entered for year (or cohort).

HKR attempt to reduce the residual or error variance u_{isg} by controlling for several different fixed effect variables. The Texas data has student test scores over several grades (4 to 7), several years (1993 to 1997), thousands of schools, and hundreds of thousands of students. The data are "stacked" so that an observation is a single gain score (or a single score and a lag score pair) within in a given grade, school, and cohort. Thus fixed effects can be estimated for various combinations of students, grades, schools, attendance zones, and years (cohorts). Specifically, four fixed effect combinations are removed in most of their 2004 and 2006 models: student, year by grade, school by grade, and attendance zone by year.

To simplify the discussion, only the math gain score models for black students will be considered here, and the attendance zone by year fixed effect is dropped (it appears to be the least important of the fixed effect variables). HKR get the following results for black student math scores in their 2004 paper (see Figure 1, next page).

Note that with neither of the large fixed effects removed, there is a small negative effect of black peers. After removing student fixed effects only, there is a strong positive effect of black peers (surprising), but after removing both student and school by grade fixed effects, the sign reverses so there is now a very strong negative effect of the same magnitude. Thus, according to the full fixed effects model, the math scores of a black student switching from an all white school to an all black school would decline by almost one third of a standard deviation in a single year. Unfortunately, the third coefficient (and possibly the second) is incorrect because of computational problems.

[2] By cohort, we mean a group of students who start grade 3 in a given year and then progress through later grades; e.g., one cohort would be students who start 3rd grade in 2001; a second cohort would be students who start 3rd grade in 2002.

Figure 1.

HKR Models:	Black peer effect[a]	Large fixed effects removed[b]
Gain score (1)	-.0007 **	None
Gain score (2)	+.0029 ***	Student
Gain score (3)	-.0030 **	Student plus school by grade

** $p \leq .01$ *** $p \leq .001$

[a] Change in black gain scores (in standard deviations) for a 1% increase in percentage of black students

[b] All models include grade-by-year dummies, school change indicators, and free lunch indicators.

Computational Procedures

Computational challenges arise whenever there are large numbers of values for a fixed effect variable like students and school by grade in the HKR analysis. When there are a limited number of values for a fixed effect variable, fixed effects are estimated by simply including the values as dummy variables in the regression. In the Texas data, however, there are over 200,000 students per cohort in more than 3000 schools. With most desk top computers, even those with 2 gigabytes of RAM, it is not possible to invert matrices that include 3000 dummy variables, much less 200,000. Other methods must be used.

The alternative to dummy variables is "demeaning." This step transforms all variables in the regression by removing each regular variable's mean on all values of the fixed effect variable. For example, if the values of the fixed effect variable s (for school) are designated by sj, then one first calculates the mean of y for all sj. Then one calculates $y\text{-}y_{sj}$ where y_{sj} are the means of y for the categories of s. The same must be done for all of the independent variables in the regression equation. It must be emphasized that when doing regressions on demeaned variables, the degrees of freedom for the standard error has been reduced by the number of values of the fixed effect variable, and the computer will not know this unless it offers special options for this technique.

The Stata system offers two commands, **areg** and the more general **xtreg**, that automates the demeaning and uses the correct degrees of freedom. The command **areg varlist, absorb(s)** command removes a single large fixed effect variable s (i.e., the dummy variable set implied by the values of s), while **xtreg** allows two or more large fixed effect variables via double demeaning. Double demeaning controls for a second fixed effect variable by demeaning the demeaned variables a second time.[3]

It is critical to understand that double demeaning is equivalent to creating a fully interacted set of dummy variables. That is, if one fixed effect variable is S and another is T, demeaning for S and then T is the same as creating dummy variables for the S x T cross-classification. Moreover, the **areg** command cannot be used to remove one fixed effect variable via the **absorb( )** option while removing another fixed effect variable via demeaned variables in the **varlist**. The program was not designed for this purpose, and using it this way will give incorrect and inconsistent coefficients. This is apparently what happened in the HKR analysis. When HKR removed both student and school by grade fixed effects, the **areg…absorb( )** command was used for one fixed effect and demeaning was used for the other.

As an illustration, the following regressions for black peer effects were computed using achievement test scores from the Charlotte-Mecklenburg school district, a subset of the North Carolina data I was using for a replication. This district has 116 elementary and middle schools in the data set, which is small enough so that that I can remove school by grade fixed effects using regular dummy variables. The regression models below are set up to replicate the HKR gain models, in that they use gain scores as the dependent variable and,

[3]An equivalent procedure is to first create means for the two cross-classified variables, remove these means for each variable in the regression, and then treat this as a single large fixed effect variable.

Figure 2.

Armor C-M Models	Black Peer Effect[a] (se)	Variations in large fixed effect computations[b]
CORRECT		
Gain score (1)	-.0010 (.0003)	No large fixed effects
Gain score (2)	+.0021 (.0018)	Student removed by **absorb**; school by grade removed by conventional dummy variables
INCORRECT		
Gain score (3)	+.0020 (.0020)	Student removed by **absorb**; school by grade removed by demeaning
Gain score (4)	-.0008 (.0003)	Student removed by demeaning; school by grade removed by **absorb**

[a] Change in black gain scores (in standard deviations) for a 1% increase in percentage of black students
[b] All models include grade-by-year dummies, school change indicators, and free lunch indicators.

in addition to the fixed effect variations shown above, they also control directly for year by grade dummies, school changes, and student free lunch (see Figure 2).

The gain score model (1) without removal of the large fixed effect variables produces a coefficient not unlike the HKR coefficient in their model (1); in fact, it is slightly larger (-.0010 vs. -.0007). The correct removal of the non-interacted student and school by grade fixed effects (Gain score 2) produces a positive black peer effect. However, when the **areg** command is used to remove two large fixed effect variables, the size and sign of the coefficient depends on which is removed by **absorb** and which is removed by demeaning, a clearly flawed procedure. The sign change is similar to that occurring between models (2) and (3) in the HKR results.

There is no way to remove two large non-interacted fixed effect variables in Stata, or any other statistical software that I know about. In the case of the HKR study (and the Armor replications) this means that student and school by grade cannot be removed simultaneously without their interaction. If they are interacted, this means estimating fixed effects for student by school by grade, which is indeterminate because it uses up all the degrees of freedom.

Finally, I have now completed replications the HKR models using data from North Carolina, South Carolina, and the national Early Childhood Longitudinal Study, controlling for school by grade fixed effects. In a private communication, Hanushek argues that this fixed effects variable is the more important control in lag models because the lagged achievement variable removes most of the student-specific variation. Generally, I find very small black peer effects, mostly not significant (Armor and Duck, 2007a, 2007b).

References

Armor, D.J. and S. Duck O'Niel (2007a). Unraveling the Effect of Black Peers on Black Achievement. Education Working Paper Archives, University of Arkansas.

Armor, D.J. and S. Duck O'Niel (2007b). The Effect of Black Peers on Black Test Scores, unpublished manuscript, School of Public Policy, George Mason University.

Hanushek, Eric A., John F. Kain, and Steven G. Rivkin (2002). New Evidence about *Brown v. Board of Education*: The Complex Effects of School Racial Composition on Achievement.

Working Paper #8799, National Bureau of Economic Research, February.

Hanushek, E.A., John Kain & Steven Rivkin.(2004, 2006) New Evidence about *Brown v. Board of Education*: The Complex Effects of School Racial Composition on Achievement Institute for Research on Poverty Discussion Paper no 1284-04; Revised March 2006.

Johnston, J. and J. DiNardo (1997). Econometric Methods, McGraw-Hill.

Teaching the Concept "Degrees of Freedom"

by Ferris J. Ritchey
fritchey@uab.edu

One of the most vexing challenges when teaching both undergraduate and graduate statistics is conveying a meaningful notion of degrees of freedom. Unlike mathematically trained students who carry rulers in their heads, many social science students do not readily think in linear terms; therefore, the word *degree* is vague. And the word *freedom* does not seem to fit the material in a course that requires so much formulated work and has greatly restricted students' leisure time. The traditional definition of degrees of freedom as *the number of independent events* always evokes the question: Independent of what? Moreover, this definition conjures up *independent* variable and this causes confusion.

I have found a way to approach the concept of degrees of freedom that answers the question "Independent of what?" without having to ask it. I propose an alternative definition of degrees of freedom as *the number of opportunities in sampling to compensate for limitations, distortions, and potential weaknesses in statistical procedures*. The following is excerpted from my text, *The Statistical Imagination: Elementary Statistics for the Social Sciences*. 2nd edition, pp. 322-324. New York: McGraw-Hill.

What Are Degrees Of Freedom?

In doing a statistical analysis on a sample, care must be taken to prevent research procedures from leading to inaccurate conclusions about the population. Every measurement instrument and statistical technique has limitations that potentially distort interpretation of data. For example, the famous Hubble Telescope (which rides on a satellite outside the atmosphere) provides distorted images because of a microscopic misalignment in the curvature of its lens. As a result, the photographic images appear blurred. The stars themselves are not blurred; the blurring is a function of the measurement instrument's limitations. The faulty telescope gets in the way of an accurate assessment of the true shapes of distant galaxies. On-site adjustments (via the space shuttle) compensate for the bent lens up to a point, but Hubble images are still not pure. The telescope has a peak degree of accuracy and this level is fixed. The conclusions drawn about the nature of its photographic subjects (stars, galaxies, quasars, etc.) are restricted by the tools and methods used to gather data. Even with computer enhancements, Hubble scientists inevitably confront a lack of flexibility in correcting the distortions of their measurement instrument. The picture they see is only a close approximation of what truly appears there.

Similarly, statistical procedures have limitations that potentially get in the way of an accurate assessment of population parameters. To estimate the spread of a sampling distribution of means, we must consider the effects of the major limitation of the mean (Chapter 4): The calculation of the mean is affected by extreme scores or outliers. This distorting effect is especially troublesome with small samples. Being aware of this limitation, we adjust calculations to account for the sensitivity of the mean to outliers, just as Hubble scientists computer-enhance their photographic images. Any statistical procedure has limits–a lack of total freedom in how it is used. We use the term *degrees of freedom* to refer to how flexible a statistical procedure is. The more degrees of freedom we have, the better, because **degrees of freedom** are *the number of opportunities in sampling to* *compensate for limitations, distortions, and potential weaknesses in statistical procedures*.

[For instance,] why are the degrees of freedom for the approximately normal *t*-distribution calculated as n - 1? For a variable, an extreme score in the sample can produce an inflated or deflated mean–one that does not reflect the true population parameter value of that variable. When the sample size is small, this distortion can be rather large. Once a high-value extreme score is randomly drawn into a small sample, there are not many opportunities left for a low-value case to be selected to pull the computed mean closer to the true population parameter. An extremely high score fixes the calculation of the mean into the high-value end of the score range of the variable. The small sample is inflexible, not free of the mean's limitation of sensitivity to extreme scores. It has few degrees of freedom.

To illustrate these principles, we noted in Chapter 7 … that an infinite set of single random digits ranges from 0 to 9 with a mean of 4.5… That is, the true population parameter, μ_X, is 4.5. Suppose, however, we did not know this, and to get an estimate of this population parameter, we sampled these digits and computed the sample mean, $\overline{X}$. Ideally, this estimate would be close to the true parameter of 4.5. This is accomplished when in the random draw process each chosen score on the high side (e.g. 9, 8, 7, or 6) is balanced out by a score on the low side (e.g., 0, 1, 2, 3, or 4). With a true parameter average of 4.5, a perfectly accurate random sample would include as many 0's as 9's because the mean of these two scores is 4.5. Similarly, this perfect sample would include as many 1's as 8's, 2's as 7's, 3's as 6's, and 4's as 5's. But suppose that our sample size is small, say, n = 5. Imagine further that the first random digit drawn for the sample is a 9, an extreme score. When 9 is added to ΣX in computing the sample mean, it is likely to cause our estimate of the parameter to be on the high side. For example, the following sequence of sample draws could occur and result in a "high-side" sample mean of 6.2:

Sampling sequence: 9, 5, 3, 8, 6

$$\overline{X} = \frac{\Sigma X}{n} = \frac{31}{5} = 6.2$$

With this small sample, after we draw the 9, it is very likely that our estimate will be high because we had only four sampling opportunities left (n - 1) to pick up a 0 to balance out the calculation of the mean. We would say that we have only 4 degrees of freedom. Drawing a 9 locks us into a high-side estimate. A sample of five is not very flexible once an extreme score enters the sample.

With a large sample, say, size 130, the drawing of a 9 early on is not as large a problem. We have 129 more chances to draw a 0 to bring the sample calculation of the mean back into the 4.5 range. With a large sample there is greater freedom of adjustment in the sampling procedure.

Another [traditional] way to look at the concept of degrees of freedom is "independence of sampling events." For instance, suppose someone said he or she collected five random digits and computed a mean of 6.2, as in the preceding illustration. If this researcher told us the values of four of the digits, we could mathematically determine the fifth. That is, if the first four digits are 9, 5, 3, and 8, for the Xs to add to 31 to produce a mean of 6.2, the last digit has to be 6. In other words, the last digit is not free to vary; its value is dependent upon how the mean is calculated. Thus, in calculating the degrees of freedom for a mean, we subtract 1 from the sample size. In this example that leaves us 4 degrees of freedom. Four of the digits are "free to vary." Degrees of freedom, then, can be viewed as the number of independent sampling events–events that are independent of the limitations of the statistical formula used.

It is only with a sampling distribution of means that degrees of freedom are calculated as n - 1. With other statistical procedures, adjustments for degrees of freedom depend on the particular limitations of a procedure. .. [Therefore, degrees of freedom are calculated different ways for various statistical procedures.]

Degrees of freedom calculations represent a recognition of the limitations of a procedure. We … use wording such as "*adjust* for degrees of freedom" and "*correct* for degrees of freedom." We often say that a particular limitation causes us to "*lose* degrees of freedom." For instance, the sensitivity of the mean to outliers causes us to lose 1 degree of freedom. Adjusting for the degrees of freedom of a procedure is an essential part of assessing sampling error. We must constantly be aware that to sample is to look through a narrow lens. We must ask: Are what we see and what is truly there one and the same? If we know our lens is blurred, we must take this into account, just as Hubble scientists "correct" their digital photographic images with computer enhancements. The calculation of degrees of freedom is our mode of correction.

Book Announcement

Forthcoming Book by Stephen L. Morgan and Christopher Winship

Morgan, Stephen L. and Christopher Winship. 2007. "Counterfactuals and Causal Inference: Methods and Principles for Social Research." Cambridge: Cambridge University Press.

The book will be released July 31, 2007.

New Faculty Member Announcement

Congratulations to Jennie Brand (jebrand@email.unc.edu). She will be starting as Assistant Professor of Sociology at UCLA this fall.

New PhD in Survey Research and Methodology

Congratulations to Mario Callegaro (mca@unlserve.unl.edu). On May 5th he obtained a PhD in Survey Research and Methodology from the University of Lincoln, NE. He is the first PhD in this program that started in the Fall of 2003.

His dissertation is titled: "Seam effects changes due to modifications in question wording and data collection strategies. A comparison of conventional questionnaire and event history calendar seam effects in the PSID."

He will start working as Survey Research Scientist for the company Knowledge Networks in Menlo Park, CA next September.

Notes from the Editor

I have changed the publication schedule to conform more closely with current practice: Winter and Summer. Submission deadlines are a few weeks prior to publication, but circumstances mandate flexibility. Despite repeated requests, submissions trickle in very slowly, so I have been postponing publication until there are enough to justify the effort. As always, I solicit section news, news about section members, brief essays, and suggestions for process, content, and format. Please send contributions, suggestions, and/or comments to me at L.raffalovich@albany.edu.The section web page is here at SUNY-Albany (http://www.albany.edu/asam). Please give us feedback.