Journal of Science and Technology, 3(2): 391-398, July 2013

LOCAL LIKELIHOOD ESTIMATION FOR THE PROPORTIONAL HAZARDS MODEL AND COMPARISON WITH GLOBAL LIKELIHOOD

MD. SANWAR HOSSAIN[1], M. ATAHARUL ISLAM[2], DELUAR J. MOLOY[*3], and M. ROUNGU AHMMAD[1]

[1]Department of Statistics, Jagannath University, Dhaka; [2]Department of Applied Statistics, East West University, Dhaka, [3]Department of Statistics, Mawlana Bhashani Science and Technology University, Tangail-1902, Bangladesh

Abstract

In this Study, we develop the idea of scatterplot smoothing using local likelihood and extend smoothing idea to other kind of data whose relationship is expressible through likelihood function. Our idea is to replace a simple parametric function and estimate the unspecified smooth function locally. This procedure is designed for nonparametric regression modeling in situations where a non- Gaussian likelihood is appropriate, such as the Proportional hazards model. The local likelihood concept is used in the Proportional hazards model for applications to the artificially generated data and also for a real life data such as the Health Retirement Survey (HRS) data. The application of local likelihood is local likelihood procedure appear to perform better result than global likelihood procedure.

Key Words: Local Likelihood, Smoothing, Neighborhood, Proportional Hazards model

Introduction

The problem of smoothing sequences of observations is important in many branches of science. Smoothing methods have been widely used to estimate trends in economic time series. Local fitting of polynomials has been used for many decades to smooth time series plots (Macauley, 1931).

Local regression is a popular form of nonparametric regression, combining excellent theoretical properties with conceptual simplicity and flexibility to find structure in many datasets. Stone (1977), Cleveland (1979), Cleveland and Devlin (1988) discuss a multivariate setting. Local regression may be viewed as a special case of the local likelihood procedure introduced by Tibshirani and Hastie (1987). This procedure is designed for nonparametric regression modeling in situations where a non- Gaussian likelihood is appropriate, such as logistic regression and proportional hazards model (Cox, 1972).

Cleveland (1979) showed with little additional cost by computing and plotting smoothed points, the visual information on a scatterplot can be greatly enhanced. Robust locally weighted regression is a method for smoothing a scatterplot, (x_i, y_i), $i = 1, 2, ..., n$, in which the fitted value at x_k is the value of a polynomial fit to the data using weighted least squares, where the weight for (x_i, y_i) is large if x_i is close to x_k and small if it is not. A robust fitting procedure is used that guards against deviant points distorting the smoothed points. Visual, computational and statistical issues of robust locally weighted regression are discussed.

[*] Author for correspondence: mdeluar@yahoo.com

Tibshirani and Hastie (1987) discussed a scatterplot is applying to data of the form $\{x_1, y_1), (x_2, y_2), ..., (x_n, y_n)\}$ and used local fitting to estimate the dependence of Y on X. A simple example is the running lines smoother, which fits a least square line to the y values falling in a window around each x value. The value of the estimated function at x is given by the value of the least squares line at x. A smoother generalizes the least squares line, which assumes that the dependence of Y on X is linear.

In this paper, we extend smoothing ideas to other kinds of data. In particular, we consider (X, Y) data whose relationship is expressible through a likelihood function and replace a simple parametric function like $\beta_0 + \beta_1$ appearing in the likelihood with an unspecified smooth function $s(x)$ and to estimate $s(x)$ locally. Take for example the situation in which is 0-1 response and x is a covariate. The usual linear logistic model assumes that , where $(p(x) = Pr(Y = 1|X = x)$. Also appearing is a smooth estimate, based on the more general model $\log\big(p(x)/(1-p(x))\big) = s(x)$, with $s(x)$ an arbitrary smooth function. As was the case in the scatterplot example, the smooth does a better job of capturing the relationship between Y and X than the line does. The smooth is produced by technique call local likelihood estimation. The basic idea is a simple extension of the local fitting technique used in a scatterplot smoothing.

Hastie and Tibshirani (1987) introduced the Local Scoring procedures which replace the linear form $\sum X_j\beta_j$ by a sum of smooth functions $\sum s_j(X_j)$. The $s_j(.)$ are unspecified functions that are estimated using scatterplot smoothers. The technique is applicable in any likelihood-based regression model; the class of Generalized Linear Models contains many of these. In this class, the Local Scoring procedure replaces the linear predictor $\eta = \sum X_j\beta_j$ by the additive predictor $\sum s_j(X_j)$; hence, the name Generalized Additive Models. Local Scoring can also be applied to non-standard models like Cox's proportional hazards model for survival data.

Cleveland and Devlin (1988) described locally weighted regression, or loess, is a way of estimating a regression surface through a multivariate smoothing procedure, fitting a function of the independent variables locally and in a moving fashion analogous to how a moving average is computed for a time series. With local fitting they can estimate a much wider class of regression surfaces than with the usual classes of parametric functions, such as polynomials. The goal of this paper is to show, through applications, how loess can be used for three purposes: data exploration, diagnostic checking of parametric models, and providing a nonparametric regression surface. Along the way, the following methodology is introduced: (a) a multivariate smoothing procedure that is an extension of univariate locally weighted regression; (b) statistical procedures that are analogous to those used in the least-squares fitting of parametric functions; (c) several graphical methods that are useful tools for understanding loess estimates and checking the assumptions on which the estimation procedure is based; and (d) the M plot, an adaptation of Mallow's C_p procedure, which provides a graphical portrayal of the trade-off between variance and bias, and which can be used to chose the amount of smoothing.

A Review of Scatterplot Smoothing

Given independent data pairs $\{x_1, y_1), (x_2, y_2), ..., (x_n, y_n)\}$, assumed to be realization of a response variable *Y* and a predictor *X*, a scatterplot smoother produces a decomposition of the form

$$y_i = s(x_i) + \in_i \tag{1}$$

Where $s(.)$ a "smooth" is a function and $\in_i$ is a residual error. We will not define exactly what "smooth" means here; vaguely speaking, we are thinking of $s(.)$ as a function less smooth than a straight line but smoother than an interpolating polynomial.

There are many ways to estimate $s(.)$. Tibshirani and Hastie (1987) used the method of "local averaging". It is motivated as follows. If we knew the joint distribution of Y and X, a reasonable way to find $s(.)$ would be to minimize $E(Y - s(X))^2$, where the expectation is taken over the joint distribution. Conditioning on $X = x$, this has solution $\hat{s}(x) = E(Y|X = x)$ for each x. In practice, we don't know the joint distribution but have only a sample from it. The idea, then, is to estimate $E(Y|X = x)$ from the data. This leads to the class of local average estimate for $s(.)$:

$$\hat{s}(x_i) = Ave_{j \in N_i} y_j \tag{2}$$

Where "*Ave*" represents some averaging operator like mean or median and *Ni* is a "neighborhood" of x_i (a set of indices of points whose x values are "close" to x_i). The same types of neighborhoods considered by Tibshirani and Hastie (1987) are symmetric nearest neighborhoods. Associated with a neighborhood is the span or window size w; this is the proportion of the total points that each neighborhood contains. Let $[x]$ represent the integer part of x and assume that $[wn]$ is odd. Then a span w symmetric nearest neighborhood contains $[wn]$ points; the *ith* point plus $([wn] - 1)/2$ points on either side of the *ith* point. Assuming that the data points are sorted by increasing x value, a formal definition is

$$N_i = \left\{ \max\left(i - \frac{[wn]-1}{2}, 1\right), \ldots, i-1, i, i+1 \ldots, \min\left(t + \frac{[wn]-1}{2}, n\right) \right\} \tag{3}$$

Note that the neighborhoods are truncated near the end points if $([wn] - 1)/2$ points are not available. The span controls the smoothness of the resulting estimate: larger spans will produce smoother (less variable) estimates but with possibly more bias. A span of $1/n$ corresponds to 1 point per neighborhood. The span is either fixed a priori or chosen adaptively from the data.

If *Ave* stands for arithmetic mean, then $\hat{s}(.)$ is the running mean, the simplest possible scatterplot smoother. The running mean is not a satisfactory smoother because it creates large biases at the endpoints and doesn't reproduce straight lines (i.e. if the data lie exactly along a straight line, the smooth of the data will not be a straight line). A slight refinement of the running average, the running lines smoother alleviates these problems. The running lines estimate is defined by

$$\hat{s}(x_i) = \hat{\beta}_{0i} + \hat{\beta}_{1i} x_i \tag{4}$$

Where $\hat{\beta}_{0i}$ and $\hat{\beta}_{1i}$ are the least squares estimates for the data points in N_i:

$$\hat{\beta}_{1i} = \frac{\sum_{j \in N_i} (x_j - x_i) y_j}{\sum_{j \in N_i} (x_j - x_i)^2} \tag{5}$$

$$\hat{\beta}_{0i} = y_i - \hat{\beta}_{1i} x \bar{x}_i$$

And $\bar{x}_i = \frac{1}{n}\sum_{j\in N_i} x_j, \bar{y}_i = \frac{1}{n}\sum_{j\in N_i} Y_j$

The running lines smooth is the most obvious generalization of the least squares line. When every neighborhood contains 100% of the data points, the smooth agrees exactly with the least squares line. For smaller spans, it produces less smooth estimates. Although very simple is its nature, the running lines smoother produces reasonable results and has the advantage that the estimates can be updated. That is, to find $\hat{s}(x_{i+1})$ from $\hat{s}(x_i)$, only an 0(1) operation is needed. This makes the entire smoothing algorithm 0(n).

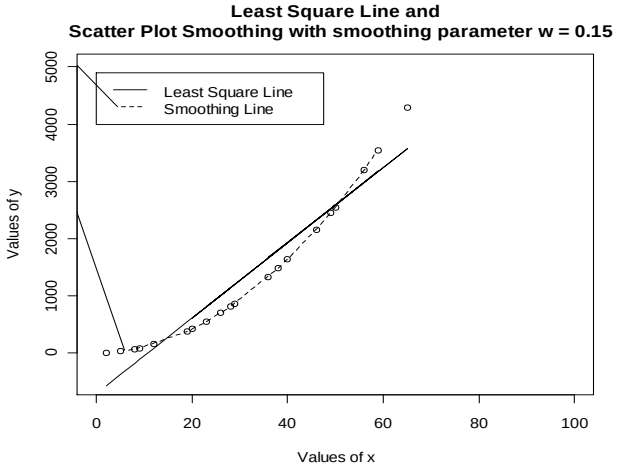

Fig. 1. Least Squares Line and Scatterplot Smooth using smoothing parameter w = 0.15

In Figure 1 the least squares line has been replaced by a "scatter plot smooth." This smooth was computed by a type of local averaging-around each *X* value a window of 3 points was formed and a least squares line was fit to the points in the window. The value of the smooth at *X* is given by the value of the "local line" at *X*. as we can see, the smooth captures the trend of the data better than the least squares line.

The Local Likelihood Generalization

Score Vector

$$U_i(\beta_i) = \left[\frac{\partial \log L_i(\beta_i)}{\partial \beta_i}\right]_{p\times 1} \tag{6}$$

Be the score vector for $x \in N_i$.

The *hth* element of the score vector will be

$$U_i(\beta_i) = \left[\frac{\partial \log L_i(\beta_i)}{\partial \beta_i}\right] = \sum_{l\in D\cap N_i}\left[xlh - \frac{\sum_{j\in r_i\cap N_i} x_{jh}\exp(x_j\beta_i)}{\sum_{j\in r_i\cap N_i}\exp(x_j\beta_i)}\right] \tag{7}$$

$$= \sum_{l\in D\cap N_i}\left[xlh - Ajh\right], h = 1,2,..., p$$

where, $A_{jh} = \frac{\sum_{j \in R_i \cap N_i} x_{jh} \exp(x_j \beta_i)}{\sum_{j \in R_i \cap N_i} \exp(x_j \beta_i)}$

Information matrix

Let

$$I_i(\beta_i) = \left[-\frac{\partial^2 \log L_i(\beta_i)}{\partial \beta_{ih} \partial \beta_{ik}} \right]_{p \times p} \tag{8}$$

Be the sample information matrix for $x \in N_i$. The $(h, k)th$ element of $I_i(\beta_i)$ is given by

$$I_{ihk}(\beta_i) = -\frac{\partial^2 \log L_i(\beta_i)}{\partial \beta_{ih} \partial \beta_{ik}} \tag{9}$$

$$= \frac{\sum_{j \in R_i \cap N_i x_j h x_{jk}} \exp(x_j \beta_i)}{\sum_{j \in R_i \cap N_i} \exp(x_j \beta_i)} - A_{jh} \times A_{jk}$$

$(h, k = 1, 2, ..., p)$

Then $\hat{\beta}$ is obtained by the iterative use of this following equation:

$$\beta_i^{(1)} = \beta_i^{(0)} + [I(\beta_i^{(0)})] - 1 U(\beta_i^{(0)}) \tag{10}$$

The process is repeated until successive β_i estimates agree to a specified extent and $U_i(\beta_i) = 0$ at convergence.

$\hat{\beta}_i$ obtained in this way is called maximum partial likelihood estimate of β_i.

Data and Application

Here we will discuss the application of Proportional hazards model for artificially generated data and Health and Retirement Survey (HRS) data for local likelihood and Global Likelihood. Normally we use the global likelihood to maximize the parameter but Tibshirani and Hastie (1987) gave the concept of Local Likelihood Estimation.

Application of Health and Retirement Survey (HRS) data of Cox proportional Hazard for Local and Global Likelihood

Cox (1972) and others have shown that this partial log-likelihood can be treated as ordinary log likelihood to derive valid (partial) MLEs of β. Therefore, we can estimate hazard ratios and confidence intervals using maximum likelihood techniques discussed previously. The only difference is that these estimates are based on the partial as opposed to the full likelihood. The partial likelihood is valid when there are no ties in the data set. That is, no two subjects have the same event time. If there are ties in the data set, the true partial log-likelihood function involves permutations and can be time-consuming to compute. In this case, either the Breslow (1974) or Efron (1977) approximations to the partial log-likelihood can be used.

Table 1. Estimate of Local Likelihood when n = 20 and smoothing parameter w = 0.75

Model No.	Gender	exp(coef(gender))	exp(-coef(gender))	se(gender)
Model 1	-1.5097567	0.2209637	4.525629	1.1429716
Model 2	-1.1331910	0.3220041	3.105550	0.9318154
Model 3	-0.8006066	0.4490565	2.226891	0.8722951
Model 4	-0.8903731	0.4105026	2.436038	0.8375012
Model 5	-0.8959169	0.4082331	2.449581	0.8902035
Model 6	-1.0088707	0.3646305	2.742502	0.8756544

Model No.	bmi	exp(coef(bmi))	exp(-coef(bmi))	se(bmi)
Model 1	0.04950753	1.050754	0.9516980	0.11520144
Model 2	0.13730188	1.147174	0.8717070	0.13860167
Model 3	0.01858298	1.018757	0.9815886	0.11846284
Model 4	0.03874773	1.039508	0.9619934	0.10297190
Model 5	0.11889196	1.126248	0.8879037	0.12950792
Model 6	0.08024652	1.083554	0.9228888	0.05997114

Table 1 shows the estimate of parameter, standard error, positive exponential of parameter and negative exponential of parameter.

Table 2. Estimate of Global Likelihood for Cox proportional hazards Model

gender	exp(coef(gender))	exp(-coef(gender))	se(gender)
-1.04605687	0.35132032	2.84640522	0.80079853

bmi	exp(coef(bmi))	exp(-coef(bmi))	se(bmi)
0.09416279	1.09873859	0.91013459	0.04661575

If we compare the Table 1 and Table 2, we see that the estimate of parameters and standard errors from locally and globally fitting of the proportional hazards model is more or less similar.

Graphical representation of Log-Likelihood

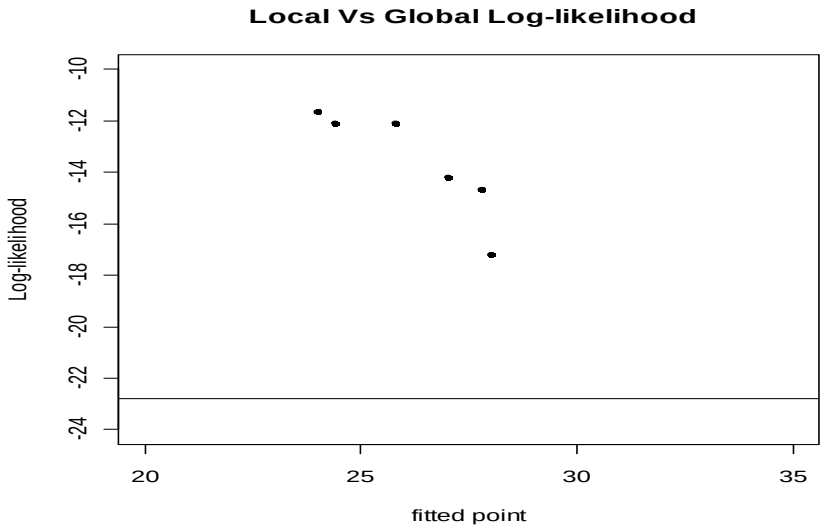

Fig. 2. Local (w = 0.75) Vs Global Log-likelihood when n = 20

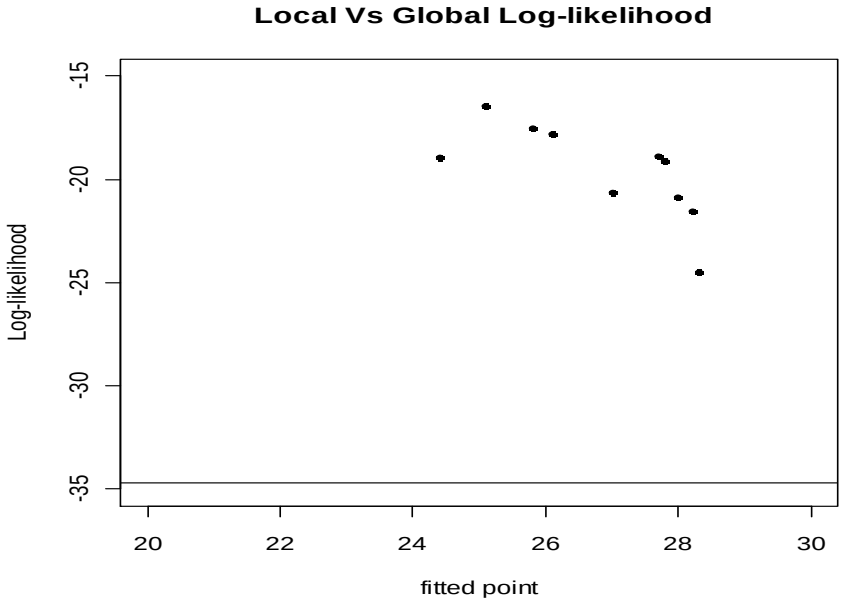

Fig. 3. Local (w = 0.70) Vs Global Log-likelihood when n = 30

Summary

Figure 2 and 3 present the graphical representation of Global and Local likelihood. Fitting local likelihood is several times for different number of observations and different smoothing parameter. X-axis represents the fitted point and Y-axis represents the Log-

likelihood. Solid circle denotes the log-likelihood against the locally fitted points. We have drawn a line parallel to the X-axis of the value of Log-likelihood for globally fitting. From the above graph we see that, the maximum value of log-likelihood is achieved when Cox-proportional hazard model is fitted locally.

Conclusion

In this study we have examined the utility of the Local likelihood approach for constructing the local likelihood. We have considered the symmetric nearest neighborhood concept. The symmetric neighborhood concept by using local fitting to estimate dependence on explanatory variable is called the running lines smoother. We have shown that locally fitted scatterplot is smoother than the least squares line as displayed in Figure 1. The smoothing scatterplot is more predictable than the least squares line.

An application of the HRS data in the proportional hazards model also gives the best fitted model than globally fitted model. For different smoothing parameter, locally fitted log likelihood is maximum than globally fitted log likelihood.

References

Breslow, N. (1974). Covariance analysis of censored survival data, Biometrics, **30:** 89-99.

Cleveland, W. S. (1979). Robust Locally Weighted Regression and Smoothing Scatterplots, *the Journal of the American Statistical Association*, **74**: 828-836.

Cleveland, W. S. & Devlin, S. J. (1988). Locally weighted regression: an approach to regression analysis by local fitting, *Journal of the American Statistical Association,* **83**: 596-610.

Cox, D. R. (1972). Regression Models and Life Tables, *The Journal of the Royal Statistical Society*, Ser. B, **34**: 187-202.

Efron, B. (1977). The efficiency of Cox's likelihood function for censored data, *The Journal of the American Statistical Association,* **72**: 557-565

Macauley, F. R. (1931). Smoothing of Time Series, *National Bureau of Economic Research,* New York

Stone, C. J. (1977). Consistent nonparametric regression (with discussion). *The Annals* of Statistics, **5:** 595-620.

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