Persistent Inequality? Answers From Hybrid Models for Longitudinal Data

Marc A. Scott and Mark S. Handcock

Sociological Methods Research 2005; 34; 3
DOI: 10.1177/0049124105277194

Persistent Inequality?

Answers From Hybrid Models for Longitudinal Data

MARC A. SCOTT
New York University

MARK S. HANDCOCK
University of Washington, Seattle

Many questions in social research must be evaluated over time. For example, in studies of intragenerational mobility, measuring opportunity for economic advancement requires longitudinal data. The authors develop and use a class of hybrid functional models to demonstrate how different models can lead to extremely different substantive conclusions. They provide guidelines for longitudinal data analyses in which variance partitions are central to the inquiry. In their analysis of the National Longitudinal Survey of Youth, the authors conclude that in a period of rising wage dispersion, the bulk of inequality is persistent over the life course. Their models provide support for the scenario in which wage inequality rises steadily while instability slowly diminishes over time. They obtain mild evidence of increased wage instability for somewhat older workers in the early 1990s, matching a recessionary trend. These findings contribute significantly to understanding wage inequality in United States over the past 25 years.

Keywords: *covariance structure; variance components; functional data analysis; wage inequality; National Longitudinal Survey*

1. INTRODUCTION

Longitudinal data require more sophisticated modeling techniques because the responses are correlated within individuals. This correlation structure can be quite complicated, and many parametric forms for

AUTHORS' NOTE: *This research was supported by the National Science Foundation under grant SES-0088061 and partially supported by the Russell Sage Foundation and the Rockefeller Foundation. We thank the editor and two anonymous reviewers for their extremely helpful comments and suggestions. We also thank Annette D. Bernhardt and Martina Morris, whose discussions and insights over the years are deeply embedded in this article.*

SOCIOLOGICAL METHODS & RESEARCH, Vol. 34, No. 1, August 2005 3-30
DOI: 10.1177/0049124105277194

covariance have been developed over the years. The most familiar of these are error-in-variables models such as compound symmetry and time-series models such as ARIMA (see Diggle, Liang, and Zeger 1994 for others). While many of these models "soak up variation" quite well and thus at least approximate the covariance structure of the process, they are rarely oriented toward answering specific hypotheses and do not always admit a meaningful variance components decomposition. That is, the parameters are often not directly interpretable with respect to specific research questions. In some instances, the covariance modeling is viewed as a nuisance that must be considered but is not of direct interest. The focus of our research efforts has been in applications in which the covariance structure is of intrinsic interest to practitioners.

In this article, we contrast a range of models for covariance structure; in particular, we relate commonly employed growth curve models to a class of models that we developed to capture population-level features of the covariance. We find dramatic differences in the economic implications of three exemplary models and discuss the role of the analyst and model in resolving such differences. We also show that a goodness-of-fit concern that arose with the more common model can be better understood using the new class of proposed models.

In the domain of wage inequality, the research questions that may be addressed by cross-sectional versus longitudinal data are fundamentally different. For example, Levy and Murnane (1992) and Danziger and Gottschalk (1993) document an increasing cross-sectional polarization in wages that began in the early 1970s and continued into the 1990s. Such findings imply that cross-sectionally, comparing, say, 1975 to 1985, the rich had become richer and the poor poorer in real terms. But were the rich in 1975 still rich in 1985? This is a very different question and may be answered only with individual-level wage data recorded over time. A more specific concern was whether the opportunity to move into higher wage brackets had become more or less available in this age of increased inequality. Some might argue, as is often done in the context of education, that as long as opportunity is available, it is a fair game or, in this case, labor market.

In the mid-1990s, the debate shifted somewhat toward these intragenerational mobility concerns. Gottschalk and Moffitt (1994)

used the Panel Study of Income Dynamics (PSID) and found that mobility in the 1980s was about at the same level as in the past. The source of increased wage polarization was an increase in short-term wage fluctuations; wages had become more unstable, adding more variation to any cross section, but mobility was intact. Bernhardt et al. (2001) used the National Longitudinal Survey of Youth (NLSY) and instead found a decrease in mobility and no change in wage instability. Both of these studies used variants of the common mixed-effects growth curve model to handle the more complex longitudinal structure. Our goal in this article is not to resolve the discrepancy between these two authors directly but to explore how the choice of model influences the findings in this very important debate.

This article is organized as follows. In section 2, we discuss the literature on longitudinal data modeling, focusing primarily on how covariation is modeled. In section 3, we introduce our modeling class, its predecessors, and related techniques. In section 4, we explore intragenerational wage inequality in the 1980s and 1990s, contrasting three exemplary modeling approaches. In section 5, we illustrate the use of hybrid models to answer specific research hypotheses. In sections 6 and 7, we summarize and discuss the implications of the findings.

2. MODELS FOR COVARIANCE

To date, a number of different approaches have been developed that are appropriate for modeling covariance structure. Random coefficient models, as described in Longford (1993), model heterogeneity via random perturbations to a known structure captured in a design matrix. Their strengths include a well-established set of available inferential techniques, provided one knows a priori the correct *form* of the design for the covariance. Nonlinear extensions of random coefficient models as well as nonparametric estimation of the coefficient densities are described in Davidian and Giltinan (1995). All these methods either assume a highly restrictive form for the covariance or leave it largely unspecified.

A different approach to covariance modeling comes from the functional data analysis literature, in which the data are sampled so

frequently that they can be taken to be functions. The framework for this approach is given in Ramsay and Silverman (1997), and key models were developed by Rice and Silverman (1991), Kneip (1994), and Besse, Cardot, and Ferraty (1997). Lindstrom (1995) combines ideas from both mixed-effects and functional approaches, while Barry (1995) provides a Bayesian model for the covariance function. Works by Ke and Wang (2001) and Guo (2002) are attempts to incorporate more flexible modeling approaches in a broader class of models; these have strong roots in the smoothing spline ANOVA (Wang 1998) and SEMOR (Lindstrom 1995) modeling classes. Brumback and Rice (1998) show that certain forms of covariance yield smoothing spline trajectory *predictions* for each subject.

The conceptual framework that motivated the development of our models is known as latent curve analysis. It began with Rao (1958) and was developed in Meredith and Tisak (1990). The Meredith and Tisak modeling class was specified quite generally; unconstrained, their latent curves resemble principal components and suffer some of the limitations of that framework. For example, there is no formal mechanism for incorporating unbalanced or incomplete records when estimating the required covariance matrix, and estimates are likely to be highly variable. Our development offers a model-based alternative. By placing a limited constraint on the curves, we eliminate some of the estimation problems associated with Meredith and Tisak and customize the modeling class so that it directly addresses specific research hypotheses. The models are conceptually rooted in the functional data analysis paradigm described in Rice and Silverman (1991), in which individual differences take the form of latent curves, constrained to lie in some function space. In the economic context, the curves represent permanent differences between individual trajectories (signal), and whatever remains can be deemed transient (noise). We will show that these function spaces can be chosen to inform substantive hypotheses and that the resulting variance components will have a useful interpretation.

In Scott and Handcock (2001a), we developed this new class of latent curve covariance models and called them "proto-splines." Scott (1998) and Scott and Handcock (2001b) developed software for maximum likelihood estimation and inference for this class.

Related work by James, Hastie, and Sugar (2000) framed similar models as principal component models in the functional data arena and applied these to biological growth data. Inference in that work was based on the bootstrap. The approaches are complementary in several ways. We developed asymptotic inference for the complete data case (approximate in the more general case), while James et al. used bootstrap for inference and emphasized application to sparsely sampled data. We established, and in this article expand, a modeling framework for relating this functional data analysis approach to social research questions. James (2002) developed related formulations for generalized linear models in the biological domain.

3. MODEL FORMULATION

Suppose that a subject i is observed at specific design points, $\vec{t}_i$, and let $Y_i(t) \equiv Y_{it}$ be the observation at time t. Our basic latent curve model is

$$Y_i(t) = \mu(t) + \sum_{\nu=1}^{K} \omega_{i\nu}\phi_\nu(t) + \epsilon_i(t), \tag{1}$$

where $\mu(t)$ is a mean function, to be discussed subsequently; $\phi_\nu(t)$ is a basis curve; $\omega_{i\nu} \sim f(\cdot)$ are individual specific coefficients for the νth curve drawn from some distribution f to be specified; and $\epsilon_i(t)$ is an error term. Specification of the basis curves ϕ_ν (called principal functions in Ramsey and Silverman 1997) is a central challenge in models of this type. They provide the foundation (literally, the design) for how individual differences are generated. Most mixed-effects models implemented in statistical software packages are special cases of this latent curve model, for which basis curves are specified prior to model estimation. We depart from this approach by estimating the basis curves concurrently with all other model parameters. We estimate ϕ_ν from the data subject to the constraint that they lie within a limited class of functions; the function class is selected so that fitted versions adjudicate between competing substantive hypotheses.

A good example of a flexible class of functions that can be used in this manner are the family of cubic splines. These smooth, piecewise polynomials are defined in a fixed interval in which one or more "knots" are placed. The knots can be thought of as joints, allowing the curve to bend up or down dramatically at these inflection points. One can test whether the process under investigation undergoes a change in trajectory by choosing knots at key points in time. The fitted latent curve sets the inflections permanently—unlike other modeling classes, the shape of the curve does not vary across subjects. Thus, it represents systematic variation or meaningful structure existing in the entire population. There are other function spaces that may be equally informative in different contexts, such as those that allow for jump processes or change points.

We now assume that the function space has been chosen and formally describe how the ϕ_ν are to be estimated. Let $\{\psi_j\}$ be an orthonormal basis to a smooth function space $\mathcal{H}$ of dimension $S \leq T$, where T is the number of distinct design points. There is no restriction on these basis functions, and in particular, they need not be a spline basis. The main idea to proto-splines is to use only a subset of the $\{\psi_j\}$ to construct each latent curve ϕ_ν. This departure from most functional analysis models forces the curves to come from a proper subspace of $\mathcal{H}$. If $\mathcal{H}$ is a spline basis, then this implies that the resulting ϕ_ν were partially formed or "proto"-splines. Most approaches to smooth estimates of covariance must impose an external orthogonality constraint between each ϕ_ν; a lack of external constraints was one of the motivations behind the proto-spline approach.

For this comparative analysis and most of the discussion, the single proto-spline model will suffice. This means that we have only to estimate a single function $\hat{\phi}_1$, and it will use all S of the available basis functions in $\{\psi_j\}$. For illustrative purposes, one can let $\{\psi_j\}$ be the orthogonal basis of second-order polynomials (over time), or $\{1, t, t^2\}$ (appropriately orthogonalized). The $\hat{\phi}_1$ estimated is analogous to a principal component, as it explains a large portion of the covariance. The latent curve is a deterministically weighted sum of basis functions.

$$\hat{\phi}_1(t) = \sum_j \hat{\eta}_j \psi_j(t), \tag{2}$$

where the $\hat{\eta}_j$ are estimated from the data, establishing the shape of the latent curve $\hat{\phi}_1$.

Restating our single curve model,

$$Y_i(t) = \mu(t) + \omega_{i1}\phi_1(t) + \epsilon_i(t). \tag{3}$$

For identifiability, the random coefficient ω_1 is presumed to be standard Gaussian, and we allow ϕ_1 to have norm other than one. Part of what makes our latent curve approach novel and more flexible is that ϕ_1 is only *constrained* —its exact form is driven entirely by the data. We attribute a hybrid interpretation to models such as these; they are not quite population-average models, nor are they individual specific, in the common usage of those terms. Here, ϕ_1 represents a population feature, as there are no subscripts for subjects in any of its components. Individual curves are built up through a shift of $\omega_{i1}\phi_1(t)$ from the mean, so that subject-level variation is wholly captured through rescaling via the random coefficient. Figure 1 depicts two different latent curves estimated from the NLSY data. The curves differ because they have been constrained to come from somewhat different function spaces. Figure 2 uses the quadratic fitted curve in Figure 1 to generate four individual-specific curves surrounding the mean process. These curves correspond to $\omega_{i1} = 2, 1, 0, -1, -2$, multiplying the single latent curve $\hat{\phi}_1$ (the mean process is thus a fifth curve). We will see that interpreting this latent curve as a feature of the population will help us differentiate findings from competing models.

We have discussed the relationship of our proto-spline models to latent curve and functional data models. Proto-splines are also related to a nonstandard class of mixed-effects models known as reduced-rank models (see James et al. 2000). The nonstandard and rather complex form the covariance takes in our class of models requires further analysis to establish the asymptotics. We explore their relationship below.

What follows is a comparison of a K proto-spline model and a standard mixed-effects model. For notational convenience, we suppress reference to time in the functions, representing $\psi_j(t)$ and $Z_i(t)$ as ψ_j and Z_i, respectively. Let $Z_i = \{\psi_1, \psi_2, \ldots, \psi_T\}$ be a design matrix constructed using the basis functions for the space $\mathcal{H}$. The coefficients η_j are assumed to be ordered so that if there are K different groups used in the model, with the kth group coefficients

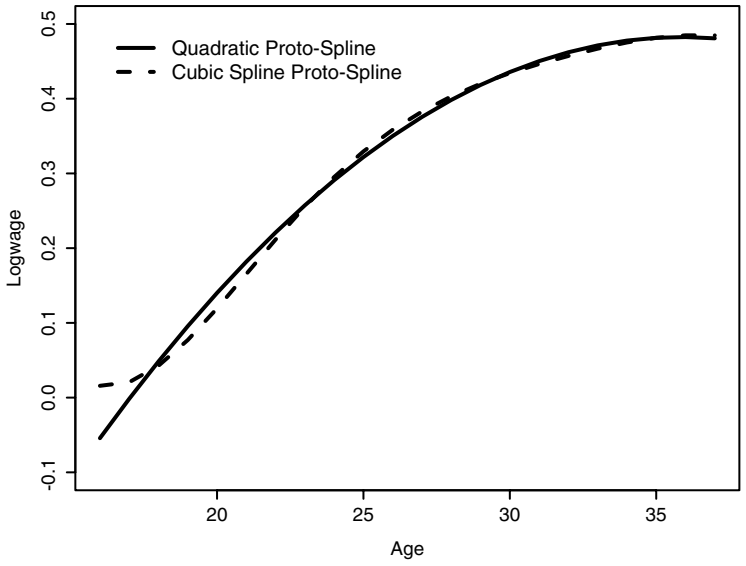

Figure 1: Proto-Spline Curves for Quadratic and Cubic Spline Model Formulations

given as $\gamma_k = (\eta_{k1}, \ldots, \eta_{kn_k})^T$, then these can be stacked into a $T \times K$ matrix $\Gamma = \bigoplus_{k=1}^{K} \gamma_k$. The difference between these two models can be understood by examining their representations. The proto-spline model class is

$$Y_i = X_i\beta + Z_i\Gamma\delta_i + \epsilon_i, \tag{4}$$

where X_i is a design matrix tracking the mean and $\delta_i \sim N_K(0, I_K)$. The likelihood-equivalent mixed-effects model is

$$Y_i = X_i\beta + Z_i\delta_i^* + \epsilon_i, \tag{5}$$

where $\delta_i^* \sim N_T(0, \Gamma\Gamma')$. Given this identification, one might suppose that through a change of variables, one could estimate this model using standard mixed-effects software. But there is no transformation of covariates here; rather, there is a transformation of covariance structure. The proto-spline formulation (2) has K random effects, while (5) has T, and this new structure is positive semi-definite (thus, the reduced rank designation). As such, the covariance associated with model (5) is not implemented in standard statistical software.

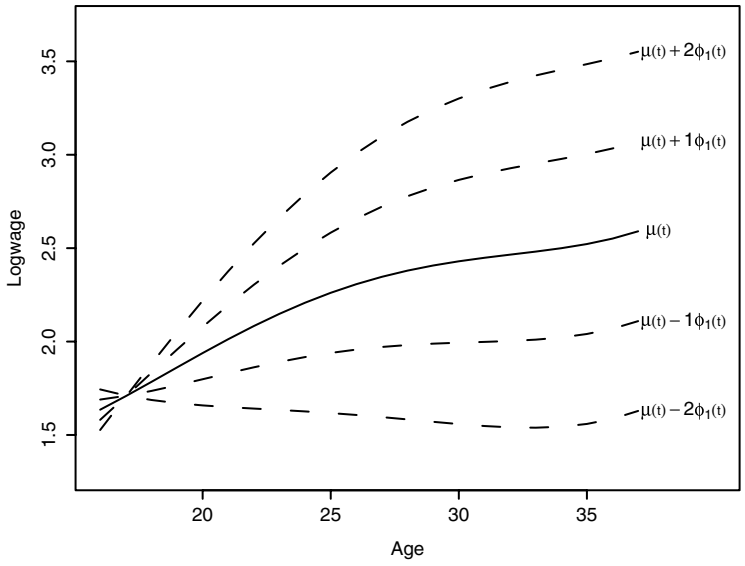

Figure 2: Five Exemplary Permanent Wage Realizations for the Quadratic Proto-Spline Model Formulation

Restricting our attention to the *single* proto-spline model ($K = 1$), formulation (2) contains a scalar random effect, δ_i, while the vector δ_i^* defined in (5) contains T effects. Although this covariance structure $\Gamma\Gamma'$ governing δ_i^* is degenerate, it does not introduce problems with inference because the degeneracy is removed in the full likelihood, once residual variation is included. We estimate this model using maximum likelihood; see James et al. (2000) for an alternative estimating procedure. If one examines the structure $\Gamma\Gamma'$ more closely, it is apparent that each element of the vector δ_i^* is linearly dependent on each of the others, so in essence, only one random effect is generated by this covariance structure when $K = 1$. The dependent relationships correspond to the relative magnitudes of the components η_{kj} of γ_k. So the likelihood-equivalent model (5) is a reduced-rank, nonstandard mixed-effects model.

Scott (1998) and Scott and Handcock (2001b) prove that for a balanced design, the maximum likelihood estimates for proto-splines are consistent and asymptotically Gaussian. We use these findings in the

estimation and approximate inference associated with our application, to which we now turn.

4. APPLICATION TO THE ANALYSIS OF WAGE INEQUALITY

To contrast the hybrid functional class of models with traditional mixed-effects models, we explore their ability to address an important question in labor economics. Returning to our motivating example, we look at inequality of wage outcomes for young workers in the United States and competing explanations for this phenomenon. The issue is whether wages have become more volatile or more permanently divergent. Rather than compare different economic periods, we focus on the variance partition itself. We use models to identify permanent and transient portions of a collection of wage trajectories and then explore how the partition evolves over time. What is surprising is that the partition depends highly on the model class being fit.

We will be investigating a data set from the NLSY: A representative sample of young men ages 14 to 21 was interviewed in 1979 and has been interviewed yearly since then, with 1994 being the last year included in our analyses. For comparability with other cohorts and studies, we also selected only non-Hispanic whites ages 14 to 21 in the first year of the survey, with a resulting sample size of 2,427. These workers were followed for a total of 16 years. On average, they worked 11 of those years, yielding close to 26,000 total observations. A detailed description of the construction of the data set is given in Bernhardt et al. (2001).

Empirically, the variance of the wage process increases as individuals age. This can be clearly seen in Figure 3, in which the total variance of wages increases from about 0.10 to 0.30 over 20 years of the life course (hourly wages have been adjusted for inflation, and the natural logarithm of these is our response variable). This increase is consistent with the scenario in which wages "fan out" over time—individuals develop an overall trajectory, and differences in those tend to grow. Part of the explanation may be that wages are much more volatile for older workers, and given labor market trends during this period, such as significant corporate downsizing, this theory cannot be dismissed without further analysis. In other words, the increase in

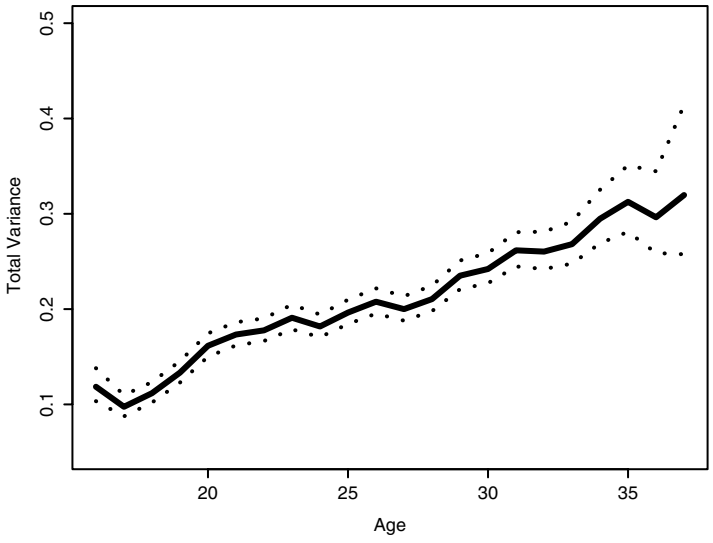

Figure 3: Empirical Variance of the Wage Process Over Time, With 95 Percent Confidence Bands

variation observed as workers age may be driven by an increase in transient wage fluctuations. This argument mirrors Gottschalk and Moffitt (1994) in their claim that increased wage volatility accounts for growth in overall wage dispersion across two time periods but this requires further investigation. To reach any conclusion about whether the bulk of variation is permanent or transient, we refer to models and their variance component decompositions. The question we address in this article is which model—or which class of models—we should choose and the implications of the choice.

We briefly discuss the mean function as it relates to our economic analysis. The covariates that are associated with a worker's trajectory, such as demographics, job acquisition, turnover, completion of educational milestones, enrollment status, and moves in and out of different industries and occupations, form a historic profile. Fixed-effects regression coefficients capture the way these covariates shape the expected (or mean) trajectory.

In the context of exploring permanent and transient variation in wages, an exogenous covariate captures permanent differences. The

clearest example is age, with the expectation that 16-year-olds earn less, on average, than 30-year-olds. Any analysis that did not model the role of age or a close proxy would seem inadequate. Can we say the same about education or experience, or the role of frequent job changing? These covariates are partially endogenous, so that extra-mean variation is correlated with them. This may be induced by an unobserved covariate, such as ability, which may be correlated with a covariate such as education. In the longitudinal setting, the endogeneity-correcting approach of Chamberlain (1984) can be used to adjust the value and interpretation of such covariates.

It is interesting to note that in ordinary regression, one is reallocating variation from unexplained to explained, while in the longitudinal setting, one is apportioning between three types: fixed (via the mean), structured variation, and unstructured. The first two are permanent, while the last is transient. Endogenous covariates complicate matters as they contribute to both structured and fixed components of variation. They unambiguously contribute to permanent variance, however. While we are concerned with the role of endogenous covariates and explore this in depth (including corrections for endogeneity) in Bernhardt et al. (2001), we restrict our attention in this analysis to a mean structure based only on the exogenous covariate age. This will in no way limit our comparison of modeling approaches, as the same mean process is being used throughout. In applications for which pure growth trajectories over time are the primary focus, questions of endogeneity are less central.

COMPARATIVE MODELS

We begin by specifying several standard mixed-effects models, with the random intercept model (RI) as a common starting point. In this model, individuals are assumed to possess an unobserved, person-specific trait that identifies the extent to which they differ from the mean response.

$$Y_i(t) = X_i(t)\beta + \alpha_i + \epsilon_i(t), \tag{6}$$

in which $X_i(t)$ captures mean effects, and α_i captures the unobserved trait. In some disciplines, an interpretation, such as "motivation," is attached to this trait. We are agnostic to such identifications and

simply view these terms as capturing some unobservable differences, which may be the aggregation of one or several characteristics, measurable or otherwise. For our purposes, we want the covariance structure induced by all these unobservables to be modeled appropriately in the permanent portion of the variation. We emphasize that α_i does not vary over time.

In this and all of our models, the mean effect chosen consists of a quartic in age. This choice was made empirically but was based as well on Murphy and Welch (1990); we tested a nonparametric specification and found the polynomial superior in terms of the Bayesian information criterion (BIC; Schwarz 1978). The dependency of covariance on proper mean specification makes this choice important (Diggle et al. 1994). The random effect α_i is assumed to be Gaussian, with constant variance. To allow for different economic trends, we add an error structure that is heteroscedastic (variance depending on the year of observation). This error structure was compared to several others and was also justifiable in terms of BIC. The heteroscedastic error structure was common to all models, and estimation was by the method of maximum likelihood.

We note that a RI model can be thought of as the most basic latent curve model in that it captures population-level structure. However, it assumes that the "shape" of that curve is known from the outset or, equivalently, $\phi_1(t) = 1$. In this model, there is nothing unknown or even estimated, with the exception of the variance of the random coefficient α_i. As mentioned above, this form assumes that individual differences exist at all ages, with the implication that one's first few jobs define one's lifetime expectations. This is clearly not the case, but perhaps the RI model can be justified on the grounds that it captures whether individual means (across time) are above or below expectations. Focusing on mean differences across time ignores the fact that these curves have more complex *trajectories*.

In the class of standard mixed-effects models, there is a hierarchy typically employed with growth curves. If the zeroth-order polynomial is insufficient, then try a first-order polynomial, and so forth. Pinheiro, Bates, and Lindstrom (1994) outline such model building in more general terms. We fit a first-order (random slope) model to this data, but we limit our discussion to the second-order random

quadratic (RQ) model, which has many of the same properties and provides a superior fit in terms of BIC. This model is

$$Y_i(t) = X_i(t)\beta + Z_i(t)\delta_i + \epsilon_i(t). \tag{7}$$

$X(t)$ is a basis for quartic polynomials, such as $(1, t, t^2, t^3, t^4)$, and $Z(t)$ is a basis for quadratic polynomials, such as $(1, t, t^2)$, and $\delta_i \sim N(0, G)$, where G is a 3×3 unspecified covariance matrix. The vector δ_i can take on any value, so in effect, every (mean zero) quadratic curve may be generated by the term $Z(t)\delta_i$. The random effects are distributionally constrained by the covariance G, so curves have very different occurrence probabilities. The mean curve $X(t)\beta$ and individual deviation $Z(t)\delta_i$ comprise the permanent portion of the wage trajectory, but these are not directly summarized in the model's covariance parameters. No single parameter is sufficient; the variances of the components of the vector δ_i cannot be fully interpreted without considering the covariance terms. Instead, a meaningful summary of model-based, age-specific permanent variance is given by the diagonal of $Z(t)\hat{G}Z(t)'$.

In contrast to the two polynomial models is the single latent curve proto-spline (SPS) model:

$$Y_i(t) = X_i(t)\beta + \omega_i\phi_1(t) + \epsilon_i(t), \tag{8}$$

with $X(t)$ as before, and $\omega_i \sim N(0, 1)$. Because $\phi_1(t)$ is a continuous curve, identical for each subject in the population, any information about systematic differences between individuals gets transmitted across time by following the shape of this curve (think of the curve as a conduit). The specification of a function space that captures a meaningful set of transmittal mechanisms remains as a crucial step in the analysis. We initially chose the function space of quadratic curves but eventually employed a cubic spline function space. Quadratics are a fairly limited family for proto-spline curves because there can be no inflection points. In these models, the structured covariance is clearly identified with permanent differences between individuals.

FITTED MODELS AND VARIANCE PARTITION SUMMARIES

We find that the maximum likelihood estimates for our three models tell somewhat different stories. The RI model estimates the

between-trajectory permanent variance to be constant at 0.1144. The RQ model estimates a variance of 0.1051 for the intercept, but this is negatively correlated (−0.673) with the slope, so individual differences are attenuated rather than exaggerated over the life course. For the SPS model, the shape of all extra-mean permanent differences is easily summarized by $\hat{\phi}_1(t)$, the common shape of variation, which is scaled by a *single* random coefficient (Figure 1). The solid curve represents the strongest single systematic variation in the stochastic process, much as the first principal component of a covariance matrix would, under the constraint that it comes from the family of quadratic curves. The dashed curve in Figure 1 is the proto-spline fit for a model that uses a cubic polynomial basis and will be discussed in section 5. A summary of the permanent variance generated over time is constructed by squaring the values of $\hat{\phi}_1(t)$ at each age.

Each model generates a different variance partition; how these change over time is important and characterized by Figures 4 and 5. Note that the cohort is tracked over 16 years, but these years vary depending on when the worker entered the survey. So someone who was 16 in 1979 is 27 in 1990, while someone who is 21 in 1979 is 27 in 1985. Here, we report the (weighted) mean variance at each age. In the NLSY design, age and cohort effects are confounded. This does not, however, lessen our ability to pick up differences in year-to-year labor markets; the variances still change over time—we simply cannot separate the market versus age effects.

Examining Figure 4, we see a clear difference in how the two standard random-effects models decompose the total variation. The RQ starts out with a permanent variance of 0.10, which then drops down to under 0.05 by age 18 and rises substantially to 0.30 by age 35. Contrast this to the RI model, which by definition sets this to a fixed level (0.11). The 95 percent confidence bounds are given for the permanent variance estimates. These are based on Normal theory asymptotics but are likely to be quite accurate given the large sample size (the median number of observations per age is about 1,500; all but three ages have more than 500 observations contributing to the estimate). The SPS model's permanent variance estimates through age 25 are extremely precise; this occurs as a result of a combination of constraints due to function space, the population-average

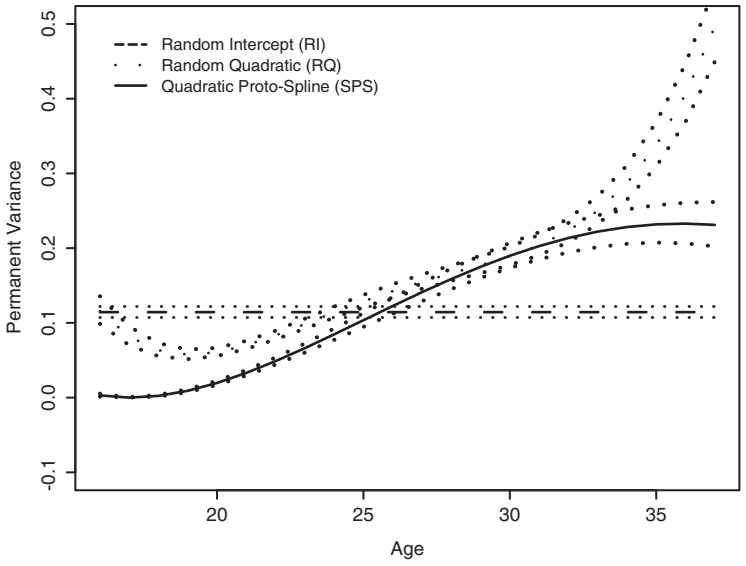

Figure 4: Permanent Variance of the Wage Process Over Time, All Three Models

nature of the latent curve, and the actual process being modeled. The latter apparently provides substantial evidence of little to no permanent variation early on. Each model's estimates are, however, conditional on the model itself and, as such, do not incorporate model uncertainty. In section 5, we compare predictions from each model to further evaluate their validity. In section 7, we revisit the issue of model uncertainty from a Bayesian perspective. The permanent variance estimates for all three models are most divergent before age 24 and after age 33, and these differences are highly significant. So we are witnessing significant differences over time in what should be labeled permanent variance.

The SPS model suggests that permanent variation is extremely small in the early labor market experience. It predicts that substantial growth in permanent differences follows, but the prediction is significantly less than the RQ model estimates at the oldest ages. We must consider the definition of "permanent" differences a bit further. The RI model implicitly defines permanent wage to be average wage, while the RQ model allows it to take any quadratic form. The SPS model requires that the wage levels at young ages inform

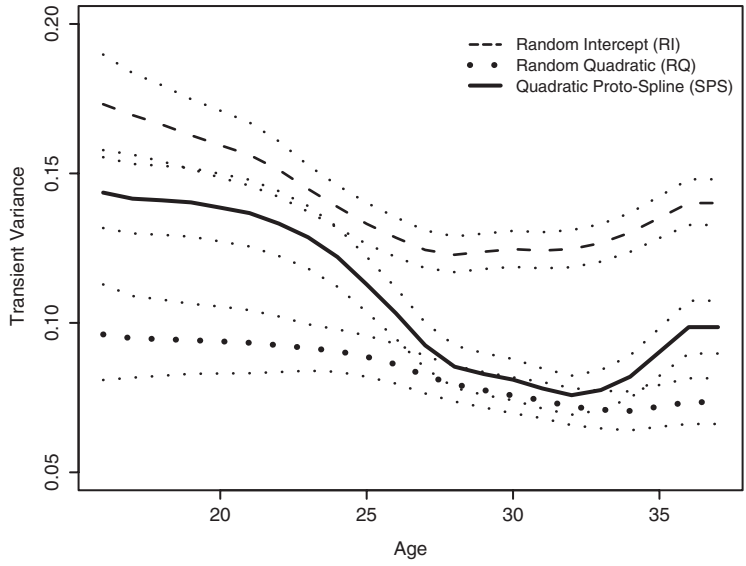

Figure 5: Transient Variance of the Wage Process Over Time, All Three Models

those at older ages via a latent curve that is based on population-level information. So the latter model amasses evidence as to how the past influences the present and future for the population as whole. The three models (and definitions) tell different stories: permanent differences are constant (RI), they dip and then grow dramatically (RQ), and they are insignificant for young workers but become more and more relevant over time (SPS). The latter is most consistent with the job churning and matching process, in which a combination of more and less successful employment experiences dominates the early period, eventually stabilizing.

The complementary transient variation is displayed in Figure 5. The RI model indicates some decline over time, while the RQ model fit shows less of this effect and is overall much lower. It is clear that the latter model apportions substantially more variation to permanent differences between workers and that this grows over time. For the RI model, there is very mild evidence of an upswing in transient variation for workers beginning at age 34, but this is not so for the RQ model. The SPS model is quite literally in between its two

competitors, with a dramatic dip in transient variation by age 27 and a significant upward tick for the oldest workers. This variance decomposition is highly consistent with the job churning and matching process, each of which is hard to measure directly. It is consistent with wages permanently fanning out and transient variance subsiding. The upswing at the older ages could indeed be an economic trend consistent with a recessionary period beginning in the early 1990s.

A goodness of fit cannot be evaluated on the variance partitions as neither is observed empirically. However, we can look at the combination of these two components and see whether they resemble the observed total variation. This is presented in Figure 6, and it is clear that the RI model is doing a poor job of fitting the shape of the empirical variation (represented by the thickest line). The RQ model shows very poor fit for older workers and somewhat poor fit for the youngest workers. The total variance is matched best by the SPS model, giving us evidence that our overall picture of the wage process is well described by the corresponding decompositions.

We summarize these findings further in section 6, but for now we describe the three different conclusions one might reach from these models as follows: Transient variance dominates (RI); permanent variance dominates while transient variance is somewhat stable—the period of job churning is not accompanied by substantially more volatility (RQ); permanent and transient variances are about of equal impact on overall variation, with transient effects dominating early and permanent effects dominating later, and the partitions are consistent with job churning and downsizing scenarios (SPS). It seems pretty clear that wages fan out permanently, but nonsubtle differences between the RQ and SPS models remain.

5. ADDRESSING RESEARCH QUESTIONS WITH HYBRID MODELS

How can we use these models to delineate between competing hypotheses? We focus on the permanent variance component and illustrate the strength of hybrid functional models in this arena by examining two features of the RQ model partitions that are highly questionable, substantively. The first is that permanent variance is higher at age 16 than at age 20. The second is that permanent

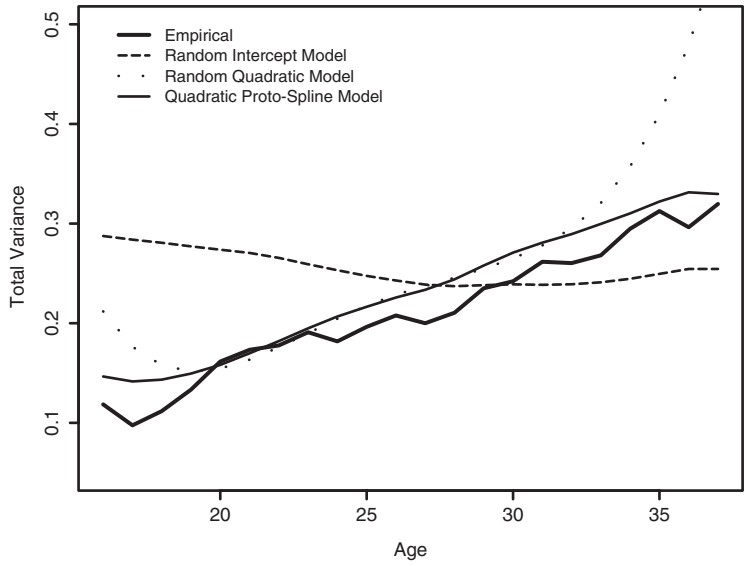

Figure 6: Total Variance of the Wage Process Over Time, Empirically, and for All Three Models

variance is well above 0.20 for workers in their mid-30s. Given the lack of fit implied by the total variation curves in Figure 6, there is evidence that both conclusions may be inaccurate.

In formulating a model that can inform the first of these concerns, we ask why would permanent variation go through a period of decline and then pick up again from ages 16 to 23? The scenario implied by the RQ model fit is likely to resemble the following. Workers who start out a bit higher than the mean see those initial benefits decline or at least stagnate over time, while workers starting out much lower apparently make back those initial losses through later growth. This is consistent with an "education effect" in which some workers experience large wage growth postschooling, surpassing the gains of the less educated (but more experienced) workers in the long run. If we consider two prototypical trajectories, one with high growth and the other beginning higher but then stagnating, the variance would indeed be described by a "crossover" period with lowest variance, and the information about initial wages would be transmitted through the churning

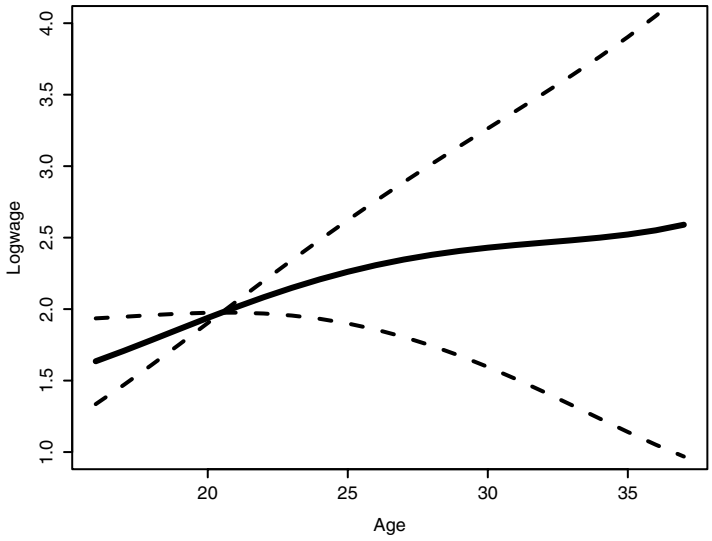

Figure 7: Two Stylized Trajectories Surrounding the Overall Mean for the RQ Model (Under This Scenario, Permanent Variance Would Reach a Minimum at Age 21)

period. A stylized depiction of this scenario is provided in Figure 7. An alternative is that there is more wage volatility in this early period, wherein workers may have much higher or lower wages initially, but these have little influence on the remaining trajectory.

We note that the "education effect" scenario suggests the presence of two distinct subgroups. These may be well identified in the data, either through covariates or as well-separated clusters. However, education effects are by no means uniform in the NLSY, and clusters are easily masked by a large overall level of variation in trajectories. A latent class extension to proto-spline models, described in Scott and Handcock (2001c), could be used to gauge support for a two-subgroup model. However, a single-curve proto-spline model, with the proper choice of function space, works as well and illustrates the utility of this modeling class.

Since individual differences in SPS models are built up through rescaled versions of a single curve, a function space that allows early wage levels to inform the remaining trajectory—without requiring

that they do so—will differentiate between the education and alternative hypotheses. We fit another SPS model using a cubic spline function space with knots chosen to allow bends at ages 18, 21, 24, and 32; the first three are critical to the hypothesis being tested, while the last will address our concern regarding variance at older ages. By allowing the proto-spline curve to bend in several locations, the point at which wage differences become *permanent* is apparent in the fitted curve $\hat{\phi}_1$. Since that curve reflects extra-mean differences, an initial drop in permanent variation would entail a change of sign from negative to positive at the crossover point (the permanent variance is captured in the square of this curve). The resulting fit from this more flexible function space is given as the dashed line in Figure 1. It fails to cross zero, even negligibly, suggesting little to no evidence for initial differences carrying forward into later wage trajectories. The transient and permanent variance prediction for the cubic spline SPS model was practically indistinguishable from the quadratic SPS, so for simplicity of presentation, we exclude the former from most model summaries.

Given our conclusions regarding the first hypothesis, we now evaluate the second: Could differences between older individuals be as large as is suggested by the RQ model? We anticipated this question in our choice of function space with the inclusion of a knot at age 32. If permanent variance were larger for older workers, the proto-spline curve should bend to reflect this. The fitted cubic spline curve in Figure 1 shows no indication of increased permanent variation for workers in their mid-30s. The evidence is thus quite substantial against the decomposition of wage variation implied by the RQ model. Moreover, a reasonable alternative—that there has been a mild increase in wage volatility for older workers—is more consistent with the downsizing trend of that period and is clearly suggested by the proto-spline fit.

We conclude that the permanent variance decomposition for the RQ model is potentially misleading, particularly at the most extreme ages. We developed the notion of the model dependence of partitions in some detail in Scott and Handcock (2001a), but in this study we go further and claim that a very common approach to variance partitioning is overly sensitive to short-term variation. The model can also be characterized as allowing many points to exert high

influence *on the covariance structure*. Some insight into the failure of RQ models, especially at the extremes, can be gained in this case.

In these data, we do not have a single complete trajectory. Instead, workers are followed for 16 years but starting at different ages between 14 and 21. We witness profiles from ages 14 to 29, 15 to 30, and so forth, with profiles from ages 21 to 36 for the oldest members of the cohort. We note that some workers are actually observed as late as age 37, and we drop any single person-period observed at age 14 or 15 because the labor market for very young workers is substantively different. The set of person-specific profiles effectively spans the full age range of 16 to 37 but is never complete for a single worker. The mixed-effects RQ model is less able to accurately borrow strength from the information from different periods to describe the whole trajectory. It cannot "patch together" the local information from parts of trajectories into a common one because it lacks a mechanism to transmit information from one period of observation to another. In contrast, the common continuous curve in proto-spline models provides such a mechanism.

Further evidence of overfitting in the RQ model involves the following comparison. We extrapolated model-based predicted curves to the full age range (16-37), using best linear unbiased predictions (BLUPs). We then used these predictions in two different ways. Using the full age range for each worker, we computed empirical permanent variance curves from the predictions. These reproduced the RQ model-based permanent variance curves very well. We then repeated this analysis but restricted the age range included for each worker to the *observed* ages for that worker. In other words, we did not extrapolate beyond the points of observation. The empirical permanent variance based on this restriction was initially much flatter and showed little of the crossover implied by the model parameters. If the RQ model were picking up something "real" in the first period, then it should have been reflected in the restricted estimates. It is more likely that by following the data too closely, the curvature at the extremes is overestimated, misclassifying the variance. Further evidence of what is a truly permanent wage in this domain comes from the cubic spline proto-spline model, which would have picked up and *transmitted* early differences through the early period of churning but did not.

TABLE 1: Model Summary

Comparison of Models			
Model	*df*	Bayesian Information Criterion (BIC)	Log Likelihood
Random intercept	14	21001	−10429
Random quadratic	19	15917	−7862
Proto-spline	16	19652	−9745

6. MODEL SUMMARY

We have evidence that hybrid functional models track the empirical variation in the process better than competing models and that they are able to address rather sophisticated questions in a substantive field. It is interesting to note that they do not measure up quite as well when standard model selection approaches using information criteria are applied. In Table 1, the BIC and log likelihoods for three of the models presented are compared. The SPS model performs better than the RI model, but the RQ model has the best overall BIC. This is partially due to divergent purposes in modeling: BIC selects a parsimonious model that best fits the observed data, while the SPS model and even the RI model capture long-term variation in a manner that assigns a meaningful, hybrid interpretation to the variance components. The SPS approach allows the shape of the person-specific effect to be more complex than a simple intercept shift, depicting the long-range dependence within trajectories in a common shape. The RQ has no such aims and in fact may be overfitting, in terms of what is substantively important.

We make a further claim that the proper comparison is between RI and SPS models. This is based not on the number of parameters in the model, which are accounted for by BIC, but on the number of *random structured components* in the model. We find very little discussion of this point in the literature. The RI and SPS models each have one random component driving the differences between individuals over time. The RQ has three. Three ‘‘random degrees of freedom’’ offer substantial flexibility that gets reflected in the log likelihood but not in the degrees of freedom accounted for by BIC.

TABLE 2: Analysis of Variance

Model-Based Extra-Mean Variance Partitions						
Source	Random Intercept	Percentage	Random Quadratic	Percentage	Proto-Spline	Percentage
Permanent variance	0.1144	45.4	0.1427	62.9	0.1166	52.1
Transient variance	0.1377	54.6	0.0842	37.1	0.1072	47.9
Total variance	0.2520	100.0	0.2268	100.0	0.2238	100.0

This may be at the cost of potential overfitting, of which we have amassed substantial direct and indirect evidence. We note that BIC may be helpful in comparisons of proto-spline models employing different function spaces or different knot choices, particularly when the remaining model structure is very similar.

In Scott and Handcock (2001a), we develop a double proto-spline model, which includes the effects of two independent latent curves in each trajectory. This model contains two random degrees of freedom (one for each curve). We fit this model to our data and found its BIC was better than that for a random slope mixed model (the natural comparison), and it approached the BIC of the RQ model, which has three random degrees of freedom.

We have repeatedly emphasized the strength of hybrid functional models to address substantive hypotheses. We now report the variance partition associated with each model. This partition reflects each model's best attempt to separate signal from noise under different definitions of signal. Age-weighted summaries of the decompositions previously reported in Figures 4 through 6 are given in Table 2. We first note that permanent variance represents the majority of variation in all but the RI model. Substantively, there is a clear need for a model that captures more than mean differences in trajectories, so the variance partition for this simpler model is suspect. What may be less immediate is the difference in the proportion of permanent variance for the RQ and SPS models, which are 63 percent and 52 percent, respectively. The original motivation for this analysis involved a comparison *between cohorts* facing very different labor markets, and any temporal change in the proportion reported above reflects a change in long-term mobility prospects for workers. Thus, the choice of model and the resulting variance partition have serious consequences.

These last two partitions show similar total variation, but both overestimate the empirical variance, which is 0.2043. This was hinted at in Figure 6, since our best models still tend to be above the empirical estimate. The cubic spline proto-spline findings, not reported above, suggest a further decrease in transient variation, to 0.0976 and a total variance of 0.2152. We take this as evidence that we could improve our modeling somewhat still.

7. DISCUSSION AND CONCLUSION

We have presented a model for the variation in longitudinal data that formalizes and extends the original idea of latent curve analysis proposed by Meredith and Tisak (1990). The approach is rooted in the theory of stochastic processes and functional data analysis. The model is hybrid in nature, with an interpretation somewhere between those of population-average and individual-specific longitudinal models. The hybrid interpretation, combined with the proper choice of function space, can help to resolve important substantive hypotheses. In a labor economic application using the NLSY, we customized our function space to investigate several hypotheses concerning permanent variance in young worker trajectories. In this context, the continuity of the curve and its interpretation as a population-level feature provided a substantively meaningful link from one period to another.

A strength of hybrid functional models, of which proto-splines are an important class, is their interpretability. The interpretation is part of a philosophically distinct modeling approach and thus not only generates new knowledge with its use but also establishes a new way to think about and "allocate" the variance components associated with longitudinal data. It is clear from our development that the choice of the function space and, in our application, the cubic spline knots are key to appropriate modeling of the covariance, and we have demonstrated the sensitivity of findings to model specification.

Given this sensitivity, further exploration of model uncertainty is recommended. What was striking about the variance decompositions for these three models was that they overlapped very little, particularly at the early and late ages, and how "certain" each model's findings were (most confidence bands were quite small and nonoverlapping

across models). A Bayesian model-averaging approach, in which model uncertainty is specified a priori, would allow us to say more about the variance partitioning and the viability of the various proposed models, a posteriori (see Hoeting et al. 1999).

Based on our comparative analyses, we conclude that the proto-spline model provides the variance decomposition most consistent with the NLSY data set and our knowledge of that economic period. Permanent variance accounts for about 52 percent of the total variation. How this evolves over time is a key component of the decomposition, and, as witnessed in Figure 4, the most plausible scenario involves steady but nonaccelerating growth in wage disparity. The change in transient variance over time, depicted in Figure 5, suggests initially high levels that drop substantially by the mid-20s. This is consistent with an initial period of "job churning" for young workers. Notably, there is an uptick in transient variance for older workers, which is consistent with the effects of firm downsizing in this period. The economic structure implied by the three models in Figures 4 and 5 is quite different, and we have amassed sufficient evidence to cast doubt on the commonly employed constant or quadratic random-effects models. To get meaningful estimates of inequality's persistence, we must look to hybrid functional models for covariation.

REFERENCES

Barry, Daniel. 1995. "A Bayesian Model for Growth Curve Analysis." *Biometrics* 51:639-55.

Bernhardt, Annette D., Martina Morris, Mark S. Handcock, and Marc A. Scott. 2001. *Divergent Paths: Economic Mobility in the New American Labor Market*. New York: Russell Sage Foundation.

Besse, Philippe, Hervé Cardot, and Frédéric Ferraty. 1997. "Simultaneous Non-Parametric Regressions of Unbalanced Longitudinal Data." *Computational Statistics and Data Analysis* 24:255-70.

Brumback, Babette A. and John A. Rice. 1998. "Smoothing Spline Models for the Analysis of Nested and Crossed Samples of Curves." *Journal of the American Statistical Association* 93:961-76.

Chamberlain, Gary. 1984. "Panel Data." In *Handbook of Econometrics*, edited by Zvi Griliches and M. D. Intriligator. Amsterdam: Elsevier Science.

Danziger, Sheldon and Peter Gottschalk. 1993. *Uneven Tides: Rising Inequality in America*. New York: Russell Sage Foundation.

Davidian, Marie and David M. Giltinan. 1995. *Nonlinear Models for Repeated Measurement Data*. London: Chapman & Hall.

Diggle, Peter J., Kung-Yee Liang, and Scott L. Zeger. 1994. *Analysis of Longitudinal Data*. Oxford, UK: Oxford University Press.

Gottschalk, Peter and Robert Moffitt. 1994. "The Growth of Earnings Instability in the US Labor Market." *Brookings Papers on Economic Activity* 2:217-72.

Guo, Wensheng. 2002. "Functional Mixed Effects Models." *Biometrics* 58:121-8.

Hoeting, Jennifer A., David Madigan, Adrian E. Raftery, and Chris T. Volinsky. 1999. "Bayesian Model Averaging: A Tutorial." *Statistical Science* 14:382-417.

James, Gareth M. 2002. "Generalized Linear Models With Functional Predictors." *Journal of the Royal Statistical Society Series B* 54:411-32.

James, Gareth M., Trevor J. Hastie, and Catherine A. Sugar. 2000. "Principal Component Models for Sparse Functional Data." *Biometrika* 87:587-602.

Ke, Chunlei and Yuedong Wang. 2001. "Semiparametric Nonlinear Mixed-Effects Models and Their Applications." *Journal of the American Statistical Association* 96:1272-98.

Kneip, Alois. 1994. "Nonparametric Estimation of Common Regressors of Similar Curve Data." *Annals of Statistics* 22:1386-1427.

Levy, Frank and Robert Murnane. 1992. "U.S. Earnings Levels and Earnings Inequality: A Review of Recent Trends and Proposed Explanations." *Journal of Economic Literature* 30:1333-81.

Lindstrom, Mary J. 1995. "Self-Modelling With Random Shift and Scale Parameters and a Free-Knot Spline Shape Function." *Statistics in Medicine* 14:2009-21.

Longford, Nicholas T. 1993. *Random Coefficient Models*. Oxford, UK: Oxford University Press.

Meredith, William and John Tisak. 1990. "Latent Curve Analysis." *Psychometrika* 55:105-22.

Murphy, Kevin and Finis Welch. 1990. "Empirical Age-Earnings Profiles." *Journal of Labor Economics* 8:202-29.

Pinheiro, José C., Douglas M. Bates, and Mary J. Lindstrom. 1994. "Model Building in Nonlinear Mixed Effects Models." Pp. 1-8 in *ASA Proceedings of the Biopharmaceutical Section*. Alexandria, VA: American Statistical Association.

Ramsay, James O. and Bernard W. Silverman. 1997. *Functional Data Analysis*. New York: Springer.

Rao, Calyampudi R. 1958. "Some Statistical Methods for Comparison of Growth Curves." *Biometrics* 14:1-17.

Rice, John A. and Bernard W. Silverman. 1991. "Estimating the Mean and Covariance Structure Nonparametrically When the Data Are Curves." *Journal of the Royal Statistical Society Series B* 53:233-43.

Schwarz, Gideon. 1978. "Estimating the Dimension of a Model." *The Annals of Statistics* 6:461-4.

Scott, Marc A. 1998. "Statistical Models for Heterogeneity in the Labor Market." Ph.D. dissertation, New York University.

Scott, Marc A. and Mark S. Handcock. 2001a. "Covariance Models for Latent Structure in Longitudinal Data." *Sociological Methodology* 31:265-304.

Scott, Marc A. and Mark S. Handcock. 2001b. "Covariance Models for Latent Structure in Longitudinal Data." Working Paper 14, Center for Statistics and the Social Sciences, University of Washington, Seattle.

Scott, Marc A. and Mark S. Handcock. 2001c. "Covariance Models for Latent Structure in Clustered Longitudinal Data." In *ASA Proceedings of the Bayesian Statistical Sciences Section*. Alexandria, VA: American Statistical Association.

Wang, Yuedong. 1998. "Smoothing Spline Models With Correlated Random Errors." *Journal of the American Statistical Association* 93:341-8.

Marc A. Scott is an assistant professor of educational statistics at the Steinhardt School, New York University. His research interests center on the development of statistical models for longitudinal data. Models for continuous and discrete outcomes are of interest both separately and in how they jointly unfold. Current research involves latent space models for the analysis of career sequences, particularly in the low-wage labor market, and the evaluation of spatial variation in resources and outcomes within the New York City public school system. Recent articles include "Competing Risks Event History Models for Educational Attainment," forthcoming in the Journal of Educational and Behavioral Statistics*, and "Change-Point Models for Growth and Decline of Lung Function in Children with Duchenne's Muscular Dystrophy," published in* Applied Statistics.

Mark S. Handcock is a professor of statistics and sociology at the University of Washington, Seattle. His research involves methodological development and is based largely on motivation from questions in the social sciences. It focuses on the development of statistical models for the analysis of social network data, spatial processes, and longitudinal data arising in labor economics. Recent applications have been to social relations networks with the objective of understanding the social determinants of HIV spread and the combination of survey and population-level information. Other applications have been to models for stream networks that combine information from multiple environmental surveys. Recent work involving social networks has been published in the Journal of the American Statistical Association *and* Nature.