MARKOV MODELS FOR LONGITUDINAL DATA FROM COMPLEX SAMPLES

Laurel A. Beckett, Rush-Presbyterian-St. Luke's; Dwight B. Brock, National Institute on Aging;
Paul A. Scherr, Centers for Disease Control; Carlos Mendes de Leon, Yale University
Laurel A. Beckett, Center for Research on Health and Aging, 710 S. Paulina, Chicago, IL 60612

KEY WORDS: pseudo MLE, transition models, logistic regression, aging

INTRODUCTION: Three practical aspects of recent epidemiologic research have led to the use of complex survey designs for long-term observational studies. First, policy development may require valid estimates for subgroups such as minorities or the very old, leading to stratified designs. Second, the statistical power in some studies may depend on the number of prevalent or incident cases identified, suggesting the use of two-phase designs that oversample from high-risk subgroups identified in the first phase. Finally, clustering may be used to reduce cost.

Recent years have seen considerable research in statistical methods for longitudinal data analysis. An overview was given in a double issue of *Statistics in Medicine* devoted to longitudinal data [1988]. The complex sampling field has undergone an expansion parallel to that for longitudinal data analysis; a technical monograph by Skinner, Holt, and Smith contains a review of the theoretical issues and an extensive bibliography [1989]. The linkage between complex sampling and longitudinal studies, however, is still in its early development. JNK Rao has noted the perils of ignoring the sampling plan in statistical analysis [1986], while other authors have argued that in some regression settings, adjustment using regression coefficients may suffice rather than a full likelihood-based approach [Korn and Graubard, 1991]. This paper examines the impact of different approaches to handling the sampling design in fitting a Markov transition model to longitudinal data collected in a large-scale aging study.

A MARKOV TRANSITION MODEL: Longitudinal studies can be used both to provide descriptions of change and to assess possible predictors of change. If the individuals in the study have been characterized by the presence or absence of a condition which may change over time, then transition models are a natural way to describe this change. Muenz and Rubinstein proposed a Markov model using a logistic link function to characterize the probability of transitions between two states as a function of covariates [1985]. A sample of n individuals is observed at up to k equally spaced times, although some individuals may have missing data at some time observation times. The individual is recorded as being in state 0, state 1, or missing at each time. In addition, a set of p fixed covariates $(x_1, \ldots, x_p)$ is observed for each individual, with no missing covariates. The underlying process is assumed to be memoryless and stationary. The model has separate logistic regression equations for the probability of transitions from state 0 to state 1, P_{01}, and the probability of transitions from state 1 to state 1, P_{11}:

$$P_{01} = (1 + e^{-(b_0 + b_1 x_1 + \cdots + b_p x_p)})^{-1} \qquad 1$$

$$P_{11} = (1 + e^{-(d_0 + d_1 x_1 + \cdots + d_p x_p)})^{-1}, \qquad 2$$

with $P_{i1} = 1 - P_{i0}$.

This model assumes that a unit increase in one of the covariates leads to a constant change in the log odds of a transition, regardless of the values of other covariates. Transition probabilities outside the range 0 to 1 are not permitted. Finally, the model permits direct comparison of the estimated effects of covariates to the results of other studies using logistic regression or Mantel-Haenszel methods.

POINT ESTIMATION: Muenz and Rubinstein proposed estimation of the logistic regression coefficients by maximum likelihood. If there are no gaps between observations (all successive pairs are exactly one time unit apart rather than two or more), the likelihood factors into two terms, one involving transitions from state 0 and depending on the b_k, and the other involving transitions from state 1 and depending on the d_k. The contribution of individual q to the log likelihood can then be written:

$$\begin{aligned} u_q = \sum_{q=1}^{n} [& m_q(00) \sum_{k=1}^{p} b_k x_k \\ & + (m_q(00) + m_q(01)) \log P_{00} \\ & + m_q(10) \sum_{k=1}^{p} d_k x_k \\ & + (m_q(10) + m_q(11)) \log P_{10}] \end{aligned} \qquad 3$$

where $m_q(ij)$ is the number of transitions individual q experienced from i to j. The complete log

ual q experienced from i to j. The complete log likelihood is given by

$$\log L = \sum_{q=1}^{n} u_q. \qquad 4$$

Maximum likelihood estimates (MLEs) of b and d can be obtained by finding the roots of equation (4) using the usual logistic regression algorithms, provided that both the total number of transitions from state 0 and the total number of transitions from state 1 exceed $p+2$ and that the design matrix of the covariates has full rank. Since this model has a general exponential family form, MLEs are best asymptotically normal. Moreover, the standard errors of the MLEs can be estimated for large n using the information matrix. In addition, if functions of the parameters such as estimated transition probabilities for specific age-sex combinations are of interest, the corresponding functions of the MLEs are also maximum likelihood.

The Markov model approach described so far could be implemented using existing software for logistic regression, provided each transition was handled as a separate observation. The situation addressed in this paper involves two complications: first, complex sampling rather than simple random sampling, and second, gaps in the data. We begin by comparing three approaches to estimation in the complex sample setting without gaps. These methods are compared in an example involving physical function in a stratified sample of persons 65 and older in New Haven, CT. We then explore possible modifications when gaps are present.

ESTIMATION FROM COMPLEX SAMPLE DESIGNS: Three possible approaches to parameter estimation in the complex sample setting are considered here. The first approach is to ignore the sampling design and to obtain MLEs of the regression parameters of the Markov model by the procedure described in the previous section, as if the study were based on a simple random sample. The parameter estimates are the roots of equation (4), and their covariance matrix can be estimated by the inverse information matrix using the second partial derivatives. There has been extensive discussion in the statistical literature about the appropriateness of this approach in regression models in the complex sample setting, the degree to which parameter estimates may be misleading, and the likely underestimation of the standard errors. DuMouchel and Duncan [1983] gave an overview for linear regression; similar considerations apply for logistic regression.

The second approach is to use the unweighted MLE procedure described in the previous section, but to add covariates such as stratum that are related to the sampling design, thus adjusting by regression for the design. Korn and Graubard [1991] suggested that this adjusts adequately for the sampling design in many cases. This approach is less appropriate when each stratum contains a few large clusters with substantial within-cluster correlation. The design effect would then be too large and would led to biases in estimating variances. They also stated that a general sufficient condition under which the weights would have little effect on bias of the estimates is: "The distribution of the outcome (disease) for given levels of the risk factor and covariates does not depend on any variables used in the sampling design or used to adjust for nonresponse." A weighted analysis may be less efficient if the stratum weights differ greatly, increasing the standard errors.

The third approach is to adjust using sampling weights, via pseudo maximum likelihood estimation. Skinner [1989] reviewed the problem of generalizing maximum likelihood estimation to complex samples in the general regression setting and noted that the full MLE would require many assumptions and would likely give a very complicated expression for the exact likelihood. An alternative is to define as the target parameter the so-called census coefficient, the solution of the population version of equation (4),

$$T(b,d) = \sum_{q=1}^{N} u_q(b,d) = 0. \qquad 5$$

The pseudo MLE is defined to be the solution of an appropriately weighted estimate of equation (5), with the individual contributions u_q to the score function calculated as in equation (3):

$$\hat{T}(b,d) = \sum_{q=1}^{n} w_q u_q = 0. \qquad 6$$

The pseudo MLE is a consistent estimator of the census coefficient, and it is asymptotically normal in the logistic regression setting. If equal weights are used, the pseudo MLE reduces to the usual MLE.

The most difficult technical problem in the complex sample setting is to obtain valid estimates of the sampling variances of the parameter estimates. In the simple random sample case, the covariance matrix of the parameter estimates can be estimated using the information matrix:

$$I(b,d) = -\partial \hat{T}(b,d)/\partial b, d \qquad 7$$

where $\hat{T}$ is the unweighted version of equation (6). The information matrix, even when weighted, is not the correct variance if the pseudo MLE is not the exact MLE [Skinner 1989]. An alternative approach is to use a Taylor series approximation for the sampling variance. This linear approximation is consistent under broad assumptions. The general form of the linearization estimator for regression models is given by

$$V(b,d) = I^{-1}V(\hat{T})I^{-1} \qquad 8$$

where I^{-1} is given by equation (7).

$$V(\hat{T}) = \sum_{h=1}^{H} \frac{n_h}{n_h - 1} \sum_{d=1}^{n_h} (z_{hd} - \bar{z}_h)(z_{hd} - \bar{z}_h)', \qquad 9$$

where z_{hd} is the sum of the $w_t u_t$ across sample units t in primary sampling unit d within stratum h.

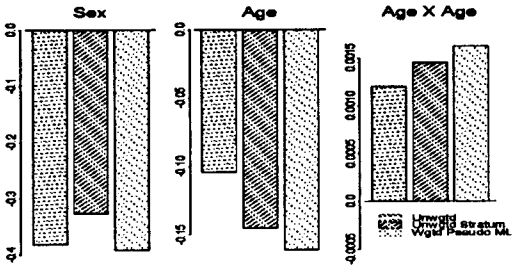
Details of this approach have been given for the logistic regression model by Roberts, Rao and Kumar [1987] and Chambless and Boyle [1985] and implemented in PC CARP [1986]. SAS PROC LOGIST [1986] has a weighted option that would give the pseudo MLE if the correct weights were used, but would not give the correct standard errors, since it uses the information matrix rather than a linearized variance estimator.

DuMouchel and Duncan [1983] took an intermediate position, suggesting that whether or not to use sampling weights in linear regression depends on the precise definition of the target parameter. They argued that, while the pseudo MLE is consistent for the census coefficient, this parameter may be difficult to interpret or misleading, especially in subgroups. On the other hand, the assumptions required for the unweighted estimates to be valid must be carefully checked and may require extra predictors, transformations, or interactions in the model.

COMPARISON IN AN EXAMPLE: The approaches described above were compared in a practical application, using 6 years of physical function data from the New Haven site of the Established Populations for Epidemiologic Studies in the Elderly (EPESE), sponsored by the National Institute on Aging [Cornoni-Huntley *et al.* 1986]. A sample of 2812 persons 65 and older was selected, stratified by sex and by three types of housing. A higher proportion of males was sampled than of females, and higher proportions of people from public housing and from private housing for the elderly were sampled than from general community housing. The frame was a 1979 utilities listing, and clusters of 12 consecutive housing units were chosen. The outcome variable for this study was self-reported impairment in mobility, a composite based on self-report of problems in any of four activities: transferring from a bed to a chair, walking across a small room, climbing stairs, or walking half a mile. Covariates were male sex, age, and a quadratic term for age.

We compared three methods for estimating the effect of the covariates on transitions to impaired mobility and recovery from impairment. Only results for transitions to impairment are presented here, but the results were similar for recovery. The first method was the usual unweighted MLE, ignoring the complex sample design. The second method was pseudo MLE, taking the design into account. The third method was the usual unweighted MLE, with indicators added for two of the three housing types in an effort to adjust for the sampling design. The data set was restricted to 1450 subjects with complete data at all 6 interviews, to avoid problems of gaps in the data, and commercial software was used (SAS PROC LOGIST and PC CARP).

Figure 1. Comparison of logistic regression coefficients from three approaches to handling complex sampling in the New Haven EPESE data: unweighted (ignoring sampling), unweighted (adjusting using housing stratum in regression), weighted (pseudo MLE).

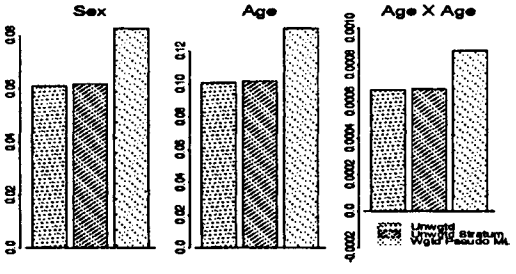


The unweighted MLEs of the logistic regression parameters differed consistently from the weighted pseudo MLEs (Figure 1). Compared to the pseudo MLEs, the unweighted estimates underestimated the quadratic effect of age, suggesting a less dramatic increase in the incidence of new impairment in mobility with increasing age. The unweighted but regression-adjusted estimates gave coefficients closer to the pseudo MLEs for the effects of age, but tended to underestimate the pro-

tective effect of male sex on incidence of impaired mobility. The estimated probabilities of incident impairment, calculated from the logistic regression coefficients, also showed systematic differences across methods, with the unweighted approaches overestimating incidence and underestimating recovery, compared to the pseudo MLE (results not shown here). These differences likely reflect the oversampling of males, the poor, and the very old.

Figure 2 shows the corresponding standard error estimates for each of the three methods. The estimated standard errors were substantially lower for both the unweighted and regression-adjusted MLEs than for the weighted pseudo MLEs (Figure 2). These biases would lead to coverage below the nominal level for confidence intervals and higher than nominal Type I error in hypothesis tests.

Figure 2. Comparison of estimated standard errors of coefficients for unweighted, regression adjusted, and pseudo MLE.



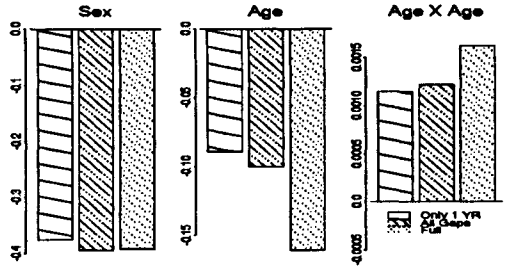
Thus, in this example it appears to be important to adjust for the sample design by appropriate weighted estimates with the corresponding linearized variance estimator.

ESTIMATION WHEN THERE ARE GAPS IN THE DATA: The likelihood formulas of equations (1) and (2), and hence the log likelihood formula of equation (3), do not give the correct formula for transitions across gaps of more than one time interval. The correct likelihood then would be the sum of the likelihoods of all possible paths. These likelihoods, log likelihoods, and the first and second derivatives can be calculated recursively, as shown in Muenz and Rubinstein [1985] and MLEs obtained. Similarly, in the complex sample setting, pseudo MLEs can be obtained by taking a weighted sum of the scores after calculating them recursively.

The commercial software available for obtaining pseudo MLEs for logistic regression does not allow for calculating likelihoods recursively across gaps. It is natural, then, to ask how much precision is lost by modifying the data set to exclude gaps of longer than one time period or by treating these gaps as if they lasted only one time period, so that commercial software can still be used.

We compared two different pseudo MLE approaches using commercial software with the full, recursion-based method pseudo MLE estimates of the Markov model parameters using a Fortran program developed for this grant. The complete New Haven data set of 2812 people was used. In the first approach, all transitions with gaps (one or more years of interview data missing between years with known mobility status) were omitted, and only one year transitions were used. In the second approach, all transitions were used, regardless of how many years had intervened between self reports of mobility status. The recursion-based pseudo MLE also used all transitions, but the likelihood was calculated recursively across all possible pathways during years with no recorded interview data.

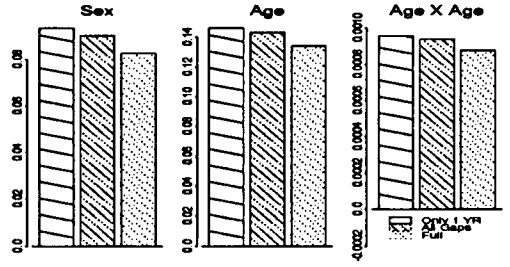
Figure 3. Comparison of regression coefficients for three methods of handling gaps in data: using only one year transitions, pretending all transitions were exactly 1 year, and estimating likelihood from recursion. All estimates were pseudo MLEs.



Omitting all but the one-year transitions led to very different estimates of the logistic regression coefficients (Figure 3), as well as underestimates of the incidence of impaired mobility and overestimates of the probability of recovery (not shown). This bias was a direct result of the tendency for longer gaps to be more likely to occur in people with health problems. Moreover, the reduction in effective sample size led to a substantial increase in standard errors (Figure 4). Including all gaps but treating them as if they lasted only one year gave standard error estimates closer to the results using the full recursive equations, but the parameter estimates still showed some differences (Figures 3

and 4).

Figure 4. Comparison of estimated standard errors for three methods of handling gaps in the data.



DISCUSSION:

This example showed that adjustment for the sampling design can make substantial differences both in parameter estimates and in standard errors for a longitudinal model. Failure to adjust adequately led to substantial biases in inferences about the population-wide effects of covariates on probabilities of change. Moreover, attempts to adjust by regression alone, without using the weighted pseudo MLE, did not adequately reflect the uncertainties of inference from the complex sample. The sampling design for the New Haven EPESE site had sampling proportions differing greatly across strata and moderate sized clusters within each sex-housing stratum; the effects would likely not have been so pronounced in a design closer to a simple random sample. Finally, while existing software could in principle be used for this Markov model by omitting some transitions or by ignoring the length of the gap, the results are sufficiently different from the full model to encourage using the full recursive likelihood when possible.

ACKNOWLEDGMENT: This research was funded by PO 2988626 ASB, CCDPHP, Centers for Disease Control and Prevention, Atlanta, Georgia.

REFERENCES:

Chambless LE, Boyle KE (1985). Maximum likelihood methods for complex sample data: logistic regression and discrete proportional hazards models. *Commun Statist Theory Meth* 14; 1377–1392.

Cornoni-Huntley J, Brock DB, Ostfeld AM, Taylor JO, Wallace RB, eds. (1993). Established populations for epidemiologic studies of the elderly: study design and methodology. *Aging Clin Exp Res* 5; 27–37.

Fuller WA (1986). PC CARP. Ames, Iowa: Statistical Laboratory, Iowa State University.

Herrell F (1986). PROC LOGIST. *SUGI Supplemental Library User's Guide.* Cary, North Carolina: SAS Institute.

Korn E, Graubard B (1991). Epidemiologic studies utilizing surveys: accounting for the sampling design. *Am J Public Health* 81; 1166–1173.

Muenz L, Rubinstein L (1985). Markov models for covariate dependence of binary sequences. *Biometrics* 41; 91–101.

Rao JNK (1986). Discussion on complex sampling session. *Proc Section on Survey Res* 46; Amer Stat Assoc.

Roberts G, Rao JNK, Kumar S (1986). Logistic regression analysis of sample survey data. *Biometrika* 74; 1–12.

Skinner CJ, Holt D, Smith TMF, eds. (1989). *Analysis of Complex Samples*, New York: Wiley Press.

Statistics in Medicine 7; (1988).