Censored Local Likelihood

Clive R. Loader [1]

October 16, 1996

Abstract

Censored survival data arises in many situations. Often, interest focuses on covariate effects: for example, does increasing a covariate reduce a patient's risk of death? One approach to analyzing this kind of data is a censored regression model, which estimates the mean survival time, conditional on the value of the covariates. Typically, one transforms the data to allow for censoring effects, and fits a regression model to the transformed data using least squares.

When survival times are not normally distributed least squares techniques may be inappropriate. In this paper we develop likelihood regression models for censored data, and fit using local polynomial methods. After appropriate changes to the likelihood to allow for censoring, many methods that have been developed for generalized linear models and local likelihood are found to carry over with minor modifications to the censored case; we study in particular fitting algorithms, cross validation, AIC and diagnostics based on simultaneous confidence bands for local coefficients. Asymptotic properties of the estimates are derived.

Key Words: Censored regression; Local likelihood; Local polynomials.

---

[1] Member of Technical Staff, Lucent Technologies, Room 2C-279, 700 Mountain Avenue, Murray Hill, NJ 07974, U.S.A. The Author acknowledges the cricinfo database, `http://www.cricket.org` from where the batting dataset was obtained.

1 Introduction

Likelihood models can be used in regression studies when ordinary least squares techniques are inappropriate; for example, count data. The book McCullagh and Nelder (1989) provides an extensive account.

When insufficient information is available for a regression model to be specified, smooth models can be fitted. This idea originated as long ago as Henderson (1924), who fitted spline models to count data using a logistic likelihood. However, smooth likelihood models have received relatively little attention until the past decade, when increased computational power has made implementation more feasible. The local likelihood procedure was first proposed in Brillinger (1977) and studied in more detail by Tibshirani and Hastie (1987), Staniswalis (1989) and others.

In this paper we study the use of local likelihood in the presence of censored observations. The estimate is defined in Section 2. Diagnostic tools for use with the estimate are discussed in Section 3. Section 4 briefly discusses computational issues and software. Some examples are presented in Section 5. Some theory and asymptotic properties of the estimate are presented in Section 6.

This work is complimentary to existing work on local polynomial methods with censored data. Tibshirani and Hastie (1987) estimated covariate effects using a localized version of the proportional hazards regression model. Gentleman and Crowley (1991) also considered the proportional hazards model, estimating both covariate effects and the cumulative hazard. Wu and Tuma (1990) considered hazard rate estimation in the presence of covariates, using a local log-linear model for the hazard rate and global linear model for the covariates. Fan and Gijbels (1994) fitted censored regression models using local polynomials. Li and Doss (1995) derive estimates of the survival, hazard and cumulative hazard functions using local linear models for covariate effects. Hjort (1997) considered the estimation of hazard rates in the presence of censoring, using a local likelihood procedure. Gray (1996) uses multivariate local regression to estimate hazard rates in the presence of covariates, using local polynomial regression. An adaptation of Hjort's local likelihood procedure to this problem is described in Chapter 7 of Loader (1996b).

2 Likelihood Models

In a likelihood regression model, we observe predictor variables $x_i \in R^d; i = 1, \ldots, n$ and responses T_i, and assume a model

$$T_i \sim f(t, \theta(x_i)),$$

where $f(t, \theta)$ is a family of densities or probability mass functions. The object is to estimate the function $\theta(x)$, and the conditional mean

$$\mu = \int t f(t, \theta) dt.$$

In many cases μ is a 1-1 function of θ, and so the two estimation problems are equivalent. The link function $g(\cdot)$ defines the relation between θ and μ; $\theta = g(\mu)$.

In the censored likelihood model, we do not observe T_i, but rather observe

$$Y_i = \begin{cases} T_i & T_i < c_i \\ c_i & T_i \geq c_i \end{cases}.$$

The censoring indicators $\delta_i = I(Y_i = c_i)$ are observed, although the censoring times c_i are usually unobserved for uncensored observations ($\delta_i = 0$). The global log-likelihood is

$$\mathcal{L}(\theta) = \sum_{i=1}^{n} l(Y_i, \theta(x_i), c_i)$$

where

$$l(y,\theta,c) = \begin{cases} \log f(y,\theta) & y < c \\ \log \int_y^\infty f(x,\theta)dx & y = c \end{cases}.$$

Let $W(\cdot)$ be a symmetric non-negative weight function; in our examples we use the tricube weights

$$W(v) = \begin{cases} (1-|v|^3)^3 & |v| \le 1 \\ 0 & |v| > 1 \end{cases}.$$

For a fitting point x, choose a bandwidth $h = h(x)$ and assign weights

$$w_i(x) = W\left(\frac{x_i - x}{h}\right)$$

Within the smoothing window, assume a local polynomial approximation of order $p \ge 0$;

$$\theta(u) \approx \langle a, A(u-x)\rangle ; x - h \le u \le x + h$$

where a is a vector of coefficients and $A(v)$ is a vector of the polynomial basis functions; in one dimension, $A(v) = (\,1 \quad v \quad \ldots \quad v^p/p!\,)^T$. The local likelihood at a point x is

$$\mathcal{L}_x(a) = \sum_{i=1}^{n} w_i(x) l(Y_i, \langle a, A(x_i - x)\rangle, c_i). \tag{1}$$

The local likelihood estimate is defined by maximizing the local likelihood.

Definition 1 (Local Likelihood Estimate) *Let $\hat{a} = \hat{a}(x)$ be the parameter vector that maximizes the local likelihood (1). The local likelihood estimate is*

$$\begin{aligned} \hat{\theta}(x) &= \langle \hat{a}, A(0)\rangle = \hat{a}_0. \\ \hat{\mu}(x) &= g^{-1}(\hat{\theta}(x)). \end{aligned}$$

Let $\dot{l}$ and $\ddot{l}$ denote the first and second partial derivatives of $l(Y,\theta,c)$ with respect to θ. Under some regularity conditions, differentiation of (1) shows the estimate $\hat{a}$

satisfies the system of equations

$$\sum_{i=1}^{n} w_i(x)\dot{l}(Y_i, \langle \hat{a}, A(x_i - x)\rangle, c_i)A(x_i - x) = 0. \tag{2}$$

These equations can be solved using the Newton-Raphson iterations;

$$\hat{a}^{(j+1)} = \hat{a}^{(j)} + \mathbf{J}^{-1}\sum_{i=1}^{n} w_i(x)A(x_i - x)\dot{l}(Y_i, \left\langle \hat{a}^{(j)}, A(x_i - x)\right\rangle, c_i)$$

where

$$\mathbf{J} = \mathbf{J}(x) = -\sum_{i=1}^{n} w_i(x)A(x_i - x)A(x_i - x)^T\ddot{l}(Y_i, \theta_i, c_i)$$

and evaluated at $\theta_i = \left\langle \hat{a}^{(j)}, A(x_i - x)\right\rangle$.

In the theory of generalized linear models, it is usually recommended that $E(\ddot{l})$ be used in place of $\ddot{l}$; this improves the stability of the Newton-Raphson algorithm and leads to more natural variance estimates. In the presence of censoring this modification is unavailable. The expected value is

$$E\ddot{l}(Y_i, \theta_i, c_i) = \int_{-\infty}^{c_i} \ddot{l}(y, \theta_i, c_i)f(y, \theta_i)dy + \ddot{l}(c_i, \theta_i, c_i)(1 - F(c_i, \theta_i)).$$

This depends on the nuisance parameters c_i, which are usually unknown for the uncensored observations, and cannot reasonably be estimated without making additional stochastic assumptions.

It is of interest to study the local likelihood further for the normal model. We suppose for simplicity the variance $\sigma^2 = 1$ is known. Under the identity link, we can show

$$\dot{l}(Y_i, \mu_i, c_i) = Y_i^* - \mu_i$$

where

$$Y_i^* = E(T_i|Y_i, c_i) = \begin{cases} Y_i & Y_i < c_i \\ \mu_i + \phi(Y_i - \mu_i)/(1 - \Phi(Y_i - \mu_i)) & Y_i = c_i \end{cases}. \tag{3}$$

It follows the model can be fitted by an alternating least squares algorithm:

1. Initialize $Y_i^* = Y_i$.

2. Obtain $\hat{\mu}(x)$ by regressing Y_i^* on x_i.

3. Update Y_i^* using (3).

4. Repeat 2 and 3 until convergence.

This leads to a localized version of the algorithm of Schmee and Hahn (1979).

In censored regression literature, nonparametric transformations are considered much more frequently than parametric transformations such as (3); see Fan and Gijbels (1994) for approaches using local linear approaches, and references therein for approaches using global regression models. This has the obvious appeal of removing normality assumptions. However, there are limitations to nonparametric approaches. If the censoring times are bounded - for example, if censoring results from termination of the study - one has no information about the upper tail of the distribution of the T_i, and the mean cannot be estimated consistently.

3 Diagnostics

In this section we develop a variety of diagnostic tools for censored local likelihood: Cross validation and AIC; simultaneous prediction bands; and residual based diagnostics. These tools are useful for a variety of different purposes. Cross validation and AIC attempt to select smoothing parameters by measuring the predictive power of the fit: how good is the fitted curve at predicting future observations? However, these and other bandwidth selectors are of limited use in exploratory data analysis, since they aren't directly related to the question of interest: what, if any, structure in the data is real? A better approach is to begin with a smoothing parameter allowing the fit some flexibility from the global model, and use other diagnostic tools to assess the fit. The prediction bands we develop provide an assessment of statistical significance of features in the dataset. Residuals are used to assess adequacy of the fit - for example, when a peak or other significant structure has been found, has the fit properly modeled the peak, or is the peak trimmed?

Cross validation for local likelihood models was considered in Firth, Glosup, and Hinkley (1991). The method extends directly to the case of censored observations.

Definition 2 (Likelihood cross validation) *Let $\hat{\theta}_{-i}(x_i)$ be the local likelihood estimate of $\theta(x)$ when (x_i, Y_i, c_i) is removed from the dataset. The Likelihood cross validation score is defined as*

$$LCV = \sum_{i=1}^{n} D(Y_i, \hat{\theta}_{-i}(x_i), c_i).$$

where

$$D(Y_i, \theta_i, c_i) = 2\left(\sup_{\theta} l(Y_i, \theta, c_i) - l(Y_i, \theta_i, c_i)\right). \tag{4}$$

In large datasets, full computation of the LCV score may be undesirable. An approximation to the cross validated fit is

$$\hat{\theta}_{-i}(x_i) \approx \hat{\theta}_i(x_i) - \text{infl}(x_i)\dot{l}(Y_i, \hat{\theta}_i, c_i)$$

where the influence function is defined by

$$\text{infl}(x) = e_1^T \mathbf{J}^{-1} e_1 W(0).$$

Using a linear approximation to the deviance gives

$$D(Y_i, \hat{\theta}_{-i}(x_i), c_i) \approx D(Y_i, \hat{\theta}(x_i), c_i) - 2\text{infl}(x_i)\dot{l}(Y_i, \hat{\theta}(x_i), c_i)^2. \tag{5}$$

The simplification of (5) using the relation $E\dot{l}(Y,\theta,c)^2 = -E\ddot{l}(Y,\theta,c)$ motivates our definition of Akaike's criterion for local likelihood.

Definition 3 *The Akaike criterion for local likelihood is*

$$AIC = \sum_{i=1}^{n} D(Y_i, \hat{\theta}(x_i)) + 2\nu_1$$

where ν_1 is the effective degrees of freedom, defined by

$$\nu_1 = -\sum_{i=1}^{n} \text{infl}(x_i) E(\ddot{l}(Y_i, \theta_i, c_i)),$$

and estimated by

$$\hat{\nu}_1 = -\sum_{i=1}^{n} \text{infl}(x_i) \ddot{l}(Y_i, \hat{\theta}_i, c_i)$$

The higher order coefficients of the parameter estimate $\hat{a}$ can be interpreted as estimates of the derivatives of $\theta(x)$. These can be used to help determine significant structure in the estimate. For example, if $\theta(x)$ is constant, then $\theta'(x) = 0$ for all x. Thus, large values of $\hat{a}_1$ are indicative of points where $\theta(x)$ is increasing.

To determine the significance of $\hat{a}_1$, we compute 95% simultaneous prediction bands for $\hat{a}_1$ under the null hypothesis that $\theta'(x) = 0$ for all x, and declare significance whenever $\hat{a}_1$ falls outside these bands. The methods developed by Sun and Loader (1994) for simultaneous confidence bands in local regression and extended to generalized linear models in Sun, Loader, and McCormick (1994) can be simply adapted to the present problem. The probability approximation used is

$$P(\sup_{a \le x \le b} \frac{\hat{a}_1(x)}{\sqrt{\text{var}(\hat{a}_1(x))}} \ge c) \approx 2(1 - \Phi(c)) + \frac{\kappa_0}{\pi} e^{-c^2/2}$$

where $\Phi(\cdot)$ is the standard normal distribution function and

$$\kappa_0 = \int_a^b \left| \frac{\partial^2}{\partial x_1 \partial x_2} \sigma(x_1, x_2) \right|_{x_1 = x_2 = x}^{1/2} dx \qquad (6)$$

and $\sigma(x_1, x_2)$ is the covariance of the process $\hat{a}_1(x)/\sqrt{\text{var}(\hat{a}_1(x))}$, approximated in Theorem 3 below. Computationally, (6) can be simplified considerably; in particular, full evaluation and differentiation of the covariance is not required. See Sun and Loader (1994) for more details.

Residuals for generalized linear models are discussed in Section 2.4 of McCullagh and Nelder (1989). In the presence of censoring, the deviance residuals are the most convenient; these are defined by

$$r_i = \sqrt{D(Y_i, \hat{\theta}(x_i), c_i)}\,\text{sgn}(\dot{l}(Y_i, \hat{\theta}(x_i), c_i))$$

where $D(y, \theta, c)$ is defined by (4) and $\text{sgn}(\,\cdot\,)$ is the sign function.

4 Implementation and Software

Any local likelihood software can be simply modified to incorporate censoring: we simply need to ensure the correct likelihood and derivatives are used for censored observations. The formulae for the influence function, variance and effective degrees of freedom are identical to those in the uncensored case.

All examples in this paper are implemented using the LOCFIT software, described in Loader (1996b). The S commands for fitting and plotting a censored likelihood model - here using the batting dataset described in Section 5 - are simply:

```
> fit <- locfit(runs~dy,cens=no,data=border,family="geom",alpha=0.6)
> plot(fit)
```

The software presently supports censored local likelihood for the Gaussian family (identity link); Poisson family (identity, log and square root links); gamma family (log and inverse links) and geometric/negative binomial family (log link). The important concavity of the log-likelihood is easily shown to hold in all cases for uncensored observations. For censored observations, convexity holds in many cases although verification becomes messy; some examples are sketched in Chapter 7 of Loader (1996b). The convexity fails in the Gamma family with the inverse link, when the shape parameter is less than 1.

5 Examples

In the game of cricket, a batsman (or batswoman) attempts to score runs by hitting balls bowled by players on the opposing side. A batsman's innings usually continues until he is 'out' - for example, a ball is caught by a player on the fielding side. Occasionally, a batsman's innings will be left unfinished, recorded as 'not out'. A common measure of the batsman's performance is the average over several innings, computed as the total number of runs scored, divided by the number of times the batsman was 'out'. Figure 1 shows the record of Australian cricketer Allan Border, who scored 11174 runs between 1977 and 1994. Of the 265 innings played, 44 were not out, giving a career average of 50.56.

To localize this computation, we model using the geometric distribution,

$$P(T_i = y) = \frac{\mu_i^y}{(1 + \mu_i)^{y+1}}$$

and the log link, $\theta_i = \log(\mu_i)$. A local quadratic model is fitted, choosing $h = h(x)$ according to a nearest neighbor rule, covering 60% of the data. The fitted curve in the top panel of Figure 1 shows an initial increase, with $\mu(x)$ peaking at about 60 in 1985, followed by a gradual decline to about 42 in 1994. In cricketing terms, this change is quite significant: batsmen typically maintain averages between 40 and

50, while 60 represents an exceptional average. The bottom panel of Figure 1 helps determine the statistical significance of the changes; here, the local slopes lie entirely within the 95% prediction bands, so we cannot conclude definitively any departure from a constant mean.

The analysis so far makes no allowance for several sources of variation, such as strength of the opposing side or condition of the playing field. Thus we might expect the data to show more variability than would be expected for the geometric model. We consider the negative binomial distribution with shape parameter w and mass function

$$P(T_i = y) = \binom{w+y-1}{y} \frac{\mu_i^y}{(1+\mu_i)^{w+y}}.$$

This distribution has mean $w\mu_i$ and variance $w\mu_i(\mu_i + 1)$. When $w < 1$, the distribution is overdispersed with respect to the geometric distribution, and the discrete hazard function decreases to an asymptote at $1/(1 + \mu_i)$. This is in qualitative agreement with behavior observed in cricket scores by Kimber and Hansford (1993).

Using a globally constant model for the dispersion parameter and the local log-quadratic model for μ, we obtain the maximum likelihood estimate $\hat{w} = 0.80$. A quantity of interest is the difference $w\hat{\mu}_1(x) - \hat{\mu}_0(x)$ where $\hat{\mu}_1(x)$ and $\hat{\mu}_0(x)$ are estimates under the negative binomial and geometric models respectively. This quantity estimates the effect of censoring on a batsman's average. From the plot in Figure 2, the censoring effect was between 1.5 and 2 runs per innings up to 1990, declining in later years where there is less censoring. From this model, we estimate that Border's average would have been between 1 and 2 runs higher, had he been able to complete all his 'not out' innings. This might come as a surprise to some cricket fans, who often claim not out innings boost, rather than hinder, a batsman's average.

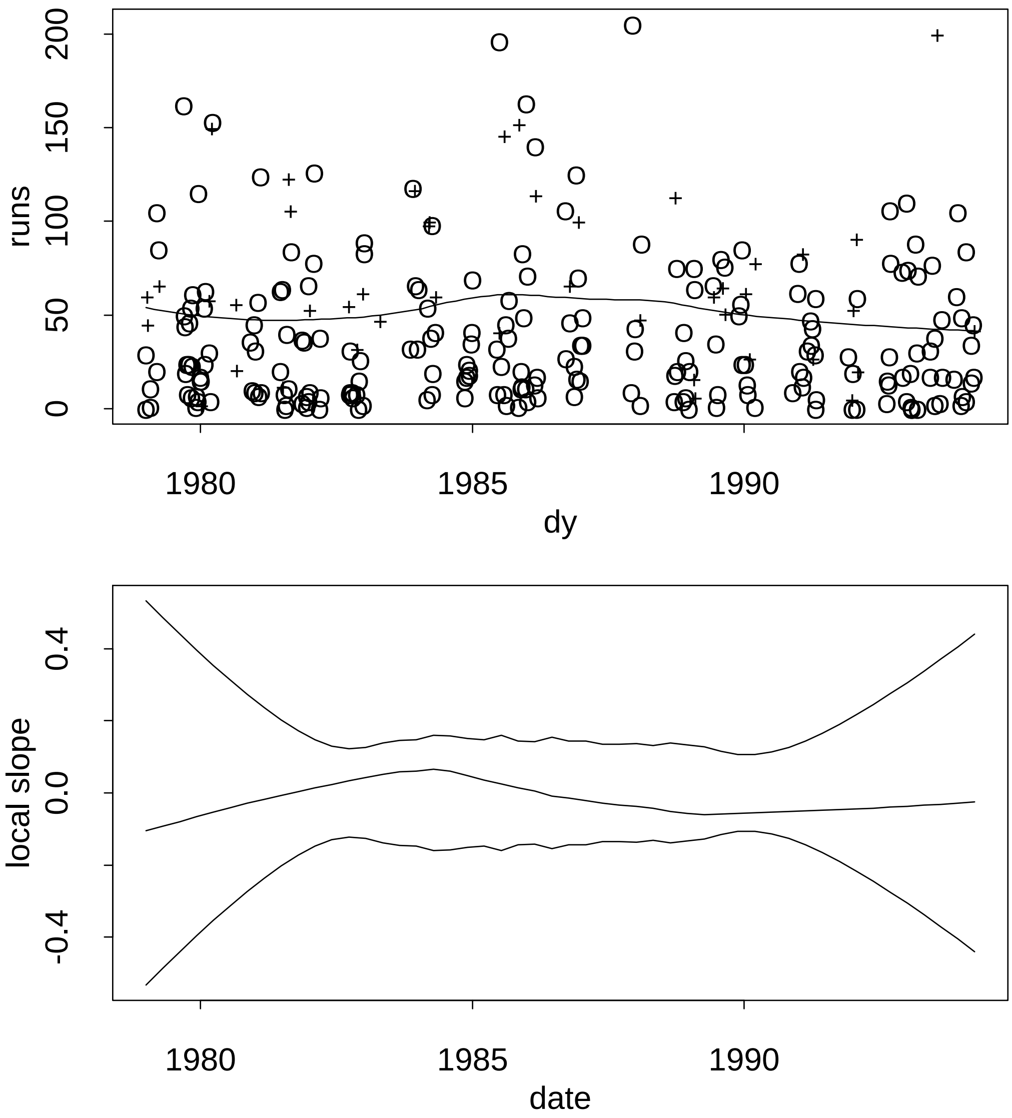

Figure 1: Batting dataset, Runs scored and smooth (top) and local slopes and prediction bands under the constant mean hypothesis (bottom). Uncensored (out) scores denoted by 'o'; censored (not out) scores by '+'.

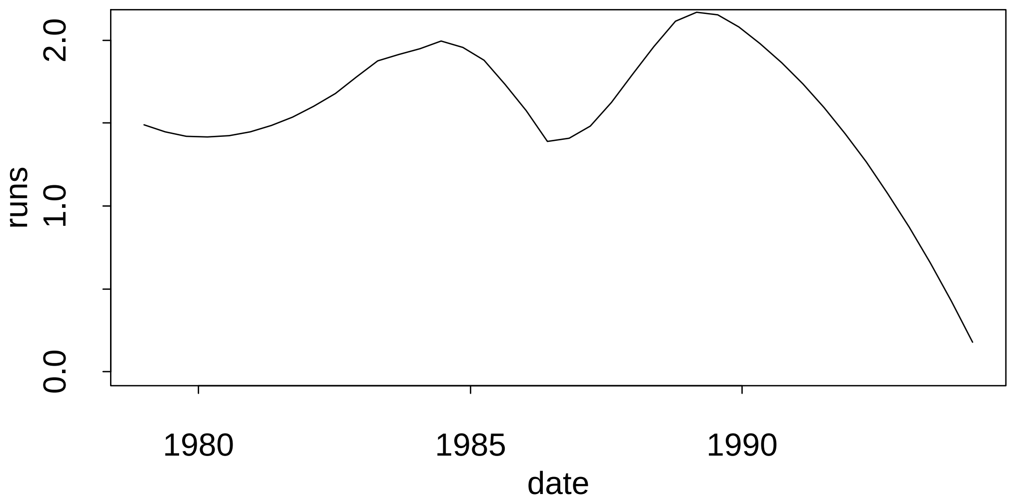

Figure 2: An estimate of the penalty, in runs per innings, caused by censoring in Border's batting performance.

Our second example is the Stanford heart transplant data (Miller and Halpern 1982), where we model the mean survival time as a function of the patient's age at transplant. We use a Weibull model with constant shape parameter,

$$f(t, a_i, b) = a_i b t^{b-1} \exp(-a_i t^b).$$

A local quadratic model, with an 80% nearest neighbor bandwidth, is used to estimate $\log(a_i)$; the maximum likelihood estimate of b is $\hat{b} = 0.625$. Figure 3 shows the results. The smooth in the top panel shows an initially constant mean survival time, followed by a sharp decrease. The bottom panel confirms the significance of the decrease, particularly between ages 45 and 55 where the local slope is outside the prediction bands.

Having determined that the decrease is significant, the question becomes whether this feature has been adequately modeled by the smooth in Figure 3. This question

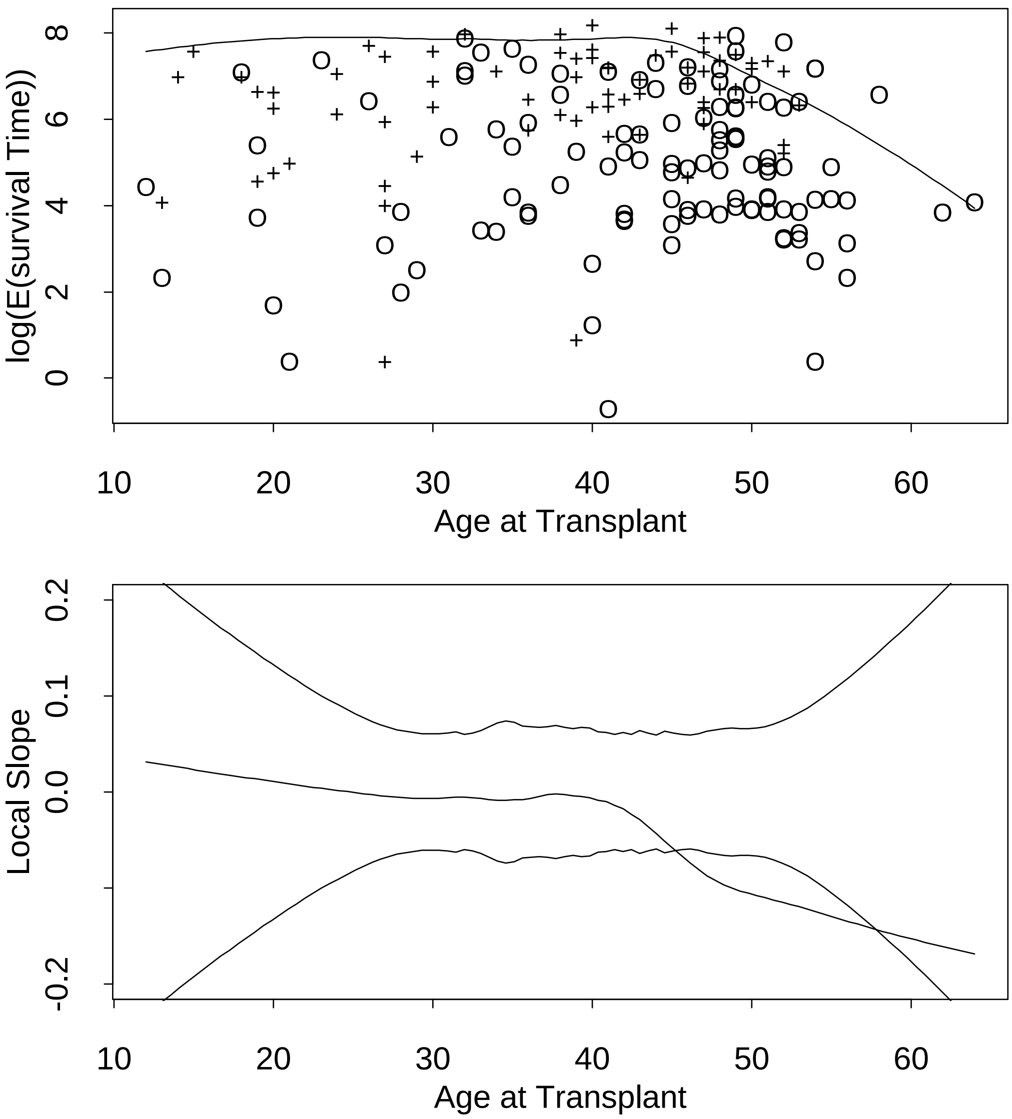

Figure 3: Heart transplant data and smooth (top) and local slopes and prediction bands under the constant mean hypothesis (bottom).

is addressed by the residual plot in Figure 4. The smoothed residuals show only minor fluctuations, and provide no strong evidence of lack of fit.

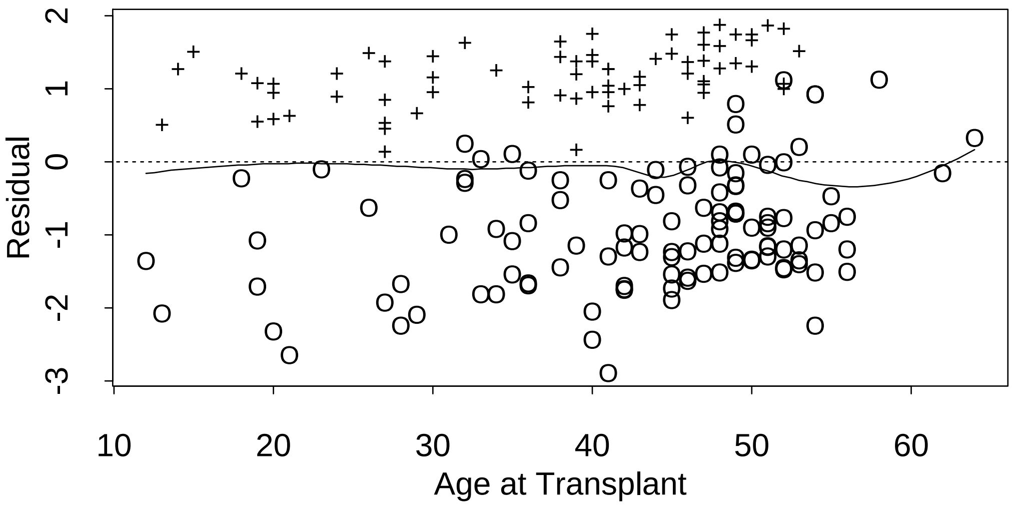

Figure 4: Smoothed residual plot for the heart transplant data. The smooth curve is a local quadratic fit, with nearest neighbor span $\alpha = 0.6$.

6 Some Theory for Local Likelihood

In this section we study some basic theoretical properties of the local likelihood estimate. Many results are similar to existing results for uncensored local regression and likelihood (Tsybakov 1986; Fan 1992; Fan, Heckman, and Wand 1995; Loader 1996a among others) so we only sketch proofs here. Theorem 1 establishes existence and uniqueness of the local likelihood estimate. Consistency is established in Theorem 2. Theorem 3 derives a bias-variance expansion.

Theorem 1 *Suppose the parameter space is an open interval $b_0 < \theta(x) < b_1$; $b_0 = -\infty$ and/or $b_1 = \infty$ are allowed. Also suppose that for all y and c, $l(y, \theta, c)$ is upper bounded and concave.*

Let $\Xi \subset \{1, \ldots, n\}$ be the set of observations with $w_i(x) > 0$ and $l(Y_i, \theta, c_i) \to -\infty$ as $\theta \downarrow b_0$ or $\theta \uparrow b_1$. Let $\mathbf{X}^*$ be a matrix with rows $A(x_i - x)^T; i \in \Xi$.

If $\mathbf{X}^*$ has full column rank, the local likelihood estimate exists and is the unique solution of the local likelihood equations (2).

Proof: Since $\mathbf{X}^*$ has full rank, the local likelihood surface $\mathcal{L}_x(a)$ is strictly concave. Hence at most one maximizer exists, and this must satisfy (2) since the parameter space is assumed to be open.

To establish existence, let $\hat{a}^{(j)}$ be a sequence of parameter vectors satisfying

$$\mathcal{L}_x(\hat{a}^{(j)}) \to \sup_a \mathcal{L}_x(a).$$

Then for each $i \in \Xi$,

$$b_0 < \liminf_{j \to \infty} \left\langle \hat{a}^{(j)}, A(x_i - x) \right\rangle \leq \limsup_{j \to \infty} \left\langle \hat{a}^{(j)}, A(x_i - x) \right\rangle < b_1$$

and the full rank of $\mathbf{X}^*$ implies $\{\hat{a}^{(j)}\}$ lie in a compact set. Thus the sequence $\{\hat{a}^{(j)}\}$ has a limit point, which maximizes the local likelihood.

$\square$

To derive asymptotic properties of the local likelihood estimate, we must make assumptions as to how the design sequence $\{x_i\}$, censoring times $\{c_i\}$ and bandwidth $h = h_n$ behave as $n \to \infty$. Our basic assumption is that there exist $m(x)$ and $G_x(c)$, both continuous in x, such that

$$\frac{1}{nh^d} \sum_{i=1}^{n} K\left(\frac{x_i - x}{h}, c_i\right) \to m(x) \int \int K(v, c) dG_x(c) dv \tag{7}$$

for any bounded continuous function $K(v,c)$ with $K(v,c)=0$ whenever $|v|>1$. We also need to assume $h \to 0$, and that $nh^{-d}E(\mathbf{J})$ has a nonsingular limit. The existence of this limit is a consequence of (7); nonsingularity excludes singular designs and some obviously silly cases such as $c_i = -\infty$ for all i.

The statement of (7) as an assumption is slightly unusual. If the x_i and c_i are random variables with marginal density $m(x)$ and conditional density $G_x(c)$, (7) is essentially a consequence of standard consistency results for kernel density estimation, provided $nh^d \to \infty$. If x_i form a grid, quadrature arguments establish (7), again requiring $nh^d \to \infty$. Boundaries are also covered, by appropriately restricting the range of v.

A consequence of (7) and $h \to 0$ is that for any absolutely integrable function K^*,

$$\frac{1}{nh^d}\sum_{i=1}^{n} K^*\left(Y_i, \frac{x_i - x}{h}, c_i\right) \to m(x)\int\int K^{**}(v,c)dG_x(c)dv \tag{8}$$

where

$$K^{**}(v,c) = \int_{-\infty}^{c} K^*(y,v,c)f(y,\theta(x))dy + K^*(c,v,c)(1-F(c,\theta(x))).$$

If K^* is square integrable, the convergence will be at rate $O_p((nh^d)^{-1/2})$.

Theorem 2 *Let* $\mathbf{H}$ *be a diagonal matrix whose entries are powers of* h*, defined so that*

$$\mathbf{H}^{-1}A(v) = A(v/h).$$

Then

$$\mathbf{H}\hat{a} \to \begin{pmatrix} \theta(x) \\ 0 \\ \vdots \\ 0 \end{pmatrix}.$$

In particular, this implies $\hat{a}$ is consistent; $\hat{\theta}(x) = \hat{a}_0 \to \theta(x)$.

Proof: Applying (8) to $K_a^*(y, v, c) = W(v)l(y, v, c)$ gives

$$\frac{1}{nh}\mathcal{L}_x(\mathbf{H}^{-1}a) \quad \to \quad m(x)\int\int K_a^{**}(v, c)dG_x(c)dv. \tag{9}$$

This holds pointwise for any a. Using the assumption of a concave log-likelihood, both sides of this equation are concave; it follows the convergence must be uniform on any bounded set; in particular in a neighborhood around the maximum. To complete the proof, we now note the left hand side of (9) is maximized at $a = \mathbf{H}\hat{a}$, and the right hand side is maximized at $a = (\,\theta(x) \quad 0 \quad \dots \quad 0\,)^T$. □

Having established consistency, we can perform a standard quadratic expansion in a neighborhood of the maximum to get a more explicit approximation to $\hat{a}$. We state the result only for the one dimensional case; in the multidimensional case similar results hold, but involve all partial derivatives of $\theta(x)$ of order $p+1$. We remark that as an informal approximation, the result is just a two term Taylor series; assumptions like (7) are needed only to ensure the random error term has the claimed size.

Theorem 3 *Fix the fitting point x. Let a^* denote a parameter vector consisting of the true derivatives of $\theta(x)$ up to order p. Then*

$$\begin{aligned}\mathbf{H}(\hat{a} - a^*) \quad = \quad & \mathbf{H}E(\mathbf{J})^{-1}\sum_{i=1}^{n} w_i(x)A(x_i - x)\dot{l}(Y_i, \theta_i, c_i) \\ & + \mathbf{H}E(\mathbf{J})^{-1}\frac{\theta^{(p+1)}(x)}{(p+1)!}\sum_{i=1}^{n} w_i(x)\,(x_i - x)^{p+1}\,A(x_i - x)v_i \\ & + o(h^{p+1}) + o_P((nh)^{-1/2}).\end{aligned}$$

where $v_i = -E(\ddot{l}(Y_i, \theta_i, c_i))$. The random term has covariance function

$$\Sigma(x, x') = \mathbf{H}E(\mathbf{J}(x))^{-1}\sum_{i=1}^{n} w_i(x)w_i(x')v_iA(x_i - x)A(x_i - x')^T E(\mathbf{J}(x'))^{-1}\mathbf{H}.$$

Proof: Let $\tilde{\theta}_i = \langle \hat{a}, A(x_i - x) \rangle$. A Taylor series expansion and the consistency result of Theorem 2 give

$$\dot{l}(Y_i, \tilde{\theta}_i, c_i) = \dot{l}(Y_i, \theta_i, c_i) + (\theta_i^* - \theta_i)\ddot{l}(Y_i, \theta_i, c_i) + (\tilde{\theta}_i - \theta_i^*)\ddot{l}(Y_i, \theta_i, c_i) + o_P(1).$$

Using the approximation

$$\theta_i^* - \theta_i = \frac{(x_i - x)^{p+1}}{(p+1)!}\theta^{(p+1)}(x) + o_P(h^{p+1})$$

uniformly on the smoothing window, and noting $\hat{\theta}_i - \theta_i^* = \langle \hat{a} - a^*, A(x_i - x) \rangle$, the local likelihood equations (2) become

$$\begin{aligned} 0 \quad = \quad & \frac{1}{nh}\mathbf{H}^{-1}\sum_{i=1}^{n} w_i(x) A(x_i - x)\dot{l}(Y_i, \theta_i, c_i) \\ & + \frac{1}{nh}\mathbf{H}^{-1}\frac{\theta^{(p+1)}(x)}{(p+1)!}\sum_{i=1}^{n} w_i(x)\,(x_i - x)^{p+1}\, A(x_i - x)\ddot{l}(Y_i, \theta_i, c_i) \\ & - \frac{1}{nh}\mathbf{H}^{-1}\mathbf{J}(\hat{a} - a^*) + o(h^p) + o_P((nh)^{-1/2}). \end{aligned}$$

Solving for $\hat{a} - a^*$ yields the result. Justification for replacing $\ddot{l}$ by $E(\ddot{l})$ is a consequence of (8). The covariance function follows by direct evaluation. □

References

Brillinger, D. R. (1977). Discussion of Stone (1977). *The Annals of Statistics* **5**, 622–623.

Fan, J. (1992). Design adaptive nonparametric regression. *Journal of the American Statistical Association* **87**, 998–1004.

Fan, J. and Gijbels, I. (1994). Censored regression: Local linear approximations and their applications. *Journal of the American Statistical Association* **89**, 560–570.

Fan, J., Heckman, N. E., and Wand, M. P. (1995). Local polynomial kernel regression for generalized linear models and quasi-likelihood functions. *Journal of the American Statistical Association* **90**, 141–150.

Firth, D., Glosup, J., and Hinkley, D. V. (1991). Model checking with nonparametric curves. *Biometrika* **78**, 245–252.

Gentleman, R. and Crowley, J. (1991). Local full likelihood estimation for the proportional hazards model. *Biometrics* **47**, 1283–1296.

Gray, R. J. (1996). Hazard regression using ordinary nonparametric regression smoothers. *Journal of Computational and Graphical Statistics* **5**, 190–207.

Henderson, R. (1924). Some points in the general theory of graduation. In *Proceedings of the International Mathematical Congress*, pp. 815–820.

Hjort, N. L. (1997). Dynamic likelihood hazard rate estimation. *Biometrika* **84**, to appear.

Kimber, A. C. and Hansford, A. R. (1993). A statistical analysis of batting in cricket. *Journal of the Royal Statistical Society, Series A* **156**, 443–455.

Li, G. and Doss, H. (1995). An approach to nonparametric regression for life history data using local linear fitting. *The Annals of Statistics* **23**, 787–823.

Loader, C. R. (1996a). Local likelihood density estimation. *The Annals of Statistics* **24**, to appear.

Loader, C. R. (1996b). *Local Regression and Likelihood.* Electronic book.
`http://cm.bell-labs.com/stat/project/locfit/index.html`

McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models.* London: Chapman and Hall.

Miller, R. and Halpern, J. (1982). Regression with censored data. *Biometrika* **69**, 521–531.

Schmee, J. and Hahn, G. J. (1979). A simple method for regression analysis with censored data (with discussion). *Technometrics* **21**, 417–434.

Staniswalis, J. G. (1989). On the kernel estimate of a regression function in likelihood based models. *Journal of the American Statistical Association* **84**, 276–283.

Stone, C. J. (1977). Consistent nonparametric regression (with discussion). *The Annals of Statistics* **5**, 595–645.

Sun, J., Loader, C., and McCormick, W. (1994). Confidence bands in generalized linear models. Technical report, Department of Statistics, University of Michigan. `http://www.stat.lsa.umich.edu/~jsun/glm.ps`

Sun, J. and Loader, C. R. (1994). Simultaneous confidence bands in linear regression and smoothing. *The Annals of Statistics* **22**, 1328–1345.

Tibshirani, R. J. and Hastie, T. J. (1987). Local likelihood estimation. *Journal of the American Statistical Association* **82**, 559–567.

Tsybakov, A. B. (1986). Robust reconstruction of functions by the local-approximation method. *Problems of Information Transmition* **22**, 69–84.

Wu, L. and Tuma, N. B. (1990). Local hazard models. *Sociological Methodology* **20**, 141–180.