9-26-2005

Marginal Regression Modeling under Irregular, Biased Sampling

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

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

1 Introduction

In a longitudinal study, of focus is to examine the association of covariate process $\{X(t),\ t \in [0,\tau]\}$ and the response process $\{Y(t),\ t \in [0,\tau]\}$, where the predetermined constant τ is the end of study or time at which the last person to follow drops out. A very important feature of longitudinal data is the actual timing of observations. The actual timing can be divided into three categories: designed sampling times that are strictly followed; designed sampling times that are not strictly followed; and observational times. An example of the second timing category, where there is an extra noise to the pre–defined sampling times, is the health research study data we analyze in Section 6. It is a study on a population of homeless people, where the investigators experienced an extreme non–compliance to the scheduled visits. An example of the third category of timing are administrative data, where the actual sampling process is not part of the study. When data are collected under the second or third category, the irregular sampling times present an extra challenge for the statistical analysis of such data set. There, the investigator has no control over the sampling times and the related frequency of sampling the individuals. These irregular sampling times can be subject-specific. That can introduce a biased sampling design for the mean–response model, when factors that are not included as covariates influence sampling times. We demonstrate the philosophy of biased sampling with a simple example taken from air pollution. Assume a binary indicator, $Z(t)$, of an asthma attack at time t. Further assume an air pollution measure at time t as a covariate $X(t)$. Let the outcome 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. The lung function measure clearly is associated both with the air pollution measure and presence or absence of the 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 characterized by present asthma attacks form the majority of the observed data. Our overall philosophy behind choosing covariates is that the mean–response model covariates X should be picked on scientific ground purpose but the covariates Z for sampling–times model should cover the true nature of the sampling process. Then, modeling response with the air pollution covariate only we obtain unbiased estimates for those who have come for a visit, primarily people who suffer from an asthma attack at that time. However, we obtain an exaggerated estimate of the lung function measure and air pollution measure association for the general public.

Under discrete time, when the sampling times come from a finite set of points, the sampling can be viewed as a missingness problem. Taxonomy of missingness, as formalized by Rubin (1976), is based on factors that drive the seeing of observations. The key issue is whether the fact that data are missing is related to the values of the variables in the data set. For a complete survey of the current methodology of missing data we refer a keen reader to Little & Rubin (2002). In discrete time models we can view biased sampling as 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 of biased sampling under discrete time are informative intermittent missingness or informative follow-up. In continuous time, when the sampling times come from an interval such as $[0,\tau]$, we can not talk about missingness in the original meaning. Here the data are missing with probability 100% as the response is observed at discrete time points, not continuously over time as a curve. Nevertheless, we can look at inclusion of the data rather than missingness of the data, from a similar perspective.

Generalized estimating equations, introduced in Liang & Zeger (1986) and denoted by GEE, are a popular estimation method. It is well known that the standard inferential approach based on the GEE yields biased inference if the sampling times depend, additionally to dependence on the mean–response model covariates, on response or other covariates related with response. Moreover, as discussed by Pepe & Anderson (1994), when using other than working independence, additional marginalization assumption on the mean of response at certain time conditional on the covariates X at that time, $E[Y(t)|X(s),\ s \in [0,\tau]] = E[Y(t)|X(t)]$, is needed. This assumption can be hard to satisfy. History of covariates can be incorporated into the current covariates. But supposing that conditioning on future covariates on top of the current ones does not alter the mean–response model is a strong assumption. Imagine that a hypertensive drug is chosen based upon current blood pressure measurement. We are interested in the association of the chosen drug and next visit blood pressure measurement. We can assume that given the current drug the mean current blood pressure is independent of the drug history. It is very unlikely though that given the mean current drug the current blood pressure and the future drug assignment are independent. However, in modeling the association of treatment and blood pressure we still want to condition only on current treatment. Fully marginal regression models, where the marginalization assumption is not required, seem an attractive alternative when analyzing longitudinal data.

Some of the statistical literature on longitudinal data with continuous sampling times can deal with informative drop-out, also called right censoring, see Wu & Carroll (1988), Diggle & Kenward (1994) and Scharfstein et al. (1999). The procedures of Troxel et al. (1998), accommodating intermittent informative missingness, is not suitable for data that occur at non-predefined, irregular times.

The approach of Lipsitz et al. (2002) is likelihood–based and thus it accommodates biased sampling when the sampling times depend on the previous value of response. It does not accommodate biased sampling when there are additional covariates, associated with response, that govern the sampling process. They separate the likelihood function into two components: one for the response process and the other for the sampling–times process, where they compute the likelihood of the time elapsed between two sampling times. In order to ignore the later process, Lipsitz et al. impose a strong sampling–times model assumption of dependence only on history of observed response measurements. In their fully parametric approach they assume that the repeated measures of response have a multivariate Gaussian distribution. Moreover, the estimation procedure relies for consistency of the estimators of interest on correct specification of the autocorrelation structure of response. Lipsitz et al. point out that "the potential bias due to misspecification of the covariance can be considerable".

D. Lin & Ying (2001) integrate counting processes techniques with longitudinal data settings under continuous time. They assume a linear regression model

$$E[Y_i(t)|X_i(t)] = \alpha_0(t) + \beta_0^T X_i(t). \tag{1·1}$$

The parameter of interest, β_0, is a p–dimensional vector. The effect of time is modeled completely non-parametrically and the intercept curve $\alpha_0(\cdot)$ is an infinite-dimensional nuisance parameter. In spite of providing a very powerful framework for incorporating sampling–times patterns into estimation of a mean–response regression model parameters, D. Lin & Ying impose assumption about the response process and the sampling–times process relationship that does not enable biased sampling. That assumption is independence of the response process and the sampling–times process conditional on mean–response model covariates. In our example of studying an association between the lung function measure

and the air pollution we would either need to include adjusting for the asthma attack or assume independence of the sampling–times process and the lung function measure conditioning on the air pollution alone.

In the model (1·1) that we adopt 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 unknown 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. The semiparametric form of the mean–response model with unspecified intercept is used as well by X. Lin & Carroll (2001a) and X. Lin & Carroll (2001b). They denote T_{ij} the j–th sampling time of individual i and assume that $\theta_0(\cdot)$ is an unspecified function of time, but smooth. The mean–response model is

$$E[Y_i(T_{ij})|X_i(T_{ij}), T_{ij}] = g\left[\theta_0(T_{ij}) + \beta_0^T X_i(T_{ij})\right]. \tag{1·2}$$

Additionally, they assume an assumption similar to Pepe & Couper (1997)

$$E[Y_i(T_{ij})|X_i(T_{ij}), T_{ij}] = E[Y_i(T_{ij})|X_i(T_{ij}), T_{ij},\ (X_i(T_{ik}), T_{ik})_{\forall k \neq j}] \tag{1·3}$$

that can be limiting as we discussed previously. X. Lin and Carroll use profile–based estimating equation for estimation of the parameter of interest and kernel estimating equation for the nonparametric estimation. They note that for longitudinal data kernel smoothing does not involve band–width selection issues only. It is a 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. The approach of X. Lin and Carroll does not handle biased sampling.

We note that our way of incorporating time, similar to Lipsitz et al. (2002), D. Lin & Ying (2001) and H. Lin et al. (2004), is a functional or process–like approach. We model the conditional mean of response for a certain individual for any time t, unlike for instance X. Lin & Carroll (2001a), we do not condition on the set of the individual's observation times. We say that at any time t the conditional mean response exists and follows our given model or that it only exists and follows the given model at the observation times but if we had observed response at a different time it would have existed and followed the given model. Conditioning on the observation times can be appropriate when the response either does not exist at other than observation times or that it does not follow the given mean model at other than observation times. Now we give an example of such situation when conditioning on observation time is reasonable whereas our process–like approach is not. Assume that we have longitudinal data consisting of birth–weight of children born to a certain population of women. Interest is in the association of a child's birth weight and his/her mother's age. It is obvious that process-like modeling is inappropriate as birth can not happen at any time but the possibility of the event is dependent on when the last event occurred. Thus the response is not defined at times within 9 months of any previous observation time.

We start with notation introduction. Then we describe the estimation under the sampling–times model and followed with a brief summary of the D. Lin & Ying (2001) estimator. In Section 5 we are suggesting a class of estimators that account for the possibility

of biased sampling under continuous time in linear regression setting with unspecified intercept. Our sampling–times model covariates are not restricted in any way. For instance, the response at a previous sampling time can be included or an average of a covariate over a subject's history. In the class of estimators we are highlighting an estimator that has certain variance–stabilizing properties. The new estimators were shown to be $\sqrt{n}$-consistent with normal limiting distribution and simple asymptotic variance.

In Section 6 we illustrate our method on a health service research study, the HUD–VASH study. The data were obtained from H. Lin at Yale University, New Haven with permission from the study primary investigator Dr. R. Rosenheck at Veterans Affairs Northeast Program Evaluation Center, West Haven. Homeless people with mental illness were randomized to three different interventions. Percentage days homeless within the last three months as an outcome variable and a handful of covariates were recorded at follow–up times. Those times were fixed by design but not followed. Intervention efficacy is the scientific question of our interest. This data set was recently used in H. Lin et al. (2004). There, they developed a class of "inverse–intensity–of–visit process–weighted" estimators in marginal regression models for longitudinal responses that might be observed in a continuous-time fashion. However, their mean–response model covariates are fixed over time. Next, their sampling–times model does not allow for pre–specified visit times. Also, their estimator is fairly complicated, involving smoothing techniques.

To investigate finite sample behavior of the proposed estimator we have performed simulation studies, reported in Section 7. They show even for moderate sample sizes that asymptotic approximations are accurate. The proposed estimators have smaller squared error compared to the standard independent GEE estimators even with known correctly specified intercept and also compared to the original Lin and Ying's estimators.

2 Notation and models

We assume a fully marginal model for mean of response $Y_i(t)$, denoted by $\mu_i(t)$, as a function of covariates $X_i(t)$ of individual $i \in \{1, \ldots, n\}$ and baseline $\alpha_0(t)$ at time $t \in [0, \tau]$. The linear regression full data model is

$$E[Y_i(t)|X_i(t)] = \alpha_0(t) + \beta_0^T X_i(t). \tag{2·1}$$

The parameter of interest, β_0, is a p dimensional vector. The effect of possibly time–varying covariates X is modeled linearly and the effect of time is modeled completely non-parametrically. The intercept curve $\alpha_0(\cdot)$ is modeling the mean of response at a certain time given covariates X at that time are zero. It is not of special interest, it is a infinite-dimensional nuisance parameter. We note that there are no assumptions about the form of the intercept function, we require neither smoothness nor continuity of that function. This concept is a generalization to longitudinal data of the intercept in cross-sectional models based on one time point only. Estimators of β_0 that we consider do not require to estimate $\alpha_0(\cdot)$ correctly for their validity. We actually totally avoid estimation of $\alpha_0(\cdot)$. On the other hand, as $\alpha_0(\cdot)$ is not estimated, prediction of the mean of response is not possible.

We do not impose distributional assumption on the response process $\{Y(t) : t \in [0, \tau]\}$ in any way. We do not need to specify the within–person auto-covariance of the response process.

We model fully marginal mean of response and thus we do not need the Pepe & Anderson (1994) assumption about modeling the mean response at time t conditional on the whole covariate process over time.

The model for response, formulated in equation (2·1), is a functional full data model. However, we assume to observe response not continuously over time but at certain observation times only. Denote for individual $i \in \{1, \ldots, n\}$ the set of observation times $\{T_{i1}, T_{i2}, \ldots, T_{iK_i}\}$ as $\mathcal{T}_i$, with $0 \leq T_{i1} < T_{i2} < \ldots < T_{iK_i} \leq \tau$. Total number of observed events of the i-th individual, K_i, is random. Denote $\mathbb{T} = \{\mathcal{T}_j,\ j = 1, \ldots, n\}$ the set of observation times across all individuals. Define $N_i(t) = \sum_{k=1}^{K_i} I(T_{ik} \leq t)$ the number of observations of individual i by time t. Further let us define $N_i(0) = 0$. The underlying uncensored process we denote as $N_i^*(\cdot)$ with $N_i(t) = N_i^*(t \wedge C_i)$, where the symbol $\wedge$ is the minimum and variable C_i is drop-out time or end of follow-up τ, whatever comes first.

We assume a marginal rate model for the uncensored observation times of each individual $i \in \{1, \ldots, n\}$ at time $t \in [0, \tau]$

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

We assume that $EN_i^*(\tau) < \infty$. $\Lambda_0(\cdot)$ is an arbitrary non-decreasing function of time t, continuous up to countably many points, in our settings a finite number of points suffices.

The comparison of the proportional rate model (2·2) to the classical Cox-type proportional mean model can be found in D. Lin et al. (2000).

We impose two crucial assumptions: non-informative drop-out for the mean of response,

$$E[Y_i(t)|X_i(t), C_i \geq t] = E[Y_i(t)|X_i(t)], \tag{2·3}$$

saying that mean of response Y at time t depends on covariates X at time t and drop-out time C through covariates X at time t only; and an independent sampling assumption,

$$E[dN_i^*(t)|Z_i(t), X_i(t), Y_i(t), C_i \geq t] = E[dN_i^*(t)|Z_i(t)], \tag{2·4}$$

saying that sampling times depend on covariates Z, X, response Y and drop-out time C through sampling times covariates Z only.

We define the at–risk process $\{\xi(t),\ t \in [0, \tau]\}$ as $\xi(t) = I(C > t)$ and assume that $\Pr(C \geq \tau) > 0$.

Note, that although response Y of the i–th individual is observed only at a set $\mathcal{T}_i$ of random times, the expectations in (2·1) and (2·2) do not condition on these times.

Additional technical assumptions are given in the Appendix.

D. Lin & Ying (2001) require an additional assumption, which does exclude biased sampling. The assumption is that the response variable is assumed independent of the sampling times given the covariates of the mean–response model. That is, the sampling times are not allowed to depend on additional covariates not in the mean–response model; covariates Z must be part of covariates X. D. Lin & Ying's independent sampling assumption is

$$E[dN_i^*(t)|Z_i(t), X_i(t), Y_i(t), C_i \geq t] = E[dN_i^*(t)|X_i(t)]. \tag{2·5}$$

The difference between our independent sampling assumption (2·4) and their independent sampling assumption (2·5) is conditioning on covariates $Z(t)$ instead of $X(t)$ on the right hand side of the equation. It is a major difference. Our philosophy behind choosing model covariates is that the mean–response model covariates X should be picked on scientific ground purpose but the covariates Z for sampling–times model should cover the true nature of the sampling process. However, when we allow the two sets of covariates to be arbitrary, we can introduce a biased sampling scheme into our mean–response model and thus we need to account for the biased sampling to obtain consistent estimators of β_0.

3 Sampling–times model

Based on the proportional rates model (2·2) and the drop-out part of assumption (2·4) the 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

$$U^{\dagger}(\gamma) = \sum_{i=1}^{n} \int_0^{\tau} \{Z_i(t) - Av_1(Z)(t;\gamma)\} dN_i(t), \tag{3·1}$$

where the weighted mean Av_1 of any variable V at time t is

$$Av_1(V)(t;\gamma) = \sum_{i=1}^{n} V_i(t) \frac{\xi_i(t) \exp\{\gamma^T Z_i(t)\}}{\sum_{j=1}^{n} \xi_j(t) \exp\{\gamma^T Z_j(t)\}}. \tag{3·2}$$

The weighted mean Av_1 has weights proportional to the probability that individual i, relative to other individuals, has an observation at time t under the sampling–times model (2·2).

Estimation of the parameter γ_0 is β_0-free. Solution of the estimating equation and derivation of asymptotic properties of the estimator are based on a zero–mean random process $\{\mathcal{M}_i(t;\gamma_0,\Lambda_0(\cdot)),\ t \in [0,\tau]\}$ defined as

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

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 asymptotic weighted mean curve of a covariate process $\{V(t),\ t \in [0,\tau]\}$ as

$$av_1(V)(t;\gamma) = \lim_{n\to\infty} Av_1(V)(t;\gamma) = \frac{E[V_1(t)\xi_1(t)\exp\{\gamma^T Z_1(t)\}]}{E[\xi_1(t)\exp\{\gamma^T Z_1(t)\}]}.$$

and a matrix A and a matrix Σ as

$$\begin{aligned} A &= \lim_{n\to\infty} E\frac{1}{n}\left[\frac{-\partial U^{\dagger}(\gamma)}{\partial\gamma}|_{\gamma_0}\right] \\ &= E\int_0^{\tau} [Z_1(t) - av_1(z)(t;\gamma_0)]^{\otimes 2}\, \xi_1(t)\exp\{\gamma_0^T Z_1(t)\} d\Lambda_0(t) \qquad (3·4) \\ \Sigma &= \lim_{n\to\infty} \mathrm{Cov}\left[\frac{1}{\sqrt{n}} U^{\dagger}(\gamma_0)\right] \\ &= E\left[\int_0^{\tau} [Z_1(t) - av_1(z)(t;\gamma_0)]\, d\mathcal{M}_1(t;\gamma_0,\Lambda_0(\cdot))\right]^{\otimes 2}. \end{aligned}$$

Notation $v^{\otimes 2}$ stands for the outer product vv^T of a vector v. The matrix A is further used in the formula for variance of estimator of β_0 in the mean–response model (2·1) that is dependent upon estimation of the parameter γ_0. The asymptotic variance of $\sqrt{n}(\hat{\gamma} - \gamma_0)$ is Γ, where $\Gamma = A^{-1}\Sigma A^{-1}$. A straightforward consistent estimator of the variance Γ is $\hat{\Gamma} = \hat{A}^{-1}\hat{\Sigma}\hat{A}^{-1}$, where

$$\begin{aligned} \hat{\Sigma} &= \frac{1}{n}\left[\sum_{i=1}^{n}\int_0^{\tau} [Z_i(t) - Av_1(Z)(t;\hat{\gamma})]\, d\mathcal{M}_i(t;\hat{\gamma},\widehat{\Lambda}(\cdot))\right]^{\otimes 2} \\ \hat{A} &= \frac{1}{n}\sum_{i=1}^{n}\int_0^{\tau} [Z_i(t) - Av_1(Z)(t;\hat{\gamma})]^{\otimes 2}\, \xi_i(t)\exp\{\hat{\gamma}^T Z_i(t)\} d\widehat{\Lambda}(t) \end{aligned}$$

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

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

See D. Lin et al. (2000) for detailed derivation of the parameter estimation in the sampling–times model (2·2) and comparison of assumptions and estimation to the widely used proportional mean model.

4 Lin and Ying's estimator

D. Lin & Ying (2001) defined the estimator of parameter β_0 in the mean–response model (2·1) as

$$\begin{aligned} \hat{\beta}(\hat{\gamma}) \quad = \quad & \left[\sum_{i=1}^{n} \int_0^\tau W(t) \left\{ X_i(t) - Av_1(X)(t;\hat{\gamma}) \right\}^{\otimes 2} dN_i(t) \right]^{-1} \times \\ & \times \sum_{i=1}^{n} \int_0^\tau W(t) \left\{ X_i(t) - Av_1(X)(t;\hat{\gamma}) \right\} \left\{ Y_i(t) - Av_1(Y^\star)(t;\hat{\gamma}) \right\} dN_i(t). \quad (4\cdot1) \end{aligned}$$

They denote $Y^\star(t)$ any approximation to the true response value at time t that has a nonrandom limit. We use nearest neighbor approximation both in our simulations and the data analysis. The equation (4·1) is resembling a least squares estimator with, at each time point t, centered both covariates X and response Y. D. Lin & Ying (2001) showed that, under the additional assumption (2·5) of no biased sampling, the estimator $\hat{\beta}(\hat{\gamma})$ is $\sqrt{n}$–consistent and asymptotically normal.

5 Estimation under biased sampling

We define a class of new estimators for a general case of dependence of the sampling–times process and the mean–response process allowing for biased sampling. Out of the class we highlight one estimator with certain variance–stabilizing properties. This " stabilized" estimator collapses down to the original D. Lin and Ying's estimator under unbiased sampling, once sampling-times model covariates Z are subset of the mean–response model covariates X.

For individual $i \in \{1, \ldots, n\}$ at time $t \in [0, \tau]$ we define the inverse weights as

$$\rho_i(t;\gamma,h) = \frac{\exp\{\gamma^T Z_i(t)\}}{h(X_i(t))}. \quad (5\cdot1)$$

The inverse weight $\rho_i(t;\gamma_0,h)$ is proportional to the probability that individual i, relative to other individuals, has an observation at time t under the sampling–times model (2·2). The weight helps us to standardize the observed data to the underlying population. The important component in the inverse weights (5·1) is the numerator; the denominator does not change the conditional expectation as long as it is a deterministic function of the mean–response models covariates X. The denominator is there only to improve the precision.

To insure nice asymptotic properties of the final estimator of β_0 we assume that $\rho_i(t;\gamma_0,h)$ is bounded away from zero and that the function $h(\cdot)$ has bounded variation. That is for some $c > 0$ for all $t \in [0, \tau]$

$$\rho_i(t;\gamma_0,h) > c$$

for all individuals $i \in \{1, \ldots, n\}$ and for some $K < \infty$

$$|h(0)| + \int_0^\infty |dh(x)| \leq K.$$

We note that ρ bounded away from zero restricts the variation of sampling probability between individuals, but not between time points. It is not necessary that $\Lambda(t)$ be continuous.

Let us define a random process $\{M_i(t) = M_i(t; \beta, \gamma, \mathcal{A}(\cdot), h(\cdot)),\ t \in [0, \tau]\}$ as

$$M_i(t) = \int_0^t \frac{1}{\rho_i(s; \gamma, h)} \left\{ \left[Y_i(s) - \beta^T X_i(s) \right] dN_i(s) - \xi_i(s) \exp\{\gamma^T Z_i(s)\} d\mathcal{A}(s) \right\}, \quad (5{\cdot}2)$$

where $\mathcal{A}(t) = \int_0^t \alpha(s) d\Lambda(s)$. We note that the process defined by equation (5·2) is a weighted version of the process used by D. Lin & Ying (2001). We claim that

$$E\left[dM_i(t; \beta_0, \gamma_0, \mathcal{A}_0(\cdot), h) | X_i(t) \right] = 0$$

for any $h(X_i(t))$. With no knowledge about the variance of response, we want to make the weights variance, $\text{var}[\rho_i(t; \gamma, h)]$, as small as possible to increase the efficiency of the estimator of β_0. 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 and the estimator using this weight a "stabilized" estimator. The best choice of δ_0 we base on an estimator of δ_0 in a proportional rate model conditioning on covariates X, similar to model (2·2). When sampling–times model covariates Z are a subset of the mean–response model covariates X, then $\rho_i(t; \gamma, h_0) = 1$ for all individuals at all times, using the independent sampling assumption (2·4).

The fundamental set of estimating equations, based on the process $\{M_i(t),\ t \in [0, \tau]\}$ as defined in equation (5·2), and its properties, is

$$\sum_{i=1}^n M_i(t; \beta, \gamma_0, \mathcal{A}(\cdot), h) = 0 \ \ \forall t \in [0, \tau] \quad (5{\cdot}3)$$

$$\sum_{i=1}^n \int_0^\tau W(t) X_i(t)\, dM_i(t; \beta, \gamma_0, \mathcal{A}(\cdot), h) = 0, \quad (5{\cdot}4)$$

where $\{W(t),\ t \in [0, \tau]\}$ is a weight process. We solve the infinite-dimensional equation (5·3) for $\{\hat{\mathcal{A}}(t),\ t \in [0, \tau]\}$. Solving equation (5·3) at time $t \in [0, \tau]$ yields

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

We define a weighted mean Av_2 for any variable V at time t as

$$Av_2(V)(t; h) = \sum_{i=1}^n V_i(t) \frac{\xi_i(t) \frac{1}{\rho_i(t;\gamma,h)} \exp\{\gamma^T Z_i(t)\}}{\sum_{j=1}^n \xi_j(t) \frac{1}{\rho_j(t;\gamma,h)} \exp\{\gamma^T Z_j(t)\}} = \sum_{i=1}^n V_i(t) \frac{\xi_i(t)\, h(X_i(t))}{\sum_{j=1}^n \xi_j(t)\, h(X_j(t))}.$$

The sample mean curve is γ–free and β–free. Replacing $\mathcal{A}(t)$ with its actual estimator obtained above, equation (5·4) yields an estimating function

$$U(\beta;\gamma_0,h)=\sum_{i=1}^{n}\int_0^{\tau}W(t)\left[X_i(t)-Av_2(X)(t;h)\right]\left[Y_i(t)-\beta^T X_i(t)\right]\frac{1}{\rho_i(t;\gamma_0,h)}dN_i(t). \quad (5\cdot5)$$

Inspired by Rotnitzky et al. (1998), we add a quantity into the estimating function that does not change its expectation at true (β_0,γ_0) but decreases its variance. We can subtract an arbitrary deterministic function of time denoted by $g(t)$ from the third right-hand side term of expression (5·5). To minimize variance of our proposed estimator, an optimal $g(t)$ is $\alpha_0(t)$, approximated by $Av_2(Y^\star)(t;h)-\beta^T Av_2(X)(t;h)$. We set the estimating function (3·1) to 0 for $\hat{\gamma}$ and substitute those into equation (5·5) as well. The final estimating equation becomes

$$\begin{aligned}U(\beta;\hat{\gamma},h) &= \sum_{i=1}^{n}\int_0^{\tau}W(t)\left[X_i(t)-Av_2(X)(t;h)\right]\times\\ &\times \left\{Y_i(t)-Av_2(Y^\star)(t;h)-\beta^T\left[X_i(t)-Av_2(X)(t;h)\right]\right\}\frac{1}{\rho_i(t;\hat{\gamma},h)}\,dN_i(t)\end{aligned} \quad (5\cdot6)$$

The proposed final estimator of β_0 from model (2·1) has the form

$$\begin{aligned}\hat{\beta}(\hat{\gamma},h) &= \left[\sum_{i=1}^{n}\int_0^{\tau}\left\{X_i(t)-Av_2(X)(t;h)\right\}^{\otimes 2}\frac{W(t)}{\rho_i(t;\hat{\gamma},h)}\,dN_i(t)\right]^{-1}\times\\ &\times \sum_{i=1}^{n}\int_0^{\tau}\left\{X_i(t)-Av_2(X)(t;h)\right\}\left\{Y_i(t)-Av_2(Y^\star)(t;h)\right\}\frac{W(t)}{\rho_i(t;\hat{\gamma},h)}\,dN_i(t)\end{aligned} \quad (5\cdot7)$$

When using the stabilized estimator, we can substituted the unknown δ_0 in (5·7) with its estimator without changing the asymptotic properties, see van der Vaart (2000) Theorem 5.31.

The asymptotic variance of $\sqrt{n}(\hat{\beta}(\hat{\gamma},h)-\beta_0)$ is $D^{-1}VD^{-1}$. The matrix of derivatives of the estimating function U with respect to the parameter of interest β is

$$\begin{aligned}D &= \lim_{n\to\infty}E\left[-\frac{1}{n}\frac{\partial U(\beta;\gamma_0,h)}{\partial\beta}|_{\beta_0}\right]\\ &= E\int_0^{\tau}w(t)\left[X_1(t)-av_2(X)(t;h)\right]^{\otimes 2}\frac{1}{\rho_1(t;\gamma_0,h)}\,dN_1(t),\end{aligned} \quad (5\cdot8)$$

where the asymptotic weighted mean av_2 for a variable V at time t is

$$av_2(V)(t;h)=\lim_{n\to\infty}Av_2(V)(t,h)=\frac{E[V_1(t)\,\xi_1(t)\,h(X_1(t))]}{E[\xi_1(t)\,h(X_1(t))]}.$$

Matrix D can be consistently estimated by

$$\hat{D}=\frac{1}{n}\sum_{i=1}^{n}\int_0^{\tau}W(t)\left[X_i(t)-Av_2(X)(t;h)\right]^{\otimes 2}\frac{1}{\rho_i(t;\hat{\gamma},h)}\,dN_i(t).$$

Further, we define the covariance matrix of the estimating function U as

$$\begin{aligned}V &= \lim_{n\to\infty}\mathrm{Cov}\left[\frac{1}{\sqrt{n}}U(\beta_0;\hat{\gamma},h)\right]=E\left[\int_0^{\tau}w(t)\left[X_1(t)-av_2(X)(t;h)\right]dR_1(t;\beta_0,\gamma_0,\mathcal{A}_0(\cdot),h)-\right.\\ &\left.-HA^{-1}\int_0^{\tau}\left[Z_1(t)-av_1(Z)(t;\gamma_0)\right]^T d\mathcal{M}_1(t;\gamma_0,\Lambda_0(\cdot))\right]^{\otimes 2},\end{aligned} \quad (5\cdot9)$$

where we account for the dependence of the estimator of β_0 on estimation of the sampling–times model parameter γ_0. Matrix A is as defined in formula (3·4) and process $\{\mathcal{M}(t),\ t \in [0,\tau]\}$ in formula (3·3). We define process $\{R(t),\ t \in [0,\tau]\}$ as

$$R_i(t;\beta,\gamma,\mathcal{A}(\cdot),h) = M_i(t;\beta,\gamma,\mathcal{A}(\cdot),h) - \int_0^t (\bar{Y}^\star(s;h) - \beta^T A v_2(X)(s;h)) \frac{1}{\rho_i(t;\gamma,h)} d\mathcal{M}_i(s;\gamma,\Lambda(\cdot))$$

and the matrix H is a matrix of derivatives of the estimating function U with respect to the sampling–times model parameter γ at the true parameter value γ_0

$$H = \lim_{n\to\infty} E\left[-\frac{1}{n}\frac{\partial U(\beta_0;\gamma,h)}{\partial\gamma}|_{\gamma_0}\right] = E\int_0^\tau w(t)\left[X_1(t) - av_2(X)(t;h)\right] \times$$
$$\times \quad \left[Y_1(t) - av_2(Y^\star)(t;h) - \beta_0^T\left[X_1(t) - av_2(X)(t;h)\right]\right] Z_1(t)\ \frac{1}{\rho_1(t;\gamma_0,h)}\ dN_1(t) \quad (5\cdot10)$$

The matrix H can be consistently estimated by

$$\begin{aligned}\hat{H} \quad = \quad & \frac{1}{n}\sum_{i=1}^n \int_0^\tau W(t)\left[X_i(t) - Av_2(X)(t;h)\right] \times \\ & \times \left[Y_i(t) - Av_2(Y^\star)(t;h) - \hat{\beta}^T\left[X_i(t) - Av_2(X)(t;h)\right]\right] Z_i(t)\frac{1}{\rho_i(t;\hat{\gamma},h)}\ dN_i(t).\end{aligned}$$

A consistent estimator of the matrix V is thus

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

6 HUD–VASH study

In 1992, the US Department of Housing and Urban Development (HUD) and the US Department of Veterans Affairs (VA) established the HUD–VA Supported Housing (HUD–VASH) program. The study took place at four sites across the country. Veterans were eligible if they were homeless at the time of outreach assessment, had been homeless for one month or longer, and had received a diagnosis of a major psychiatric disorder or an alcohol or drug abuse disorder. All veterans provided written informed consent to participate in the study. The 460 homeless veterans were randomly assigned to one of three 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 scientific question was 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 three

Table 1: HUD–VASH: quantiles of number of follow–up visits per individual by treatment arm.

	minimum	25%	median	75%	maximum
HUD–VASH	1	7	9	10	12
case management	1	5	7	9	12
standard care	1	3	6	8	12

months. The data collected at baseline were income, an indicator of receiving any social security or VA benefits and a Lehman measure of the quality of life. 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 three months. However, subjects often missed assessment and came between scheduled interviews. Concern was raised that there as an association between the visit process and the outcome process. For detailed study description see Rosenheck et al. (2003). This paper also addresses cost-effectiveness considerations for the three interventions.

In the analysis of the data, we set τ to 48 months based on the span of the observed data and $C_i = \tau$ for all individuals $i \in \{1, \ldots, 460\}$. That means that we do not allow anybody to drop–out of the study sooner than at the 48 months. There is not any drop–out by protocol that would exclude certain individuals after study beginning and if no event occurs by the study end we consider that just an intermittent missing data. The 460 individuals made a total of 2855 follow–up visits by 48 months since randomization. Quantiles of the total counts of follow–up visits per treatment arm, shown in Table 1, suggest highest follow–up for the HUD–VASH intervention group, lower for the case–management group and lowest for the standard VA care. Figure 1 shows the primary outcome of percentage days 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. The HUD–VASH intervention group has the highest level of follow–up visits and the standard care group the lowest level of visiting.

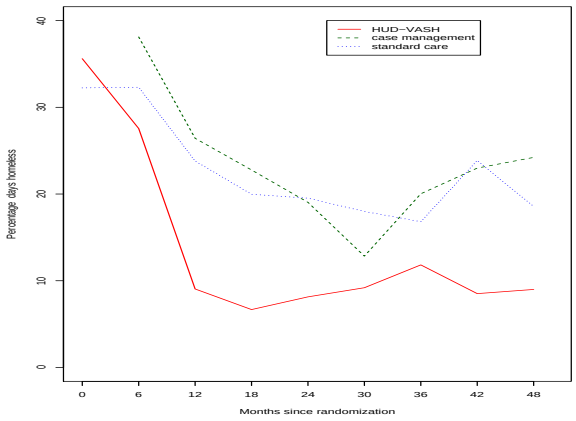
To answer the question of efficacy of intervention we model the percentage days homeless during the last three months, denoted as PH, as a function of treatment assignment. We consider a linear regression model

$$\begin{aligned} E\left[PH_i(t)|Trt_i\right] = \alpha_0(t) \quad &+ \quad \beta_{01}\, I(Trt_i = \text{HUD–VASH}) + \\ &+ \quad \beta_{02}\, I(Trt_i = \text{case management}). \end{aligned} \tag{6·1}$$

The mean–response model (6·1) assumes a shift of the mean curve of days homeless within the last three months across the interventions. The standard VA homeless service arm is the reference group. The primary parameter of interest is the 2-dimensional parameter vector $\beta_0 = (\beta_{01}, \beta_{02})^T$.

The sampling–times model we define in equation (6·2) as a proportional rate model. The covariates were suggested by the primary investigator Dr. Rosenheck. The time–invariant predictors of timing of visits are intervention assignment, income at baseline, in thousands of dollars, denoted as IB, an indicator of receiving any social security or VA benefits at baseline, BB and a Lehman measure of the quality of life at baseline. Time–varying predictors for the sampling–times model are percentage homeless approximated

Figure 1: HUD–VASH: averaged percentage days homeless during the last three months by treatment arm.



by previous value carried forward, denoted by $PH^{\star}$, and cumulative number of visits so far, denoted by N_{-}.

$$E\left[dN_i^{\star}(t)|Trt_i, IB_i, BB_i, PH_i^{\star}(t), QLB_i, N_{-i}(t)\right] = \exp\left\{\gamma_{01} I(Trt_i = \text{HUD–VASH}) + \right.$$
$$+ \gamma_{02} I(Trt_i = \text{case management}) + \gamma_{03} IB_i + \gamma_{04} BB_i + \gamma_{05} PH_i^{\star}(t) + \gamma_{06} QLB_i +$$
$$\left. + \gamma_{07} N_{-i}(t) + \gamma_{08} N_{-}^{\text{HUD–VASH}}{}_i(t) + \gamma_{09} N_{-}^{\text{case management}}{}_i(t) \right\} d\Lambda_0(t) \qquad (6{\cdot}2)$$

We estimate the parameter γ_0 of dimension 9. Parameter estimates, as shown in Table 2, suggest that higher intensity of visiting is associated with lower baseline income, lower baseline quality of life, receiving any social or VA benefits at baseline, having higher approximated percentage days homeless and higher cumulative number of visits so far, differentiated by treatment arm. 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, with the same quality of life baseline measure, indicator of social or VA benefits at baseline, approximated percentage days homeless and having the same number of visits so far, assuming that the number of visits so far is in the range of (0, 11). 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, with the same quality of life baseline measure, indicator of social or VA benefits at baseline, approximated percentage days homeless and having the same number of visits so far, ranging from 0 to 11.

On the 5% level we see a statistically significant difference of proportion of days homeless within the last 3 months between the HUD–VASH intervention arm and standard VA care arm, favoring the HUD–VASH intervention. At any time the percentage days homeless during the last three months averaged in the HUD–VASH treatment arm is by 10.344% lower than in the case management treatment arm. The 95% confidence interval is $(3.6\%, 17.0\%)$. Scientifically a 10% decrease in the proportion of days homeless is significant. The estimate of β_{02} suggests increase of proportion days homeless comparing the case management group to the standard VA care. However, on 5% statistical significance level 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. We

Table 2: HUD–VASH: parameter estimates and their standard errors in the intensity rate model (6·2) for sampling times.

	$\widehat{\gamma_0}$	$SE(\widehat{\gamma_0})$
HUD–VASH	0.359	0.044
case management	0.217	0.054
IB	-0.172	0.482
BB	0.104	0.041
$PH^{\star}$	0.001	0.001
QLB	-0.007	0.019
N_{-}	0.044	0.015
$N_{-}^{\text{HUD–VASH}}$	-0.018	0.016
$N_{-}^{\text{case management}}$	-0.014	0.023

note that the sign of that estimate is surprisingly positive. These findings support the conclusion that setting aside housing resources is necessary and sufficient for facilitating exit from homelessness in this population.

Table 3: HUD–VASH: estimates of primary parameter of interest (β_{01}, β_{02}) and their standard errors in the mean–response model (6·1).

	Intervention group	
	HUD–VASH	Case management
$\widehat{\beta_0}$	-10.344	0.594
$SE(\widehat{\beta_0})$	3.411	5.843

We contrast our results to those obtained when fitting the linear regression model (6·1) but not accounting for the biased sampling. There, we compute the Lin and Ying's estimates as well as the naive estimates that would have very often be used, based on the GEE. There the parametric model is

$$\begin{aligned} E\left[PH_i(t)|Trt_i\right] &= \boldsymbol{\beta_{00}} f(t) + \beta_{01}\, I(Trt_i = \text{HUD–VASH}) + \\ &\quad + \beta_{02}\, I(Trt_i = \text{case management}), \end{aligned} \tag{6·3}$$

where $f(t)$ is a natural cubic spline with 4 degrees of freedom. Both naive parameter estimates, shown in Table 4, suggest qualitatively the same answer. However, we see in both a decrease in favoring the HUD–VASH treatment and also increase of disliking the case management care. Fitting the sampling–times model (6·2) 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. We conclude that the data tend to be biased upwards. We note that we do not see any substantial efficiency loss when in the mean–response model we leave the intercept unspecified as in equation (6·1) versus using the natural spline in equation (6·3).

7 Simulations

We consider a semi-parametric additive marginal model

$$E\left[Y(t)|X_1(t)\right] = \alpha_0(t) + \beta_{01} X_1(t), \tag{7·1}$$

Table 4: HUD–VASH: Lin & Ying estimates in model (6·1) and naive GEE estimates in model (6·3) of the primary parameter of interest $\beta_0 = (\beta_{01}, \beta_{02})^T$. Standard errors are included in parenthesis.

	Intervention group	
$\widehat{\beta_0}$ $(SE(\widehat{\beta_0}))$	HUD–VASH	Case management
LY	-8.141 (3.001)	2.054 (4.978)
GEE	-7.571 (2.841)	5.141 (4.446)

where the mean–response model covariate is X_1. Estimation of parameter β_{01} is of our major interest, quantifying the association between the response and the covariate X_1. The model (7·1) arises from a random effect model such as

$$Y(t) = \alpha_0(t) + \beta_{01}X_1(t) + \beta_{02}\left(Z_2(t) - E[Z_2(t)|X_1(t)]\right) + \epsilon(t). \tag{7·2}$$

We note that we do not want to confuse the reader with confounding in the mean–response model (7·1) and thus in model (7·2) we subtract the expectation of the covariate Z_2 conditional upon covariate X_1. Model (7·1) is desirable for instance when the covariate Z_2 is modifying the association of response and covariate X_1. To link the motivating example of biased data, the covariate X_1 is the air pollution measure and covariate Z_2 is the indicator of an asthma attack. However, the sampling–times are driven by both covariates $Z_1 = X_1$ and Z_2. The random effect, hidden in the error term ϵ, is used to introduce autocorrelation. We impose a Normal distribution on the error ϵ_i of the form $\epsilon_i|\phi_i \sim N(\phi_i, \sigma^2_\epsilon)$, with mean ϕ_i being an $N(0, \sigma^2_\phi)$ random variable. As ϕ_i is fixed for a person, errors and thus responses on the same subject are positively correlated in time whenever $\sigma^2_\phi > 0$. We take $\sigma_\phi = 0.2$ and $\sigma_\epsilon = 0.1$, resulting in error variance 0.05 and correlation 0.8 for any two time points.

The following functions were considered as the baseline predictor $\alpha_0(t)$: a nonlinear trend $\sqrt{t}$, a sine–wave $\sin(t)$, and three fairly extreme functions $\exp\left(range\,|\sin(t)|\right)$, $\sin(peak\ t)$ and $\exp(range\,|\sin(peak\ t)|)$. Parameter *range* controls the extreme size of the intercept values and was set to 2. The peakedness parameter *peak* was set to 3. In Figure 2 we plot the five functions considered with the specific parameters. The nonlinear $\sqrt{t}$ trend and sine wave were considered in D. Lin & Ying (2001) in simplified models. The two exponential baseline curves should mimic air pollution data with the signal order of magnitude smaller than confounders, as found in Dominici (2004). We do not assume any 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 function.

Covariate X_1 is chosen to be Bernoulli distributed with 0.5 success probability at any time, demonstrating a treatment when $X_1(t) = 1$ and placebo when $X_1(t) = 0$. The second covariate Z_2 is dependent upon the first covariate. If a person is not at certain time t on treatment, then $Z_2(t)$ is normally distributed with mean and variance four. If a person is at certain time on a treatment, then $Z_2(t)$ is normally distributed with mean two and variance one. The intention in these settings is to model a reducing effect of treatment on values of the second covariate. Discretization of continuous time is based on a grid of 100 per a time unit. Parameters β_{01} and β_{02} were set to 1 and 3, respectively.

For the sampling–times model, the observation times follow a random-effect Poisson counting process with intensity $\lambda_i(t) = \eta_i \exp\{\gamma_{01}Z_{1i}(t) + \gamma_{02}Z_{2i}(t)\}$. The random effect η_i is Gamma distributed with mean $\mu_\eta = 1$ and variance $\sigma^2_\eta = 0.01$. Thus for each individual the times of observations are positively correlated. Parameter γ_1 we set to -0.2, γ_2 to 0.3,

Table 5: Quantiles of number of follow–up visits per individual in simulations.

	minimum	25%	median	75%	maximum
$\tau = 1$	1	1	2	3	13
$\tau = 4$	1	6	8	10	29

making a person on a treatment at time t less likely to have an observation at that time.

The censoring variable C is distributed uniformly on the interval $(\tau/2, \tau)$. Setting with τ to 1 and 4 should demonstrate cases of a few and many observations per person. The resulting quantiles of number of observations taken over all individuals and simulations are in Table 5. The weight W is set to 1 over the entire time span $[0, \tau]$.

We present bias, sampling standard error, SSE, and sampling mean of estimated standard errors, SEE, of the estimates $\hat{\beta}_{01}$ taken over 1000 simulations. We also present two measures of squared errors comparison among two estimation approaches, denoted for convenience by relative efficiency RE. RE I is based on mean of the ratio of empirical mean squared error of estimate (4·1) of β_{01} over empirical mean squared error of GEE estimate of β_{01}. RE II is based on empirical median of ratios of squared errors, a more robust efficiency estimate motivated by Pitman closeness. We report 95% sampling coverage probability. With 1000 simulations the precision of the coverage probability is about 1.4%. Number of individuals in a sample is set to 20, 50, 100 and 200.

We compare the proposed estimator with the independent GEE estimator assuming a working independence and a known intercept as well as the original Lin and Ying's estimator. In these biased settings the proposed estimator can account for the biased sampling and thus is still consistent. However, both the GEE estimator and the original LY estimator are inconsistent.

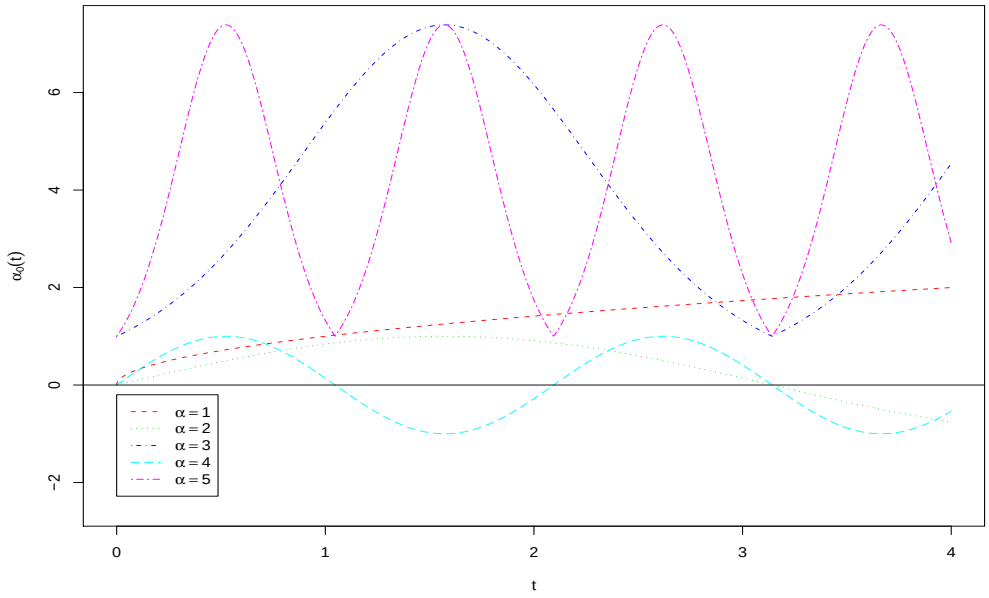
Simulation results for sample size 50 we show in Table 6 and for sample size Table 7. Findings from those tables are consistent for sample sizes of 20 and 100 individuals as well, those tables are not shown here. The bias estimate of the proposed estimator is always negligible relative to the sampling standard error, SSE, for all intercept functions and the two scenarios of numbers of observations per individual. The model based variance of the proposed estimator of the coefficient β_{01}, SEE, is usually very slightly underestimating the true variance of the estimator of β_{01}, SSE. This discrepancy between SEE and SSE is diminishing with increasing sample size. We do not see any pattern regarding the intercept function. The 95% coverage probability of the proposed estimator, CP, is ranging from 92% to 94% for studies with 50 individuals and from 94% to 95% for studies with 200 individuals.

Both RE I and RE II favor the proposed estimator over the two naive estimators. RE II, based on Pitman closeness, favors our proposed estimator more than RE I. Based on median RE II is robust to ratios of squared errors that are large in rare cases. As we expected, the RE of the proposed IIRR estimator over GEE estimator with provided intercept as an offset is smaller that the RE of the proposed IIRR estimator over the LY estimator. The reason is the provided intercept for the GEE estimator. In our studies that drawback of both the LY estimator and our proposed IIRR estimator of increased variance was largely overcome by the decreased bias term. With more repeated measurements per person as well as larger sample size are both RE measures increasing. Our proposed estimator performs well under moderate and small sample size, including the estimator of its variance. It performs superior to the GEE estimator and the LY estimator.

In other simulations under unbiased sampling we addressed the possible loss of precision

under extreme intercept functions in the LY estimators (and thus the IIRR estimators) compared to the GEE estimators. Under those scenarios even a slight misspecification of the intercept function in the GEE estimator produced however much greater squared error in the GEE estimate than the loss of precision due to unspecified intercept. We concluded there that the precision loss was hugely overweighted by unbiasedness.

Figure 2: Intercept functional forms: $\alpha = 1$ for $\sqrt{t}$, $\alpha = 2$ for $\sin(t)$, $\alpha = 3$ for $\exp\left(range\,|\sin(t)|\right)$, $\alpha = 4$ for $\sin(peak\ t)$ and $\alpha = 5$ for $\exp(range\,|\sin(peak\ t)|)$.



8 Discussion

In this paper, we propose a class of consistent and asymptotically normal estimators for the parameters of group contrasts in the linear regression model for longitudinal data under biased sampling that occurs at continuous time. We call our estimators "inverse–intensity rate–ratio–weighted" estimators, abbreviated as IIRR estimators.

Under biased sampling, the mean–response process and the sampling–times process are allowed to be correlated conditional on the mean–response model covariates. We believe that this biased sampling setting is very appealing when longitudinal data are collected at irregular, possibly subject–specific times. Those can arise as a result of noncompliance to the scheduled visit times or as a result of observational nature of the data where there is no visit schedule.

Our mean–response model is a fully marginal model, where we model response at any time as a function of covariates at that time point only, not requiring the Pepe & Anderson (1994) assumption. We adapt a process–like approach where we say that the response follows the given model at any time, regardless whether an observation was collected or not.

The sampling–times model we use, namely a proportional rate model, is a semiparametric model allowing for flexibility. However, we note that for validity of our proposed

Table 6: Statistics for the proposed estimator of parameter β_{01} in model (7·1) under biased sampling for sample size 50.

					GEE		LY	
	Bias	SSE	SEE	CP	RE I	RE II	RE I	RE II
$\alpha = 1, \tau = 1$	−0.111	1.188	1.096	0.92	1.8	2.5	4.5	9.6
$\alpha = 2, \tau = 1$	−0.106	1.151	1.091	0.92	1.8	2.5	4.8	8.8
$\alpha = 3, \tau = 1$	−0.126	1.191	1.104	0.93	1.7	2.3	4.6	8.6
$\alpha = 4, \tau = 1$	−0.136	1.196	1.091	0.93	1.7	2.3	4.4	8.8
$\alpha = 5, \tau = 1$	−0.055	1.205	1.151	0.94	1.8	2.2	4.4	8.4
$\alpha = 1, \tau = 4$	−0.065	0.626	0.609	0.94	2.3	4.6	14.5	28.5
$\alpha = 2, \tau = 4$	−0.071	0.639	0.615	0.92	2.2	4.7	14.2	29.8
$\alpha = 3, \tau = 4$	−0.080	0.637	0.620	0.93	2.2	4.5	14.2	31.4
$\alpha = 4, \tau = 4$	−0.110	0.620	0.602	0.93	2.3	4.7	14.8	33.4
$\alpha = 5, \tau = 4$	−0.083	0.635	0.628	0.94	2.1	4.5	14.4	32.2

Table 7: Statistics for the proposed estimator of parameter β_{01} in model (7·1) under biased sampling for sample size 200.

					GEE		LY	
	Bias	SSE	SEE	CP	RE I.1	RE II.1	RE I.2	RE II.2
$\alpha = 1, \tau = 1$	−0.073	0.572	0.575	0.94	2.6	5.0	21.5	42.9
$\alpha = 2, \tau = 1$	−0.052	0.575	0.573	0.95	2.7	5.5	21.0	44.3
$\alpha = 3, \tau = 1$	−0.039	0.593	0.568	0.94	2.4	5.5	19.7	45.0
$\alpha = 4, \tau = 1$	−0.045	0.600	0.572	0.94	2.5	5.2	19.4	42.7
$\alpha = 5, \tau = 1$	−0.027	0.580	0.595	0.94	2.7	5.9	20.8	48.9
$\alpha = 1, \tau = 4$	−0.013	0.320	0.312	0.94	8.1	18.6	66.7	152.9
$\alpha = 2, \tau = 4$	−0.024	0.316	0.311	0.94	8.2	18.6	68.7	164.3
$\alpha = 3, \tau = 4$	−0.032	0.324	0.314	0.94	7.8	17.7	64.5	144.7
$\alpha = 4, \tau = 4$	−0.038	0.314	0.312	0.94	8.1	18.4	69.5	157.6
$\alpha = 5, \tau = 4$	−0.003	0.318	0.320	0.95	8.3	18.5	67.5	149.2

estimators it is extremely important to specify the sampling–times model correctly. The parametric part of the model can be specified different and, after modification of the definition of the inverse weights, we would still obtain correct inference. We will further study the robustness of our estimation approach under sampling–times model misspecification where a multiplicative model is not suitable. We note that a positive probability of having an observation at certain time, which means spikes in the intensity function, is manageable by our approach, once all individuals share those time points. The covariates of the sampling–times model should be strongly predictive of the sampling times, whereas choice of the covariates of the mean–response model should be governed by the scientific question.

Our mean–response model is a semiparametric model, where we do not specify the intercept function. Naturally, efficiency of estimation under the semiparametric mean–response models can be lower than under a correct parametric mean–response model. In simulations we have seen, however, that under those scenarios where most efficiency was lost, a huge amount of accuracy was lost when an incorrect parametric intercept was used. Also, elsewhere we provide an estimation approach that accommodates biased

sampling using a parametric model and thus reader has the option of a parametric versus semiparametric model choice.

We do not impose any distributional assumptions on the response process. Covariates in the mean–response model can contain lagged covariate values. Moreover, covariates in the sampling–times model can contain lagged response values.

We require a non–informative drop–out for the mean–response model. This assumption might seem to be a strong one. However, if there is still a slightest chance that a person has not completely dropped out but we just temporarily do not see any observations, we can handle that as biased sampling.

Survival analysis with time–varying covariates requires that the entire covariate process is observed. It is not a coincidence that our approach requires the same. To be specific, we need covariate values on any individual at any observation time in the sample as long as the individual is still under follow–up. There is not a simple solution to that data collection issue. Approximations of the covariate processes cause biased estimators. Subcohort sampling techniques are not trivially applied in practice. H. Lin et al. (2004) solve the same problem in an effectively equivalent way by defining a sampling–times model that conditions on the last observed value of covariates.

Separation of the sampling–times model and the mean–response model as two distinct models enables to perform model checking separately. A range of model checking techniques of the sampling–times model (2·2) was suggested in Section 4 of D. Lin et al. (2000) using certain cumulative sums of residuals based on the process $\{\mathcal{M}(t),\ t \in [0,\tau]\}$ defined in equation (3·3). We keep in mind that those residuals are not martingales and thus present a technical challenge to construct formal tests. Those model checking techniques include both graphical and numerical inspections for functional form of covariates, exponential link function and proportional rates assumptions. For the last we can plot Schoenfeld residuals against time with a fitted smoother, just as described in Grambsch & Therneau (1994) checking proportional hazard assumption. Also an omnibus test for checking the overall fit of the model was constructed. We would like to address similar matter in our future work for the mean–response model checking. We believe that the residuals based on the process $\{M(t),\ t \in [0,\tau]\}$ could be used for that in a similar fashion.

9 Appendix

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 $t \in [0,\tau]$ for individuals 1 through n. The counting uncensored process of events at the end of follow-up τ, $N_i^\star(\tau)$, is required to be bounded by a constant. Both mean–response model covariates X_i and sampling–times model covariates Z_i need to have a bounded total variations by a constant for all individuals. That is $|Z_{ji}(0)| + \int_0^\tau |dZ_{ji}(t)| \leq K,\ j = 1,\ldots,g$ and $|X_{ji}(0)| + \int_0^\tau |dX_{ji}(t)| \leq K,\ j = 1,\ldots,p$. The function $h(\cdot)$ also needs to have bounded variation.

The inverse weight $\rho_i(t;\gamma,h)$ needs to bounded away from zero. The weight function $W(\cdot)$ is assumed to be a difference of two monotone functions, each of which converges to a deterministic function. We denote the asymptotic limit of $W(\cdot)$ by $w(\cdot)$.

Implementation of the estimation procedure

The above estimation procedure can be implemented in S-plus/R with relative ease. The sampling–times model can be fitted by function *coxph* to obtain estimates of γ_0 and δ_0. Function *lm* being applied to centered, by subtracting the mean curve, covariates X and centered and weighted by inverse of ρ response Y provides the estimate of β_0. The standard errors of $\hat{\beta}(\hat{\gamma}, \hat{\delta})$ can be obtained by bootstrapping or implementing the asymptotic formulas for variance. We will add an implicit function into R to provide the estimate of β_0 and its variance directly based on row data.

Large Sample Theory

Based on the asymptotic theory as established in D. Lin & Ying (2001) using monotone functions and "manageable processes" tools, as described in Pollard (1990) and Bilias et al. (1997), we derive the large sample properties of the proposed estimator. The estimating function (5·6) at point $(\beta_0; \gamma_0, h)$ can be written as

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

Further $\frac{1}{\sqrt{n}}U(\beta_0;\gamma_0,h)$ is asymptotically equivalent to

$$\begin{aligned}
&\frac{1}{\sqrt{n}}\int_0^t w(s)\sum_{i=1}^{n} X_i(s)\,dM_i(s;\beta_0,\gamma_0,\mathcal{A}_0(\cdot),h) - \\
&-\frac{1}{\sqrt{n}}\int_0^t w(s)\left[av_2(Y^\star)(s;h) - \beta_0^T av_2(X)(s;h)\right]\sum_{i=1}^{n}\rho_i(s;\gamma_0,h)X_i(s)\,d\mathcal{M}_i(s;\gamma_0,\Lambda_0(\cdot)) - \\
&-\frac{1}{\sqrt{n}}\int_0^t w(s)av_2(X)(s;h)\;\sum_{i=1}^{n} dM_i(s;\beta_0,\gamma_0,\mathcal{A}_0(\cdot),h) + \\
&+\frac{1}{\sqrt{n}}\int_0^t w(s)av_2(X)(s;h)\left[av_2(Y^\star)(s;h) - \beta_0^T av_2(X)(s;h)\right]\;\sum_{i=1}^{n}\rho_i(s;\gamma_0,h)d\mathcal{M}_i(s;\gamma_0,\Lambda_0(\cdot)).
\end{aligned}$$

A sequence of Taylor series expansions yield

$$\frac{1}{\sqrt{n}}U(\beta_0;\hat{\gamma},h) = \frac{1}{\sqrt{n}}U(\beta_0;\gamma_0,h) - \frac{1}{n}\frac{\partial U(\beta_0;\gamma,h)}{\partial\gamma}\,|_{\gamma^\circ}\left(\frac{1}{n}\frac{\partial U^\dagger(\gamma)}{\partial\gamma}\,|_{\gamma^\star}\right)^{-1}\frac{1}{\sqrt{n}}U^\dagger(\gamma_0)\;(9{\cdot}1)$$

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

Estimating function $\frac{1}{\sqrt{n}}U(\beta_0;\hat{\gamma},h)$ is asymptotically equivalent to

$$\begin{aligned}
\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\quad &\left\{\int_0^\tau w(t)\,[X_i(t) - av_2(X)(t;h)]\,[dM_i(t;\beta_0,\gamma_0,\mathcal{A}_0(\cdot),h) - \right. \\
&-\left(av_2(Y^\star)(t;h) - \beta_0^T av_2(X)(t;h)\right)\rho_i(t;\gamma_0,h)d\mathcal{M}_i(t;\gamma_0,\Lambda_0(\cdot))\big] - \\
&\left.- HA^{-1}\int_0^\tau [Z_i(t) - av_1(Z)(t;\gamma_0)]^T\,d\mathcal{M}_i(t;\gamma_0,\Lambda_0(\cdot))\right\},
\end{aligned}$$

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

Based on equation (9·1)

$$\begin{aligned}\sqrt{n}(\hat{\beta}(\hat{\gamma},h)-\beta_0) &= \left(-\frac{1}{n}\frac{\partial U(\beta;\hat{\gamma},h)}{\partial\beta}\mid_{\beta^\star}\right)^{-1}\frac{1}{\sqrt{n}}U(\beta_0;\hat{\gamma},h)\\ &= \left(-\frac{1}{n}\frac{\partial U(\beta;\hat{\gamma},h)}{\partial\beta}\mid_{\beta^\star}\right)^{-1}\times\left[\frac{1}{\sqrt{n}}U(\beta_0;\gamma_0,h)-\right.\\ &\quad\left.-\frac{1}{n}\frac{\partial U(\beta_0;\gamma,h)}{\partial\gamma}\mid_{\gamma^\circ}\left(\frac{1}{n}\frac{\partial U^\dagger(\gamma)}{\partial\gamma}\mid_{\gamma^\star}\right)^{-1}\frac{1}{\sqrt{n}}U^\dagger(\gamma_0)\right] \qquad (9{\cdot}2)\end{aligned}$$

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

$$\begin{aligned}&\frac{1}{\sqrt{n}}\sum_{i=1}^{n}D^{-1}\left[\int_0^\tau w(t)\left[X_i(t)-av_2(X)(t;h)\right]\left[dM_i(t;\beta_0,\gamma_0,\mathcal{A}_0(\cdot),h)-\right.\right.\\ &\left.-(av_2(Y^\star)(t;h)-\beta_0^T av_2(X)(t;h))\rho_i(t;\gamma_0,h)d\mathcal{M}_i(t;\gamma_0,\Lambda_0(\cdot))\right]-\\ &\left.-HA^{-1}\int_0^\tau\left[Z_i(t)-av_1(Z)(t;\gamma_0)\right]^T d\mathcal{M}_i(t;\gamma_0,\Lambda_0(\cdot))\right].\end{aligned}$$

It is a sum of mean zero i.i.d. random vectors for any function $h(X_i(\cdot))$. Using arguments similar to Liang & Zeger (1986), asymptotically we obtain equivalent expressions when using a random variable $\hat{\delta}$ instead of fixed δ_0 when talking about the stabilized estimator.

This plus consistency of $\hat{\beta}(\hat{\gamma},h)$ and of $\hat{D}$ yields that $\sqrt{n}\left(\hat{\beta}(\hat{\gamma},\hat{\delta})-\beta_0\right)$ is asymptotically normal with a consistent estimator of the asymptotic variance being $\hat{D}^{-1}\hat{V}\hat{D}^{-1}$ where we use $\hat{\delta}$ instead the unknown δ_0.

References

Bilias, Y., Gu, M., & Ying, Z. (1997). Toward a general asymptotic theory for cox model with staggered entry. *Annals of Statistics* **25**, 662–682.

Diggle, P. & Kenward, M. G. (1994). Informative drop-out in longitudinal data analysis. *Applied Statistics* **43**, 49–93.

Dominici, F. (2004). *Health effects of air pollution: statistical challenges, findings, and policy implications.* Department of Biostatistics, Bloomberg School of Public Health, Johns Hopkins University, http://www.biostat.jpsph.edu/fdominici/Stanfordtalk.pdf.

Grambsch, P. M. & Therneau, T. M. (1994). Proportinal hazards tests and diagnostics based on weighted residuals. *Biometrika* **81**, 515–526.

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.

Liang, K. Y. & Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika* **73**, 13–22.

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.

Lipsitz, S. R., Fitzmaurice, G. M., Ibrahim, J. G., Gelber, R., & Lipshultz, S. (2002). Parameter estimation in longitudinal studies with outcome-dependent follow-up. *Biometrics* **58**, 621–630.

Little, R. J. A. & Rubin, D. B. (2002). *Statistical analysis with missing data*. Wiley Series in Probability and Statistics.

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.

Pepe, M. S. & Couper, D. (1997). Modeling partly conditional means with longitudinal data. *Journal of the American Statistical Association* **92**, 991–998.

Pollard, D. (1990). *Empirical processes: theory and applications*. Hayward: Institute of Mathematical Statistics.

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.

Rotnitzky, A., Robins, J. M., & Scharfstein, D. O. (1998). Semiparametric regression for repeated outcomes with nonignorable nonresponse. *Journal of the American Statistical Association* **93**, 1321–1339.

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

Scharfstein, D. O., Rotnitzky, A., & Robins, J. M. (1999). Adjusting for nonignorable drop-out using semiparametric nonresponse models. *Journal of the American Statistical Association* **94**, 1096–1120.

Troxel, A. B., Lipsitz, S. R., & Harrington, D. P. (1998). Marginal models for the analysis of longitudinal measurements with nonignorable non-monotone missing data. *Biometrika* **85**, 661–672.

van der Vaart, A. W. (2000). *Asymptotic statistics*. Cambridge University Press.

Wu, M. C. & Carroll, R. J. (1988). Estimation and comparison of changes in the presence of informative right censoring by modeling the censoring process. *Biometrics* **44**, 175–188.