9-26-2005

Semiparametric Loglinear Regression for Longitudinal Measurements Subject to Irregular, Biased Follow-up

Petra Buzkova
University of North Carolina, buzkova@u.washington.edu

Thomas Lumley
University of Washington, tlumley@u.washington.edu

1 Introduction

Longitudinal data often are irregularly spaced and the follow-up times can vary from person to person. Moreover, those times often are continuous, not restricted to a predetermined set of times. Biased sampling under continuous times occurs when there is no effective control of the follow-up times. That happens when the scheduled follow-up times are not strictly followed, as in the health services research study we analyze, or when follow-up times are not scheduled at all, as in administrative data where the follow-up times are just observational times. We demonstrate the philosophy of biased sampling with a simple example taken from air pollution. Assume an air pollution measure at time t as a covariate $X(t)$. Let the outcome $Y(t)$ be a lung function measure, such as FEV1, the volume exhaled during the first second of a forced expiratory maneuver started from the level of total lung capacity. Scientists are interested in quantifying the association of FEV and air pollution. Further assume a binary indicator, $Z(t)$, of an asthma attack at time t. The lung function measure clearly is associated both with the air pollution measure and presence or absence of an asthma attack. Also, the occurrence of an asthma attack may be related to the the air pollution measure. Assume that a person with an asthma attack searches for medical help more often and that the person has a lower lung function measure. So, data on individuals with present asthma attacks form the majority of the observed data. If modeling FEV with the air pollution covariate we obtain unbiased estimates for those who have come for a visit, primarily people who suffer from an asthma attack on that day. Then, we obtain an exaggerated estimate of the association of the lung function measure and the air pollution measure for the general public. The bias can be overcome by including $Z(t)$ as a covariate, but this estimates the effect of X adjusted for Z, which may not be our targeted inference.

In a longitudinal study, we wish to examine the association of covariate process $\{X_i(t),\ t \in [0,\tau]\}$ and the response process $\{Y_i(t),\ t \in [0,\tau]\}$, where the predetermined constant τ is the end of study. Longitudinal data can be analyzed within the framework of fully marginal regression models. By "fully marginal" models we mean those that do

not require the assumption

$$E[Y_i(t)|X_i(t)] = E[Y_i(t)|X_i(\cdot)].$$

In particular, we do not require that the relationship between $Y_i(t)$ and future $X_i(\cdot)$ be specified, in contrast to GEE and many likelihood methods, see Pepe & Anderson (1994) or Pan et al. (2000).

We focus on a semiparametric outcome model where the intercept function $\alpha_0(\cdot)$ is an unspecified arbitrary function of time. The reason why non-parametric modeling of the intercept is attractive is that the effect of time may be complicated and it would be better modeled non-parametrically in order to avoid model misspecification. This concept is generalization to longitudinal data of the intercept in cross-sectional models based on one observation time point only. There, the intercept is, however, a one–dimensional unknown parameter, whereas here in longitudinal setting it is an infinitely–dimensional parameter. We note that this non-parametric intercept modeling is not needed in discrete times models with a small set of possible sampling times. There, we can add a sampling–time–specific parameter, resulting in a fully parametric model. A semiparametric regression with unspecified intercept is used for instance in Lin & Carroll (2001b) and Lin & Carroll (2001a). In their approach they use profile–based estimating equation for estimation of the parameter of interest and kernel estimating equation for the nonparametric estimation. We note that for longitudinal data kernel smoothing does not involve band–width selection issues only. It is a very hard task to provide a $\sqrt{n}$–consistent estimator there, achieved either by artificially under–smoothing or using a working independence in the profile–kernel estimating equations. They do not address biased sampling. Lin & Ying (2001) integrated counting processes techniques into analysis of longitudinal data under continuous time. They assume a linear regression model with unspecified intercept. They also claim that the parametric specification for the baseline mean function of the response variable over time is hard as " This can be a difficult task in practice". Bůžková & Lumley (2005b) generalized the approach of Lin

& Ying (2001) to biased sampling.

Recently, H. Lin et al. (2004) developed a class of weighted estimators in marginal generalized regression models for longitudinal responses that might be observed in a continuous-time fashion. Their outcome model covariates are fixed over time. Their sampling-times model requires the estimation of a smooth hazard rate. This complicates estimation, but more importantly, rules out sampling times that have some positive probability, as occurs when there is partial compliance with a discrete set of planned observation times. Bůžková & Lumley (2005a) proposed a similar approach that relies on a proportional rates model for sampling times but avoids the need for a smooth hazard function. This estimator is simple and easily implemented and choosing a log-link provides a parametric counterpart of the estimator proposed in this paper.

We suggest a class of estimators that in loglinear semiparametric models account for the possibility of biased sampling due to follow-up dependent on outcome or outcome-related auxiliary variables. The loglinear models are suitable for Poisson, Gamma or even Binomial data. Covariates in both the outcome model and the observation-times model are not restricted in any way. For instance, the response at previous sampling time can be included or an average of a covariate over a subject's history.

Notation and the proposed models are described in Section 2. Section 3 discusses estimation and inference. We illustrate our methods on a health services research study, called HUD–VASH study, in Section 4. In Section 5 we report simulation studies and in Section 6 we describe a second estimation approach for the loglinear semiparametric model and give a few concluding remarks.

2 Notation and models

We assume a fully marginal mean model for response Y_i as a function of covariates X_i of individual $i \in \{1, \ldots, n\}$ at time $t \in [0, \tau]$. Thus we model $E[Y_i(t)|X_i(t)]$, denoted

by $\mu_i(t)$. We consider the full data semi–parametric log–link model

$$\log\left(\mu_i(t)\right) = \alpha_0(t) + \beta_0^T X_i(t). \tag{2·1}$$

The effect of time-varying covariates X_i is modeled parametrically in a linear way. The parameter of major interest, β_0, is a p-dimensional vector. The intercept curve $\alpha_0(\cdot)$ is not of special interest, it is a infinite-dimensional nuisance parameter, whose estimation we avoid.

The model for response, formulated in equation (2·1), is a full data model. However, we assume to observe response not continuously over time but at certain observation times only. Denote the set of observation times $\{T_{i1}, T_{i2}, \ldots, T_{iK_i}\}$ for individual $i \in \{1, \ldots, n\}$ as $\mathcal{T}_i$, with $0 \leq T_{i1} < T_{i2} < \ldots < T_{iK_i} \leq \tau$. K_i is a random total number of observed events of the i-th individual. Denote $\mathbb{T} = \{\mathcal{T}_j,\ j = 1, \ldots, n\}$ the set of sample's observation times. Define $N_i(t) = \sum_{k=1}^{K_i} I(T_{ik} \leq t)$ the counting process of number of events of individual i by time t. The underlying uncensored process denote as $N_i^*(\cdot)$, $N_i(t) = N_i^*(t \wedge C_i)$ where C_i is drop-out time or end of follow-up τ, whatever comes first. We assume a marginal rate model for uncensored observation times of each individual $i \in \{1, \ldots, n\}$ at time $t \in [0, \tau]$:

$$E[dN_i^*(t)|Z_i(t)] = \exp\{\gamma_0^T Z_i(t)\} d\Lambda_0(t). \tag{2·2}$$

The cumulative intensity $\Lambda_0(\cdot)$ is an arbitrary non-decreasing function of time t, continuous up to countably many points, in our settings a finite number of points suffices.

We write $\xi_i(t) = I(C_i > t)$ for the at-risk process based on the drop-out C_i and assume that $\Pr(C_i \geq \tau) > 0$. We define two weighted averages of a variable V at time

t, Av_1 and Av_2, as

$$Av_1(V)(t;\beta,h) = \sum_{i=1}^{n} V_i(t)\frac{\xi_i(t)h(X_i(t))\exp\{\beta^T X_i(t)\}}{\sum_{j=1}^{n}\xi_j(t)h(X_j(t))\exp\{\beta^T X_j(t)\}} \quad (2{\cdot}3)$$

$$Av_2(V)(t;\gamma) = \sum_{i=1}^{n} V_i(t)\frac{\xi_i(t)\exp\{\gamma^T Z_i(t)\}}{\sum_{j=1}^{n}\xi_j(t)\exp\{\gamma^T Z_j(t)\}}. \quad (2{\cdot}4)$$

The weighted average Av_1 has weights proportional to a function $h(\cdot)$ and $\exp\{\beta^T X_i(t)\}$. It it a multiplicative centering of the variable V at time t. The weighted average Av_2, being the expected value for a sampled person at time t, has weights proportional to the probability of the individual having an observation at time t, based on the observation-times model (2·2).

There are two crucial assumptions characterizing the models. They are non-informative drop-out for the mean of response,

$$E[Y_i(t)|X_i(t), C_i \geq t] = E[Y_i(t)|X_i(t)], \quad (2{\cdot}5)$$

saying that $EY_i(t)$ depends on covariates $X_i(t)$ and drop-out C_i through covariates $X_i(t)$ only, and independent sampling assumption,

$$E[dN_i^*(t)|Z_i(t), X_i(t), Y_i(t), C_i \geq t] = E[dN_i^*(t)|Z_i(t)], \quad (2{\cdot}6)$$

saying that sampling times depend on covariates $Z_i(t), X_i(t)$, on response $Y_i(t)$ and drop-out C_i through covariates $Z_i(t)$ only.

The end of follow-up, τ, is a constant. Note, that although response Y_i is observed only at random times T_{ij}, the expectations in (2·1), (2·5) and (2·6) do not condition on those times. Additional technical assumptions are given in the Appendix.

3 Estimation

3·1 *Observation-times model*

Based on the proportional rates model (2·2) and the drop-out part of assumption (2·6), parameter vector γ_0 of length g can be consistently estimated by $\hat{\gamma}$, the solution to a set of estimating equations $U^\dagger(\hat{\gamma}) = 0$. The estimating function $U^\dagger(\gamma)$ is defined as

$$U^\dagger(\gamma) = \sum_{i=1}^{n} \int_0^\tau \{Z_i(t) - Av_2(Z)(t;\gamma)\}\, dN_i(t), \tag{3·1}$$

where Av_2 is defined in equation (2·4). Solution of (3·1) and derivation of asymptotic properties of the estimator are based on a zero mean random process $\{\mathcal{M}_i(t;\gamma_0,\Lambda_0),\ t \in [0,\tau]\}$ defined as

$$\mathcal{M}_i(t;\gamma,\Lambda) = N_i(t) - \int_0^t \xi_i(s)\exp\{\gamma^T Z_i(s)\}\, d\Lambda(s). \tag{3·2}$$

Though the estimating function (3·1) is the same as under the Cox proportional hazards model, the asymptotic variance is different due to imposing weaker assumptions in the proportional rate model (2·2). Define the limit of the weighted average Av_2 of a variable V at time t as

$$Av_2(v)(t;\gamma) = \lim_{n\to\infty} Av_2(V)(t,\gamma) = \frac{E[V_1(t)\xi_1(t)\exp\{\gamma^T Z_1(t)\}]}{E[\xi_1(t)\exp\{\gamma^T Z_1(t)\}]}$$

and denote $X^{\otimes 2} = XX^T$ the outer product of X. The asymptotic variance of $\sqrt{n}(\hat{\gamma}-\gamma_0)$ is Γ, $\Gamma = A^{-1}\Sigma A^{-1}$, where

$$\begin{aligned} A &= \lim_{n\to\infty} E\frac{1}{n}\left[\frac{-\partial U^\dagger(\gamma)}{\partial\gamma}|_{\gamma_0}\right] \\ &= E\int_0^\tau [Z_1(t) - Av_2(z)(t;\gamma_0)]^{\otimes 2}\,\xi_1(t)\exp\{\gamma_0^T Z_1(t)\}\, d\Lambda_0(t) \qquad (3·3) \\ \Sigma &= \lim_{n\to\infty} \mathrm{Cov}\left[\frac{1}{\sqrt{n}}U^\dagger(\gamma_0)\right] = E\left\{\left[\int_0^\tau [Z_1(t) - Av_2(z)(t;\gamma_0)]\, d\mathcal{M}_1(t;\gamma_0,\Lambda_0)\right]^{\otimes 2}\right\}. \end{aligned}$$

In deriving the properties of the estimators in the mean-response model we will need only matrix A and therefore we refer a reader to Lin et al. (2000) for detailed derivation of the variance of the estimator of γ_0. There is a straightforward consistent estimator of A

$$\hat{A} \;=\; \frac{1}{n}\sum_{i=1}^{n}\int_0^\tau [Z_i(t) - Av_2(Z)(t;\hat{\gamma})]^{\otimes 2}\,\xi_i(t)\exp\{\hat{\gamma}^T Z_i(t)\}\,d\hat{\Lambda}(t)$$

with Aalen–Breslow estimator of $\Lambda_0(t)$

$$\hat{\Lambda}(t) = \sum_{i=1}^{n}\int_0^t \frac{dN_i(s)}{\sum_{j=1}^{n}\xi_j(s)\exp\{\hat{\gamma}^T Z_j(s)\}}.$$

3·2 *Biased sampling*

To adjust for the biased sampling we use an approach based on inverse intensity rate ratio weighting. For individual $i \in \{1,\ldots,n\}$ at time $t \in [0,\tau]$ define inverse weights $\rho_i(t;\gamma,h)$ as

$$\rho_i(t;\gamma,h) = \frac{\exp\{\gamma^T Z_i(t)\}}{h(X_i(t))}. \tag{3·4}$$

In the numerator we can include any function $h(\cdot)$ that is a deterministic function of the mean–response model covariates $X_i(t)$. The weight is proportional to the inverse of the probability of individual i having an observation at time t. Ideally we would like the weight to be proportional to inverse of variance of response. With no knowledge about the variance of response, we want to make the weights variance, $\mathrm{var}[\rho_i(t;\gamma,h)^{-1}]$, as small as possible to increase estimators efficiency. Motivated by Hernán et al. (2002) we try to find a function $h(\cdot)$ that decreases the variability of the weights. We choose

$$h_0(X_i(t)) = \exp\{\delta_0^T X_i(t)\}$$

and we call the inverse weight $\rho_i(t;\gamma,h_0)$ a stabilizing inverse weight. The best choice of δ_0 we base on an estimator of δ_0 in a proportional rate model similar to model (2·2)but condition on covariates X instead of Z. When observation-times model covariates $Z_i(t)$ are a subset of the mean–response model covariates $X_i(t)$, for all individuals at all times, then $\rho_i(t)=1$, using the independent sampling assumption (2·6).

3·3 *Estimation in the loglinear model*

Motivated by the generalized estimating equations, we develop a class of estimators defined as solution to unbiased estimating equations in the model (2·1). To compute the estimates we do not require to use any smoothing techniques for consistent estimation of the nuisance function $\alpha_0(t)$ or the baseline intensity $d\Lambda_0(t)$. For individual $i\in\{1,\ldots,n\}$ let us define the process $\{M(t;\mathcal{A},\beta,\gamma,h),\ t\in[0,\tau]\}$

$$\begin{aligned} M_i(t;\mathcal{A},\beta,\gamma,h) &= \int_0^t \frac{1}{\rho_i(t;\gamma,h)}\{Y_i(s)\,dN_i(s)- \\ &\quad -\exp\{\beta^T X_i(s)\}\xi_i(s)\exp\{\gamma^T Z_i(s)\}\,d\mathcal{A}(s;\alpha,\Lambda)\big\} \end{aligned} \tag{3·5}$$

with $\mathcal{A}(\cdot)$ defined as

$$\mathcal{A}(t;\alpha,\Lambda) = \int_0^t \exp\{\alpha(s)\}\,d\Lambda(s). \tag{3·6}$$

The $dM_i(t;\mathcal{A}_0,\beta_0,\gamma_0,h)$ has a mean zero conditional on covariates $X_i(t)$ for any deterministic function $h(\cdot)$. The fundamental set of estimating equations is

$$\sum_{i=1}^n M_i(t;\mathcal{A},\beta,\gamma,h) = 0\ \ \forall t\in[0,\tau] \tag{3·7}$$

$$\sum_{i=1}^n \int_0^\tau W(t)X_i(t)dM_i(t;\mathcal{A},\beta,\gamma,h) = 0, \tag{3·8}$$

where $W(\cdot)$ is a weight process over time. We notice that equation (3·7) actually is an infinite-dimensional set of equations as those are defined for all times $t\in[0,\tau]$. Solving

equation (3·7) for any $t \in [0, \tau]$ yields

$$\hat{\mathcal{A}}(t) = \sum_{i=1}^{n} \int_0^t \frac{Y_i(s) \frac{1}{\rho_i(s;\gamma,h)} dN_i(s)}{\sum_{j=1}^{n} \frac{\xi_j(s)}{\rho_j(s;\gamma,h)} \exp\{\gamma_0^T Z_j(s)\} \exp\{\beta^T X_j(s)\}}.$$

Plugging $\hat{\mathcal{A}}(\cdot)$ into equation (3·8), the estimating function of an estimator of β_0 is

$$U(\beta;\gamma,h) \;=\; \sum_{i=1}^{n} \int_0^\tau W(t) \left[X_i(t) - Av_1(X)(t;\beta,h)\right] Y_i(t) \frac{1}{\rho_i(t;\gamma,h)} dN_i(t), \quad (3{\cdot}9)$$

where the mean curve $Av_1(\cdot)$ is as defined in equation (2·3). We can plug in any deterministic function of time $\zeta(t)$ in the way shown below without changing the mean value of the estimating function (3·9) at the true points β_0, γ_0. So the enriched estimating function has the form

$$\begin{aligned} U(\beta;\gamma,h) \;=\; & \sum_{i=1}^{n} \int_0^\tau W(t) \left[X_i(t) - Av_1(X)(t;\beta,h)\right] \times \\ & \times \left[Y_i(t) - \zeta(t) \exp\{\beta^T X_i(t)\}\right] \frac{1}{\rho_i(t;\gamma,h)} \, dN_i(t). \end{aligned}$$

Function $\zeta(t)$ is an arbitrary deterministic function of time. We gain precision if $\zeta(t) = \exp\{\alpha_0\}$. Thus, $\zeta(t)$ can be estimated as

$$\hat{\zeta}(t) = \sum_{i=1}^{n} \frac{Y_i(t)}{\exp\{\beta^T X_i(t)\}} \frac{\xi_i(t) h(X_i(t)) \exp\{\beta^T X_i(t)\}}{\sum_{j=1}^{n} \xi_j(t) h(X_j(t)) \exp\{\beta^T X_j(t)\}}.$$

Further we approximate the unknown outcome $Y_j(t)$ for $t \in \mathbb{T} - \{T_{jk}, k = 1, \ldots, K_j\}$. The approximated response is denoted by $Y^\star$, being for instance nearest neighbor or some more sophisticated approximation or smoothing method. We emphasize that for validity of the estimator of $\hat{\beta}(\hat{\gamma}, h)$ we do not need any good estimator of $\alpha_0(t)$ or any good approximation of $Y_j(t)$ at $t \in \mathbb{T} - \mathcal{T}_j$. Bad estimators of those can worsen precision only, they do not affect the validity of the estimator $\hat{\beta}(\hat{\gamma}, h)$. The final estimating

function defining the estimator $\hat{\beta}(\hat{\gamma}, h)$ is

$$\begin{aligned} U(\beta; \hat{\gamma}, h) &= \sum_{i=1}^{n} \int_0^\tau W(t) \left[X_i(t) - Av_1(X)(t; \beta, h) \right] \times \\ &\quad \times \left[Y_i(t) - \exp\{\beta^T X_i(t)\} Av_1 \left(\frac{Y^\star}{\exp\{\beta^T X\}} \right) (t; \beta, h) \right] \frac{1}{\rho_i(t; \hat{\gamma}, h)} \, dN_i(t) \end{aligned} \tag{3.10}$$

Solution to $U(\beta; \hat{\gamma}, h) = 0$ is $\sqrt{n}$-consistent and asymptotically normal. First define

$$\begin{aligned} R_i(t; \mathcal{A}, \beta, \gamma, h) &= M_i(t; \mathcal{A}, \beta, \gamma, h) - \\ &\quad - \int_0^t \exp\{\beta^T X_i(s)\} Av_1 \left(\frac{Y^\star}{\exp\{\beta^T X\}} \right) (s; \beta, h) \, \frac{1}{\rho_i(s, \gamma, h)} d\mathcal{M}_i(s; \gamma, \Lambda) \end{aligned}$$

and

$$\begin{aligned} H &\equiv \lim_{n \to \infty} E \left[-\frac{1}{n} \frac{\partial U(\beta_0; \gamma, h)}{\partial \gamma} |_{\gamma_0} \right] = E \int_0^\tau w(t) \left[X_1(t) - Av_1(x)(t; \beta_0, h) \right] \times \\ &\quad \times \left[Y_1(t) - \exp\{\beta_0^T X_1(t)\} Av_1 \left(\frac{y^\star}{\exp\{\beta^T x\}} \right) (t; \beta_0, h) \right] Z_1(t) \frac{1}{\rho_1(t; \gamma_0, h)} \, dN_1(t). \end{aligned}$$

The asymptotic variance of $\sqrt{n}(\hat{\beta}(\hat{\gamma}, h) - \beta_0)$ is $D^{-1} V D^{-1}$. The covariance matrix V is defined as

$$\begin{aligned} V &\equiv \lim_{n \to \infty} \mathrm{Cov} \left[\frac{1}{\sqrt{n}} U(\beta_0; \hat{\gamma}, h) \right] \\ &= E \left[\int_0^\tau w(t) \left[X_1(t) - Av_1(x)(t; \beta_0, h) \right] dR_1(t; \mathcal{A}_0, \beta_0, \gamma_0, h) - \right. \\ &\quad \left. - H A^{-1} \int_0^\tau \left[Z_1(t) - Av_2(z)(t; \gamma_0) \right]^T d\mathcal{M}_1(t; \gamma_0, \Lambda_0) \right]^{\otimes 2} \end{aligned}$$

and the matrix D of derivatives as

$$\begin{aligned} D &\equiv \lim_{n \to \infty} E \left[-\frac{1}{n} \frac{\partial U(\beta; \gamma_0, h)}{\partial \beta} |_{\beta_0} \right] \\ &= E \int_0^\tau w(t) \left[Av_1 \left(x^{\otimes 2} \right) (t; \beta_0, h) - \left\{ Av_1(x) \right\}^{\otimes 2} (t; \beta_0, h) \right] Y_1(t) \frac{1}{\rho_1(t; \gamma_0, h)} \, dN_1(t). \end{aligned}$$

We note that in the covariance matrix V we account for estimation of γ_0. Matrix H and D can be consistently estimated by

$$\begin{aligned}\hat{H} &= \sum_{i=1}^{n}\int_0^\tau W(t)\left[X_i(t) - Av_1(X)(t;\hat{\beta},h)\right]\times \\ &\quad \times\left[Y_i(t) - \exp\{\hat{\beta}^T X_i(t)\}Av_1\left(\frac{Y^\star}{\exp\{\beta^T X\}}\right)(t;\hat{\beta},h)\right] Z_i(t)\frac{1}{\rho_i(t;\hat{\gamma},h)}\,dN_i(t) \\ \hat{D} &= \frac{1}{n}\sum_{i=1}^{n}\int_0^\tau W(t)\left[Av_1\left(X^{\otimes 2}\right)(t;\hat{\beta},h) - \{Av_1(X)\}^{\otimes 2}(t;\hat{\beta},h)\right] Y_i(t)\frac{1}{\rho_i(t;\hat{\gamma},h)}dN_i(t).\end{aligned}$$

A consistent estimator of V is

$$\begin{aligned}\hat{V} &= \frac{1}{n}\sum_{i=1}^{n}\left[\int_0^\tau W(t)\left[X_i(t) - Av_1(X)(t;\hat{\beta},h)\right]d\hat{M}_i(t) - \right. \\ &\quad \left. - \hat{H}_2\hat{A}^{-1}\int_0^\tau \left[Z_i(t) - Av_2(Z)(t;\hat{\gamma})\right]^T d\hat{\mathcal{M}}_i(t)\right]^{\otimes 2},\end{aligned}$$

where

$$d\hat{\mathcal{M}}_i(t) = N_i(t) - \int_0^t \xi_i(s)\exp\{\hat{\gamma}^T Z_i(s)\}\,d\hat{\Lambda}(s).$$

4 HUD–VASH study data analysis

In 1992, the US Department of Housing and Urban Development (HUD) and the US Department of of Veterans Affairs (VA) established the HUD–VA Supported Housing (HUD–VASH) program. Veterans were eligible if they were literally homeless at the time of outreach assessment, had been homeless for 1 month or longer, and had received a diagnosis of a major psychiatric disorder or an alcohol or drug abuse disorder. The 460 homeless veterans were randomly assigned to 1 of 3 intervention groups: HUD–VASH intervention consisting of case management and housing vouchers (182 individuals); case management (90 individuals); standard VA homeless services (188 individuals). Vouchers authorized payment of a standardized local fair–market rent less 30% of the individual beneficiary's income. The important question to be answered by the program is whether setting aside housing resources is either necessary or sufficient for facilitating

exit from homelessness in this population. The primary outcome was percentage of days homeless during the last 3 months. Auxiliary time–dependent variables collected during the study were income in the past three months and whether social security or VA benefits were received during the past three months. Follow–up interviews were scheduled for every 3 months. However, subjects often missed assessment and came between scheduled interviews. Concern is raised that there as an association between the follow-up process and the outcome process. For detailed study description see Rosenheck et al. (2003).

In the analysis of the data, we set τ to 48 months and $C_i = \tau$ for all individuals $i \in \{1, \ldots, 460\}$. The 460 individuals made a total of 2855 follow–up visits by 48 months since randomization. The HUD–VASH intervention group has the highest level of follow–up visits and the standard care group the lowest level of visiting. Figure 1 shows the primary outcome of percentage homeless during the last 3 months specific for each treatment group. The time discretization is based on 6 months intervals. A crude view at the data suggests that the HUD–VASH intervention is more effective in reducing homelessness that the other two interventions that appear comparable.

To answer the question of efficacy of intervention we model the percentage days homeless during the last three months, denoted as PH, as a function of treatment assignment. We consider a semiparametric log–link model

$$\begin{aligned} \log E\left[PH_i(t)|Trt_i\right] &= \alpha_0(t) + \beta_{01} I(Trt_i = \text{HUD–VASH}) \\ &+ \beta_{02} I(Trt_i = \text{case management}), \end{aligned} \tag{4·1}$$

where the estimation of the parameters β_{01} and β_{02} is of primary scientific interest. The intercept $\alpha_0(t)$ is a nuisance parameter.

The covariates of the observation-times model (4·2) were suggested by the primary investigator. The time–invariant predictors of timing of visits are intervention assignment (HUD-VASH arm estimate 0.359, case management arm estimate 0.217, compared to the standard VA care arm), income at baseline (in thousands of dollars, denoted as IB,

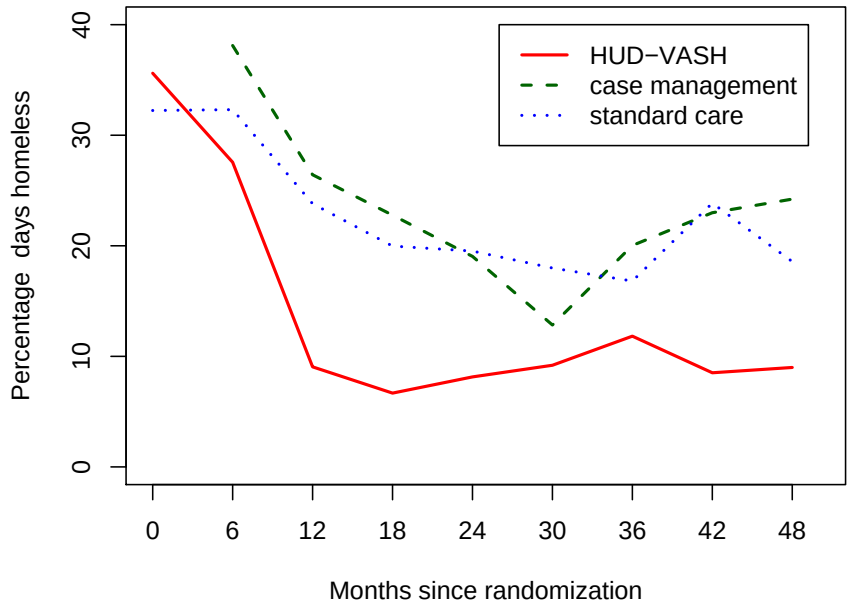

Figure 1: HUD–VASH: averaged outcome in each treatment arm.

estimate of -0.172), an indicator of receiving any social security or VA benefits at baseline (BB, estimated as 0.104) and the quality of life at baseline (denoted by QLB, estimate of -0.007). Time time–varying predictors for the observation-times model are percentage homeless approximated by previous value carried forward (estimated by 0.001), denoted by $PH^{\star}$, and cumulative number of visits so far, denoted by $N\_$, stratified by arm (estimated to be 0.044 for the standard VA care, additional -0.018 for the HUD-VASH arm and -0.014 for the case management arm). Parameter estimates suggest that higher intensity of visiting is associated with lower baseline income, receiving any social or VA benefits at baseline, having higher approximated percentage days homeless and higher cumulative number of visits so far. At any time, individual in the HUD–VASH intervention arm is more likely to have a visit than an individual under only case management, comparing two individuals on the same level of baseline income, indicator of social or

VA benefits at baseline, approximated percentage days homeless and having the same number of visits so far. Similarly, individual under case management is more likely to have a visit than an individual on standard care, comparing two individuals on the same level of baseline income, indicator of social or VA benefits at baseline, approximated percentage days homeless and having the same number of visits so far.

$$E\left[dN_i^\star(t)|Trt_i, IB_i, BB_i, PH_i^\star(t), QLB_i, N_{-i}(t)\right] = \exp\left\{\gamma_{01} I(Trt_i = \text{HUD–VASH}) + \right.$$
$$+ \gamma_{02} I(Trt_i = \text{case management}) + \gamma_{03} IB_i + \gamma_{04} BB_i + \gamma_{05} PH_i^\star(t) + \gamma_{06} QLB_i +$$
$$\left. + \gamma_{07} N_{-i}(t) + \gamma_{08} N_-^{\text{HUD–VASH}}{}_i(t) + \gamma_{09} N_-^{\text{case management}}{}_i(t) \right\} d\Lambda_0(t) \qquad (4{\cdot}2)$$

The semiparametric log–link model (4·1) suggests that at any time the ratio of the expected percentage days homeless within the last 3 months is 0.491 for the HUD–VASH treatment arm compared to the standard VA care arm. The 95% confidence interval is $(0.351, 0.686)$. Though $\hat{\beta}_{02}$ suggests increase of proportion days homeless comparing the case management group to the standard VA care, we did not have enough power to find evidence that the case management treatment resulted differentially than the standard VA care on the percentage days homeless on 5% statistical significance level. The 95% confidence interval for the ratio of mean percentage of days homelss, comparing the case management group to the standard VA care group, is $(0.725, 1.587)$, point estimate 1.073. Table 1 provides the characteristics of the raw estimates of β_{01} and β_{02}.

Table 1: HUD–VASH: estimates of primary parameter of interest (β_{01}, β_{02}) for model (4·1). The point estimate, its standard error, the standardized estimate and 95% confidence intervals are provided.

	$\hat{\beta}$	SE($\hat{\beta}$)	Z-statistic	95% CI
HUD–VASH	-0.712	0.171	-4.164	(-1.047, -0.377)
case management	0.070	0.200	0.350	(-0.322, 0.462)

For comparison and actual evidence for existence of biased sampling we computed the naive estimates, when not adjusting for biased sampling. We assume a log–link

parametric model

$$\begin{aligned} \log E\left[PH_i(t)|Trt_i\right] &= \boldsymbol{\beta_{00}} f(t) + \beta_{01} I(Trt_i = \text{HUD–VASH}) \\ &+ \beta_{02} I(Trt_i = \text{case management}), \end{aligned} \tag{4·3}$$

where $f(t)$ is as before a natural cubic spline with 4 degrees of freedom. We compute the estimates of the parameters of model (4·3) using a GEE with log–link and independent working correlation matrix. The naive parameter estimates, shown in Table 2, suggest qualitatively the same answer. However, we see a decrease in favoring the HUD–VASH treatment (point estimate of the ratio is 0.537, 95% confidence interval (0.355, 0.813)) and also increase of disliking the case management care (point estimate 1.097, 95% confidence interval (0.604 , 1.989)). Fitting the observation-times model (4·2) we learned that individuals who were worse off, which is those with more homelessness, lower baseline income and receiving baseline benefits, tended to have increased intensity of visiting, resulting in an upward bias. This biasness is different for the treatment arms, as suggested by fitting the observation-times model (4·2).

Table 2: HUD–VASH: naive GEE estimates of primary parameter of interest (β_{01}, β_{02}) for model (4·3). The point estimate, its standard error, the standardized estimate and 95% confidence intervals are provided.

	$\hat{\beta}$	SE($\hat{\beta}$)	Z-statistic	95% CI
HUD–VASH	-0.621	0.211	-2.943	(-1.035, -0.207)
case management	0.092	0.304	0.303	(-0.504, 0.688)

5 Simulations

Assume a random effect semiparametric Poisson-Normal model for response Y with covariate X_1, that is model for each $i \in \{1, \ldots, n\}$ and time $t \in [0, \tau]$

$$E\left[Y_i(t)|X_{1i}(t), \phi_i\right] = \exp\{\alpha_0(t) + \beta_{01} X_{1i}(t) + \phi_i\}. \tag{5·1}$$

Random effect ϕ_i, $\phi_i \sim N(0, \sigma_\phi^2)$ is used to introduce autocorrelation. As ϕ_i is fixed for a person, means and thus responses on the same subject are correlated (positively) in time when $\sigma_\phi > 0$. Model (5·1) can be obtained by marginalization from a model

$$E\left[Y_i(t)|X_{1i}(t), Z_{2i}(t), \phi_i\right] = \frac{\exp\{\alpha_0(t) + \beta_{01}X_{1i}(t) + \beta_{02}Z_{2i}(t) + \phi_i\}}{E\left[\exp\{\beta_{02}Z_{2i}(t)\}|X_{1i}(t)\right]}, \tag{5·2}$$

where the denominator is included in order to avoid confounding of the mean–response model by the covariate Z_2. The marginal mean model is

$$\mu_i(t) = E\left[Y_i(t)|X_{1i}(t)\right] = \exp\{\alpha_0(t) + \beta_{01}X_{1i}(t) + \sigma_\phi^2/2\}. \tag{5·3}$$

Marginally this count response distribution has no closed form. The following functions were considered as the baseline predictor $\alpha_0(t)$: first $0.1\sqrt{t}$, second $0.1\sin(t)$, third $0.1\exp\left\{range\,|\sin(t)|\right\}$, fourth $0.1\sin(peak\ t)$ and fifth $0.1\exp\{range\ |\sin(peak\ t)|\}$. Parameter *range* controls the extreme size of the intercept values and was set to 2. The peakedness parameter *peak* was set to 3. See Figure 2 illustrating the five functions considered with the specific parameters. The nonlinear trend and sine wave were considered in Lin & Ying (2001) in simplified models. We do not assume specification of the intercept function, therefore these various cases are used to demonstrate the estimator performance under a range of various scenarios of the baseline predictor.

There are two covariates in the sampling times model. First one represents treatment and is changing over time randomly for each individual, $X_1 \sim Bernoulli(0.5)$. This covariate is the mean model covariate as well. Second covariate X_2 is dependent upon the first covariate. If a person is not at certain time t on treatment ($X_1(t) = 0$), then $X_2(t)$ is normally distributed with mean and variance four. If a person is at certain time on a treatment ($X_1(t) = 1$), then $X_2(t)$ is normally distributed with mean two and variance one. The intention in these settings is to model an effect of treatment to reduce values of the second covariate. Parameters of interest β_{01}, β_{02} were set to 0.5 and -1, respectively. Discretization of continuous time is based on a grid of 100 per a time unit.

Table 3: Quartiles of number of observations per individual.

	min	25	50	75	max
$\tau = 2$	1	1	2	3	10
$\tau = 8$	1	5	8	9	25

Censoring variable C is Uniform between $\tau/2$ and τ, with τ picked 2 or 8, representing the median number of observations per individual. These settings should demonstrate cases of a few and many observations per person. See table (3) for the distribution of number of events per person.

The observation times follow a random-effect Poisson counting process with intensity $\lambda_i(t) = \eta_i \exp\{\gamma_{01} X_{1i}(t) + \gamma_{02} X_{2i}(t)\}$.Random effect $\eta_i \sim \Gamma(\mu_\eta = 1, \sigma^2_\eta = 0.01)$. We set γ_1 is set to -0.2, γ_2 to 0.3. Thus a person on a treatment is less likely to be sampled. The parameter of putting weight to time is set to 1 for all time.

Along with our proposed estimate of β_0, independent GEE estimates are computed as well for comparison of the new estimator with this widely used one. We are assuming independent covariance structure to avoid eventual bias of the GEE estimator due to modeling not the marginal mean $E[Y_i(t)|X_i(t)]$ but $E[Y_i(t)|X_i(s), s \in \{T_{i1}, \ldots, T_{iK_i}\}]$ instead. To accommodate the intercept and the additional term $\exp\{\sigma^2_\phi\}$, we include as an offset into GEE. For details on fitting a marginal model to mixed effects log linear regression data via GEE see Grömping (1996).

Table 4: Statistics for estimator of β_{01} in the semiparametric log–link model (5·3) under biased sampling for sample size 50.

	Bias	SSE	SEE	CP	M I	M II
$\alpha = 1, \tau = 2$	−0.034	0.377	0.333	0.91	1.28	1.61
$\alpha = 2, \tau = 2$	−0.029	0.375	0.339	0.90	1.22	1.50
$\alpha = 3, \tau = 2$	−0.029	0.355	0.325	0.91	1.24	1.82
$\alpha = 4, \tau = 2$	−0.028	0.381	0.343	0.91	1.17	1.53
$\alpha = 5, \tau = 2$	−0.042	0.350	0.321	0.92	1.39	1.87
$\alpha = 1, \tau = 8$	−0.012	0.203	0.208	0.95	2.74	4.68
$\alpha = 2, \tau = 8$	−0.008	0.220	0.211	0.93	2.36	3.68
$\alpha = 3, \tau = 8$	−0.023	0.206	0.202	0.94	2.70	4.54
$\alpha = 4, \tau = 8$	−0.017	0.211	0.210	0.94	2.59	4.35
$\alpha = 5, \tau = 8$	−0.021	0.193	0.203	0.95	3.06	5.23

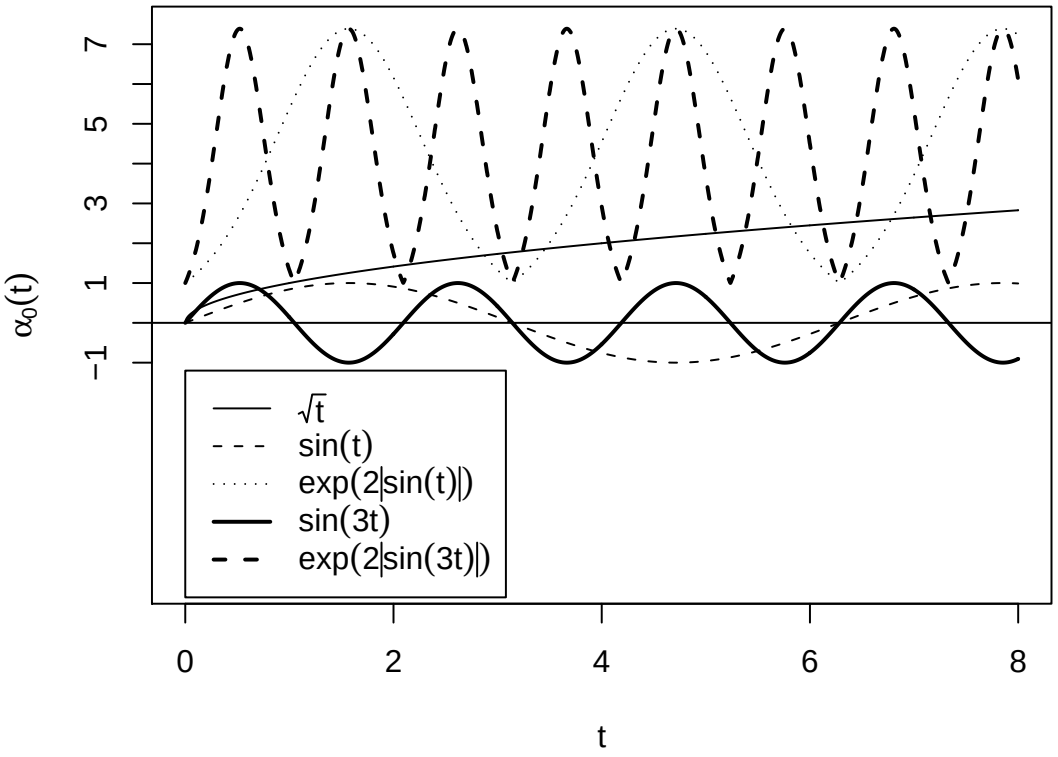

Figure 2: Intercept functional forms.

We present bias, sampling standard error of $\hat{\beta}_0$ and sampling mean of estimated standard errors taken over 1000 simulations. We also present two comparison measures of squared errors. Measure M I is based on mean of the ratio of empirical mean squared error of the new estimate of β_0 over empirical mean squared error of GEE estimate of β_0. Measure M II is based on empirical median of ratios of squared errors, being a more robust efficiency estimate motivated by Pitman closeness. We report 95% sampling coverage probability with precision of 1.4%. Number of individuals in a sample is set to 50 and 200. Results for 20 and 100 individuals per sample are not shown but are

Table 5: Statistics for estimator of β_{01} in the semiparametric log–link model (5·3) under biased sampling for sample size 200.

	Bias	SSE	SEE	CP	M I	M II
$\alpha = 1, \tau = 2$	−0.016	0.176	0.194	0.97	3.55	5.68
$\alpha = 2, \tau = 2$	−0.021	0.186	0.195	0.96	3.21	5.11
$\alpha = 3, \tau = 2$	−0.007	0.136	0.145	0.94	3.40	5.18
$\alpha = 4, \tau = 2$	0.010	0.183	0.195	0.96	3.16	5.98
$\alpha = 5, \tau = 2$	−0.011	0.185	0.189	0.95	3.18	5.41
$\alpha = 1, \tau = 8$	−0.016	0.121	0.128	0.96	8.67	16.05
$\alpha = 2, \tau = 8$	−0.005	0.109	0.116	0.97	8.16	13.82
$\alpha = 3, \tau = 8$	−0.013	0.103	0.112	0.97	8.99	16.03
$\alpha = 4, \tau = 8$	−0.001	0.105	0.117	0.98	8.76	14.09
$\alpha = 5, \tau = 8$	−0.005	0.097	0.113	0.99	9.95	17.90

consistent with those shown here.

Tables 4 and 5 provide summaries for the estimator with sample sizes of 50 and 200 after excluding 0.5% of simulations providing the most outlying estimates. For both types of proposed estimators the bias estimate is always (for all considered scenarios of various τ, n, $\alpha_0(\cdot)$,) negligible relative to the sampling standard error (SSE). Comparing SSE and SEE, the model based variance of the estimator of coefficients β_0 (SEE) is usually slightly underestimating the true variance of the estimator of β (SSE). With larger n this discrepancy is decreasing. Both types of estimators have large outliers compared to the GEE estimator. Originally, the mean squared error comparison measure M I favored the biased GEE to our approaches, due to the outliers in both our methods under both sample sizes. The measure M II, that is robust to outliers, however favored under both sample sizes largely our new approaches. The magnitude of that favor increases with sample size. If we exclude 0.5% of the simulations with the largest outliers, even the measure M I will mostly favor our approaches. The M I bigger than one means that the mean squared error of our new estimator is smaller than the mean squred error of the GEE estimator. So, the variance induced by unspecified intercept of our estimator is ovecome by the bias of the GEE estimator. Coverage probability (CP) is improving with increasing sample size as well as larger number of observations per person. It is not showing any trend with respect to the choice of intercept function.

6 Discussion

6·1 *Required Information*

In survival analysis settings, assuming knowledge of the covariate process $\{Z(t),\ t \in [0,\tau]\}$ at all times up to dropout is a standard assumption. On the contrary, in longitudinal analysis settings, the usual assumption is that covariate process $\{X(t),\ t \in [0,\tau]\}$ is known at observation times only. Our estimator, combining survival analysis and longitudinal analysis approaches together, requires knowledge of both covariate processes continuously at all times up to dropout. Approximations of the covariate process cause our estimators to be biased. However, subcohort sampling techniques enable our estimators to stay consistent after paying certain precession price.

6·2 *Terminology*

In discrete times models, where observation times come from a finite set of points, biased sampling can be viewed as a missingness problem. We base our stratification of missingness pattern on the typical one introduced by Rubin (1976). We can talk there about biased sampling being equivalent to missingness at random given covariates X and Z. It is informative missingness given covariates X only. Other terms that are being used in that situation are informative inter-mitten missingness or informative follow-up. However, in continuous observation-times settings we do not want to talk about how it relates to the missingness classification. Here the data are missing with probability of 100% as the data are observed at discrete time points, not continuously over time as curves.

6·3 *Another estimation approach*

Motivated by the linear regression approach, we developed an additional class of estimators suitable for the semiparametric loglinear model (2·1). We define a weighted average

$\tilde{A}v_1$ that has weights proportional to a function $h(t)$.

$$\tilde{A}v_1(V)(t;h) = \sum_{i=1}^{n} V_i(t) \frac{\xi_i(t)h(X_i(t))}{\sum_{j=1}^{n} \xi_j(t)h(X_j(t))} \quad (6{\cdot}1)$$

The construction of the estimator is based on a process

$$\tilde{M}_i(t;\mathcal{A},\beta,\gamma,h) = \int_0^t \frac{1}{\rho_i(s;\gamma,h)} \left\{ \frac{Y_i(s)}{\exp\{\beta^T X_i(s)\}} dN_i(s) - \right. \\ \left. -\xi_i(s) \exp\{\gamma^T Z_i(s)\}\, d\mathcal{A}(s;\alpha,\Lambda) \right\}. \quad (6{\cdot}2)$$

Following the steps of derivation of the estimating equation (3·10) we derive the estimating function

$$\tilde{U}(\beta;\hat{\gamma},h) = \sum_{i=1}^{n} \int_0^\tau W(t) \left[X_i(t) - \tilde{A}v_1(X)(t;h) \right] \times \\ \times \left[\frac{Y_i(t)}{\exp\{\beta^T X_i(t)\}} - \tilde{A}v_1 \left(\frac{Y^\star}{\exp\{\beta^T X\}} \right)(t;h) \right] \frac{1}{\rho_i(t;\hat{\gamma},h)}\, dN_i(t) \quad (6{\cdot}3)$$

Solution to $\tilde{U}(\beta;\hat{\gamma},h) = 0$ is $\sqrt{n}$-consistent and asymptotically normal random vector. Throughout all our simulations we have however discovered an inferior behavior of this "tilde" class of estimators compared to the proposed one.

Table 6: Relative efficiencies for the two types of estimators for log–link models.

	RE I	RE II	RE I	RE II
	sample size 50		sample size 200	
$\alpha=1,\tau=2$	1.360	0.947	4.464	0.846
$\alpha=2,\tau=2$	1.543	0.926	4.434	0.803
$\alpha=3,\tau=2$	1.432	0.901	6.655	0.831
$\alpha=4,\tau=2$	1.361	0.920	7.646	0.829
$\alpha=5,\tau=2$	1.749	0.884	5.134	0.853
$\alpha=1,\tau=8$	2.058	0.887	12.885	0.911
$\alpha=2,\tau=8$	2.437	0.875	18.445	0.855
$\alpha=3,\tau=8$	2.562	0.906	20.825	0.851
$\alpha=4,\tau=8$	3.240	0.919	14.263	0.856
$\alpha=5,\tau=8$	2.953	0.910	15.732	0.782

Figure 3 generated from simulation with sample size 200, α one and τ two shows

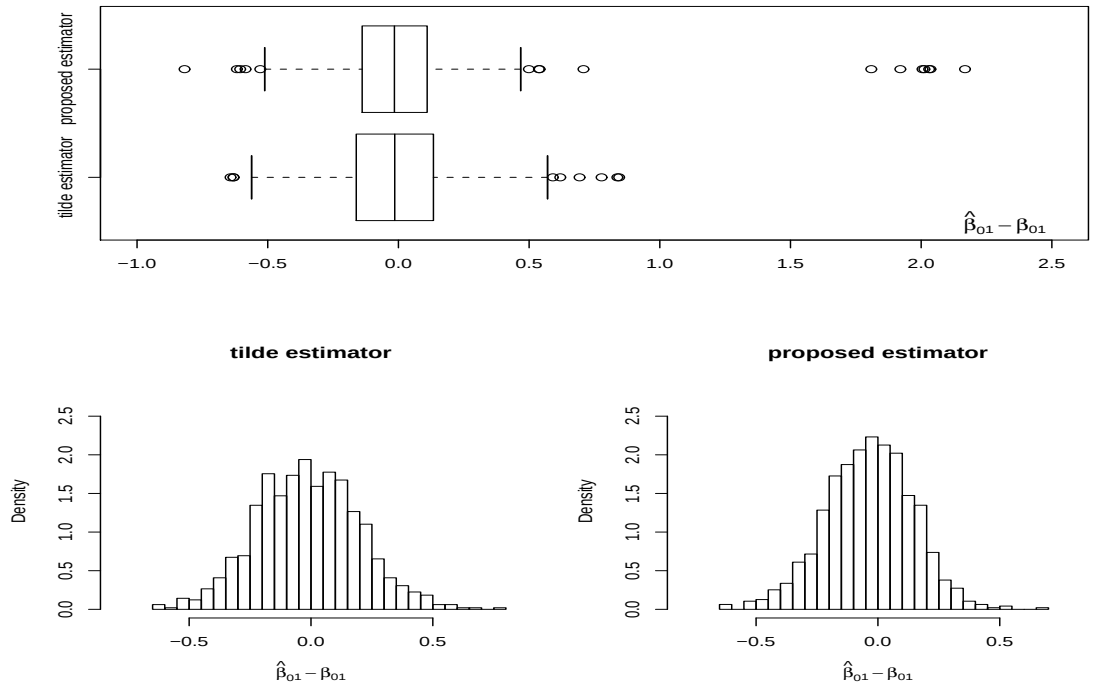

Figure 3: Comparison of $\hat{\beta}_{01} - \beta_{01}$ under the two approaches for log–link models. Box plot is shown in the upper part and histograms of a region $(-0.8, 0.8)$ in the lower part.

that the variability of the proposed estimates is generally smaller than the variability of the tilde estimates. However, the proposed estimates relatively rarely (0.5%) reach very extreme values. All simulations that we considered showed consistent pattern to the one shown in Figure 3. The squared error comparison measures M become measures of relative efficiency. Relative efficiencies type I and II differ, see Table 6. Relative efficiency type I largely favors the tilde estimator, due to the rare but very extreme values. Relative efficiency of type II, which is a more robust measure, favors the mainly

proposed estimator.

Implementing the tilde estimation in the HUD–VASH data analysis, our findings are consistent with simulations finding. Estimates of both parameters of interest have larger standard errors under the tilde approach. The point estimate of $\exp\{\beta_{01}\}$ is 0.48, with 95% CI (0.29, 0.79). The point estimate of $\exp\{\beta_{02}\}$ is 1.07, with 95% CI (0.52, 2.18).

6·4 *Independent observation times*

We would also like to point out that the "Independent observation times" version of our estimator, similar to Lin & Ying (2001), exists as well and we feel that the contribution to log linear models allowing for the unspecified intercept alone is worthy noting, regardless of biased or unbiased sampling. We think that the process-like centering of both the response and covariates is a unique tool for an elegant solution to the intercept issue.

Acknowledgment

The HUD–VASH study data were obtained from H. Lin at Yale University, New Haven with permission from the study primary investigator R. Rosenheck at Veterans Affairs Northeast Program Evaluation Center, West Haven.

Appendix

A *Large Sample Theory*

We base our derivations on asymptotic theory established in Lin & Ying (2001) using monotone functions and "manageable processes" tools. Estimating function U defined in (3·10) at the true values β_0, γ_0 can be rewritten as

$$U(\beta_0; \gamma_0, h) \;=\; \sum_{i=1}^{n} \int_0^{\tau} W(t)\,[X_i(t) - Av_2(X)(t; \beta, h)]\, dR_i(t; \mathcal{A}_0, \beta_0, \gamma_0, h).$$

By arguments similar to those in Appendix of Lin & Ying (2001), $\frac{1}{\sqrt{n}}U(\beta_0;\gamma_0,h)$ is asymptotically equivalent to

$$\frac{1}{\sqrt{n}}\int_0^\tau w(t)\left[\sum_{i=1}^n X_i(t)dM_i(t;\mathcal{A}_0,\beta_0,\gamma_0,h) - Av_1(x)(t;\beta_0,h)\sum_{i=1}^n dM_i(t;\mathcal{A}_0,\beta_0,\gamma_0,h)-\right.$$
$$-Av_1\left(\frac{y^\star}{\exp\{\beta_0^T x\}}\right)(t;\beta_0,h)\ \sum_{i=1}^n X_i(t)\exp\{\beta_0^T X_i(t)\}\frac{1}{\rho_i(t;\gamma_0,h)}d\mathcal{M}_i(t;\gamma_0,\Lambda_0)+$$
$$\left.+Av_1(x)(t;\beta_0,h)Av_1\left(\frac{y^\star}{\exp\{\beta_0^T x\}}\right)(t;\beta_0,h)\ \sum_{i=1}^n \exp\{\beta_0^T X_i(t)\}\frac{1}{\rho_i(t;\gamma_0,h)}d\mathcal{M}_i(t;\gamma_0,\Lambda_0)\right].$$

After applying Taylor's expansion several times, we obtain

$$\frac{1}{\sqrt{n}}U(\beta_0;\hat\gamma,h) = \frac{1}{\sqrt{n}}U(\beta_0;\gamma_0,h) - \frac{1}{n}\frac{\partial U(\beta_0;\gamma,h)}{\partial\gamma}\mid_{\gamma^\circ}\left(\frac{1}{n}\frac{\partial U^\dagger(\gamma)}{\partial\gamma}\mid_{\gamma^\star}\right)^{-1}\frac{1}{\sqrt{n}}U^\dagger(\gamma_0) \quad \text{(A1)}$$

with γ° and $\gamma^\star$ being on the line segment between γ_0 and $\hat\gamma$.

Equation (A1) is asymptotically equivalent to

$$\frac{1}{\sqrt{n}}\sum_{i=1}^n\left[\int_0^\tau w(t)\left[X_i(t)-Av_1(x)(t;\beta_0,h)\right]\left[dM_i(t;\mathcal{A}_0,\beta_0,\gamma_0,h)-\right.\right.$$
$$\left.-Av_1\left(\frac{y^\star}{\exp\{\beta_0^T x\}}\right)(t;\beta_0,h)\ \exp\{\beta_0^T X_i(t)\}\frac{1}{\rho_i(t;\gamma_0,h)}d\mathcal{M}_i(t;\gamma_0,\Lambda_0)\right]-$$
$$\left.-HA^{-1}\int_0^\tau\left[Z_i(t)-Av_2(z)(t;\gamma_0)\right]^T d\mathcal{M}_i(t;\gamma_0,\Lambda_0)\right]$$

which is a sum of n independent identically distributed mean zero random vectors.

So $\sqrt{n}(\hat\beta(\hat\gamma,h)-\beta_0)$ can be expressed as

$$\begin{aligned}\sqrt{n}(\hat\beta(\hat\gamma,h)-\beta_0) &= \left(-\frac{1}{n}\frac{\partial U(\beta;\hat\gamma,h)}{\partial\beta}\mid_{\beta^\star}\right)^{-1}\frac{1}{\sqrt{n}}U(\beta_0;\hat\gamma,h)\\ &= \left(-\frac{1}{n}\frac{\partial U(\beta;\hat\gamma,h)}{\partial\beta}\mid_{\beta^\star}\right)^{-1}\ \times\ \left[\frac{1}{\sqrt{n}}U(\beta_0;\gamma_0,h)-\right.\\ &\quad \left.-\frac{1}{n}\frac{\partial U(\beta_0;\gamma,h)}{\partial\gamma}\mid_{\gamma^\circ}\left(\frac{1}{n}\frac{\partial U^\dagger(\gamma)}{\partial\gamma}\mid_{\gamma^\star}\right)^{-1}\frac{1}{\sqrt{n}}U^\dagger(\gamma_0)\right] \quad \text{(A2)}\end{aligned}$$

and thus $\sqrt{n}(\hat{\beta}(\hat{\gamma},h)-\beta_0)$ is asymptotically equivalent to

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n} D^{-1}\left[\int_0^\tau w(t)\left[X_i(t)-Av_1(x)(t;\beta_0,h)\right]\right.$$
$$\times\left[dM_i(t;\mathcal{A}_0,\beta_0,\gamma_0,h)-\rho_i(t;\gamma_0,h)Av_1\left(\frac{y^\star}{\exp\{\beta_0^T x\}}\right)(t;\beta_0,h)\,d\mathcal{M}_i(t;\gamma_0,\Lambda_0)\right]-$$
$$\left.-HA^{-1}\int_0^\tau\left[Z_i(t)-Av_2(z)(t;\gamma_0)\right]^T d\mathcal{M}_i(t;\gamma_0,\Lambda_0)\right]$$

which is a sum of mean zero i.i.d. random vectors.

This plus consistency of $\hat{\beta}(\hat{\gamma},\hat{\delta})$ and of $\hat{D}$ when the estimator of δ_0 is used yields that $n^{\frac{1}{2}}\left(\hat{\beta}(\hat{\gamma},\hat{\delta})-\beta_0\right)$ is asymptotically normal with a consistent estimate of the asymptotic variance being $\hat{D}^{-1}\hat{V}\hat{D}^{-1}$.

B *Assumptions*

We assume that $(Y_i(\cdot),X_i(\cdot),Z_i(\cdot),N_i^*(\cdot),\xi_i(\cdot))$ are i.i.d. quintuples of random processes over time t for individuals 1 through n. The counting uncensored process of events at the end of follow-up τ, $N_i(\tau)$, is required to be bounded by a constant. Both mean response model covariates X_i and observation-times model covariates Z_i need to have bounded total variations by a constant for all individuals $i=1,\ldots,n$. That is $|Z_{ji}(0)|+\int_0^\tau|dZ_{ji}(t)|\le K,\ j=1,\ldots,g$ and $|X_{ji}(0)|+\int_0^\tau|dX_{ji}(t)|\le K,\ j=1,\ldots,p$. Total number of observations per individual i, denoted by K_i, is bounded. The weight function $W(\cdot)$ is a difference of two monotone functions, each of which converges to a deterministic function. We denote the limit of $W(\cdot)$ by $w(\cdot)$. We assume that the function $h(\cdot)$ has bounded variation.

C *Implementation of the estimation procedure*

The estimating procedures mentioned in this paper can be implemented in S-plus/R with relative ease. The observation-times model (2·2) can be fitted by function *coxph* in package *survival* in order to obtain the estimate of γ_0. In the log–link mean–response

models (2·1), we need to solve the estimating equations as specified in equations (6·3) and (3·10). We can solve those by the *optim* function from package *stats*. Because of biased sampling we also need to re-weight the response by the inverse probability weights ρ. The standard errors of the estimate of β_0 in all cases can be obtained by bootstrapping or implementing the sandwich estimates provided. We plan on adding an implicit function into R that would conveniently provide the estimates and their characteristics.

References

Bůžková, P. & Lumley, T. (2005a). Longitudinal data analysis for generalized linear models under irregular, biased sampling: Situations with follow-up dependent on outcome or auxiliary variables. *Journal of the Royal Statistical Society, B Series* (submitted).

Bůžková, P. & Lumley, T. (2005b). Marginal regression modeling under irregular, biased sampling. *Journal of the American Statistical Association* (submitted).

Grömping, U. (1996). A note on fitting a marginal model to mixed effects log-linear regression data via gee. *Biometrics* **52**, 280–285.

Hernán, M. A., Brumback, B. A., & Robins, J. M. (2002). Estimating the causal effect of zidovudine on CD4 count with a marginal structural model for repeated measures. *Statistics in Medicine* **21**, 1689–1709.

Lin, D. Y., Wei, L. J., Yang, I., & Ying, Z. (2000). Semiparametric regression for the mean and rate functions of recurrent events. *Journal of the Royal Statistical Society, B Series* **62**, 711–730.

Lin, D. Y. & Ying, Z. (2001). Semiparametric and nonparametric regression analysis of longitudinal data. *Journal of the American Statistical Association* **96**, 103–126.

Lin, H., Scharfstein, D. O., & Rosenheck, R. A. (2004). Analysis of longitudinal data with irregular, outcome-dependent follow-up. *Journal of the Royal Statistical Society, Series B* **66**, 791–813.

Lin, X. & Carroll, R. J. (2001a). Semiparametric regression for clustered data. *Biometrika* **88**, 1179–1185.

Lin, X. & Carroll, R. J. (2001b). Semiparametric regression for clustered data using generalized estimating equations. *Journal of the American Statistical Association* **96**, 1045–1056.

Pan, W., Louis, T. A., & Connett, J. E. (2000). A note on marginal linear regression with correlated response data. *The American Statistician* **54**, 191–195.

Pepe, M. S. & Anderson, G. L. (1994). A cautionary note on inference for marginal regression models with longitudinal data and general correlated response data. *Communications in Statistics - Simulation* **23**, 939–951.

Rosenheck, R., Kasprow, W., Frisman, L., & Liu-Mares, W. (2003). Cost–effectiveness of supported housing for homeless persons with mental illness. *Archives of General Psychiatry* **60**, 940–951.

Rubin, D. B. (1976). Inference and missing data. *Biometrika* **63**, 581–592.