Approaches to Modeling Self-rated Health in Longitudinal Studies: Best Practices and Recommendations for Multilevel Models

Isaac Sasson
Department of Sociology and Population Research Center
University of Texas at Austin
305 E. 23rd Street, Stop G1800
Austin, TX 78712
Email: isasson@prc.utexas.edu

December 18, 2012

Word count: 8,672 (including text, notes, and references)

Approaches to Modeling Self-rated Health in Longitudinal Studies: Best Practices and Recommendations for Multilevel Models

Abstract

Self-rated health (SRH) is a key measure in the study of population health with proven external validity in predicting mortality. Nevertheless, failing to address the measure's ordinal scale in statistical analyses poses a potential threat to internal validity. Despite the advent of rich panel data, sociologists have generally been slow in adopting longitudinal methods for ordinal outcomes, and many discard valuable information in favor of simpler methods. This paper reviews and contrasts several approaches to modeling SRH in longitudinal studies under the generalized linear mixed model framework. Model performance is compared (e.g., linear versus nonlinear, conditional versus marginal) using simulation and data from the Health and Retirement Study. Findings suggest that conditional cumulative-logit models provide more statistical power than their linear counterparts, but result in similar substantive conclusions. By contrast, dichotomizing SRH significantly reduces power and is ill-advised. The paper concludes with recommendations for modeling ordinal outcomes in longitudinal studies.

Introduction

Self-rated health (SRH) is perhaps the most commonly studied health measure among demographers, epidemiologists, and sociologists of health. It has been utilized in numerous studies for both descriptive and inferential purposes and shown to have significant external validity with respect to predicting mortality (Kaplan and Camacho 1983; Idler and Benyamini 1997; DeSalvo et al. 2006). Despite the increasing availability of rich longitudinal data, a review of the literature suggests that researchers often opt for cross-sectional analyses of SRH using a single wave of data or focus on change in scores between only two waves (e.g., Hughes and Waite 2009; Baker et al. 2001; Hughes et al. 2007; Luoh and Herzog 2002). Such practices discard an enormous amount of useful data that are central in describing complex health trajectories over time. The use of longitudinal data is of paramount importance for studying population level health, as cross-sectional data seriously underestimate the deterioration of health with aging and fail to reflect the progression of SRH over the life course (Orfila et al. 2000).

One of the principal reasons for the underutilization of longitudinal data, at least in the case of SRH, is the lack of consensus about appropriate methodology for longitudinal ordinal outcomes. Even when elaborate statistical methods are used to model SRH longitudinally, limited attention is explicitly given to its ordinal scale and researchers often default to methods designed for continuous outcomes (Benyamini et al. 2009; Liu 2012; Meadows 2009; Sacker, Worts, and McDonough 2011; Wilmoth, London, and Parker 2010). The use of ad hoc methodology not only threatens the validity of results, but also adds considerable difficulty when comparing results across studies.

This paper aims to alleviate such ambiguities by reviewing and comparing several leading approaches to modeling SRH in longitudinal studies. First, I discuss the unique characteristics of SRH and key concerns when modeling ordinal outcomes in cross-sectional studies. Second, I review a general framework for modeling longitudinal data, with particular attention to ordinal outcomes. Models for longitudinal data are classified using a two-dimensional scheme: linear versus nonlinear (e.g., logit) and marginal versus conditional (i.e., mixed models). Third, I compare the performance of several modeling approaches using both data from the Health and Retirement Study and simulation with known population parameters.

Statistical methods can serve multiple purposes including description (data reduction), prediction, and explanation (inference) – each having distinct implications for modeling practices, with the latter typically being of most interest to social scientists (Shmueli 2010). Model selection is particularly difficult across different model families and assessing model fit depends, at least in part, on the analyst's goals. Thus, in comparing longitudinal approaches to modeling SRH, I chose to emphasize statistical inference and substantive interpretation over other goals of statistical analysis. Specifically, models are evaluated on the basis of making correct inference and their statistical power.

Finally, I provide a practical discussion for researchers on the trade-offs associated with each modeling approach. The practice of dichotomizing SRH is also evaluated in the context of longitudinal studies and commented on. Despite the focus on SRH, much of the following discussion can easily extend to other ordinal outcomes with similar characteristics.

Characteristics of Self-rated Health

The measurement of SRH is not entirely consistent across studies. Differences exist in the number of levels measured (Eriksson, Undèn, and Elofsson 2001), respondents' frame of reference (Bailis, Segall, and Chipperfield 2003; Krause and Jay 1994), and health perceptions across cultures and social groups (Jylhä et al. 1998). In terms of statistical modeling, however, these differences are of little importance. For illustrative purposes, I adopt the Health and Retirement Study version: "Would you say your health is excellent, very good, good, fair, or poor?" (HRS 1992). Three important characteristics of SRH are relevant to its statistical analysis as a dependent variable:

1. Discrete – while self-rated health is assumed to represent a latent (continuous) construct, it is typically measured discretely on an ordinal scale (e.g., Poor-Fair-Good-Very good-Excellent).
2. Reversible – fortunately, and despite our long-term mortal expectation, health can in fact improve rather than simply decline over the life course. That is, in terms of measurement, the transition from each category of SRH to any other category is permissible. Transitions between categories over time do not have to be consecutive either, so that a hypothetical person can transition from having "Very Good" to "Poor" health without crossing categories in between. Note that this is not a matter of unobservable data but part of the natural process of change in health.
3. Asymmetric/monotonic – this property separates SRH from a Likert-type scale (e.g., agree-neutral-disagree) in that it has no meaningful central category of

reference. Scale asymmetry is not to be confused with *empirical* asymmetry, whereby the distribution of SRH is often skewed, at least in the general population, as majority of people are relatively healthy throughout most of their lives. Skewness in the distribution of SRH, of course, also has implications to statistical analyses (e.g., methods based on normal approximation) aside from scale asymmetry.

The inherent characteristics of SRH should be taken into consideration when constructing statistical models of health over time. While cross-sectional studies generally seem to adhere to these attributes, longitudinal studies are far from showing convergence in methodology. The following section discusses methodological concerns in cross-sectional models of ordinal outcomes, many of which carry over to the longitudinal context.

Modeling Ordinal Outcomes in Cross-Sectional Studies

Ordinal outcomes are often modeled using methods designed for nominal or interval scales, with the former ignoring any information given by the ordering of the categories and the latter treating them as arbitrarily (most often evenly) spaced. When applied to ordinal variables, multinomial models tend to have less power than ordinal methods in detecting associations (Agresti 2010:3), as they involve additional parameters and, consequently, fewer degrees of freedom in statistical inference. More often, however, researchers choose to collapse categories or simply dichotomize the outcome variable. When categories represent arbitrary cut-points of an unobserved continuous scale, as in

the case of SRH, then reducing the number of categories can result in bias and loss of power (Ananth and Kleinbaum 1997; Agresti 2010:38).

Methods designed for continuous data (i.e., OLS[1] regression and its derivatives) may perform better than nominal models with regard to SRH, but they too carry notable shortcomings. Primarily, these methods assign a score to each category assuming some meaningful difference-value between them. For example, by using the standard scoring of SRH (ranging from 1 to 5) we implicitly assume that moving from "Excellent" to "Very Good" health has an equivalent meaning as moving from "Fair" to "Poor" health, as the intervals between all categories are the same. While the choice of scores is not limited to linearity (e.g., one can choose {1, 2, 4, 7, 8} rather than 1-5), this choice nonetheless requires justification and typically introduces a degree of arbitrariness to the outcome. OLS models with ordinal outcomes are also known to suffer from floor and ceiling effects, leading to biased parameter estimates that exceed the original scale, and causing residuals to be correlated with predictor variables (Agresti 2010:5). These in turn may result in "false-positive" detection of interaction terms between covariates.

Generalized linear models (GLM) extend ordinary regression models by allowing the dependent variable to follow a distribution other than normal and the link function to be other than identity. A linear combination of predictors now relates to some function of the mean (e.g., logit, probit, etc.) rather than to the mean itself. Most commonly, in the case of ordinal models, these correspond to the multinomial distribution (a multivariate GLM) and the logit link. Fullerton (2009) provides a thorough typology of ordered logistic models based on the choice of numerator and denominator in the logit link and the proportional odds assumption (i.e., that regression coefficients are constant across

levels of SRH). According to the typology, the **cumulative** logit link function models the probability of being at *or below* each category relative to all categories above it. The cumulative approach is often justified by assuming that an unobserved continuous variable underlies the observed categorical variable (Agresti 2010:53). Alternatively, the **stage** (also known as continuation-ratio) approach models the probability of being in a stage (category) compared to stages above it. In this approach one must pass through categories successively and irreversibly, such as stages in educational attainment. Finally, the **adjacent-category** approach compares a single category to another category of choice (strictly, this method is nominal rather than ordered) and is most useful when a meaningful midpoint exists for the outcome (e.g., an "agree-neutral-disagree" formulation). Table 1 summarizes the corresponding logit functions for each of the three model types.

Table 1: logit functions for common ordinal models, with J-1 cut-points for J categories.

Cumulative	Stage	Adjacent-Category[a]
$Log\left(\frac{\Pr(Y \leq j)}{\Pr(Y > j)}\right)$	$Log\left(\frac{\Pr(Y = j)}{\Pr(Y > j)}\right)$	$Log\left(\frac{\Pr(Y = j)}{\Pr(Y = j')}\right)$

[a] In the adjacent-logit model j' is the reference category.

A major advantage of the cumulative-logit approach, given that the proportional odds assumption holds, is that the original latent variable directly relates to the linear predictor (i.e., $E[Y^*] = \mathbf{X}\beta$). Contrary to the OLS model, regression coefficients derived from the cumulative-logit model are invariant to the choice of response categories

(McCullagh 1980; Agresti 2010:56). Thus, by using an ordinal model we rid ourselves from choosing – and having to justify – arbitrary values for the outcome variable.

Given the characteristics of SRH (monotonic scale, arbitrary cut-points, representing a latent construct), the cumulative-logit approach seems to fit most among the GLM link functions listed above. This is the approach I adopt for the rest of the paper and extend to the longitudinal context.

Longitudinal Models for Ordinal Outcomes

Repeated measures of individuals' health status over time are generally treated within one of two major frameworks: multilevel models (MLM hereafter; also known as HLM for hierarchical linear models), and latent growth models (LGM) stemming from the SEM tradition (Meredith and Tisak 1990). While the following discussion emphasizes the MLM specification, equivalent models can generally be implemented under the LGM framework. As others have noted (e.g., Jackson 2010), the distinction between the two frameworks is becoming increasingly vague and software-dependent with respect to longitudinal models and similar methodological concerns apply.

Under the MLM framework, SRH measurements are considered nested observations within individuals (level-1), while the individuals are considered independent, randomly selected blocks (level-2). The two-level model can be broken down into two conceptual stages corresponding to each model level: a) estimating growth parameters (intercept, slope, etc.) for each subject to summarize individual SRH trajectories; b) making inference about variation in growth parameters between subjects.

Thus, SRH is the level-1 dependent variable and growth parameters (e.g., random intercept, random slope) are the level-2 dependent variables.

In the same way that GLM extend OLS regression to discrete outcomes, generalized linear mixed models (GLMM) extend linear MLM. However, since there is more than one random variable in the multilevel model, assumptions are now made separately at each model level (e.g., choosing a random distribution and link function for each random variable). A two-level GLMM with random-intercept and random-slope follows[2]:

Level-1: $$g(\cdot)_{it} = \gamma_{0i} + \gamma_{1i} T_{it} + \sum_{p} \gamma_{pi} Z_{pit} \quad (1)$$

Level-2: $$\gamma_{0i} = \beta_{00} + \sum_{k} \beta_{0k} X_{ki} + u_{0i} \quad (2)$$

$$\gamma_{1i} = \beta_{10} + \sum_{m} \beta_{1m} X_{mi} + u_{1i} \quad (3)$$

$$\gamma_{pi} = \beta_{p0} \quad (4)$$

where, in equation (1), $g(\cdot)_{it}$ is a function of individual *i*'s expected outcome at time *t* (conditional on the random effects), γ_{0i} and γ_{1i} are random effects (i.e., growth parameters), T_{it} is time measurement for individual *i* on occasion *t*, and γ_{pi} are level-1 fixed-effects corresponding to *p* time-varying covariates, Z_{pit}. Equations (2) and (3) can be interpreted as in simple regression models, with random effects regressed on person-level covariates (X_i) that are fixed across measurements (e.g., race and gender). Equation (4) specifies additional fixed effects that can enter the model with time-varying covariates at level-1.

Of more interest is level-1, where the choice of a link function, $g(\cdot)$, and a random distribution for SRH *conditional* on random effects, allows us to generalize the

multilevel model to discrete outcomes. When $g(\cdot)$ is the identity link and the conditional probability distribution is assumed normal, we end up with the conventional linear MLM. Instead, choosing the cumulative-logit link function and a multinomial distribution for the level-1 outcome leads us to a GLMM. Regardless of the assumptions made at level-1, it is common to operate under the tractable assumption that level-2 random effects follow a multivariate (here bivariate) normal distribution. Even when the random effects distribution is misspecified, consistent and asymptotically normal parameter estimates can be obtained for linear multilevel models (Verbeke and Lesaffre 1997). However, with GLMM, misspecification of random effects may have more severe consequences (Hartford and Davidian 2000; Litiere, Alonso, and Molenberghs 2007).

Next, I present three competing longitudinal models for ordinal outcomes: (A) linear (normal) multilevel model; (B) conditional ordered-logit model; (C) marginal ordered-logit model with correlated observations. After introducing the models, I discuss methodological concerns that relate to all three of them.

Model A: Linear Multilevel

The linear multilevel model is commonly used in the literature for modeling longitudinal outcomes; it assumes the identity link function and normally distributed error terms in Equation (1), and a bivariate normal distribution for the random effects in Equations (2) and (3). With respect to SRH, the random-intercept suggests that individuals vary in their initial health status, and the random-slope that they also vary in the rate of change in health over time.

Applying the linear multilevel model to non-normal outcomes may bias our inference on estimated regression coefficients. ML estimates of fixed effects are asymptotically normal even when the error-terms are non-normal, although departure from normality requires larger sample sizes. Estimates of standard errors, on the other hand, may be biased downwards, inflating test statistics and providing a false sense of confidence in hypothesis tests. Thus, robust inference procedures are needed in order to draw correct conclusions with non-normal data. In the case of single parameter tests (e.g., the Wald test), this generally corresponds to Huber-White robust standard errors (Maas & Hox 2004). Tests for overall model fit also need to be adjusted for departure from normality. For models estimated using ML, Satorra and Bentler (1994) suggested a scaling factor for model chi-square statistics (known as the SB-scaled chi square test statistic[3]).

Model B: Conditional Ordered-Logit

As previously mentioned, GLMM extend standard GLM to include random effects as well as fixed effects in the linear combination of predictors. Equation (1) can be modified to incorporate a variety of link functions and probability distributions (as far as they can be estimated) to describe the outcome variable at hand. Given the characteristics of SRH, the cumulative-logit function and the multinomial distribution make the natural choice for level-1 in the multilevel model. Thus, we can substitute Equation (1) with:

$$Log\left(\frac{\Pr(Y_{it} \leq j)}{\Pr(Y_{it} > j)}\right) = \gamma_{0i} + \gamma_{1i}T_{it} + \sum_{p}\gamma_{pi}Z_{pit} \qquad (5)$$

In addition, the intercept term, β_{00}, in equation (2) is substituted by a separate intercept, β_{j00}, for each of the J SRH categories. The ordinal model estimates J-1 logits

concurrently, corresponding to the number of cut-points between SRH categories. Note that this specification takes the proportional odds form, assuming that fixed-effects are constant across all categories of SRH. Since growth parameters are considered latent variables, they are assumed to follow a bivariate normal distribution just as in the linear multilevel model.

Model B is termed *conditional* to emphasize that its fixed effects are interpreted conditional on random effects. In other words, effects are person-specific rather than population-averaged (this is in fact true for Model A too). When the link function is nonlinear, such as the logit link, and when considerable variation exists between individuals' health trajectories, conditional effects differ in magnitude from marginal (i.e., population-averaged) effects. Since the attenuation of marginal effects is accompanied by attenuation of standard errors, inferential statistics in both models is generally similar (Agresti 2002:501). However, conditional and marginal models differ substantially in interpretation, as discussed next.

Model C: Marginal Ordered-Logit

Marginal models are distinctly different from the previous models, as they are not truly multilevel and do not include random effects. In fact, they are specified as conventional GLM with the exception of relaxing the assumption of independence of observations. This type of model is termed marginal, as opposed to conditional, as it models the marginal distribution of Y_{it} averaged over all individuals. A general specification of a marginal linear model using the cumulative logit link is as follows:

$$Log\left(\frac{\Pr(Y_{it} \le j)}{\Pr(Y_{it} > j)}\right) = \alpha_j + \sum_k \beta_{0k} X_{ki} + \left(\beta_{10} + \sum_m \beta_{1m} X_{mi}\right) T_{it} \qquad (6)$$

Equation (6), however, does not explicitly reflect potential dependence among observations within clusters. Modeling the data under the assumption of complete independence would still result in unbiased regression coefficient estimates, but will likely produce underestimated standard errors (which in turn affect inference).

Technically, specifying the correlation structure of within-cluster observations depends on the choice of software package and method of estimation. One approach is using ML estimation with robust (sandwich) estimators for standard errors that account for within-person clustering (Rabe-Hesketh and Skrondal 2008:300). Alternatively, when feasible, *quasi-likelihood* estimation (i.e., the GEE method) can be used by specifying a working correlation matrix that captures the within-cluster dependence (Agresti 2010:268). The method of estimation has important implications for missing data: while ML estimation generally relies on the assumption that observations are missing at random (MAR), quasi-likelihood methods make the stricter assumption that data are missing completely at random (MCAR) (Agresti 2002:501).

Marginal and Conditional Means in GLMM

Most importantly, marginal and conditional models differ in the expected values they produce and their substantive interpretation. As the name implies, conditional models yield person-specific expected values conditional on both fixed *and* random effects. For SRH, these are interpreted as the expected health trajectory for a person located at a specific place in the distribution of random effects (i.e., with below or above average initial health status and below or above average rate of change in health status). Marginal models, on the other hand, give rise to population-averaged trajectories conditional on

fixed effects alone. In effect, they average out individual heterogeneity in SRH. The difference between the two model types is made clear when written explicitly.

Consider a general GLM for longitudinal data[4]:

$$g(\mu_{it}) = \mathbf{x}_{it}'\beta \tag{7}$$

where $\mathbf{x}_{it}$ is a vector of covariate and $\boldsymbol{\beta}$ a vector of coefficients. The marginal mean for an individual with particular covariate values at time *t*, is simply:

$$\mu_{it} = E[Y_{it}] = g^{-1}(\mathbf{x}_{it}'\beta) \tag{8}$$

Now, consider a GLMM with additional random effects (Agresti 2002, 492):

$$g(\mu_{it}^{C}) = \mathbf{x}_{it}'\beta + \mathbf{z}_{it}'\mathbf{u}_{i} \tag{9}$$

where $\mathbf{z}_{it}$ is a vector of covariates and $\mathbf{u}_{i}$ a vector of random effects. Then, the conditional mean, μ_{it}^{C}, is given by:

$$\mu_{it}^{C} = E[Y_{it} \mid \mathbf{u}_{i}] = g^{-1}(\mathbf{x}_{it}'\beta + \mathbf{z}_{it}'\mathbf{u}_{i}) \tag{10}$$

And the marginal mean, μ_{it}^{M}, is:

$$\mu_{it}^{M} = E[Y_{it}] = E_{\mathbf{U}}[E(Y_{it} \mid \mathbf{u}_{i})] = \int g^{-1}(\mathbf{x}_{it}'\beta + \mathbf{z}_{it}'\mathbf{u}_{i}) f(\mathbf{u}_{i}) d\mathbf{u}_{i} = h(\mu_{it}^{C}) \tag{11}$$

The last equation clearly shows that in a GLMM, the marginal and conditional means are not equivalent and that the former is a function of the latter (specifically, averaging over the individual heterogeneity reflected in random effects).

Note that when *g* is the identity link, the conditional mean becomes an additive function of fixed and random effects (Ritz and Spiegelman 2004), which implies a special case where:

$$\mu_{it}^{M} = E_{\mathbf{U}}[E(Y_{it} \mid \mathbf{u}_{i})] = E_{\mathbf{U}}[\mathbf{x}_{it}'\beta + \mathbf{z}_{it}'\mathbf{u}_{i}] = \mathbf{x}_{it}'\beta + \mathbf{z}_{it}'E[\mathbf{u}_{i}] = \mathbf{x}_{it}'\beta \tag{12}$$

Thus, in the linear case, when random effects are at their zero means, we have that $\mu_{it}^{C} = \mu_{it}^{M} = \mathbf{x}_{it}'\beta$. However, this is generally not the case with other link functions, including as the logit.

Additional Considerations with Longitudinal Models

Statistical modeling requires making assumptions and models for longitudinal data are no different. However, the empirical literature is often mute with respect to many of the assumptions involved in longitudinal data analysis, regardless of the scale of measurement (continuous/ordinal) or framework (MLM/LGM) used. Whether implicitly or explicitly disclosed, several aspects should be considered systematically throughout the modeling process:

1) *Probability distributions and link functions*. As noted above, in the multilevel context probability distributions and link functions are chosen separately at each model level. Since growth parameters in level-2 are generally considered latent variables, a common choice is the multivariate normal distribution along with the canonical identity link (albeit other possibilities exist). Level-1 assumptions can more directly accommodate the observed outcome's (e.g., SRH for individual *i* at time *t*) characteristics. As with GLM for cross-sectional data, one can choose a normal distribution and the identity link for a conventional continuous interpretation; alternatively, the characteristics of SRH can be addressed explicitly by choosing a multinomial distribution and the cumulative-logit link.
2) *Functional form of the dependent variable trajectory*. Rather than imposing an arbitrary functional form (linear, quadratic, cubic, piecewise, etc.) on individuals'

health trajectories over time, this choice should stem either from theory or from empirical results. Under the MLM framework, random effects can be tested sequentially for best model fit and included or omitted as necessary. Discontinuities can be tested using piecewise models when theory dictates such effects are plausible (for example, an abrupt deterioration in health following a stressful event). Similarly, under the LGM framework, one can extend beyond linear trajectories by adding factors that correspond to higher polynomials and test model fit. A more parsimonious alternative with LGM is the unspecified two-factor model (Duncan, Duncan, and Strycker 2006:31), which includes a factor for the intercept, specified as usual, and a second latent factor with only the first and second loadings fixed while any additional loadings are estimated freely. Since loadings on the second factor are now unrestricted to linearity, the resulting curve may follow more complex forms.

3) *Level-1 variance-covariance matrix.* Observations nested within individuals are unlikely to be independent of each other and the longitudinal model should account for significant departure from independence. For example, repeated SRH measurements over time are likely to be correlated within-person. The variance of SRH is also likely to increase over time, as individuals in the sample age and health disparities accumulate. The inclusion of person-specific random effects explicitly allows for within-person heteroscedasticity and autocorrelation (Singer and Willet 2003:84). This is evident in the composite error term, when equations (2) and (3) are substituted into equation (1). If within-person measurements are

still hypothesized to be correlated *conditional* on random effects, a more complex variance-covariance structure for level-1 can be specified.

4) *Level-2 variance-covariance matrix*. In growth curve models, variance components are of interest, as they indicate the amount of heterogeneity in individuals' smoothed growth trajectories, conditional on other covariates (Singer & Willet 2003:93). The covariance between growth parameters can also be estimated and often has a meaningful interpretation. For example, a significant negative correlation between SRH random-intercept and random-slope may indicate that respondents with high initial health status tend to show greater decline in health over time. When few growth parameters are included in the model the choice of an unstructured covariance matrix results in estimating few additional parameters. Depending on theory or empirical findings, especially with more complex functional forms of health trajectories, it is possible to impose restrictions on the covariance structure such as independence of growth parameters or equality in variance components.
5) *Time metric*. Researchers often use a variable indicating data wave as the temporal covariate in the model. Doing so implicitly assumes a discrete metric of time, where subjects are measured at exactly the same intervals. However, in large surveys it is typically the case that measurement intervals vary significantly between respondents. If more refined data are available, such as interview dates at each wave, a "continuous" measure of time may be calculated. Variables other than data wave, such as respondent's age, may be used as temporal measures in

the model. In general, the MLM framework better lends itself to modeling asynchronous measurements compared to the LGM framework (Jackson 2010).

6) *Model Estimation*. Estimation can be quite difficult with longitudinal data, especially with nonlinear models. For conditional models, Maximum Likelihood methods are generally used and are known to have several desirable properties, such as consistency and asymptotic normality (Singer and Willett 2003:65). In addition, ML estimates are unbiased even in presence of missing values as long as they are missing at random (Schafer and Graham 2002). However, when extended to include nonlinear link functions and non-normal distributions, the likelihood function cannot be evaluated directly and numerical methods are used for approximation. This process can become computationally cumbersome even with a moderate sample size. Alternatively, marginal models average the random effects across all individuals and are much more efficient in computation time. These models can be estimated using quasi-likelihood methods (using the GEE approach) without making explicit assumptions on the distribution of random effects (Zeger and Liang 1986). However, quasi-likelihood methods require the stricter assumption that data are missing completely at random (MCAR) (Agresti 2010:312). Furthermore, the interpretation of marginal models is fundamentally different from that of conditional models, as I illustrate in the next section. In general, marginal means can be extracted from conditional models, as the latter contain more information than do marginal models (Agresti 2002:500).
7) *Inference on fixed and random effects*. Single parameter tests for fixed effects in multilevel models are generally comparable to conventional regression analysis.

> When based on ML estimation these test statistics are asymptotically normal (such as the Wald test). Likelihood ratio tests can be used for testing random effects or overall model fit, at least with nested models. When using linear MLM/LGM with discrete outcomes, inference should be adjusted for departure from non-normality. Inference on random effects is more elusive and considered highly sensitive to imbalanced designs (Singer and Willett 2003:73).

Finally, a word about the choice between MLM and LGM is in order. Both frameworks permit the modeling of longitudinal ordinal outcomes, but may differ in flexibility. For example, MLM may be preferable over LGM when dealing with an asynchronous study design, where respondents are not measured concurrently or when there is particular interest in modeling the impact of specific life events that are experienced by individuals at different times. Conversely, LGM allow for more flexible error-structures and for elaborate models where growth factors serve as both dependent and independent variables. When trajectories cannot be easily approximated by a known mathematical function, LGM has the benefit of using unspecified growth factors to estimate trajectories freely. A more comprehensive discussion of the differences between MLM and LGM is available elsewhere (Ghisletta and Lindenberger 2004). As mentioned earlier, these differences are somewhat software dependent and may diminish even further in the future (Jackson 2010).

Example from the Health and Retirement Study

The cumulative advantage hypothesis suggests that health disparities between socioeconomic (SES) groups will widen over time and especially at old age. Indeed,

multiple studies have found diverging SRH trajectories by education, income, and race/ethnicity with age (Lynch 2003; Mirowsky and Ross 2008; Ross and Wu 1996; Shuey and Willson 2008; Willson, Shuey, and Elder 2007). In order to illustrate models A-C, I use longitudinal data from the Health and Retirement Study (RAND 2011)[5] to estimate SRH trajectories by education (<HS = less than high school; HS/GED = high school diploma or GED; >HS = some college or higher), net of other socio-demographic factors (age at baseline, gender, race, and Hispanic origin). For the sake of simplicity, all analyses in this section are unweighted and unadjusted for the Health and Retirement Study's complex survey design.

The Health and Retirement Study is a nationally representative, multi-stage probability sample of non-institutionalized older adults in the contiguous U.S. and their spouses. Respondents of the first birth cohort (1931-1941) were first surveyed in 1992 and then repeatedly at two-year intervals (the current sample includes nine waves of data up to 2008). Sample descriptive statistics are shown in Table 2.

[Table 2 here]

Being impartial with respect to the functional form of SRH over time, I first estimated an unspecified two-factor LGM (not shown here), which suggested that individual SRH progresses linearly over time. Thus, a random-intercept and random-slope model seems appropriate (a test of the random-slope model against a simpler random-intercept model also proved significant). Four models were estimated[6] for the sample data: (A1) linear multilevel model; (A2) linear multilevel model with robust standard errors; (B) conditional ordered-logit; (C) marginal ordered-logit. Inference on

fixed effects, if not coefficient estimates directly, can then be compared across all models. Results are summarized in Table 3.

[Table 3 here]

As expected, coefficient estimates are identical across models A1 and A2, with only standard errors differing as reflected in the associated p-values. In both models, age at baseline, Hispanic origin, and race (self-identified as black or other), but not gender, have significant negative effects on initial SRH. Relative to the reference category (less than high school education), having a high school diploma or GED is associated with .59 higher SRH at baseline; having some college education or higher is associated with .99 higher SRH at baseline.

For the average person, with mean initial SRH and a mean rate of change, SRH is expected to decline linearly over time at a rate of .033 per year. While the effects of age, gender, race (black only), and education on the rate of change in SRH are statistically significant, due to the large sample size, they are generally quite small even as they accumulate over several years (see Figure 1a). The effects of education on the slope of SRH suggest, perhaps, a slight convergence over time, rather than divergence. Note that the effect of Hispanic origin on the slope of SRH is slightly below the .05 level in Model A1, but slightly above the significance level in Model A2 (.04 and 058, respectively).

Coefficient estimated cannot be compared directly between the linear MLM (model A) and the nonlinear models (Models B and C). However, they can be compared with respect to substantive conclusions such as direction and significance of fixed effects, and, perhaps more importantly, with respect to predicted means and probabilities.

Fixed effects on the intercept of SRH (i.e., initial health status) in Model B and C resemble those of Models A1 and A2 in both statistical significance and direction. This is not the case, however, with respect to fixed effects on the slope of SRH. For example, the effect of age at baseline on the slope of SRH is positive and significant in Models A1, A2, and C, but negative and non-significant in Model B. Nevertheless, the effect is negligible in terms of magnitude and substantive interpretation, and may show significance simply due to the large sample size. Thus, it is better to interpret effects across models according to their theoretical and substantive importance, rather than simply based on significance level alone.

According to Model A1, women's SRH declines slower than does the slope for men. However, net of other factors, the expected gender difference in SRH amounts to a mere .064 after 16 years of follow-up. Similarly, in Model B, the probability of women having "Very good" or "Excellent" health is 2.5% higher for women relative to men after 16 years from baseline. Interestingly, the same coefficient is an order of magnitude smaller in Model C (compared to Model B) and is not statistically significant. This suggests that, at the population level, we cannot conclude that a gender difference exists in the change of SRH over time.

[Figure 1 here]

Going back to the original question of SRH trajectories by education, Figure 1 shows the predicted trajectories for Models A1 and B. Figure 1a suggests that the average non-Hispanic white male with less than high school, with mean initial SRH and mean rate of change, is expected to report an SRH score of 2.99 ("Good") at age 55 and a score of 2.46 (between "Good" and "Fair") 16 years later. A similar respondent with some

college education or higher, is expected to report a score of 3.98 at baseline and 3.30 at the end of the follow-up period. Consistently in between these two trajectories is the high school graduate. It is clear that while the three individuals' health status is stratified by education at baseline, the difference between individuals (assuming categories SRH are evenly spread apart) remains about the same throughout the study period.

Figure 1b tells a similar story by showing the predicted probability of having "Very good" or "Excellent" health for comparable individuals. At age 55, there is a .86 probability for a highly educated non-Hispanic white male to report one of those higher SRH categories, compared to only .19 for a similar person with less than high school education. Sixteen years later these probabilities decline to .39 and .05, respectively.

Due to the nonlinearity of Model B, it is difficult to detect visually whether the slopes of SRH significantly and substantively differ by education. This is one apparent shortcoming relative to the linear model. On the other hand, results from Model B are more readily interpretable relative to Model A, in which predicted SRH scores do not necessarily match distinct categories (i.e., what does a predicted score of 2.46 actually mean?).

[Figure 2 here]

Figure 2 illustrates the fundamental difference between conditional (person-specific) and marginal (population-averaged) models. Figure 2a shows the predicted probabilities of the different SRH category at each wave for Model B. These probabilities are conditional on fixed *and* random effects. Thus, Figure 2a can be interpreted as the "prognosis" of a non-Hispanic, white male, aged 55 at baseline and who graduated from high school, whose initial SRH and rate of change are average or typical among people

like him. Such a person is most likely (89%) to report "Good" or "Very good" health at age 55, with a low chance of being in "Excellent" health (7%) and virtually no chance of reporting "Poor" health (.27%).

This is in sharp contrast to the picture painted in Figure 2b. Model C predicts that, for the HRS cohort at large, about 4 percent of non-Hispanic white, male high school graduates report "Poor" health at age 55; Twenty percent are predicted to report "Excellent" health. Both of these probabilities are significantly higher than the individual prediction derived from Model B, for a person of "typical" baseline health status and "typical" rate of change in health.

Taken together, these results suggest that the choice between conditional and marginal models should largely be driven by the researcher's substantive interests. The choice between linear and nonlinear models, on the other hand, is less clear. While both models resulted in similar (though not identical) substantive conclusions with the Health and Retirement Study sample, it is difficult to say which is preferable. Unfortunately, there is no simple test statistic or procedure to choose between models based on different probability distributions. Instead, simulation can be used to explore which model is superior with respect to a limited set of criteria (e.g., statistical power).

Model Performance in Simulation

Model comparison is a critical aspect of statistical modeling in sociological research (Weakliem 2004). Selection between competing models is often based on goodness-of-fit statistics such as R^2; for complex nested models, under maximum likelihood estimation, relative goodness-of-fit can be compared using likelihood ratio tests. With non-nested

complex models, selection can prove even more difficult as model fit statistics do not naturally arise to allow formal tests. Nevertheless, several information criteria have been developed for such cases. These include AIC and BIC, differing in their penalty for additional model parameters (Singer and Willet 2003:120), with the former oriented toward predictive accuracy and the latter toward explanatory goodness-of-fit (Shmueli 2010). Unfortunately, all of those methods share a common dependency on the model likelihood function.

By contrast, the models reviewed in this paper are based on different probability distributions (e.g., multinomial, normal); in other cases, such as the marginal cumulative-logit model estimated using GEE, the likelihood is not evaluated at all. Conventional approaches to model selection abandon us and alternative benchmarks are needed.

In what follows, I rely on randomly generated data with known population parameters to compare statistical power in detecting fixed-effects across Models A-C[7]. With respect to regression coefficients, power can be defined as the probability of observing a statistically significant coefficient estimate when the true coefficient differs from zero (given some α level). With complex models, power calculations can easily become intractable and simulation is preferred over analytic approaches. Furthermore, simulation provides an ideal case where data are completely balanced and there are no missing values. In the context of simulation, power can be obtained by drawing multiple random samples from a specified population and repeatedly estimating the various models. Power is then estimated as the proportion of samples in which a particular test was significant.

Since SRH is assumed to represent a latent variable, a continuous outcome was generated for each subject *i* at each time *t* from a normal distribution. Random effects (intercept and slope) were drawn for each subject *i* from a bivariate-normal distribution with a modest negative correlation. Two uncorrelated predictor variables, one dichotomous and one continuous, were incorporated at level-2 (with no time-varying covariates at level-1). The true population model (see Table 4 for a list of population parameters) in composite form is:

$$Y_{it} = (\beta_{00} + \beta_{01}X_{1i} + \beta_{02}X_{2i} + (\beta_{10} + \beta_{11}X_{1i} + \beta_{12}X_{2i})T_{it}) + (u_{0i} + u_{1i}T_{it} + \varepsilon_{it}) \qquad (13)$$

The outcome variable was then categorized using four levels and four models are estimated: Models A-C, as specified above, and Model D, in which the outcome is further dichotomized. Model D follows the same specification as does Model C in Equation 6, but estimated for a dichotomous outcome using the GEE method.

[Table 4 here]

Since longitudinal studies are characterized by repeated measures for each person, the question of sample size becomes more complicated than in cross-sectional studies. One has to consider the number of subjects *and* the number of observations per subject. For this reason, I constructed power surfaces (rather than power curves) in Figures 3-5, with multiple combinations of the two factors. The number of individuals (100, 200, and 300) is plotted on the horizontal axis; the number of observations per subject varies from three to five and is plotted on the vertical axis. Since all models were reasonably powerful with respect to fixed effects on the random-intercept (i.e., β_{01} and β_{02}) they are not shown here. Instead, Figures 3-5 show power plots for fixed effects on the random-slope (i.e., β_{10}, β_{11} and β_{12})[8].

[Figure 3-5 here]

Figure 3 suggests that the models are generally comparable in performance with respect to β_{10}, with some advantage to Model A (linear MLM). Recall that β_{10} corresponds to the average effect of time on the outcome variable (person-specific effect in Models A & B, and population averaged effect in Models C & D) when other level-2 covariates are set to zero. Furthermore, power increases both as a function of the number of individuals and the number of observations per individual (this is reflected in darker shades directed toward the top-right corner of each square plot).

Figures 4 & 5, describing results for β_{11} and β_{12} (the fixed-effects on the rate of change over time), reveal some marked differences between the models. Model D, the marginal logit with dichotomous outcome, has practically no power at all, regardless of sample size. The ordinal Models B & C perform better than the linear MLM (Model A). Overall, Model B performs better than any of the other models and is preferable with respect to statistical power. Note that in Figures 4 & 5 power tends to increase as the number of subjects increases, but not as much as a function of the number observations per subject (especially in Model A). This result may not be surprising considering that β_{11} and β_{12} relate to level-2 covariates that remain fixed within-person across repeated measurements.

Discussion

Longitudinal data are increasingly available and elaborate statistical methods are needed to fully utilize them (for example, ten waves of panel data are currently available for the original Health and Retirement Study cohort). Regretfully, researchers often opt for

cross-sectional analyses or rely on difference scores between two waves while discarding enormous amounts of additional data. Multilevel modeling offers a powerful framework for handling the full range of data across multiple waves, and is particularly useful for describing individual trajectories over time. However, linear multilevel models may not be ideal for all outcomes. Studies of self-rated health commonly neglect the measure's unique characteristics and instead rely on assigning numerical scores to ordered categories. Instead, generalizations of the multilevel model to discrete outcomes constitute a viable alternative.

Since sociologists and other social scientists are generally concerned with explanation over prediction or data reduction (Shmueli 2010), I chose to emphasize substantive interpretation and statistical power when comparing models for longitudinal data. Competing longitudinal models for SRH were juxtaposed along two dimensions: linear versus nonlinear and conditional versus marginal. Results from simulation suggest that conditional ordered-logit models are generally more powerful than linear (normal) multilevel models in detecting fixed effects on growth parameters (in particular, the random-slope). However, while the conditional ordered-logit model outperforms all other models, it is also the most computationally cumbersome.

By contrast, marginal ordered-logit models require little computation time but still perform reasonably well with respect to statistical power. However, predicted means derived from marginal models refer to population-averages. With nonlinear models, marginal and conditional means can show marked differences. In this regard, marginal models are unfit for inference on individual (person-specific) trajectories in much the same way that results from repeated cross-sections are. Since every marginal model

stems, at least implicitly, from a particular conditional model (Lee and Nelder 2004), marginal means and trajectories can generally be inferred from conditional models (but not the opposite). In addition, estimation of marginal models using quasi-likelihood methods is less robust to missing data and assumes that data are missing completely at random (MCAR).

Conditional models are better fit for describing and making inferences about individual trajectories over time. However, interpretation should be conducted with care: predicted probabilities or means, such as SRH scores, are conditional on person-specific effects (i.e., random intercept and slope). When the random components are set to zero, the interpretation of trajectories generally applies to a "typical" individual with mean initial SRH and mean rate of change in SRH (conditional on other covariates).

Choosing between the linear and ordered-logit conditional models is more difficult. Clearly, the linear model is easier to estimate and readily interpretable. However, it can potentially suffer from floor and ceiling effects and has less statistical power compared to its ordinal counterpart. At the same time, I find that in the case of SRH, as illustrated with a large sample from the Health and Retirement Study, it leads to similar substantive conclusions as does the ordered-logit conditional model. Its main shortcoming is that predicted SRH scores have no immediate interpretation, as do predicted probabilities in the ordered-logit model.

As an aside, it is worth making a note on the practice of dichotomizing ordinal variables. Previous studies have suggested that dichotomizing SRH results in minimal reduction in efficiency (e.g., Manor, Matthews, and Power 2000); results in this study

point to the contrary. The simulation results clearly show that collapsing categories, at least with the marginal-logit model, leads to a serious loss of power and is ill-advised.

Based on findings from this study, I propose the following guidelines for modeling self-rated health and similar ordinal outcomes in longitudinal studies (while some may seem trivial they are still worth repeating):

1. Use all of your data. Don't discard valuable information for the sake of simpler methods, although simpler methods may be preferable given the same use of data.
2. Let substantive interpretation lead the choice between marginal (population-averaged) and conditional (person-specific) models.
3. When estimating marginal models with missing data use maximum likelihood estimation. Alternatively, estimate a corresponding conditional model using maximum likelihood and derive from it the marginal means.
4. When comparing models with different probability distributions (e.g., normal, binomial/multinomial), translate effects to predicted means and probabilities rather than rely on statistical significance alone.
5. If power is an issue, use a conditional ordered-logit model over a linear multilevel model. Don't dichotomize your outcome.
6. To save computation time, start with a linear multilevel model for exploratory to perfect your model specification. Second, try robust inference to account for non-normality. Finally, try the more complex ordered-logit model.

In conclusion, I would like to echo Alan Agresti (2010:5) in commenting that "strict adherence to operations that utilize only the ordering in ordinal scales limits the scope of useful methodology." That is, despite the superiority of some methods over others in

ideal scenarios, real data often confront us with complex circumstances – be it missing data, complex survey designs, or model misspecification – and no single model should be trusted blindly.

Notes

[1] Here and throughout the paper OLS is used colloquially to mean simple and multiple regression models with normally distributed errors, and does not refer strictly to the least-squares estimation method.

[2] Although the model can be extended to include additional random effects, equations 1-4 specify the more common random-intercept and random-slope model.

[3] Note that the difference of two SB-scaled chi-square statistics does not itself follow a chi-square distribution, and has to be adjusted in order to properly conduct the chi-square difference test for nested models (see Satorra & Bentler 2001).

[4] Strictly, g is a monotonic differentiable function and the response variable follows a distribution from the natural exponential family (Agresti 2002:116).

[5] The HRS (Health and Retirement Study) is sponsored by the National Institute on Aging (grant number NIA U01AG009740) and is conducted by the University of Michigan.

[6] All models were estimated in Stata: Model A1 using the `xtmixed` command; Model A2 using `xtmixed` with robust standard errors; Model B using `gllamm`, a user-written program (Rabe-Hesketh, Skrondal, and Pickles 2004); and Model C using `ologit` with robust standard errors for clustered observations (using sandwich estimator). Computation time for Model B was exceptionally long, taking nearly 26 hours on a standard desktop PC. By contrast, all other models were estimated in less than a minute.

[7] Simulation was conducted in Stata; Full code (adapted from Feiveson 2009) is available in supplementary material.

[8] At three observations per subject, a few iterations of Model A failed to obtain standard-errors; Model D failed to converge on a few iterations with 100 subjects and five observations per subject. These estimation errors, however, have only minor effects on results shown in Figures 3-5.

References

Agresti, Alan. 2002. *Categorical Data Analysis*. Hoboken, NJ: Wiley.

Agresti, Alan. 2010. *Analysis of Ordinal Categorical Data*. Hoboken, NJ: Wiley.

Ananth, Cande and David G. Kleinbaum. 1997. "Regression Models for Ordinal Responses: A Review of Methods and Applications." *International Journal of Epidemiology* 26(6):1323-33.

Bailis, Daniel S., Alexander Segall, and Judith G. Chipperfield. 2003. "Two Views of Self-rated General Health Status." *Social Science & Medicine* 56:203-17.

Baker, David W., Joseph J. Sudano, Jeffrey M. Albert, Elaine A. Borawski, and Avi Dor. 2001. "Lack of Health Insurance and Decline in Overall Health in Late Middle Age." *The New England Journal of Medicine* 345(15):1106-12.

Benyamini, Yael, Tsachi Ein-Dor, Karni, Ginzburg, and Zahava Solomon. 2009. "Trajectories of Self-Rated Health among Veterans: A Latent Growth Curve Analysis of the Impact of Posttraumatic Symptom." *Psychosomatic Medicine* 71:345-52.

DeSalvo, Karen B., Nicole Bolser, Kristi Reynolds, Jiang He, and Paul Muntner. 2006. "Mortality Prediction with a Single General Self-rated Health Question: A Meta-Analysis." *Journal of General Internal Medicine* 21:267-75.

Duncan, Terry E., Susan C. Duncan, and Lisa A. Strycker. 2006. *An Introduction to Latent Variable Growth Curve Modeling*. Mahwah, NJ: Lawrence Erlbaum Associates.

Eriksson, Ingeborg, Anna-Lena Undèn, and Stig Elofsson. 2001. “Self-rated Health. Comparisons between Three Different Measures. Results from a Population Study.” *International Journal of Epidemiology* 30:326-33.

Feiveson, Alan. 2009. “Calculating Power by Simulation.” College Station, TX: StataCorp LP. Retrieved May 1, 2012 (http://www.stata.com/support/faqs/statistics/power-by-simulation/).

Fullerton, Andrew S. 2009. “A Conceptual Framework for Ordered Logistic Regression Models.” *Sociological Methods & Research* 38(3):306-47.

Ghisletta, Paolo and Ulman Lindenberger. 2004. “Static and Dynamic Longitudinal Structural Analyses of Cognitive Changes in Old Age.” *Gerontology* 50:12–16.

Hartford, Alan and Marie Davidian. 2000. “Consequences of Misspecifying Assumptions in Nonlinear Mixed Effects Models.” *Computational Statistics & Data Analysis* 34:139-64.

Health and Retirement Study. 1992. “Biennial Interview Questionnaire. Section B: Health Status.” Ann Arbor, MI: University of Michigan. Retrieved May 1, 2012 (http://hrsonline.isr.umich.edu/modules/meta/1992/core/qnaire/online/A03_B.pdf)

Hughes, Mary Elizabeth and Linda J. Waite. 2009. “Marital Biography and Health at Mid-Life.” *Journal of Health and Social Behavior* 50(3):245-60.

Hughes, Mary Elizabeth, Linda J. Waite, Tracey A. LaPierre, and Ye Luo. 2007. “All in the Family: The Impact of Caring for Grandchildren on Grandparents’ Health.” *Journal of Gerontology: Social Sciences* 62B(2):S108-19.

Idler, Ellen L. and Yael Benyamini. 1997. “Self-rated Health and Mortality: A Review of Twenty-Seven Community Studies.” *Journal of Health and Social Behavior* 38:21-37.

Jackson, Dennis L. 2010. “Reporting Results of Latent Growth Modeling and Multilevel Modeling Analyses: Some Recommendations for Rehabilitation Psychology.” *Rehabilitation Psychology* 55(3):272-85.

Jylhä, Marja, Jack M. Guralnik, Luigi Ferrucci, Jukka Jokela, and Eino Heikkinen. 1998. “Is Self-rated Health Comparable across Cultures and Genders?” *Journal of Gerontology: Social Sciences* 53B:S144-S152.

Kaplan, George A. and Terry Camacho. 1983. “Perceived Health and Mortality: A Nine-Year Follow-Up of the Human Population Laboratory Cohort.” *American Journal of Epidemiology* 117(3):292-304.

Krause, Neal M. and Gina M. Jay, 1994. “What Do Global Self-rated Health Items Measure?” *Medical Care* 32(9):930-42.

Lee, Youngjo and John A. Nelder. 2004. “Conditional and Marginal Models: Another View.” *Statistical Science* 19(2):219-38.

Litiere, Saskia, Ariel Alonso, and Geert Molenberghs. 2007 “Type I and Type II Error Under Random-Effects Misspecification in Generalized Linear Mixed Models.” *Biometrics* 63:1038-44.

Liu, Hui. 2012. “Marital Dissolution and Self-rated Health: Age Trajectories and Birth Cohort Variations.” *Social Science & Medicine* 74:1107-16.

Luoh, Ming-Ching and A. Regula Herzog. 2002. “Individual Consequences of Volunteer and Paid Work in Old Age: Health and Mortality.” *Journal of Health and Social Behavior* 43(4):490-509.

Lynch, Scott .M. 2003. “Cohort and Life-course Patterns in the Relationship between Education and Health: A Hierarchical Approach.” *Demography* 40(2):309-331.

Maas, Cora J. M. and Joop J. Hox. 2004. “The Influence of Violations of Assumptions on Multilevel Parameter Estimates and Their Standard Errors.” *Computational Statistics & Data Analysis* 46:427–440.

Manor, Orly, Sharon Matthews, and Chris Power. 2000. “Dichotomous or Categorical Response? Analysing Self-rated Health and Lifetime Social Class.” *International Journal of Epidemiology* 29:149-57.

McCullagh, Peter. 1980. “Regression Models for Ordinal Data.” *Journal of the Royal Statistical Society: Series B* 42(2):109-42.

Meadows, Sarah O. 2009. “Family Structure and Fathers’ Well-Being: Trajectories of Mental Health and Self-Rated Health.” *Journal of Health and Social Behavior* 50:115-31.

Meredith, William and John Tisak. 1990. “Latent Curve Analysis.” *Psychometrika* 55(1):107-22.

Mirowsky, John and Catherine E. Ross. 2008. “Education and Self-rated Health: Cumulative Advantage and Its Rising Importance.” *Research on Aging* 30(1):93-122.

Orfila, Francesc, Montserrat Ferrer, Rosa Lamarca, and Jordi Alonso. 2000. “Evolution of Self-rated Health Status in the Elderly: Cross-Sectional vs. Longitudinal Estimates.” *Journal of Clinical Epidemiology* 53:563-70.

Rabe-Hesketh, Sophia and Andres Skrondal. 2008. *Multilevel and Longitudinal Modeling Using Stata*. College Station, TX: Stata Press.

Rabe-Hesketh, Sophia, Andres Skrondal, and Andrew Pickles. 2004. “Generalized Multilevel Structural Equation Modeling.” *Psychometrika* 69(2):167-90.

RAND HRS Data, Version K. 2011. Santa Monica, CA: RAND Center for the Study of Aging.

Ritz, John and Donna Spiegelman. 2004. “Equivalence of Conditional and Marginal Regression Models for Clustered and Longitudinal Data.” *Statistical Methods in Medical Research* 13:309-23.

Ross, Catherine E. and Chia-Ling Wu. 1996. “Education, Age, and the Cumulative Advantage in Health.” *Journal of Health and Social Behavior* 37(1):104-120.

Sacker, Amanda, Diana Worts, and Peggy McDonough. 2011. “Social Influences on Trajectories of Self-rated Health: Evidence from Britain, Germany, Denmark and the USA.” *Journal of Epidemiology and Community Health* 65:130-36.

Satorra, Albert and Peter M. Bentler. 1994. “Corrections to Test Statistics and Standard Errors in Covariance Structure Analysis.” Pp. 399-419 in *Latent Variables Analysis: Applications for Developmental Research*, edited by Alexander von Eye and Clifford C. Clogg. Thousand Oaks, CA: Sage.

Satorra, Albert and Peter M. Bentler. 2001. “A Scaled Difference Chi-Square Test Statistic for Moment Structure Analysis.” *Psychometrika* 66(4):507-14.

Schafer, Joseph L. and John W. Graham. 2002. “Missing Data: Our View of the State of the Art.” *Psychological Methods* 7(2):147-77.

Shmueli, Galit. 2010. “To Explain or to Predict?” *Statistical Science* 25(3):289-310.

Shuey, Kim M. and Andrea E. Willson, 2008. “Cumulative Disadvantage and Black-White Disparities in Life-Course Health Trajectories.” *Research on Aging* 30(2):200-25.

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

Verbeke, Geert and Emmanuel Lesaffre. 1997. “The Effect of Misspecifying the Random-Effects Distribution in Linear Mixed Models for Longitudinal Data.” *Computational Statistics & Data Analysis* 23:541-56.

Weakliem, David L. “Introduction to the Special Issue on Model Selection.” *Sociological Methods & Research* 33(2):167-87.

Willson, Andrea E., Kim M. Shuey, and Glen H. Elder, Jr. 2007. “Cumulative Advantage Processes as Mechanisms of Inequality in Life Course Health.” *American Journal of Sociology* 112(6):1886-1924.

Wilmoth, Janet M., Andrew S. London, and Wendy M. Parker. 2010. “Military Service and Men’s Health Trajectories in Later Life.” *Journal of Gerontology: Social Sciences* 65B(6):744-55.

Zeger, Scott L. and Kung-Yee Liang, 1986. “Longitudinal Data Analysis for Discrete and Continuous Outcomes.” *Biometrics* 42(1):121-30.

Table 2: Sample descriptive statistics, Health and Retirement Study

Variable	N	Mean/Proportion	SD
Age (at baseline)	12,651	55.26	5.67
Female	12,652	0.54	0.50
Hispanic	12,642	0.09	0.29
White	12,652	0.80	0.40
Black	12,652	0.17	0.37
Other race	12,652	0.04	0.19
Education			
Less than high school	12,652	0.27	0.44
High school / GED	12,652	0.53	0.50
Some college or higher	12,652	0.20	0.40
Self-rated health			
1992	12,652	3.42	1.21
1994	11,419	3.35	1.17
1996	10,770	3.37	1.15
1998	10,238	3.16	1.15
2000	9,626	3.24	1.14
2002	9,202	3.20	1.11
2004	8,768	3.12	1.12
2006	8,249	3.11	1.11
2008	7,837	3.05	1.09

Table 3: Self-rated health model comparison, Health and Retirement Study 1992-2008

Variable	**Model A1**	**Model A2**	**Model B**	**Model C**
Effect on intercept				
Intercept1	**2.990** (<.001)	**2.990** (<.001)	**-4.075** (<.001)	**-2.189** (<.001)
Intercept2	-	-	**-1.391** (<.001)	**-0.718** (<.001)
Intercept3	-	-	**1.418** (<.001)	**0.740** (<.001)
Intercept4	-	-	**4.419** (<.001)	**2.378** (<.001)
Age	**-0.017** (<.001)	**-0.017** (<.001)	**-0.056** (<.001)	**-0.031** (<.001)
Female	0.009 (.648)	0.009 (.648)	0.045 (.458)	0.035 (.291)
Hispanic	**-0.333** (<.001)	**-0.333** (<.001)	**-1.024** (<.001)	**-0.583** (<.001)
Race				
White	ref	ref	ref	ref
Black	**-0.403** (<.001)	**-0.403** (<.001)	**-1.304** (<.001)	**-0.722** (<.001)
Other race	**-0.166** (.001)	**-0.166** (.001)	**-0.536** (.001)	**-0.284** (.002)
Education				
<HS	ref	ref	ref	ref
HS/GED	**0.590** (<.001)	**0.590** (<.001)	**1.840** (<.001)	**0.997** (<.001)
>HS	**0.991** (<.001)	**0.991** (<.001)	**3.195** (<.001)	**1.715** (<.001)
Effect on slope				
Intercept	**-0.033** (<.001)	**-0.033** (<.001)	**-0.098** (<.001)	**-0.033** (<.001)
Age	**0.000** (.031)	**0.000** (.035)	-0.001 (.188)	**0.001** (.002)
Female	**0.004** (.005)	**0.004** (.005)	**0.010** (.022)	-0.001 (.802)
Hispanic	**0.005** (.040)	0.005 (.058)	**0.019** (.012)	-0.001 (.900)
Race				
White	ref	ref	ref	ref
Black	**0.009** (<.001)	**0.009** (<.001)	**0.031** (<.001)	**0.016** (<.001)
Other race	0.001 (.809)	0.001 (.819)	0.002 (.879)	0.007 (.311)
Education				
<HS	ref	ref	ref	ref
HS/GED	**-0.005** (.002)	**-0.005** (.003)	**-0.019** (<.001)	**-0.011** (.001)
>HS	**-0.009** (<.001)	**-0.009** (<.001)	**-0.040** (<.001)	**-0.026** (<.001)
Variance components				
Var(intercept)	0.827	0.827	8.621	-
Var(slope)	0.002	0.002	0.021	-
Cov(intercept,slope)	-0.017	-0.017	-0.196	-
Residual variance	0.431	0.431	-	-

Models: A1 = Linear MLM; A2 = Robust MLM; B = Conditional Ordered-Logit; C = Marginal Ordered-Logit.

a Coefficient estimates significant at (two-tailed) α=.05 are in bold.

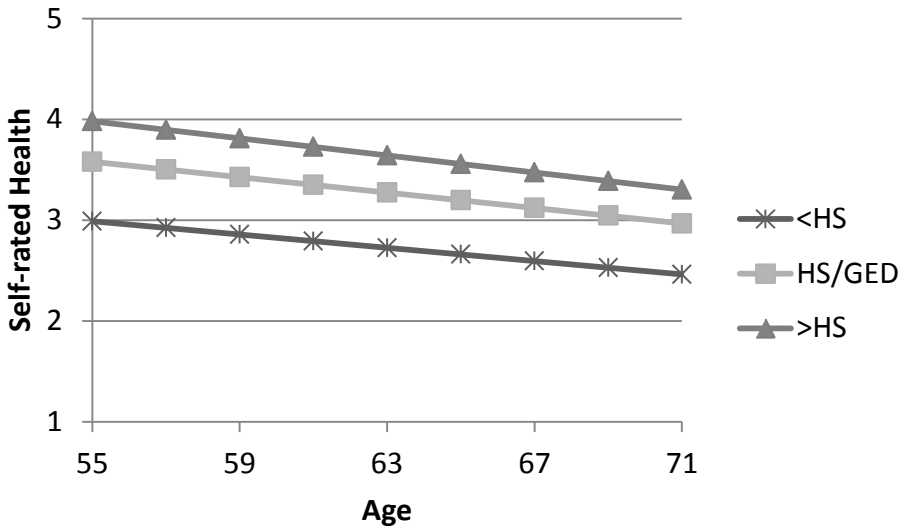
b p-values in parentheses.

Table 4: List of population parameters used in simulation

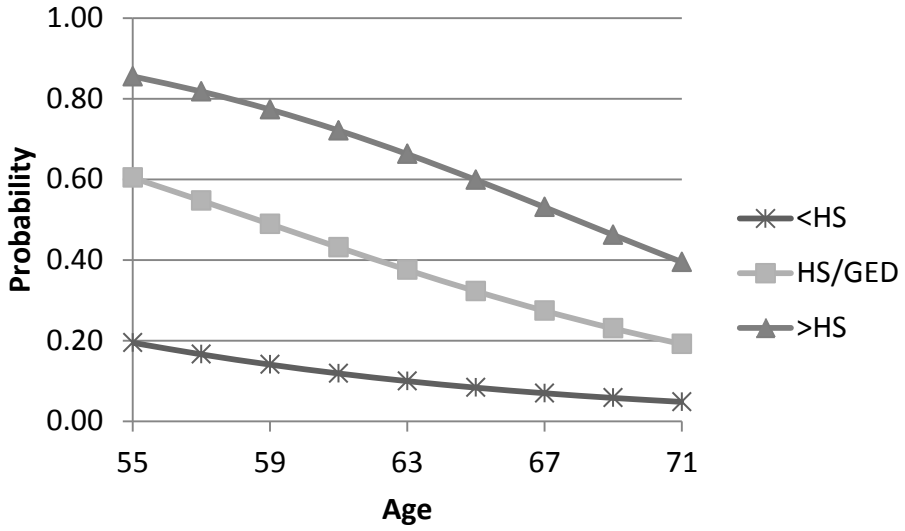
Regression Coefficients		Variance Components	
Parameter	Value	Parameter	Value (SD)
β_{00}	10	σ_{ε}	15
β_{01}	-16	σ_{00}	20
β_{02}	0.6	σ_{11}	7
β_{10}	4	ρ_{01}	-0.15[a]
β_{11}	-4		
β_{12}	0.06		

[a] Correlation coefficient

Figure 1: Predicted trajectories of self-rated health, Health and Retirement Study 1992-2008*



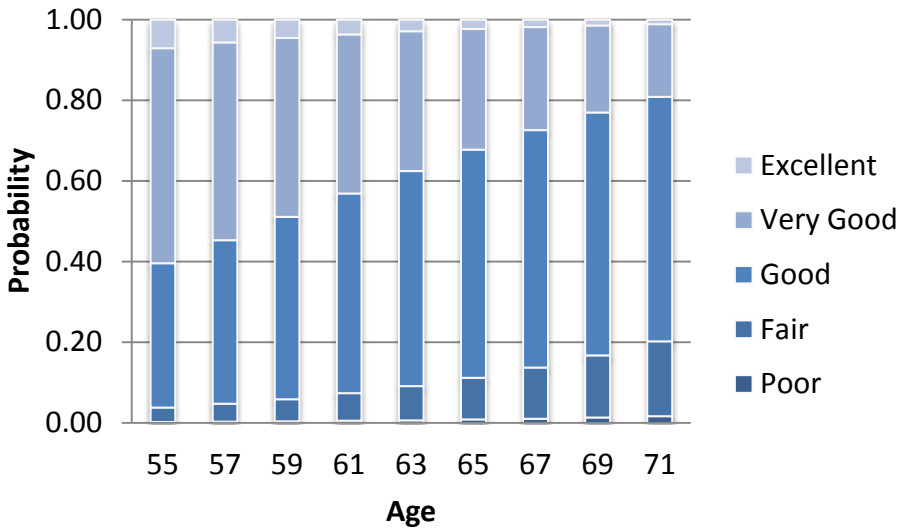
a) Predicted self-rated health, Model A: Linear multilevel model



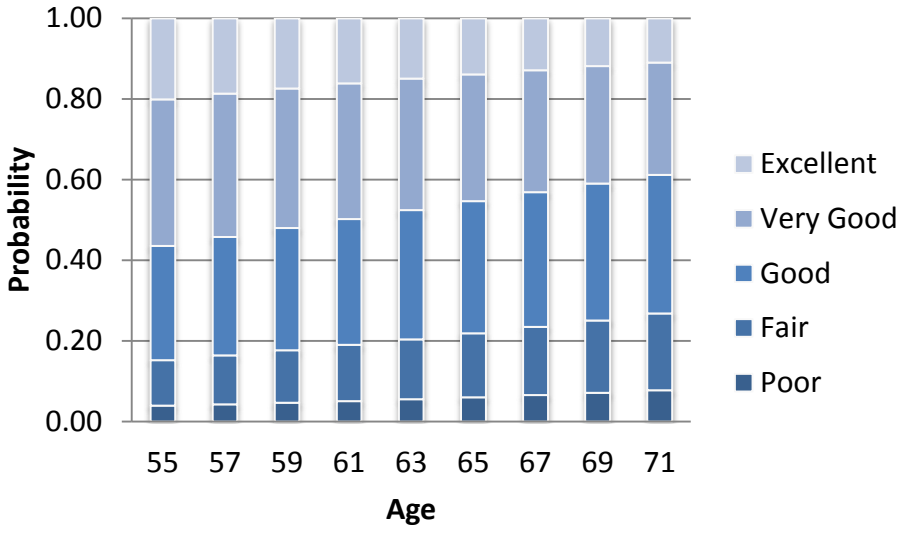
b) Predicted probability of "Very good" or "Excellent" health, Model B: Conditional ordered-logit

* Predicted trajectories for non-Hispanic white males, aged 55 at baseline, with random effects set at their zero means.

Figure 2: Predicted conditional and marginal probabilities of self-rated health categories over time; Results from ordered-logit regression*



a) Model B: Conditional ordered-logit



b) Model C: Marginal ordered-logit

* Predicted probabilities for non-Hispanic white males, aged 55 at baseline; in Model B, predicted probabilities are conditional on random effects set to their zero means.

Figure 3: Model comparison. Power surfaces for β_{10} at α=0.05.

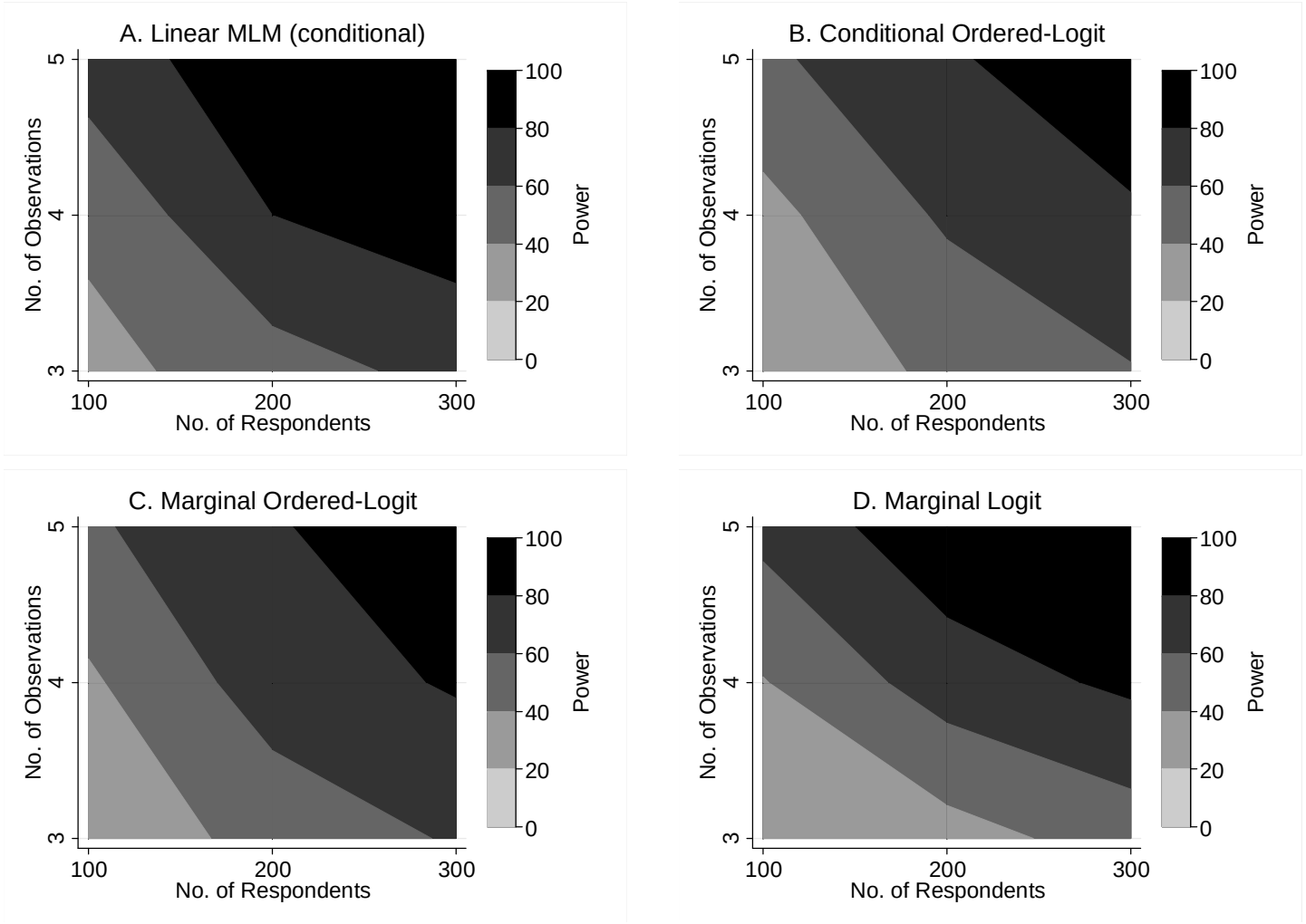


Figure 4: Model comparison. Power surfaces for β_{11} at α=0.05.

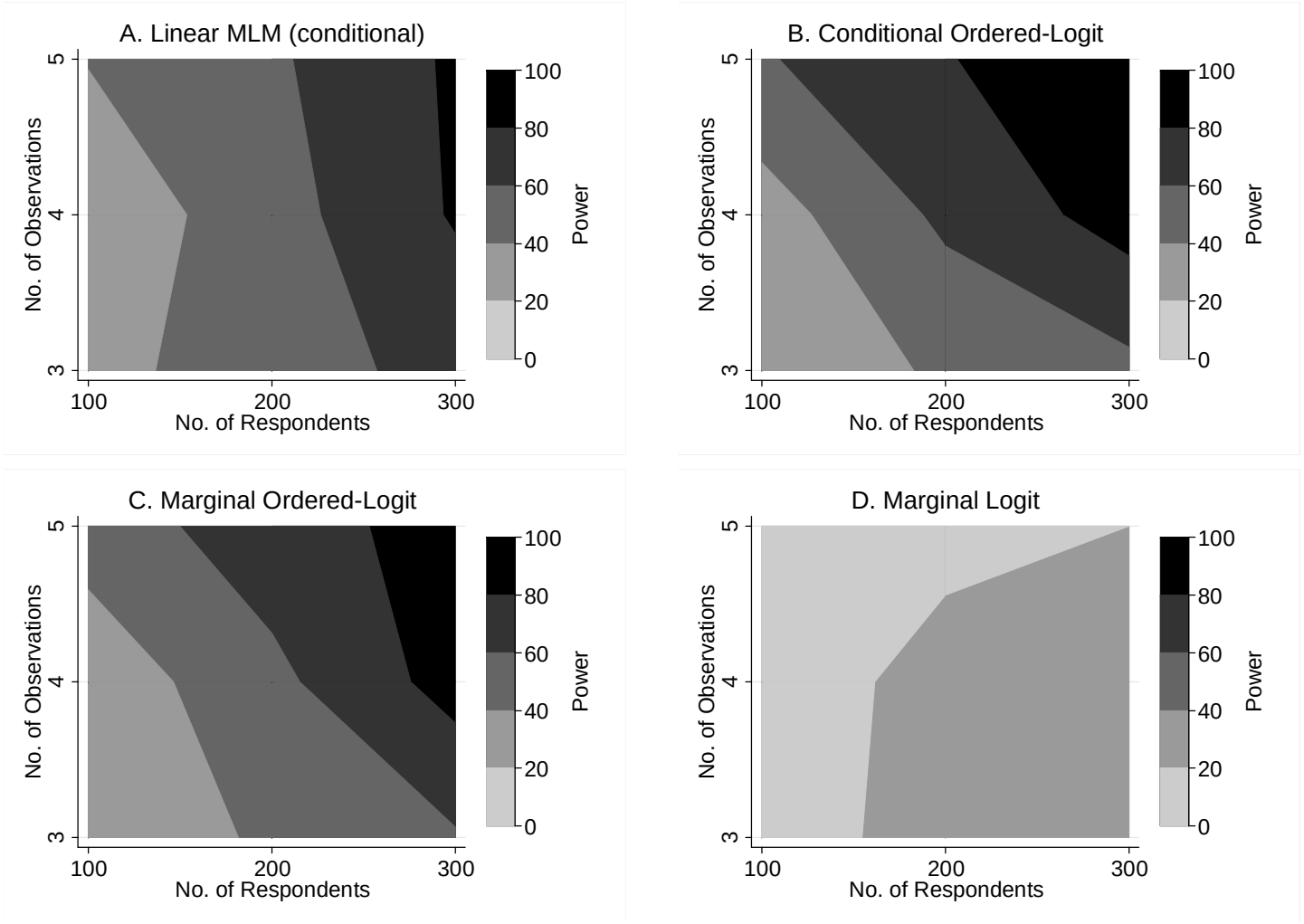


Figure 5: Model comparison. Power surfaces for β_{12} at α=0.05.