SLAC REPORT-275
UC-32
(M)

LOCAL LIKELIHOOD ESTIMATION*

ROBERT JOHN TIBSHIRANI

Stanford Linear Accelerator Center
Stanford University
Stanford, California 94305

December 1984

Prepared for the Department of Energy
under contract number DE-AC03-76SF00515

Printed in the United States of America. Available from the National Technical Information Service, U.S. Department of Commerce, 5285 Port Royal Road, Springfield, Virginia 22161. Price: Printed Copy A04; Microfiche A01.

*Ph.D Dissertation

Abstract

Given scatterplot data $\{(X_1,Y_1),(X_2,Y_2),...(X_n,Y_n)\}$, Y being a response and X a predictor, *a scatterplot smoother* uses *local averaging* to estimate the dependence of Y on X. A simple example is the *running lines smoother*, which fits a least squares line to the the Y values falling in a window around each X value. A smoother generalizes the least squares line, which assumes the dependence of Y on X is linear.

In this work, we extend the idea of local averaging to likelihood-based regression models. One application is in the class of *generalized linear models* (Nelder and Wedderburn (1972)) We enlarge this class by replacing the covariate form $x\beta$ with an unspecified smooth function $s(x)$. This function is estimated from the data by a technique we call "*Local Likelihood Estimation*"— a type of local averaging. Multiple covariates are incorporated through a forward stepwise algorithm.

We also apply the local likelihood technique to the proportional hazards model of Cox (1972), for censored data. The proportional hazards assumption $\lambda(t\,|x) = \lambda_0(t)\exp(x\beta)$ is replaced by $\lambda(t\,|x) = \lambda_0(t)\exp(s(x))$, and the function $s(x)$ is estimated from the data by local likelihood estimation.

In a number of real data examples, the local likelihood technique proves to be effective in uncovering non-linear dependencies.

Finally, we give some asymptotic results for local likelihood estimates and provide some methods for inference.

Work supported by the Department of Energy under contracts DE-AC03-76SF00515 and DE-AT03-81-ER10843, and by the Office of Naval Research under contract ONR N00014-81-K-0340, and by the U.S. Army Research Office under contract DAAG29-82-K-0056.

Acknowledgments

First and foremost, I wish to express my great appreciation to my adviser Bradley Efron. With his rare insight and constant encouragement, he has truly been an inspiration. Brad has taught me many things, not the least of which is to take the science of statistics seriously.

I am very lucky to have met Trevor Hastie here at Stanford— I now have a talented collaborator, and more importantly, a friend for life. (you too Lynda!).

Good research requires a stimulating and friendly atmosphere. We have such an atmosphere at Sequoia Hall, thanks to everyone— faculty, staff and students, all held together by our indomitable leader, Judi Davis. Soon after my arrival, Andreas Buja, Jerome Friedman and Werner Stuetzle made a big impact on me, and changed the way that I thought about applied statistics. I would especially like to thank Jerry for his generous support the past three years. I have also benefitted greatly from interaction with my classmates. I am fortunate to have worked with (and struggled along with) students of such high calibre.

I would also like to thank David Andrews and Paul Corey for sparking my interest in statistics with their generous enthusiasm. I thank David again for encouraging me to come to Stanford.

I have also been a member of the the Computation Research Group at SLAC for the past 3 years, and I would like to thank everyone there for making it an enjoyable and productive experience.

Thanks also to the Natural Sciences and Engineering Research Council of Canada, the Department of Energy, the Office of Naval Research and the Army Research Office for their financial support.

My appreciation is extended to Brad, Jerry and Paul Switzer for their thoughtful readings of this dissertation.

Lastly, thank you to Cheryl and Charlie for your unflagging love and devotion.

TABLE OF CONTENTS

Chapter 1

Introduction

Figure (1.1) contains 100 data pairs along with the least squares line summarizing the relationship of a response (say Y) and a covariate (X). In Figure (1.2) , the least squares line has been replaced by a "scatterplot smooth." This smooth was computed by a type of local averaging— around each X value a window of 20 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 reason is simple— the smooth doesn't make a rigid assumption about the form of the relationship between Y and X.

In recent years, there has been a great deal of interest in scatterplot smoothing by local averaging (see for example Cleveland(1979) and Friedman and Stuetzle(1981)) and the availability of fast computers has been essential in this development. These smooths are useful as a descriptive tool (as we have seen above) and also as building blocks for non-parametric regression models. Important developments in the latter area can be found in Friedman and Stuetzle (1981) and Brieman and Friedman(1982).

In this dissertation we explore an application of smoothing ideas to other kinds of data. In particular, we consider (X,Y) data whose relationship is expressible through a likelihood function. Take for example the situation in which Y is a 0-1 response and X is a covariate. For such a data set, Figure (1.3) shows the logistic regression line, estimated by maximum likelihood. On the same plot, the observed logits are shown. (Since we can't take the logit of 0 or 1, the Y's were grouped first). In Figure (1.4) , the line has been replaced by a smooth. 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. In Figures (1.5) and (1.6) , we see another example. Here our data is survival data and hence Y is a (possibly censored)

lifetime. Figure (1.5) shows the estimated log relative risk line given by Cox's proportional hazards model. In Figure (1.6) , the line has been replaced by a "log relative risk smooth".

The smooths in Figures (1.4) and (1.6) were obtained from a procedure we call **"local likelihood"** estimation. The basic idea is simple extension of the local averaging technique used in scatterplot smoothing. Given a global method for estimating a linear response (e.g. maximum likelihood estimation in the linear logistic model), we apply it locally, estimating a separate line in a window around each x value. The value of the estimated line at x is the estimate of the smooth response function at x.

By varying the window size, we can control the smoothness of the estimated function. The larger the windows, the smoother the estimated function. When each window contains 100% of the data, the local likelihood procedure corresponds exactly to the global linear method. Hence local likelihood generalizes linear likelihood estimation.

This dissertation is devoted to the study of local likelihood. We describe the method in general, showing how smooths like those in Figures (1.4) and (1.6) are obtained, and we will study some of its theoretical properties. In the exponential family, the local likelihood method extends the class of *generalized linear models* (Nelder and Wedderburn (1972)) by allowing covariates to enter the link function in a non-linear fashion. We investigate the linear logistic model, a member of this class, and its extension. We also explore in depth the application of the method to the proportional hazards model. This model was the motivating example behind local likelihood.

The chapters are organized as follows. Chapter 2 defines the local likelihood method and discusses the estimation procedure. Both the exponential and non-exponential family set-ups are described; included is a short discussion of the application to the Cox model. We also discuss a forward stepwise algorithm for building multiple covariate models. Chapter 3 describes in detail the application of the local likelihood procedure to Cox's proportional hazards model. We discuss a number of topics: bootstrapping the models, robustifying the fit, and assessing goodness of fit. We also present a number of simulations designed to study the bias properties of the procedure, and finally, some real data examples are given. Chapter 4 contains a short description of the application of local likelihood to the logistic

regression model for binary data. The discussion is brief, referring the reader to Hastie and Tibshirani (1984) for further details,

Chapter 5 provides some asymptotic results for local likelihood estimates in the exponential family. Consistency and efficiency of the estimates are discussed. We conjecture (without proof) similar results for the proportional hazards model.

In the last chapter (6), we address two important questions: 1) how many parameters are used up by a local likelihood smooth? and 2) is it reasonable to use Akaike's Information Criterion to choose the window size? We give approximate answers to these questions, backing up our claims with a simulation study.

Figure (1.1)

Least Squares Line

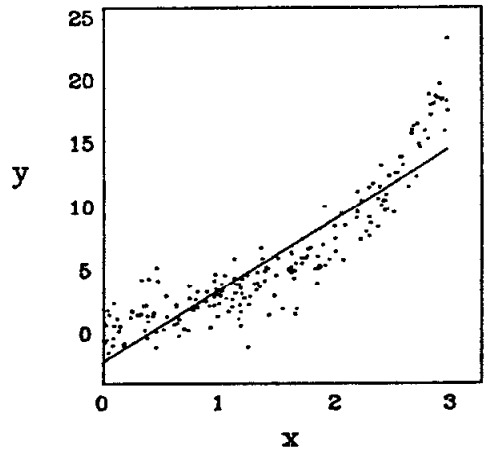


Figure (1.2)

Scatterplot Smooth

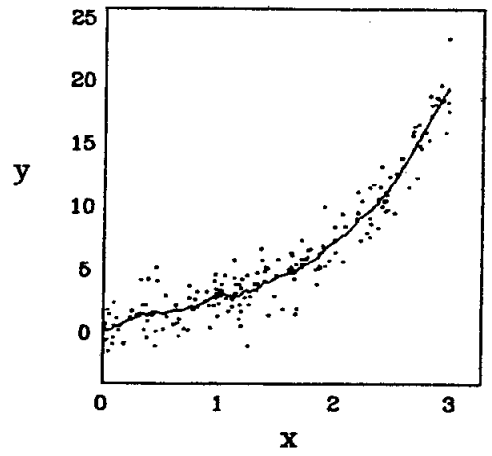


Figure (1.3)

Logistic Line

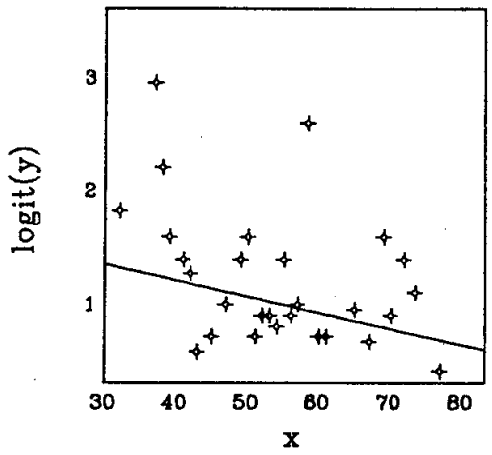


Figure (1.4)

Local Likelihood Logistic Smooth

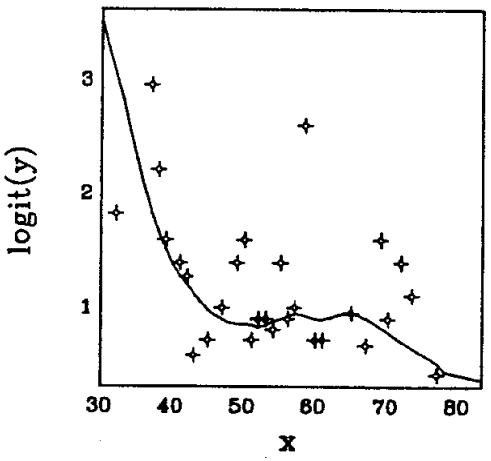


Figure (1.5)

Relative Risk Line

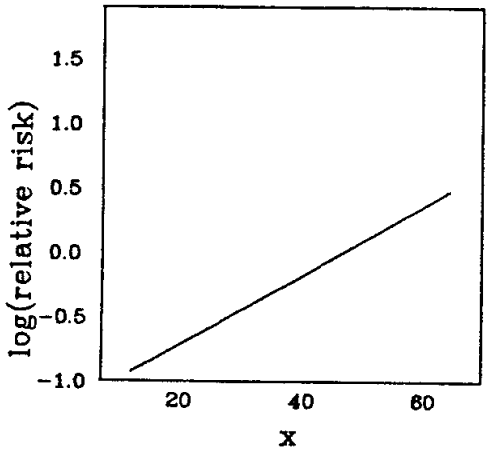
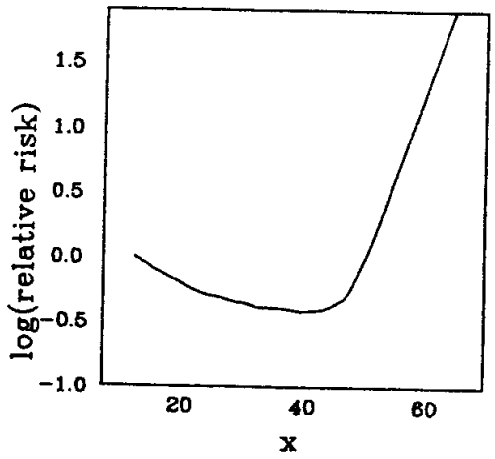


Figure (1.6)

Local Likelihood Relative Risk Smooth



Chapter 2

Local Likelihood— A description

2.1. Introduction.

In this chapter we introduce the local likelihood idea. Since local likelihood estimation is a generalization of scatterplot smoothing, we begin with a review of the latter.

2.2. A Review of Scatterplot Smoothing.

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

$$y_i = s(x_i) + \epsilon_i \tag{2.1}$$

Here $s(\cdot)$ is a "smooth" function and ϵ_i is a residual error. We won't define exactly what "smooth" means here; vaguely speaking, we're thinking of $s(\cdot)$ as a function less smooth than a straight line but smoother than an interpolating polynomial.

There are many ways to estimate $s(\cdot)$— we'll concentrate here on 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 $s(\cdot)$ would be to minimize $E(Y - s(X))^2$, where the expectation is taken over this joint distribution. Conditioning on $X = x$, this has solution $\hat{s}(x) = E(Y \mid X = x)$ for each x. In practice, we don't know this joint distribution but have only a sample from it. The idea, then, is to estimate $E(Y \mid X = x)$ from the data. This leads to the class of *local average* estimates for $s(\cdot)$:

$$\hat{s}(x_i) = \mathbf{Ave}_{j \in N_i} y_j \tag{2.2}$$

where "**Ave**" represents some averaging operator like mean or median, and N_i is a "neighborhood" of x_i (a set of indices of points whose x values are "close" to x_i). The only type of neighborhoods we'll consider in this dissertation are *symmetric nearest neighborhoods*. Assuming that the data points are sorted by increasing x value, these are defined by:

$$N_i = \{max(i - \frac{k-1}{2}, 1), ..., i-1, i, i+1, ...min(i + \frac{k-1}{2}, n)\} \tag{2.3}$$

The parameter k is called the *span* of the smoother and controls the smoothness of the resulting estimate. The value of k must be chosen in some way from the data.

If **Ave** stands for arithmetic mean, then $\hat{s}(\cdot)$ is the *running mean*, the simplest possible scatterplot smoother. The running mean is not a satisfactory smoother because it creates large biasses 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

$$\hat{s}(x_i) = \hat{\beta}_{0i} + \hat{\beta}_{1i} x_i \tag{2.4}$$

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

$$\begin{aligned} \hat{\beta}_{1i} &= \frac{\sum_{j \in N_i}(x_j - \bar{x}_i)y_j}{\sum_{j \in N_i}(x_j - \bar{x}_i)^2} \\ \hat{\beta}_{0i} &= \bar{y}_i - \hat{\beta}_{1i}\bar{x}_i \end{aligned} \tag{2.5}$$

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 it 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 a $O(1)$ operation is needed. This reduces the overall algorithm from $O(n^2)$ to $O(n)$.

Interpolation can be used to provide an estimate of $s(\cdot)$ at X values not occurring in the sample.

2.3. Local Gaussian Smoothing.

Since least squares estimation corresponds to maximum likelihood when the data are Gaussian, it is not surprising that the running lines smoother can be described as a "running maximum likelihood" method for Gaussian data. Assume as before that

$$y_i = s(x_i) + \epsilon_i \tag{2.6}$$

and in addition that the $\epsilon_i \sim$ i.i.d $\mathcal{N}(0, \sigma^2)$. Then for x in a neighborhood N_i of x_i, a reasonable approximation to $s(x)$ is

$$s(x) \approx \beta_{0i} + \beta_{1i} x \tag{2.7}$$

Considering only the points in N_i, the maximum likelihood estimates of β_{0i} and β_{1i} are given by (2.5) . Based on (2.7) , this gives as an estimate of $s(x_i)$:

$$\hat{s}(x_i) = \hat{\beta}_{0i} + \hat{\beta}_{1i} x_i \tag{2.8}$$

Hence running lines smoothing corresponds to finding approximate maximum likelihood estimates in a neighborhood around each data point.

We call this type of estimation *"LOCAL LIKELIHOOD ESTIMATION"* or *"LOCAL LIKELIHOOD "* for short. In this dissertation, we extend the idea of local likelihood to non-Gaussian likelihoods. It can be applied in principal to any situation in which the effect of a covariate is modelled through a likelihood. In fact, as will see in the proportional hazards model, the "likelihood" doesn't even have to be a likelihood in the strict sense.

2.4. Local Likelihood: General Definition.

Suppose we have n data tuples of the form (y_i, x_i, e_i), where y is a response variable, x is a covariate or predictor variable, and e is a vector containing any additional information. (In censored data problems, e would indicate whether y is censored. In many problems (like regression), e is empty.) Suppose that modelling considerations lead to maximization of a function of the form

$$L(\beta_0, \beta_1) = g(y_1, y_2, \ldots y_n, \theta_1, \theta_2, \ldots \theta_n, e_1, e_2, \ldots e_n) \tag{2.9}$$

where $\theta_i = \beta_0 + \beta_1 x_i$. For example, $L(\beta_0, \beta_1)$ could be a likelihood function and the estimates maximizing $L(\beta_0, \beta_1)$ would be the maximum likelihood estimates. The *LOCAL LIKELIHOOD* method replaces $\beta_0 + \beta_1 x_i$ with an arbitrary smooth function $s(x_i)$:

$$L(s(x_1), s(x_2), \ldots s(x_n)) = g(y_1, y_2, \ldots, y_n, \theta_1, \theta_2, \ldots \theta_n, e_1, e_2, \ldots e_n) \tag{2.10}$$

with $\theta_i = s(x_i)$. The problem is to estimate $s(\cdot)$ at the points $\{x_1, x_2, \ldots x_n\}$. Maximization of $L(s(x_1), s(x_2), \ldots s(x_n))$ results in an unsatisfactory estimate due to overfitting. In many situations, it simply reproduces the data. As an alternative, we define the *local likelihood* estimate of $s(x_i)$ as

$$\hat{s}(x_i) = \hat{\beta}_{0i} + \hat{\beta}_{1i} x_i \tag{2.11}$$

where $\hat{\beta}_{0i}$ and $\hat{\beta}_{1i}$ maximize the local likelihood:

$$L_i(\beta_{0i}, \beta_{1i}) = g(\{y_j, \beta_{0i} + \beta_{1i} x_j, e_j\}, j \in N_i) \tag{2.12}$$

The local likelihood procedure produces a smooth estimate of the curve $s(\cdot)$ at the points $\{x_1, x_2, \ldots x_n\}$. It avoids overfitting by averaging over neighborhoods. The width of the neighborhoods (the span) controls the smoothness of the resulting estimate— larger spans will tend to produce smoother curves.

The function $L(\beta_0, \beta_1)$ need not be a likelihood, (in Cox's model it is a "partial likelihood"), but in any case, we call this procedure "Local Likelihood" estimation.

2.5. Local Likelihood— Definition in the i.i.d. Case.

In the i.i.d case, we observe n independent data pairs $\{(x_1, y_1), \ldots (x_n, y_n)\}$ and we assume that given $X = x$, Y has density

$$Y \mid x \sim f(Y, \theta) \tag{2.13}$$

where $\theta = s(x)$. The likelihood is given by:

$$L(s(x_1), s(x_2), \ldots s(x_n)) = \prod_1^n f(y_j, \theta_j) \tag{2.14}$$

where $\theta_j = s(x_j)$.

The local likelihood estimate of $s(x_i)$ is

$$\hat{s}(x_i) = \hat{\beta}_{0i} + \hat{\beta}_{1i}x_i \tag{2.15}$$

where $\hat{\beta}_{0i}$ and $\hat{\beta}_{1i}$ maximize the local likelihood:

$$L_i = \prod_{j \in N_i} f(y_j, \beta_{0i} + \beta_{1i}x_j) \tag{2.16}$$

2.6. Asymptotic Properties of Local Likelihood Estimates.

Other than the fact that it produces smooth estimates, why is the local likelihood procedure reasonable? On a heuristic level, it's easy to see that for well behaved $s(\cdot)$ functions, $\hat{s}(\cdot)$ will be consistent for $s(\cdot)$. Consider a fixed point x_i. As $n \to \infty$ and the neighborhoods shrink in such a way that k_n, the span for sample size n, goes to infinity, while the width of the neighborhood goes to zero, we have $\hat{\beta}_{0i} \to \beta_{0i}$, $\hat{\beta}_{1i} \to \beta_{1i}$, ($\beta_{0i}$ and β_{1i} being the true slope and intercept) and the error in approximation (2.7) goes to zero. Hence $\hat{s}(x_i)$ will converge to $s(x_i)$. In addition to consistency, local likelihood estimates enjoy (in a weak sense) the optimality properties of maximum likelihood estimates. They are asymptotically normal and first order efficient with respect to sample size k_n. These properties are established in Chapter 5.

2.7. The Bias—Variance Tradeoff.

The span parameter controls the smoothness of the estimated function. Larger spans will tend to produce smoother, less variable estimates, but these estimates will tend to be biassed if the underlying function is non-linear. Conversely, smaller spans will produce less biassed but more variable estimates. A data-based criterion is therefore needed to select the span that best trades off these two factors for a given data set. We describe such a criterion in Section 2.13.

2.8. Computation of Local Likelihood Estimates.

To find each $\hat{\beta}_i = (\beta_{0i}, \beta_{1i})$ we use a Newton-Raphson search. Let $U_i(\beta^0)$ be the 2 by 1 score vector with jth entry

$$U_i(\beta^0) = \left(\frac{\partial log L_i}{\partial \beta_{ji}}\right)_{\beta=\beta^0} \tag{2.17}$$

and $I_i(\beta_0)$ be the 2 by 2 observed information matrix with jkth entry

$$I_i(\beta^0) = -\left(\frac{\partial^2 log L_i}{\partial \beta_{ji}\beta_{ki}}\right)_{\beta=\beta^0} \tag{2.18}$$

for the ith local likelihood both evaluated at some point β^0. Then given an initial guess $\hat{\beta}_i^{init}$, the Newton-Raphson method produces the new trial value:

$$\hat{\beta}_i^{new} = \hat{\beta}_i^{init} + I(\hat{\beta}_i^{init})^{-1}U(\hat{\beta}_i^{init}) \tag{2.19}$$

This procedure is iterated until convergence. It is used to find $\hat{\beta}_i$ (and hence $\hat{s}(x_i)$) for each neighborhood, going in order as i runs from 1 to n. The local likelihood estimate $\hat{\beta}_i$ is used as a starting value for the maximation of L_{i+1}; because the estimates don't tend to differ much from one neighborhood to the next, convergence is typically achieved in 1 or 2 iterations.

2.9. Exponential Family Case.

A special case of the above occurs when f is a member of the exponential family. Then the log likelihood has the form

$$\log L = \sum_{1}^{n} \{\{y_j\theta_j - b(\theta_j) - c(y_j, \sigma)\}/\sigma^2\} \tag{2.20}$$

where $\theta_j = s(x_j)$ and σ is a scale parameter. If σ is unknown, (2.20) is not generally an exponential family but the estimation procedure we will describe is unchanged because the score function for θ doesn't involve σ.

The local log likelihood is:

$$\log L_i = \sum_{j \in N_i} \{\{y_j\theta_{ij} - b(\theta_{ij}) - c(y_j, \sigma)\}/\sigma^2\} \tag{2.21}$$

where $\theta_{ij} = \beta_{0i} + \beta_{1i}x_j$. Letting X represent the n by 2 design matrix with first column $(1, 1, ...1)^t$ and second column $(x_1, x_2, ...x_n)^t$, and letting $W = diag\{I(j \in N_i)\}$, the local score function has the simple form

$$U_i(\beta_i) = X^tW(y - b'(X\beta)) \tag{2.22}$$

The observed information is $I(\beta_i) = X^tWb''(X\beta_i)X$ and the Newton-Raphson step is:

$$\hat{\beta}_i^{new} = \hat{\beta}^{init} + I^{-1}(\hat{\beta}^{init})X^tW(y - b'(X\hat{\beta}^{init})) \tag{2.23}.$$

In the above, we have modelled the natural parameter θ. We could just as well model some other parameter (like $E(y)$); in any specific problem, there may be reasons to prefer one parametrization to another. For example, in the binary response problem, it is more convenient to model the natural parameter $\log \frac{p}{1-p}$ then the expectation p because the latter would require that the estimated smooth stay between 0 and 1.

2.10. Relationship to Generalized Linear Models.

Model (2.20) can be viewed as a extension of the class of *generalized linear models* (Nelder and Wedderburn (1972)). A generalized linear model is defined by $Y \mid x \sim f(Y, \theta)$ and $E(Y) = g(\beta_0 + \beta_1 x)$, where f has the exponential form (2.20) . If g (the "link function") is invertible, this corresponds to $g^{-1}(E(Y)) = \beta_0 + \beta_1 x$. In the local likelihood set-up, we have generalized $\beta_0 + \beta_1 x$ to $s(x)$.

2.11. Number of Parameters— "Degrees of Freedom".

In Chapter 6, we discuss an approximate method for determining how many independent parameters a local likelihood smooth is really fitting. Since the local likelihood estimate produces a function smoother than the data, we would expect that it uses less than n independent parameters. This is the case. Consider a scatterplot smoother with span s. Such a smoother is linear in that the fit $\hat{y}$ can be written as $P(s)y$ where $P(s)$ is a *smoother matrix*. $P(s)$ will depend on the set of x values observed, as well as the span. In traditional linear least squares estimation, $P(s)$ is the hat matrix $X(X^tX)^{-1}X^t$. We

show in Chapter 6 that for a scatterplot smoother with span s, the number of degrees of freedom used up is $trace(P(s))$. (This result and related results are also given in Cleveland (1979)). We also show that for *any* local likelihood fit (in the exponential family), with span s, the number of degrees of freedom is about $trace(P(s))$. Thus, although the matrix $P(s)$ is only used in the estimation process of the Gaussian local likelihood model, (and not in the estimation of other local likelihood models), the *trace* of this matrix turns out to be the relevant quantity nonetheless. Note that this generalizes the result in linear estimation, in which $P(s)$ is an idempotent projection matrix and hence $trace(P(s)) = rank(P(s)) = p$, the rank of the column space of X.

The quantity $trace(P(s))$ turns out to be significantly less than n. In an example given in Chapter 6 with 100 data points and $s = .5$, $trace(P(s))$ is 3.65. Thus we are really fitting only 3.65 "parameters".

2.12. Application to Censored Data and the Cox Model.

In the censored data problem we observe data triples $(y_i, x_i, \delta_i), i = 1, 2, ...n$ where δ_i indicates whether or not the response y_i is censored. The proportional hazards model of Cox(1972) models the relationship between y and x by assuming that x acts on the hazard function in a multiplicative way:

$$\lambda(y \mid x) = \lambda_0(y)e^{\beta x} \tag{2.24}$$

where $\lambda_0(y)$ is an unspecified function. This assumption allows β to be estimated independently of $\lambda_0(y)$ by maximizing the *partial likelihood*:

$$PL = \prod_{i \in D} \frac{e^{\beta x_i}}{\sum_{j \in R_i} e^{\beta x_j}} \tag{2.25}$$

where D is the set of indices of the uncensored y's and R_i is the risk set prior to y_i. The local likelihood generalization of (2.24) is

$$\lambda(y \mid x) = \lambda_0(y)e^{s(x)} \tag{2.26}$$

and the local likelihood estimate of $s(x_i)$ is given by $s(x_i) = \hat{\beta}_{0i} + \hat{\beta}_{1i}x_i$ where $\hat{\beta}_{1i}$ maximizes

the *Local Partial Likelihood*

$$PL_i = \prod_{l \in D \cap N_i} \frac{e^{\beta_{1i} x_l}}{\sum_{j \in R_l \cap N_i} e^{\beta_{1i} x_j}} \tag{2.27}$$

The estimation of β_{0i} is a little trickier— we'll be discussing this in detail in Chapter 3.

As mentioned earlier, the Cox model was the motivating example behind the local likelihood idea. It is a good example of a situation in which the response enters into the model in an implicit way. Because of this, it would be difficult to smooth the responses explicitly; the local likelihood technique produces smooth estimates while staying within the likelihood framework.

2.13. Span Selection.

The estimation of a local likelihood smooth requires the choice of a span size. In scatterplot smoothing, one popular method for choosing the span size is *cross-validation*. For a number a trial spans, smooths are estimated leaving out each data point one by one. A cross-validation sum of squares is calculated and the span having the smallest value is selected. This is detailed in Friedman and Stuetzle (1981).

In the local likelihood problem, cross-validation turns out to be very computationally expensive. As an alternative, we explore in this dissertion the use of *Akaike's Information Criterion* (*AIC*). In fitting a generalized linear model with maximized likelihood L and p independent parameters, the *AIC* is defined by:

$$AIC = -2 \log L + 2p \tag{2.28}$$

The first term measures the goodness of fit of the model, while the second term penalizes the number of parameters used. Hence the *AIC* attempts to tradeoff variability and bias.

We make use of the *AIC* criterion for selecting the span of the local likelihood smooth. Using $trace(P(s))$ as an approximate number of degrees of freedom, the span size s is selected to minimize *AIC* based on the value of the global likelihood (2.10) . In a number of examples, we'll see that this procedure chooses reasonable span sizes, producing estimates that aren't too jagged nor too biassed. We justify this use of *AIC* is Chapter 6.

2.14. Weight Functions.

The local likelihood procedure implicitly attachs weights to the observations: a weight of 1 for the observations in the current window and 0 otherwise. The weight function or kernel is therefore rectangular. In principle, one could use a more smoothly descending weight function— this would require specification of a weighted likelihood. While we don't discuss this problem in general, we consider it in the context of the proportional hazards model (Chapter 3).

2.15. Multiple Covariates and Backfitting.

The above discussion shows how the local likelihood idea can be used to estimate the smooth for a single covariate. If more than one covariate is available, the model takes the form $\theta = \sum_{j=1}^{p} s_j(\cdot)$. To estimate the $s_j(\cdot)$'s, a forward stepwise algorithm is used, analogous to a forward stepwise regression algorithm. The algorithm proceeds by smoothing on each variable, and selecting the smooth that most improves the fit. When one smooth is selected, the remaining variables are smoothed and the one that most improves the fit is chosen. The process is repeated until no new variable can significantly improve the fit.

Now suppose that this procedure selects a smooth $\hat{s}_1(\cdot)$ at the first step and a smooth $\hat{s}_2(\cdot)$ at the second step. Then the smooth $\hat{s}_1(\cdot)$ may not be "optimal" given that $\hat{s}_2(\cdot)$ is in the model. Hence it is desirable to re-estimate $\hat{s}_1(\cdot)$ to accomodate $\hat{s}_2(\cdot)$. Now given the adjusted estimate $\hat{s}_1^*(\cdot)$, we can adjust $\hat{s}_2(\cdot)$ and so on, iterating until convergence. This process is called "backfitting", (Friedman and Stuetzle(1982)). In general, (with more than 2 smooths), whenever a new smooth is entered into the model, the smooths already in the model are adjusted to accommodate the new smooth. Specifically, all but one of the smooths are held constant and the remaining smooth is re-estimated. This is done for each smooth in turn until the fit no longer improves by a significant amount. As an example, suppose a new smooth $\hat{s}_{r+1}(\cdot)$ is added to a model containing smooths $\hat{s}_1(\cdot), \ldots \hat{s}_r(\cdot)$. Then the backfitting procedure would consist of estimating $s_j(\cdot)$ in the model

$$\theta = \sum_{k \neq j} \hat{s}_k(\cdot) + s_j(\cdot) \tag{2.29}$$

treating $\sum_{k \neq j} \hat{s}_k(\cdot)$ as a constant. This is done for j running from 1 to $r+1$.

We have no proof of convergence for the backfitting algorithm, although it has converged in all the examples that we've tried. In a simple linear regression framework, with p (possibly non-orthogonal) covariates $x_1, x_2, \ldots x_p$, one can show that backfitting converges to the correct answer (Stuetzle (1983), personal communication). That is, if we project the current residual vector onto each covariate in turn, the residual vector converges to the correct residual vector i.e. the response minus the projection of the response onto the column space of $x_1, x_2, \ldots x_p$.

2.16. How do we select covariates for the model?.

This question can be addressed through examination of $-2 \log L(\hat{\mu})$, but it is customary in generalized linear modelling to work with an equivalent measure, the "deviance". The deviance is $2 \log(L(y)/L(\hat{\mu}))$ which equals to $-2 \log L(\hat{\mu}) + \textit{constant}$. At each stage, then, we find the smooth that decreases the deviance the most. This smooth is then added to the model if the decrease in the deviance is large compared to the number of "parameters" used up by the smooth.

2.17. The Scale Parameter in the Exponential Family case.

The exponential form (2.20) may or may not contain an unknown scale parameter, but in any case, the likelihood estimation procedure is unchanged because the score doesn't involve the scale. An estimate of scale is needed, however, if the deviance is to be used to assess importance of model terms. As is true for standard generalized linear models, we would fit some maximal model and use the mean deviance as our estimate of scale, This could be used to form a "scaled deviance", proceeding thereafter as if the scale were known.

In the only exponential family model we discuss (the logistic model) the scale is a function of the mean, so this issue doesn't arise. Hence we will not go into scale estimation in this dissertation.

2.18. Generalizations of Local Likelihood.

The local likelihood models described here can be generalized to multiparameter models. Each parameter would be modelled in the form $s(x)$ and separate smooths would be estimates for each.

The additive model could also be generalized by a model of the form $\theta_i = f(\sum_{j=1}^{p} s_j(x_{ij}))$ and estimated by expanding f in a one term Taylor series. This is the idea used in the Predictive ACE procedure (Friedman and Owen (1984)).

This generalizations will not be pursued in this dissertation but in subsequent research.

Chapter 3

Application to the Cox Model

In this chapter we describe the application of the local likelihood technique to Cox's proportional hazards model for survival data. We begin with a general description of the problem.

3.1. Introduction.

In the past twelve years a number of methods have been suggested for the analysis of regression data in which the response variable is subject to right censoring. The most common application is in the study of survival in clinical trials. Patients in such trials often survive to the end of the study period or are lost to followup— their survival time is said to be "censored". Formally, we cannot observe their survival time T, but instead we observe only $Y = \min(T, C)$, where C is the patient's censoring time. A set of measurements(covariates) $\boldsymbol{z}$ is available for each patient and the goal is to investigate the relationship between T and $\boldsymbol{z}$. Typically, a large proportion of the responses are censored, so standard regression techniques cannot be used.

D.R. Cox (1972) proposed an elegant solution to the problem, introducing what is now known as the proportional hazards model. This model assumes that the hazards of individuals with different covariates are related in a multiplicative way, that is

$$\lambda(t\,|\boldsymbol{z}) = \lambda_0(t)\;\theta(\boldsymbol{z},\boldsymbol{\beta}) \tag{3.1}$$

where $\lambda(t\,|\boldsymbol{z})$ is the hazard function at covariate level $\boldsymbol{z}$ defined by

$$\lambda(t\,|\boldsymbol{z}) = \lim_{\Delta t \to 0} \frac{P(t \le T < t + \Delta t\,|T \ge t, \boldsymbol{z})}{\Delta t}, \tag{3.2}$$

$\theta(\boldsymbol{z},\boldsymbol{\beta})$ is a parametric function and $\lambda_0(t)$ is an unspecified function. Since the ratio of $\lambda(t\,|\boldsymbol{z}_1)$ and $\lambda(t\,|\boldsymbol{z}_2)$ is the relative risk between the covariate levels $\boldsymbol{z}_1$ and $\boldsymbol{z}_2$, we see that assumption (3.1) is equivalent to an assumption that the relative risk between two covariate levels does not vary with time.

To ensure that $\lambda(t\,|\boldsymbol{z})$ remains non-negative, Cox suggested the parameterization $\theta(\boldsymbol{z},\boldsymbol{\beta}) = \exp(\boldsymbol{z}\cdot\boldsymbol{\beta})$ where $\boldsymbol{z}\cdot\boldsymbol{\beta}$ denotes inner product. This gives the most widely used proportional hazards model

$$\lambda(t\,|\boldsymbol{z}) = \lambda_0(t)\exp(\boldsymbol{z}\cdot\boldsymbol{\beta}) \tag{3.3}$$

The advantage of assumption (3.1) is that $\boldsymbol{\beta}$ can be estimated without specification of $\lambda_0(t)$ by maximizing the "partial likelihood"

$$PL = \prod_i \frac{\exp(\boldsymbol{z}_i\cdot\boldsymbol{\beta})}{\sum_{j\in R(t_i)}\exp(\boldsymbol{z}_j\cdot\boldsymbol{\beta})} \tag{3.4}$$

where $R(t_i)$ is the risk set at time $t_i - 0$. As the name implies, (3.4) is not a true likelihood, but work by Cox(1975), Efron(1977), and Oakes(1977) has shown that the estimate $\hat{\boldsymbol{\beta}}$ from (3.4) is consistent and nearly fully efficient. Estimation of $\boldsymbol{\beta}$ from the partial likelihood depends only on the ranks of the survival times; this non-parametric aspect along with the free form of $\lambda_0(t)$ are the main reasons for the model's popularity.

The proportional hazards model as defined above makes two important assumptions: proportionality of hazards, and the parametric form $\exp(\boldsymbol{z}\cdot\boldsymbol{\beta})$ for $\theta(\boldsymbol{z},\boldsymbol{\beta})$. Using the local likelihood technique described in Chapter 2, we will develop an algorithm for estimating a proportional hazards model which does not require the specification of a parametric form for $\theta(\boldsymbol{z},\boldsymbol{\beta})$. The hazard is modelled as

$$\lambda(t\,|\boldsymbol{z}) = \lambda_0(t)\;\exp\Big(\sum_{j=1}^{p} s_j(z_j)\Big) \tag{3.5}$$

where the $s_j(\cdot)$'s are general smooth functions that are estimated from the data.

We will first discuss in detail the estimate for one covariate; later on we'll describe a forward stepwise algorithm for the full model (3.5) .

3.2. Estimation of a Single Relative Risk Function.

Suppose n items are placed on test and give rise to (possibly censored) observation times $\{y_1, y_2, \ldots y_n\}$ with associated (fixed) covariates $\{x_1 \leq x_2 \ldots \leq x_n\}$. (The y_i's are in order of increasing x_i). Assume for now that the y_i's are distinct– the case of ties will be discussed later. Let D be the set of indices of the failures among the y_i's, let δ_i be 1 if item i fails and 0 otherwise. To facilitate construction of a partial likelihood, we will make the usual assumption of non-informative censoring (see Kalbfleisch and Prentice(1980)).

The model we assume for the hazard is

$$\lambda(t\,|x) = \lambda_0(t)\exp(s(x)) \tag{3.6}$$

where $s(x)$ is some smooth function of x. Clearly we have no information about $s(x)$ at x-values not occurring in the sample, so estimation of $s(x)$ involves estimation of the n parameters $\{s(x_1), s(x_2), \ldots s(x_n)\}$.

The partial likelihood of the data is

$$PL = \prod_{i\in D} \frac{\exp(s(x_i))}{\sum_{j\in R_i}\exp(s(x_j))} \tag{3.7}$$

where $R_i = \{j\,|y_j \geq y_i\}$, the risk set at time $y_i - 0$. Notice that the terms in the product are in order of increasing x_i. The partial likelihood is usually written with terms ordered by t_i (see Cox(1972)); the x order will make the notation simpler for our purposes.

To estimate $s(x_1), s(x_2), \ldots s(x_n)$, we apply the local likelihood technique introduced

in Chapter 2. As before, let N_i be a symmetric neighborhood around x_i:

$$N_i = \{max(i - \frac{k-1}{2}), 1), \ldots i-1, i, i+1, \ldots min(i + \frac{k-1}{2}, n)\} \tag{3.8}$$

For $x \in N_i$ we assume $s(x) \approx \alpha_i + x\beta_i$, and the local partial likelihood for the data in N_i is

$$PL_i = \prod_{l\in D\cap N_i} \frac{\exp(\alpha_i + x_l\beta_i)}{\sum_{j\in R_i\cap N_i}\exp(\alpha_i + x_j\beta_i)} \tag{3.9}$$

To estimate α_i and β_i, we maximize PL_i. Note, however, that α_i is not estimable from PL_i since the $exp(\alpha_i)$ terms cancel one another giving

$$PL_i = \prod_{l\in D\cap N_i} \frac{\exp(x_l\beta_i)}{\sum_{j\in R_i\cap N_i}\exp(x_j\beta_i)} \tag{3.10}$$

Let $\hat{\beta}_i$ maximize $L_i(\cdot)$. Although α_i (thus $s(x_i)$) is not estimable locally, we can use the slope estimates $\{\hat{\beta}_1, \ldots \hat{\beta}_2\}$ to estimate $\{s(x_1), \ldots s(x_n)\}$, as follows. We have $s(x_i) = \int_c^{x_i} s'(t)dt$ and $s'(x) = \beta_i$ for $x \in N_i$, hence to estimate $s(x_i)$ we can use any estimate of $\int_c^{x_i} s'(t)dt$ based on $(x_1, \hat{\beta}_1), \ldots (x_n, \hat{\beta}_n)$. Before discussing some particular integral estimators, it is important to note that the choice of c is arbritrary, reflecting the fact that $s(x)$ is only determined up to an additive constant. Substitution of $s(x) + c$ for $s(x)$ in (3.6) doesn't change the model because the factor e^c can be absorbed into the arbritrary function $\lambda_0(t)$. For simplicity, then, we define $c = x_1$, so that $s(x_1) = 0$.

To estimate $\int_{x_1}^{x_i} s'(t)dt$, we can use the simple rectangular rule defined by

$$\hat{s}(x_i) = \sum_1^i (x_j - x_{j-1}) * \hat{\beta}_i \tag{3.11}$$

for $i > 1$ and $\hat{s}(x_1) = 0$. This could also be written as $\hat{s}(x_i) = (x_i - x_{i-1}) * \hat{\beta}_i$, so that the rectangular rule constructs the estimate $\hat{s}(\cdot)$ by joining each line segment to the previous one, with prescribed slope $\hat{\beta}_i$

For greater accuracy, we instead use the trapezoidal rule defined by

$$\hat{s}(x_i) = \sum_1^i (x_j - x_{j-1}) * \frac{(\hat{\beta}_i + \hat{\beta}_{i-1})}{2} \tag{3.12}$$

for $i > 1$ and $\hat{s}(x_1) = 0$.

The procedure is summarized in the following algorithm:

Relative Risk Smoother

For i=1 to n

Find $\hat{\beta}_i$ that maximizes $PL_i(\cdot)$

End For

$\hat{s}(x_1) = 0$

For i=2 to n

$\hat{s}(x_i) = \sum_1^i (x_j - x_{j-1}) * \frac{(\hat{\beta}_i + \hat{\beta}_{i-1})}{2}$

End For

Output $\{\hat{s}(x_1), \hat{s}(x_2), \ldots \hat{s}(x_n)\}$

3.2.1. Selection of the Span

The span is chosen to minimize an approximate AIC ("Akaike's Information Criterion") given by

$$AIC = -2 \log PL + 2trace(P(s)) \tag{3.13}$$

PL is the value of the overall partial likelihood (expression (3.7))) and $P(s)$ is the local linear smoothing matrix of span (s), based on the observed x values. The first term measures fit of the model, and the second term penalizes model complexity. AIC is minimized over spans .3,.4,.5,.6,and .7 of n. The use of AIC is discussed in Chapter 6.

3.2.2. Significance of a Smooth and "Degrees of Freedom"

In proportional hazard modelling, the "deviance" has no obvious analogue, so one works directly with $-2 \log PL$ to assess significance of a smooth. The "degrees of freedom" of the smooth are more difficult to obtain, however. The simulation study in section 5 shows

that the formula $trace(P)$ is biassed downward for the Cox model. Therefore we find the mean deviance decrease by the simulation technique described in that section. The *trace* formula is still adequate for span selection, however, since biasses will tend to cancel out in comparing two spans.

Before illustrating the technique with real data example, we first make a few remarks.

Remarks

- For data with tied t_i values, we use the approximation suggested by Peto(1972) and Breslow(1974) for the partial likelihood

$$PL \approx \prod_i \frac{\exp(z_i \cdot \beta)}{(\sum_{j \in R(t_i)} \exp(z_j \cdot \beta))^{d_i}} \tag{3.14}$$

  where d_i equals the number of failures at t_i and z_i equals the sum of z_i's for items failing at t_i. This approximation is used for each of the partial likelihoods $PL_i(\cdot)$.

- For data with tied x values, two things are done. First, each neighborhood is expanded (if necessary) to ensure that if a point j is in a given neighborhood, so is any other point k having $x_k = x_j$. This makes the estimation procedure invariant to the incoming order of the data points. Secondly, the smooths for each of the tied values are averaged and each smooth value is assigned the average. That is, if $x_j = x_{j+1} \ldots = x_{j+m}$, then for each $j \le i \le j+m$, $\hat{s}(x_i)$ is assigned the value $\sum_j^{j+m} \hat{s}(x_i)/(m+1)$.

- When the span size is expressed as a fraction f, the actual span used is the largest odd integer less than or equal to fn.

3.2.3. Example 1: The Stanford Heart Transplant Data

The first example that we will use for illustration of this technique is the Stanford Heart Transplant Data, as reported by Miller and Halpern(1983). There are 157 observations consisting of survival time after transplant and two covariates: age (in years) at time of

transplant and T5 mismatch score. Figure (3.1) shows a plot of survival time vs age, with squares representing failures and plusses representing censored times. The procedure chose a span size of .7 and produced the smooth shown in Figure (3.2) . The actual estimate of relative risk ($exp(\hat{s}(\cdot))$) is shown in Figure (3.3) . A summary of the results is shown in Table 3.1.

Table 3.1. Stanford Heart Transplant Data

Analysis of Age

Model	*−2 Log Likelihood*	*Number of Parameters*
Null	902.40	0
Age (linear)	894.82	1
Age + Age^2	886.24	2
Age (smooth, span .7)	884.65	2.95
Piecewise linear	885.40	2

The smooth reduced $-2\log PL$ from a null value of 902.40 to 884.65. For comparison, a standard proportional hazards model with a single term for age produced a value of 894.82 for $-2\log PL$ and the addition of a quadratic term for age reduced it to 886.24. The resulting quadratic function is shown in Figure (3.2) (broken line). The smooth in Figure (3.2) suggests that the relative risk before age 45 is approximately constant, while the quadratic curve, perhaps misleadingly, indicates a decrease in risk before age 45. We note that the smooth produces a smaller value of $-2\log PL$ (by 1.6) but uses .95 more "parameters".

Based on Figure (3.2) , we tried to summarize $\hat{s}(\cdot)$ by a piecewise linear covariate $z = -.2$ for $age < 44$ and $z = .12 \cdot age - 5.5$ for $age > 44$. Using z as a covariate in a model of the form $\lambda_0(t)\exp(\beta\hat{s}(x))$, a standard computer program for fitting proportional hazards models produced a value of 885.40 for $-2\log PL$. This provides further evidence that the quadratic shape for the relative risk may not be realistic.

3.3. A Forward Stepwise Algorithm.

In this section we describe a forward stepwise algorithm for the case of more than one covariate, using the relative risk smoother of Section 3.2.

The algorithm proceeds by smoothing on each variable, and selecting the smooth that most improves the fit. When one smooth is selected, the remaining variables are smoothed and the one that most improves the fit is chosen. The process is repeated until no new variable can significantly improve the fit.

The only "non-standard" aspect of the algorithm is the process of "backfitting", as used by Friedman and Stuetzle(1982). Whenever a new smooth is entered into the model, the smooths already in the model are adjusted to accommodate the new smooth. Specifically, all but one of the smooths are held constant and the remaining smooth is re-estimated. This is done for each smooth in turn until the fit no longer improves by a significant amount. As an example, suppose a new smooth $\hat{s}_{r+1}(x_{r+1})$ is added to a model containing smooths $\hat{s}_1(x_1),\ldots\hat{s}_r(x_r)$. Then the backfitting procedure would consist of estimating $s_j(x_j)$ in the model

$$\lambda(t\,|\,x) = \lambda_0(t)\exp\Big(\sum_{k\neq j}\hat{s}_k(x_k) + s_j(x_j)\Big) \tag{3.15}$$

treating $\sum_{k\neq j}\hat{s}_k(x_k)$ as a constant. This is done for j running from 1 to $r+1$.

An outline of this algorithm is:

Forward Stepwise Algorithm

While (not all variables have been selected)

Find the smooth that decreases $-2\log PL$ the most

If decrease < threshold1 exit

If current model contains more than one smooth

Backfit smooths until decrease in $-2\log PL$ < threshold2

End While

The output of the algorithm is $\{\hat{s}_{11},\ldots\hat{s}_{1n}\},\ldots\ldots\{\hat{s}_{h1},\ldots\hat{s}_{hn}\}$ where h is the number of smooths selected.

3.3.1. Stanford Heart Transplant Data: Age and T5

The forward stepwise algorithm was run on the Stanford Heart Transplant data described in Example 1. A plot of log survival time versus T5 mismatch score is shown in Figure (3.4) . The smooths for each variable separately are shown in Figures (3.2) and (3.5) . Threshold1 was set to zero to allow both variables to enter. Threshold2 was .01. The results are summarized in Table 3.2.

Table 3.2. Stanford Heart Transplant Data

Analysis of Age and T5

Model	*−2Log Likelihood*	*Number of Parameters*
Null	902.40	0
T5 (smooth, span= .7)	899.99	2.68
Age + T5	882.53	2.95 + 2.68
Age + T5 (backfit)	882.52	2.95 + 2.68

Age was entered first, then T5 mismatch score. The smooth for T5 is shown in Figure (3.6) . Backfitting had only a negligible effect, so the smooth for age was virtually identical to Figure (3.2) . The results indicate that the effect of T5, after adjusting for age, is very slight.

3.3.2. Example 2: Mouse Leukemia Data

Kalbfleisch and Prentice (1980) analyzed the results of a study designed to examine the genetic and viral factors that may influence the developement of spontaneous leukemia in AKR mice. The original data set contains 204 observations, with six covariates and 2

causes of death(cancerous and non-cancerous) measured. Kalbfleisch and Prentice perform a number of analyses; we will follow one of them here, using any death as the endpoint and the four covariates:

z_1: antibody level (%)

z_2: Gpd-1 phenotype

z_3: sex(1=male, 2=female)

z_4: coat colour

Antibody level took on continuous values, although about half of the mice had a value of 0. The other three covariates were binary. Of the 204 observations, 4 had missing values and were discarded.

Table 3.3 shows the results of forward stepwise local likelihood estimation applied to these data.

Table 3.3. Mouse Leukemia Data

Multivariate Analysis

Model	*−2Log Likelihood*	*Number of Parameters*
Null	1189.06	0
Antibody (smooth, span= .5)	1173.98	1.85
Antibody+Gpd-1	1170.90	1.85 + 1
Antibody (linear)	1183.16	1
Antibody (linear + quadratic)	1183.07	2
Piecewise linear	1177.34	2

Each of GPD-1, sex and coat color were modelled with a single parameter. Antibody was the most important factor, reducing $-2\log PL$ by 15.08. Gpd-1 was next in importance but not significant at 95%. A graph of the estimated smooth for antibody in shown in Figure (3.7) (the smooth values were not joined so that the distribution of antibody levels could

be seen). It is markedly non-linear, changing slope at antibody level =7.5%. Also included in Table 3.3 are linear and quadratic terms for antibody. Even with a quadratic term, the fit of the parametric Cox model is significantly worse than the local likelihood smooth.

Based on Figure (3.7), a piecewise linear covariate was created by joining each of the left and rightmost smooth values to the bending point by straight lines. $-2 \log PL$ for this covariate was 1177.34, still significantly worse than the smooth model. This indicates that the bowed shape of the smooth between antibody levels 7.5% and 80% is supported by the data.

3.4. Further Topics.

3.4.1. Computational Considerations

A Newton-Raphson search is used to find the slope estimate $\hat{\beta}_i$ for each neighborhood. This means that an $O(k_n)$ operation is required for each neighborhood, making the entire procedure $O(n^2)$ (assuming $k_n \sim n$). This is not a problem for moderate n (say $n \sim 200$) because the final estimate for the ith neighborhood is an excellent starting value for the $i+1$st neighborhood. Typically, convergence is obtained in 2-3 iterations. As an example, the smooth in Example 1 required .67 sec on an IBM 3081.

For larger data sets, we speed up the procedure by calculating the fit only every mth point; this reduces the running time by about a factor of m. The smooths for the remaining x-values are obtained by interpolation.

The scatterplot smoother of Friedman and Stuetzle (1981) uses updating formula to achieve an $O(n)$ algorithm. We have been unable to obtain such formulae for this problem because of the non-linear nature of the estimation.

3.4.2. Categorical Variables

Since it doesn't make sense to estimate a smooth for a covariate taking on unordered discrete values, such variables are treated in the standard way. If the covariate takes on J values, $J-1$ dummy variables are created and a slope parameter is estimated for each. Hence in analyzing a data set containing both continuous and categorical variables, a smooth is estimated for each continuous covariate and slope parameters are estimated for the categorical variables.

3.4.3. Examining Goodness of Fit

In fitting a standard (linear) proportional hazards model to a set of data, the goodness of fit of the model should be examined. The overall question is: does the model fit? If not, it might be because a) the covariate effects are non-linear, b) additional covariates are required, or c) the proportional hazards assumption is unrealistic. The local likelihood extension of the model solves (a). The algorithm automatically finds the best functional form for each covariate effect.

Problems (b) and (c) are difficult to answer for the local likelihood model, just as they are for the standard proportional hazards model. For assessing the appropiateness of the proportional hazards assumption, a number of approaches are available. If the covariates are discrete, an estimate of the log hazard can be plotted for each subgroup; these should be approximately parallel if the proportional hazards assumption holds. For continuous covariates (clearly of interest here), matters are more difficult. A number of methods have been suggested, but none, in our opinion, are very effective. Kay(1977) utilizes residuals of the form $\hat{\Lambda}_0(y_i)\exp(x_i\hat{\beta}_i)$, which have a censored exponential distribution (with mean 1) if the model is correct. The value of these residuals is extremely questionable, however. Crowley and Hu(1977) point out that if a null model ($\beta = 0$) is fit, (and there is no censoring) the residuals will have *exactly* an exponential(1) distribution, no matter what the true model is. In a later discussion, Crowley and Stormer (1983) confirm this by simulation but do

suggest that a plot of the residuals versus a new covariate may still be of use in assessing the importance of the covariate.

Another way to check the proportional hazards assumption is by partitioning the covariate and time space, and comparing the observed to expected number of failures in each cell. The expected number of failures can be computed using the estimated survivor function

$$\hat{S}(t\,|\,x) = \hat{S}_0(t)^{\exp(\hat{s}(x))}$$

where

$$\hat{S}_0(t) = \exp(-\hat{\Lambda}_0(t))$$

and

$$\hat{\Lambda}_0(t) = \sum_{i\,|\,y_i<t} \frac{\delta_i}{\sum_{j\in R_i}\exp(\hat{s}(x_j))} \tag{3.16}$$

This is by direct analogy to the estimator for the standard proportional hazards model (see Kalbfleisch and Prentice (1980), pg 116). Schoenfield(1982) suggests a more complicated version of this procedure, and he provides a chi-square type statistic for testing goodness of fit.

For multiple covariates, this type of goodness of fit procedure would be ineffective because many of the cells would be empty or near empty. A more promising idea would be to insert a term like $x \log t$ and check if the fit is substantially improved. This was suggested by Cox(1972). Unfortunately, time-dependent covariates haven't been implemented in our procedure because of the additional computational cost (see section 3.4.8.)

As a final comment, it is important to mention the paradoxical nature of this problem. It is the non-parametric element of the proportional hazards model (the arbitrary baseline hazard) that makes the goodness of fit difficult to assess. In a sense, goodness of fit should not be as big a concern as it is in other regression models.

3.4.4. Bootstrapping the models

To assess the variability of an estimated relative risk curve, the bootstrap (Efron (1979)) can be applied. As in the regression modelling, there are (at least) two ways to bootstrap: we can resample the triples (y_i, x_i, δ_i) or the resample the residuals (r_i, δ_i) (where $r_i = \hat{\Lambda}_0(y_i)\exp(\hat{s}(x_i)))$ and add them back to the fitted model. As in the regression case, the second method assumes that the fitted model is correct.

The results for these two bootstrap methods applied to Example 1 are shown in Figures (3.8) and (3.9) . 20 bootstraps were computed for each method. In Figure (3.8) , the curves have considerable variability in the low and high age groups; in Figure (3.9) , there is less overall variability. The use of the bootstrap for the proportional hazard model requires further study; Efron(1980) looks at the bootstrap for the Kaplan-Meier curve.

3.4.5. Case Control Data and a Comparison to Thomas' Method

Thomas (1983) provides a method of finding the maximum likelihood estimate of $r(x)$ in the proportional hazards model $\lambda(t\,|\,x) = \lambda_0(t)r(x)$ subject to $\hat{r}(x)$ monotone in x. The algorithm is extremely complex and not fully understood by this author. It produces a step function $\hat{r}(\cdot)$ with steps occurring only at some of the failures.

Thomas applied his algorithm to a data set consisting of 215 lung cancer cases, each matched with 5 controls, sampled from a large cohort of Quebec chrysotile miners and millers (see Liddell, McDonald and Thomas). The covariate of interest was total dust exposure. The effect of various levels of dust exposure was desired so that industry standards could be established.

In order to handle case control data of this type, only a small change is required in the local likelihood procedure. The local partial likelihood simply becomes a partial likelihood for case-control data. This, in turn, is the same as the partial likelihood for prospective data, except that each risk set consists of a case and its associated controls (see Prentice and Breslow (1978) for details). It turns out that in the modified local likelihood procedure,

a case-control set only enters into the partial likelihood for a given neighborhood if the case and at least one control exist in the neighborhood.

Figure (3.10) shows the results of the various estimation procedures applied to the lung cancer data.* The solid line is the local likelihood smooth $exp(\hat{s}(\cdot))$, and the step function (dashed line) is Thomas monotone m.l.e. The functions are in qualitative agreement, with the monotone m.l.e suffering from its jagged shape.

The advantages of the local likelihood procedure over Thomas' method are clear. The monotone m.l.e is not smooth and is forced to be monotone. As well, Thomas' procedure can handle only one covariate. The local likelihood procedure suffers from none of these problems.

3.4.6. A Bias Study

In this section we discuss a number of simulations designed investigate how well the procedure estimates the true underlying function. In particular, we want to find out how much it underestimates curvature for larger spans, especially at the endpoints.

A sample of 200 X values were generated from $U(-1,1)$, and survival times T were generated from the model $\log T = 5 + 4x^2 + \epsilon$ where ϵ had the extreme value distribution $exp(\epsilon - exp(\epsilon))$. This corresponds to the hazard model $\lambda(t\,|\,x) = exp(-5 - 4x^2)$. Censoring times C were then generated from $U(0,11)$, and the observed response was $Y = \min(T, C)$. This resulted in an average censoring rate of 51 percent. Figure (3.11) shows one sample generated in this way, and and Figures (3.12) - (3.16) show the local likelihood estimated smooths for spans .3 to .7 along with the true function (broken line). Since the functions are determined only up to an additive constant, they were translated to have the same mean over the range of x. Our aim here was to found out how well the procedure reproduces

* Unfortunately, we could only obtain a slightly smaller data set from Thomas, consisting of 188 of the 215 case-control groups. The local likelihood procedure was applied to this reduced data set, while Thomas' procedure was applied to the full data set

curvature in the middle of the covariate range (so that endpoint effects don't enter in). We see that the estimates are quite jagged for smaller spans, fairly accurate for medium spans, but underestimate the curvature for span .7. gets to .7. Figure (3.17) shows the average of 20 replications (with the same set of x values) allowing the procedure to choose the span by the *AIC* criterion. The average smooth captures the shape of the true function remarkably well.

Next, we investigated the effect of endpoint bias. We generated data from the same model as above, except that X was $U(-1,.5)$ (We cut off the X range so that the true function would be non-linear near an endpoint.) The local likelihood smooths for spans .3 to .7 are shown in Figures (3.18) to (3.22) , along with the true underlying function (broken line). We see severe biasses for the smaller spans, with a span of .7 performing the best. Figure (3.23) shows the average of 20 replications, allowing the procedure to choose the span. The average smooth underestimates the curvature, but reproduces the function quite well.

We conclude from this modest study that the local likelihood procedure may have low bias, with a tendency to underestimate curvature slightly at the endpoints. A more ambitious study would investigate the effects of sample size, censoring rate and covariate distribution.

3.4.7. A Robust Fit

There are two types of influential points that can create problems in regression modelling: outliers in time space and outliers in covariate space. The first type are not as much of a problem here because the partial likelihood depends only on the ranks of the survival times. Still, Cain and Lange (1983) give an example in which a few large survival times have a large effect on the regression coefficient.

Outliers in covariate space are potentially more dangerous. Because of the local nature of the fitting, it will not be as much a problem in the local likelihood model as it is in the linear proportional hazards model, but with spans as large as .7n, it is still a concern.

A simple modification of the fitting procedure can help reduce the effect of covariate

outliers in both the standard and local likelihood proportional hazard models. The idea is to downweight observations based on their distance from the "center" of the data. This idea is exploited in the bounded influence regression literature (see Krasker and Welch (1973) and the references therein). In order to define a "weighted" partial likelihood estimate, we need to define the partial likelihood for a sample with weights w_i on (y_i, x_i, δ_i), $i = 1, 2, \ldots n$. $(\sum_1^n w_i = n)$. It is natural to require that when the w_i's are integers, the weighted partial likelihood should exactly coincide with the partial likelihood for a sample with w_i copies of point i. A form suggested by Cain and Lange almost satisfies this requirement:

$$PL^{\boldsymbol{w}} = \prod_{i \in D} \left(\frac{exp(\boldsymbol{x}_i \beta)}{\sum_{j \in R_i} w_i exp(\boldsymbol{x}_j \beta)} \right)^{w_i} \tag{3.17}$$

When the weights are integers, this reduces not to the exact partial likelihood for the corresponding sample, but to the standard approximation for tied data given in Section 3.2. As long as each w_i is small compared to $\sum_{j \in R_i} w_j$, (as it will be in our case) this approximation is adequate. When the original data contains ties, we can modify (3.17) :

$$PL^{\boldsymbol{w}} = \prod_{i \in D} \frac{exp(\sum_1^{d_i} \boldsymbol{x}_j w_j \beta)}{[\sum_{j \in R_i} w_i exp(\boldsymbol{x}_j \beta)]^{\sum_1^{d_i} w_l}} \tag{3.18}$$

where d_i is the number of failures at y_i. Expression (3.18) reduces the correct (approximate) partial likelihood when the weights are integers.

Maximation of (3.18) with appropriate weights provides a more robust fitting procedure. Let $\boldsymbol{x}^c$ be some "center" of the covariate space and let v_j be some scaled measure of distance of $\boldsymbol{x}_j$ from $\boldsymbol{x}^c$. Then a reasonable choice of weights is $w_j \sim e^{-v_j}$. For the linear proportional hazards model, it would be natural to choose $\boldsymbol{x}^c = \bar{\boldsymbol{x}}$ and

$$v_j = \boldsymbol{x}_j^t (X^t X)^{-1} \boldsymbol{x}_j \tag{3.19}$$

In the univariate case, this reduces to

$$v_j = \frac{(x_j - \bar{x})^2}{\sum_1^n (x_j - \bar{x})^2} + \frac{1}{n} \tag{3.20}$$

For the local likelihood extension of the model, we can use partial likelihood form

(3.18) in each neighborhood, and weights proportional e^{-v_j} where

$$v_j = \frac{(x_j - x_i)^2}{\sum_1^n (x_j - x_i)^2} + \frac{1}{k_n} \tag{3.21}$$

Note that x_i is used as the center of the neighborhood instead of the mean— this ensures that points near the ends receive large weights in their own neighborhoods.

Figure (3.24) shows the robust version of the local likelihood procedure applied to age variable (solid line). The smooth looks very similar to the unweighted (0-1 weights) smooth (broken line); this is not surprising since there are no outlying ages in the sample. Figure (3.25) shows the unweighted smooth (broken line) applied to the sample after having moved a failure at the highest age (62) to 92 (only the portion of the the smooth from ages 12 to 62 is shown). The weighted smooth (dotted line) looks much like the weighted smooth applied to the original data (solid line, same as solid line in Figure (3.24)). The downweighting has successfully reduced the effect of the outlying point on the overall smooth. Of course, the weighting scheme described here could be applied within the parametric setting, but we haven't pursued this.

The "robustifying" scheme discussed here is important if the local partial likelihood procedure is to be used in "auto-pilot" mode; alternatively, since each covariate is fit separately, a simple scatterplot of y versus each covariate should reveal any large outliers in covariate space.

In the theoretical investigations of the following chapters we'll restrict attention (for simplicity) to unweighted smoothing procedures.

3.4.8. Extending the Model

There are (at least) two ways that the model could be extended. The first way would be to allow time-dependent covariates. In principle, this would be straightforward; as in the standard proportional hazards model, one would simply insert the "current" covariate values when constructing each term of the partial likelihood. There may be computational

problems with this, however. With fixed covariates, the risk sets can be computed by "stripping off" each failure or censoring as they occur. With time-dependent covariates, however, the risk sets must be recomputed for each failure time. This would increase the cost by about a factor of n. We haven't tried implementing time-dependent covariates; this may be pursued in subsequent research.

Another way to generalize the model is to allow linear combinations of covariates to enter into the model. The form of the model would be

$$\lambda(t \mid \mathbf{z}) = \lambda_0(t) \exp(\sum(s(\boldsymbol{\alpha}_i \cdot \mathbf{z}))) \tag{3.22}$$

The vectors $\boldsymbol{\alpha}_i$ could be found by a numerical search. This is the "Projection Pursuit Regression" idea introduced by Friedman and Stuetzle(1981). Besides the obvious computational cost, this model would suffer from a lack of interpretability.

Figure (1)

Heart Translant Data— Age

Diamond: uncensored, Plus: censored

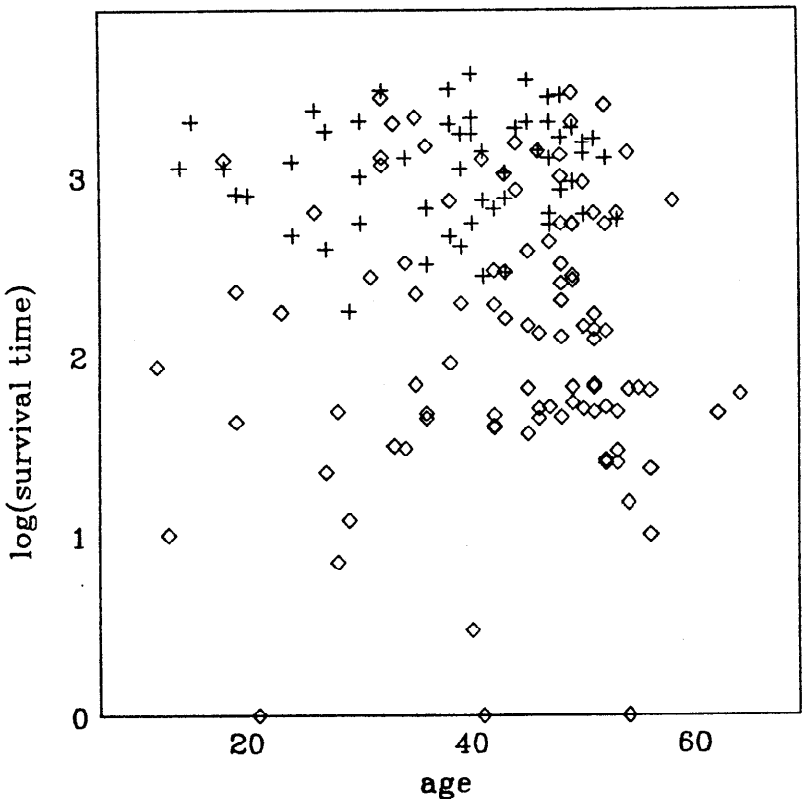


Figure (2)

Local Likelihood Estimate for Age

Solid line: L.L smooth, Broken line: quadratic fit

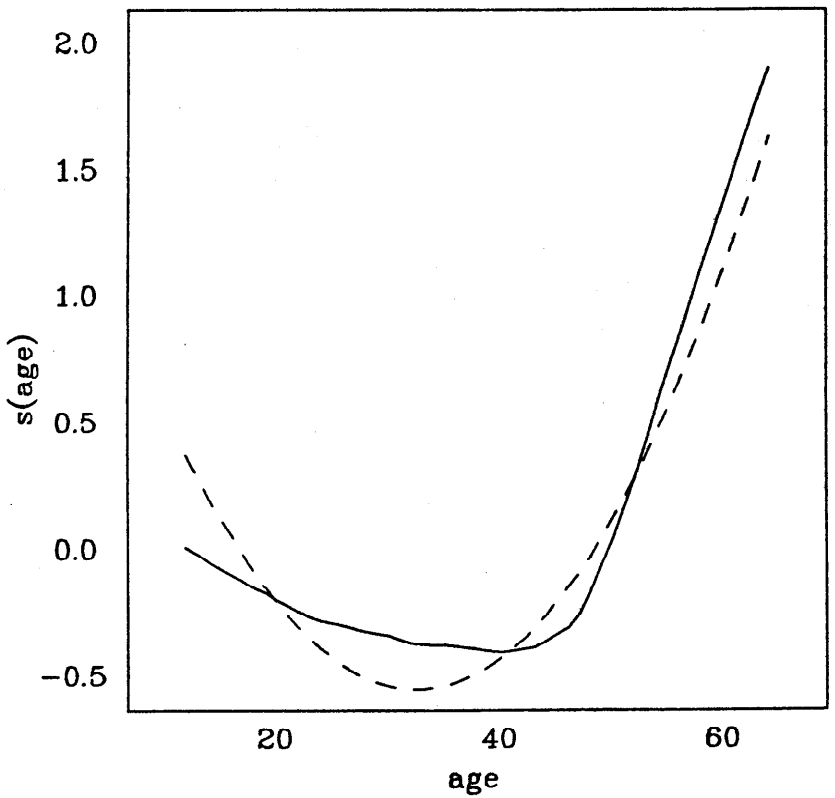


Figure (3)

Local Likelihood Estimate of Relative risk for Age

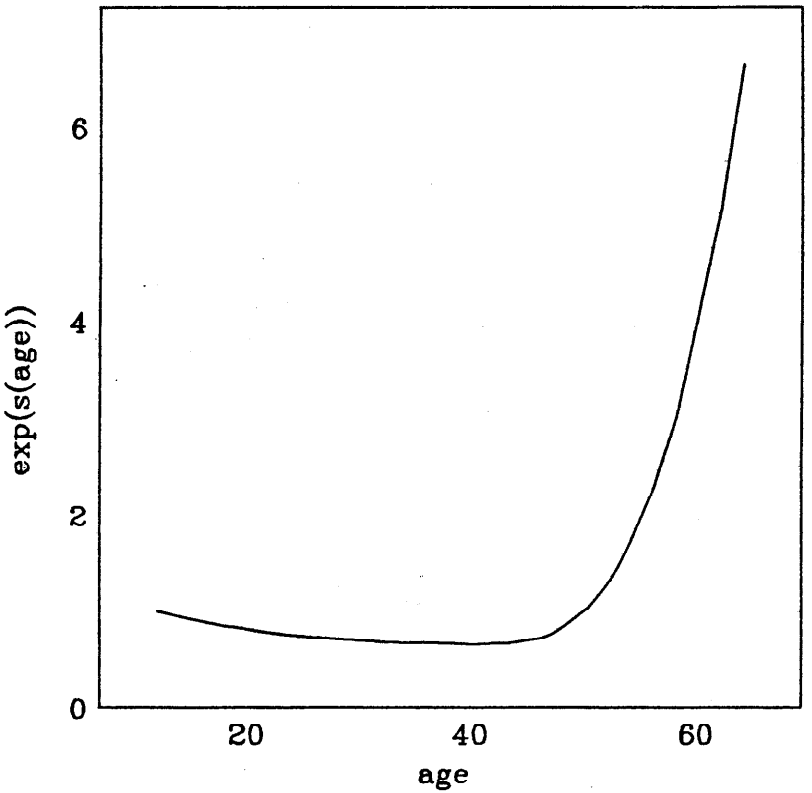


Figure (4)

T5 Mismatch Score

Diamond: uncensored, Plus: censored

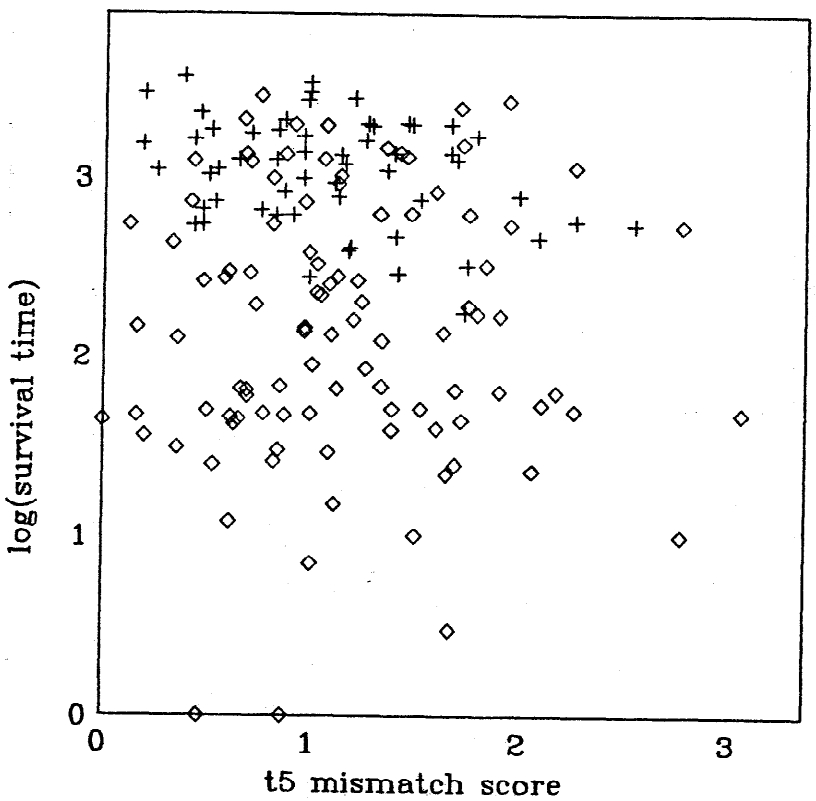


Figure (5)

Local Likelihood Smooth for T5 Mismatch Score

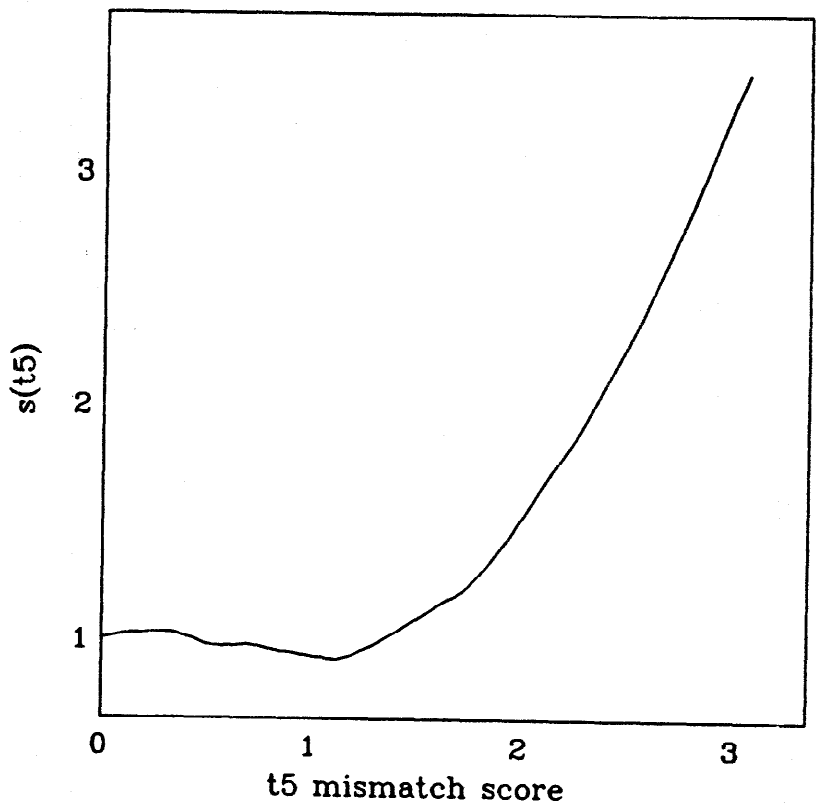


Figure (6)

T5 smooth with age in the model

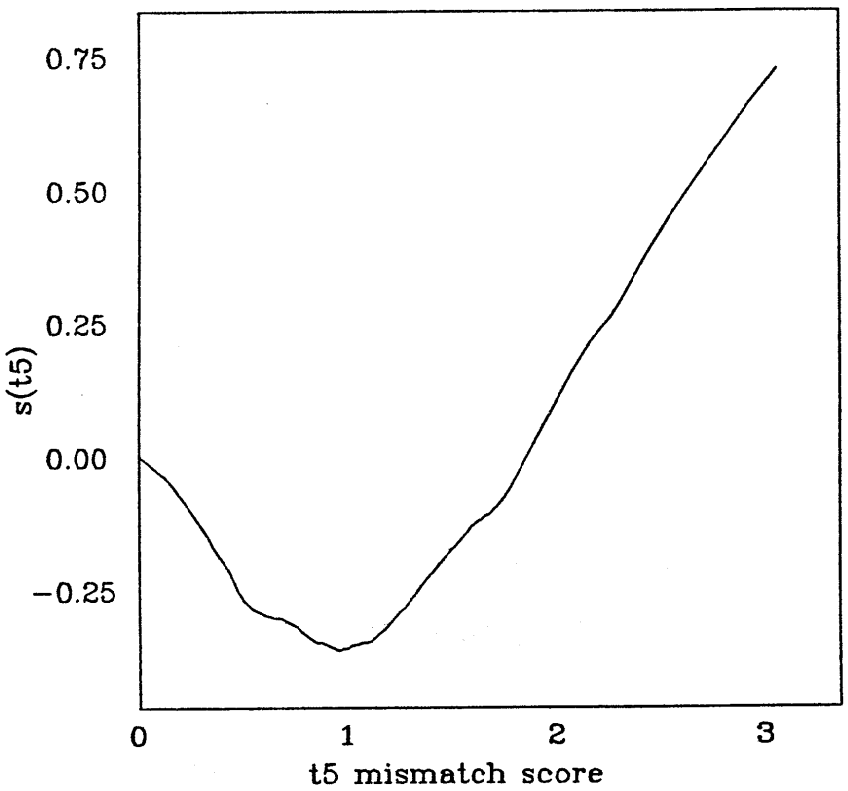


Figure (7)

Mouse Leukemia Data: Smooth for Antibody

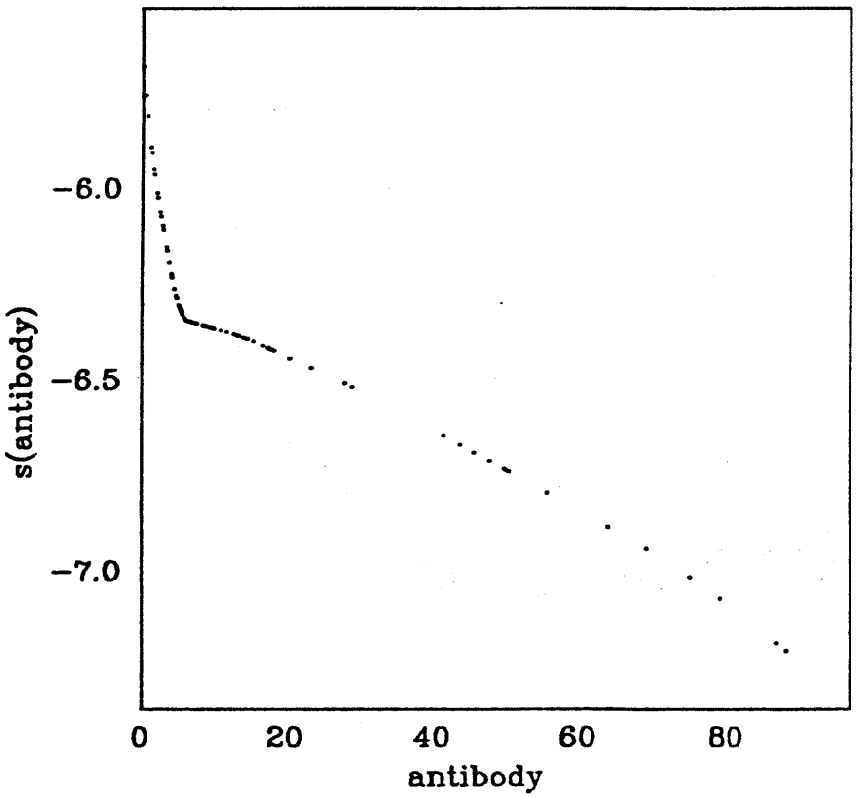


Figure (8)

Bootstrap smooths (Resampling (y_i, x_i, δ_i))

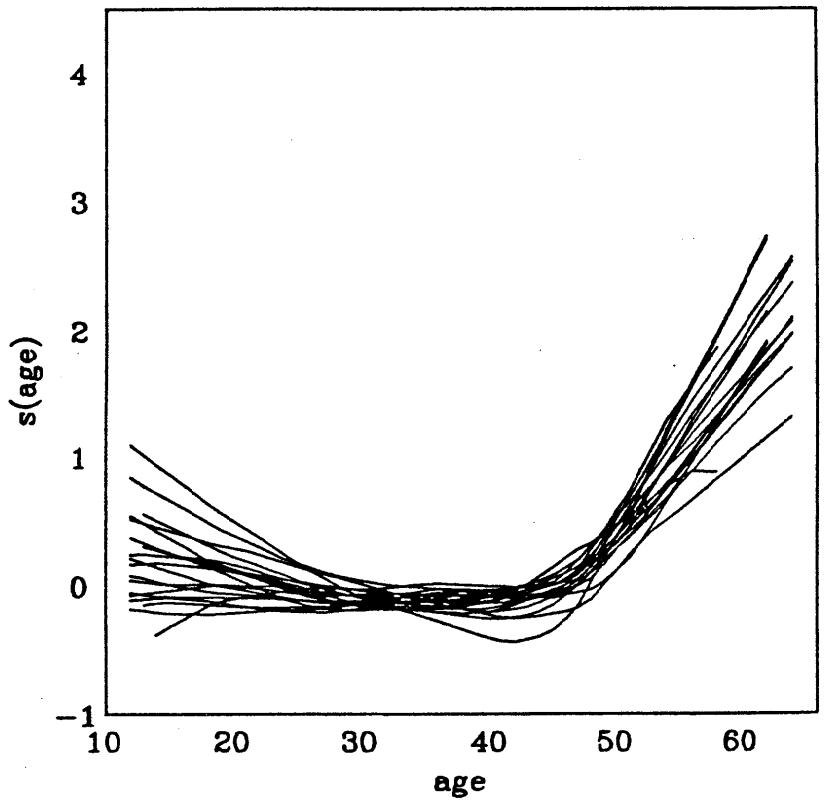


Figure (9)

Bootstrap smooths (Resampling residuals)

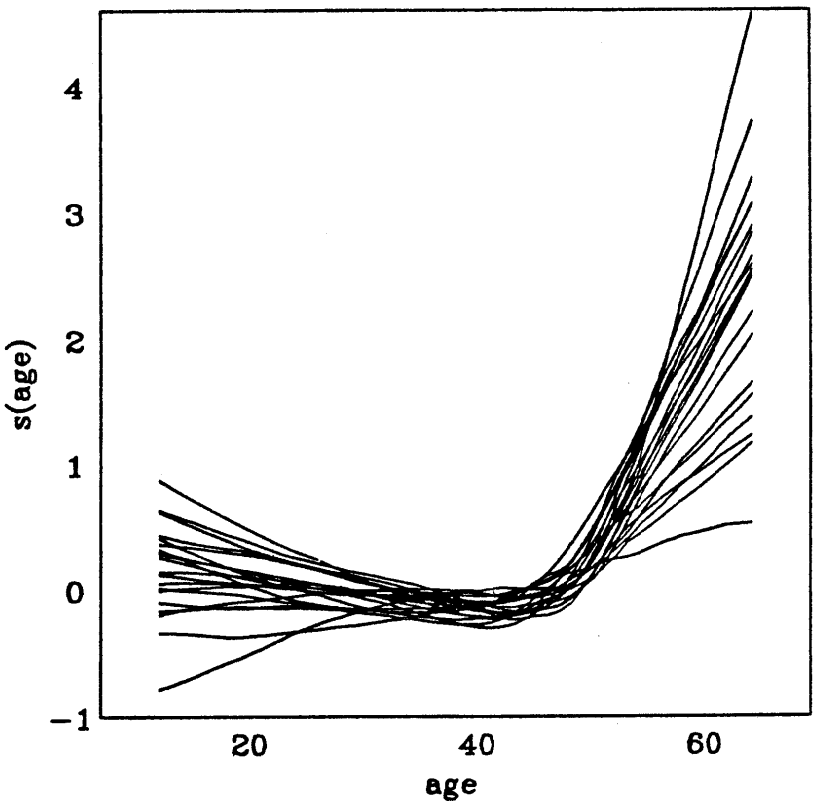


Figure (10)

Estimates of Relative Risk for Lung Cancer Data

Solid line: L.L smooth, Broken line: Monotone m.l.e

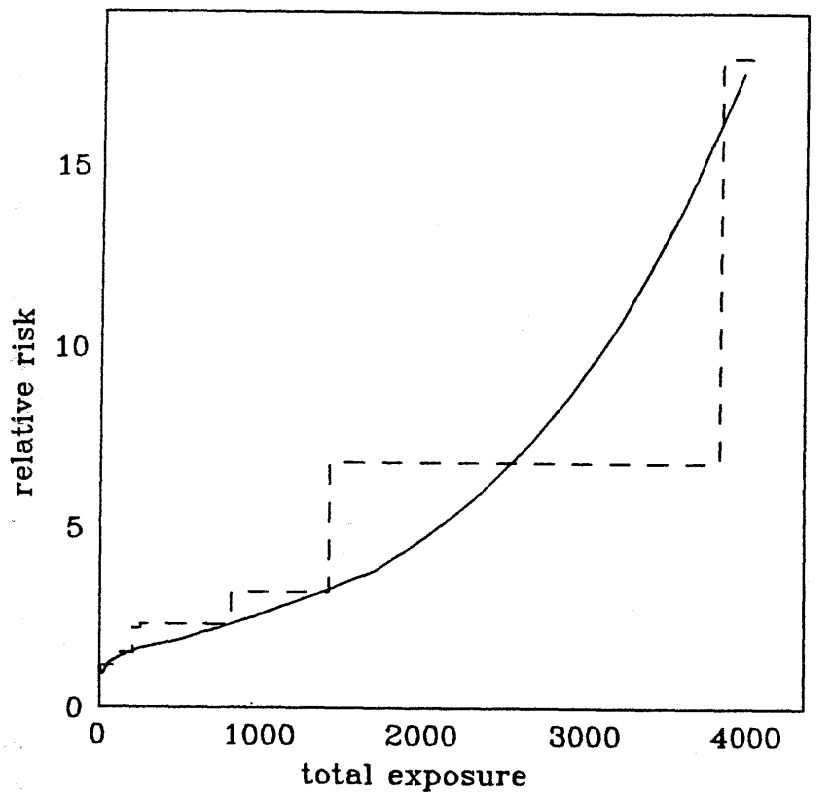


Figure (11)

Quadratic data

Square: uncensored, Plus: censored

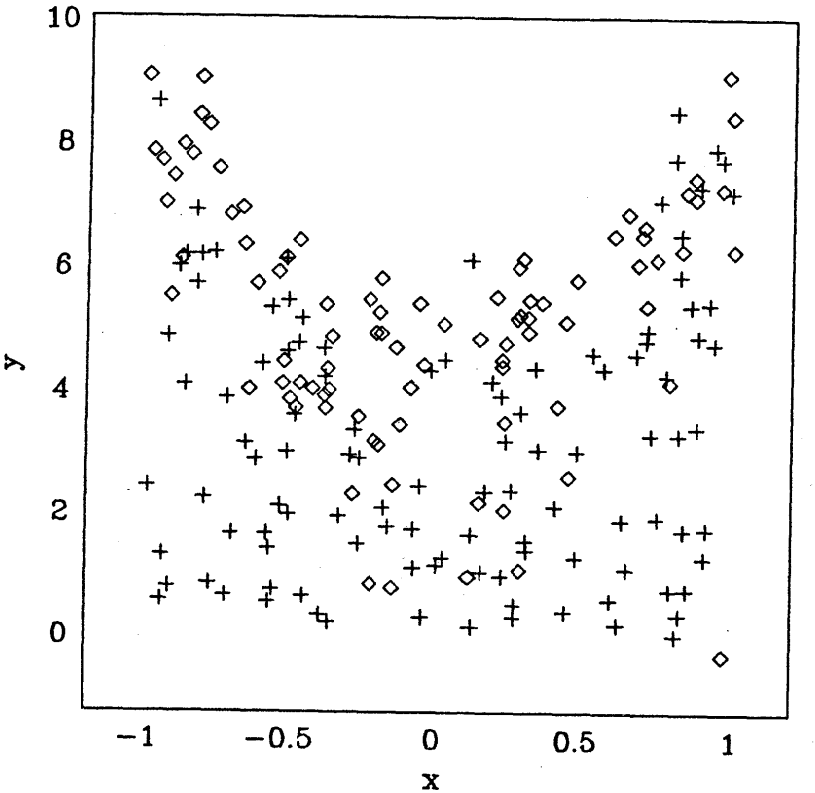


Figure (12)

Local likelihood fit: span .3

Solid line: L.L fit, Broken line: true quadratic function

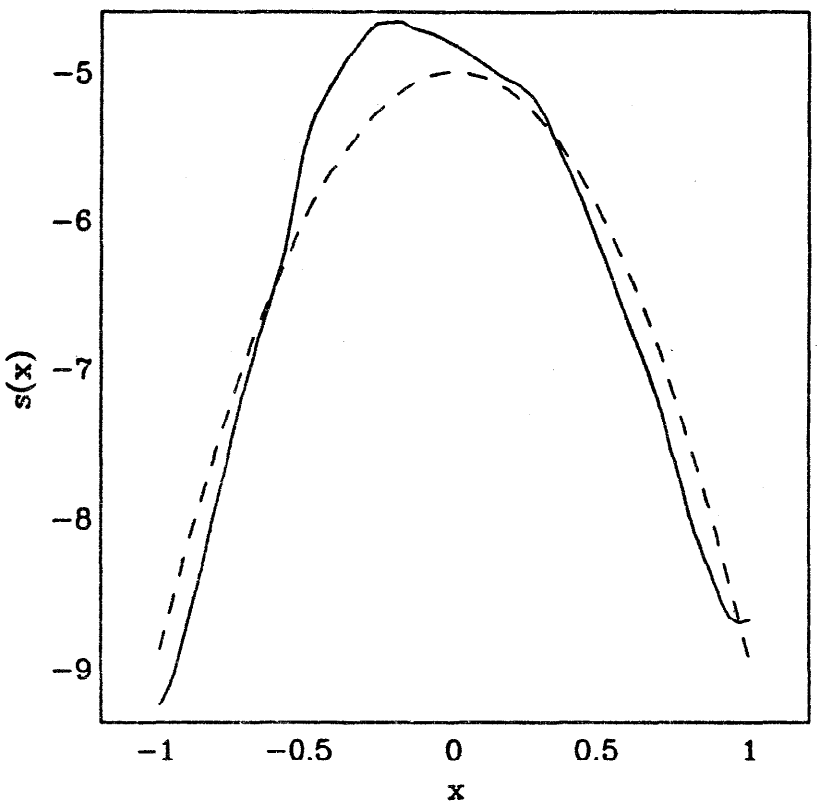


Figure (13)

Local likelihood fit: span .4

Solid line: L.L fit, Broken line: true quadratic function

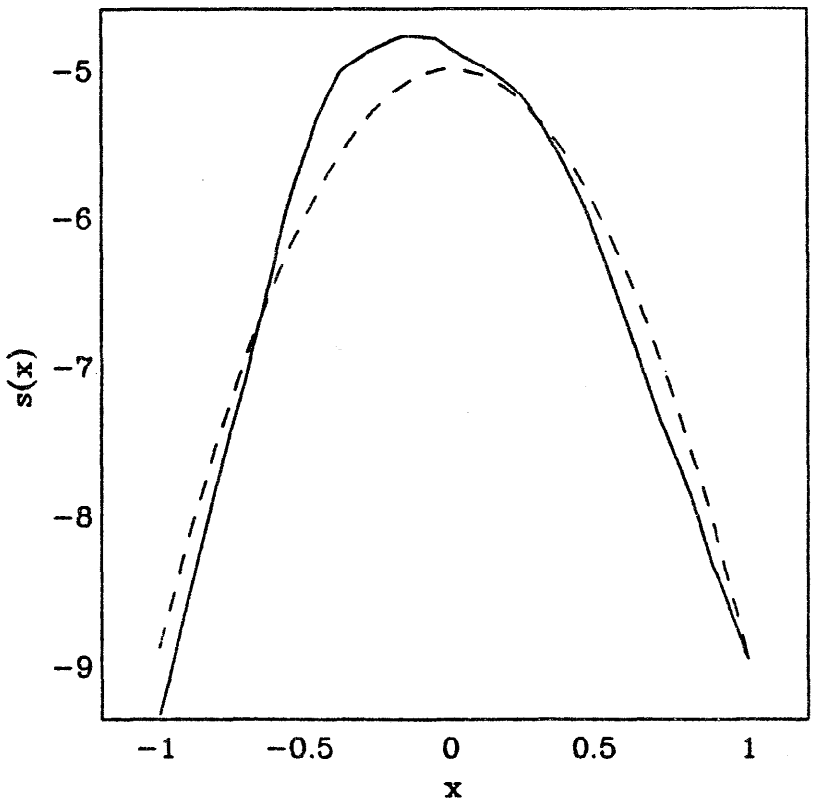


Figure (14)

Local likelihood fit: span .5

Solid line: L.L fit, Broken line: true quadratic function

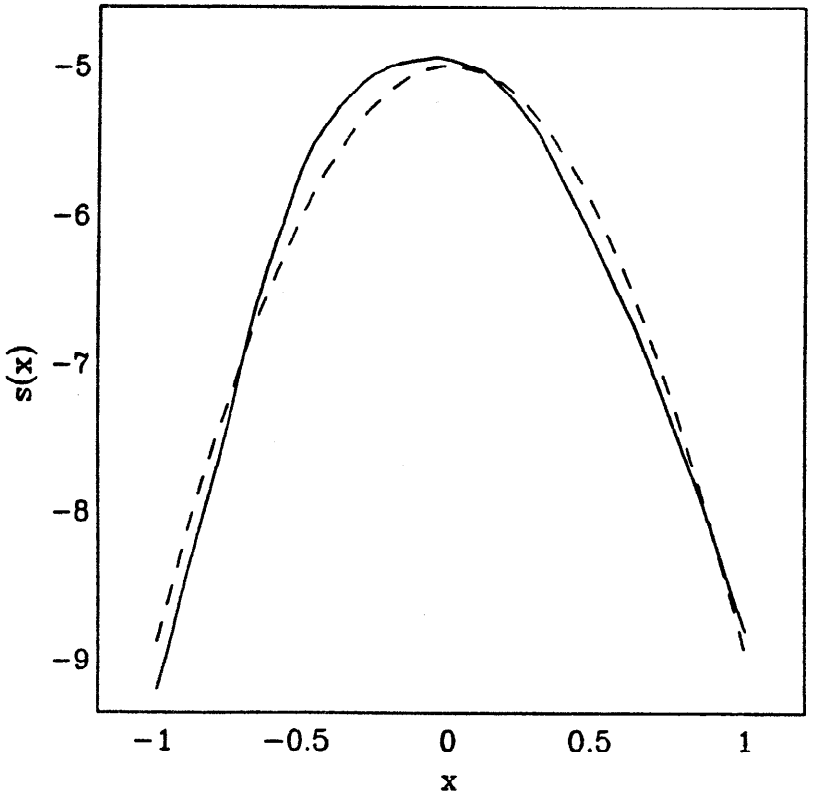


Figure (15)

Local likelihood fit: span .6

Solid line: L.L fit, Broken line: true quadratic function

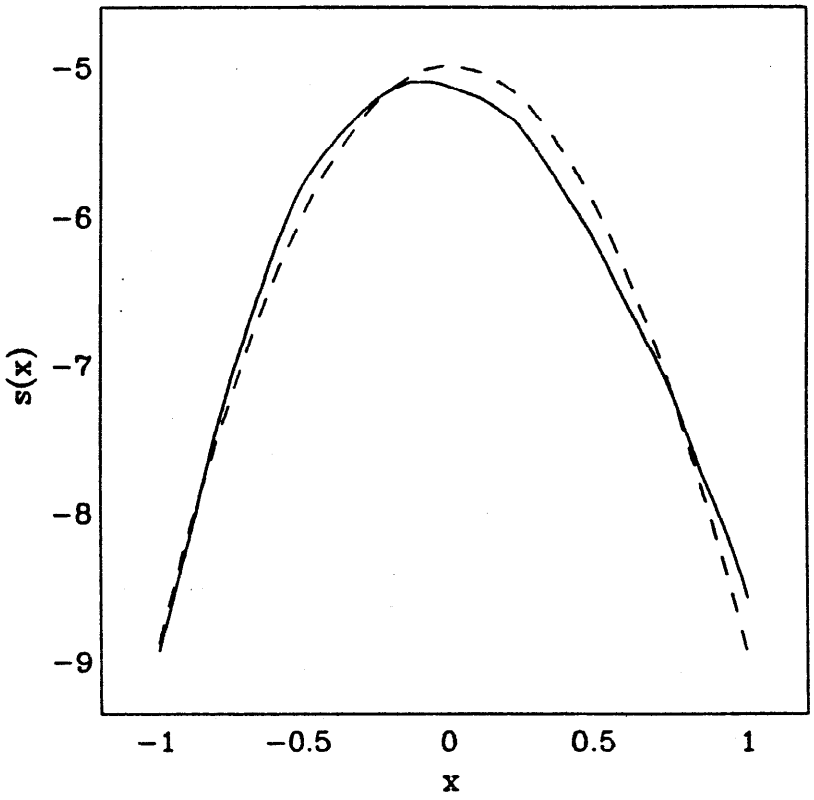


Figure (16)

Local likelihood fit: span .7

Solid line: L.L fit, Broken line: true quadratic function

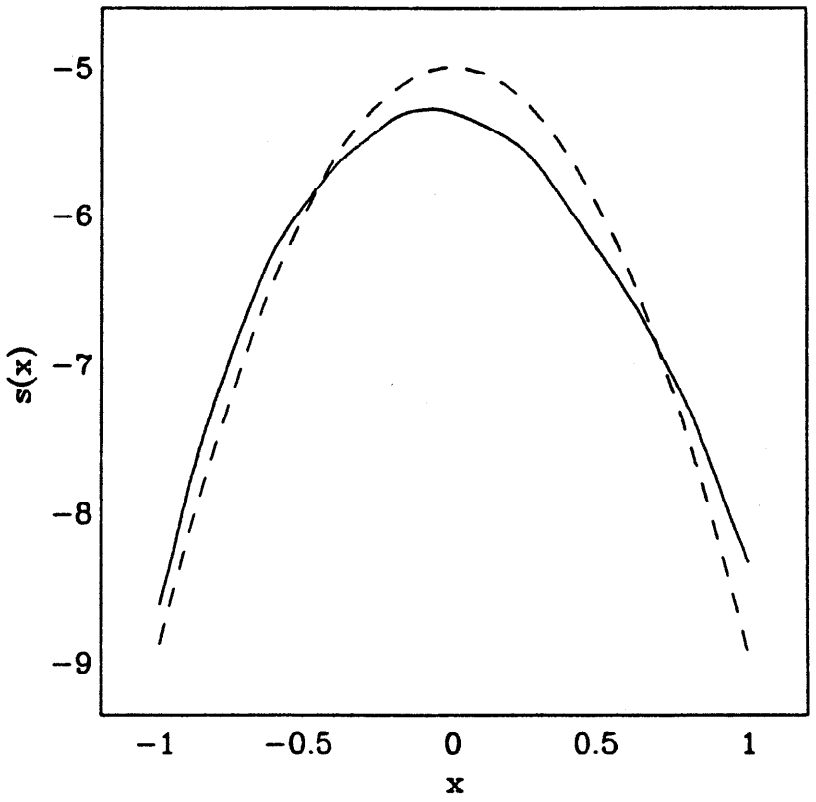


Figure (17)

Average of 20 Local likelihood fits, varying span

Solid line: L.L fit, Broken line: true quadratic function

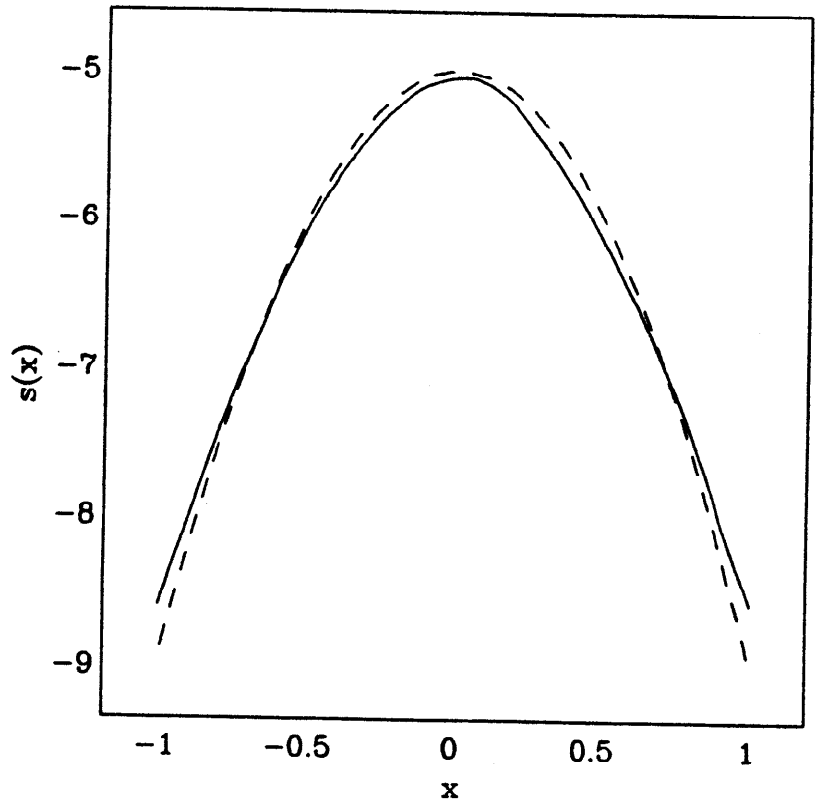


Figure (18)

Local likelihood fit: span .3

Solid line: L.L fit, Broken line: true quadratic function

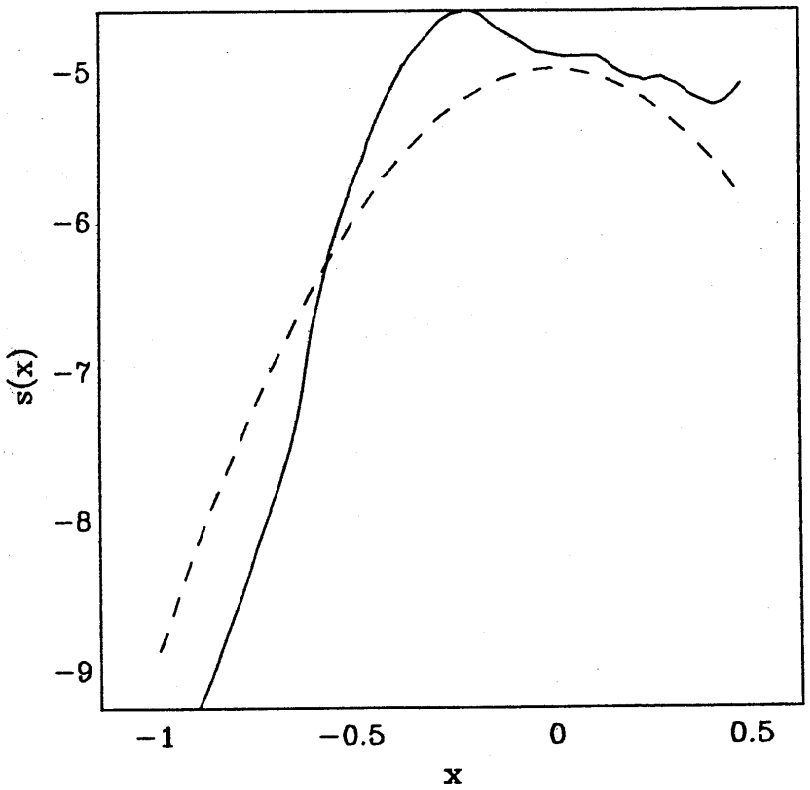


Figure (19)

Local likelihood fit: span .4

Solid line: L.L fit, Broken line: true quadratic function

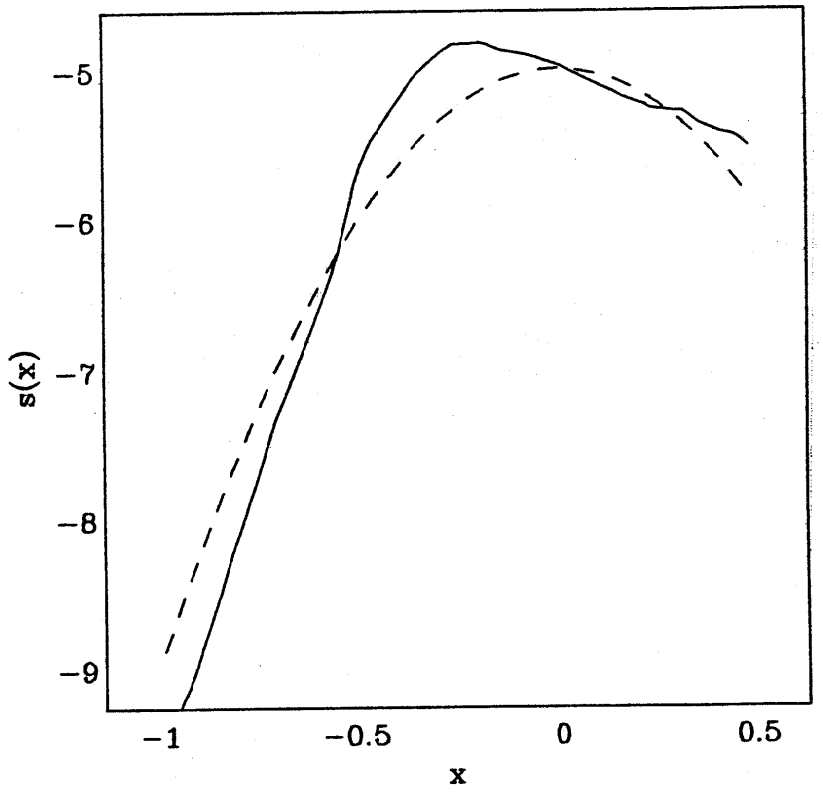


Figure (20)

Local likelihood fit: span .5

Solid line: L.L fit, Broken line: true quadratic function

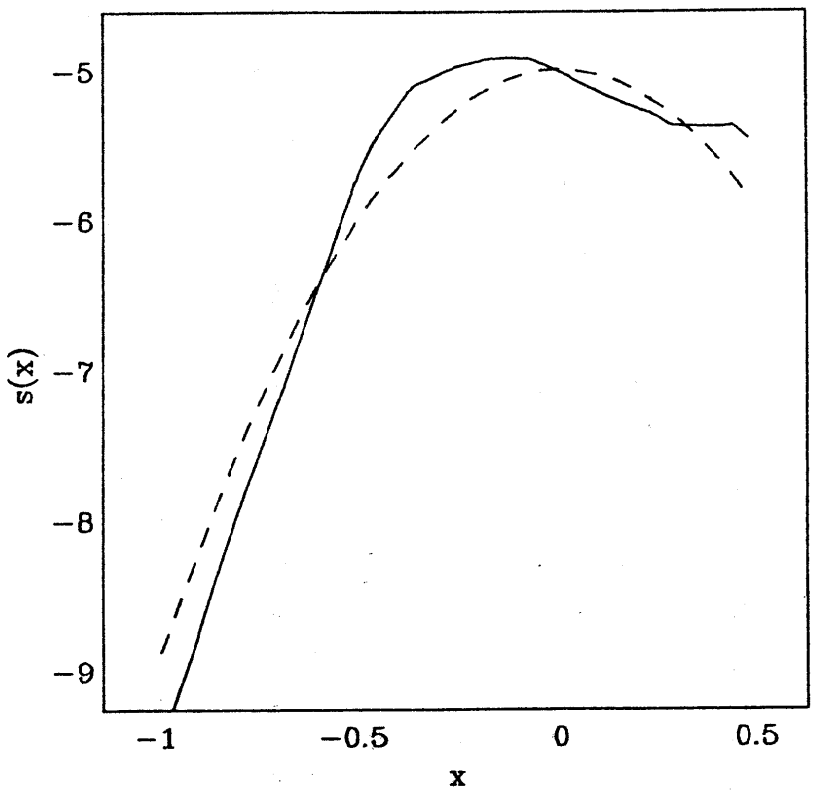


Figure (21)

Local likelihood fit: span .6

Solid line: L.L fit, Broken line: true quadratic function

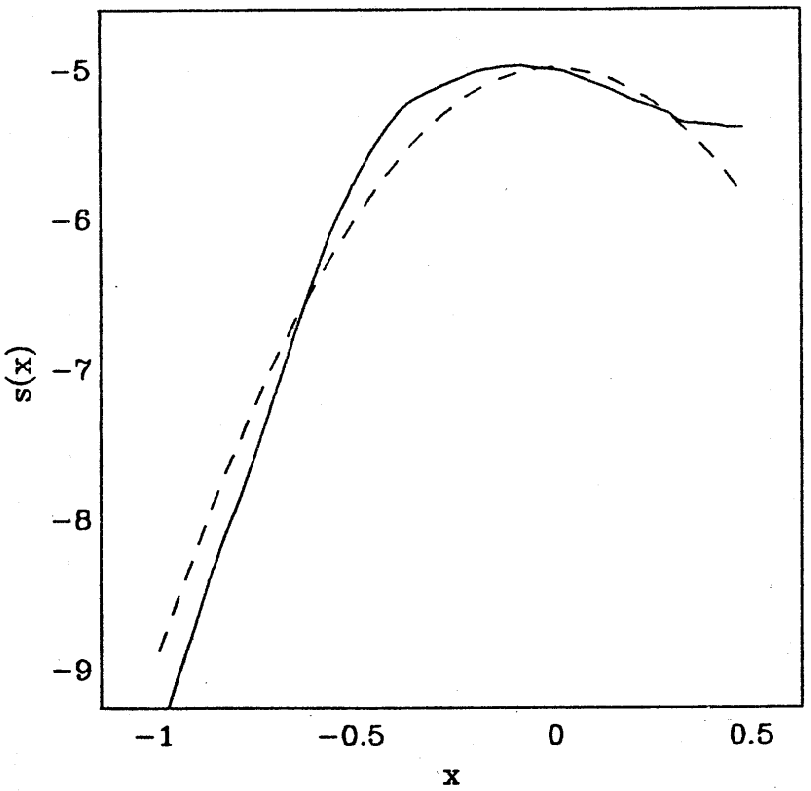


Figure (22)

Local likelihood fit: span .7

Solid line: L.L fit, Broken line: true quadratic function

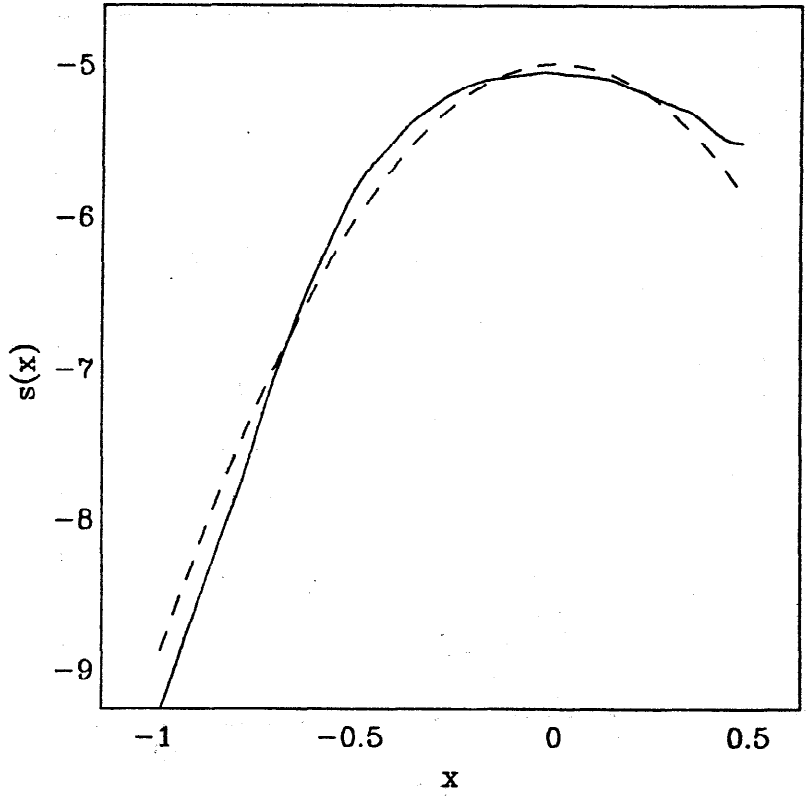


Figure (23)

Average of 20 Local likelihood fits, varying span

Solid line: L.L fit, Broken line: true quadratic function

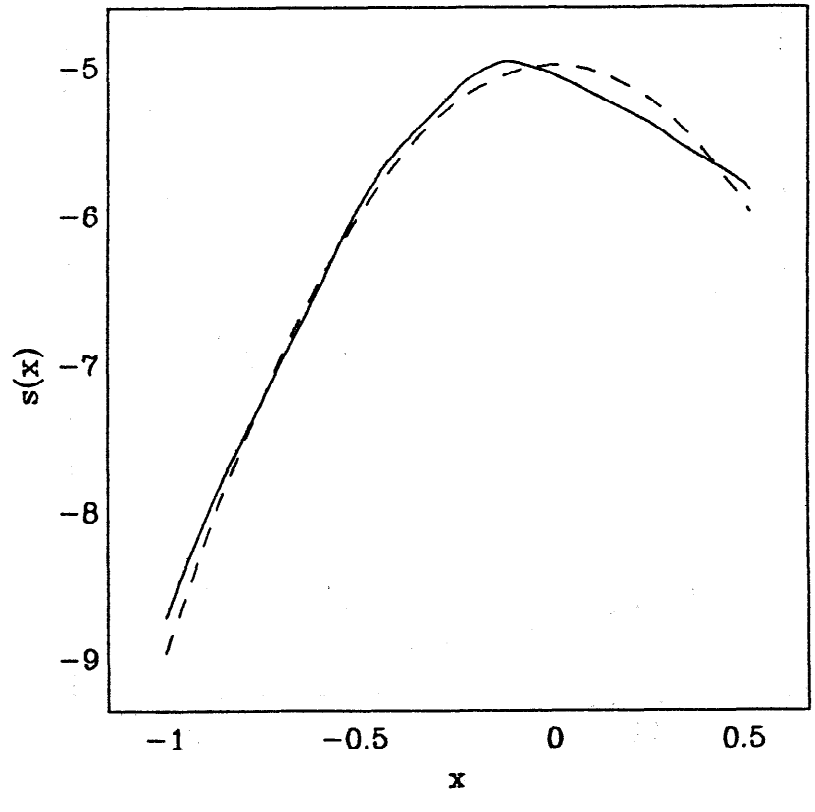


Figure (24)

Solid line: Weighted smooth: no outlier

Broken Line: Unweighted smooth, no outlier

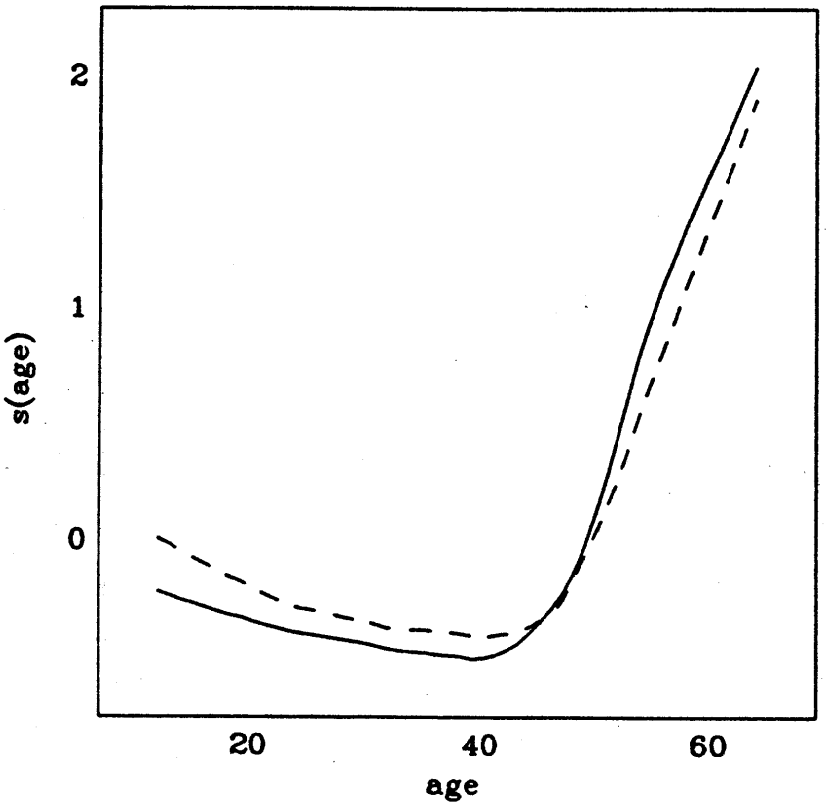
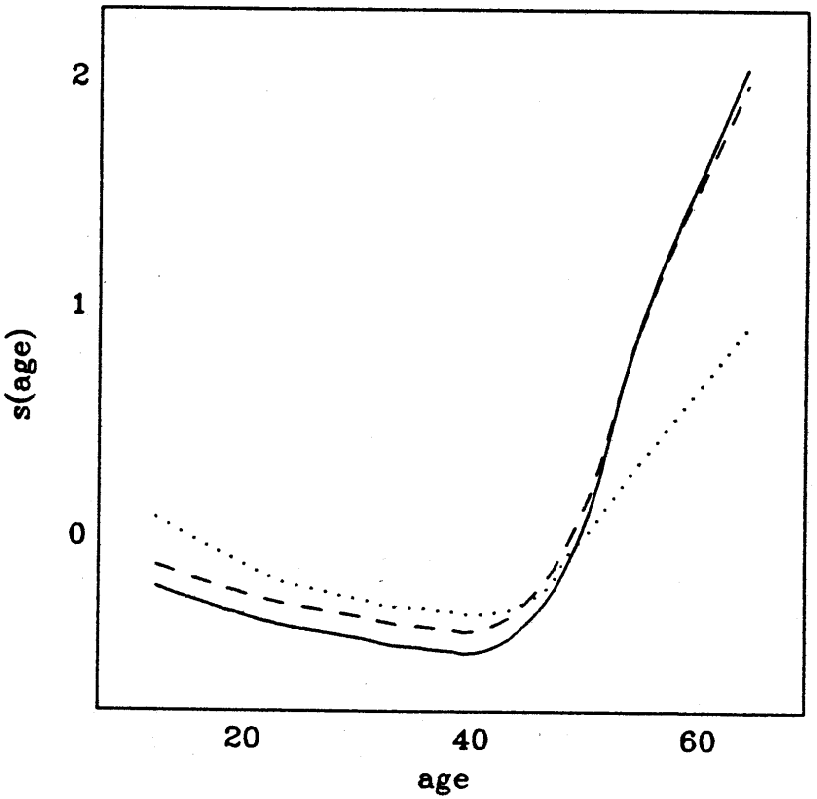


Figure (25)

Solid line: Weighted smooth, no outlier

Broken and dotted lines: Weighted and unweighted smooths with outlier



Chapter 4

Application to the Logistic Model

4.1. Introduction.

In Chapter 2, we discussed how the local likelihood technique could be applied to any generalized linear model. Probably the most commonly used such model (besides the normal regression model) is the linear logistic model for binary data. In this chapter, we'll illustrate the local likelihood procedure in this setting. Further discussion can be found in Hastie (1983) and Hastie and Tibshirani(1984).

4.2. The Problem and a Review of the Linear Logistic Model.

We have data of the form $\{(y_1,x_1),(y_2,x_2),...(y_n,x_n)\}$ where the response y is 0 or 1 and x is an explanatory variable. The observations are assumed to be independent. The problem is to investigate the dependence of y on x.

Let $x=(1,x)$ and let $p(x)=P(y=1\mid x)$. The log likelihood of the data is

$$\log L=\sum_{j=1}^{n}\{y_i\log p_j+(1-y_j)\log(1-p_j)\} \tag{4.1}$$

where $p_j=p(x_j)$. Letting X represent the matrix with jth row equal to $(1,x_j)$, the score equation has the form

$$X^t(y-p)=0 \tag{4.2}$$

The *linear logistic* model assumes that

$$logit\ p(x)=x^t\beta \tag{4.3}$$

Written as a function of β, the log likelihood is

$$\log L(\beta)=\sum_{j=1}^{n}\{y_ix_i^t\beta-\log(1+e^{x_i^t\beta})\} \tag{4.4}$$

A Newton-Raphson procedure is typically used to find $\hat\beta$. The expected information matrix is

$$I(\beta)=X^tDiag\{p_j(1-p_j)\}X \tag{4.5}$$

and the Newton-Raphson iteration has the form

$$\hat\beta_{new}=\hat\beta_{old}+I^{-1}(\beta_{old})X^t(y-\hat p_{old}) \tag{4.6}$$

4.3. The Local Likelihood Generalization.

The formulation of section 2.3 for generalized linear models can be applied directly. Instead of assuming a linear form for $logit\ p(x)$, we assume

$$logit\ p(x)=s(x) \tag{4.7}$$

The local likelihood for x_i is

$$\log L_i(\beta_i)=\sum_{j\in N_i}\{y_jx_j^t\beta_i-\log(1+e^{x_j^t\beta_i})\} \tag{4.8}$$

Letting $\hat\beta_i$ maximize $\log L_i(\beta_i)$, the local likelihood estimate of $s(x_i)$ is $\hat s(x_i)=x_i^t\hat\beta_i$.

4.3.1. Span Selection and Multiple Covariates

As discussed in Section 2.3, the span k is chosen to minimize an approximate AIC criterion:

$$AIC=-2\log L+2\ trace(P(s)) \tag{4.9}$$

With multiple covariates, the model takes the form

$$logit\ p(x)=\sum_{1}^{p}s(\cdot) \tag{4.10}$$

A forward stepwise algorithm is used to select covariates, and backfitting is performed whenever two or more smooths have entered the model. The significance of a smooth can

be judged by comparing of the decrease in $-2\log L$ ("the deviance") to the number of degrees of freedom $trace(P(s))$.

4.4. An Example: Breast Cancer Data.

A study conducted between 1958 and 1970 at the University of Chicago's Billings Hospital concerned the survival of patients who had undergone surgery for breast cancer (Haberman (1976)). There are 306 observations on 4 variables.

y=1 if patient survived $\geq$ 5 years, 0 otherwise

x_1=age of patient at time of operation

x_2=year of operation

x_3=number of positive axillary nodes detected

The local likelihood procedure applied to all three covariates produced the smooths shown in Figures (4.1) , (4.2) , and (4.3) . Table 1 shows the decrease in deviance due to each variable.

Table 4.1. Analysis of Breast Cancer Data

Model	*Deviance*	*Number of Parameters*
Constant	353.67	1
# of nodes(span= .5)	319.8	2.4
# of nodes + Age(span = .6)	310.45	2.4 + 2.4
# of nodes + Age + Yr of oper (span= .5)	307.67	2.4 + 2.5 + 2.4

Age and number of nodes are important, year of operation is not. The final model has a deviance of 307.74 on (306-2.41-2.54-2.41)=298.54 degrees of freedom.

Landwehr et al (1984) analyzed this data set to explore the usefulness of partial residual plots in identifying parametric forms of covariate effects. Their final model was

$$logit\ p(x) = \beta_0 + x_1\beta_1 + x_1^2\beta_2 + x_1^3\beta_3 + x_2\beta_4 + x_1x_2\beta_5 + (log(1+x_3))\beta_6 \qquad (4.11)$$

The deviance of this model is 302.3 on 299 degrees of freedom. The fitted terms for each covariate are super-imposed on Figures (4.1) , (4.2) , and (4.3) (broken lines). The functions for x_1 and x_3 are very similar; they differ for x_2, but the overall effect of this variable is very small.

Hastie (1984) and Hastie and Tibshirani(1984) discuss the relative merits of the local likelihood and partial residual plot procedures. They give two reasons to suggest why the local likelihood procedure is preferable:

- The partial residual technique, in suggesting the parametric form for a covariate effect, relies on the assumption that the covariate forms for other effects are correct. Indeed these effects are usually assumed to be linear. The local likelihood procedure finds the best functional form for all covariates simultaneously.

- The partial residual technique requires quite a bit of ingenuity in identifying the various covariate effects. The local likelihood procedure, on the other hand, is automatic.

4.5. Comparison to the Scatterplot Smoothing Approach.

The local likelihood method extends the linear logistic model through a type of local averaging within the likelihood framework. Computationally, it would seem simpler to ignore the fact that the y's are 0's and 1's and apply scatterplot smoothing techniques directly. This works fine for a single covariate: a scatterplot smooth of y on x_1 is shown in Figure (4.4) . On the same figure, the estimated local likelihood probability smooth $\exp(\hat{s}(x_1)/(1+\exp(\hat{s}(x_1))$ is shown (broken line). Not surprisingly, the two smooths are similar.

With multiple covariates, the local likelihood approach is more attractive for precisely the same reasons that the linear logistic model has gained popularity. In fitting a model of the form $y = \sum_1^p s(x_i)$, one would have to ensure that the fitted probabilities lie between 0 and 1. This would require some sort of truncation of the smooths. On the other hand, the local likelihood approach models *logit p* so the fitted probabilities are always between 0 and 1. Secondly, the local likelihood approach produces an additive model on the logit

scale. A large body of literature suggests that for many types of data, effects are more likely to be additive on the logit scale than on the probability scale. One could try to adapt the regression approach by grouping the y's then using the *logit* of the grouped values as responses. This would likely produce similar results to the local likelihood approach *if* the information loss due to grouping wasn't too large. More details can be found in Hastie and Tibshirani(1984).

Figure (1)

Estimates for Age

Solid line: L.L smooth, Broken line: parametric function

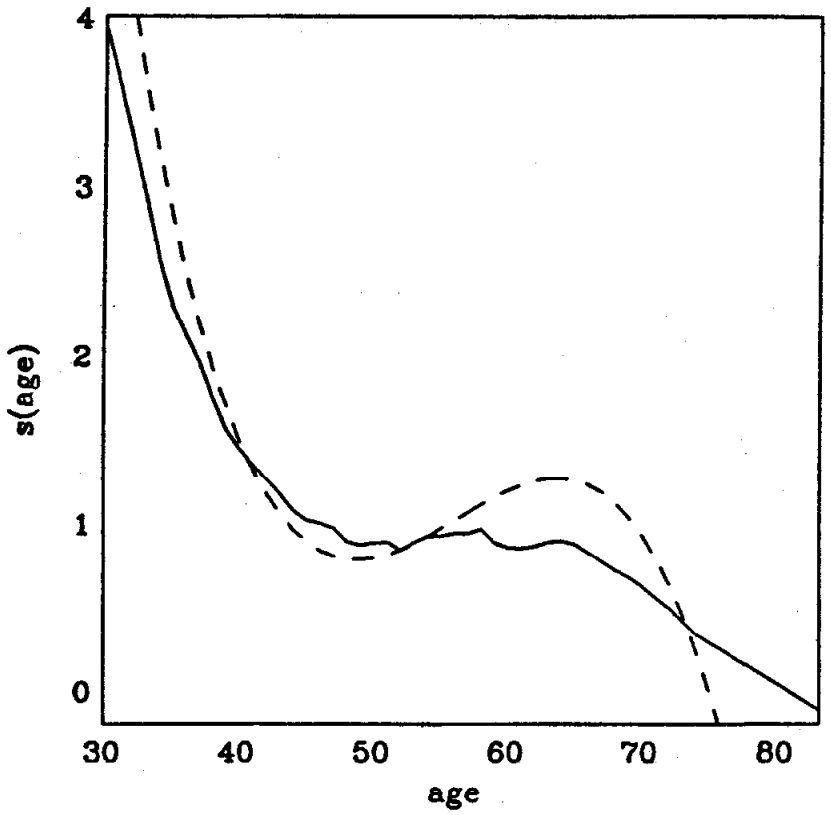


Figure (2)

Estimates for Year of operation

Solid line: L.L smooth, Broken line: parametric function

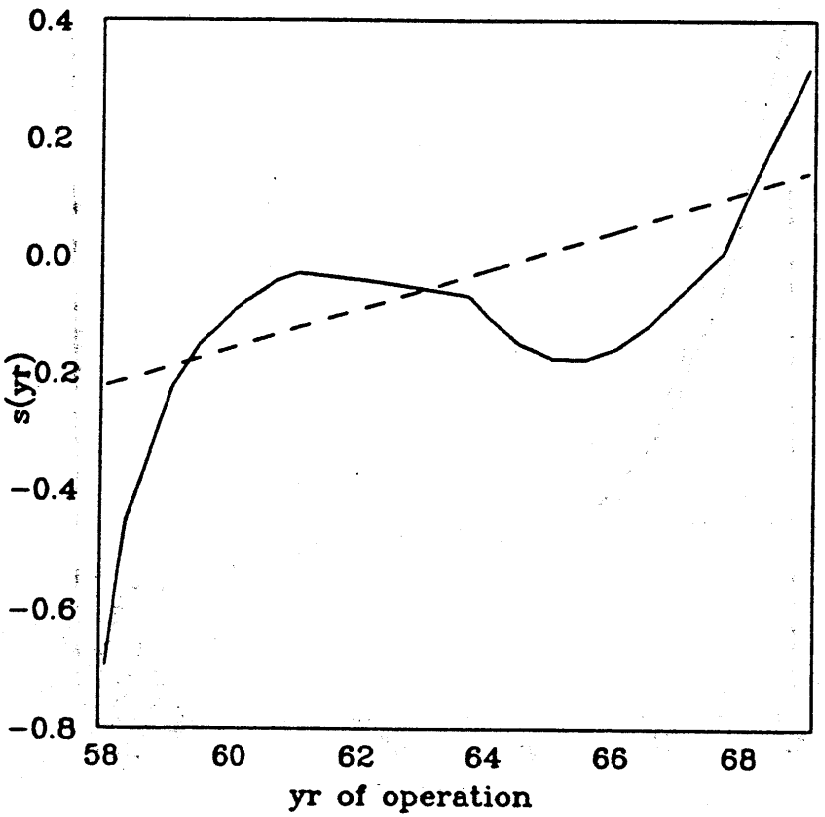


Figure (3)

Estimates for # of nodes

Solid line: L.L smooth, Broken line: parametric function

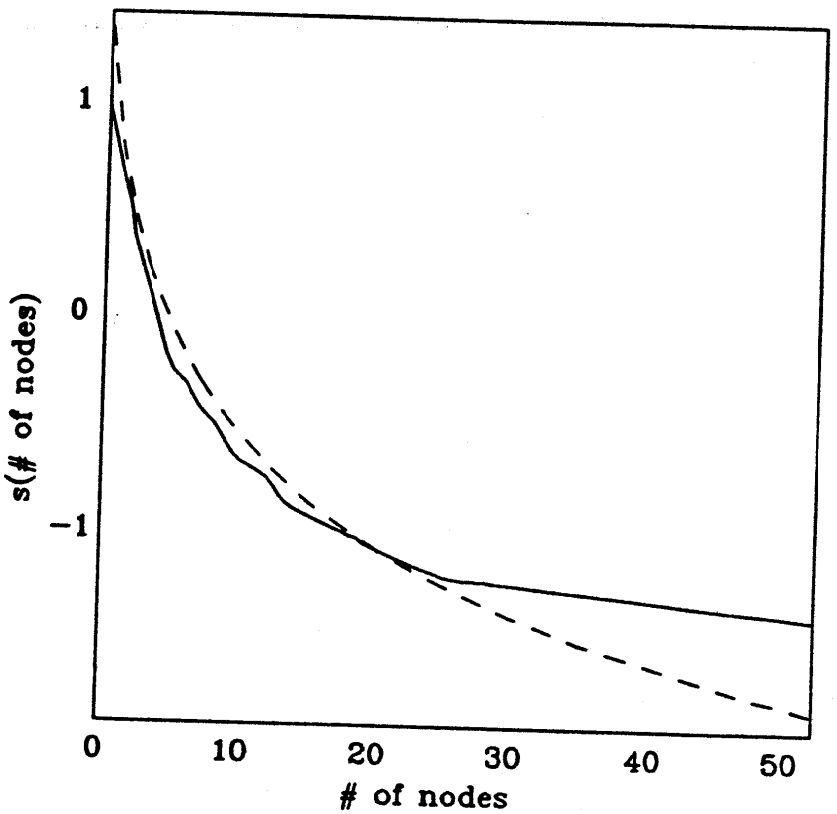
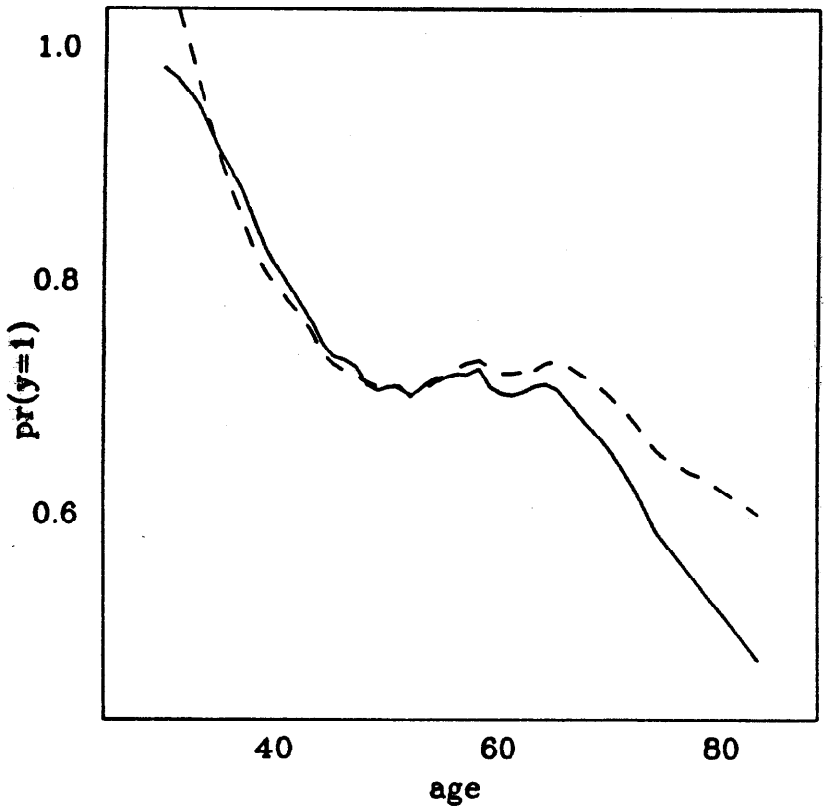


Figure (4)

Estimates for Age

Solid line: L.L smooth, Broken line: Scatterplot smooth



Chapter 5

Asymptotic Theory For Local Likelihood Estimates

5.1. Introduction.

In this chapter we show that, in a sense to be made precise, local likelihood estimates possess the asymptotic optimality properties of maximum likelihood estimates. We'll consider only the exponential family case; with appropriate conditions, the results should be generalizable to any regular family. At the end of the chapter, we conjecture (without proof) a result for the proportional hazards model.

5.2. Local Likelihood Estimates in the Exponential Family.

Since LLE's are just maximum likelihood estimates calculated locally, we can derive results for LLE's by modifying standard MLE theory to account for the local nature of the estimation. We begin with the MLE theory for generalized linear models provided by McCullagh (1983) and modify it appropriately.

5.2.1. A Review of Results for Generalised Linear Models

A special case of McCullagh's results is the following. Suppose $Y_1, \ldots Y_n$ are independent random variables with density

$$Y_i \sim \exp\{\{y_i\theta_i - b(\theta_i) - c(y_i, \sigma)\}/\sigma^2\} \tag{5.1}$$

We assume that $\boldsymbol{\theta}$, the parameter of interest, is expressible in the form $\boldsymbol{\theta} = X\boldsymbol{\beta}$ where X is a fixed n by p design matrix and $\boldsymbol{\beta}$ is a p-dimensional parameter. Letting $\boldsymbol{u}(\boldsymbol{\beta}) = E(\boldsymbol{Y}) =$

$b'(\theta)$, the score function has the form

$$U_\beta = X^t(Y - u(\beta))/\sigma^2 \tag{5.2}$$

Let $i_\beta = Var(U_\beta) = \sigma^2 X^t V X$ where $V\sigma^2 = cov(Y) = b''(\theta)\sigma^2$, and denote by $u_\beta(\beta, y) = X^t(y - u(\beta))/\sigma^2$ the score equation used to determine $\hat{\beta}$. We make the following regularity assumptions:

$$\frac{i_\beta}{n} \to C > 0 \tag{A1}$$

$$E|Y_i|^3 \leq M_1 < \infty \ \forall i \tag{A2}$$

$$|x_{ij}| \leq M_2 < \infty \ \forall i, j \tag{A3}$$

$$u_\beta(\beta, y) \ \textit{has a continuous 2nd derivative} \tag{A4}$$

With these assumptions, the following results can be derived:

$$n^{-1/2}U_\beta \sim N_p(0, \sigma^2 i_\beta/n) + O_p(n^{-1/2}) \tag{5.3}$$

$$E(\hat{\beta} - \beta) = O(n^{-1}) \tag{5.4}$$

and

$$n^{1/2}(\hat{\beta} - \beta) \sim N_p(0, n\sigma^2 i_\beta^{-1}) + O_p(n^{-1/2}) \tag{5.5}$$

The proof of these results follow the standard proofs for MLE's. The score U_β is a sum of independent random variables with mean 0 and covariance i_β. Assumptions (A1)—(A3) ensure that Liaponouv's condition is satisfied and the central limit theorem implies (5.3) . Expanding $u_\beta(\hat{\beta}, y) = 0$ around β gives

$$0 = u_\beta(\beta, y) - I_{\beta^*}(\hat{\beta} - \beta) \tag{5.6}$$

where I_{β^*} is minus the second derivative matrix evaluated at a point β^* lying on the line segment joining β and $\hat{\beta}$. Now $I_\beta = O_p(n^{1/2})$, $i_\beta = O(n)$ and $I_\beta - i_\beta = O_p(n^{1/2})$ with $E(I_\beta) = i_\beta$. Since $U_\beta = O_p(n^{1/2})$, we see from (5.6) that there exists a root $\hat{\beta}$ satisfying $\hat{\beta} - \beta = O_p(n^{-1/2})$ and

$$\hat{\beta} - \beta = I_{\beta^*}^{-1}U_\beta = i_{\beta^*}^{-1}U_\beta + O_p(n^{-1}) = i_\beta^{-1}U_\beta + O_p(n^{-1}) \tag{5.7}$$

(assuming $i_\beta^{-1} = O(n^{-1})$ Taking expectations in (5.7) gives (5.4) ; (5.5) follows by combining (5.3) and (5.7) , and applying Slutsky's theorem.

5.3. Some Remarks.

- McCullagh starts with the more general score equation $D^tV^-(Y - u(\beta)) = 0$ where $D = du/d\beta$ and $V = Cov(Y)\sigma^2$. (This reduces to the form $X^t(Y - u(\beta))$ when the link function is such that $\theta = X\beta$). From this he proves consistency and asymptotic normality of the estimate $\hat{\beta}$. Also, he notes that to obtain the asymptotic results, it is not necessary to assume a form for the likelihood: one need only assume that the score equation has the form $D^tV^-(Y - u(\beta))$. Since this equation only depends on the first two moments of Y, there can be more than one likelihood giving the same score equation. McCullagh calls any likelihood giving this score function a "quasi-likelihood". If Y is in the exponential family and the log-likelihood is linear in y, then the likelihood and quasi-likelihood correspond. In other cases, there can be more than one likelihood resulting in the same quasi-likelihood. In this event, the quasi-likelihood estimate may not equal the MLE, but it is still consistently and efficiently estimates the true parameter. According to McCullagh, "quasi-likelihood" estimation could be useful in a situation in which one isn't willing to assume a specific form for the likelihood, but is willing to specify a relationship between the mean and variance.

- McCullagh's results as stated in their full generality seem to be wrong; he doesn't assume that the Y_i's are independent and allows a general covariance structure V. In this case, the score function U_β is no longer a sum of independent random variables and asymptotic normality doesn't necessarily hold.

5.3.1. Local Likelihood Estimation

Consider initially a sample of size n containing an observation at a point x_0. We shall establish asymptotic properties of the LLE at x_0. We assume that Y_i is distributed according to the exponential family (5.1) , with $\theta_i = s(x_i)$. Let k_n be the number of points in the neighborhood N_0^n used for estimating $s(x_0)$. Assume that as $n \to \infty$, $k_n \to \infty$, but

the neighborhood shrinks so that $max_{(i,j \in N_0^n)} |x_i - x_j| = o(k_n^{-1/2})$. We argue below that for estimation of the slope and intercept of the line tangent to $s(\cdot)$ at x_0, the LLE is consistent and asymptotically normal, and has the efficiency of a MLE based on sample size k_n. This implies that, for estimation of $s(x_0)$, the LLE has minimum asymptotic mean squared error among all estimates based on k_n observations.

In this set-up, $p = 2$ and $X = (1, x)$. The score function for the local likelihood at x_0 is

$$U_\beta = X^t W(Y - u(\beta))/\sigma^2 \tag{5.8}$$

where $u(\beta) = b'(X\beta)$, and $W = Diag\{I(i \in N_0^n)\}_{n \times n}$.

Let $\beta = (\beta_1, \beta_2)$ be the coefficents of the line tangent to $s(\cdot)$ at x_0 i.e. $\beta_2 = s'(x_0)$ and $\beta_1 = s(x_0) - \beta_2 x_0$. We make assumptions (A1)—(A4) as well as the following:

$$|s(\cdot)| \leq M_3 < \infty \tag{A5}$$

$$s'(\cdot) \text{ exists with } |s'(\cdot)| \leq M_4 < \infty \tag{A6}$$

$$b''(\cdot) \text{ exists with } |b''(\cdot)| \leq M_5 < \infty \tag{A7}$$

Under these conditions, the following results obtain:

$$k_n^{-1/2} U_\beta \sim \mathcal{N}_2(0, \sigma^2 i_\beta / k_n) + O_p(k_n^{-1/2}) \tag{5.9}$$

$$E(\hat{\beta} - \beta) = O(k_n^{-1}) \tag{5.10}$$

and

$$k_n^{1/2}(\hat{\beta} - \beta) \sim \mathcal{N}_2(0, k_n \sigma^2 i_\beta^{-1}) + O_p(k_n^{-1/2}) \tag{5.11}$$

These imply the following results for the local likelihood estimate $\hat{s}(x_0) = \hat{\beta}_1 + \hat{\beta}_2 x_0$:

$$E(\hat{s}(x_0) - s(x_0)) = O(k_n^{-1}) \tag{5.12}$$

and

$$k_n^{1/2}(\hat{s}(x_0) - s(x_0)) \sim \mathcal{N}(0, k_n \sigma^2 A) + O_p(k_n^{-1/2}) \tag{5.13}$$

where $A = (1\ x_0) i_\beta^{-1} (1\ x_0)^t$.

To prove (5.9) — (5.11) , we note that by Liaponouv's theorem $k_n^{-1/2} X^t W(Y - E(Y)) \rightarrow \mathcal{N}_2(0, \sigma^2 i_\beta / k_n) + O_p(k_n^{-1/2})$. Hence, we need only show that

$$k_n^{-1/2} X^t W(E(Y) - u(\beta)) \rightarrow 0 \tag{5.14}$$

Equation (5.14) implies that $k_n^{-1/2} X^t W(Y - u(\beta))$ has the same limiting distribution as $k_n^{-1/2} X^t W(Y - E(Y))$ and the results (5.9) — (5.11) then follow from those of the previous section, with n replaced by k_n.

To establish relation (5.14) , we expand each term of $E(Y_i) - u_i(\beta) = b'(s(x_i)) - b'(\beta_1 + \beta_2 x_i)$ in a one term Taylor series as follows:

$$\begin{aligned} b'(s(x_i)) &= b'(s(x_0) + (x_i - x_0)s'(h_1)) \\ &= b'(s(x_0)) + (x_i - x_0)s'(h_1)b''(h_2) \end{aligned} \tag{5.15}$$

$$\begin{aligned} b'(\beta_1 + \beta_2 x_i) &= b'(\beta_1 + \beta_2 x_0 + \beta_2(x_i - x_0)) \\ &= b'(\beta_1 + \beta_2 x_0) + \beta_2(x_i - x_0)b''(h_3) \\ &= b'(s(x_0)) + \beta_2(x_i - x_0)b''(h_3) \end{aligned} \tag{5.16}$$

In the above, $|h_1 - x_0| \leq |x_i - x_0|$, $|h_2 - s(x_0)| \leq |(x_i - x_0)s'(h_1)|$, and $|h_3 - s(x_0)| \leq |\beta_2(x_i - x_0)|$. Combining (5.15) and (5.16) we have

$$b'(s(x_i)) - b'(\beta_1 + \beta_2 x_i) = (x_i - x_0)[s'(h_1)b''(h_2) + \beta_2 b''(h_3)] \tag{5.17}$$

Now $|x_i - x_0| = o(k_n^{-1/2})$ (by assumption) and the remaining terms are bounded. Hence (5.14) is established.

5.3.2. Some Remarks

- As in the global case studied by McCullagh, the preceding results don't require the assumption that Y has distribution of the exponential family (5.1) . We can make

the weaker assumption that the score function has the form $U_\beta = X^t W(Y - u(\beta))/\sigma^2$ with $E(Y) = u(\beta) = b'(s(x))$.

- The results above assumed that the maximum distance between any two points a neighborhood goes down at the rate $o(k_n^{-1/2})$. In the local likelihood procedure, the span is chosen to minimize an Akaike-type criterion. In principle, then, one should show that selecting the span in this way results in the correct order of shrinkage of the neighborhood. We haven't pursued this, however,

- We have established convergence results for the estimate of a single value of the smooth function. With more work, one could presumably show convergence of the entire estimated function to a Gaussian process. Again, we have decided not to go into this.

5.4. Asymptotics for the Proportional Hazards Model.

In this section, we conjecture an asymptotic result for the local likelihood procedure in the proportional hazards model.

Suppose n items are placed on test and give rise to (possibly censored) observation times $\{y_1, y_2, ...y_n\}$ with associated (fixed) covariates $\{x_1, x_2..., x_n\}$. Let $\delta_i = 0$ if y_i is censored and 1 if y_i is uncensored, and following Tsiatis(1980), we assume that the triples (y_i, x_i, δ_i) are i.i.d. Let D the set of indices of the failures among the y_i's. To facilitate construction of a partial likelihood, we will make the usual assumption that the censoring mechanism is non-informative (see Kalbfleisch and Prentice(1980)).

Under the model

$$\lambda(t \mid x) = \lambda_0(t) \exp(x\beta) \tag{5.18}$$

the partial likelihood is

$$PL = \prod_{l \in D} \frac{e^{\beta x_l}}{\sum_{j \in R_l} e^{\beta x_j}} \tag{5.19}$$

and the score function is

$$u(\beta) = \sum_{l \in D} \left(x_l - \frac{\sum_{j \in R_l} x_j e^{\beta x_j}}{\sum_{j \in R_l} e^{\beta x_j}} \right) \tag{5.20}$$

Tsiatis shows that there exist a consistent root $\hat{\beta}$ of the score equation with the following asymptotic behaviour:

$$n^{1/2}(\hat{\beta} - \beta) \rightarrow \mathcal{N}(0, v) \tag{5.21}$$

where $v = \int_0^{T_0} -dQ Var(z \mid R(t))$, $Q(t) = P(t \geq t, \delta = 1)$, and T_0 is an upper bound on Y.

In the local likelihood framework, we assume that the hazard has the form

$$\lambda(t \mid x) = \lambda_0(t) \exp(s(x)) \tag{5.22}$$

where $s(x)$ is some smooth function of x. The derivative $s'(x_0)$ at some fixed point x_0 is estimated by $\hat{\beta}_0$ maximizing the local partial likelihood

$$PL_0 = \prod_{l \in D \cap N_0^n} \frac{e^{\beta_0 x_l}}{\sum_{j \in R_l \cap N_0^n} e^{\beta_0 x_j}} \tag{5.23}$$

The local score equation is

$$u_0(\beta_0) = \sum_{l \in D \cap N_0^n} \left(x_l - \frac{\sum_{j \in R_l \cap N_0^n} x_j e^{\beta_0 x_j}}{\sum_{j \in R_l \cap N_0^n} e^{\beta_0 x_l}} \right) \tag{5.24}$$

As in the exponential family case, we assume that as $n \rightarrow \infty$, $k_n \rightarrow \infty$ and $max_{\{i,j \in N_0^n\}} |x_i - x_j| = o(k_n^{-1/2})$. A reasonable conjecture, under regularity conditions on $s(\cdot)$, is that the local score equation has a consistent root $\hat{\beta}_0$, and asymptotically

$$k_n^{1/2}(\hat{\beta}_0 - s'(x_0)) \rightarrow \mathcal{N}(0, v) \tag{5.25}$$

where v is defined above. Fixing $\hat{s}(x') = s(x') = 0$ for some x', a result like (5.21) could also be obtained for $\hat{s}(x_0)$. This would require a convergence proof for the integral estimator $s(x_0) = \int_{x'}^{x_0} s'(t)dt$, and hence consideration of the simultaneous estimation of $s(\cdot)$ at $x_1, x_2, \ldots x_n$ We will not attempt to prove these results; the simpler case treated by Tsiatis is quite involved. Recently, more general results for the proportional hazards model (not requiring that the triples (y_i, x_i, δ_i) be i.i.d) have been obtained using a martingale approach by Anderson and Gill (1982). A modification of those results to local likelihood estimation should also be possible.

Chapter 6

Degrees of Freedom and AIC approximations

6.1. Introduction.

In the chapter, we provide justifications for 1) the formula *degrees of freedom* = *trace(smoother matrix)* and 2) the use of AIC, in the local likelihood procedure. We also provide a number of simulations to support our claims. As in chapter 4, we concentrate on the exponential family case, although our simulations suggest that similar results are true in the proportional hazards model as well.

The actual result that we derive is the following. Consider two local likelihood fits $\hat{\mathbf{y}}_1$ and $\hat{\mathbf{y}}_2$ with corresponding smoother matrices P_1 and P_2. (By "corresponding smoother matrix" P, we mean that if $\hat{\mathbf{y}}$ is based on a set of X value $\mathbf{x}$ and span s, then P is the matrix producing locally linear least squares fits of span s, based on $\mathbf{x}$.) If on the average, the two smoothers produce the same fit, then the difference in deviance between the two fits has expected value $[trace(P_2) - trace(P_1)]$. Thus we can think of $trace(P)$ as the number of degrees of freedom used up by the smoother based on P. This generalizes the standard hypothesis testing set-up of linear estimation, in which we have two nested fits and we consider the difference in deviance when the smaller model is correct.

We will discuss the scatterplot (Gaussian Likelihood) case first, for which this result is exact. Then we will show that the result holds approximately for local likelihood estimation in the exponential family.

With these results, we then provide a justification for using the *AIC* procedure for span selection.

Before starting, we will review some results on the distribution of quadratic forms.

6.2. The Distribution of Quadratic Forms.

Suppose $\mathbf{y}$ is a random n vector with mean η and variance V. Then if A is an n by n real symmetric matrix, it can be shown that

$$E(\mathbf{y}^t A\mathbf{y}) = \eta^t A\eta + trace(AV) \tag{6.1}$$

If in addition $\mathbf{y}$ has a multivariate normal distribution, then

$$(\mathbf{y}-\eta)^t A(\mathbf{y}-\eta) \sim \sum_1^n \lambda_i \chi_1^2 \tag{6.2}$$

where the λ_i's are the eigenvalues of AV. In particular, this implies

$$Var[(\mathbf{y}^t-\eta)A(\mathbf{y}-\eta)] = 2\sum_1^n \lambda_i^2 \tag{6.3}$$

These results can be found in many books; see for example Guttman (1983).

Now suppose A is not symmetric. Then we can replace A by a symmetric matrix as follows:

$$\mathbf{y}^t A\mathbf{y} = \frac{1}{2}\{\mathbf{y}^t A\mathbf{y} + \mathbf{y}^t A^t\mathbf{y}\} = \mathbf{y}^t A^*\mathbf{y} \tag{6.4}$$

where $A^* = \frac{1}{2}\{A + A^t\}$. From this, we see that (6.1) holds for non-symmetric A's since $\eta^t A^*\eta = \eta^t A\eta$ and $trace(A^*V) = trace(AV)$. And corresponding to (6.2) we have

$$(\mathbf{y}-\eta)^t A(\mathbf{y}-\eta) \sim \sum_1^n \alpha_i \chi_1^2 \tag{6.5}$$

where α_i are the eigenvalues of A^*. Finally, $\sum_1^n \alpha_i^2 = trace(A^{*2}) = trace(A^2) = \sum_1^n \lambda_i^2$. Hence (6.3) is true for non-symmetric matrices as well.

6.3. The Decrease in Residual Sum of Squares.

6.3.1. Linear Regression

Here we review a familiar set-up. Suppose we have a response vector $\mathbf{y}$ with $E\mathbf{y} = \mathbf{f}$ and $Var(\mathbf{y}) = I$. A matrix of covariate values X is available (assumed to have 1's in the first column) and we postulate two models for $E\mathbf{y}$: M_1 and M_2. The two models are such

that the linear space $\mathcal{L}_1$ specified by M_1 is a linear subspace of the space $\mathcal{L}_2$ specified by M_2. A example of this set-up is $M_1 : E\mathbf{y} = \alpha\mathbf{1}$ and $M_2 : E\mathbf{y} = \mathbf{X}\boldsymbol{\beta}$ respectively. Let $R(\mathbf{y}, \hat{\mathbf{y}}) = (\mathbf{y} - \hat{\mathbf{y}})^t(\mathbf{y} - \hat{\mathbf{y}})$ be the residual sum of squares associated with a fit $\hat{\mathbf{y}}$. We are interested in the following problem. If $\mathbf{f} \in \mathcal{L}_1$, what is the distribution of $R(\mathbf{y}, \hat{\mathbf{y}}_1) - R(\mathbf{y}, \hat{\mathbf{y}}_2)$?

Let P_1 and P_2 be the matrices that project onto the spaces $\mathcal{L}_1$ and $\mathcal{L}_2$ respectively.

In order to analyze $R(\mathbf{y}, \hat{\mathbf{y}}_1) - R(\mathbf{y}, \hat{\mathbf{y}}_2)$, we need the following pythagorean type relation:

$$R(\mathbf{y}, \hat{\mathbf{y}}_1) = R(\mathbf{y}, \mathbf{f}) - R(\hat{\mathbf{y}}_1, \mathbf{f}) \tag{6.6}$$

This is easily established by writing $(\mathbf{y} - \mathbf{f})^t(\mathbf{y} - \mathbf{f})$ as $(\mathbf{y} - \hat{\mathbf{y}}_1 + \hat{\mathbf{y}}_1 - \mathbf{f})^t(\mathbf{y} - \hat{\mathbf{y}}_1 + \hat{\mathbf{y}}_1 - \mathbf{f})$ and expanding. The corresponding result is also true for $R(\mathbf{y}, \hat{\mathbf{y}}_2)$ and combining these, we obtain

$$R(\mathbf{y}, \hat{\mathbf{y}}_1) - R(\mathbf{y}, \hat{\mathbf{y}}_2) = R(\hat{\mathbf{y}}_2, \mathbf{f}) - R(\hat{\mathbf{y}}_1, \mathbf{f}) \tag{6.7}$$

This has expected value

$$E(R(\mathbf{y}, \hat{\mathbf{y}}_1) - R(\mathbf{y}, \hat{\mathbf{y}}_2)) = trace(P_2) - trace(P_1) \tag{6.8}$$

If the rank of P_i is p_i, this is simply $p_2 - p_1$, a familiar result. Hence the expected decrease in residual sum of squares equals the excess number of parameters fit. Note also that $E(R(\hat{\mathbf{y}}_2, \mathbf{f})) = trace(Var(\hat{\mathbf{y}}_2)) = trace(P_2)$, and similarly for $\hat{\mathbf{y}}_1$. Hence the expected decrease in residual sum of squares also equals the increase in total variance of the fitted values.

The pythagorean relation (6.6) is a special case of an information decomposition valid in any exponential family. The general result is known as Simon's theorem (Simon (1973)). Note also that (6.6) holds for any $\mathbf{f} \in \mathcal{L}_1$, not just $\mathbf{f} = E\mathbf{y}$. We will use relations similar to (6.6) in analyzing the scatterplot and local likelihood procedures in the next sections.

Finally, if we assume further that $\mathbf{y} \sim \mathcal{N}(\mathbf{f}, I)$, then

$$R(\mathbf{y}, \hat{\mathbf{y}}_1) - R(\mathbf{y}, \hat{\mathbf{y}}_2) \sim \chi^2_{p_2 - p_1} \tag{6.9}$$

6.3.2. Scatterplot Smoothers

Consider now the case where we have a single covariate $\mathbf{x} = (x_1, x_2, \ldots x_n)$, and the global least squares fits are replaced by a scatterplot smoothing fits. We will restrict our attention to linear smoothers, so that the output $\hat{\mathbf{y}}$ can be written as $\hat{\mathbf{y}} = P\mathbf{y}$, where P is a "smoother matrix". An example of such a P, and the one we have in mind, is the matrix that produces local least squares fits, as discussed in Chapter 2. This matrix will depend on the set of covariate values $x_1, x_2, \ldots x_n$ and on the span of the smoother. Given $x_1, x_2, \ldots x_n$ and a smoothing algorithm, it is easy to produce P: the ith row of P is the output of the smoother applied to the ith unit vector. Such a P is not idempotent and hence not a projection matrix. We will call a matrix P producing local least squares fits a "local linear smoother matrix".

Given two smoothers P_1 and P_2, producing fit vectors $\hat{\mathbf{y}}_1$ and $\hat{\mathbf{y}}_2$, we ask the same question as in the previous section: what is the distribution of $R(\mathbf{y}, \hat{\mathbf{y}}_1) - R(\mathbf{y}, \hat{\mathbf{y}}_2)$? Here we are thinking of a situation in which the smoother P_2 is more complex than the smoother P_1. For example, we might have $P_1\mathbf{y} = \bar{y}\mathbf{1}$ and $P_2\mathbf{y} = \text{smooth of } \mathbf{y}$. Then the quantity $R(\mathbf{y}, \hat{\mathbf{y}}_1) - R(\mathbf{y}, \hat{\mathbf{y}}_2)$ would be of interest in assessing the importance of the smooth P_2. In the previous section, we assumed that the smaller model was correct, i.e. $\mathbf{f} \in \mathcal{L}_1$. Here we have not assumed any "models"; the appropriate assumption is: $E\hat{\mathbf{y}}_1 \approx E\hat{\mathbf{y}}_2$, so that $P_1\mathbf{f} = P_2\mathbf{f}$. This says that the smoother P_2 produces the same fit on the average as P_1.

First, we require a pythagorean relation like (6.6) . Letting, $\mathbf{h} = P_1\mathbf{f} = P_2\mathbf{f}$, it is easy to show that

$$E(R(\mathbf{y}, \hat{\mathbf{y}}_1)) = E(R(\mathbf{y}, \mathbf{h})) - E(R(\hat{\mathbf{y}}_1, \mathbf{h})) \tag{6.10}$$

$$E(R(\mathbf{y}, \hat{\mathbf{y}}_2)) = E(R(\mathbf{y}, \mathbf{h})) - E(R(\hat{\mathbf{y}}_2, \mathbf{h})) \tag{6.11}$$

the cross-product term in each case being $E(\mathbf{y}(I - P_i)^t(P_i\mathbf{y} - \mathbf{h}) = \mathbf{f}^t(I - P_i)^tP_i\mathbf{f} - \mathbf{f}^t(I - P_i)\mathbf{h} = 0$. Combining (6.10) and (6.11) , and using the fact that for local linear smoother matrices P, $trace(P^tP) = trace(P)$, we obtain

$$\begin{aligned} E(R(\mathbf{y}, \hat{\mathbf{y}}_1) - R(\mathbf{y}, \hat{\mathbf{y}}_2)) &= E(R(\hat{\mathbf{y}}_2, \mathbf{h})) - E(R(\hat{\mathbf{y}}_1, \mathbf{h})) \\ &= trace(P_2) - trace(P_1) \end{aligned} \tag{6.12}$$

the same as in the least squares setting.

If we assume that $\mathbf{y} \sim \mathcal{N}(\mathbf{f}, I)$, we can find (approximately) the distribution of $R(\mathbf{y}, \hat{\mathbf{y}}_1) - R(\mathbf{y}, \hat{\mathbf{y}}_2)$. It's easier to work directly with $R(\mathbf{y}, \hat{\mathbf{y}}_1) - R(\mathbf{y}, \hat{\mathbf{y}}_2)$ as follows:

$$\begin{aligned} R(\mathbf{y}, \hat{\mathbf{y}}_1) - R(\mathbf{y}, \hat{\mathbf{y}}_2) &= \mathbf{y}^t(I - P_1)^t(I - P_1)\mathbf{y} - \mathbf{y}^t(I - P_2)^t(I - P_2)\mathbf{y} \\ &= (\mathbf{y} - \mathbf{f})^t A(\mathbf{y} - \mathbf{f}) + 2(\mathbf{f} - \mathbf{h})^t(P_1 - P_2)\mathbf{y} \end{aligned} \tag{6.13}$$

where $A = \{P_1^t P_1 - P_1^t - P_1 - (P_2^t P_2 - P_2^t - P_2)\}$. If we ignore the second term in (6.13) (it is zero in expectation) and let $\{\lambda_i\}$ and $\{\alpha_i\}$ be the eigenvalues of A and $\frac{1}{2}(A + A^t)$ respectively, then

$$R(\mathbf{y}, \hat{\mathbf{y}}_1) - R(\mathbf{y}, \hat{\mathbf{y}}_2) \sim \sum_1^n \alpha_i \chi_1^2 \ \textit{(approximately)} \tag{6.14}$$

and

$$Var(R(\mathbf{y}, \hat{\mathbf{y}}_1) - R(\mathbf{y}, \hat{\mathbf{y}}_2)) \approx 2\sum_1^n \lambda_i^2 \tag{6.15}$$

6.4. The Decrease in Deviance.

In the previous section, we derived the exact result $E(R(\mathbf{y}, \hat{\mathbf{y}}_1) - R(\mathbf{y}, \hat{\mathbf{y}}_2)) = trace(P_2) - trace(P_1)$ for local linear scatterplot smoothing. In this section, we show that an analogous result is approximately true for local likelihood procedures in the exponential family. In this more general setting, the deviance takes the place of the residual sum of squares. The scatterplot smoothing case, of course, corresponds to local likelihood fitting with a Gaussian likelihood, and the residual sum of squares is the deviance for a Gaussian Likelihood.

In order to derive a deviance approximation, we will first obtain a relation similar to (6.10) and (6.11) for exponential families.

6.4.1. Pythagorean Relations for the Deviance

We assume that the Y_i's are independent with density of the exponential family form

$$g_{\theta_i}(y_i) = \exp\{\{y_i\theta_i - b(\theta_i) - c(y_i, \sigma)\}/\sigma^2\} \tag{6.16}$$

with respect to some carrier measure. The scale parameter σ plays no special role and is assumed to be 1.

Let $k_{\boldsymbol{\theta}}(y) = \prod_1^n g_{\theta_i}(y_i)$, and let $b(\boldsymbol{\theta}) = (b(\theta_1), b(\theta_2), \ldots b(\theta_n))$. The density $k_{\boldsymbol{\theta}}(\mathbf{y})$ can be indexed by the natural parameter $\boldsymbol{\theta}$ or the expectation parameter $\boldsymbol{\mu} = E_{\boldsymbol{\theta}}\mathbf{y} = b'(\boldsymbol{\theta})$. We will write $b'^{-1}(\boldsymbol{\mu})$ as $\boldsymbol{\theta}_{\boldsymbol{\mu}}$, and let $\Sigma_{\boldsymbol{\mu}}$ be the (diagonal) covariance matrix of the Y_i's. The quantity (twice) Kullback-Leibler distance between $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$ is defined by

$$I(\boldsymbol{\mu}_1, \boldsymbol{\mu}_2) = 2E_{\boldsymbol{\mu}_1} \log \frac{k_{\boldsymbol{\mu}_1}(\mathbf{y})}{k_{\boldsymbol{\mu}_2}(\mathbf{y})} \tag{6.17}$$

We will call this the "deviance"— a short aside will clarify why we are allowed to do this. The deviance is defined in the generalized linear model literature as $D(\mathbf{y}, \boldsymbol{\mu}) = -2\log[k_{\boldsymbol{\mu}}(\mathbf{y})/k_{\mathbf{y}}(\mathbf{y})]$. Hoeffding's theorem (see Efron (1977)) states that in the exponential family $I(\mathbf{y}, \boldsymbol{\mu}) = D(\mathbf{y}, \boldsymbol{\mu})$. Note that $I(\boldsymbol{\mu}_1, \boldsymbol{\mu}_2)$ is not in general equal to $D(\boldsymbol{\mu}_1, \boldsymbol{\mu}_2)$. They are equal when $\boldsymbol{\mu}_1 = \mathbf{y}$ (as above) and also when $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$ represent nested fits (see Simon's theorem in Efron (1977)). In other cases, $D(\boldsymbol{\mu}_1, \boldsymbol{\mu}_2)$ is an estimate of $I(\boldsymbol{\mu}_1, \boldsymbol{\mu}_2)$. Since the result that we seek to establish involves quantities of the form $I(\mathbf{y}, \boldsymbol{\mu})$ in the exponential family, we shall use the term "deviance".

Let $\hat{\mathbf{y}}$ be some fit vector and let $\mathbf{h}$ be such that $E(\boldsymbol{\theta}_{\hat{\mathbf{y}}}) = \boldsymbol{\theta}_{\mathbf{h}}$. Then we have

$$I(\mathbf{y}, \mathbf{h}) = 2(\boldsymbol{\theta}_{\mathbf{y}} - \boldsymbol{\theta}_{\mathbf{h}})^t\mathbf{y} - 2\sum_1^n (b(\theta_{y_i}) - b(\theta_{h_i})) \tag{6.18}$$

$$I(\hat{\mathbf{y}}, \mathbf{h}) = 2(\boldsymbol{\theta}_{\hat{\mathbf{y}}} - \boldsymbol{\theta}_{\mathbf{h}})^t\hat{\mathbf{y}} - 2\sum_1^n (b(\theta_{\hat{y}_i}) - b(\theta_{h_i})) \tag{6.19}$$

and

$$I(\mathbf{y}, \hat{\mathbf{y}}) = 2(\boldsymbol{\theta}_{\mathbf{y}} - \boldsymbol{\theta}_{\hat{\mathbf{y}}})^t\mathbf{y} - 2\sum_1^n (b(\theta_{y_i}) - b(\theta_{\hat{y}_i})). \tag{6.20}$$

Hence

$$I(\mathbf{y}, \mathbf{h}) = I(\mathbf{y}, \hat{\mathbf{y}}) + I(\hat{\mathbf{y}}, \mathbf{h}) + \Delta \tag{6.21}$$

where $\Delta = 2(\boldsymbol{\theta}_{\hat{\mathbf{y}}} - \boldsymbol{\theta}_{\mathbf{h}})^t(\mathbf{y} - \hat{\mathbf{y}})$.

What can we say about Δ? We'll examine the global and local cases separately.

Δ for the Global (linear) model

If $\boldsymbol{\theta}$ is modelled as $\boldsymbol{\theta} = X\beta$, $\boldsymbol{\theta}_{\hat{\boldsymbol{h}}} \subset \mathcal{L}_{col}(X)$ (say $\boldsymbol{\theta}_{\hat{\boldsymbol{h}}} = X\beta_0$) and $\hat{\boldsymbol{y}}$ is the m.l.e., then the score equation is

$$X^t(\boldsymbol{y} - \hat{\boldsymbol{y}}) = 0 \tag{6.22}$$

Hence

$$\begin{aligned}\Delta =& 2(\boldsymbol{\theta}_{\hat{\boldsymbol{y}}} - \boldsymbol{\theta}_{\hat{\boldsymbol{h}}})^t(\boldsymbol{y} - \hat{\boldsymbol{y}}) \\ =& 2(\hat{\beta} - \beta_0)^t X^t(\boldsymbol{y} - \hat{\boldsymbol{y}}) \\ =& 0\end{aligned} \tag{6.23}$$

Thus for linear models, the pythagorean relationship

$$I(\boldsymbol{y}, \hat{\boldsymbol{h}}) = I(\boldsymbol{y}, \hat{\boldsymbol{y}}) + I(\hat{\boldsymbol{y}}, \hat{\boldsymbol{h}}) \tag{6.24}$$

holds exactly. As mentioned in the previous section, this result is a special case of Simon's theorem (see also Efron (1975)).

Δ for the Local Likelihood Model

If instead $\boldsymbol{\theta}$ is estimated by local likelihood, we no longer have $\Delta = 0$. We assume that as before that $E(\boldsymbol{\theta}_{\hat{\boldsymbol{y}}}) = \boldsymbol{\theta}_{\hat{\boldsymbol{h}}}$. Consider first the Gaussian case, for which $\mu = \boldsymbol{\theta}$. Then $\hat{\boldsymbol{y}} = P\boldsymbol{y}$ where P is a local linear smoother matrix, and

$$\Delta = 2(P\boldsymbol{y} - \hat{\boldsymbol{h}})^t(\boldsymbol{y} - P\boldsymbol{y}) \tag{6.25}$$

This is exactly the cross product term discussed below (6.11) , which we proved has expectation 0. Since in the Gaussian case, $I(\mu_1, \mu_2) = R(\mu_1, \mu_2)$, this re-verifies (6.10) and (6.11) .

When the link function $b(\boldsymbol{\theta})$ is non-linear, $E(\Delta) \neq 0$. But we can see that it will be small, for the following reason. The local score equation to determine $\hat{\boldsymbol{y}}_i$ is $X_i^t(\boldsymbol{y}_i - \hat{\boldsymbol{y}}_i) = 0$ where X_i is the design matrix for the *ith* neighborhood and $\boldsymbol{y}_i$ and $\hat{\boldsymbol{y}}_i$ are the response and fit vectors for the *ith* neighborhood. We also have the local approximations $\boldsymbol{\theta}_{\hat{\boldsymbol{y}}} = X_i\beta \approx \boldsymbol{\theta}_{\hat{\boldsymbol{h}}} = X_i\beta_0$ (say). Hence each element of $\Delta = 2(\boldsymbol{\theta}_{\hat{\boldsymbol{y}}} - \boldsymbol{\theta}_{\hat{\boldsymbol{h}}})^t(\boldsymbol{y} - \hat{\boldsymbol{y}})$ is approximately 0. In the deviance approximations of the next section, we will therefore assume $E(\Delta) = 0$.

6.4.2. The Deviance Approximations

The Global (Linear) Model

We postulate two nested linear models for $\boldsymbol{\theta}$, $M_1 : \boldsymbol{\theta} = X_1\beta$ and $M_2 : \boldsymbol{\theta} = X_2\beta$, with $\mathcal{L}_{col}(X_1) \in \mathcal{L}_{col}(X_2)$. Let $\hat{\boldsymbol{y}}_1$ and $\hat{\boldsymbol{y}}_2$ be the m.l.e.'s under M_1 and M_2 and let $\boldsymbol{\theta}_{\hat{\boldsymbol{h}}} \in \mathcal{L}_{col}(X_1)$. Let $\text{rank}(X_1) = p_1$, $\text{rank}(X_2) = p_2$.

Consider the difference in deviance between the two fitted models $I(\boldsymbol{y}, \hat{\boldsymbol{y}}_1) - I(\boldsymbol{y}, \hat{\boldsymbol{y}}_2)$ Using the result of the previous section, we have

$$\begin{aligned}I(\boldsymbol{y}, \hat{\boldsymbol{y}}_1) - I(\boldsymbol{y}, \hat{\boldsymbol{y}}_2) =& I(\boldsymbol{y}, \hat{\boldsymbol{h}}) - I(\hat{\boldsymbol{y}}_1, \hat{\boldsymbol{h}}) - [(I(\boldsymbol{y}, \hat{\boldsymbol{h}}) - I(\hat{\boldsymbol{y}}_2, \hat{\boldsymbol{h}})] \\ =& I(\hat{\boldsymbol{y}}_2, \hat{\boldsymbol{h}}) - I(\hat{\boldsymbol{y}}_1, \hat{\boldsymbol{h}})\end{aligned} \tag{6.26}$$

A taylor series expansion gives

$$I(\hat{\boldsymbol{y}}_2, \hat{\boldsymbol{h}}) \approx (\hat{\boldsymbol{y}}_2 - \hat{\boldsymbol{h}})\Sigma_{\hat{\boldsymbol{h}}}^{-1}(\hat{\boldsymbol{y}}_2 - \hat{\boldsymbol{h}}) \tag{6.27}$$

Also

$$Var(X_2\hat{\beta}) \approx X_2(X_2^t\Sigma_{\hat{\boldsymbol{h}}}X_2)^{-1}X_2^t \tag{6.28}$$

so that

$$Var(\hat{\boldsymbol{y}}_2) \approx \Sigma_{\hat{\boldsymbol{h}}}X_2(X_2^t\Sigma_{\hat{\boldsymbol{h}}}X_2)^{-1}X_2^t\Sigma_{\hat{\boldsymbol{h}}} \tag{6.29}$$

Thus

$$\begin{aligned}E(I(\hat{\boldsymbol{y}}_2, \hat{\boldsymbol{h}})) \approx& E(\hat{\boldsymbol{y}}_2 - \hat{\boldsymbol{h}})\Sigma_{\hat{\boldsymbol{h}}}^{-1}(\hat{\boldsymbol{y}}_2 - \hat{\boldsymbol{h}}) \\ =& trace(\Sigma_{\hat{\boldsymbol{h}}}^{-1}\Sigma_{\hat{\boldsymbol{h}}}X_2(X_2^t\Sigma_{\hat{\boldsymbol{h}}}X_2)^{-1}X_2^t\Sigma_{\hat{\boldsymbol{h}}}) \\ =& trace(I_{p_2}) = p_2\end{aligned} \tag{6.30}$$

In exactly the same way we get $E(I(\hat{\boldsymbol{y}}_1, \hat{\boldsymbol{h}})) \approx p_1$ and hence

$$E(I(\boldsymbol{y}, \hat{\boldsymbol{y}}_1) - I(\boldsymbol{y}, \hat{\boldsymbol{y}}_2)) \approx p_2 - p_1 \tag{6.31}$$

This is not surprising, of course, since Wald's theorem tells us that $I(\boldsymbol{y}, \hat{\boldsymbol{y}}_1) - I(\boldsymbol{y}, \hat{\boldsymbol{y}}_2) \rightarrow \chi^2_{p_2-p_1}$.

The Local Likelihood Model

Now consider the case where $\hat{\mathbf{y}}_1$ and $\hat{\mathbf{y}}_2$ represent local likelihood fits. We have a single covariate $x_1, x_2, \ldots x_n$, and let $\mathbf{x}_i = (1\ x_i)^t$. We assume $E(\boldsymbol{\theta}_{\hat{\mathbf{y}}_1}) = E(\boldsymbol{\theta}_{\hat{\mathbf{y}}_2}) = \boldsymbol{\theta}_h$. Letting $v_i = Var(\text{ith element of } \hat{\mathbf{y}}_2)$ and $\sigma_i = ii\text{th entry of } \Sigma_h$, we have

$$E(I(\hat{\mathbf{y}}_2, h)) \approx E(\hat{\mathbf{y}}_2 - h)\Sigma_h^{-1}(\hat{\mathbf{y}}_2 - h) = \sum_1^n v_i\sigma_i^{-1} \tag{6.32}$$

using the fact that $E(\hat{\mathbf{y}}_2) \approx h$. Now

$$v_i \approx \sigma_i X_i(X_i^t\Sigma_h^i X_i)^{-1}X_i^t\sigma_i \tag{6.33}$$

where X_i and Σ_h^i are the design and covariance matrices for the ith neighborhood. Hence

$$\begin{aligned} E(I(\hat{\mathbf{y}}_2, h)) &\approx \sum_1^n \sigma_i\mathbf{x}_i(X_i^t\Sigma_h^i X_i)^{-1}\mathbf{x}_i^t\sigma_i\sigma_i^{-1} \\ &= \sum_1^n \sigma_i\mathbf{x}_i(X_i^t\Sigma_h^i X_i)^{-1}\mathbf{x}_i^t \end{aligned} \tag{6.34}$$

Now in a given neighborhood, σ_i can be taken as approximately constant, so we have

$$E(I(\hat{\mathbf{y}}_2, h)) \approx \sum_1^n \mathbf{x}_i(X_i^tX_i)^{-1}\mathbf{x}_i^t = trace(P_2) \tag{6.35}$$

Similarly,

$$E(I(\hat{\mathbf{y}}_1, h)) \approx trace(P_1) \tag{6.36}$$

Hence for local likelihood smoothers we have

$$E(I(\mathbf{y}, \hat{\mathbf{y}}_1) - I(\mathbf{y}, \hat{\mathbf{y}}_2)) \approx trace(P_2) - trace(P_1) \tag{6.37}$$

In the preceding "derivation", we have made a number of approximations, and it's important to find out how accurate the formula $trace(P_2) - trace(P_1)$ really is. In the next section, we describe a simulation study to investigate this.

Finally, we note that the actual distribution of the decrease is more difficult to obtain. Even in the simple scatterplot case, we have seen that this distribution is NOT chi-squared but a weighted linear combination of χ_1^2's. In the general local likelihood case, we have not succeeded in obtaining a workable approximation for this distribution. The simulations

of the next section show that, at least for small samples, the distribution of the deviance decrease is quite a bit more spread out that the corresponding chi-squared distribution.

6.5. Degrees of Freedom Simulations.

Table 1 shows the results of a modest simulation study designed to check the accuracy of the formula $E(I(\mathbf{y}, \hat{\mathbf{y}}_1) - I(\mathbf{y}, \hat{\mathbf{y}}_2)) = trace(P_2) - trace(P_1)$.

Table 1. Results of Degrees of Freedom Simulation

Entries are *mean(variance) of deviance decrease*

	Span				
Source	.3	.4	.5	.6	.7
(1) $Trace(P) - 1$	4.09(10.14)	3.32(8.07)	2.65(6.15)	2.34(5.09)	2.16(4.27)
(2) *Scatterplot Smooth(y normal)*	4.14(10.00)	3.39(7.75)	2.61(6.03)	2.31(5.08)	2.09(4.32)
(3) *Scatterplot Smooth(y uniform)*	4.19(10.06)	3.45(8.50)	2.77(6.52)	2.41(5.79)	2.21(4.99)
(4) *Logistic Model(constant vs smooth)*	4.34(13.47)	3.40(11.62)	2.72(9.12)	2.28(7.51)	2.17(6.28)
(5) *Logistic Model(linear vs smooth)*	3.29(11.71)	2.25(8.25)	1.63(6.21)	1.29(4.58)	1.12(2.89)
(6) *Cox Model(no censoring)*	5.58(13.37)	4.24(8.99)	3.63(7.52)	3.12(6.25)	2.71(5.48)
(7) *Cox Model(40% censoring)*	5.36(13.54)	4.16(9.04)	3.62(6.98)	3.13(5.86)	2.73(5.20)

The numbers in the table were obtained as follows. 100 x values were generated from $N(0,1)$ and fixed for the entire table. Given these x values, we constructed the local linear smoother matrices for the indicated spans, and the trace of each matrix (minus 1) is shown in line (1). The numbers in parentheses are variance estimates based on formula (6.15) .

Consider for example the entry 4.09 in the top left hand corner. According to the preceding derivation, this should be an estimate of the expected decrease in deviance due to fitting a local likelihood model with that span .3 versus a model with only a constant.

To obtain line (2), we generated 100 y_i's from $N(0,1)$ and computed $R(\mathbf{y}, \bar{y}1) - R(\mathbf{y}, \hat{\mathbf{y}})$, $\hat{\mathbf{y}}$ being the fit from a scatterplot smoother ($\hat{\mathbf{y}} = P\mathbf{y}$) with span as shown. Line(2) shows the mean and variance from 500 such repetitions of this process.

Line (3) was obtained in same way as line (2), except that the y_i's were generated from uniform $(-\sqrt{3}, \sqrt{3})$, the range chosen so that $Var(y_i) = 1$.

To obtain line(4), we generated 100 y_i's from $binomial(1, 1/2)$ and fit a smooth logistic model with spans of .3 to .7. The numbers show the mean and variance of $I(\mathbf{y}, \hat{y}1) - I(\mathbf{y}, \hat{\mathbf{y}})$ over 500 repetitions.

Line (5) was generated in a similar fashion as line (4), showing instead the mean and variance of $I(\mathbf{y}, \hat{\mathbf{y}}_l) - I(\mathbf{y}, \hat{\mathbf{y}})$, $\hat{\mathbf{y}}_l$ being the linear logistic fit, with y_i generated from a linear logistic model, $P(y_i = 1 \,|\, x) = e^{.2x}/(1 + e^{.2x})$.

Lines (6) and (7) show simulation results for the Cox model. 100 y values were generated according to $y = exp(1 + \epsilon)$, where ϵ had an extreme value distribution. This corresponds to a constant hazard (exponential) model. For line (6), no censoring was applied. For line (7), censoring variables c_i were generated from e^{6u}, $u \sim \mathcal{U}(0, 1)$. This produced a censoring rate of about 40%. A smooth Cox model was fit and the quantity $-2 \log L(null\ model) - (-2 \log L(smooth))$ was computed. Lines (6) and (7) show the mean and variance of this quantity over 500 repetitions.

The results give fairly strong support to the approximation $E(I(\mathbf{y}, \hat{\mathbf{y}}_1) - I(\mathbf{y}, \hat{\mathbf{y}}_2)) = trace(P_2) - trace(P_1)$. Lines (2) and (3) agree well with (1), not surprising since the approximation is exact for scatterplot smoothers. Line (4) also is in good agreement, with a small upward bias for smaller spans. Line (5) should be 1 less than line (1), (since the global linear fit uses 2 degrees of freedom) and the results indicate that. In examining the Cox results, we must remember that there is no constant in the model, so lines (6) and (7) should be 1 greater than line (1). This is roughly the case, with a downward trend in the higher spans.

The variance results are a little unsettling. In general, the variances will depend on the higher (> 2) moments of the distribution of y; the variance in line (1) was derived assuming y_i was $\mathcal{N}(0, 1)$. The variances in line (2) of course are in agreement with line (1), but those in line (3) and especially line (4) are higher. The variances in lines (4)— (6) are not comparable to line (1), since they are based on different model comparisons. We can see, however, that the variance to mean ratio is often greater than 2 (the ratio for a chi-square variate).

We conclude from these simulations that the approximation $E(I(\mathbf{y}, \hat{\mathbf{y}}_1) - I(\mathbf{y}, \hat{\mathbf{y}}_2)) = trace(P_2) - trace(P_1)$ is satisfactory as a rough rule of thumb. We do note, however, that the distribution of this decrease is more spread out than a chi-square variate with the corresponding degrees of freedom, so that tests based on the percentile points will be too liberal.

Finally, it is important to mention that the above simulations were relatively inexpensive on a large computer. Hence for a given data set it may be feasible to get "exactly" the distribution of the decrease by simulation.

6.6. Akaike's Information Criterion(AIC) For Span Selection.

Using the results of the previous section, we show in this section that it's reasonable to use an AIC criterion to choose the span in the local likelihood estimation procedure.

Let's briefly review the AIC for a parametric model. Given a model k_μ, suppose we can choose among maximum likelihood estimates $\hat{\mu}_1, \hat{\mu}_2, \ldots \hat{\mu}_l$ based on $p_1, p_2, \ldots p_k$ degrees of freedom respectively. Suppose also that each model can be considered a sub-model of a true model k_{μ_0}. Then Akaike's information criterion (AIC) (Akaike 1973) specifies that we should choose the model that minimizes

$$AIC = -2 \log k_{\hat{\mu}_i}(\hat{\mathbf{y}}) + 2p_i \tag{6.38}$$

where $\log k_{\hat{\mu}_i}(\hat{\mathbf{y}})$ is the value of the likelihood at $\hat{\mu}_i$.

Akaike derived the AIC by showing that $E(AIC) \approx E(I(\mu_0, \hat{\mu}_i)) + constant$. Hence the model that minimizes AIC approximately minimizes the expected Kullback-Leibler distance from the true model.

From the form of the AIC, it is clear that it attempts to trade-off goodness of fit of the model with model complexity. Not surprisingly, it turns out to be identical to Mallow's C_p in the linear regression setting and asymptotically equivalent to the cross-validated likelihood technique in general (see Stone (1977) for these results).

In the local likelihood procedure, we propose choice of the span parameter s to minimize

$$AIC = -2\log k_{\hat{y}(w)}(\hat{y}) + 2\ trace\ P(w)) \tag{6.39}$$

where $P(w)$ denotes the smoother matrix producing local linear fits with span w, and $\hat{y}(w)$ denotes the corresponding fitted values. This makes sense intuitively: as the span w increases, $-2\log k_{\hat{y}(w)}(\hat{y})$ will increase but the degrees of freedom $trace\ P(w)$ will decrease. Hence the AIC will trade off lack of fit with complexity of the smooth.

In what follows, we will show that the AIC is reasonable in the local likelihood setting, in that it approximately equals a measure of expected distance to the true model. The logic of the derivation follows that of Akaike (1973). Consider the exponential family set-up of section 6.4.1. Using the notation of that section, we let P be a local linear smoother matrix corresponding to some span and $\hat{y}$ be the estimated fit vector (dropping the argument (w) for convenience). Let h be such that $E(\theta_{\hat{y}}) = \theta_h$, and further let μ_0 be the "true model" in that $Ey = \mu_0$. We require an estimate of the Kullback-Leibler distance $I(\mu_0, \hat{y})$. The following pythagorean relation is easy to derive from the results of section 6.4.1

$$E(I(\mu_0, \hat{y})) = E(I(\mu_0, h)) + E(I(h, \hat{y})), \tag{6.40}$$

the cross-product term being $(\theta_h - \theta_{\hat{y}})'(\mu_0 - h)$, with expectation 0. The second term is the sum of the variance of the $\hat{y}_i$'s and equals approximately $trace(P)$. To get an estimate of the first term, we expand $E(I(y, \hat{y}))$ as

$$\begin{aligned} E(I(y, \hat{y})) &= E(I(y, h)) - E(I(\hat{y}, h)) \\ &= [E(I(y, \mu_0) + E(I(\mu_0, h)] - E(I(\hat{y}, h)) \end{aligned} \tag{6.41}$$

Thus $E(I(y, \hat{y})) \approx (n + E(I(\mu_0, h))] - trace(P)$, or $E(I(\mu_0, h)) \approx E(I(y, \hat{y})) - n + trace(P)$. Combining this with (6.14) we obtain

$$E(I(\mu_0, \hat{y})) \approx E(I(y, \hat{y})) - n + 2\ trace(P) \tag{6.42}$$

Noting that $I(y, \hat{y})) = -2\log k_{\hat{y}(w)}(\hat{y}) + constant$ is constant for all spans, we arrive at the AIC criterion (6.39) .

Finally, we display in Figure 6.1 the idea behind this derivation. The distance $I(\mu_0, \hat{y})$ is estimated by $I(\hat{y}, h)$ plus an estimate of $I(\mu_0, h)$ derived from $I(y, \hat{y})$.

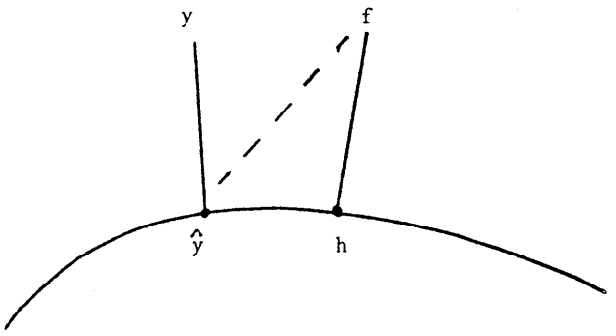


Figure 6.1. The AIC Picture

Chapter 7

Closing Remarks

In this dissertation, we have introduced and studied a non-parametric procedure that generalizes likelihood-based regression models. This procedure is potentially useful as a tool for exploratory data analysis as well as for building non-parametric regression models.

The algorithms are computationally expensive but with the tremendous increase in the speed of computers, this should become less and less of a problem. In a few years, in fact, the local likelihood procedure could very well run comfortably on a personal desktop computer.

The local likelihood approach is quite different from "quasi-linear" methods like partial residuals. These latter methods start with a linear model, then look for systematic deviations from it. The local likelihood approach, on the other hand, is fully non-parametric; in a sense, it abandons the linear model completely. Further research and experience will determine under what circumstances each of the approaches is more effective.

We plan to continue this research in a number of other areas. Graphical display is one such area. The local likelihood procedure could form part of a motion graphics package to display and analyze multivariate data. This could be especially useful for binary or censored data. Another area of interest is the application of the bootstrap to the procedure. The problem of obtaining confidence bands is closely tied in with this.

Finally, and perhaps most importantly, we plan to use the the local likelihood procedure in our data analyses. This should help point to ways in which the procedure can be improved, and ultimately, determine its value.

REFERENCES

Akaike, H. (1973) Information theory and an extension of the entropy maximization principle. 2nd International Symposium on information theory, pp 267-281.

Anderson, P. and Gill, R. (1982) Cox's regression model for counting processes: a large sample study. Ann. Stat. 10, 4, 1100-1120.

Breiman, L. and Friedman J.H. (1982). Estimating optimal correlations for multiple regression and correlation. Stanford U. tech. rep. Orion 010.

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

Cain, K. and Lange, N. (1984) Estimating case influence for proportional hazards regression models with censoring. Tech rep 320Z, Dept. of Biostatistics,Dana-Farber Cancer Institute.

Cleveland, W.S. (1979). Robust locally weighted regression and smoothing scatterplots. J. Amer. Statist. Assoc., 74 828-836.

Cox, D.R. (1972). Regression models and life tables. J. Roy. Stat. Soc. B, 34, 187-202.

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

Crowley, J. and Hu, M.(1977). Covariance Analysis of heart transplant survival data. J. Amer. Statist. Assoc. 72, 27-36.

Efron, B. (1975). The Geometry of Exponential Families. Ann. Stat. 6, No 2, 362-376

Efron, B. (1977). The Efficiency of Cox's likelihood function for censored data. J. Amer. Statist. Assoc. 72, 557-565.

Efron, B. (1979). Bootstrap Methods: another look at the Jackknife. Ann. Stat 7, pp 1-26.

Efron, B. (1980) Censored data and the bootstrap. J. Amer. Stat. Assoc. 76, 312-19.

Friedman, J.H., and Stuetzle, W. (1981). Projection pursuit regression. J. Amer. Statist. Assoc. 76, 817-823.

Friedman, J.H., and Stuetzle, W. (1982). Smoothing of scatterplots. Stanford Univ. technical report - Orion 003.

Friedman, J.H., and Owen, A. Predictive ACE. In preparation.

Guttman, I. (1983) Linear Models. Wiley, New York.

Haberman, S. (1976) Generalized Residuals for Log Linear Models. Proc. of 9th Int. Biost. Conf, Boston 104-122.

Hastie, T. (1983). Non-parametric logistic regression. Stanford University Technical report ORION 016.

Hastie, T. (1984) Discussion of "Graphical Methods for assessing logistic regression models", by Landwehr et al, J. Amer. Stat. Assoc. 79, 61-63.

Hastie, T., and Tibshirani, R. (1984). Generalized Additive Models. Deparment of Statistics, Stanford University Tech. rep 002.

Kalbfleisch, J.D., and Prentice, R.L. (1980). The Statistical analysis of failure time data. Wiley, New York.

Kay, R. (1977). Proportional hazard models and the analysis of censored survival data. J. Roy. Stat. Soc. C., 26, 227-237.

Krasker, W. and Welsch, R. (1982) Efficient bounded influence regression using alternate definitions of sensitivity. J. Amer. Stat. Assoc. 77, 595-605

Landwehr, J., Pregibon, D. and Shoemaker, A. (1984) Graphical methods for assessing logistic regression models. J. Amer. Stat. Assoc., 79, 61-63.

McCullagh, P. (1983). Quasi Likelihood Functions. Ann. Stat. 11, 59-67.

Miller, R.G. and Halpern, J. (1982). Regression with censored data. Biometrika 69, 3, 521-31.

Nelder, J.A. and Wedderburn, R.W.M. (1972). Generalized Linear Models. J. Roy. Stat. Soc. A, 135, 370-384.

Oakes, D. (1977). The asymptotic information in censored survival data. Biometrika 674,

441-448.

Peto, R. (1972). Discussion on Professor Cox's paper. J. Roy. Stat. Soc. B, 34, 205-207.

Prentice, R. and Breslow, N. (1978). Retrospective Studies and Failure Time Models. Biometrika 65, 153-158.

Simon, G. (1973). Additivity of information in exponential family probability laws. J. Amer. Stat. Assoc. 68, 478-482.

Stone, M. (1977) As asymptotic choice of model by cross-validation and Akaike's criterion. J. Roy Stat. Soc B., No 1, vol 7, 44-47.

Thomas, D. (1983) Non-parametric estimation and tests of fit for dose response relations. Biometrics, Vol 39, No 1, 263-268.

Tsiatis, A. (1981). A large sample study of Cox's regression model. Vol 9, No 1, 93-108.