AUSTRIAN JOURNAL OF STATISTICS
Volume 35 (2006), Number 2&3, 143–155

Regression Model Fitting for the Interval Censored 1 Responses

Hira L. Koul and Tingting Yi
Michigan State University, U.S.A.

Abstract: In the interval censored case 1 data, an event occurrence time is unobservable, but one observes an inspection time and whether the event has occurred prior to this time or not. Such data is also known as the interval censored case 1 data. It is of interest to assess the effect of a covariate on the event occurrence time variable. This note constructs tests for fitting a class of parametric regression models to the regression function of the log of the event occurrence time variable when the data are interval censored case 1 and when the error distribution is known. These tests are based on a certain martingale transform of a marked empirical process. They are asymptotically distribution free in the sense that their asymptotic null distributions neither depends on the null model nor on any of the distributions of the covariate, the inspection time or error variables. However, the test statistic itself depends on the error distribution. Some simulation studies assessing some finite sample level and power behavior of some of the proposed tests are also given.

Keywords: Marked Empirical Process.

1 Introduction

The focus of this paper is to develop tests of lack-of-fit of a regression model when the response variable is subject to interval censoring case 1. This kind of data occurs frequently in clinical trials and longitudinal studies. For example, a patient is given a diagnostic test to detect whether the patient has the disease or not. In this case the time of the onset of the disease Y^0 is un-observable. If the disease is found to be present then one only knows that $Y^0 \leq T^0$, where T^0 is the time the test is administered. In other words in this situation one observes (δ, T^0) where $\delta := I(Y^0 \leq T^0)$, with $I(A)$ denoting the indicator of the event A. This type of data is also known as the current status data.

Hoel and Walburg (1972), Finkelstein and Wolfe (1985), Finkelstein (1986), Diamond et al. (1986), Diamond and McDonald (1991), Keiding (1991), among others, contain several examples of interval censoring case 1 data sets from clinical, tumorigenicity and demographic studies. The recent review article by Jewell and Laan (2004) contains some additional applications to health related studies.

Now suppose one is interested in assessing the effect of a covariate Z on the time of the onset of the disease, e.g., age of the patient. One way to proceed is to use the classical regression analysis where one regresses $Y := \log Y^0$ on Z but one observes only (δ, T) with $T = \log T^0$. But then the question of which regression model to choose from a possible class of given models becomes relevant.

More precisely, assume Y has finite expectation and let $\mu(z) := \mathrm{E}(Y|Z = z)$ denote the regression function. Let $\mathcal{M} = \{m_\theta(z) : z \in \mathbb{R}\,,\ \theta \in \Theta\}$ be a given parametric family

of functions, where Θ is a subset of the q-dimensional Euclidean space $\mathbb{R}^q$. This class of functions represents a possible class of regression models and the problem of interest is to test the hypothesis

$$H_0 : \mu(z) = m_{\theta_0}(z)\,, \quad \text{for some } \theta_0 \in \Theta\,, \forall z \in \mathbb{R}\,,$$

based on n i.i.d. observations $X_i = (\delta_i, T_i, Z_i)$, $1 \le i \le n$ on (δ, T, Z), where $\delta = I(Y \le T)$. The alternative of interest is that H_0 is not true.

In the case Y_i's are fully observable tests for the lack-of-fit hypothesis H_0 have been based on the marked residual empirical process

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{Y_i - m_{\theta_n}(Z_i)}{\hat{\sigma}_n(Z_i)}\, I(Z_i \le z)\,, \quad z \in \mathbb{R}\,, \tag{1}$$

where θ_n is a $\sqrt{n}$-consistent estimator of θ_0 under the null hypothesis, and $\hat{\sigma}_n^2(z)$ is a consistent estimator of the conditional variance of $Y - m_\theta(Z)$, given $Z = z$, under H_0, cf., An and Cheng (1991), Stute (1997), and Stute et al. (1998), among others. The latter paper shows that the tests based on its innovation martingale transforms are asymptotically distribution free.

This paper develops an analog of this transformation for the current status response data when the distribution Z is not known but that of the error variable and T is known. Even in this case no rigorous testing procedure is available in the literature at the present time. The next section discusses the main results about testing a simple hypothesis, the composite hypothesis H_0 and the needed assumptions. We also briefly discuss a modification of the proposed test when the distribution of T is not known. Section 3 contains a simulation study illustrating the finite sample behavior of level and power of some proposed tests.

2 Main Results

In this section we shall construct an analog of the process given at (1) that is suitable for the current status data and that will be useful for testing simple and composite hypotheses discussed in subsections 2.1 and 2.2, respectively.

Since the Y_i's are not observable, we need to replace them in (1) by $\hat{Y}_i$, a copy of

$$Y^* = \mathrm{E}(Y|\delta, T, Z) = \delta \mathrm{E}(Y|\delta, T, Z) + (1-\delta)\mathrm{E}(Y|\delta, T, Z)\,. \tag{2}$$

To proceed further, let F denote the d.f. of the error $\varepsilon := Y - \mu(Z)$. Assume that

$$\begin{aligned} &F \text{ is continuous, } 0 < F(y) < 1, \text{ for all } y \in \mathbb{R}, \mathrm{E}\varepsilon = 0, \mathrm{E}\varepsilon^2 < \infty. \\ &\varepsilon \text{ is conditionally independent of } T, \text{ given } Z, T \text{ and } \varepsilon \text{ are independent of } Z. \end{aligned} \tag{3}$$

Then, with $\bar{F} := 1 - F$, we obtain,

$$\mathrm{E}(Y|\delta = 1, T = t, Z = z) = \frac{\int_{-\infty}^{t} y dF(y - \mu(z))}{F(t - \mu(z))} \tag{4}$$

$$
\begin{aligned}
&= \frac{\int_{-\infty}^{t-\mu(z)} y dF(y)}{F(t-\mu(z))} + \mu(z)\,, \\
\mathrm{E}(Y|\delta=0, T=t, Z=z) &= \frac{\int_{t}^{\infty} y dF(y-\mu(z))}{1-F(t-\mu(z))} \\
&= \frac{\int_{t-\mu(z)}^{\infty} y dF(y)}{\bar F(t-\mu(z))} + \mu(z)\,, \qquad t, z \in \mathbb{R}\,.
\end{aligned}
$$

Let

$$
\begin{aligned}
R(d,t,z) &:= \mathrm{E}(Y|\delta=d, T=t, Z=z) - \mu(z)\,, \\
\bar F(y) &:= 1-F(y)\,, \quad L(y) := F(y)\bar F(y)\,, \\
\nu(y) &:= \int_{-\infty}^{y} x dF(x)\,, \quad y \in \mathbb{R}\,, \\
\sigma^2(z) &:= Var(R(\delta,T,Z)|Z=z)\,, \qquad d=0,1;\ t,z \in \mathbb{R}\,.
\end{aligned}
$$

From (2), (4), and the fact $\nu(\infty)=0$, we obtain

$$
\begin{aligned}
R(d,t,z) &= d\,\frac{\int_{-\infty}^{t-\mu(z)} y dF(y)}{F(t-\mu(z))} + (1-d)\,\frac{\int_{t-\mu(z)}^{\infty} y dF(y)}{\bar F(t-\mu(z))} \\
&= \nu(t-\mu(z))\left[\frac{d}{F(t-\mu(z))} - \frac{1-d}{\bar F(t-\mu(z))}\right] \\
&= \frac{\nu(t-\mu(z))[d-F(t-\mu(z))]}{L(t-\mu(z))}\,.
\end{aligned}
$$

By the conditional independence of ε and T, given Z, and the independence of T and Z, $\mathrm{E}\{R(\delta,T,Z)|Z\}=0$ and

$$
\begin{aligned}
0 < \sigma^2(z) = \mathrm{E}\left\{\frac{\nu^2(T-\mu(z))}{L(T-\mu(z))}\right\} < \infty\,, \qquad \forall z \in \mathbb{R}\,, \\
\mathrm{E}\sigma^2(Z) < \infty\,, \qquad \text{by the assumption } \mathrm{E}\varepsilon^2 < \infty.
\end{aligned}
$$

The entities $R(\delta_i,T_i,Z_i)/\sigma(Z_i)$ play the role of the standardized residuals in the current status data.

2.1 Tests of a Simple Hypothesis

To test the simple hypothesis $\tilde H_0$: $\mu(z)=\mu_0(z)$, $z \in \mathbb{R}$, where μ_0 is a known function, the analogue of the process (1) suitable here would be

$$
V_n^0(z) := \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{R_0(\delta_i,T_i,Z_i)}{\sigma_0(Z_i)}\, I(Z_i \le z)\,, \qquad z \in \mathbb{R}\,,
$$

where R_0, σ_0 are the above R, σ functions with μ replaced by μ_0.

Let Ψ and G denote the d.f.'s of T and Z, respectively, and B denote the standard Brownian motion on $[0,\infty)$. Using an argument similar to the one used in Stute

(1997), it can be verified that under (3) and $\tilde{H}_0$, V_n^0 converges weakly to $B \circ G$, in $D[-\infty, \infty]$ and uniform metric. Thus, for example the test that would reject $\tilde{H}$ whenever $K_n := \sup_{z\in\mathbb{R}} |V_n^0(z)| > b_\alpha$, where b_α is the $100(1-\alpha)$th percentile of the distribution of $\sup_{0\le t\le 1} |B(t)|$, would have the asymptotic size α.

Consistency. Let $\mu_1(z)$ be an alternative regression function and consider the problem of testing the simple hypothesis $\tilde{H}$: $\mu(z) = \mu_0(z)$ against the alternative $\tilde{H}_1$: $\mu(z) = \mu_1(z)$, for all z. Let

$$\begin{aligned} w(t,z) &:= \frac{1}{\sigma_0(z)} \frac{\nu(t-\mu_0(z))}{L(t-\mu_0(z))}\,, \\ \Delta(t,z) &:= F(t-\mu_1(z)) - F(t-\mu_0(z))\,, \quad t,z \in \mathbb{R}\,. \end{aligned}$$

Then one can rewrite $V_n^0(z) = V_n^1(z) + n^{1/2} D_n(z)$, where

$$\begin{aligned} V_n^1(z) &:= \frac{1}{\sqrt{n}} \sum_{i=1}^n w(T_i, Z_i) \left[\delta_i - F(T_i - \mu_1(Z_i))\right] \, I(Z_i \le z) \\ D_n(z) &:= \frac{1}{n} \sum_{i=1}^n w(T_i, Z_i) \Delta(T_i, Z_i) \, I(Z_i \le z)\,. \end{aligned}$$

Let P_1 and E_1 signify the probability measure and expectation under H_1. Observe that the distributions of T and Z are not affected by the choice of the regression function. Assume that F and μ_1 satisfy

$$\text{(a)} \quad \mathrm{E}\left(w(T,Z)|\Delta(T,Z)|\right) < \infty\,, \qquad \text{(b)} \quad G(\mu_1(Z) \ne \mu_0(Z)) > 0\,. \tag{5}$$

In view of this fact, (5(b)), and F strictly increasing on $\mathbb{R}$ we obtain that $\mathrm{E}D_n(z) = \mathrm{E}w(T,Z)\Delta(T,Z)\, I(Z \le z) \ne 0$, for at least one z. Hence, by the classical Glivenko-Cantelli type argument, $\sup_z |D_n(z)| \to \sup_z |\mathrm{E}w(T,Z)\Delta(T,Z)I(Z \le z)|$, a.s., which is not equal to zero.

Next, assume additionally that μ_1 satisfies

$$0 < \mathrm{E}_1 w^2(T,Z) \left[\delta - F(T - \mu_1(Z))\right]^2 < \infty\,. \tag{6}$$

Let $K(z) := \mathrm{E}_1 w^2(T,Z) \left[\delta - F(T-\mu_1(Z))\right]^2 I(Z \le z)$. Then, $\mathrm{E}_1 V_n^1(z) \equiv 0$, and a direct calculation shows that for all $-\infty \le z_1 < z < z_2 \le \infty$,

$$\begin{aligned} &\mathrm{E}_1 \Big\{ (V_n^1(z) - V_n^1(z_1))(V_n^1(z_2) - V_n^1(z)) \Big\}^2 \\ &= \frac{n-1}{n} [K(z) - K(z_1)][K(z_2) - K(z)] \le (K(z_2) - K(z_1))^2\,. \end{aligned}$$

This fact, in view of Theorem 15.6 in (Billingsley, 1968), implies that V_n^1 converges weakly to $B(K)$ in $D[-\infty, \infty]$ and uniform metric.

Summarizing, if F is strictly increasing on $\mathbb{R}$, then the test that rejects $\tilde{H}$: $\mu(z) = \mu_0(z)$ for all z, whenever $\sup_z |V_n^0(z)| \ge b_\alpha$ is consistent against all those alternatives $\mu_1(z)$ that satisfy (5) and (6).

By the Cauchy-Schwarz inequality applied twice, once to the conditional expectation, given Z, and once to the expectation with respect to the distribution of Z, it can be seen that (5)(a) is implied by

$$\mathrm{E}\frac{\Delta^2(T,Z)}{L(T-\mu_0(Z))} < \infty\,, \tag{7}$$

This condition in turn is satisfied for example when the support of F is $\mathbb{R}$ and that of $T-\mu_0(Z)$ is an interval (a,b), $-\infty < a < b < \infty$, since in this case the left hand side of (7) is bounded above by $\{F(a)\bar{F}(b)\}^{-1}$. In this case we also have the condition (6) satisfied, for the integral in (6) equals

$$\mathrm{E}\frac{\nu^2(T-\mu_0(Z))L(T-\mu_1(Z))}{\sigma_0^2(Z)L^2(T-\mu_0(Z))} \le \{4F(a)\bar{F}(b)\}^{-1}\,.$$

Another example where the condition (7) holds is when F is $N(0,\sigma^2)$ distribution function and T is a $N(0,\tau^2)$ random variable, with $\sigma > \tau$.

2.2 Tests for H_0

To discuss the more interesting problem of testing H_0, we proceed as follows. For convenience, let P_θ denote the joint distribution of (δ,T,Z) when $\mu = m_\theta$, and E_θ and Var_θ denote the corresponding mean and variance operations. Let R_θ, σ_θ stand for R, σ when $\mu = m_\theta$ and θ_n denote a $\sqrt{n}$-consistent estimator of θ_0, under H_0, based on (δ_i,T_i,Z_i); $1 \le i \le n$. See Section 3 below on how to obtain such an estimator of θ_0. Tests of H_0 will be based on the process $\tilde{V}_n(z) := V_n(z,\theta_n)$, where

$$V_n(z,\theta) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{R_\theta(\delta_i,T_i,Z_i)}{\sigma_\theta(Z_i)}\,I(Z_i \le z)\,, \qquad z \in \mathbb{R}\,, \theta \in \Theta\,.$$

To analyze asymptotic behavior of $\tilde{V}_n$, we need to make the following assumptions.

$$\sqrt{n}\|\theta_n - \theta_0\| = O_p(1)\,, \quad (P_{\theta_0}) \tag{8}$$

$$F \text{ has zero mean finite variance and a continuous density } f \tag{9}$$

$$m_\theta \text{ is differentiable in a neighborhood of } \theta_0 \text{ with its } q\times 1 \text{ vector of first derivatives } \dot{m}_{\theta_0}, \text{ such that } \forall\ 0 < b < \infty \tag{10}$$

$$\sup_{1\le i\le n\,,\ n^{1/2}\|\theta-\theta_0\|\le b}\frac{n^{1/2}|m_\theta(Z_i) - m_{\theta_0}(Z_i) - (\theta-\theta_0)'\dot{m}_{\theta_0}(Z_i)|}{\sigma_{\theta_0}(Z_i)} = o_p(1)\,.$$

$$\sup_{1\le i\le n\,,\ n^{1/2}\|\theta-\theta_0\|\le b}\left|\frac{\sigma_{\theta_0}(Z_i)}{\sigma_\theta(Z_i)} - 1\right| = o_p(1)\,. \tag{11}$$

$$\mathrm{E}\left\|\frac{\dot{m}_{\theta_0}(Z)}{\sigma_{\theta_0}(Z)}\right\|^2 < \infty\,. \tag{12}$$

$$\mathrm{E}\frac{|\nu(T-m_{\theta_0}(Z))|\ f(T-m_{\theta_0}(Z))}{L(T-m_{\theta_0}(Z))} < \infty\,. \tag{13}$$

Now, let $\dot{R}_\theta$, $\dot{\sigma}_\theta$ denote the vectors of the first derivatives of R_θ, σ_θ with respect to θ, and $\dot{r}_\theta$ a similar entity for the ratio $r_\theta := R_\theta/\sigma_\theta$. The existence of these entities is guaranteed by the assumptions (9) and (10). Direct calculations show that under these assumptions, for θ in a neighborhood of θ_0 and with $x = t - m_\theta(z)$,

$$\dot{R}_\theta(d,t,z) \;=\; \Big[\frac{-xf(x)[d-F(x)]+\nu(x)f(x)}{L(x)}+\nu(x)[d-F(x)]\frac{f(x)(1-2F(x))}{L^2(x)}\Big]\,\dot{m}_\theta(z)\,,$$

$$\dot{r}_\theta(d,t,z) := \frac{\partial}{\partial\theta}\Big(\frac{R_\theta(d,t,z)}{\sigma_\theta(z)}\Big) = \frac{\dot{R}_\theta(d,t,z)}{\sigma_\theta(z)} - \frac{\dot{\sigma}_\theta}{\sigma_\theta}(z)\frac{R_\theta(d,t,z)}{\sigma_\theta(z)}\,.$$

Let, for a $z \in \mathbb{R}$,

$$h_\theta(z) := \frac{1}{\sigma_\theta(z)}\mathrm{E}_\theta\dot{R}_\theta(\delta,T,z)\,, \qquad \ell_\theta(z,\Psi) := \mathrm{E}\left(\frac{\nu(T-m_\theta(z))\;f(T-m_\theta(z))}{L(T-m_\theta(z))}\right)\,.$$

Observe that because $R_{\theta_0}(\delta,T,Z)$ is conditionally centered under H_0, given Z, we obtain $\mathrm{E}_{\theta_0}\dot{r}_{\theta_0}(\delta,T,z) = h_{\theta_0}(z)$. Also, by the assumed independence,

$$\mathrm{E}\Big(\dot{R}_{\theta_0}(\delta,T,Z)|T=t,Z=z\Big) = \frac{\nu(t-m_{\theta_0}(z))\;f(t-m_{\theta_0}(z))}{L(t-m_{\theta_0}(z))}\;\dot{m}_{\theta_0}(z)\,,$$

$$\mathrm{E}_{\theta_0}\dot{R}_{\theta_0}(\delta,T,z) = \ell_{\theta_0}(z,\Psi)\dot{m}_{\theta_0}(z)\,, \quad h_{\theta_0}(z) = \ell_{\theta_0}(z,\Psi)\frac{\dot{m}_{\theta_0}(z)}{\sigma_{\theta_0}(z)}\,.$$

Next, let

$$D_\theta(z) := \mathrm{E}_\theta\Big\{\frac{\dot{R}_\theta(\delta,T,Z)}{\sigma_\theta(Z)}\;I(Z\le z)\Big\}\,, \qquad z\in\mathbb{R}\,,\theta\in\Theta\,.$$

Note that by (13), $\sup_z|D_{\theta_0}(z)| < \infty$. By the independence of T and Z,

$$D_{\theta_0}(z) = \int_{-\infty}^{z} h_{\theta_0}(u)dG(u) = \mathrm{E}_{\theta_0}\dot{r}_{\theta_0}(\delta,T,Z)\;I(Z\le z)\,, \quad z\in\mathbb{R}\,. \tag{14}$$

Now, rewrite

$$\tilde{V}_n(z) = V_n(z,\theta_0) - n^{1/2}(\theta_n-\theta_0)\;n^{-1}\sum_{i=1}^{n}\dot{r}_{\theta_0}(\delta_i,T_i,Z_i)\;I(Z_i\le z)$$
$$-n^{-1/2}\sum_{i=1}^{n}[r_{\theta_n}-r_{\theta_0}-(\theta_n-\theta_0)'\dot{r}_{\theta_0}](\delta_i,T_i,Z_i)\;I(Z_i\le z)\,.$$

The model assumptions and (9)-(13) imply that

$$\max_{1\le i\le n}\Big|n^{1/2}[r_{\theta_n}-r_{\theta_0}-(\theta_n-\theta_0)'\dot{r}_{\theta_0}](\delta_i,T_i,Z_i)\Big| = o_p(1)\,, \quad (H_0)\,.$$

A Glivenko-Cantelli type argument and the LLN's implies that under H_0,

$$\sup_{z\in\mathbb{R}}\Big|n^{-1}\sum_{i=1}^{n}\dot{r}_{\theta_0}(\delta_i,T_i,Z_i)\;I(Z_i\le z) - D_{\theta_0}(z)\Big| = o_p(1)\,.$$

These facts and a routine argument yield the following

Theorem 1. Under the assumptions (3), (8) – (13), we obtain that uniformly in $z \in \mathbb{R}$, under P_{θ_0},

$$\tilde{V}_n(z) = V_n(z, \theta_0) - n^{1/2}(\theta_n - \theta_0)' D_{\theta_0}(z) + o_p(1)\,.$$

Moreover, $V_n(\cdot, \theta_0) \Longrightarrow B(G(\cdot))$, in $D[-\infty, \infty]$, with respect to the uniform metric.

Next, we develop an analog of the linear transformation of Stute et al. (1998). For any matrix D, let D' denote its transpose, and let

$$\begin{aligned} A_{\theta_0}(z) &= \int_z^\infty h_{\theta_0}(u) h_{\theta_0}(u)' dG(u)\,, \\ &= \int_z^\infty \ell^2_{\theta_0}(u, \Psi) \frac{\dot{m}_{\theta_0}(u) \dot{m}_{\theta_0}(u)'}{\sigma^2_{\theta_0}(u)} dG(u)\,, \qquad z \in \mathbb{R}\,. \end{aligned}$$

Note that this is a nonnegative definite $q \times q$-matrix. But we shall assume that $A_{\theta_0}(z_0)$ is nonsingular for some $z_0 \in \mathbb{R}$. Then $A_{\theta_0}(z)$ is positive definite for all $z \le z_0$. Write $A^{-1}_{\theta_0}(z_0)$ for the inverse $(A_{\theta_0}(z_0))^{-1}$ and define the linear functional transform

$$Q(\varphi)(z) = \varphi(z) - \int_{-\infty}^z h_{\theta_0}(z_1)' A^{-1}_{\theta_0}(z_1) \left[\int_{z_1}^\infty h_{\theta_0}(z_2) \varphi(dz_2) \right] dG(z_1)\,, \qquad z \le z_0\,.$$

When we apply Q to Brownian motion $B \circ G$, the inner integral needs to be interpreted as a stochastic integral.

Observe that (14) readily implies $Q(D'_{\theta_0} U) = 0$, for all $U \in \mathbb{R}^q$. Arguing as in Stute et al. (1998), one can also verify that Q maps $B \circ G$ to $B \circ G$. Consequently, we have $Q(B \circ G + D'_{\theta_0} U) = Q(B \circ G) = B \circ G$, for any $U \in \mathbb{R}^q$.

These observations together with Theorem 1 suggest that under H_0, $Q\tilde{V}_n$ would also converge weakly to $B \circ G$. But the transformation Q depends on the unknown parameters θ_0, Ψ and G. Let h_n, A_n, and σ_n denote the h_{θ_0}, A_{θ_0}, and σ_{θ_0} after θ_0, Ψ, and G are replaced by θ_n, and the empirical distribution functions Ψ_n of T_i's and G_n of Z_i's, respectively, in there. Define the estimate of Q to be

$$Q_n(\varphi)(z) = \varphi(z) - \int_{-\infty}^z h_n(z_1)' A^{-1}_n(z_1) \left[\int_{z_1}^\infty h_n(z_2) \varphi(dz_2) \right] dG_n(z_1)\,, \qquad z \le z_0\,.$$

To verify the weak convergence of $Q_n \tilde{V}_n$ we need the following additional conditions on h_θ. For every $k < \infty$,

$$\begin{aligned} &\sup_{n^{1/2}\|\theta - \theta_0\| \le k\,,\ 1 \le i \le n} \|h_\theta(Z_i) - h_{\theta_0}(Z_i)\| = o_p(1)\,, \qquad (15) \\ &\mathrm{E}\|h_{\theta_0 + n^{-1/2} t_2}(Z) - h_{\theta_0 + n^{-1/2} t_1}(Z)\|^2 \le C\, \|t_2 - t_1\|^2\,, \end{aligned}$$

for all t_1, t_2 in the ball $\{t \in \mathbb{R}^q\,;\ \|t\| \le k\}$, and for some constant C which may depend on θ_0 and F,Ψ, and G. Using the methods of proof of Koul and Stute (1999), one can verify that under the above assumed conditions and under H_0, $Q_n \tilde{V}_n \Longrightarrow B \circ G$ in $D[-\infty, z_0]$

and uniform metric. Hence, tests based on any continuous functional of this process will be asymptotically distribution free

Note that because h_θ involves $\dot{m}_\theta$ and σ_θ, the condition (15) imposes additional smoothness conditions on these functions. As an example, if F is a normal distribution and Ψ has a Lipschitz Lebegue density then (15) is impled by

$$\sup_{n^{1/2}\|\theta-\theta_0\|\le k,\ 1\le i\le n} \|\dot{m}_\theta(Z_i) - \dot{m}_{\theta_0}(Z_i)\| = o_p(1)\,,$$
$$\mathrm{E}\|\dot{m}_{\theta_0+n^{-1/2}t_2}(Z) - \dot{m}_{\theta_0+n^{-1/2}t_1}(Z)\|^2 \le C\,\|t_2-t_1\|^2\,.$$

3 Estimation of θ

In order to apply the above results, it is important to have a $n^{1/2}$-consistent estimator of θ_0 under H_0. (Li and Zhang, 1998) constructed M-estimators of the regression coefficients in a linear regression model with interval censored data and when the error distribution function is unknown. Since here F is assumed to be known and since their estimator is computationally much more involved, one may instead use the conditional least square estimator defined by

$$\hat{\theta}_{lse} = \mathrm{argmin}_{\theta\in\Theta} \sum_{i=1}^{n} [\delta_i - F(T_i - m_\theta(Z_i))]^2\ .$$

Assume that F has a continuously differentiable density f and

$$\Sigma_{\theta_0} := \mathrm{E}\Big(f^2(T - m_{\theta_0}(Z))\ \dot{m}_{\theta_0}(Z)\dot{m}_{\theta_0}(Z)'\Big)$$

is positive definite. In addition assume that $\dot{m}_\theta$ is continuously differentiable with the matrix of derivatives $\ddot{m}_{\theta_0}(z)$ satisfying $\|\ddot{m}_{\theta_0}(z)\| \le M_{\theta_0}(z)$, with $\int M_{\theta_0}(z)dG(z) < \infty$. Then using the classical Cramér type of argument one can verify that

$$n^{\frac{1}{2}}(\hat{\theta}_{lse}-\theta_0) = \Sigma_{\theta_0}^{-1} n^{-\frac{1}{2}} \sum_{i=1}^{n} [\delta_i - F(T_i - m_{\theta_0}(Z_i))] f(T_i - m_{\theta_0}(Z_i))\dot{m}_{\theta_0}(Z_i) + o_p(1)\,,\ (P_{\theta_0})\,.$$

Consequently, under H_0,

$$n^{1/2}(\hat{\theta}_{lse} - \theta_0) \to \mathcal{N}_q(0, \Omega_0)\,,$$
$$\Omega_0 := \Sigma_{\theta_0}^{-1} M_0 \Sigma_{\theta_0}^{-1}\,,$$
$$M_0 := \mathrm{E}\Big\{ (F\bar{F}f^2)(T - m_{\theta_0}(Z))\ \dot{m}_{\theta_0}(Z)\dot{m}_{\theta_0}(Z)'\Big\}\,.$$

See, e.g., Liese and Vajda (2004) for a general method of proving asymptotic normality in nonlinear regression models.

Unknown Distribution of T. The only place where the distribution function of T appears in the process V_n is via the conditional variance $\sigma^2_{\theta_n}(Z_i)$. An obvious estimate of this entity is

$$s_n^2(Z_i) := n^{-1} \sum_{j=1}^{n} \frac{\nu^2(T_j - m_{\theta_n}(Z_i))}{L(T_j - m_{\theta_n}(Z_i))}\,.$$

Then the tests of H_0 can be based on the process $\hat{V}_n$, the analog of $\tilde{V}_n$ where $\sigma_{\theta_n}(Z_i)$ is replaced by $s_n(Z_i)$. The test of $\tilde{H}_0$ can be modified similarly. The asymptotic properties of these tests will be similar to those of the above, but are not discussed here.

Unknown F. In principle the above testing procedure can be adapted to the case when F is unknown by replacing F in $\tilde{V}_n$ and Q_n by its nonparametric conditional maximum likelihood estimator (NPMLE), given Z_i, $1 \le i \le n$. This estimator is defined to be

$$\text{argmax}_F \sum_{i=1}^{n} \left[\delta_i \log F(T_i - m_{\hat{\theta}}(Z_i)) + (1-\delta_i) \log \bar{F}(T_i - m_{\hat{\theta}}(Z_i)) \right],$$

where $\hat{\theta}$ is a $n^{1/2}$-consistent estimator of θ_0 under H_0. Such an estimator of F can be computed by using the techniques described in Ayer et al. (1955) or Groeneboom and Wellner (1992). However, the asymptotic properties like the weak convergence under H_0 of the so modified $\tilde{V}_n$ appears to be intractable at the present time. Such properties are also not available for the NPMLE of F even when there are no nuisance parameters to be estimated.

There also does not appear to be any readily available $n^{1/2}$-consistent estimator of θ for a general nonlinear m_θ when F is unknown. For linear m_θ, one could use estimators proposed by Li and Zhang (1998). The results of Klein and Spady (1993) may be found useful for a general nonlinear m_θ.

4 A Simulation

Here we report results of a finite sample simulation. For simplicity we took $\mathcal{M}$ to be simple linear regression model. Thus $q = 1$ and $m_\theta(z) = \theta z$, $\theta \in \mathbb{R}$. In this case then several entities simplify as follows. Let $Z_{(1)} \le Z_{(2)} \le \cdots \le Z_{(n)}$ denote the ordered Z_i's and $T_{(i)}$'s, $\delta_{(i)}$'s denote the corresponding T_i's and δ_i's. Also, let $\ell_{nj}(z) \equiv \ell_{\theta_n}(Z_{(j)}, \Psi_n)$, $R_{nj} := R_{\theta_n}(\delta_{(j)}, T_{(j)}, Z_{(j)})$, $\sigma_{nj} := \sigma_n(Z_{(j)})$, and $A_{nj} := A_n(Z_{(j)})$, where now

$$A_n(z) := \frac{1}{n} \sum_{i=1}^{n} \frac{\ell_n^2(Z_{(i)})}{\sigma_n^2(Z_{(i)})} Z_{(i)}^2 \; I(Z_{(i)} \ge z)\,.$$

Then

$$\begin{aligned} Q_n \tilde{V}_n(z) &= \tilde{V}_n(z) - \frac{1}{n} \sum_{i=1}^{n} \frac{Z_{(i)} \ell_{ni}}{A_{ni} \sigma_{ni}} \frac{1}{n^{1/2}} \sum_{j=1}^{n} \frac{Z_{(j)} \ell_{nj} R_{nj}}{\sigma_{nj}^2} \; I(Z_{(j)} \wedge z \ge Z_{(i)}) \\ &= \frac{1}{n^{1/2}} \sum_{j=1}^{n} \left\{ I(Z_{(j)} \le z) - \frac{1}{n} \sum_{i=1}^{j} \frac{Z_{(i)} Z_{(j)} \ell_{ni} \ell_{nj}}{A_{ni} \sigma_{ni} \sigma_{nj}} \; I(z \ge Z_{(i)}) \right\} \frac{R_{nj}}{\sigma_{nj}}\,. \end{aligned}$$

Here $A_{\theta_0}(z)$ and $A_n(z)$ are positive definite for all $z \in \mathbb{R}$ and the statistics

$$K_n := \sup_{z \in \mathbb{R}} |V_n^0(z)|\,, \qquad \hat{K}_n := \sup_{z \in \mathbb{R}} |Q_n \tilde{V}_n(z)|$$

Table 1: Empirical sizes and powers of the K_n test, $F = F_1 = \text{DE}(0, \beta)$

		$\alpha = 0.1$		$\alpha = 0.05$		$\alpha = 0.01$	
a	β	$n = 100$	$n = 200$	$n = 100$	$n = 200$	$n = 100$	$n = 200$
0	1	0.083	0.095	0.038	0.053	0.018	0.021
	2	0.090	0.091	0.047	0.044	0.017	0.018
	3	0.097	0.115	0.052	0.040	0.031	0.020
1	1	0.350	0.574	0.231	0.433	0.147	0.340
	2	0.205	0.342	0.135	0.246	0.075	0.172
	3	0.146	0.242	0.082	0.170	0.046	0.102
2	1	0.747	0.968	0.644	0.939	0.537	0.904
	2	0.508	0.792	0.388	0.691	0.264	0.591
	3	0.345	0.576	0.241	0.458	0.159	0.349

are well defined. The proposed test rejects $\tilde{H}$ (H_0) whenever $K_n > b_\alpha$, $\hat{K}_n > b_\alpha$, where b_α is the $100(1-\alpha)$th percentile of the distribution of $\sup_{0\le t\le 1} |B(t)|$. Note that

$$K_n = \frac{1}{\sqrt{n}} \max_{1\le j\le n} \Big| \sum_{k=1}^{j} \frac{R_0(\delta_{(k)}, T_{(k)}, Z_{(k)})}{\sigma_0(Z_{(k)})} \Big|,$$

$$\hat{K}_n = \frac{1}{\sqrt{n}} \max_{1\le k\le n} \Big| \sum_{j=1}^{k} \Big[1 - \frac{1}{n} \sum_{i=1}^{j} \frac{Z_{(i)} Z_{(j)} \ell_{ni} \ell_{nj}}{A_{ni} \sigma_{ni} \sigma_{nj}}\Big] \frac{R_{nj}}{\sigma_{nj}} \Big|.$$

In our simulations we generated Z_i's from the uniform distribution on the interval $[0, 1]$ and ε_i's independently from following three densities.

$$\text{DE}(0, \beta) : f_1(x) := \frac{1}{2\beta} e^{-|x|/\beta};$$

$$\text{Logistic}(0, \beta) : f_2(x) := \frac{e^{-x/\beta}}{\beta(1 + e^{-x/\beta})^2};$$

$$N(0, \sigma^2) : f_2(x) := \frac{1}{\sqrt{2\pi}\sigma} e^{-x^2/\sigma^2},\ x \in \mathbb{R}.$$

Finally, Y_i's were generated according to the model $Y_i = 3Z_i + aZ_i^2 + \varepsilon_i$, $1 \le i \le n$. The censoring variables T_i's were generated from the uniform distribution on the interval $[0, 3]$. Hence $\tilde{H}$: $\mu(z) = 3z$ and H_0: $\mu \in \mathcal{M}$ hold with $\theta_0 = 3$ if and only if $a = 0$.

We computed the empirical sizes and powers for different values of a and different error distributions. The results represent the Monte Carlo levels when $a = 0$ and the Monte Carlo powers when $a \neq 0$. The test K_n was simulated for the sample $n = 100, 200$ at all three error distributions while $\hat{K}_n$ was simulated for $n = 200, 400$ at only the two error distribution functions, $F = F_1$, $F = F_3$. All simulations are based on 1000 replications. The entries in the tables are obtained as the $\#(K_n > b_\alpha)/1000$ and $\#(\hat{K}_n > b_\alpha)/1000$, where the b_α's are obtained from the distribution of $\{\sup |B(t)|;\ 0 \le t \le 1\}$. In particular $b_{0.1} = 1.96$, $b_{0.05} = 2.2414$, and $b_{0.025} = 2.5$.

From these simulations, we see that the empirical size is close to the nominal level for the samples sizes of 200 and larger when $\alpha = .1, .05$, while for $\alpha = .01$ there is more

Table 2: Empirical sizes and powers of the K_n test, $F = F_2 = \text{Logistic}(0, \beta)$

		$\alpha = 0.1$		$\alpha = 0.05$		$\alpha = 0.01$	
a	β	$n = 100$	$n = 200$	$n = 100$	$n = 200$	$n = 100$	$n = 200$
0	1	0.088	0.095	0.043	0.043	0.022	0.021
	2	0.092	0.098	0.046	0.048	0.021	0.022
	3	0.101	0.079	0.047	0.040	0.021	0.021
1	1	0.262	0.488	0.162	0.348	0.108	0.252
	2	0.134	0.237	0.090	0.150	0.054	0.104
	3	0.131	0.149	0.077	0.085	0.045	0.052
2	1	0.641	0.922	0.527	0.869	0.406	0.784
	2	0.327	0.568	0.221	0.455	0.146	0.341
	3	0.209	0.348	0.130	0.242	0.083	0.166

Table 3: Empirical sizes and powers of the K_n test, $F = F_3 = \text{Normal}(0, \sigma^2)$

		$\alpha = 0.1$		$\alpha = 0.05$		$\alpha = 0.01$	
a	σ	$n = 100$	$n = 200$	$n = 100$	$n = 200$	$n = 100$	$n = 200$
0	1	0.095	0.083	0.052	0.035	0.026	0.019
	2	0.097	0.085	0.041	0.046	0.020	0.020
	3	0.091	0.117	0.041	0.061	0.020	0.036
1	1	0.416	0.742	0.313	0.630	0.233	0.053
	2	0.208	0.385	0.135	0.270	0.089	0.187
	3	0.155	0.260	0.085	0.173	0.045	0.107
2	1	0.890	0.987	0.808	0.976	0.722	0.964
	2	0.581	0.870	0.472	0.782	0.363	0.686
	3	0.363	0.639	0.265	0.528	0.186	0.423

Table 4: Empirical sizes and powers of the $\hat{K}_n$ test, $F = F_1 = \text{DE}(0, \beta)$

		$\alpha = 0.1$		$\alpha = 0.05$		$\alpha = 0.01$	
a	β	$n = 200$	$n = 400$	$n = 200$	$n = 400$	$n = 200$	$n = 400$
0	1	0.089	0.109	0.037	0.054	0.016	0.024
	2	0.058	0.114	0.025	0.059	0.012	0.032
	3	0.077	0.100	0.035	0.047	0.016	0.023
1	1	0.106	0.169	0.051	0.089	0.017	0.050
	2	0.080	0.135	0.040	0.067	0.019	0.040
	3	0.081	0.122	0.042	0.058	0.021	0.025
3	1	0.177	0.477	0.091	0.301	0.042	0.171
	2	0.191	0.355	0.098	0.233	0.050	0.144
	3	0.161	0.286	0.082	0.175	0.042	0.109
5	1	0.217	0.534	0.090	0.343	0.042	0.217
	2	0.274	0.548	0.165	0.401	0.082	0.263
	3	0.231	0.470	0.135	0.323	0.080	0.203

Table 5: Empirical sizes and powers of the $\hat{K}_n$ test, $F = F_3 = \mathrm{N}(0, \sigma^2)$

		$\alpha = 0.1$		$\alpha = 0.05$		$\alpha = 0.01$	
a	σ	$n = 200$	$n = 400$	$n = 200$	$n = 400$	$n = 200$	$n = 400$
0	1	0.054	0.104	0.020	0.049	0.006	0.025
	2	0.057	0.101	0.025	0.052	0.009	0.017
	3	0.072	0.152	0.029	0.081	0.013	0.033
1	1	0.058	0.107	0.022	0.048	0.008	0.017
	2	0.127	0.171	0.065	0.093	0.028	0.057
	3	0.119	0.229	0.052	0.123	0.026	0.077
3	1	0.225	0.509	0.124	0.344	0.056	0.204
	2	0.211	0.472	0.122	0.290	0.064	0.195
	3	0.195	0.344	0.113	0.219	0.056	0.124
5	1	0.305	0.619	0.169	0.448	0.087	0.302
	2	0.288	0.541	0.019	0.414	0.110	0.291
	3	0.243	0.396	0.157	0.279	0.091	0.192

variability, depending on the tails of the error distribution. In particular the scale parameters σ and β appear to have influence on the Monte Carlo level. Under the alternatives, the power decreases as β or σ increases, while it increases as a increases and sample size increases.

References

An, H. Z., and Cheng, B. (1991). A Kolmogorov-Smirnov type statistic with application to test for nonlinearity in time series. *International Statistical Reviews*, *59*, 287-307.

Ayer, M., Brunk, M. D., Ewing, G. M., Reid, W. T., and Silverman, E. (1955). An empirical distribution function for sampling with incomplete information. *Annals of Mathematical Statistics*, *26*, 641-647.

Billingsley, P. (1968). *Convergence of Probability Measures*. New York: J. Wiley.

Diamond, I. D., and McDonald, J. W. (1991). Analysis of Current Status Data. In J. Trussell, R. Hankinson, and J. Tilton (Eds.), *Demographic Applications of Event History Analysis* (p. 231-252). Oxford University Press.

Diamond, I. D., McDonald, J. W., and Shah, I. H. (1986). Proportional hazards models for current status data: application to the study of differentials in age at weaning in Pakistan. *Demography*, *23*, 607-620.

Finkelstein, D. M. (1986). A proportional hazards model for interval-censored failure time data. *Biometrics*, *42*, 845-854.

Finkelstein, D. M., and Wolfe, R. A. (1985). A semiparametric model for regression analysis of interval-censored failure time data. *Biometrics*, *41*, 933-945.

Groeneboom, P., and Wellner, J. A. (1992). Information Bounds and Nonparametric Maximum Likelihood Estimation. In *DMV Seminar* (Vol. 19). Basel: Birkhäuser Verlag.

Hoel, D. G., and Walburg, H. E. (1972). Statistical analysis of survival experiment. *Journal of National Cancer Institute*, *49*, 361-372.

Jewell, N. P., and Laan, M. van der. (2004). Current status data: review, recent developments and open problems. In *Advances in Survival Analysis* (Vol. 23, p. 625-642). Amsterdam: Elsevier.

Keiding, N. (1991). Age-specific incidence and prevalence: A statistical perspective (with discussion). *Journal of Royal Statistical Society*, *154*, 371-412.

Klein, R. W., and Spady, R. H. (1993). An efficient semiparametric estimator for binary response models. *Econometrica*, *61*, 387-421.

Koul, H. L., and Stute, W. (1999). Nonparametric model checks for time series. *Annals of Statistics*, *27*, 204-236.

Li, G., and Zhang, C. H. (1998). Linear regression with interval censored data. *Annals of Statistics*, *26*.

Liese, F., and Vajda, I. (2004). A general asymptotic theory of M-estimators II. *Mathematical Methods in Statistics*, *13*(1), 82-95.

Stute, W. (1997). Nonparametric model checks for regression. *Annals of Statistics*, *25*, 613-641.

Stute, W., Thies, S., and Zhu, L. X. (1998). Model checks for regression: an innovation process approach. *Annals of Statistics*, *26*, 1916-1934.

Authors' address:

Hira L. Koul and Tingting Yi
Michigan State University
Statistics and Probability
A435 Wells Hall
East Lansing, Michigan USA

Tel. 517-353-7170
Fax: 517-432-1405
E-Mail: koul@stt.msu.edu