SLAC PUB-3160
June 1983
(M)

NON-PARAMETRIC LOGISTIC REGRESSION

Trevor J. Hastie
Computation Research Group
Stanford Linear Accelerator Center
and Department of Statistics
Stanford University

Abstract

Linear logistic regression models the expectation of a dichotomous response variable with the model $\ln(p(\mathbf{x})/(1-p(\mathbf{x}))) = \mathbf{x}'\mathbf{a}$. Often the assumption of linearity is violated, and alternative forms are sought. In this article we deal with models of the form $\ln(p(\mathbf{x})/(1-p(\mathbf{x}))) = \sum \phi_i(x_i)$ where the ϕ_i are general smooth functions of the explanatory variables. Estimation is achieved using local maximum likelihood. The technique is illustrated with two examples, and is compared to existing techniques such as partial residual plots.

Keywords and phrases: Logistic regression, smooth, non-linear, non-parametric, local likelihood.

(Submitted to Applied Statistics)

This work was 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.

1. INTRODUCTION.

An important statistical problem is that of regressing a binary response variable on a set of predictor variables. This has a special application in medical diagnosis problems and risk analysis. An example treated later deals with the survival of a patient after surgery for breast cancer. The response variable y is coded 1 if the patient survived after a specified period, else 0. For each patient we also have a vector $\mathbf{x}$ of explanatory or predictor variables such as age, year of operation and possibly some prognostic factors. We have a sample of such patients which we use to model the probability of the binary response as a function of the explanatory variables. We can also invisage using the resultant model to predict the response of future cases for which the response is not yet known.

Specifically, we wish to estimate $p(\mathbf{x}) \equiv p(y = 1 \mid \mathbf{x}) = 1 - p(y = 0 \mid \mathbf{x})$ for any vector $\mathbf{x}$. A standard approach to the problem is the linear logistic regression model

$$p(\mathbf{x}) = \frac{e^{\mathbf{x}'\mathbf{a}}}{1 + e^{\mathbf{x}'\mathbf{a}}}$$

or

$$\begin{aligned} \ln \frac{p(\mathbf{x})}{1 - p(\mathbf{x})} &= \text{logit } p(\mathbf{x}) \\ &= \mathbf{x}'\mathbf{a} \end{aligned} \tag{1.1}$$

In words this says that the log-odds of surviving are linear in the predictor variables. The unknown parameters $\mathbf{a}$ can be found by maximum likelihood and tests of significance can be based on the likelihood ratio statistic (Cox, (1970)). The procedure is well implemented in the package GLIM (Baker and Nelder, (1978)). Another approach is to make the assumption that the predictor variables are jointly normal with the same covariance matrix in each group (y=1 or y=0) but with different mean vectors. For such a model the log-odds are once again linear in $\mathbf{x}$ and the parameters are functions of the parameters of the normal distributions. This is known as Fisher's linear discriminant function. (Lachenbruch, (1975)).

The logit form in (1.1) guarantees that the estimated probabilities are positive and in the interval [0,1]; It is also the form of the natural parameter for the Binomial distribution in the Exponential family.

An often unjustified and misleading assumption is that logit $p(\mathbf{x})$ is *linear* in $\mathbf{x}$. The effect of a predictor may be felt only for a portion of its range. e.g. The effect of age on

the risk of heart disease may only be prominent after 35 years and a linear term would tend to smooth over this effect. Sometimes this linear effect is all that is required in terms of predictive ability; from a data analytic viewpoint, however, a linear term might be inappropriate and lead to the wrong interpretation.

In order to generalize (1.1), we propose the model

$$\text{logit } p(\mathbf{x}) = \sum_{i=1}^{p} \phi_i(x_i) \tag{1.2}$$

where $\phi_i(x_i)$ is an unspecified non-parametric smooth function of x_i, and p is the dimension of x. The estimation is performed using the local likelihood technique introduced by Tibshirani (1982) in the context of censored data and the proportional hazards model.

In section 2 we discuss in detail the estimation of the smooth functions and give some ideas on the *degrees of freedom* as well as inference.

In section 3 we look at two examples. One consists of generated data where we know the true function. The other is an analysis of breast cancer data.

In section 4 we compare the technique with the partial residual plots of Pregibon (1981) and also with the smoothing techniques of Henry (1983).

2. ESTIMATION

2.1 The linear logistic model

We first consider the linear case in which logit $p(\mathbf{x}) = \mathbf{x}'\mathbf{a}$. The log-likelihood for n independent observations $(y_1, \mathbf{x}_1), \ldots, (y_n, \mathbf{x}_n)$ is

$$\begin{aligned} l(\mathbf{a}) &= \sum_{i=1}^{n} [y_i \ln p_i + (1 - y_i) \ln(1 - p_i)] \\ &= \sum_{i=1}^{n} [y_i \mathbf{x}_i'\mathbf{a} - \ln(1 + e^{\mathbf{x}_i'\mathbf{a}})] \end{aligned} \tag{2.1.1}$$

where $p_i = p(\mathbf{x}_i)$. Let X be a $n \times p$ matrix of the predictor variables, y an n vector of responses and p an n vector of model probabilities with ith element p_i. The maximum

likelihood estimate $\hat{\mathbf{a}}$ maximizes (2.1.1) and the score function is given by

$$X^t(\mathbf{y} - \hat{\mathbf{p}}) = 0 \tag{2.1.2}$$

where

$$\hat{p}_i = \frac{e^{\mathbf{x}_i^t \hat{\mathbf{a}}}}{1 + e^{\mathbf{x}_i^t \hat{\mathbf{a}}}}$$

The expected information matrix is given by

$$I(\mathbf{a}) = X^t V X$$

where V is a diagonal matrix with ith entry $p_i(1 - p_i)$. The Newton-Raphson iterative procedure can be used to solve the non-linear system (2.1.2) with the estimate at the $(t+1)$ st iteration

$$\hat{\mathbf{a}}(t+1) = \hat{\mathbf{a}}(t) + I^{-1}(\hat{\mathbf{a}}(t)) X^t(\mathbf{y} - \hat{\mathbf{p}}(t)) \tag{2.1.3}$$

See, for example, Landwehr, Pregibon and Shoemaker, (1982).

2.2 The non-linear model with one predictor

In this section we show how to estimate the model logit $p(x) = \phi(x)$ where x is a scalar predictor variable. Let the sample points $x_1, x_2, \ldots, x_n$ be sorted in ascending order.

We wish the estimate at each point x_i to exhibit the local behaviour of the response. We thus consider only those points within a certain neighborhood of x_i and base the estimation on them. The neighborhood is defined in terms of a span, which is a proportion of the sample. Usually we take half the span to the left, and half to the right of x_i. At the endpoints we have to consider asymmetric neighborhoods. Consider then the *local likelihood* for span $s, (s \in (0,1])$, at point i given by

$$l(\mathbf{a}(i), i, s) = \sum_{j=l(i,s)}^{r(i,s)} [y_j a_0(i) + y_j x_j a_1(i) - \ln(1 + e^{a_0(i) + a_1(i) x_j})] \tag{2.2.1}$$

where

$$l(i,s) = \max(0, i - [\frac{ns}{2}])$$
$$r(i,s) = \min(n, i + [\frac{ns}{2}])$$

Let $\hat{\mathbf{a}}(i)$ maximize (2.2.1) and define

$$\begin{aligned}\hat{\phi}(x_i) &= \hat{\phi}_i \\ &= \hat{\mathbf{a}}_0(i) + \hat{\mathbf{a}}_1(i)x_i\end{aligned} \tag{2.2.2}$$

The estimate of $\phi(x_i)$ is only affected by the $ns/2$ nearest neighbors to the left and $ns/2$ to the right, and thus exhibits local properties of the data.

As we move to estimate $\phi(x_{i+1})$, point $l(i,s)$ leaves the likelihood and point $r(i,s)+1$ enters it, and thus the likelihood does not change much. As a consequence, $\hat{\mathbf{a}}(i+1)$ is not much different from $\hat{\mathbf{a}}(i)$, and hence $\hat{\phi}(x_{i+1})$ is not much different from $\hat{\phi}(x_i)$. This results in a smooth estimated curve $\hat{\phi}(.)$. As s increases towards 1, $\hat{\phi}$ will get smoother and in the limit is the usual straight line.

Each local likelihood is maximized using the above iterative procedure and can be time consuming. However, $\hat{\mathbf{a}}(i)$ is an excellent starting value for the $(i+1)$ st local likelihood and convergence is usually achieved in 1 or 2 iterations.

2.3 The non-linear model with more than one predictor

The procedure here is related to the *backfitting* algorithm applied to additive models in Friedman and Stuetzle (1982), and adapted by Tibshirani (1982) for local likelihood estimation in the Cox model.

Suppose we are given $\phi_1(.), \ldots, \phi_{p-1}(.)$ and let

$$\Phi_{(p)}(\mathbf{x}_j) = \sum_{i=1}^{p-1} \phi_i(x_{ji})$$

where $\mathbf{x}_j' = (x_{j1}, x_{j2}, \ldots, x_{jp})$. We need to estimate $\phi_p(x_{jp})$. We can write the local likelihood as

$$l(\mathbf{a}(i), i, s) = \sum_{j=l(i,s)}^{r(i,s)} \left(y_j \Phi_{(p)}(\mathbf{x}_j) + y_j a_0(i) + y_j x_{jp} a_1(i) - \ln(1 + e^{\Phi_{(p)}(\mathbf{x}_j) + a_0(i) + a_1(i) x_{jp}}) \right) \tag{2.3.1}$$

where the data is sorted according to x_{ip}. The score function is now

$$\frac{\partial l(\mathbf{a}(i), i, s)}{\partial \mathbf{a}(i)} = \begin{pmatrix} \sum_{i=l(i,s)}^{r(i,s)} (y_j - \hat{p}_j) \\ \sum_{i=l(i,s)}^{r(i,s)} (y_j x_{jp} - \hat{p}_j x_{jp}) \end{pmatrix} \tag{2.3.2}$$

where logit $(\hat{p}_j) = \Phi_{(p)}(x_j) + \hat{a}_0(i) + \hat{a}_1(i)x_{jp}$. The local information is defined similarly. Thus $\hat{\phi}_p(.)$ can be found using the Newton-Raphson procedure as before. The backfitting algorithm is now given:

Initialize: Set $\hat{\phi}_j^{old}(.) \equiv 0 \forall j, \quad \Phi_{(1)}(.) \equiv 0, \quad k = 0$

Loop: $k = k \pmod{p} + 1$ until convergence.

1) $\Phi_{(k)}(\mathbf{x}_j) = \sum_{l=1}^{p} \phi_l^{old}(x_{jl})$

2) find $\hat{\phi}_k^{new}(.)$ as outlined above.

3) replace $\hat{\phi}_k^{old}(.)$ with $\hat{\phi}_k^{new}(.)$

4) test fit for convergence using

$$dev(\hat{\mathbf{p}}, \mathbf{y}) = -2 \sum_{i=1}^{n} \{y_i \ln \hat{p}_i + (1 - y_i) \ln(1 - \hat{p}_i)\}$$

where logit $\hat{p}_i = \Phi_{(k)}(\mathbf{x}_i) + \hat{\phi}_k^{new}(x_{ik})$

End Loop

The quantity *dev* defined above is the analogue of the residual sum of squares in regression. It is the deviance (Nelder and Wedderburn, (1972)) and is minus two times the maximized log-likelihood. As yet no proof of convergence has been found, although the procedure has converged in all the examples considered by the author.

2.4 Degrees of Freedom and inference

In classical linear logistic regression the maximized liklihood ratios give us a χ^2 test for nested hypotheses. Specifically, let H_0 be the hypothesis that a p dimensional parameter vector $\mathbf{a}$ lies in a q dimensional subspace $\mathbf{a}^*$. We calculate $dev(\hat{\mathbf{a}}^*) - dev(\hat{\mathbf{a}})$ which is distributed as a χ^2_{p-q} variate if H_0 is true. In particular, if H_0 specifies $\mathbf{a} = 0$, then $dev(0) - dev(\hat{\mathbf{a}}) \sim \chi^2_p$ if H_0 is true (Cox, (1970)).

This suggests an *ad hoc* procedure for estimating the complexity or number of parameters in the fit obtained by smoothing against a single predictor variable x. Suppose that $\sum y_i = n_1$. Generate a random sample of size n from a Bernoulli(n_1/n) distribution and

assign them to a vector $\mathbf{y}_k$. Find $\hat{\phi}_k(x)$ and hence $dev(\hat{\mathbf{p}}_k, \mathbf{y}_k)$. Calculate the deviance for the model with only a constant term $dev(n_{1k}/n_k \mathbf{1}, \mathbf{y}_k)$, where n_{1k} is the number of ones observed in replicate k and $n_{1k}/n_k\mathbf{1}$ is the maximum likelihood estimate for $\mathbf{p}$ in the constant model. Finally, let the difference between the two be

$$dev(\hat{\phi}_k) \equiv dev(n_{1k}/n_k \mathbf{1}, \mathbf{y}_k) - dev(\hat{\mathbf{p}}_k, \mathbf{y}_k) \tag{2.4.1}$$

Repeat this a number of times and obtain the mean, variance and quantiles of the $dev(\hat{\phi}_k)$. Let the approximate number of parameters be the mean. The idea is that if the deviances really had a χ^2 distribution, the mean would be the appropriate quantity to use.

The following simple example demonstrates the procedure. 200 values of x were generated from a standard normal distribution. 200 values of y were generated repeatedly 20 times from a bernoulli($\frac{1}{2}$). Each such $\mathbf{y}$ vector was smoothed against $\mathbf{x}$ using the above procedures with spans of $.2, .3, \ldots, .6$. The whole procedure was repeated for 10 different random $\mathbf{x}$ vectors and the results were pooled yielding 200 replications per span. The means and variances of $dev(\hat{\phi}_k)$ for all the spans are given in table (2.4.1).

Table 2.4.1

Means and Variances of $dev(\hat{\phi}_k)$ for different spans

	Spans				
	.2	.3	.4	.5	.6
$\frac{1}{\theta}$	5.0	3.3	2.5	2.0	1.6
Ave(dev)	5.6	3.3	2.0	1.3	1.1
Var(dev)	12.9	8.4	6.1	3.7	3.4
Ratio	.43	.39	.33	.36	.33

Table 2.4.2

Quantiles of $dev(\hat{\phi}_k)$ and appropriate

Chi-square and Gamma distributions

Span==.3

	Quantiles				
Distribution	.50	.75	.90	.95	.99
χ^2_3	2.37	4.11	6.25	7.81	11.3
$dev(\hat{\phi})$	2.69	4.84	7.08	8.75	14.7
χ^2_4	3.36	5.39	7.78	9.49	13.3
$\chi^2_{3.26}$	2.62	4.44	6.65	8.26	11.9
$Gamma(1.27, 0.39)$	2.45	4.50	7.08	8.95	13.6

It would seem that the relationship

$$E(dev(\hat{\phi})) \approx \frac{1}{\theta}$$

is roughly satisfied. For a span of 1 we are back to a linear function then it is satisfied exactly.

The mean-variance ratios given in table (2.4.1) are not in general $\frac{1}{2}$ as would be the case if they had a χ^2 distribution. Furthermore, in table (2.4.2) we examine the quantiles for the case when the span is .3 and compare them to the appropriate χ^2 quantiles. The correspondence is fair but it turns out that we can improve the situation.

Two approaches were considered. Firstly we matched the moments to a Gamma distribution and compared the quantiles. The correspondence is closer than for the χ^2 and is also displayed in table (2.4.2). The other approach was to scale the deviances so that the mean/variance ratio was $\frac{1}{2}$. This is then matched to a χ^2 distribution with appropriate degrees of freedom and displayed in table (2.4.3).

It is not suggested that the Monte-Carlo type experiments are performed each time with real data, since they are computationally expensive. They are given here to support the rule of thumb given above, and give a rough idea about how inference could be performed.

Table 2.4.3
Percentiles of scaled $dev(\hat{\phi}_k)$
and Chi-square distribution
Span=.3

			Quantile		
Distribution	.50	.75	.90	.95	.99
$\chi^2_{2.53}$	1.90	3.48	5.47	6.95	11.8
scaled $dev(\hat{\phi})$	2.10	3.77	5.51	6.81	11.5

3. EXAMPLES

3.1 Simulated data

The first example is simulated data. Two predictor variables were generated independently for each case from a uniform distribution on (-1,1). The values of y were generated

from the bernoulli distribution with mean $p(\mathbf{x})$ where

$$\text{logit } p(\mathbf{x}) = x_1 + 2\sin(\pi x_2).$$

200 such observation vectors $(y_i, \mathbf{x}_i)$ were independently generated. The procedure converged in 3 iterations of the backfitting algorithm with a span of .5 to the solution given in figures (3.1.1) and (3.1.2). The continuous curve is the true function and the points represent the estimated function.

3.2 A real example

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 four variables.

y_i = 1 if patient i survived 5 years or longer
= 0 otherwise
x_{i1} = age of patient i at time of operation
x_{i2} = year of operation i (minus 1900)
x_{i3} = number of positive axillary nodes detected in patient i

The *linear* logistic model yielded $dev = 328.25$ with 302 degrees of freedom (dof). Landwehr et al (1982) analyzed this data set and in particular considered partial residual plots in order to identify the fundamental form in which terms should appear. Their final model was

$$\text{logit } p(\mathbf{x}) = \beta_0 + x_1\beta_1 + x_1^2\beta_2 + x_1^3\beta_3 + x_2\beta_4 + x_1x_2\beta_5 + (\log(1+x_3))\beta_6 \qquad (3.2.1)$$

with a dev of 302.3 on 299 dof. We fitted the additive model

$$\text{logit } p(\mathbf{x}) = \sum_{i=1}^{3} \phi_i(x_i) \qquad (3.2.2)$$

with a span of .5 at each stage. The results are displayed in figures (3.2.1),...,(3.2.3). The iterations converged to a dev of 307.37 with an *estimated* dof of $301.41 = 306 - 5.59 = 306 - (1 + 1.70 + 1.42 + 1.47)$. In model (3.2.1) an interaction term for x_1 and x_2 is included. This caused the dev to drop only 1.6 in our model, so we left it out.

We estimated the dof in a similar fashion to the technique used above. For each of the predictor variables, a small Monte-Carlo experiment was performed. For each of 100

replications, a sample of size 306 was drawn independently from a Bernoulli distribution with parameter $p = 225/306$. This corresponds to the complete independence model since there were 225 observed positive responses in the sample. The generated responses were smoothed against the predictor variable in question, and the resultant value of $dev(\hat{\phi})$ was recorded. This is once again a difference of deviances as in (2.4.1). The estimated dof for the predictor is the mean of the 100 replicates of $dev(\hat{\phi})$. We also add in 1 for the constant term.

The shape of the estimated curves immediately give us an idea of the functional form in which each predictor appears in the model. In particular the Clemenson's hook of x_1 (Landwehr et. al. 1982) is directly modelled without resorting to a cubic term, as is the logarithmic form of x_3.

4. DISCUSSION

The method described above gives a direct way of identifying nonlinear effects in the logistic regression model. Clearly these ideas are not restricted to logit models, but can be applied to any parametric model in which estimation is done by maximum likelihood. In the GLIM package, the range of models include loglinear models for contingency tables (Poisson or multinomial data), linear regression models for normal errors, the above logistic regression models as well as components of variance models (Gamma variates).

The regression situation has long been solved using the variety of scatterplot smoothers available. These could not be used directly for the logistic model since the logit transform is not defined for the 0-1 data. That is why maximum likelihood is an attractive framework for estimation, since the observations appear implicitly in the estimates.

In the case of loglinear models, the technique is not really useful since the predictor variables are usually discrete and unordered, and the concept of *local* no longer has meaning. In those situations, the best nonlinear model is found anyway, since the model fits a constant for each category.

Landwehr et. al. (1982) use partial residual plots to identify non-linearities. These will not be described in detail here. Partial residual plots are used in linear regression models and *may*, in some situations, identify the form of the nonlinear effect of a particular predictor variable. In order to generalize the ideas to logistic regression, one exploits the

relationship between the Newton-Raphson updating equations and generalized least squares (details can be found in the above reference, and in Nelder and Wedderburn, (1972)). The local likelihood method should always identify the nonlinear form. The partial residual plots are still useful in their own right, since they identify outliers.

Henry, (1983) has considered methods for directly smoothing the 0-1 response variables. A weighted form of conditional expectation is used (nearest neighbor averaging) and the posterior probabilities are modelled directly. One disadvantage is that the logit form is no longer there to guarantee that the estimates are in $[0, 1]$. This provides problems especially when more than one term is in the model. Since we model the logit of p, the terms in the model, be they linear or non-linear, can take on any finite values. A major advantage is that the smoothing models are computationally much cheaper since the estimate at point i can be updated to yield the estimate at point $(i + 1)$. This is not possible using the local likelihood technique.

Logistic regression is a widely used technique, and it seems appropriate to have at hand a method for identifying non-linearities within that framework. Often simple parametric forms are suggested, and one can then return to the standard models. With the rapid improvements in computing power, techniques which would previously have been impossible are now becoming computationally viable.

Acknowledgements

I wish to thank Rob Tibshirani for his many ideas on the topic and for his constant support. I also wish to thank Rob and Werner Stuetzle for their careful reading and useful comments on earlier drafts.

REFERENCES

Baker, R.J. and Nelder, J.A. (1978). *The GLIM System-Release 3.* Distributed by the Numerical Algorithms Group: Oxford.

Cox, D.R. (1970). *Analysis of Binary Data*, London: Chapman and Hall.

Friedman, J.H. and Stuetzle, W. (1980).*Projection pursuit classification.* Unpublished manuscript.

Friedman, J.H. and Stuetzle, W. (1982). *Smoothing of Scatterplots,* Dept. of Statistics Tech. Rept. Orion 3, Stanford University.

Haberman, S.J. (1976). *Generalized Residuals for Log-Linear Models.* Proc. 9^{th} Int'l Biometrics Conference, Boston, **104-122.**

Henry, D. (1983) *Projection Pursuit Odds-Ratio Regression and Classification* (in press)

Lachenbruch, P.A.(1975) *Discriminant Analysis*, New York: Hafner Press.

Landwehr, J.M., Pregibon, D and Shoemaker, A.C. (1982). *Graphical Methods for assesing Logistic Regression Models.* (in press.)

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

Pregibon, D. (1981). *Logistic Regression Diagnostics*, Annals of Statistics **9**, 705-724.

Tibshirani, R. (1982). *Non-Parametric Estimation of Relative Risk.* Submitted to the Journal of the American Statistical Association.

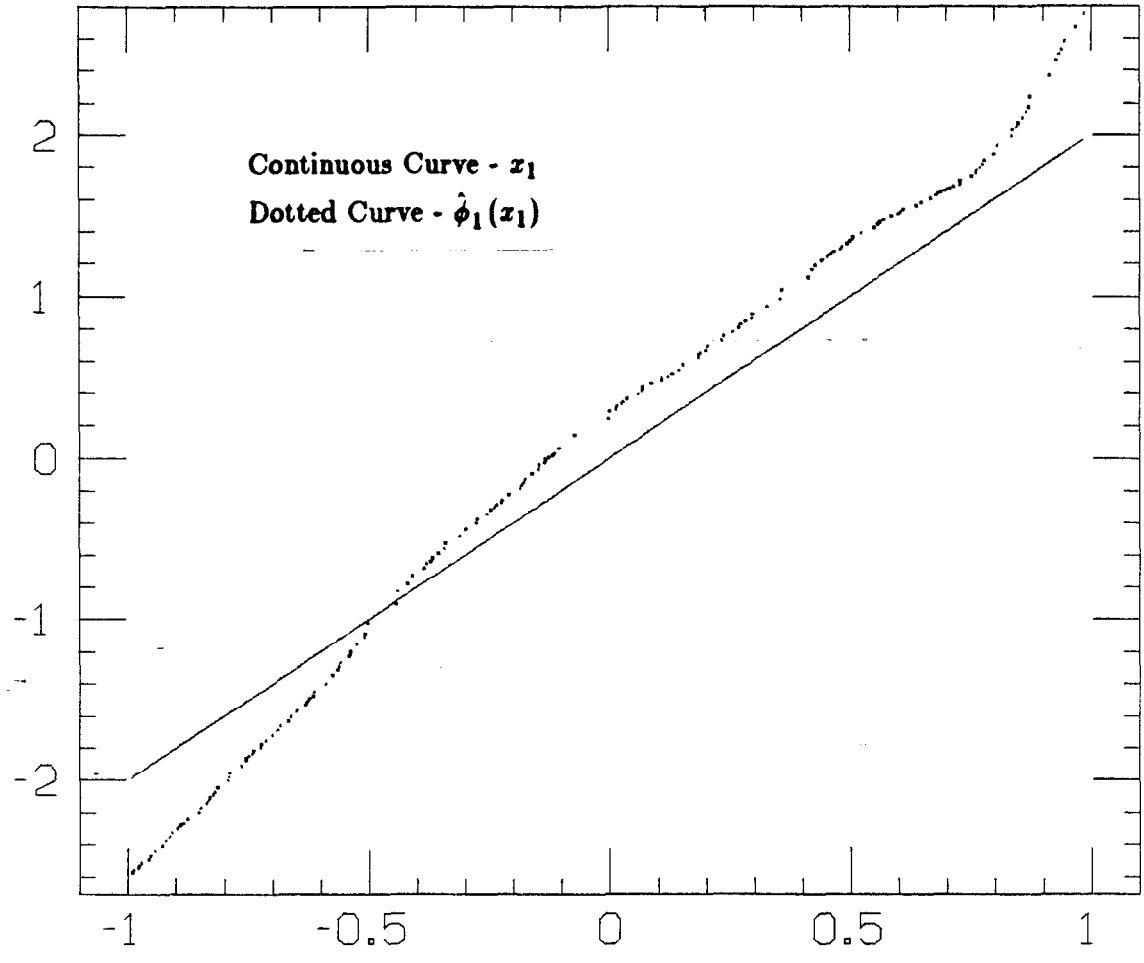


Figure 3.1.1

The true function and fitted function for the simulated data. Variable number 1 in the true model logit $p(\mathbf{x}) = x_1 + 2\sin(\pi x_2)$. The span of the local likelihood smoother is .5.

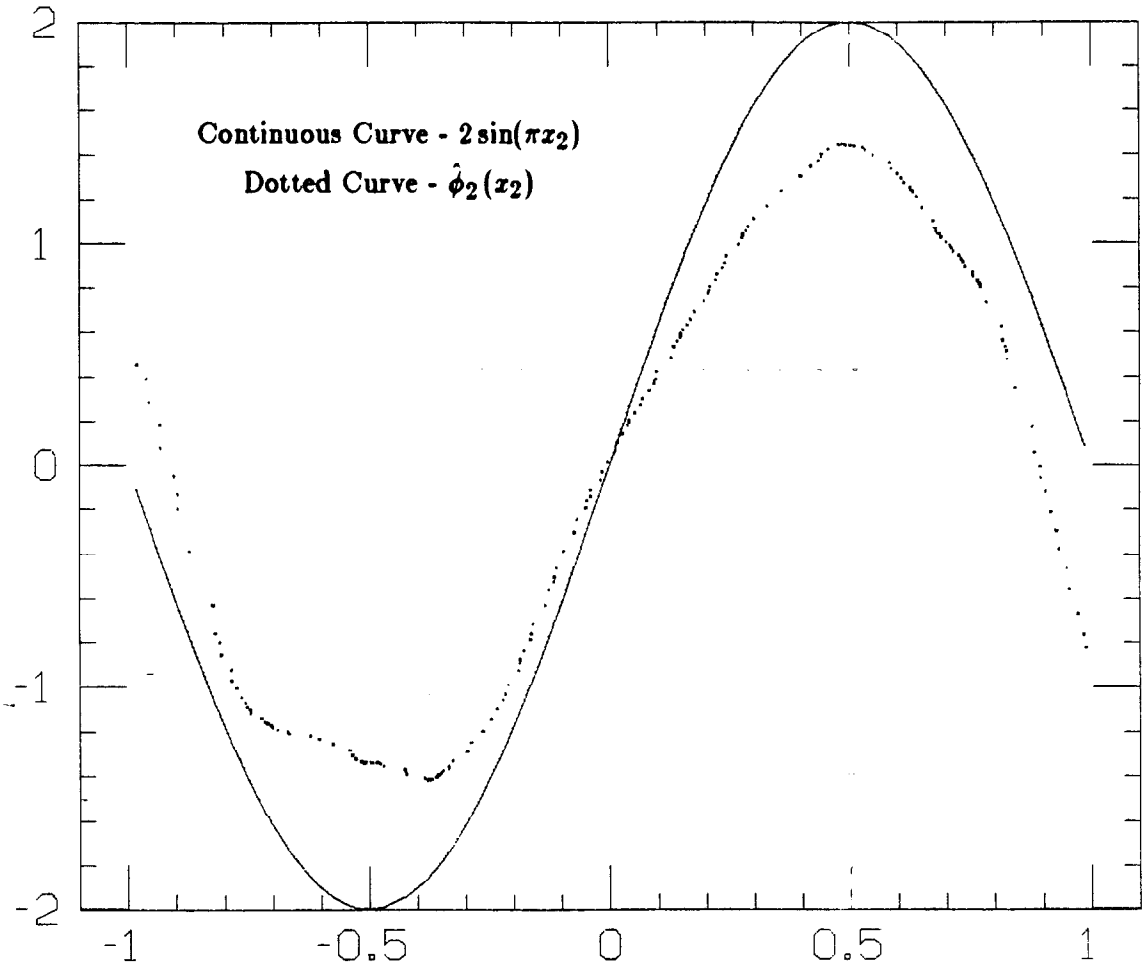
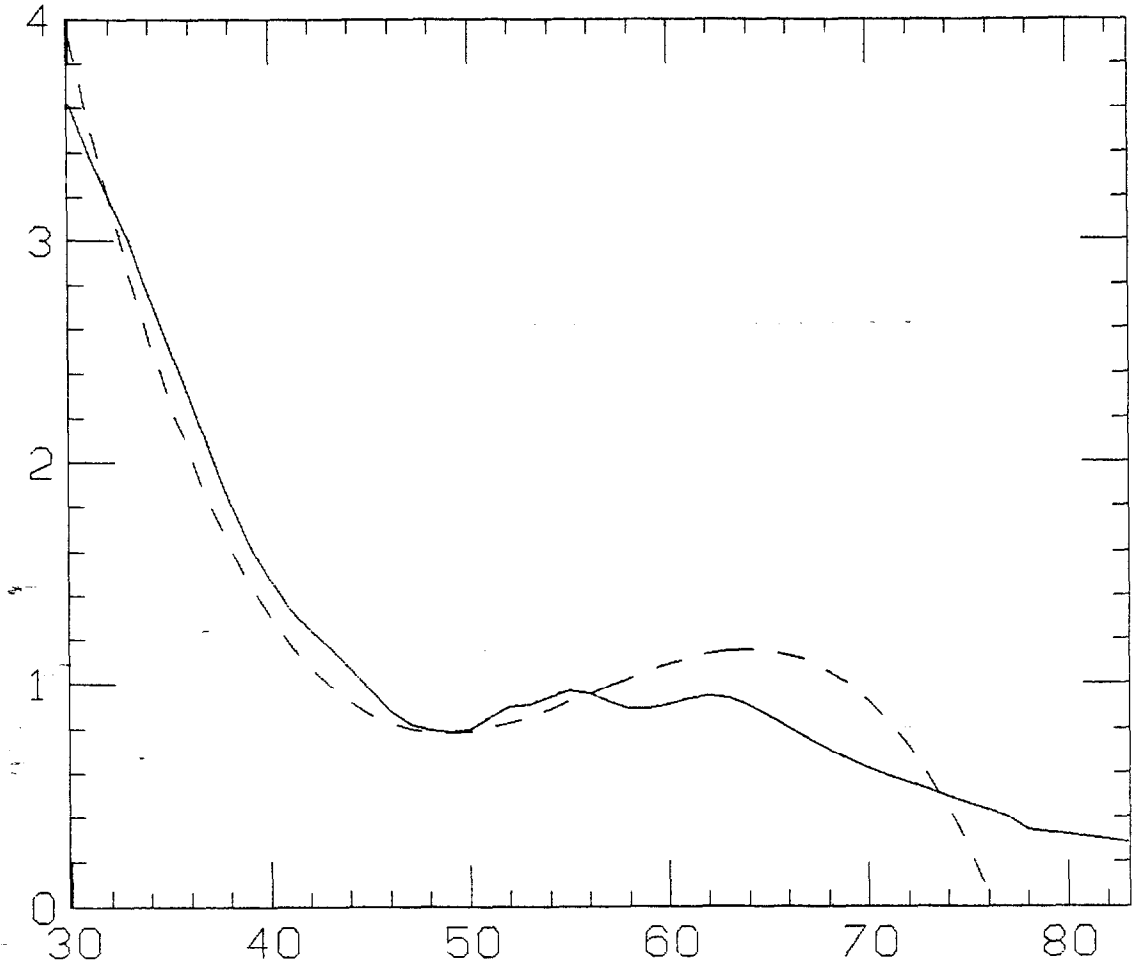


Figure 3.1.2

The true function and fitted function for the simulated data. Variable number 2 in the true model logit $p(\mathbf{x}) = x_1 + 2\sin(\pi x_2)$. The span of the local likelihood smoother is .5.



x_1 age of patient

Figure 3.2.1

The fitted function $\hat{\phi}_1(x_1)$ for the breast cancer data. The span of the local likelihood smoother is .5. The estimated *dof* is 1.70 and the *dev* is 307.37.

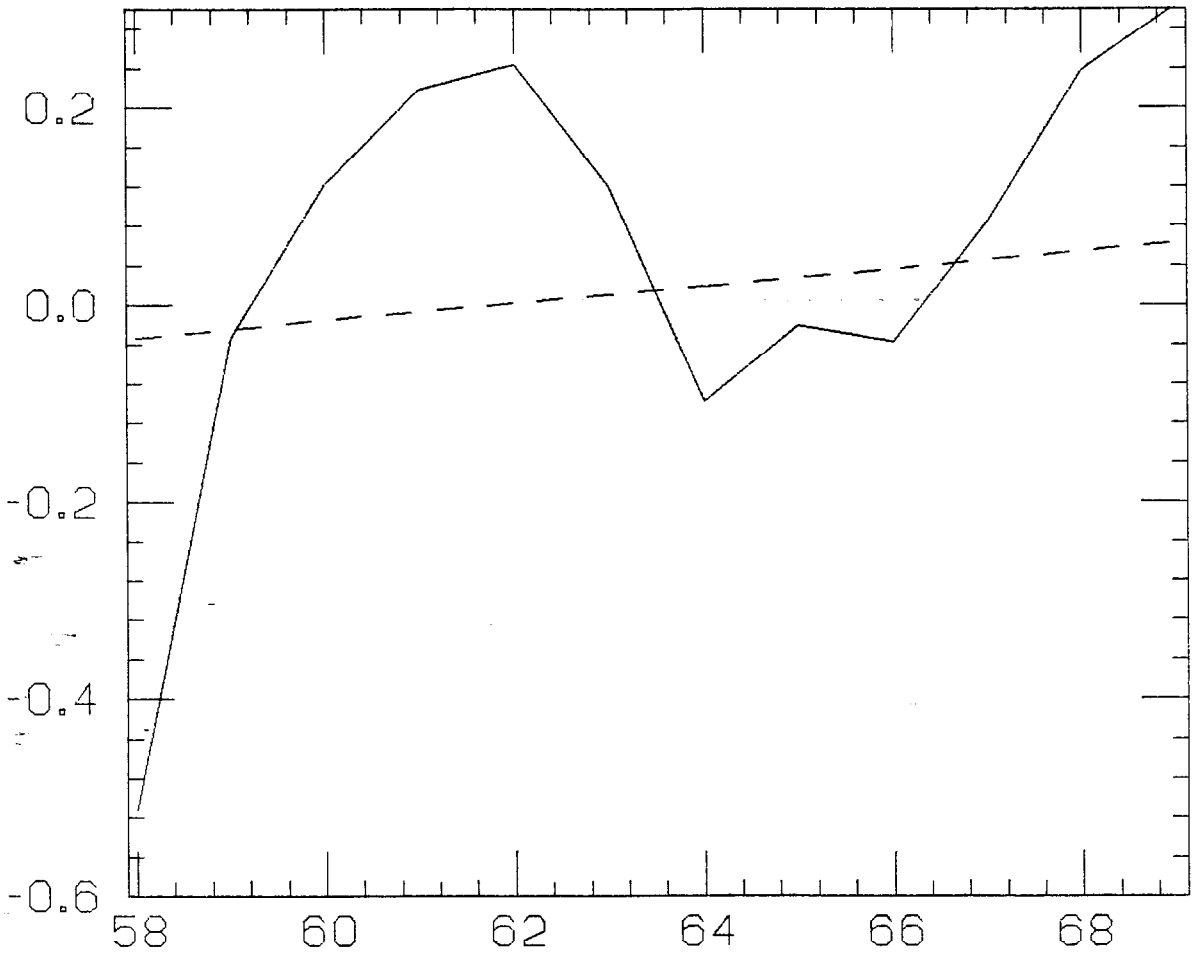


Figure 3.2.2

The fitted function $\hat{\phi}_2(x_2)$ for the breast cancer data. The span of the local likelihood smoother is .5. The estimated *dof* is 1.42 and the *dev* is 307.37.