Using Longitudinal Complex Survey Data

Mary E. Thompson

Department of Statistics and Actuarial Science, University of Waterloo, Waterloo, Canada N2L 3G1; email: methomps@uwaterloo.ca

Annu. Rev. Stat. Appl. 2015. 2:305–20

The *Annual Review of Statistics and Its Application* is online at statistics.annualreviews.org

This article's doi:
10.1146/annurev-statistics-010814-020403

Keywords

complex sample design, time-in-sample effects, attrition bias, causal inference

Abstract

Common features of longitudinal surveys are complex sampling designs, which must be maintained and extended over time; measurement errors, including memory errors; panel conditioning or time-in-sample effects; and dropout or attrition. In the analysis of longitudinal survey data, both the theory of complex samples and the theory of longitudinal data analysis must be combined. This article reviews the purposes of longitudinal surveys and the kinds of analyses that are commonly used to address the questions these surveys are designed to answer. In it, I discuss approaches to incorporating the complex designs in inference, as well as the complications introduced by time-in-sample effects and by nonignorable attrition. I also outline the use and limitations of longitudinal survey data in supporting causal inference and conclude with some summary remarks.

1. INTRODUCTION

The purpose of this article is to review the field of longitudinal survey data analysis and to outline several areas for ongoing and future research. A comprehensive look at recent work in the use of longitudinal complex survey data could profitably begin with the book *Methodology of Longitudinal Surveys*, edited by Peter Lynn (2009). This volume contains the proceedings of a conference that addressed not only the theory of sampling, estimation, and accounting for non-sampling errors but also the practical and ethical issues involved in surveying longitudinally.

Longitudinal surveys by definition attempt to interview a sample of respondents repeatedly over time, usually in what are called waves or cycles. There are several reasons for undertaking such studies, of which I have identified five.

First, some of the largest and most ambitious longitudinal surveys have been undertaken to observe people over a lengthy period to see how risk factors at baseline predict longer-term behavior, health, and mortality. Examples in the United States include the Framingham Heart Study (Mahmood et al. 2013), the National Long-Term Care Survey (Manton 1988), and the National Children's Study (Guttmacher et al. 2013). In other countries, notable examples are the Dunedin Multidisciplinary Health and Development Study (referred to as the Dunedin Study) in New Zealand, in which newborns recruited in 1972 and 1973 have been followed for over 40 years (DMHDRU 2014); the National Population Health Survey in Canada (Tambay & Catlin 1995); the Understanding Society survey in the United Kingdom (ESRC 2014); and the Survey of Health, Ageing and Retirement in Europe (SHARE Proj. 2014), which is one of several similar surveys on aging around the world. The same kind of motivation is behind some longitudinal studies of shorter duration, such as the Galveston Bay Recovery Study (see Pietrzak et al. 2013), which was designed to assess the recovery of a local population following a natural disaster, Hurricane Ike, in September 2008.

Second, longitudinal surveys are often designed to provide efficient estimates of change in the population. Classic examples include the monthly Canadian Labour Force Survey, in which one-sixth of the sample is replaced each month, allowing efficient estimation of month-to-month changes, and the United States Current Population Survey, in which respondents are included in the survey for four months, excluded for the next eight months, and included again for a final four-month period, enabling precise estimates of not only month-to-month but also year-to-year change. Underlying the purpose of this design is a model in which an individual's employment status response is imagined to depend on calendar time (and the economic environment), as well as on the characteristics of the individual. Observing the responses of the same people at successive time points allows estimation of the time dependence while controlling for the personal characteristics of the respondents. The mathematics of estimating the current mean and change of mean in finite population sampling appears early in the sampling literature (see, for example, Patterson 1950). The principle for longitudinal data in general is discussed by Diggle et al. (2013).

The capacity to provide efficient estimates of change is valuable for a third purpose, namely, the evaluation of interventions (Piesse et al. 2009). For example, in a so-called natural experiment (Shadish et al. 2002), a ban on smoking in public might be introduced in one jurisdiction but not in another similar jurisdiction. Changes in public opinion can be measured and compared in the two jurisdictions using parallel longitudinal surveys before and after the introduction of the law. In the International Tobacco Control (ITC) Policy Evaluation survey in Ireland and the United Kingdom, a random sample of smokers in each country was interviewed before and after the March 2004 introduction of the smoke-free law in Irish workplaces. Public approval of a ban on smoking in restaurants and pubs increased in the samples in both Ireland and the United Kingdom, but the increase was steeper in Ireland. If we assume that both samples were similarly affected by panel conditioning and attrition, it is plausible that the difference reflected a greater change in opinion associated with the intervention in Ireland (Fong et al. 2006).

A fourth and related reason for conducting longitudinal surveys is the desire to explore and test causal hypotheses. One of the criteria proposed by Bradford Hill (1965) for the support of a causal hypothesis by observational data is that the putative cause tends to precede the outcome in time. A related criterion is that the effect "behaves appropriately when the potential cause is applied, removed, and then reinstated" (Cox 1992, p. 292). Applying such criteria requires the observation of individuals over a period of time. Moreover, as Cox (1992) has remarked, an event that can be observed as occurring later in time than the potential cause can be ruled out as a confounder. Thus, although observational studies cannot establish causality in the same way as randomized controlled experiments, longitudinal data can lend support to causal hypotheses or help rule out alternatives. In addition, careful observation and questioning of respondents can help to elucidate a causal mechanism.

Note that so-called instrumental variables are used in several disciplines to assist in measuring hypothesized causal effects (see, for example, Angrist et al. 1996). A variable is more plausibly an instrumental variable (a kind of exogenous trigger for the supposed cause) if it can be seen to be operating before the supposed cause, which in turn occurs before the effect.

Finally, a fifth, and more mundane, reason for conducting longitudinal studies is to reduce the cost of data collection for a survey designed to monitor economic or societal progress over time. In the twenty-first century, the recruitment of survey respondents has become increasingly difficult and expensive. Once these respondents have been recruited, however, it may be possible and fruitful to question them on several occasions. Data collection in later waves or cycles can therefore be much less expensive than recruitment and initial interviewing.

The main areas of ongoing and future research in the field of longitudinal surveys can be divided into two classes: One involves the practical aspects of implementing longitudinal surveys, and one relates to the associated statistical methods. One practical problem involves choosing the sampling frame or frames for a survey: Should a telephone number frame, an address frame, or a "rich" frame (with auxiliary data) from an administrative data source or an earlier survey be the basis of sampling? A second is that of determining an appropriate number of cycles or waves for the survey, as well as the time intervals between them: What frequency of interviewing will provide the needed measurements? A third problem is keeping rates of respondent recruitment and retention high: Can the respondent burden be reduced by innovative designs and methods of data collection? What is the best way to compensate respondents? Are there ways to increase the probability of staying in touch with respondents? Fourth, survey designers must develop appropriate instruments to answer the research questions: What are the essential questions for the subject matter and the analysis? Finally, researchers face the problem of reducing or accounting for measurement error: How can the accuracy of recall be increased? Can embedded experiments be used to help account for changes in the mode of data collection?

These practical problems are pervasive, and most are not unique to longitudinal surveys. They are manifested in new ways as survey culture and data collection technology evolve. One issue does however arise from the temporal nature of the data, or the fact that successive observations on a unit are associated with points in time. These time points may be widely separated, and some types of analyses may require inference about what happens between them. It is frequently the case that, although the stage or state of each respondent at each wave is determinable (possibly subject to some measurement error or misclassification), the transition times are either unknown or known only to within certain intervals. Survey developers often pay little attention to the desirability of collecting event times. Thus, the times of events between waves are often interval censored, and at the beginning of the survey, the time of entry into the current state is unknown, making the time of sojourn in the current state left censored.

The statistical methodology problems that arise in longitudinal survey analysis include the following: (*a*) designing the sample selection to address the research questions, (*b*) adapting complex survey methods to analyses with complex models, (*c*) accounting for nonresponse and attrition, (*d*) enhancing support for causal inference, and (*e*) developing and using data visualization techniques.

This article focuses on the latter group of problems. Section 2 describes the types of analysis commonly used for longitudinal data, and a subsequent short section is devoted to each problem area described above. I then conclude with a short list of summary points.

2. TYPES OF ANALYSIS OF LONGITUDINAL DATA

A common approach to modeling longitudinal survey data involves marginal repeated measures analyses, often using a variant of the generalized estimating equation (GEE) approach described by Liang & Zeger (1986). A GEE model is a generalized linear model for a response variable expressed in terms of time-invariant or time-varying predictors, assuming a working covariance or correlation structure to account for the dependence of responses of an individual over time. This approach allows the estimation of the marginal mean function of the response variable as a function of the survey wave. Although such models do not necessarily have well-defined joint distributions for the response variable values when the response is non-Gaussian, this method of analysis is robust in the following sense: Point estimators of regression coefficients are consistent if the estimating equations are in fact unbiased, and interval estimates have close to nominal coverage if the sandwich estimator of variance is used for the parameter estimates.

A common extension of a marginal mean repeated measures analysis is to model the response at time t in terms of both the response at time $t-1$ and other variables. A special case models the response transitions from one time point to the next and is aimed at measuring the associations of changes in response and covariates (see Diggle et al. 2013).

A contrasting approach to modeling dependence over time is the use of regression models with individual-level (and perhaps higher-level) random effects that are constant over time (Skinner & Holmes 2003) or that follow a stochastic model such as a time series (Feder et al. 2000). In these models, regression coefficients measure the averages of individual dependences rather than marginal dependences. A special case is a latent curve analysis, in which the model for the response includes a linear or piecewise linear function of time with a random intercept and coefficients (see, for example, Yong et al. 2012).

Interest is growing in modeling trajectories of health status over time. In the Galveston Bay Recovery Study, for example, Pietrzak et al. (2013) use latent growth mixture models to analyze the trajectories of scores on a posttraumatic stress disorder scale over time. Assuming that each respondent belongs to one of several classes of trajectory, they find the best fit for a three-class solution. In analyzing data from the National Long-Term Care Survey, Manrique-Vallier (2014) introduces mixed-membership trajectory models that assume the existence of a small number of ideal trajectories defining classes and that allow each individual to belong simultaneously to more than one class to varying degrees.

Finally, if the observations are collected often and at regular intervals, survival and event history analyses may be possible. For example, in a survey of a labor force one can model the spells of unemployment of a worker (Hadjucek & Lawless 2013).

A complex survey design has an impact on all of these analyses. Even if a survey were free from problems such as nonresponse, attrition, or measurement error, the sampling design might be informative, in the sense that parameters of the design are associated with either the response variables or, more generally, the response variables conditional on explanatory variables. For example, if one eligible respondent is selected at random in each sampled household, individuals in larger households will have smaller probabilities of inclusion, and a response variable of interest associated with household size would also be associated with design inclusion probabilities. Thus, estimates and estimating functions may be subject to bias unless the design is taken into account, for example via the use of inverse inclusion probability weights or strategic incorporation of household size in the model. Clustering of the sample requires either the addition of cluster random effects in the model or some kind of robust design-based standard error estimation that accounts for the structure of the sample. Some papers that are specifically aimed at accounting for complex survey designs from a design-based perspective are those by Binder (1992), Lin (2000), Boudreau & Lawless (2006), and Rubin-Bleuer (2011) for proportional hazard models, for which a rigorous treatment requires martingale or empirical process methods in the proofs of asymptotic properties; papers by Sutradhar & Kovacevic (2000), Skinner & Vieira (2007), Vieira & Skinner (2008), Roberts et al. (2009), and Carrillo et al. (2010) address this topic for repeated measures models with specified mean and covariance structures.

Longitudinal data sets are typically supplied with survey weights, and design-based methods using sample estimating functions that are weighted accordingly have become part of complex survey software. Such methods are thus more accessible to researchers outside statistics than are methods that require modeling the design and substantive variables (response and explanatory variables) jointly. The latter methods are appealing and typically efficient, but, in many cases, they have yet to be provided with a user interface that makes them accessible to a large community of researchers.

3. DESIGNING THE SAMPLE SELECTION

An article by Smith et al. (2009) provides a comprehensive look at sample design for longitudinal surveys. Survey designers must decide upon the following design features at baseline: the sample sizes in various subsets of the population, the choice of the sampling frame, the sampling units, and the method of selecting which units to approach. Researchers must also plan the frequency of interviewing, or the time intervals between returning to the respondents, as well as the total number of waves. The available auxiliary data, the budget, and the possible modes of data collection will influence these decisions. Above all, these decisions should be guided by the research questions and by the types of analyses to be conducted.

In addition to the correlation of responses over time, the effects of sample clustering in a multistage design should be taken into account when determining sample sizes. Often, the most convenient way to estimate power for complex analyses may be through simulation (Arnold 2011).

The key decisions for waves subsequent to the first are the extent to which dropouts or other respondents are to be replaced, how their replacements are to be chosen, and the extent to which the sample size should be increased. For some purposes, such as examining how risk factors at baseline predict longer-term behavior, health, and mortality, the ideal is no attrition and the ability to follow every respondent until the end of the study. In such studies, trying to ensure long-term participation may be more important than adding to the sample. For other purposes, such as the evaluation of interventions in the environment, however, continuing with the same sample (minus the dropouts) might not be desirable even if complete retention were possible. Aging and "conditioning" of the sample participants might result in a single cohort being representative only of itself after a small number of waves. In fact, replacing a substantial portion of the sample with new recruits at each wave in such a scenario can be very valuable, as doing so allows one to estimate the state of the population at each wave, accounting for attrition (Deng et al. 2013), and model the time-in-sample effect described by Thompson et al. (2005).

3.1. Accounting for Conditioning Through Sample Replenishment

The phenomenon known as panel conditioning, sometimes called the time-in-sample effect, is a fundamental feature of longitudinal surveys. A special case of this phenomenon has long been known as a component of rotation group bias (Bailar 1975) and continues to be observed and studied today in such contexts as the Current Population Survey in the United States. In general, the time-in-sample effect can be regarded as the impact participation in the survey has on responses. For example, in the ITC surveys, several kinds of variables are particularly prone to such effects. The values of responses to questions such as “In the past six months, how often have you noticed . . . ?” tend to decline markedly after the first time someone is interviewed. Such declines are probably at least in part a result of memory properties, such as differential recall bias: After the first interview, previous interviews act as milestones that can help to establish the timing of past events. For example, **Figure 1** depicts data from the ITC Canada survey (Driezen & Thompson 2011), which, under an assumption that the response variable does not predict future attrition, would appear to suggest a time-in-sample effect.

In another kind of pattern, the reported incidence of smoking cessation is greater among ITC respondents in the waves after recruitment than it is in the general population of smokers. One explanation for this phenomenon might be that being in the survey can support a resolve to quit. That is, this may be an example of a situation in which participation in the survey can alter the life course or underlying state of a participant. An alternative explanation is that “satisficing” (Krosnick 1991) may occur as respondents realize that quitters have a shorter survey to deal with.

One might also suspect that individuals who are less likely to try to quit smoking are more likely to drop out of the survey. If this is the case, failing to account for this association will also induce

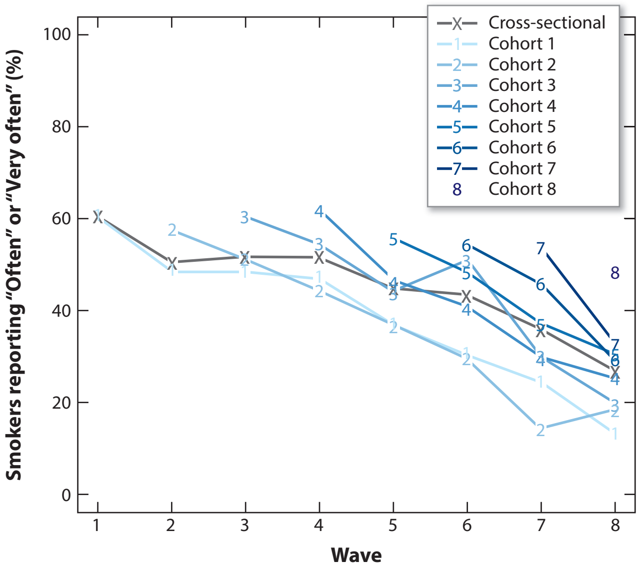

Figure 1

Weighted proportions of smokers reporting noticing warning labels on cigarette packages “Often” or “Very often” in the cohorts of the International Tobacco Control (ITC) Canada survey. Cohort c comprises the respondents recruited within wave c, for c = 1, ...

Table 1 Fictitious set of estimates of proportions of smokers noticing cigarette package warning labels as a function of response pattern

σ	$t_0(\sigma)$	$\ell(\sigma)$	$\bar{y}_1(\sigma)$	$\bar{y}_2(\sigma)$	$\bar{y}_3(\sigma)$
{1}	1	1	0.73		
{2}	2	1		0.69	
{3}	3	1			0.70
{1, 2}	1	2	0.75	0.65	
{2, 3}	2	2		0.76	0.64
{1, 2, 3}	1	3	0.80	0.68	0.61

an apparent increase in cessation rates in a cohort. In a longitudinal survey, the contributions of attrition and time-in-sample are difficult, if not impossible, to separate without strong model assumptions. If the sample is replenished (forming a new cohort) using the same sampling design at each wave, however, one can to some extent separate the effects of a changing environment from time-in-sample and attrition effects.

Suppose a longitudinal sample is observed at equally spaced waves $1, 2, \ldots, T$ and attrition is monotone: Dropouts do not return to being interviewed at later times. Then, one can define a response pattern σ as a set of indices of the waves at which a participant could respond before dropping out. For any response variable Y and wave t, one can compute a (weighted) estimator of the mean of Y at wave t based on respondents with response pattern σ, provided σ actually includes t, yielding a collection of estimators $\bar{y}_t(\sigma)$. A response pattern σ can be characterized by the pair $(t_0(\sigma), \ell(\sigma))$, where $t_0(\sigma)$ is the initial time of σ and $\ell(\sigma)$ is the length of σ.

Suppose the following functional form fits well:

$$\bar{y}_t(\sigma) \simeq \mu(t) + b(t - t_0(\sigma)) + d_{t_0(\sigma)}(\ell(\sigma)),$$

where $\mu(t)$ is a mean of $\bar{y}_t(\sigma)$ over all σ containing t, and $b(.)$ and $d_{t_0(\sigma)}(.)$ are functions.

The first term $\mu(t)$ represents the potential effects of environment. The next term is a time-in-sample effect, for fixed length of σ, and the last term is a kind of future attrition effect for fixed $t_0(\sigma)$. If $\bar{y}_t(\sigma)$ has this kind of form, the effects of environment, time-in-sample, and future attrition are observable and, to some extent, interpretable. Together with a model for the generation of σ, this construction would resemble the pattern-mixture models presented by Little (1993).

For example, **Table 1** gives a fictitious set of estimates of the proportion of smokers noticing cigarette package warning labels often or very often, controlling for demographics and cigarettes smoked per day. For each value of σ, the estimate decreases with the time in sample; the estimate at the initial point of σ tends to increase with the length of σ.

A logistic GEE version of the model to be applied might specify

$$\text{logit}[\text{Prob}(Y_{it} = 1)] = \mu(t) + \mathbf{x}'_{it}\beta + \gamma_b TIS_{it} + \delta_d \ell(\sigma(i)),$$

where $\mathbf{x}$ is a vector of covariates, TIS_{it} is the time-in-sample for respondent i at wave t, and $\sigma(i)$ is the response pattern of respondent i. The variables $\mu(t)$, TIS, and $\ell(\sigma(.))$ could be taken to be categorical. In the hypothetical example above without covariates, controlling for TIS and $\ell(\sigma(i))$ would lead to estimates of $\mu(t)$ that change little with t.

4. COMPLEX SURVEY METHODS FOR COMPLEX MODELS

The difference between complex survey analyses and other analyses is the necessity of accounting for complex sampling designs in the former. This necessity is not absolute; an analysis based on an erroneous simple random sampling assumption that happens to produce consistent point estimates and uncertainty measures will likely be preferable to an analysis that incorporates the design because the former will be more efficient. As a simple example, consider a population y with values that are independently and identically distributed (i.i.d.). If the unweighted sample estimating function

$$\sum_{i \in s} (y_i - \theta)$$

for the mean of y is unbiased (despite its dependence on s) with respect to the model for the Y values, it will be more efficient and thus preferable to the estimating function

$$\sum_{i \in s} \left(\frac{y_i - \theta}{\pi_i} \right).$$

The latter function, however, is unbiased with respect to the model and design combined and gives consistent estimates when the informativeness of the sampling design makes the former estimating function biased. In general, situations in which survey statisticians consider it safe to ignore the sampling design are rare.

There are two main approaches to incorporating the sampling design in the analysis. One is a design-based approach, typified by the second equation above, in which the sample estimating functions are design-unbiased estimators of a population-level estimating function that is unbiased under the model. Uncertainty estimates in this approach are also design based, so as to provide consistent estimates for the mean-squared error of the estimate under the model and design combined. For example, in the analysis proposed by Roberts et al. (2009), an estimating function bootstrap is employed to arrive at the design-based uncertainty measures. Depending on the basic estimating functions, this method can be referred to as a pseudo-likelihood (see, for example, Vieira & Skinner 2008) or pseudo-GEE (Carrillo et al. 2010).

The other approach to incorporating the design is to model both the design (and its variables) and the generation of the substantive variables together. For example, Pfeffermann & Sverchkov (1999) have tackled this problem by modeling the marginal distributions of the substantive variables assuming a particular informative model for the design inclusion probabilities. Feder et al. (2000) used random effects to model the contribution design clustering makes to uncertainty estimation and estimated substantive parameters using a maximum likelihood approach.

In the design-based approach, the problem of choosing weights to be used with various kinds of longitudinal analyses is still a topic of research. It is useful to try to keep in mind the following aspects of weighting:

1. Inflation weights, which are essentially the reciprocals of inclusion probabilities that have been adjusted for nonresponse and calibrated to known population totals, are usually interpreted in terms of the representation of a certain population at a certain point in time; that is, w_i is the number of members of a finite population at a certain time point represented by individual i.
2. For analytic purposes, the weights need not be the inflation weights.
3. For analytic purposes, the sample estimating function is a weighted sum of estimating function terms, and the weights should be such that the design expectation of the sample estimating function is a model-unbiased estimating function.
4. Intuitively, for efficiency of estimation and testing, the weights should be chosen such that similar individuals with similar response patterns have similar weights.

Suppose a single cohort representing the population at wave 1 is recruited and is observed for three waves with no dropout. Suppose the analysis is such that the sample estimating function is a weighted sum over individuals of complex individual-level terms, and an individual term cannot be separated into unbiased subterms that are calculable from wave 1 data, wave 2 data, and wave 3 data. It then seems most appropriate to use the cross-sectional weights for wave 1 and to consider the represented population to be the population of wave 1, observed over the three time points, with each member having a three-wave contribution to the population estimating function.

As a second scenario, suppose that the sample estimating function for a single cohort and no dropout does have terms that are separable, allowing the potential for different weights to be given to the three subterms of a single individual-level term. One example of such a scenario is a GEE analysis in which the working covariance matrix is completely specified. Suppose that the population is evolving, and that for each wave, there is a set of cross-sectional weights with which the sample members represent the population in that wave. In that case, there could be an argument for weighting the wave j contribution of the ith individual to the estimating function by the cross-sectional wave j weight of that individual. Such a weighting would make the sample estimating function unbiased for a kind of average of the full population single-wave estimating function subterms at the three waves. Most statistical software for GEE modeling currently requires that the individual weight remain constant over waves and thus would implement point and variance estimation, as in the first scenario described in the previous paragraph. SUDAAN (from Research Triangle Institute International) allows the weight of the individual to vary over waves. Running both analyses (one with constant weights and one with varying weights) can serve as a diagnostic. For example, in an analysis on waves 1, 2, and 3 of one ITC survey, the mean of a price variable was found to be significantly different for women and men in the constant weight analysis, but not in the varying weight analysis. Examination of the data showed that although the weights generally increased from the first to the third wave, the difference in mean by sex tended to diminish; including a wave $\times$ sex interaction brought the two fitted models close together.

Carrillo & Karr (2013) consider the problem of combining data from parallel cohorts in GEE analyses of data from the Survey of Doctoral Recipients. At each wave, each respondent has a cross-sectional weight representing the population of eligible people at the time of the wave. The authors advocate weighting the GEE estimating function term for individual i at wave j by the cross-sectional wave j weight of individual i, as in the second scenario above. This approach is easily adapted to accommodate attrition. However, one issue to consider when combining cohorts is the possible disparity in inflation weights. If, for example, a cohort starting at wave 2 has a much smaller sampling rate than the cohort starting at wave 1, the inflation weights of the wave 2 cohort will be much larger than those of the wave 1 cohort. According to the principle that similar individuals with similar response patterns should have similar weights, the appropriate wave 2 cross-sectional weights for a combined analysis may not necessarily be the inflation weights [for a related discussion, see Skinner & Mason (2012)].

5. ACCOUNTING FOR NONRESPONSE AND ATTRITION

There is a vast literature on accounting for nonresponse, missingness, or attrition both in longitudinal data and in the special case of survey data. Item nonresponse is an important subject, leading to some interesting problems in longitudinal data imputation, such as the event history application described by Wang et al. (2011). Bayesian methods of accounting for missing values in responses and covariates are well developed (Daniels & Hogan 2008). In this section, however, we focus on the literature on unit nonresponse: situations in which a potential respondent is not recruited or lost to follow-up or those in which a whole wave of data is missing for a recruited respondent.

Nonresponse at recruitment is analogous to nonresponse in surveys generally. However, wave nonresponse and attrition are special features of longitudinal studies. There are several reasons why researchers lose contact with a respondent in a longitudinal survey. Death or incapacity is one, and in many studies it can be regarded as an outcome rather than nonresponse. In other cases, the respondent moves out of the region or domain of interest and is no longer eligible for inclusion. In still other cases, the respondent may remain eligible but cannot be found or reached. Finally, researchers may be able to locate and contact eligible respondents but find them unwilling or unable to participate further.

We can almost never say that attrition occurs completely at random, in the sense of being independent of the responses and explanatory variables (Little & Rubin 2002). The phenomenon that younger people and people of lower educational achievement are harder to recruit to surveys is well known, and when these individuals are recruited, they are harder to retain. Even when controlling for background demographic variables, we seldom have attrition independent of the responses. For example, loss of interest in a survey may be associated with a downturn in fortune or health. Individuals or families with young children or other new responsibilities may find the task of survey completion too onerous to bear.

Most treatments for longitudinal surveys treat response versus nonresponse essentially as a second phase of sampling. The original sample is a probability sample from the population, for which the sampling probabilities, and in particular the inclusion probabilities, are known. The original sampling may or may not be informative. Nonresponse means that we observe a subsample, but the probability of the subsample given the sample is not known and must be modeled and estimated. This subsampling may also be informative and, if so, nonignorable in analyses. Not surprisingly, much of the theory about the analysis of survey data, which was developed for use with probability sampling designs, can be extended to the case of sampling with nonresponse in a two-phase paradigm.

Analyses involving complex survey data require that we account for the sampling design, and, as already noted, design-based approaches do this by including design weights in estimators or estimating functions. The weights must be adjusted in cases of attrition. If the individual estimating function terms are not separable, the individual can be weighted by the inverse or reciprocal of the product of the inclusion probability with the conditional probability of the response pattern, given inclusion. If the individual estimating function terms are separable, it might be fruitful to consider adjusting the weights for successive subterms. This is particularly true if the parameters being estimated are functions of time.

For example, Lawless (2003) looks at weighting for nonparametric (Kaplan–Meier) estimation of a survivor function in the case in which survival times are subject to censoring. He shows that, in general, weights related to both censoring and the sampling plan are needed for consistent estimation of the survivor function, and thus it is appropriate to use weights that depend on both the respondent and the time point. The arguments have been developed further in more recent work (namely, Hajducek & Lawless 2013).

In the alternative approach of modeling the substantive variables and the design variables jointly, one strategy is to add the response probabilities to the set of variables being modeled and then to estimate the parameters of interest with likelihood or Bayesian methods. This strategy can be relatively efficient, particularly for cases in which a mechanism for response-dependent sampling and missingness is well understood (see, for example, Lawless et al. 1999). However, some variations on this theme do not require full modeling. For example, Eideh & Nathan (1999) adapt a method proposed by Pfeffermann & Sverchkov (1999) and base inference on the marginal sample distribution for the substantive variables and dropout indicator. The imputation methods described by Zhang & Little (2009) use a model for the missing substantive variable data that is parametric in substantive variables and nonparametric in the propensity to respond; the latter relationship in their methodology is modeled by a penalized spline.

6. ENHANCING SUPPORT FOR CAUSAL INFERENCE

In the analysis of multivariate dependence, we often have in mind a graphical model (Lauritzen 2001) that describes a causal mechanism and that is Markovian and unidirectional provided that all covariates are accounted for. How can longitudinal survey data assist in establishing the validity of such a model or measuring its parts? As statisticians, we accept that causation can rarely be established without randomized controlled trials in which (*a*) treatments can be manipulated in a controlled manner and their outcomes can be observed, and (*b*) randomization guarantees that only the randomization device causes the treatment allocation. At best, observational data from longitudinal surveys can be used only to support or rule out causal hypotheses.

Longitudinal survey data can aid in model validation or measurement in the following ways:

1. Longitudinal surveys can be used to investigate whether certain events tend to precede certain other events. For example, in an illustration of a method for estimating the joint distribution of interval-censored event times, Pantoja-Galicia et al. (2009) showed that in data from the Canadian National Population Health Survey, attempts to quit smoking tend to increase following the start of pregnancy.
2. Longitudinal surveys can be used to evaluate interventions in "natural experiments" (Shadish et al. 2002). Often the first link in a hypothesized causal chain, termed a proximal outcome, is fairly easy to observe. A significant divergence between the responses of the treatment group and those of the control group, if one is observed, may (or may not) be due to the intervention. A lack of apparent divergence, however, suggests that the intervention has not had a discernible effect.
3. Longitudinal surveys can evaluate separate pieces of a causal chain and thus form part of a causal inference strategy. For example, it is well known that among cigarette smokers, an intention to quit at wave $t-1$ is a good predictor of an attempt to quit by wave t. If this relationship appears to be about the same at different times, in different places, and regardless of policy environments, the transtheoretical model of stages of change, a kind of causal chain model, is supported. Policy intervention then aims to try to accelerate the process by shortening the period of contemplation of change—the time that elapses before an intention is formed. Surveys may be used to elucidate the part of the chain that falls between registering new knowledge of the health consequences of smoking and developing an intention to quit.

The use of longitudinal survey data to support causal inference is hampered by the fact that although observed chains of occurrences may be causal in nature, these observations rely on information from respondents to measure environmental and other influences on their behavior, which may also stem from their behavior. In modeling reported cigarette consumption, for example, an endogenous variable or internal covariate is the price of cigarettes according to the respondent: An increase in cigarette price may lead a consumer to reduce consumption, whereas a consumer who has decided to reduce consumption may elect to purchase a more expensive variety. Similarly, the relationship between attempts to quit smoking and noticing antismoking advertising may be bidirectional. Inferences from these kinds of associations have to be predicated on a theory of the underlying processes.

A mediational model in its simplest form is a directed acyclic graph

$$X \mapsto M \to Y$$

that represents part of a causal mechanism. The interpretation of such a model is that conditional on background variables C, which are not shown for simplicity of the diagram, manipulation of the exposure X impacts the outcome Y by means of its impact on the mediator M. If M is a full mediator, then given C, Y is conditionally independent of X, given M. If M is not a full mediator, meaning that manipulation of exposure X can also lead directly to Y without involving M, then an arrow from X to M must be added to the diagram. Once the diagram is assumed, the estimation of the model parameters from data is straightforward. However, if the data are from an observational study, the interpretation of the parameters and the model as causal requires strong assumptions. VanderWeele & Vansteelandt (2009) have provided a comprehensive discussion of these assumptions, which are typically unverifiable: Observational data do not suffice to establish causality. At the same time, data from a longitudinal survey could discredit M as a mediator if the study found that M tends to precede X, and such data could support the notion of M as a mediator if it tends to follow X. The estimation of the model parameters from longitudinal survey data may therefore be useful as part of a larger body of evidence.

A randomized trial with X as treatment can establish X as a cause of Y, but it cannot necessarily explain how X causes Y. Evidence for the mechanism must come from other trials (Spencer et al. 2005) and observational studies of one kind or another, including those in which temporal sequences can be seen.

One somewhat surprising finding of the ITC surveys of smokers is that avoiding warning labels on cigarette packages by covering them up at wave t is positively associated with making an attempt to quit by wave $t + 1$, after controlling for demographics and cigarettes smoked per day. An increase in such avoidance of warning labels is well established as a proximal or nearly proximal response to the introduction of labels that are difficult to look at, but it is not obvious a priori whether an increase in avoidance should be followed by an increase or decrease in the likelihood of attempting to quit. At the same time, avoiding the sight of a warning label is not the same as avoiding thoughts about its message. One of the questions in the questionnaire asks about whether the warning label on a cigarette package makes the smoker think about the health risks of smoking, and the corresponding variable appears to be positively associated with avoidance. Thus, this variable is a plausible mediator for the association between avoidance and subsequent attempts to quit smoking, and the longitudinal structure of the data allows researchers to examine such a hypothesis. These relationships appear in work by Yong et al. (2014) and are the subject of ongoing study of the psychological mechanism(s) underlying smoking cessation.

7. USING DATA VISUALIZATION TECHNIQUES

Data visualization techniques can be helpful in the exploration of many types of data, including survey data. For longitudinal data with numerical or continuous measurements, trajectory plots in which the data from individual respondents are plotted as piecewise continuous lines against a time axis, sometimes called spaghetti plots, are particularly valuable. The lines in these plots can be colored differently or plotted separately, corresponding to the categories in any classification: strata, age–sex groups, or response patterns. A variant of these plots can be used to visualize and evaluate the development of survey weights (see Gelman 2007).

The older technique of plotting mean lines and confidence bands over time is still an important tool. For example, such plots for sample subgroups crossed with response patterns can facilitate the visual assessment of future attrition effects and time-in-sample effects.

The visual exploration of dependences can proceed with a low, medium, or high number of dimensions, as seen for example in work by Oldford & Waddell (2011) with non-survey data. Various techniques can be used to represent weighting in the data, such as varying the point size (bubble plots) or depth of color. Alternatively, including the weights themselves or a transformation of them among the variables to be examined can help to explain differences between weighted and unweighted analyses.

Geographical mapping can be an effective way of visualizing survey data, and this technique is an active area of research. For example, levels of important variables can be indicated by colors on the maps, and the locations of sample points with their weights can be indicated by bubbles or other objects. Density and intensity estimation based on survey data provide a picture of spatial trends. For example, Sangalli et al. (2013) develop a spatial spline method for analyzing data distributed over irregularly shaped spatial domains with complex boundaries and irregularities, and they use this method to produce a display of census data for the island of Montréal. Longitudinal data pose the challenge of portraying changes over time in a way that provides useful insights.

8. SUMMARY AND CONCLUSIONS

This article has reviewed several aspects of the design and use of longitudinal surveys, and it has identified some areas of current and future research. Some broad conclusions can be drawn, which are given in the following list of Summary Points. Future research will address other problems arising out of new challenges and opportunities in survey data collection and the computing environment. Some of these areas are listed below as Future Issues.

SUMMARY POINTS

1. The research questions and the plans for analyses should be articulated as far as possible at the design stage of a longitudinal survey, so as to inform both the sampling design and the design of the questionnaire. Questions that help determine the timing of events, as well as those that ask the respondent to attribute reasons for actions, can be helpful in elucidating the mechanisms underlying various events.
2. A longitudinal survey with a single cohort cannot reliably be used to estimate the prevalence of various characteristics of the population over time. Owing to conditioning and nonignorable dropout, after a couple of waves, the purely longitudinal sample is representative only of itself or of a hypothetical population of longitudinal survey responders. A replenished longitudinal sample, which consists of several cohorts, permits evaluation of and accounting for these sources of error, along with representation of the population over time.
3. Methods for analyzing longitudinal survey data are either (*a*) design-based and model-assisted using a pseudo-likelihood or pseudo-GEE, or (*b*) model-based with joint distribution of design variables and substantive variables. The ambitions of users of survey data to analyze complex models that take account of design features tend to exceed the capabilities of readily available statistical software. Work is needed to extend the modeling repertoire available to the subject matter researcher.
4. Modeling nonresponse and devising imputation and weighting strategies to deal with it continues to be an important area of research.
5. Longitudinal survey data can support causal inference within an assumed theory. However, these data can rarely establish causality in the absence of other kinds of evidence.
6. Data visualization will be an increasingly important tool not only for the analysis of relationships among substantive variables but also to evaluate the effects of design features on the analyses.

FUTURE ISSUES

1. Survey response rates are declining with changes in communications technology, and survey fieldwork is correspondingly becoming more expensive. More researchers are adopting designs with non-probability sampling of respondents, necessitating the development of new best practices for design and analysis.
2. Advances in data science have made possible the use of administrative data to supplement longitudinal survey data. Methods need to be developed for the use of imperfectly linked administrative records and supplementary data of varying quality.
3. Increasing demands for timeliness of survey results, as well as associated challenges posed by the storage and retention of data collected more frequently, will drive the development of new data processing and analysis software.
4. New methods, once fully tested and found to be practically useful, will have to be made accessible to researchers in the health and social sciences and to producers of official statistics.

DISCLOSURE STATEMENT

The author is not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

ACKNOWLEDGMENTS

This work is partially supported by a grant from the Natural Sciences and Engineering Research Council of Canada. The International Tobacco Control project is supported by grants from the National Cancer Institute of the United States (P01 CA138389) and the Canadian Institutes of Health Research (MOP-115016). Thanks are extended to an anonymous reviewer for very helpful comments.

LITERATURE CITED

Angrist JD, Imbens GW, Rubin DB. 1996. Identification of causal effects using instrumental variables. *J. Am. Stat. Assoc.* 91:444–55

Arnold BF, Hogan DR, Colford JM, Hubbard AE. 2011. Simulation methods to estimate design power: an overview for applied research. *BMC Med. Res. Methodol.* 11:94

Bailar BA. 1975. The effects of rotation group bias on estimates from panel surveys. *J. Am. Stat. Assoc.* 70:23–30

Binder DA. 1992. Fitting Cox's proportional hazards model from survey data. *Biometrika* 79:139–47

Boudreau C, Lawless JF. 2006. Survival analysis based on the proportional hazards model and survey data. *Can. J. Stat.* 34:203–16

Bradford Hill A. 1965. The environment and disease: causation or association. *Proc. R. Soc. Med.* 58:295–300

Carrillo IA, Chen J, Wu C. 2010. The pseudo-GEE approach to the analysis of longitudinal surveys. *Can. J. Stat.* 38:540–54

Carrillo IA, Karr AF. 2013. Combining cohorts in longitudinal surveys. *Surv. Methodol.* 39:149–82

Cox DR. 1992. Causality: some statistical aspects. *J. R. Stat. Soc. Ser. A* 155:291–301

Daniels MJ, Hogan JW. 2008. *Missing Data in Longitudinal Studies: Strategies for Bayesian Modeling and Sensitivity Analysis*. Boca Raton, FL: Chapman & Hall

Deng Y, Hillygus S, Reiter JP, Si Y, Zheng S. 2013. Handling attrition in longitudinal studies: the case for refreshment samples. *Stat. Sci.* 28:238–56

Diggle PJ, Heagerty P, Liang K-Y, Zeger SL. 2013. *Analysis of Longitudinal Data*. Oxford, UK: Oxford Univ. Press. 2nd ed.

DMHDRU (Dunedin Multidisc. Health Dev. Res. Unit). 2014. *The Science of Us: The 1,000 Most Studied People in the World*. Dunedin, NZ: DMHDRU

Driezen P, Thompson ME. 2011. *Comparing Policy Measures Across Multiple ITC Countries: Adjusting for Time-in-Sample*. Tech. Rep., Dec. 13, Int. Tob. Control Policy Eval. Proj. **http://www.itcproject.org/files/ITC_Technical_Report_time-in-sample-adjustment_Dec2011.pdf**

Eideh AAH, Nathan G. 2009. Joint treatment of nonignorable dropout and informative sampling for longitudinal survey data. See Lynn 2009, pp. 251–64

Elliott MR, Raghunathan TE, Li Y. 2010. Bayesian inference for causal mediation effects using principal stratification with dichotomous mediators and outcomes. *Biostatistics* 11:353–72

ESRC (Econ. Soc. Res. Counc.). 2014. *Understanding Society: The UK Longitudinal Household Study*. **https://www.understandingsociety.ac.uk/**

Feder M, Nathan G, Pfeffermann D. 2000. Multilevel modelling of complex survey longitudinal data with time varying random effects. *Surv. Methodol.* 26:53–65

Fong GT, Hyland A, Borland R, Hammond D, Hastings G, et al. 2006. Reductions in tobacco smoke pollution and increases in support for smoke-free public places following the implementation of comprehensive smoke-free workplace legislation in the Republic of Ireland: findings from the ITC Ireland/UK Survey. *Tob. Control* 15(Suppl. 3):51–58

Gelman A. 2007. Struggles with survey weighting and regression modelling. *Stat. Sci.* 22:155–64

Guttmacher AE, Hirschfeld S, Collins FS. 2013. The National Children's Study—a proposed plan. *N. Engl. J. Med.* 369:1873–75

Hajducek DM, Lawless JF 2013. Estimation of finite population duration distributions from longitudinal survey panels with intermittent followup. *Lifetime Data Anal.* 19:371–92

Krosnick J. 1991. Response strategies for coping with the cognitive demands of attitude measures in surveys. *Appl. Cogn. Psychol.* 5:213–36

Lauritzen SL. 2001. Causal inference from graphical models. In *Complex Stochastic Systems*, ed. OE Barndorff-Nielsen, DR Cox, C Klüppelberg, pp. 63–107. Boca Raton, FL: Chapman & Hall

Lawless JF. 2003. *Censoring and weighting in survival estimation from survey data*. Presented at Stat. Soc. Can. Annu. Meet., Jun. 8–11, Halifax, Can. **http://www.ssc.ca/survey/documents/SSC2003_J_Lawless.pdf**

Lawless JF, Wild C, Kalbfleisch JD. 2009. Estimation for response-selective and missing data problems in regression. *J. R. Stat. Soc. B* 61:413–38

Liang KY, Zeger SL. 1986. Longitudinal data analysis using generalized linear models. *Biometrika* 73:13–22

Lin DY. 2000. On fitting Cox's proportional hazards model to survey data. *Biometrika* 87:37–48

Little R, Zhang G. 2009. Robust likelihood-based analysis of longitudinal survey data with missing values. See Lynn 2009, pp. 317–32

Little RJA. 1993. Pattern-mixture models for multivariate incomplete data. *J. Am. Stat. Assoc.* 88:125–34

Little RJA, Rubin DB. 2002. *Statistical Analysis with Missing Data*. New York: John Wiley & Sons. 2nd ed.

Lynn P, ed. 2009. *Methodology of Longitudinal Surveys*. Chichester, UK: John Wiley & Sons

Mahmood SS, Levy D, Vasan RS, Wang TJ. 2013. The Framingham Heart Study and the epidemiology of cardiovascular disease: a historical perspective. *Lancet* 383:999–1008

Manrique-Vallier D. 2014. Longitudinal mixed membership trajectory models for disability survey data. arXiv:1309.2324 [stat.AP]

Manton KG. 1988. A longitudinal study of functional change and mortality in the United States. *Gerontology* 43:153–61

Manton KG, Stallard E, Woodbury MA. 1991. A multivariate event history model based upon fuzzy states: estimation from longitudinal surveys with informative nonresponse. *J. Off. Stat.* 7:261–93

Oldford RW, Waddell A. 2011. Visual clustering of high-dimensional data by navigating low-dimensional spaces. *Proc. 58th World Stat. Congr. Int. Stat. Inst., Dublin, Aug. 21–26*, pp. 3294–303. The Hague, Neth.: ISI. **http://2011.isiproceedings.org/papers/650370.pdf**

Pantoja-Galicia N, Kovacevic M, Thompson ME. 2009. Assessing the temporal association of events using complex longitudinal surveys. See Lynn 2009, pp. 333–50

Patterson HD. 1950. Sampling on successive occasions with partial replacement of units. *J. R. Stat. Soc. B* 12:241–55

Pfeffermann D, Sverchkov M. 1999. Parametric and semi-parametric estimation of regression models fitted to survey data. *Sankhya Ind. J. Stat. B* 61:166–86

Piesse A, Judkins D, Kalton G. 2009. Using longitudinal surveys to evaluate interventions. See Lynn 2009, pp. 303–16

Pietrzak RH, Van Ness PH, Fried TR, Galea S, Norris FH. 2013. Trajectories of posttraumatic stress symptomatology in older persons affected by a large-magnitude disaster. *J. Psychiatr. Res.* 47:520–26

Roberts G, Ren Q, Rao JNK. 2009. Using marginal mean models for data from longitudinal surveys with a complex design: some advances in methods. See Lynn 2009, pp. 351–66

Rubin-Bleuer S. 2011. The proportional hazards model for survey data from independent and clustered superpopulations. *J. Multivariate Anal.* 102:884–95

Sangalli LM, Ramsay JO, Ramsay TO. 2013. Spatial spline regression models. *J. R. Stat. Soc. B* 75:681–703

Shadish WR, Cook TD, Campbell DT. 2002. *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*. Boston: Houghton Mifflin

SHARE Proj. (Surv. Health Ageing Retire. Eur. Proj.). 2014. *Survey of Health, Ageing and Retirement in Europe*. **http://www.share-project.org**

Skinner CJ, Holmes DJ. 2003. Random effects models for longitudinal survey data. In *Analysis of Survey Data*, ed. R Chambers, CJ Skinner, pp. 205–19. Chichester, UK: John Wiley & Sons

Skinner CJ, Mason B. 2012. Weighting in the regression analysis of survey data with a cross-national application. *Can. J. Stat.* 40:697–711

Skinner CJ, Vieira MDT. 2007. Variance estimation in the analysis of clustered longitudinal survey data. *Surv. Methodol.* 33:3–12

Smith P, Lynn P, Elliot D. 2009. Sample design for longitudinal surveys. See Lynn 2009, pp. 21–34

Spencer SJ, Zanna MP, Fong GT. 2005. Establishing a causal chain: why experiments are often more effective than mediational analyses in examining psychological processes. *J. Personal. Soc. Psychol.* 89:845–51

Sutradhar B, Kovacevic M. 2000. Analyzing ordinal longitudinal survey data: Generalized estimating equations approach. *Biometrika* 87:837–48

Tambay JL, Catlin G. 1995. Sample design of the National Population Health Survey. *Health Rep.* 7:29–38. Statistics Canada, Cat. No. 82-003

Thompson ME, Boudreau C, Driezen P. 2005. *Incorporating time-in-sample in longitudinal survey models*. Presented at Stat. Can. Symp. Methodol. Chall. Future Inf. Needs, Oct.

Verhagen J, Fox JP. 2013. Longitudinal measurement in health-related surveys. A Bayesian joint growth model for multivariate ordinal responses. *Stat. Med.* 32:2988–3005

VanderWeele TJ, Vansteelandt S. 2009. Conceptual issues concerning mediation, interventions and composition. *Stat. Interface* 2:457–68

Vieira MDT, Skinner CJ. 2008. Estimating models for panel survey data under complex sampling. *J. Off. Stat.* 24:343–64

Wang C, Little R, Nan B, Harlow SD. 2011. A hot-deck multiple imputation procedure for gaps in longitudinal recurrent event histories. *Biometrics* 67:1573–82

Yong H-H, Borland R, Thrasher JF, Thompson ME. 2012. Stability of cigarette consumption over time among continuing smokers: a latent growth curve analysis. *Nicotine Tob. Res.* 14:531–39

Yong H-H, Borland R, Thrasher JF, Thompson ME, Nagelhout GE, et al. 2014. Mediational pathways of the impact of cigarette warning labels on quit attempts. *Health Psych*. 33:1410–20

Annual Review of Statistics and Its Application

Volume 2, 2015

Contents