Statistica Sinica 1(1991), 19-32

REGRESSION MODELS FOR RIGHT TRUNCATED DATA WITH APPLICATIONS TO AIDS INCUBATION TIMES AND REPORTING LAGS

J. D. Kalbfleisch and J. F. Lawless

University of Waterloo

Abstract: This study of right truncated data was motivated by problems in which individuals can experience two events in time and where the distribution of the time between events, the lag, is of interest. If individuals come under observation at the occurrence of the second event and the time of the first event is retrospectively ascertained, the observed lags are right truncated. Two examples from data on AIDS and HIV infection are considered.

Regression models are specified which lead to simple tests and estimation of covariate effects based on right truncated data. The regression models are simply expressed in terms of a reverse time hazard function in both the discrete and continuous cases. In the continuous case, the proportional hazards model for the reverse time yields a power model for the c.d.f. and a simple parameter interpretation. A complementary log-log link in the discrete case admits the same interpretation. Inference techniques based on the full likelihood in the discrete case or the partial likelihood in the continuous case are developed and illustrated in the two examples mentioned above.

Key words and phrases: truncated data, reverse time hazards, regression models for truncated data.

1. Introduction

Truncated data arise in many contexts. This investigation, however, was motivated by problems in which individuals can experience two events. The first, termed the *initiating event*, occurs at random at time t, and the second, termed the *consequent event*, occurs at time $t + x \geq t$. The time x between the two events is termed the *lag*. We focus on situations in which individuals are observed at the occurrence of the consequent event and, at that time, the time of the initiating event is ascertained. Estimation of the distribution of the lag is of interest, but observations arise from truncated versions of this distribution.

Examples of this type of data arise in many contexts. In insurance applications a liability claim may arise as a consequence of an incident at time t. The

claim, however, is not reported to the insurer until time $t + x$. This is referred to as the incurred but not yet reported problem (see e.g. Kaminsky (1987)). Two examples discussed further in this paper arise in connection with Acquired Immune Deficiency Syndrome (AIDS). In the first, t represents the time of diagnosis of an AIDS case and $t + x$ is the time it is reported to the organization responsible for surveillance. Estimation of the distribution of the reporting lag x is important in estimating current numbers of cases (see e.g. Morgan and Curran (1986), Harris (1987), Karon et al. (1989), Brookmeyer and Damiano (1989), Zeger et al. (1989)). In the second example, t is the time of infection with the human immunodeficiency virus (HIV) that is a cause of AIDS, and $t + x$ is the time of diagnosis of AIDS. Estimation of the distribution of the incubation time x is important in estimating the number of infectives in the population. When the mode of infection is by blood transfusion, for example, the time of infection t is sometimes ascertainable when the case is diagnosed at $t + x$. In all of these examples, the lag data are right truncated in that individuals are observed only if they experience the consequent event prior to some specified calendar time.

In this paper, we consider a response X (e.g. a lag) and a covariate vector z; the distribution of X given z has c.d.f. $F(x|z) = Pr\{X \leq x|z\}$. We assume that there is a truncation mechanism operating such that each individual has a truncation time τ and (x, τ) is observed if $x \leq \tau$. It is assumed that the distribution of x given τ and z has c.d.f.

$$F(x|\tau, z) = \frac{F(x|z)}{F(\tau|z)}, \quad 0 \leq x \leq \tau. \tag{1}$$

The data consist of triples (x_i, τ_i, z_i), $i = 1, \ldots, n$, for n individuals for whom $x_i \leq \tau_i$. We assume that the process generating the observations is such that the pairs (x_i, τ_i), $i = 1, \ldots, n$, are conditionally independent given $z_1, \ldots, z_n$ and n.

Figure 1 illustrates the situation for initiating and consequent events with events observed if and only if $t + x \leq T$. Here an individual with initiating event at time t_i has $\tau_i = T - t_i$ and the data consist of (x_i, τ_i) for those individuals with consequent event prior to time T. Covariates may specify gender, age at the initiating event, time of the initiating event, etc. It is also convenient to consider a plot of x_i versus τ_i as in Figure 2 which illustrates the same data as Figure 1. This graphical display carries over to the more general situations in which truncated data occur.

One purpose of this paper is to discuss regression models and related semiparametric methods for right truncated data. A second objective is to develop tests concerning the independence of x and τ. Lynden-Bell (1971), Woodroofe (1985), Wang et al. (1986), Lagakos et al. (1988), Kalbfleisch and Lawless (1989b) and others have studied nonparametric estimation from truncated data in the

absence of covariates, and we review and consolidate important results in Section 2. Sections 3 and 4 introduce regression models for discrete and continuous time. In Section 5, quasi-stationarity of lag distributions is discussed, and the regression methods are used to assess the stationarity of AIDS reporting lag and incubation distributions.

2. Estimation of the Lag Distribution

Suppose that $x_1, \ldots, x_n$ are independent observations from the truncated distributions defined by $\tau_1, \ldots, \tau_n$ where, for the moment, we assume a homogeneous population with no covariates z_i. The likelihood function can be written as

$$\prod_{i=1}^{n} dF(x_i)/F(\tau_i) \tag{2}$$

and we consider nonparametric estimation of F.

As in Lagakos et al. (1988), it is convenient to define the reverse time hazard function (r.t.h.) as $g_x = Pr\{X = x | X \leq x\}$, $x = 0, 1, 2, \ldots$, in the discrete integer-valued case and

$$g(x) = \lim_{\Delta x \to 0} Pr\{X \in (x - \Delta x, x] | X \leq x\}/\Delta x \tag{3}$$

in the continuous case. The reverse time cumulative hazard can be defined as $G(x) = \int_x^\infty g(u)\, du$ or $G(x) = \sum_{u>x} g_u$. It can be seen that

$$\begin{aligned} F(x)/F(\tau) &= P_{(x,\tau]}[1 + dG(u)] \\ &= \lim \prod_{i=1}^{M} \{1 + [G(u_i) - G(u_{i-1})]\}, \end{aligned} \tag{4}$$

where $x = u_0 < u_1 < \ldots < u_M = \tau$ and the limit is taken as $M \to \infty$ and $\max(u_i - u_{i-1}) \to 0$. The product integral (4) thus defined reduces to

$$F(x)/F(\tau) = \exp\left\{-\int_x^\tau g(u)\, du\right\}, \quad 0 \leq x \leq \tau, \tag{5}$$

in the continuous case and

$$F(x)/F(\tau) = \prod_{u=x+1}^{\tau} (1 - g_u), \quad x = 0, 1, \ldots, \tau - 1, \tag{6}$$

in the discrete (integer valued) case. It follows immediately that, in either the discrete or continuous cases, the likelihood (2) can be rewritten as a function of

$dG(x), 0 \leq x \leq \tau^* = \max\{\tau_i\}$ and consequently only the reverse time hazards are strictly identifiable with truncated data.

If the data are discrete, the likelihood (2) can be written as

$$\prod_{i=1}^{n}\left\{g_{x_i}\prod_{u=x_i+1}^{\tau_i}(1-g_u)\right\} = \prod_{j=1}^{\tau^*}\{g_j^{d_j}(1-g_j)^{n_j-d_j}\}, \tag{7}$$

where $d_j = \sum I(x_i = j)$ and $n_j = \sum I(x_i \leq j \leq \tau_i)$. Note that d_j and n_j are respectively the numbers of observed responses exactly equal to j and the number of observed responses less than or equal to j where the truncation is such that a response of j could have been observed. It follows immediately that the m.l.e. satisfies

$$\hat{g}_j = d_j/n_j, \quad j \leq \tau^*, \tag{8}$$

and corresponding estimates of $F(x)/F(\tau)$ can be obtained from (6). The quantities d_j and n_j are illustrated in Figure 2.

There is a very strong analogy with nonparametric estimation of the survivor function based on left truncated data. In fact, by inverting the time scale this analogy can be further exploited (see, for example, Lagakos et al. (1988)). In the continuous case, the estimator of $-dG(x)$ analogous to the Nelson-Aalen estimator (see e.g. Andersen and Borgan (1985)) of the cumulative hazard function is

$$-d\hat{G}(x_j^*) = d_j^*/n_j^*, \tag{9}$$

where $x_1^* < x_2^* < \cdots < x_k^* \leq \tau^*$ are the distinct responses, $d_1^*, \ldots, d_k^*$ are the associated multiplicities and $n_j^* = \#\{i : x_i \leq x_j^* \leq \tau_i\}$. It follows from (4) that

$$\begin{aligned}\hat{F}(x)/\hat{F}(\tau) &= P_{(x,\tau]}[1 + d\hat{G}(u)] \\ &= \prod_{j:x_j^*>x}(1 - d_j^*/n_j^*).\end{aligned} \tag{10}$$

The discrete result above is a straightforward extension of (12) in Kalbfleisch and Lawless (1989b). Continuous versions have been given by Lagakos et al. (1988), Lynden-Bell (1971), Woodroofe (1985) and Wang et al. (1986). Harris (1987) discussed an algorithm to compute estimates which give a special case of (8) but did not notice the existence of a closed form solution. Similarly, Brookmeyer and Damiano (1989) used a special case of (2) but did not notice that closed form m.l.e.'s can be obtained.

As in Lagakos et al. (1988) and Keiding and Gill (1990), asymptotic results for the continuous case can be obtained by appealing to results from counting

processes and martingales. In so doing, it is necessary to consider the process in reverse time so that one considers the counting process

$$N^*(u) = \sum I(x_i \geq \tau^* - u), \quad 0 \leq u \leq \tau^*.$$

There is, then, a formal equivalence between right truncation as here and results obtained in the context of a standard survival model with left truncation.

3. Discrete Time Regression Models

Consider now the discrete time model where x and τ are integer valued and suppose that, given covariates z, the r.t.h. is $g(x|z) = Pr\{X = x|X \leq x, z\}$, $x = 0, 1, 2, \ldots$. A natural class of regression models to consider is of the form

$$\psi(g(x|z)) = \psi(g_0(x)) + z'\beta, \quad x = 1, 2, \ldots, \tag{11}$$

where ψ is a specified 1:1 map of [0,1] onto $[-\infty, \infty]$ and z and β are column vectors of covariates and regression parameters respectively. This model is analogous to those used in the analysis of right censored discrete survival data (e.g. Lawless (1982), §7.3.2). The baseline r.t.h., $g_0(x)$, is left arbitrary here but could be modeled further.

Two natural choices for $\psi(\cdot)$ are the logistic,

$$\psi(u) = logit(u) = \log\left(\frac{u}{1-u}\right), \tag{12}$$

and the complementary log-log,

$$\psi(u) = \log\{-\log(1-u)\}. \tag{13}$$

For the latter model, the c.d.f. of x given z can be written as

$$F(x|z) = F_0(x)^{\exp(z'\beta)}, \tag{14}$$

where $F_0(x)$ is the baseline c.d.f. (at $z = 0$) corresponding to $g_0(x)$,

$$F_0(x) = \prod_{u=x+1}^{\infty} \{1 - g_0(u)\},$$

and this allows a simple interpretation of β in terms of its effect on the c.d.f. The same relationship (14) also holds for truncated c.d.f.'s in that

$$F(x|z)/F(\tau|z) = \{F_0(x)/F_0(\tau)\}^{\exp(z'\beta)}.$$

For (12), on the other hand, there appears to be no similar interpretation in terms of c.d.f.'s.

For the model (11) with $\psi(\cdot)$ given by (12), there exists a partial likelihood for β analogous to that given by Cox (1972, 1975). Specifically, components of the partial likelihood are constructed by considering, at each time u, the conditional probability that the observed individuals fail given the number of failures at that time and the set of individuals that fail at times less than or equal to u and have truncation times of at least u. The partial likelihood is obtained as a product over $u = \tau^*, \tau^* - 1, \ldots, 1$. In the next section, the construction of the partial likelihood in a continuous model is considered. The extension to the discrete case is straightforward. As noted above, however, the model (11) with $\psi(\cdot)$ given by (13) is more satisfactory from the viewpoint of parameter interpretation, but for this model no partial likelihood exists. In what follows, therefore, we consider the full likelihood as providing the basis for tests and estimation with general $\psi(\cdot)$.

Let $\chi = \psi^{-1}$ and denote $\psi(g_0(x)) = \theta_x$ and $\theta = (\theta_1, \ldots, \theta_{\tau^*})$. Then, the likelihood based on data (x_i, τ_i, z_i), $i = 1, \ldots, n$, is

$$\begin{aligned} L(\theta, \beta) &= \prod_{i=1}^{n} g(x_i|z_i) \prod_{u=x_i+1}^{\tau_i} \{1 - g(u|z_i)\} \\ &= \prod_{u=1}^{\tau^*} \left\{ \prod_{j \in D_u} g(u|z_j) \prod_{j \in R_u - D_u} [1 - g(u|z_j)] \right\} \\ &= \prod_{u=1}^{\tau^*} \left\{ \left[\prod_{j \in D_u} \chi_{uj} \prod_{j \in R_u - D_u} (1 - \chi_{uj}) \right] \right\}, \end{aligned}$$

where $D_u = \{i : x_i = u\}$, $R_u = \{i : x_i \leq u \leq \tau_i\}$, and $\chi_{uj} = \chi(\theta_u + z_j'\beta)$. It follows that the score vector has components

$$\frac{\partial}{\partial \theta_u} \log L = \sum_{j \in D_u} \frac{\chi'_{uj}}{\chi_{uj}(1 - \chi_{uj})} - \sum_{j \in R_u} \frac{\chi'_{uj}}{1 - \chi_{uj}}, \quad u = 1, \ldots, \tau^*,$$

and

$$\frac{\partial}{\partial \beta} \log L = \sum_{u=1}^{\tau^*} \sum_{j \in D_u} \frac{z_j \chi'_{uj}}{\chi_{uj}(1 - \chi_{uj})} - \sum_{j \in R_u} \frac{z_j \chi'_{uj}}{1 - \chi_{uj}},$$

where $\chi'_{uj} = \partial \chi(\theta_u + z_j'\beta)/\partial \theta_u$. Lawless (1982, §7.3.2) and Kalbfleisch and Prentice (1980, §4.6) discuss joint estimation of θ, β by solving the associated m.l. equations in a related problem.

Simple tests of the hypothesis $\beta = 0$ can be based on the β score evaluated at $\beta = 0$ and the m.l.e. of θ at $\beta = 0$. This can be shown to give

$$U = \frac{\partial}{\partial \beta} \log L|_{\beta=0,\hat{\theta}(0)}$$
$$= \sum_{u=1}^{\tau^*} \hat{w}(u) \sum_{j \in D_u} (z_j - \bar{z}_u), \tag{15}$$

where $\hat{w}(u) = \hat{\chi}'_u/\{\hat{\chi}_u(1 - \hat{\chi}_u)\psi'(\hat{\chi}_u)\}$, $\hat{\chi}_u = d_u/n_u = \chi(\hat{\theta}_u(0))$, $\hat{\chi}'_u = \chi'_{uj}|_{\beta=0,\hat{\theta}(0)}$, and $\bar{z}_u = \sum_{j \in R_u} z_j/n_u$. If ψ is given by (12), $\hat{w}(u) = 1$, whereas if ψ is given by (13), $\hat{w}(u) = \log(1 - \hat{\chi}_u)/\hat{\chi}_u$. The variance of U can be evaluated by obtaining the Fisher information matrix and determining the appropriate submatrix in its inverse. An alternative and very simple approach is to use

$$V = \sum_{u=1}^{\tau^*} \hat{w}(u)^2 \text{var} \left\{ \sum_{j \in D_u} (z_j - \bar{z}_u) | n_u, R_u, \beta = 0 \right\} |_{\hat{\theta}(0)}$$
$$= \sum_{u=1}^{\tau^*} \hat{w}(u)^2 \frac{d_u(n_u - d_u)}{n_u(n_u - 1)} \sum_{j \in R_u} (z_j - \bar{z}_u)(z_j - \bar{z}_u)'. \tag{16}$$

The variance computation in the uth term of (16) is conditional upon the set R_u and the fact that exactly d_u of them fail. A statistic which can be used to assess the hypothesis $\beta = 0$ is provided by

$$U'V^-U \tag{17}$$

which has an asymptotic chi-squared distribution with degrees of freedom given by the rank of V.

Special cases of (15), (16) and (17) are easily obtained. Consider, for example, the $(m + 1)$-sample test where $z_{ij} = 1$ if the ith item is in the jth sample, and 0 otherwise, $j = 1, \ldots, m$. In this case, the statistic U has rth element

$$U_r = \sum_{u=1}^{\tau^*} \hat{w}(u)\{d_{ru} - n_{ru}d_u/n_u\}, \quad r = 1, \ldots, m, \tag{18}$$

where $d_{ru} = \#\{i : x_i = u, z_{ir} = 1\}$ and $n_{ru} = \#\{i : x_i \leq u \leq \tau_i, z_{ir} = 1\}$. The matrix V has r, s element

$$V_{rs} = \sum_{u=1}^{\tau^*} \hat{w}(u)^2 \frac{d_u(n_u - d_u)}{n_u(n_u - 1)}\{\delta_{rs}n_{ru} - n_{ru}n_{su}/n_u\}, \tag{19}$$

where $\delta_{rs} = I(r = s)$. The test statistic (17) has m degrees of freedom.

In the context of initiating and consequent events described in Section 1, hypotheses of stationarity for the lag distribution are of importance. To examine stationarity, we might identify components of z with intervals of time for the initiating events. This is discussed in Section 5.

4. Continuous Regression Models

With continuous right truncated data, it is natural to consider a proportional relationship for the reverse time hazards. Thus, we suppose that the r.t.h. satisfies the relationship

$$g(x; z) = g_0(x) \exp(z'\beta) \tag{20}$$

which corresponds to a power model for the c.d.f. of the response x,

$$F(x|z) = F_0(x)^{\exp(z'\beta)},$$

where $F_0(x) = \exp\{-\int_x^\infty g_0(u)\,du\}$.

Let $N_i^*(u) = I(x_i \geq \tau^* - u)$ and $Y_i(u) = I(x_i \leq \tau^* - u < \tau_i)$. Let $N^*(u) = (N_1^*(u), \ldots, N_n^*(u))$ and $Y(u) = (Y_1(u), \ldots, Y_n(u))$. A partial likelihood for β can be based on $Pr\{dN^*(u) | \sum dN_i^*(u), \mathcal{F}_{u-}\}$ where $\mathcal{F}_u = \sigma\{N^*(s), Y(s) : 0 \leq s \leq u\}$ is a nested sequence of σ-fields that specifies the sequence of responses and truncations that equal or exceed $x^* - u$. The approximate partial likelihood for β is

$$L(\beta) = \prod_{h=1}^{m} \frac{\exp(s_h'\beta)}{\{\sum_{l \in R_h} \exp(z_l'\beta)\}^{d_h^*}}, \tag{21}$$

where $0 < x_1^* < \cdots < x_m^* < \tau^*$ are the observed responses with multiplicities $d_1^*, \ldots, d_m^*$, $R_h = \{i : Y_i(x_h^*) = 1\}$, $s_h = \sum_{i \in D_h} z_i$ and $D_h = \{i : dN_i^*(x_h^*) = 1\}$. The expression (21) is an exact partial likelihood if $d_h^* = 1$, $h = 1, \ldots, m$ and $m = n$, and is precisely of the form of the usual partial likelihood in the proportional hazards model. If ties are numerous, as in the first example of Section 5, analyses are better based on the discrete model (11). Asymptotic properties of the partial likelihood (21) can be derived by arguments analogous to those which can be found in the literature on the analysis of failure time data (for review see Andersen and Borgan (1985)).

The score statistic for a test of $\beta = 0$ is

$$\frac{\partial \log L(\beta)}{\partial \beta}\Big|_{\beta=0} = \sum_{h=1}^{m} \sum_{j \in D_h} (z_j - \bar{z}_h), \tag{22}$$

where $\bar{z}_h = \sum_{j \in R_h} z_j / n_h^*$ and $n_h^* = |R_h|$. This is of the same form as the statistic (15) in the discrete model with $\hat{w}(u) = 1$ or $\psi(\cdot)$ given by (12). An estimate of the variance of (22) can be obtained from the observed information matrix based on (21) at $\beta = 0$. Alternatively, (16) gives a valid estimate of the variance of (22) and coincides with the information when all d_u^*'s are 1.

Again, in the context of initiating and consequent events, the arguments above have simple interpretations from reversing the time scale. Lagakos et al. (1988) exploit this and develop two-sample rank tests based on continuous data.

5. Testing Quasi-Stationarity with Application to AIDS Reporting Lags and Incubation Times

In some contexts it is of considerable interest whether the response x and truncation variable τ are independent. In the context of initiating and consequent events, for example, the question arises as to whether the lag distribution is stationary over the interval $[a_1, a_2]$ for the initiating event. That is, is $f(x|t) = f(x)$ independent of t for $t \in [a_1, a_2]$? When the observations x_i are truncated at $\tau_i = T - t_i$, all that we can examine is whether the r.t.h. is independent of t. That is, we consider

$$g(x|t) = g(x), \quad x \leq T - t, \; t \in [a_1, a_2]. \tag{23}$$

The property (23) we term *quasi-stationarity* over $[a_1, a_2]$; stationarity of $f(x|t)$ over $[a_1, a_2]$ implies (23), but (23) obviously does not imply stationarity.

Tests for quasi-stationarity can be obtained by using the regression models of Sections 3 and 4 and allowing the covariate z to incorporate functions of t. First we note, however, that in the discrete case an omnibus test of (23) versus a general alternative, $g(x|t) = g_t(x)$, $x \leq T - t$, $t \in [a_1, a_2]$ can be obtained by using the likelihood ratio statistic from the likelihood (7). This is

$$\Lambda = 2 \sum_{t=a_1}^{a_2} \sum_{x=0}^{T-t} d_{tx} \log(d_{tx}/e_{tx}), \tag{24}$$

where $d_{tx} = \sum I(t_i = t, x_i = x)$ and

$$e_{tx} = d_t \{ \hat{f}(x) / \hat{F}(T - t) \}, \tag{25}$$

where $d_t = \sum_{x=0}^{T-t} d_{tx}$ and the bracketed term in (25) is obtained from (6) and (8). If the d_x's are large and $f(x) > 0$ for $x = 0, 1, \ldots, \tau^*$, Λ is approximately χ^2 with degrees of freedom $(a_2 - a_1)(2T - a_1 - a_2 - 1)/2$. We note that for the case where $a_1 = 1$ and $a_2 = T$ the test above is just the $(T - 1)(T - 2)/2$

degrees of freedom test for quasi-independence in a triangular contingency table (cf. Bishop et al. (1975), p.187ff).

Omnibus statistics like (24) are generally not very powerful. If systematic differences among truncated distributions and the reverse time hazard functions are expected, more sensitive tests may be developed by using the regression models of Sections 3 and 4, provided that the types of departures from quasi-stationarity that are envisaged are reasonably well represented by models of the form (11) or (20) with covariates suitably defined as functions of t. We illustrate by considering the two problems associated with the AIDS data mentioned in Section 1.

Example 1: Report Lags for TA AIDS Cases.

Between April 1983 and June 1988, 1760 cases of transfusion-associated (TA) AIDS cases had been reported to the Centres for Disease Control (CDC) in Atlanta, Georgia. To assess stationarity of the reporting lag distribution, we compared the lag times of the 526 patients diagnosed before March 1986 to the lag times of the 1,234 patients diagnosed during or after March 1986. We used $z_i = I(\text{March 1986} \leq t_i \leq \text{June 1988})$ with discrete time regression models (11), since report lags are recorded in months and there are many values with large multiplicities. This yields the score statistic (18) and variance estimate (19) corresponding to $m = 1$; this statistic has also been considered by Lagakos et al. (1988). The logit model (12) gives $\hat{w}(u) = 1$ and an observed value of the test statistic (17) of $(6.77)^2$. The log-log model (13) gives $\hat{w}(u) = (n_u/d_u)\log(1 - d_u/n_u)$ and an observed value for (17) of $(7.04)^2$. Both tests provide a strong indication of non-stationarity with larger reverse time hazards for cases diagnosed later. This implies smaller conditional c.d.f.'s for this group or in other words, a tendency toward longer reporting lags for the more recently diagnosed cases. Figure 3 gives plots of estimated conditional c.d.f.'s $F(x)/F(26)$ for the two groups (see Section 2) where $x = 0, 1, \ldots, 26$ represents the reporting lag in months. (Note that g_x for $x = 1, \ldots, 26$ is all that is estimable for the cases diagnosed between April 1986 and June 1988). The smaller conditional c.d.f. for later diagnoses is apparent in Figure 3. We remark, in passing, that report lags for these transfusion-associated AIDS cases tend to be somewhat longer than for the set of all AIDS cases that are reported to the CDC; see Brookmeyer and Damiano (1989, Table 1) for some estimates based on all cases.

Example 2: Incubation Distributions for TA AIDS Cases.

Of the 1760 cases reported and used in Example 1 above, 976 cases had unambiguous information on date of HIV infection by transfusion and these cases are used for this example. Observations on the incubation time are truncated by $\tau_i = T - t_i - r_i$ where T represents June 30, 1988, t_i is the date of infection and

r_i is the reporting lag. All times are measured in months. The validity of the truncation model discussed here requires that the reporting lag be stationary and we note that Example 1 provides evidence that this assumption is not entirely satisfactory. We discuss this point further below.

We consider continuous time models of the form (20) with (a) $z = t$ and (b) $z = \log(t + 1)$, where t is the time elapsed from the start of 1978. Since times are recorded in months, statistics based on (21), which include adjustments for ties, are used. Use of a partial likelihood based on the discrete time model (11) with ψ given by (12) yields the same results.

Earlier analyses of these data indicate possible differences in incubation distributions according to the age of the patient (cf. Kalbfleisch and Lawless (1989 a,b)). We therefore tested quasi-stationarity separately for four age groups: 0-12 years of age, 13-49 years of age, 50-69 years of age, and 70 years of age and over, where "age" refers to age at the time of HIV infection. The results are summarized in Table 1, which shows the approximate standard normal statistics based on (22). Only in the age group 50-69 is there evidence of non-stationarity and this is in the direction of a tendency toward shorter incubation times for the more recently infected patients. This is in agreement with the analysis of Tsai (1990), who uses a different kind of test to assess stationarity.

Table 1. Standardized score statistics for quasi-stationarity of TA AIDS incubation distributions

Age Group	(a)$z = t$	(b)$z = \log(t + 1)$
0-12($n = 86$)	$-.45$	$-.74$
13-49($n = 335$)	-1.17	-1.56
50-69($n = 463$)	-2.57	-3.74
$\geq$ 70($n = 92$)	-1.54	-1.81

We reiterate our earlier comment that our analysis assumes that the reporting lag is stationary. Example 1 provided evidence to the contrary, and it is possible that the tendency toward shorter incubation times seen here is in fact a reflection of the non-stationarity in the reporting lag distribution; a tendency toward longer reporting lags in later diagnoses would bias estimates of incubation time probabilities for these patients down somewhat. More complex modelling and an analysis that allows for simultaneous non-stationarity of both incubation times and reporting lags can be based on the multiplicative Poisson models described by Kalbfleisch and Lawless (1989a; 1989b, Section 6.1). The regression methods of the present paper do not lend themselves to detailed analyses involv-

ing more than two types of events or investigations of the non-stationarity of two or more lag distributions simultaneously.

6. Concluding Remarks

For problems involving right truncated data, regression analysis is conveniently approached through models based on reverse time hazard functions. The methods introduced here allow the easy comparison of truncated distributions for different populations. One important application is the assessment of quasi-stationarity of lag distributions which arise in connection with pairs of events, as described in Section 1. The fact that the regression models are based on reverse time hazards rather than the usual forward time hazards commonly used with lifetime data does not seem to us a drawback. The latter models do not lead to simple analyses of right truncated data. The former do, and they often provide flexible and realistic representations of covariate effects for truncated distributions; with proportional r.t.h.'s, they also yield simple interpretations for regression parameters.

Acknowledgements

We wish to thank a referee for very helpful comments on an earlier version of this paper. This research was supported by grants from the Natural Sciences and Engineering Research Council of Canada to both authors, and grant number 1-R01-DA 04722 from the National Institute on Drug Abuse, co-ordinated by the Societal Institute of the Mathematical Sciences.

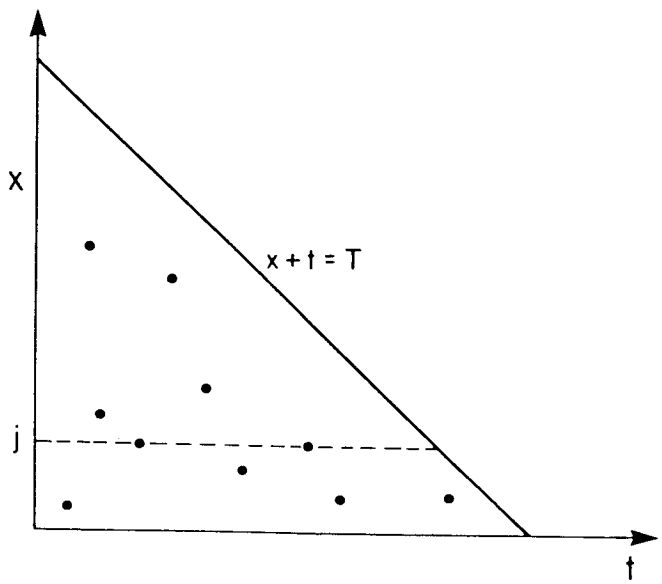

Figure 1. Times of initiating event (t) and lag (x) with truncation at $t + x = T$.

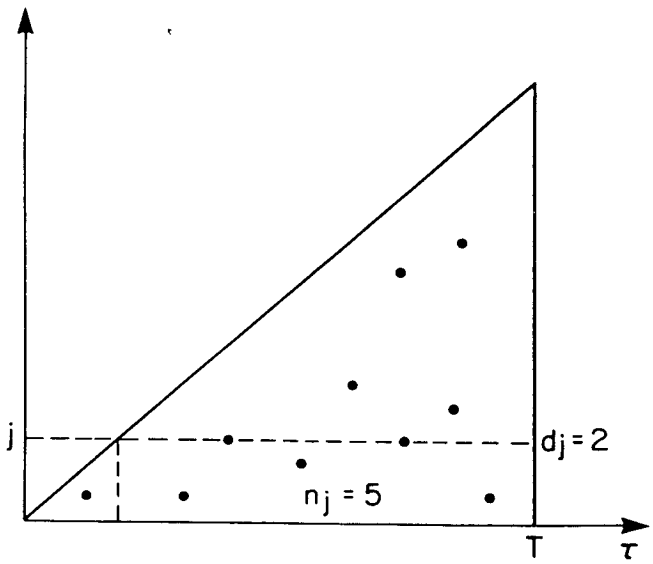

Figure 2. Lag time (x) plotted against corresponding truncation time (τ).

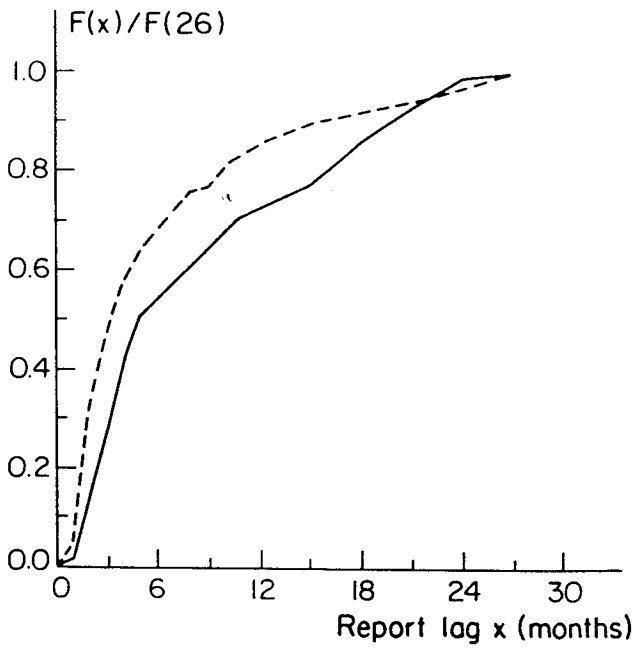

Figure 3. Estimates of TA AIDS truncated report lag distributions.

Cases diagnosed between April 1983 and March 1986 - - - - -;
Cases diagnosed after March 1986 ————————.

References

Andersen, P. K. and Borgan, D. (1985). Counting process models for life history data: A review (with discussion). *Scand. J. Statist.* **12**, 97-158.

Bishop, Y. M. M., Fienberg, S. E. and Holland, P. W. (1975). *Discrete Multivariate Analysis.* MIT Press, Cambridge, Mass.

Brookmeyer, R. and Damiano, A. (1989). Statistical methods for short-term projection of AIDS incidence. *Statistics in Medicine* **8**, 23-34.

Cox, D. R. (1972). Regression models and life-tables (with discussion). *J. Roy. Statist. Soc. Ser. B* **34**, 187-220.

Cox, D. R. (1975). Partial likelihood. *Biometrika* **62**, 269-276.

Harris, J. E. (1987). Delay in reporting Acquired Immune Deficiency Syndrome (AIDS). National Bureau of Economic Research Working Paper No. 2278.

Kalbfleisch, J. D. and Lawless, J. F. (1989a). Estimating the incubation time distribution and expected number of cases for transfusion-associated Acquired Immune Deficiency Syndrome. *Transfusion* **29**, 672-676.

Kalbfleisch, J. D. and Lawless, J. F. (1989b). Inference based on retrospective ascertainment: An analysis of the data on transfusion-related AIDS. *J. Amer. Statist. Assoc.* **84**, 360-372.

Kalbfleisch, J. D. and Prentice, R. L. (1980). *The Statistical Analysis of Failure Time Data.* John Wiley, New York.

Kaminsky, K. S. (1987). Prediction of IBNR claim counts by modelling the distribution of report lags. *Insurance Math. Econom.* **6**, 151-159.

Karon, J. M., Devine, D. J. and Morgan, W. M. (1989). Predicting AIDS incidence by extrapolating from recent trends. Unpublished MS.

Keiding, N. and Gill, R. (1990). Random truncation models and Markov processes. *Ann. Statist.* **18**, 582-602.

Lagakos, S. W., Barraj, L. M. and De Gruttola, V. (1988). Nonparametric analysis of truncated survival data, with application to AIDS. *Biometrika* **75**, 515-523.

Lawless, J. F. (1982). *Statistical Models and Methods for Lifetime Data.* John Wiley, New York.

Lynden-Bell, D. (1971). A method of allowing for known observational selection in small samples applied to 3CR quasars. *Monthly Notices Roy. Astronom. Soc.* **155**, 95-118.

Morgan, W. H. and Curran, J. W. (1986). Acquired immunodeficiency syndrome: current and future trends. *Public Health Reports* **101**, 459-465.

Tsai, W. Y. (1990). Testing the assumption of independence between truncated time and failure time. *Biometrika* **77**, 169-178.

Wang, M. C., Jewell, N. P. and Tsai, W. Y. (1986). Asymptotic properties of the product limit estimate under random truncation. *Ann. Statist.* **14**, 1597-1605.

Woodroofe, M. (1985). Estimating a distribution function with truncated data. *Ann. Statist.* **13**, 163-177.

Zeger, S. L., Lee, L. C. and Diggle, P. J. (1989). Statistical methods for monitoring the AIDS epidemic. *Statistics in Medicine* **8**, 3-21.

Department of Statistics and Actuarial Science, University of Waterloo, Waterloo, N2L 3G1, Canada.

(Received October 1989; accepted June 1990)