The Report Committee for Isaac Sasson
Certifies that this is the approved version of the following report:

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

APPROVED BY
SUPERVISING COMMITTEE:

Supervisor:

Daniel A. Powers

Debra J. Umberson

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

by

Isaac Sasson, B.S.; M.A.

Report

Presented to the Faculty of the Graduate School of

The University of Texas at Austin

in Partial Fulfillment

of the Requirements

for the Degree of

Master of Science in Statistics

The University of Texas at Austin

May 2012

Acknowledgements

I would like to thank Dr. Daniel Powers for his advice and dedicated supervision throughout the writing of this report. My deepest gratitude goes to Dr. Debra Umberson for her ongoing feedback, extended support during my graduate training, and openness to my research interests.

Additional thanks go to Dr. Mark Hayward, Dr. John Mirowsky, and several graduate students and post-doctoral fellows at the University of Texas at Austin Population Research Center, for their insightful comments on earlier versions of this manuscript.

May 2012

Abstract

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

Isaac Sasson, M.S.Stat.

The University of Texas at Austin, 2012

Supervisor: Daniel A. Powers

Self-rated health (SRH) is an outcome commonly studied by demographers, epidemiologists, and sociologists of health, typically measured using an ordinal scale. SRH is analyzed in cross-sectional and longitudinal studies for both descriptive and inferential purposes, and has been shown to have significant validity with regard to predicting mortality. Despite the wide spread use of this measure, only limited attention is explicitly given to its unique attributes in the case of longitudinal studies. While self-rated health is assumed to represent a latent continuous and dynamic process, SRH is actually measured discretely and asymmetrically. Thus, the validity of methods ignoring the scale of measurement remains questionable. We compare three approaches to modeling SRH with repeated measures over time: linear multilevel models (MLM or LGM), including corrections for non-normality; and marginal and conditional ordered-

logit models for longitudinal data. The models are compared using simulated data and illustrated with results from the Health and Retirement Study. We find that marginal and conditional models result in very different interpretations, but that conditional linear and non-linear models result in similar substantive conclusions, albeit with some loss of power in the linear case. In conclusion, we suggest guidelines for modeling self-rated health and similar ordinal outcomes in longitudinal studies.

Table of Contents

List of Tables

List of Figures

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 hundreds of cross-sectional and longitudinal studies for both descriptive and inferential purposes and has significant validity with regard to predicting mortality (Kaplan & Camacho 1983; Idler & Benyamini 1997; DeSalvo et al. 2006). Despite the ubiquity of SRH as an outcome variable, only limited attention is explicitly given to its typical five-point ordinal scale – especially in the case of longitudinal studies. Longitudinal data are 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). However, the lack of consensus in longitudinal methodology with regard to ordinal outcomes not only threatens the validity of results, but also adds considerable difficulty to the comparison of results across studies. Our aim in this paper is threefold: 1) to promote the use of repeated SRH measurements over time whenever such data are available, as it is key to understanding health trajectories over the life-course; 2) address conceptual and methodological considerations in modeling SRH in longitudinal studies; 3) using both simulated and real data, compare the performance of several existing longitudinal modeling approaches for ordinal outcomes and SRH in particular.

While we assume some familiarity with many of the statistical methods discussed (e.g., GLM, hierarchical/multilevel modeling), we review them specifically in the context of longitudinal ordinal outcomes. First, we discuss the characteristics of SRH as well as

assumptions and choices to be made when modeling SRH over time. Second, we suggest three competing approaches for the task: (A) the linear (normal) multilevel model for change, including robust inference for non-normality; (B) conditional ordered-logit model; and (C) marginal ordered-logit model allowing for correlated observations. These model families stem from both the hierarchical linear modeling (HLM/MLM) framework and the latent growth curve modeling (LGM) framework, when generalized to discrete outcomes. Third, using both simulated and real data from the Health and Retirement Study we compare performance in key points of interest to applied researchers across the competing models. Finally, we provide guidelines for modeling SRH from the viewpoint of a life-course perspective. While the methods presented in this paper are not strictly novel, a review of the literature suggests that they are seldom used with respect to SRH even in face of rich longitudinal data. Instead, many researchers choose to utilize only a single wave of data in cross-sectional analysis, or opt for using two waves with difference scores. Such practices discard an enormous amount of useful data that is central for describing complex health trajectories over time (for example, for members of the first Health and Retirement Study cohort up to nine measurements have been collected). Despite our focus on SRH, the following approaches and considerations could easily extend to other ordinal outcomes with similar characteristics.

CHARACTERISTICS OF SELF-RATED HEALTH

Measurement of SRH is not entirely consistent across studies. Differences exist in the number of levels measured, in questionnaire wording, and in respondent interpretation. These issues have been thoroughly discussed elsewhere (Krause & Jay 1994; Eriksson et al. 2001; Bailis et al. 2003). However, in terms of statistical modeling these differences are of little importance. For illustrative purposes we adopt the Health and Retirement Study version: "Would you say your health is excellent, very good, good, fair, or poor?" Regardless of these inconsistencies, we identify three important characteristics of SRH that directly pertain to its statistical analysis:

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 jump from "Very Good" health 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.

To distinguish from the apparent asymmetry of the scale, we also note that SRH is often *empirically* asymmetric. Since the majority of people are relatively healthy throughout most of their lives, the distribution of SRH in the general population is expected to be skewed.

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. In the following sections we discuss competing approaches to modeling self-rated health in longitudinal studies, compare their performance, and provide guidelines for future studies.

APPROACHES TO MODELING ORDINAL OUTCOMES

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. Multinomial models in particular, when applied to ordinal variables, tend to have less power than ordinal methods in detecting associations as they involve additional parameters (and consequently fewer degrees of freedom in statistical inference) (Agresti 2010). More often, however, researchers choose to collapse categories to dichotomize the outcome. When categories represent arbitrary cut-points of an unobserved continuous scale – as in SRH – then reducing the number of categories can result in bias and loss of power (Ananth & Kleinbaum 1997; Agresti 2010).

Methods designed for continuous data (e.g., 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 assume that moving from "Excellent" to "Very Good" health has an equivalent meaning as moving from "Fair" to "Poor" health, as the intervals between these 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.

[1] OLS in this paper is used as shorthand for simple and multiple regression models with normally distributed errors, and does not refer strictly to the least-squares estimation method.

OLS models for 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). 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 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). With respect to the link function, the **cumulative** approach assumes an underlying continuous variable that is measured discretely using several arbitrary categories. It models the probability of being at *or below* each category compared to all categories above it. Alternatively, the **stage** 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
$Log\left(\frac{\Pr(Y \leq j)}{\Pr(Y > j)}\right)$	$Log\left(\frac{\Pr(Y = j)}{\Pr(Y > j)}\right)$	$Log\left(\frac{\Pr(Y = j)}{\Pr(Y = j')}\right)$

* 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, with this model specification regression coefficient estimates are invariant to the choice of response categories (McCullagh 1980; Agresti 2010). 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 (asymmetric, arbitrary categories with an underlying construct) the cumulative-logit approach appears to fit best among the aforementioned GLMs; this is the approach we adopt here and extend to the longitudinal context.

GENERAL FRAMEWORK FOR MODELING LONGITUDINAL DATA

The aforementioned modeling approaches for ordinal data can be extended to the case of longitudinal designs. 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 & Tisak 1990). While we emphasize the MLM specification in subsequent models, equivalent models can generally be obtained using LGM. As noted by others (see Jackson 2010), the distinction between the two approaches is becoming increasingly vague and software-dependent. Whichever framework is used for longitudinal modeling, similar conceptual and methodological concerns apply. SRH measurements are considered nested observations within individuals (level-1) while the individuals themselves are considered independent, randomly-chosen blocks (level-2).

Conceptually we can divide the multilevel model into two steps corresponding to the two model levels: a) estimating growth parameters (intercept, slope, etc.) for each subject to summarize individual SRH trajectories; b) making inference about variation in growth parameters between individuals. We emphasize this distinction as the outcome in each model level is fundamentally different: SRH in level-1 and growth parameters in level-2. When generalizing multilevel models to include ordinal outcomes, decisions are to be made *separately* at each model level (e.g., choosing a random distribution and link function for each of the outcomes).

A two-level model specification with random-intercept and random-slope is as follows[2]:

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

$$\text{Level-2:} \qquad \gamma_{0i} = \beta_{00} + \sum_{k} \beta_{0k} X_{ki} + u_{0i} \tag{2}$$

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

$$\gamma_{pi} = \beta_{p0} \tag{4}$$

where, in equation (1), $g(\cdot)_{ij}$ 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_{ij} 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) simply 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.

[2] Although the model can be extended to include additional random effects, we focus on the more common random-intercept and random-slope model specified here.

Instead, choosing the cumulative-logit function and the multinomial distribution for SRH leads us to a generalized linear mixed model (GLMM). Note that regardless of our choice of probability distribution and link function in level-1, it is common to operate under the tractable assumption that the 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 & Lesaffre 1997). However, with GLMMs misspecification of random effects may have more severe consequences (Hartford & Davidian 2000; Litiere et al. 2007).

While not unique to a particular scale (continuous vs. ordinal) or framework (MLM vs. LGM), the following considerations are pertinent to longitudinal modeling:

1) *Probability distributions and link functions*. As previously noted, in the multilevel context distributions and link functions are chosen separately at each level. Since growth parameters in level-2 are generally considered latent factors, the optimal choice is often a multivariate normal distribution with the canonical identity link (albeit other possibilities exist). Choices regarding the outcome in level-1 (SRH) are more important. As in the GLM case 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 using the multinomial distribution with the cumulative-logit link.
2) *Trajectory functional form*. Rather than imposing an arbitrary functional form (e.g., linear, quadratic, cubic, piecewise, etc.) on individuals' health trajectories

over time, this choice should stem from both theory and empirical results. In the MLM framework random effects can be tested sequentially (with linear, quadratic or even cubic terms for time) and omitted when they are not statistically significant. Similarly, discontinuities can be tested using piecewise models when theory dictates such effects are plausible (for example, an abrupt deterioration in health following a specific event over the life-course). Similarly, in the LGM framework one can extend beyond linear trajectories by adding factors that correspond to higher polynomials and testing for model fit to select between them. A more parsimonious alternative with LGM is the unspecified two-factor model (Duncan et al. 2006, 31), including a factor for the intercept, specified as usual, and a second factor with only the first and second loadings fixed and any additional loadings 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 any departure from independence. Specifically, with repeated measurements a person's SRH level at one point is likely to affect sequential SRH measurements. It is also likely that the variance of SRH should increase over time as individuals in the sample age and health disparities accumulate. However, the inclusion of person-specific random effects explicitly allows for within person heteroskedasticity and autocorrelation (Singer & Willet 2003). 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 of random effects are of interest as they indicate the amount of residual variance, conditional on other covariates, between individuals' smoothed growth trajectories (Singer & Willet 2003). The covariance between growth parameters (e.g., intercept and slope) can also be estimated and often has a meaningful interpretation. For example, a significant negative correlation between SRH intercept and 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, we may choose 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 computed. 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 become quite difficult with longitudinal data, especially with non-linear models. For conditional models, as specified earlier, Maximum Likelihood methods are generally used and are known to have several desirable properties such as consistency and asymptotic normality (Singer & Willett 2003, 65). In addition, ML estimates are unbiased even in presence of missing values as long as they are missing at random (Schafer & Graham 2002). However, when extended to include non-linear link functions and non-normal distributions, the likelihood function cannot be evaluated directly and numerical methods are often used for approximation. This process can become computationally cumbersome even with moderately sized datasets. 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 & 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 we discuss later. Marginal means can generally be extracted from conditional models, but the reverse in not possible. Ultimately,

conditional models contain more information than 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. Similarly, tests for several parameters or relative model fit (with nested models) can be based on χ^2 statistics (Kline 2005). When using the linear multilevel model under the LGM framework adjustments can be made to accommodate departure from non-normality, as is the case with discrete outcomes. Inference on random effects is more elusive and considered highly sensitive to imbalanced designs (Singer & 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 of each of the aforementioned considerations. 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 allows for more flexible error-structures and the use of more elaborate models where growth factors are both outcomes and predictors. As previously noted, 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 & Lindenberger 2004). It should be noted that these differences are somewhat software dependent and may diminish even further in the future (Jackson 2010).

LONGITUDINAL MODELS FOR ORDINAL OUTCOMES

We present three competing longitudinal models for ordinal outcomes and compare their performance in testing the effects of explanatory variables on baseline and rate of change of SRH: A) linear (normal) multilevel model; B) conditional ordered-logit model; C) marginal ordered-logit model with correlated observations. We begin by specifying each model and its underlying assumptions.

Model A: Linear Multilevel

The linear multilevel model is commonly used in the literature for modeling longitudinal outcomes using either the MLM or LGM approach. As noted earlier, in its basic form the linear MLM/LGM specification assumes the identity link function and a normal distribution for the error term in equation (1), and a bivariate normal distribution for the random effects in equations (2) and (3). With 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 (this is verified in later analyses).

Applying the continuous multilevel model to a non-normal outcome variable may bias our inference on estimated regression coefficients. ML estimates of fixed effects are asymptotically normal even when the error-terms are non-normal (though departure from normality requires larger sample sizes). However, estimates of standard errors may be biased downwards and in turn affect test statistics. Robust inference procedures (Satorra 1992) may be needed to draw correct conclusions with non-normal data: in the case of single parameter tests (e.g., the Wald test) robust standard errors can be obtained; when

assessing overall model fit (or testing several parameters at once) test statistics can be adjusted. Software packages often report the model χ^2 statistic (again, under ML estimation) which can be used to test nested models against each other. The model χ^2 statistic also relies on normality assumptions and can be adjusted downward to account for kurtosis using the Satorra-Bentler (1994) scaled χ^2 statistic[3].

Model B: Conditional Ordered-Logit

A generalized linear mixed model (GLMM) extends standard GLMs to include random effects as well as fixed effects in the linear combination of predictors. As shown in equation (1), a variety of link functions and probability distributions can be used (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 the level-1 equation, substituting it with equation (5):

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

Additionally, 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 for each time point, corresponding to the number of cut-points between SRH categories. Note that this specification takes the proportional odds form (McCullagh

[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).

1980), 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.

We refer to model B as 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 non-linear, such as the logit, and when considerable variation exists between individuals' health trajectories, conditional effects differ in magnitude from marginal (i.e., population-averaged) effects (Agresti 2002, 501). Since the attenuation of marginal effects is accompanied by attenuation of standard-errors, inferential statistics in both models is generally similar. With respect to prediction, the distinction between conditional and marginal means is crucial for researchers interested in a life-course interpretation of SRH, as the former produces an individual trajectory interpretation while the latter yields a population-level interpretation (see Appendix A). We illustrate this point with real data in the following sections.

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 GLMs 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_t – averaged over all individuals. A general specification of a marginal linear model using the cumulative logit link is as follows:

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

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 estimates of the regression coefficients, but will likely produce underestimated standard errors which in turn affect substantive 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 the appropriate covariance structure (Rabe-Hesketh & Skrondal 2008). Alternatively, when feasible, *quasi-likelihood* estimation (such as the GEE method) can be used by specifying a working correlation matrix that captures the within-cluster dependence (Agresti 2010, 268). While a discussion of these methods is beyond the scope of this paper, one important caveat regarding missing data is in order: while ML estimation makes the assumption that observations are missing at random (MAR), quasi-likelihood methods make the stricter assumption that data are missing *completely* at random (MCAR).

MODEL PERFORMANCE IN SIMULATION

Model comparison is a critical aspect of statistical modeling. Selection between competing models is often based on goodness-of-fit statistics such as R^2, measuring the proportion of explained variance in simple models; for complex nested models, under maximum likelihood estimation, relative goodness-of-fit can be compared using the chi-square difference statistic (Kline 2005, 146). 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, most notably AIC and BIC which differ in their penalty for additional parameters (Singer & Willet 2003, 120). Unfortunately, these methods share a common dependency on the model likelihood function.

The models we review here are based on different probability distributions, namely the binomial, multinomial, and normal; in other cases, such as the marginal cumulative-logit model under quasi-likelihood estimation, the likelihood is not evaluated at all. Conventional approaches to model selection abandon us and alternative benchmarks are needed. In this section we use randomly generated data with known population parameters, to compare the competing models with respect to power in detecting fixed effects. This is an ideal case where data are completely balanced and no data are missing. In the next section we compare model performance against real data along differences in substantive interpretation.

In terms of regression fixed effects, power is defined as the probability of observing a statistically significant coefficient estimate when the true coefficient differs

from zero. With complex models power calculations can easily become intractable, and simulation is preferred over analytic approaches. 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.

We follow this approach with all models specified above. Since SRH is assumed to represent a latent construct, we first generate a continuous outcome from a multilevel normal distribution, where repeated measurements are nested within individuals. Random effects (intercept and slope) are drawn from a bivariate-normal distribution with a slight negative correlation. Two uncorrelated predictor variables – one binary and one continuous – appear in level-2; time-varying covariates are not included. The true population model (see Appendix B for details), in composite form, is as follows:

$$Y_{it} = \left(\beta_{00} + \beta_{01} X_{1i} + \beta_{02} X_{2i} + \left(\beta_{10} + \beta_{11} X_{1i} + \beta_{12} X_{2i}\right) T_{it}\right) + \left(u_{0i} + u_{1i} T_{it} + \varepsilon_{it}\right) \qquad (7)$$

The outcome variable is then categorized using four levels and four models are estimated: the linear multilevel model (A), and conditional ordered-logit model (B), and the marginal ordered-logit model (C). In addition, the outcome is further dichotomized as a binary variable and a fourth model is estimated – a marginal logit model (D) – that is similar to model C with fewer J categories.

Since longitudinal studies are characterized by repeated measures for each person, the question of sample size becomes more complicated than in simple models. One has to consider the number of individuals *and* the number of observations per individual separately. For this reason, we construct power surfaces (rather than power curves) in

Figures 1-3, with multiple combinations of the two factors. The number of individuals, plotted on the horizontal axis, ranges from 100 to 300; on the vertical axis we compare 3-5 repeated measurements per subject. Since all models were reasonably powerful with respect to the random intercept fixed effects (i.e., β_{01} and β_{02}), they are not shown here. Instead, we focus on the random slope fixed effects (β_{10}, β_{11} and β_{12}). One important caveat is that results for the linear multilevel model were less stable at lower numbers of measurements per subject (in particular, Obs=3), as estimation procedures failed to converge more often.

The models are generally comparable in performance with respect to β_{10}, (Fig. 1), with a slight advantage for Models A and D. When other "level-2" covariates are set to zero, β_{10} indicates the average effect of time on SRH. Moreover, power increases both as a function of the number of individuals and the number of observations per individual (this is reflected in lighter shades directed toward the top-right corner). Results with respect to β_{11} and β_{12} (Fig. 2-3) reveal some marked differences between the models. Model D, the marginal logit (using a binary outcome), has practically no power at all – regardless of sample size. The ordinal models B and C perform better than the linear multilevel (A). Overall, model B has more power than any of the other models and is preferable. Note that in this case power tends to increase as the number of individuals increases, but no as much with the number of repeated observations per person. This result is not surprising considering that β_{11} and β_{12} relate to level-2 covariates which remain fixed across repeated measurements.

Figure 1: Model comparison – power surfaces for β_{10} at α=0.05; Power indicated by color.

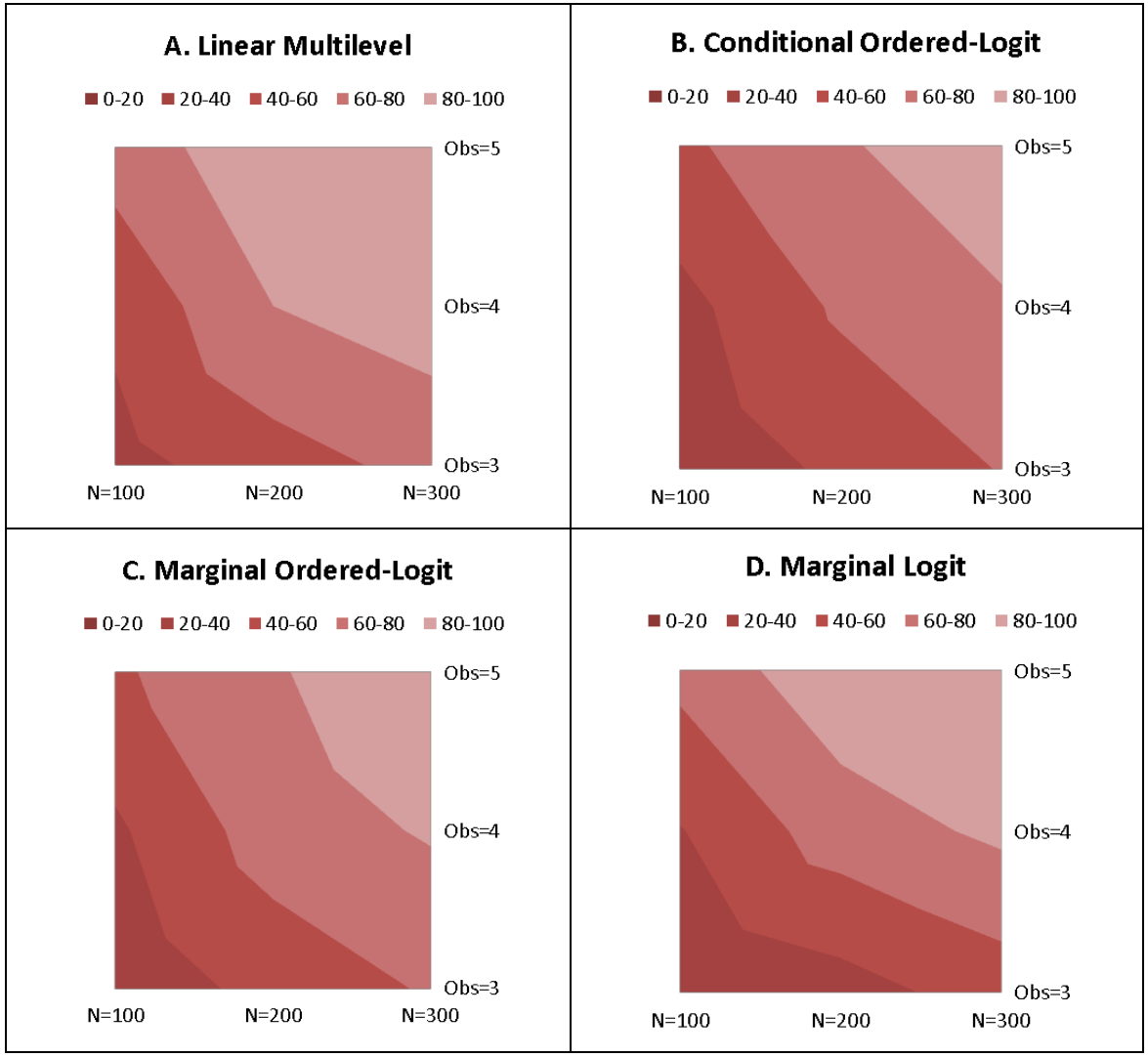


Figure 2 Model comparison – power surfaces for β_{11} at α=0.05; Power indicated by color.

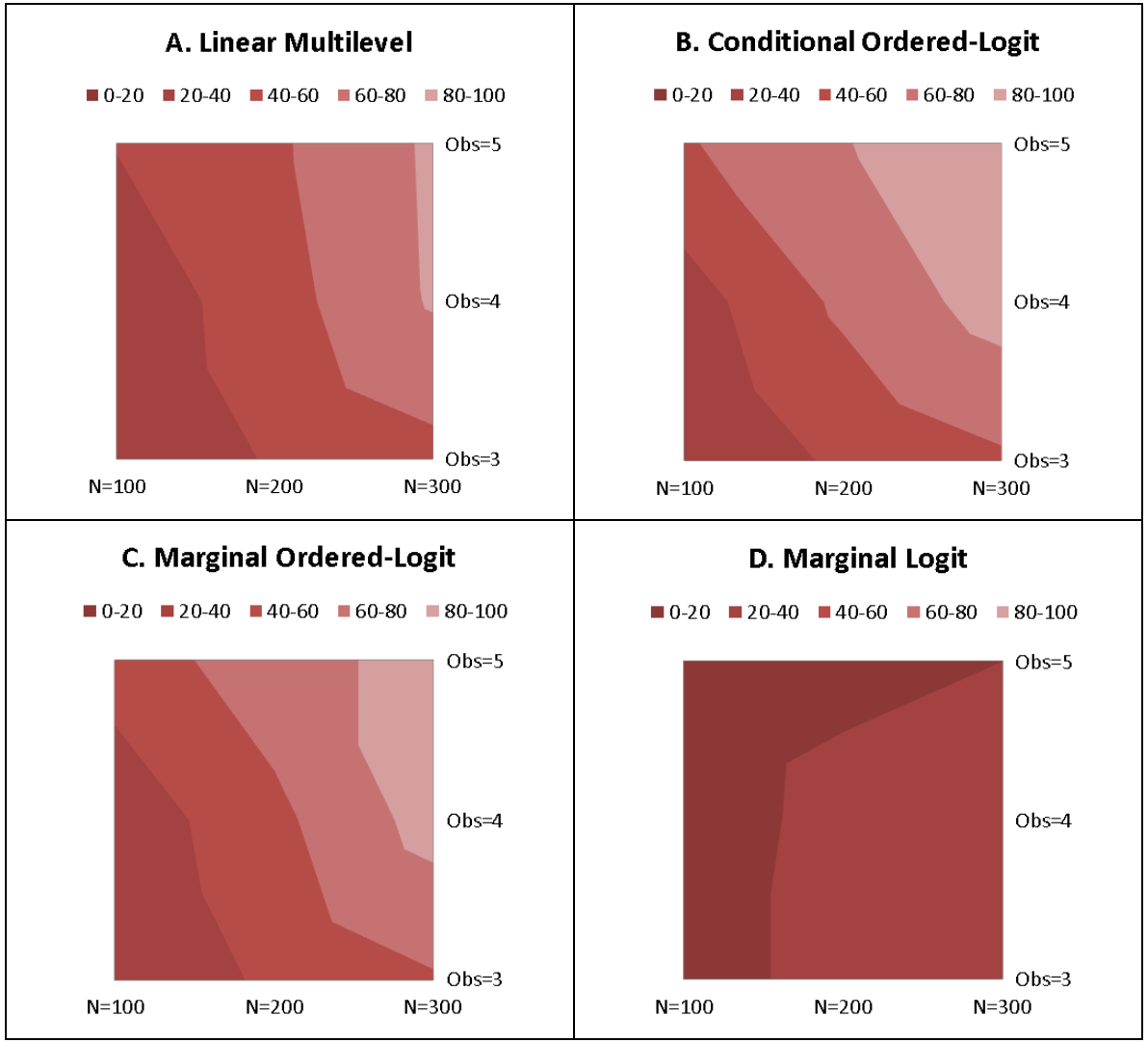
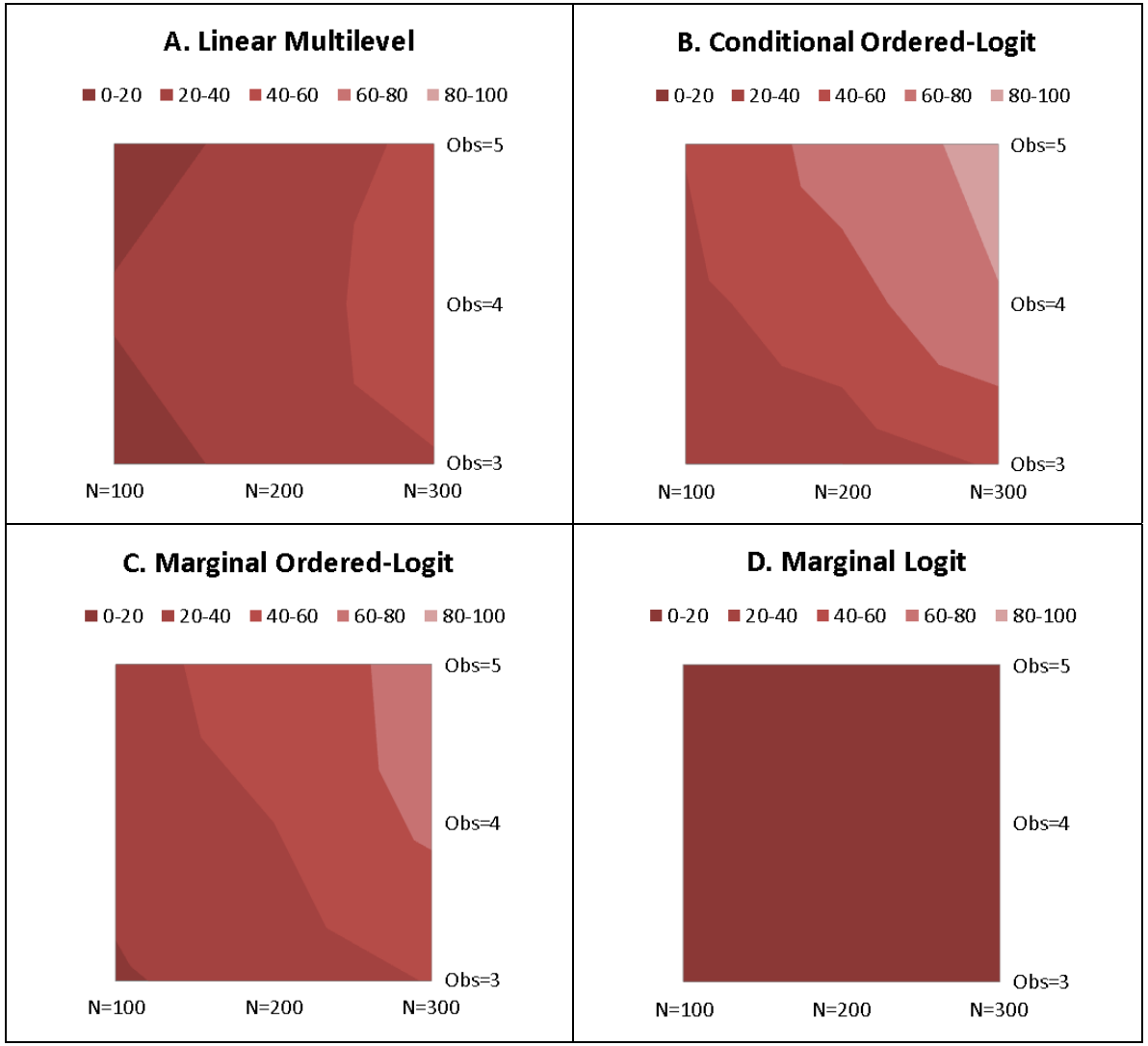


Figure 3: Model comparison – power surfaces for β_{12} at α=0.05; Power indicated by color.



RESULTS FROM THE HEALTH AND RETIREMENT STUDY

We use data from the Health and Retirement Study (HRS) to illustrate models A-C. The HRS is a nationally representative survey based on multi-stage probability sampling of U.S. adults (born 1931-1941) and their spouses. Respondents were first surveyed in 1992 and then repeatedly at two-year intervals through 2008. From a life-course perspective, we are interested in how health evolves over time at the individual level; in particular, whether SRH trajectories differ by race and gender controlling for age and education. To reduce computation time 10% (N=716) of the original sample is randomly selected, but all nine measurements per respondent are kept (see table 2 for descriptive statistics). Since we are impartial with respect to functional form, we first estimate an unspecified two-factor LGM (not shown here), which suggests that SRH progresses linearly over time. Consequently, we assume that a random intercept – random slope model structure is appropriate. In addition to models A-C, linear LGMs equivalent to model A are estimated – first with standard ML estimation and then with robust estimation of standard errors (models A2 and A3 respectively). Inference on fixed effects, if not coefficient estimates, can then be compared across all models. Results are summarized in Table 3.

Table 2: Descriptive statistics for Health and Retirement Study 10% Sample

Variable	N	Mean/Prop.	SD
Age	716	55.63	3.16
Black	716	0.17	0.38
Female	716	0.53	0.50
Female x Black	716	0.10	0.30
Education			
< High School	716	0.20	0.40
High School	716	0.40	0.49
Some College	716	0.22	0.41
College or Higher	716	0.18	0.38
Self-Rated Health			
1992	694	3.56	1.12
1994	638	3.50	1.11
1996	605	3.50	1.11
1998	577	3.32	1.07
2000	548	3.40	1.05
2002	532	3.28	1.01
2004	514	3.23	1.09
2006	486	3.23	1.05
2008	473	3.14	1.01

Table 3: Self-rated health model comparison, Health and Retirement Study 10% sample

	Model A1		Model A2		Model A3		Model B		Model C	
	Estimate	p-value	Estimate	p-value	Estimate	p-value	Estimate	p-value	Estimate	p-value
Effect on intercept										
Intercept1	-		-		-		-6.174	<0.001	-3.365	
Intercept2	-		-		-		-3.539	<0.001	-1.972	
Intercept3	-		-		-		-0.462	0.080	-0.355	
Intercept4	-		-		-		2.728	<0.001	1.424	
Intercept	3.600	<0.001	3.592	<0.001	3.592	<0.001	-		-	
Age	-0.038	0.001	-0.039	0.001	-0.039	0.001	-0.136	0.001	-0.071	0.001
Black	-0.425	0.005	-0.417	0.005	-0.417	0.013	-1.410	0.006	-0.827	0.008
Female	0.100	0.214	0.101	0.208	0.101	0.208	0.368	0.181	0.122	0.433
Female x Black	-0.300	0.131	-0.314	0.114	-0.314	0.135	-1.095	0.103	-0.501	0.187
< High School	-0.465	<0.001	-0.458	<0.001	-0.458	<0.001	-1.508	<0.001	-0.850	<0.001
High School	(ref)		(ref)		(ref)		(ref)		(ref)	
Some College	0.231	0.016	0.236	0.014	0.236	0.014	0.798	0.015	0.456	0.014
College or Higher	0.288	0.006	0.299	0.004	0.299	0.002	0.990	0.006	0.505	0.006
Effect on slope										
Intercept	-0.040	<0.001	-0.040	<0.001	-0.040	<0.001	-0.129	<0.001	-0.061	<0.001
Age	0.002	0.043	0.002	0.031	0.002	0.024	0.006	0.018	0.005	0.001
Black	0.016	0.138	0.015	0.165	0.015	0.188	0.059	0.094	0.034	0.144
Female	0.007	0.213	0.007	0.228	0.007	0.229	0.019	0.311	0.004	0.715
Female x Black	-0.001	0.931	0.000	0.974	0.000	0.976	0.008	0.861	0.008	0.783
< High School	-0.004	0.552	-0.004	0.576	-0.004	0.600	-0.013	0.571	0.011	0.478
High School	(ref)		(ref)		(ref)		(ref)		(ref)	
Some College	-0.010	0.120	-0.011	0.108	-0.011	0.096	-0.039	0.072	-0.008	0.567
College or Higher	0.003	0.698	0.002	0.740	0.002	0.738	0.003	0.914	0.008	0.598

(Table 3 continued)

Variance components					
Level-1 residual variance	0.395	-	-	-	-
Var(intercept)	0.754	0.716	0.716	8.738	-
Var(slope)	0.002	0.001	0.001	0.019	-
Cov(intercept, slope)	-0.014	-0.011	-0.011	-0.192	-

Models: A1 = Linear MLM; A2 = Linear LGM; A3 = Adjusted Linear LGM; B = Conditional Ordered-Logit; C = Marginal Ordered-Logit.

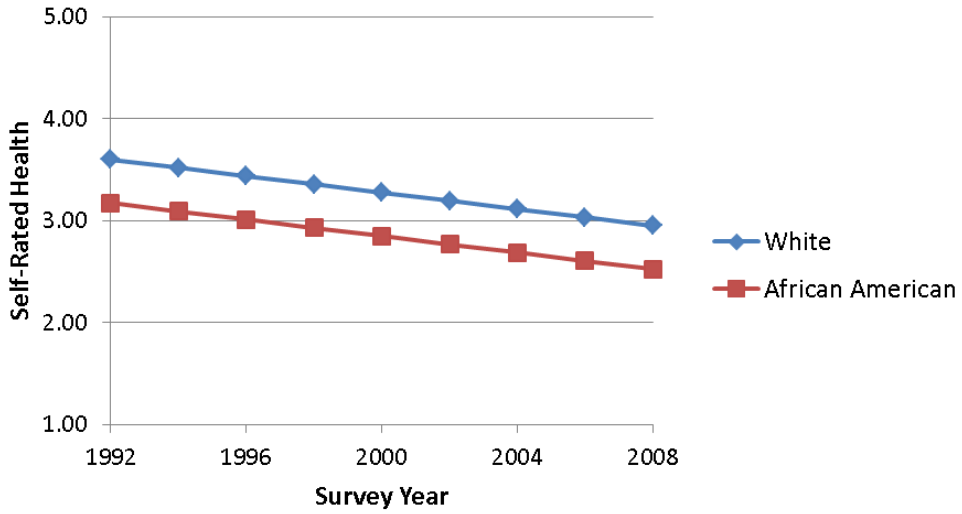
As expected, coefficient estimates and variance components are nearly identical across the MLM and LGM models (A1 and A2), when equivalently specified, and so are the standard errors and p-values. Model A3, an LGM with robust corrections for non-normality, shows similar coefficient estimates with only slightly elevated standard errors. Conclusions regarding fixed effects are unchanged, suggesting that departure from normality is not a major concern. With models A, B, and C fixed-effect coefficients cannot be directly compared; instead we comment on inferential conclusions. In all models, the effects of age and education on the random intercept are significant at the 0.05 level, but not the effect of gender or the interaction between gender and race. With respect to fixed effects on the random slope, only age has a statistically significant coefficient – this is consistent across models. However, there is significantly wider variation in p-values. These results echo our findings in the previous section: power is generally consistent across models with respect to fixed effects on the random intercept, but varies with respect to fixed effects on the random slope.

Since no gender differences are found, only race differences on initial SRH are interpreted. Figure 4 summarizes results from models A and B, when age is centered at the sample mean (55.5) and education at high-school level: in panel 4(a) blacks start with lower SRH than whites but trajectories progress in parallel over time; in panel 4(b) we collapse the probabilities (post-estimation, not to confuse with a binary logit model) of having "Good" health or below, to reveal a similar trend. Note that trajectories in panel 4(b) do not run parallel despite the fact that race has no effect on the random slope, due to the non-linear nature of the model. Since both models in Figure 4 are conditional models,

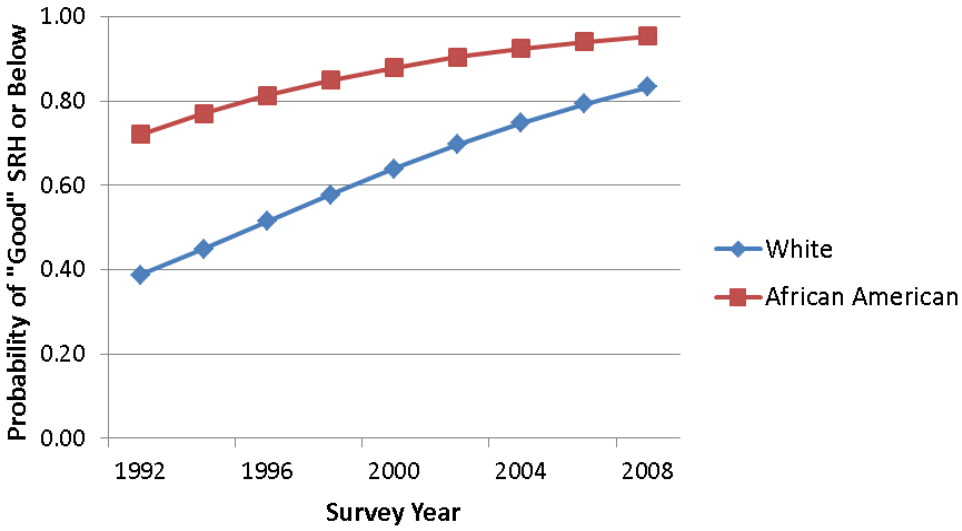
they should be interpreted as individual trajectories given fixed levels of age and education, but also conditional on random effects (in these plots, random effects are set at zero). The interpretation that follows is for individuals with *average* initial SRH and *average* rate of change in SRH, not all individuals in the sample or the population we wish to generalize to.

In Figure 5 models B and C are compared with probabilities for all SRH categories. Despite the similarities in inference shown above, model interpretation is fundamentally different. In panel 5(a) individual health trajectories for whites are conditional on fixed and random effects (again, random effects are set to zero). In panel 5(b) trajectories reflect the population-averaged trajectories, similar to cross-sectional results. Some marked differences between models exist: the probability of an individual reporting "Good" health in 2008 approaches 65% (conditional on mean random effects), but the population proportion in "Good" health is estimated at 38%. Overall, we observe a higher proportion of the population at the extremes (i.e., "Poor" and "Excellent" health) compared to an individual's probability of being in these categories (conditional on starting with mean initial SRH and having a mean rate of change in SRH). Clearly, a marginal model interpretation fails to capture how health unfolds over time at the individual level.

Figure 4: Predicted self-rated health trajectories from linear and non-linear conditional models.

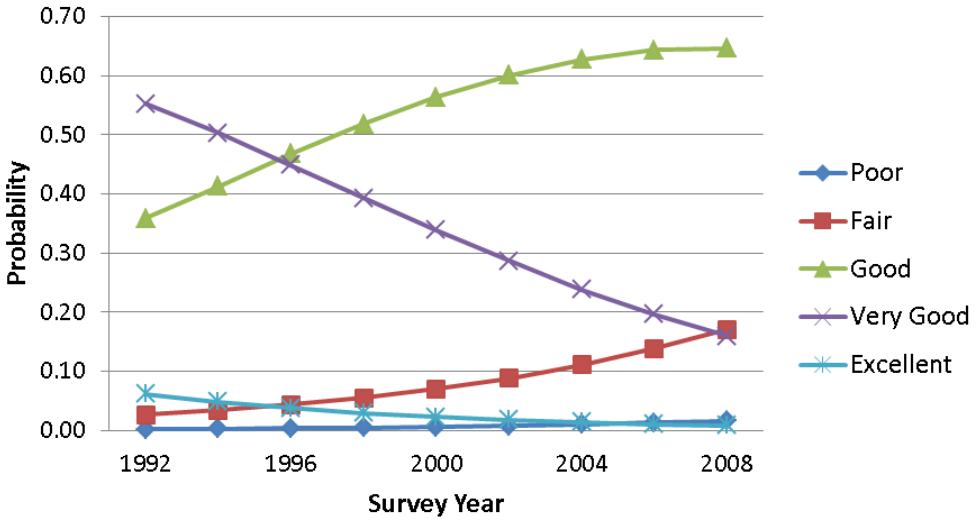


(a) Model A. Trajectories of mean SRH by race; random effects set at zero.

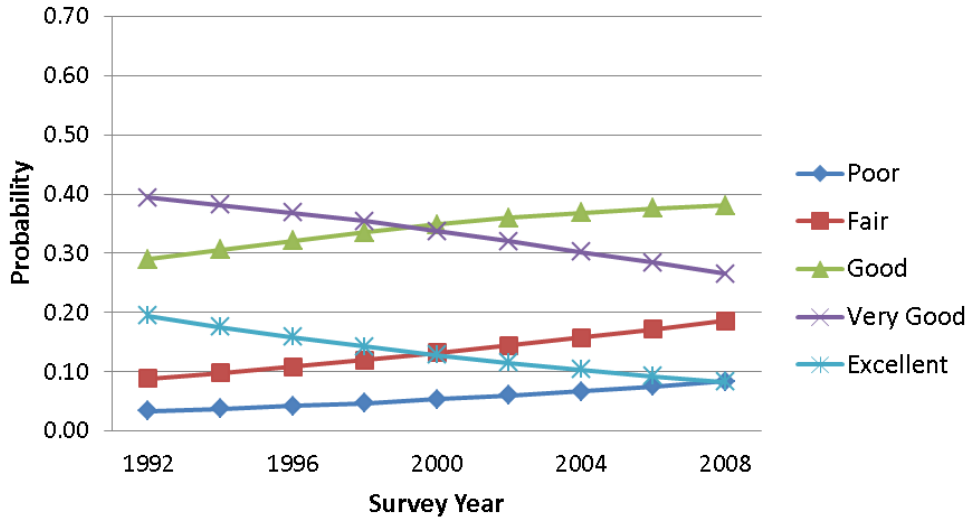


(b) Model B. Trajectories of the probability of reporting “Good” SRH or below by race; random effects set at zero.

Figure 5: Conditional and marginal trajectories of self-rated health from non-linear models.



(a) Model B. Individual SRH trajectory for whites; obtained from conditional ordered-logit model with random effects set at zero.



(b) Model C. Population-averaged SRH trajectory for whites; obtained from marginal ordered-logit model.

DISCUSSION

Longitudinal data are increasingly available and elaborate statistical methods are needed to fully exploit them. Regretfully, researchers often opt for cross-sectional analysis or focus on difference scores between two waves, while discarding enormous amounts of 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 better alternative.

In this paper competing longitudinal models for SRH are juxtaposed along two axes: conditional vs. marginal, and continuous vs. ordinal. Using simulations with randomly generated data, we show that multilevel ordered-logit models are generally more powerful than linear (normal) multilevel models in detecting fixed effects on growth parameters (in particular, the random-slope). Specifically, the conditional ordered-logit out performs all other models, but is computationally cumbersome. While others have suggested that dichotomizing SRH results in minimal reduction in efficiency (Manor et al. 2000), our results show that, in the longitudinal context, collapsing categories results in serious loss of power and is ill-advised. We further compare the models against real data and find consistent inferential conclusions with respect to fixed effects.

While marginal ordered-models may perform reasonably well and require little computation time, it is crucial to note that their interpretation refers to the population-averaged effects rather than individual life-course experiences. With non-linear models marginal and conditional predictions can show marked differences. In this regard, marginal models are unfit for inference on life-course trajectories just as repeated cross-sectional results are. It is also worth noting that the estimation of marginal models, if based on quasi-likelihood methods, is less robust to existence of missing data. In particular, it requires that data are missing completely at random (MCAR).

Conditional models are best fit for describing and making inference on individual trajectories over time. However, interpretation should be conducted with care: predicted 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 individuals with mean initial SRH and mean rate of change in SRH, given additional covariate values with fixed effects. Again, with non-linear models the mean person-level trajectory can dramatically differ from the population-level trajectory. Implicitly, every marginal model stems from a particular conditional model (Lee & Nelder 2004). Marginal (i.e., population-averaged) trajectories can generally be inferred from conditional models, but the opposite is not true.

Since studies of health outcomes typically focus on inference on fixed effects, we find that all modeling approaches (marginal or conditional, continuous or discrete) produce similar results with a large enough sample size. However, the models differ in power and interpretation. When interest lies in individual health trajectories conditional

models should be used. With respect to SRH, we recommend using linear multilevel models for exploratory analysis; some software packages allow adjustments to standard errors to account for departure from normality. For smaller sample sizes where power is a major concern, conditional ordered-logit models are preferable. Finally, we would like to echo 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 values, multi-stage survey designs, or model misspecification. Flexibility and caution are necessary with all models.

APPENDIX A: MARGINAL AND CONDITIONAL MEANS IN GLMM

A general GLM for longitudinal data can be specified as follows[4]:

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

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:

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

If the model includes random effects, as in the case of generalized linear mixed models, the marginal mean is the average of all conditional means (that is, conditional on random effects). An appropriate GLMM for longitudinal data can be expressed as (Agresti 2002, 492):

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

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:

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

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

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

This equation makes it clear 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

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

individual heterogeneity reflected in random effects). When g is the identity link, the conditional mean becomes an additive function of fixed and random effects (Ritz & Spiegelman 2004), which implies a special case where:

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

Thus, in the linear case, when the random effects are set to zero (typically at their mean), we have that $\mu_{it}^{C} = \mu_{it}^{M} = \mathbf{x}_{it}'\beta$.

APPENDIX B: POWER ANALYSIS THROUGH SIMULATION

Simulation was used to perform power analysis with Models A-D. Observations were drawn from the population specified in equation (7) using a pseudorandom number generator. The values chosen for the population parameters are shown in Table 4.

Table 4: List of population parameters and values used in simulation

Regression Coefficients		Variance Terms	
Parameter	Value	Parameter	Value
β_{00}	10	σ_ε	15
β_{01}	-16	σ_0	20
β_{02}	0.6	σ_1	7
β_{10}	4	ρ_{01}	-0.15
β_{11}	-4		
β_{12}	0.06		

The analysis was conducted in the statistical package Stata, and the accompanying user-written program GLLAMM (Rabe-Hesketh et al. 2004) was used to estimate the conditional cumulative-logit models.

First, two independent covariates, one continuous and one binary, were randomly generated for 300 observations:

```
set seed 2375
set obs 300
generate id = _n
generate x1 = rbinomial(1,0.5)
generate x2 = 100*runiform()
```

Since covariates are assumed fixed in regression analysis, the same set was used in all subsequent simulations.

As a second step, random effects were drawn for each of the 300 observations (representing individuals) from a bivariate normal distribution with means 0 and covariance matrix with the parameters specified in Table 4. Finally, a random sample of dependent variables was drawn from the population specified in equation (7), using the previously drawn covariates and random effects. Based on the random intercept – random slope population model multiple observations were drawn for each individual.

In order to assess statistical power 100 samples were drawn and each of the four models estimated using the same data. A count of the number of statistically significant effects was kept for each of the fixed effects. This procedure was repeated for 100, 200, and 300 individuals with 3, 4, and 5 observations each.

Due to the unusually long computation time required to estimate the conditional cumulative-logit model (Model B), especially when iterated repeatedly with multiple samples, only 100 samples were generated for each of the nine scenarios. However, the entire procedure was repeated for models A, C, and D with 1,000 samples in each scenario with almost identical results. Thus, we feel confident that 100 samples provide enough information to construct the power surfaces in Figures 1-3. Similarly, the procedure was repeated for a different set of random covariates with the same substantive conclusions.

The cut-points used in categorizing the dependent variable were the following: $0=\{y < 0\}$, $1=\{0 \leq y < 40\}$, $2=\{40 \leq y < 80\}$, and $3=\{y \geq 80\}$. In Model D, the dependent variable was dichotomized as $0=\{y < 50\}$, $1=\{y \geq 50\}$.

Below is the full code:

```
set seed 6598         // seed for replication
local N = 1000        // number of samples to generate
local nfe = 5         // number of fixed effects estimated
set obs `N'
scalar n_obs = 3      // number of observations per individual

// save p-values for 5 fixed effects from each sample
forvalues i = 1(1)`nfe' {
     generate pval`i' = .
}

local j = 0
while `j' < `N' {
    local j=`j'+1
    preserve

     // use sample of 100, 200, or 300 individuals
     use "SimData100.dta", clear

     // set population parameters
     scalar b00 = 10
     scalar b01 = -16
     scalar b02 = 0.6

     scalar b10 = 4
     scalar b11 = -4
     scalar b12 = 0.06

     scalar sigma_e = 15
     scalar sigma00 = 20
     scalar sigma11 = 7
     scalar sigcorr = -0.15

     // draw random effects from bivariate normal
     matrix msig = (sigma00, sigma11)
     matrix c = (1, sigcorr \ sigcorr, 1)
     drawnorm nu0 nu1, sds(msig) corr(c)
     quietly replace nu0 = b00 + b01*x1 + b02*x2 + nu0
     quietly replace nu1 = b10 + b11*x1 + b12*x2 + nu1

     // generate nested observations
     quietly expand n_obs
     sort id
     quietly egen t = seq(), f(0) t(4) by(id)

     // generate random-slope interaction terms
     quietly gen t_x1 = t * x1
     quietly gen t_x2 = t * x2
```

```
      // generate continuous outcome
      quietly generate resid = rnormal(0, sigma_e)
      quietly generate y = nu0 + nu1*t + resid

      // categorize outcome to 2 and 4 levels
      quietly egen d = cut(y), at(-1000,0,40,80,1000) icodes
      quietly egen b = cut(y), at(-1000,50,1000) icodes
```

[INSERT MODEL ESTIMATION PROCEDURE HERE]

```
      // save estimated coefficients and covariance-matrix
      matrix b=e(b)
      matrix V=e(V)

      // store estimates and p-values
      forvalues i = 1(1)`nfe' {
            scalar b`i'=b[1,`i']
            scalar tv`i'=b[1,`i']/sqrt(V[`i',`i'])
            scalar pval`i' = 2*(1-normal(abs(tv`i')))
      }

      restore
      forvalues i = 1(1)`nfe' {
            quietly replace pval`i'=scalar(pval`i') in `j'
      }

      // indicate successful simulation
    _dots `j' 0
}

// count number of significat runs under alpha=0.05
scalar alpha = 0.05
forvalues i = 1(1)`nfe' {
      count if pval`i' < alpha
}
```

The following estimation procedures were inserted in the code:

Model A: Linear Multilevel

```
      capture noisily {
            xtmixed d t x1 x2 t_x1 t_x2 ///
                  || id: t, cov(unstructured) mle
            display r(_rc)
            }
```

Model B: Conditional Cumulative-Logit

```
gen cons=1
eq alpha: cons
eq beta: t
capture noisily {
     quietly gllamm d t x1 x2 t_x1 t_x2, ///
          i(id) nrf(2) eqs(alpha beta) eform ///
          link(ologit) family(binom) adapt nip(5,5)
     display r(_rc)
     }
```

Model C: Marginal Cumulative Logit

```
ologit d t x1 x2 t_x1 t_x2, or vce(cluster id)
```

Model D: Marginal Logit

```
capture noisily {
     xtgee b t x1 x2 t_x1 t_x2, ///
          eform link(logit) family(binom) ///
          corr(unstructured) i(id) t(t) vce(robust)
     display r(_rc)
}
```

REFERENCES

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

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

Ananth, C., & Kleinbaum, D. G., 1997. "Regression Models for Ordinal Responses: A Review of Methods and Applications." *International Journal of Epidemiology* 26:1323-33.

Bailis, D. S., Segall, A., & Chipperfield, J. G. 2003. "Two Views of Self-Rated General Health Status." *Social Science & Medicine* 56:203-17.

DeSalvo et al. 2006. "Mortality Prediction with a Single General Self-Rated Health Question: A Meta-Analysis." *Journal of General Internal Medicine* 21:267-75.

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

Eriksson, I., Unden, A. L., & Elofsson, S. 2001. "Self-Rated Health. Comparisons between Three Different Measures. Results from a Population Study." *International Journal of Epidemiology* 30:326-33.

Fullerton, A. S., 2009. "A Conceptual Framework for Ordered Logistic Regression Models." *Sociological Methods & Research* 38:306-47.

Ghisletta, P., & Lindenberger, U. 2004. "Static and dynamic longitudinal structural analyses of cognitive changes in old age." *Gerontology* 50:12–16.

Hartford, A., & Davidian, M. 2000. "Consequences of Misspecifying Assumptions in Nonlinear Mixed Effects Models." *Computational Statistics & Data Analysis* 34:139-64.

Idler, E. L.. & Benyamini, Y. 1997. "Self-Rated Health and Mortality: A Review of Twenty-Seven Community Studies." *Journal of Health and Social Behavior* 38:21-37.

Jackson, D. L. 2010. "Reporting Results of Latent Growth Modeling and Multilevel Modeling Analyses: Some Recommendations for Rehabilitation Psychology." *Rehabilitation Psychology* 55:272-85.

Kaplan, G. A., & Camacho, T. 1983. "Perceived Health and Mortality: A Nine-Year Follow-Up of the Human Population Laboratory Cohort." *American Journal of Epidemiology* 117:292–304.

Kline, R. B., 2005. *Principles and Practice of Structural Equation Modeling*. New York, NY: Guilford Press.

Krause, N. M., & Jay, G. M., 1994. "What Do Global Self-Rated Health Items Measure?" *Medical Care* 32:930-42.

Lee, Y., & Nelder, J. A. 2004. "Conditional and Marginal Models: Another View." *Statistical Science* 19:219-38.

Litiere, S., Alonso, A., & Molenberghs, G. 2007 "Type I and Type II Error Under Random-Effects Misspecification in Generalized Linear Mixed Models." *Biometrics* 63:1038-44.

Manor, O., Matthews, S., & Power, C. 2000. "Dichotomous or Categorical Response? Analysing Self-Rated Health and Lifetime Social Class." *International Journal of Epidemiology* 29:149-57.

McCullagh, P. 1980. "Regression Models for Ordinal Data." *Journal of the Royal Statistical Society: Series B* 42:109-42.

Meredith, W., & Tisak, J. 1990. "Latent Curve Analysis." *Psychometrika* 55:107-22.

Orfila, F., Ferrer, M., Lamarca, R., & Alonso, J. 2000. "Evolution of Self-Rated Health Status in the Elderly: Cross-Sectional vs. Longitudinal Estimates." *Journal of Clinical Epidemiology* 53:563-70.

Rabe-Hesketh, S., & Skrondal, A. 2008. *Multilevel and Longitudinal Modeling Using Stata*. College Station, TX: Stata Press.

Rabe-Hesketh, S., Skrondal, A., & Pickles, A. 2004. "Generalized Multilevel Structural Equation Modeling." *Psychometrika* 69:167-90.

Ritz, J., & Spiegelman, D. 2004. "Equivalence of Conditional and Marginal Regression Models for Clustered and Longitudinal Data." *Statistical Methods in Medical Research* 13:309-23.

Satorra, A. 1992. "Asymptotic Robust Inference in the Analysis of Mean and Covariance Structures." *Sociological Methodology* 22:249-78.

Satorra, A., & Bentler, EM. 1994. "Corrections to Test Statistics and Standard Errors in Covariance Structure Analysis." In *Latent Variables Analysis: Applications for Developmental Research*, edited by A. von Eye & C.C. Clogg, 399-419. Thousand Oaks, CA: Sage.

Satorra, A. & Bentler, P. M. 2001. "A Scaled Difference Chi-Square Test Statistic for Moment Structure Analysis." *Psychometrika* 66:507-14.

Schafer, J. L.. & Graham, J. W. 2002. “Missing Data: Our View of the State of the Art.” *Psychological Methods* 7:147-77.

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

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

Zeger, S. L., & Liang, K. Y., 1986. “Longitudinal Data Analysis for Discrete and Continuous Outcomes.” *Biometrics* 42:121-30.