J. R. Statist. Soc. B (1995)
57, *No.* 3, *pp.* 519–532

Nonparametric Transformations for Both Sides of a Regression Model

By DOUGLAS NYCHKA† and DAVID RUPPERT

North Carolina State University, Raleigh, USA

[Received May 1992. Final revision July 1994]

SUMMARY

One way to model heteroscedasticity and skewness of the error distribution in regression is to transform both sides of the regression equation. If it is possible to transform the regression equation to result in normally distributed errors, then we can obtain more efficient parameter estimates and valid prediction intervals. One problem with this approach is that the choice of transformation is usually restricted to the power or shifted power family. Often there is no scientific basis for this model and the limited flexibility of this parametric family may miss important features of the distribution. A more comprehensive approach is to estimate the transformation by using nonparametric methods based on maximizing a penalized likelihood function. This maximization problem leads naturally to a spline estimate for the log-derivative of the transformation. An algorithm for computing the estimate is given along with results on the existence and uniqueness of the estimate. This nonparametric method is illustrated by using a non-normally distributed fisheries data set.

Keywords: MAXIMIZATION; NONPARAMETRIC TRANSFORMATION; PENALIZED LIKELIHOOD; SMOOTHING SPLINES; TRANSFORM BOTH SIDES METHOD

1. INTRODUCTION

The presence of heteroscedasticity or non-normal errors is a typical problem encountered in regression analysis. Consider the model $Y_k = f(X_k,\ \beta) + e_k$, $1 \leqslant k \leqslant N$, where X_k and Y_k are independent and dependent variables related by a parametric function $f(\ ,\ \beta)$. The random components $\{e_k\}$, $1 \leqslant k \leqslant N$, are assumed to be independent and to have a median of 0 but need not be normally, or even be identically, distributed. If the parameters in this model are estimated under the assumption of normally distributed errors, however, then the efficiency of the estimates may be low and prediction intervals derived from the estimated model may be inaccurate.

One approach used to account for departures from normality is to transform the dependent variable. Non-linear transformations often yield a symmetric distribution for the errors and when heteroscedasticity is linked to the mean level a suitable transformation may also induce a constant variance. Simply transforming the dependent variable can make it difficult to interpret the regression equation with respect to a transformed response. This mismatch is especially a problem when the relationship among dependent and independent variables in the original scale is suggested by a scientific theory or previous empirical work. In this case transforming just the dependent variable destroys the functional relationship between the

†*Address for correspondence*: Department of Statistics, North Carolina State University, Box 8203, Raleigh, NC 27695-8203, USA.
E-mail: nychka@stat.ncsu.edu

0035–9246/95/57519

dependent variable and the model predictions. A natural solution to this problem is the transform both sides (TBS) operation on the regression equation, i.e. the transformed dependent variable is modelled by the transformed prediction equation. Letting H denote a transformation, the TBS model is

$$H(Y_k) = H\{f_k(\beta)\} + \epsilon_k,$$

where $f_k(\beta) \equiv f(X_k, \beta)$ and it is assumed that $\{\epsilon_k\}$ are independently distributed, $N\{0, V_k(\delta)\}$ random variables. Here $V_k(\delta) = V(X_k, \delta)$ is a known scale function that depends on the variance parameter vector δ. Together V_k and δ model any remaining heteroscedasticity *after* transformation, in particular heteroscedasticity not linked to the mean level. If the transformation is monotonic, then the original prediction equation remains a model for the median response independently of the choice of the transformation. Also note that prediction intervals can be constructed in the original scale based on normal theory for the transformed equation.

The transformations considered in regression analysis are usually limited to a power transformation or a shifted power (SP) transformation. For $u + \lambda_2 \geqslant 0$ define

$$H(u, \lambda) = \begin{cases} \{(u + \lambda_2)^{\lambda_1} - 1\}/\lambda_1 & \lambda_1 \neq 0, \\ \log(u + \lambda_2) & \lambda_1 = 0. \end{cases}$$

Since the choice of transformation is largely empirical it is important to consider the sensitivity of $\hat{\beta}$ to H. One problem with using parametric transformations is the difficulty in extending H beyond an SP transformation. Thus, it is not easy to assess the effect of more flexible transformations on the regression parameters or on prediction intervals in the TBS model. Rather than to create more complicated transformations based on parametric expressions, we believe that it is more efficient to consider a nonparametric method of determining H. This approach not only solves the problem of how to augment the SP transformation family but also introduces more objectivity into the choice of transformation.

Given a parametric model for f, H and the variance function V, we can estimate the parameters by maximum likelihood (Carroll and Ruppert, 1988). This paper considers estimates of the model parameters and a nonparametric transformation by maximizing a penalized likelihood. Let $L(\theta, H)$ denote the log-likelihood of the observations where for convenience the parameters β and δ have been stacked into a single vector θ. A penalized likelihood for this model is

$$L_P(\theta, H) = L(\theta, H) - \rho J(H), \qquad \rho > 0. \tag{1.1}$$

Here $J(H)$ is a positive functional that quantifies the roughness of the transformation whereas ρ is the relative weight between the roughness penalty and the unconstrained log-likelihood of the data. The smoothing parameter ρ controls the amount of flexibility in the resulting estimate of H and provides a means for varying the complexity of the transformation. Typically, as ρ approaches ∞, H will have the form of a simple parametric function, such as a power transformation. For the TBS model, as ρ approaches 0, H will tend to a rougher more complicated function that can reflect the influence of small subsets of the data. Of course in practice we would expect the best choice for H to lie somewhere in between these two extremes.

By focusing on maximum penalized likelihood estimates (MPLEs) it is natural to formulate estimates of H in terms of smoothing splines. To ensure monotonicity of H it is convenient to represent the transformation in terms of its log-derivative g, and the roughness penalty $J(H)$ will be based on the second derivative of g. By considering an approximation to the integrals in the likelihood, which we believe to be highly accurate, it is possible to compute the MPLE of the regression model and H by an iterative procedure that relies on the Fisher method of scoring and fitting weighted, cubic smoothing splines. This algorithm appears to be stable. When applied to non-normal data sets the results give some insight into the sensitivity of prediction intervals with respect to H.

The next section outlines the method used to estimate the nonparametric TBS model. Section 3 reviews some relevant aspects of cubic smoothing splines and gives the details of computing the nonparametric portion of the model. Transformations and regression functions are estimated for a biological example and these results are reported in Section 4. Section 5 discusses the existence of maximizers of the penalized likelihood for the TBS model. The reader should note that a longer version of this paper is available (Nychka and Ruppert, 1992) that gives another example of fitting a TBS model to data. This example includes a variance function and a shift parameter (λ_2). Also this paper gives the proofs for the theorems cited in Section 5.

2. MAXIMIZING TRANSFORM BOTH SIDES PENALIZED LIKELIHOOD

2.1. *Transform Both Sides Likelihood and Roughness Penalty*

Under the assumptions discussed in Section 1, the log-likelihood for the observations is given by

$$L(\theta, H) = \sum_{k=1}^{n} (-\tfrac{1}{2}\,[H(Y_k) - H\{f_k(\beta)\}]^2/V_k(\delta) + \log H'(Y_k) - \tfrac{1}{2}\log V_k(\delta)) + C. \tag{2.1}$$

The penalty function for H is defined through the log-derivative. Set $g(u) = \log H'(u)$. By expressing the penalized likelihood with respect to g we can estimate the model by an unconstrained maximization problem, since any choice of g implies a monotonic H. The roughness penalty that is used in this work is

$$J(H) = \int_a^b g''^2\,\mathrm{d}u \tag{2.2}$$

where a and b are chosen such that $\{Y_k\} \subset [a, b]$. Using this parameterization for the transformation, the penalized likelihood is

$$L_{\mathrm{P}}(\theta, H) = \frac{1}{2}\sum_{k=1}^{n}\left\{-\left(\int_{A_k} \exp g(u)\,\mathrm{d}u\right)^2\bigg/ V_k(\delta) + 2g(Y_k) - \log V_k(\delta)\right\}$$
$$- \rho\int_a^b g''^2\,\mathrm{d}u + C \tag{2.3}$$

where A_k is the interval with end points at Y_k and $f(X_k, \beta)$.

A penalized likelihood has the property that as $\rho \to \infty$ the estimate of H will be the transformation that maximizes the likelihood subject to the constraint $J(H) = 0$.

The set of transformations such that $J(H) = 0$ will be referred to as the null space and is a simple parametric family. If $J(H) = 0$ then g must be a linear function, say $g(u) = \gamma_1 + \gamma_2 u$, and thus $H(u) - H(u_0) = (\exp \gamma_1/\gamma_2)\{\exp(\gamma_2 u) - \exp(\gamma_2 u_0)\}$. This parametric family may not seem unusual if we initially transform both sides of the regression equation by using the log-function and then consider H as an additional transformation. In this case u is actually in a log-scale with respect to the original observations. Setting $u = \log \omega$, then $H(u) = H(\log \omega) = C_1 \omega^{\gamma_2} + C_2$ where C_1 and C_2 depend on γ and U_0. Thus, if the initial model is 'preloaded' with a log-transformation, the null space of the roughness penalty consists of power transformations.

2.2. *Iterative Scheme for Computing Maximizer*

The penalized likelihood can be maximized by an iterative procedure that alternates between maximizing L_P over θ and then H. This separation is particularly suited to the TBS model because usually the estimates of the regression parameters are not very sensitive to the choice of transformation. Accordingly let θ_k and H_k be the estimates at the kth iteration. Given starting estimates θ_0, and $H_0(u) = u$,

```
     Do {
(1)       $\theta_{k+1}$ = maximizer of $L_p(\theta, H_k)$ for $\theta \in \mathscr{R}^p$
(2)       $H_{k+1}$ = maximizer of $L_p(\theta_{k+1}, H)$ for $J(H) < \infty$
     }
Until (convergence)
```

If this procedure does converge then $(\theta_\infty, H_\infty)$ will be a local maximum of L_p.

Implementing step (1) in the loop is the straightforward problem of computing a maximum likelihood estimate for a parametric model. For a fixed H, L_P can be maximized by using Fisher's method of scoring. The second step is more difficult as it involves a functional maximization and is the subject of Section 3.

2.3. *Connections with ACE, AVAS and LMS*

Two other methods that estimate a nonparametric transformation for regression data are ACE (Breiman and Friedman, 1985) and AVAS (Tibshirani, 1988). For bivariate data (X_k, Y_k) the ACE procedure estimates H and f to minimize the transformed residual sum of squares $\Sigma\{H(Y_k) - f(X_k)\}^2$ subject to the variance of $\{H(Y_k)\}$ being equal to 1. The AVAS procedure estimates H as the asymptotic variance stabilizing transformation. In either of these methods the estimation procedure is not likelihood based and different transformations are applied to the two sides of the regression equation. In fact, one version of ACE does not constrain f or H to be monotonic.

The advantage of ACE and AVAS is their power for fitting flexible models in the absence of any scientific guidance for formulating a parametric regression function. Our spline–TBS method has a very different advantage: the ability to transform Y and yet to preserve a model based on scientific theory. The algorithm for computing these estimates is similar to that outlined above in that we alternate between optimizing with respect to the regression estimate and the transformation.

A slightly different approach to modelling departures from normality is to assume that, conditional on X, Y can be transformed by the power transformation to a normal distribution (Green, 1988). Cole and Green (1992) have suggested a transformation (in our notation) of the form

$$H(u, X) = \frac{\{u/M(X)\}^{L(X)} - 1}{L(X)\,S(X)}$$

where M is a mean function and S is a scale function. Under the assumption that L, M and S are smooth functions Cole and Green estimated these functions by using a penalized likelihood. One advantage of this formulation is that the maximization of the penalized likelihood for each of the component functions is numerically simpler than the transformation spline described above. However, it is not clear how this approach can be extended to a TBS model. Since the TBS model only involves one nonparametric function it may provide a more parsimonious representation of the conditional distribution than the L- and S-curves.

3. ESTIMATING NONPARAMETRIC TRANSFORMATION

3.1. *Smoothing Splines*

Consider the nominal smoothing problem $Z_k = g(u_k) + e_k$, for $1 \leqslant k \leqslant N$, where e_k are assumed to be independent $N(0, 1/w_k)$. For $\rho > 0$ a cubic (weighted) smoothing spline $\hat{g}$ is defined as the function that minimizes

$$\mathscr{L}(g) = \frac{1}{N} \sum_{k=1}^{N} \{Z_k - g(u_k)\}^2 w_k + \rho \int_a^b g''^2 \,\mathrm{d}u \tag{3.1}$$

over all g such that

$$\int_a^b g''^2 \,\mathrm{d}u < \infty.$$

It is well known that $\hat{g}$ is a piecewise cubic polynomial with join points at the unique values of $\{u_k\}$. The spline solution can be parameterized by the value of the function at the points $\{u_k\}$. Let $\mathbf{g}^{\mathrm{T}} = \{g(u_1), \ldots, g(u_N)\}$ and using this parameterization and the knowledge of the functional form of the solution the minimization problem (3.1) is equivalent to

$$\min_{\mathbf{g} \in \mathbf{R}^N} \left\{ \frac{1}{N} \sum_{k=1}^{N} (Z_k - g_k)^2 w_k + \rho \mathbf{g}^{\mathrm{T}} R \mathbf{g} \right\}. \tag{3.2}$$

Here R is a matrix derived from the roughness integral that only depends on $\{u_k\}$. Differentiating with respect to $\mathbf{g}$, the minimizing vector is the solution to the linear system

$$-2(Z_k - g_k)w_k + 2\rho[R\mathbf{g}]_k = 0, \qquad 1 \leqslant k \leqslant N. \tag{3.3}$$

The reader is referred to Eubank (1988) for a derivation of this estimate and Hutchinson and de Hoog (1985) for some background on its computation.

An important feature of this characterization of a spline is that the abstract minimization problem (3.1) can be related to solving a finite dimensional problem. This reduction will also hold for more complicated models. Suppose that $Q(g)$ is a continuous functional that only depends on g through the evaluation functionals $\{g(u_k)\colon 1 \leqslant k \leqslant N\}$. If

$$Q(g) + \rho \int_a^b g''^{\,2}\,\mathrm{d}u$$

has a minimizer over g such that $\int_a^b g''^{\,2}\,\mathrm{d}u < \infty$, then the solution will be a piecewise cubic polynomial with the same boundary conditions described above. A proof of this result is given by O'Sullivan *et al.* (1986) and this fact will be used in characterizing the spline estimate of the transformation.

3.2. *Approximate Penalized Likelihood*

It is difficult to maximize the penalized likelihood exactly because of the way that H appears in the likelihood function. Our approach is to consider an accurate approximation to this likelihood that is better suited for computation. By definition, H is the indefinite integral of $\exp g$ and our numerical strategy is to approximate these integrals by sums. Let $f_k \equiv f_k(\theta) \equiv f_k(\beta)$ and $V_k \equiv V_k(\delta)$. Choose a sufficiently fine mesh of points $u_1 < u_2 < \ldots < u_N$ so that $\{Y_k\}$ is contained in $[u_1, u_N]$. Let $g_j = g(u_j)$ and choose quadrature weights W_{kj} so that

$$\int_{f_k}^{Y_k} \exp g(u)\,\mathrm{d}u \approx \sum_{j=1}^{N} W_{kj} \exp g_j \tag{3.4}$$

and interpolation weights $\{\delta_{kj}\}$ such that

$$g(Y_k) \approx \sum_{j=1}^{N} \delta_{kj} g_j. \tag{3.5}$$

The mesh size is arbitrary and can be chosen to achieve any desired level of accuracy for these approximations. However, because g is expected to be smooth, the quadrature formula should be accurate and the interpolation error should be small even for modest mesh sizes. If the observations $\{Y_k\}$ are included as a subset of the mesh then the interpolation error is 0.

For fixed θ this discretization suggests an approximate penalized log-likelihood:

$$L_{\mathrm{PA}}(g) = \sum_{k=1}^{n} \left[\left\{ \sum_{j=1}^{N} W_{kj} \exp g(u_j) \right\}^2 \Big/ 2V_k + \sum_{j=1}^{N} \delta_{kj}\, g(u_j) \right] - \rho \int_a^b g''^{\,2}\,\mathrm{d}u. \tag{3.6}$$

If we identify the first expression with the non-linear functional $-Q(g)$, then it is clear that Q will only depend on g through the evaluation of this function at the mesh points. By definition the roughness penalty in terms of g is just the usual penalty based on the integrated, squared second derivative. Thus $L_{\mathrm{PA}}(g)$ has the form

$$-Q(g) - \rho \int_a^b g''^{\,2}\,\mathrm{d}u$$

and the maximizer of L_{PA} for $\int_a^b g''^{\,2}\,\mathrm{d}u < \infty$, if it exists, will be a piecewise cubic polynomial with knots at $\{u_j\}$. Moreover, because the functional form of the solution is known, it is enough to identify the solution at the mesh points. Therefore we re-express equation (3.6) in vector notation. Let $\mathbf{g} = (g(u_1), \ldots, g(u_N))^{\mathrm{T}}$ and $h_j = \exp g_j$; then

$$L_{\mathrm{PA}}(\mathbf{g}) = -\tfrac{1}{2}\,\mathbf{h}^{\mathrm{T}}\Omega\mathbf{h} + \mathbf{D}^{\mathrm{T}}\mathbf{g}^{\mathrm{T}} - \rho\mathbf{g}^{\mathrm{T}}R\mathbf{g} \tag{3.7}$$

where

$$\Omega = W^{\mathrm{T}}\,\mathrm{diag}(V)\,W,$$

where W is the matrix with entries w_{kj}, $\mathbf{D} = (D_1, \ldots, D_N)^{\mathrm{T}}$ where

$$D_j = \sum_{k=1}^{N} \delta_{kj}$$

and R is the roughness penalty matrix as described in Section 3.1. Setting partial derivatives of equation (3.7) equal to 0 gives a system of non-linear equations that are necessary for any extremal point of the approximate penalized likelihood:

$$\frac{\partial}{\partial g_j} L_{\mathrm{PA}}(\mathbf{g}) = -h_j[\Omega\mathbf{h}]_j + D_j - 2\rho[R\mathbf{g}]_j, \qquad 1 \leqslant j \leqslant N. \tag{3.8}$$

Under fairly weak assumptions a unique maximizer of L_{PA} will exist (see Section 5) and thus the non-linear system at equation (3.8) gives sufficient conditions for a maximizer. The next section describes an iterative method to solve this system based on ordinary cubic smoothing spline algorithms.

3.3. *Maximizing L_{PA} for Fixed Regression Model*

Although the transformation spline is a non-linear function of the observed data, it is possible to linearize the defining system about a previous estimate so that it resembles system (3.3). This suggests an iterative algorithm where at each stage efficient cubic smoothing spline algorithms are used to compute a solution to an approximate linear system.

The computational strategy is to approximate the first two terms of equation (3.8) so that they can be expressed as $-2(Z_j - g_j)w_j$ (compare equation (3.3)) for some choice of pseudodata Z and weights w. This process involves two steps, a diagonalization of the system so that the jth equation only depends on h_j and $[R\mathbf{g}]_j$ and a linearization of h_j with respect to g_j. (See Appendix A.) With this linearization–diagonalization of equation (3.8) the following algorithm is proposed for solving the system.

```
Determine quadrature mesh $\{u_j\}$ $1 \leqslant j \leqslant N$
Compute $\Omega$ and $\mathbf{D}$
Initialize: $g^{\mathrm{OLD}} \equiv 0$
Do {
Compute $\mathbf{Z}$ and $\mathbf{w}$ based on $g^{\mathrm{OLD}}$, $\Omega$, $\mathbf{D}$
Compute $g^{\mathrm{NEW}}$, a cubic smoothing spline for $\{u_j, Z_j\}$ with weights $\{w_j\}$
$g^{\mathrm{OLD}} = g^{\mathrm{NEW}}$
}
Until (convergence)
```

When computing the spline–TBS estimate, this algorithm is used in step (2) of the algorithm of Section 2.2.

3.4. *Practical Implementation of Algorithm*

An algorithm for estimating the transformation was written as a Fortran subroutine and called from within the S computing environment (Becker *et al.*, 1988). Let $[a, b]$ denote the interval where a and b are the minimum and maximum values of Y_k respectively. In specifying the approximate likelihood the quadrature mesh was taken as the *union* of 150 equally spaced points in $[a, b]$ and the unique values of the dependent variable. This choice is feasible for moderate size data sets and eliminates the error in the likelihood approximation due to evaluating g at $\{Y_k\}$. Although there is some subjectivity with this choice, we did not find that the range of $[a, b]$ influenced the resulting estimate. This insensitivity is due to the fact that outside the range of Y_k and f_k the likelihood does not contribute any information about g. Thus the second derivative roughness penalty will tend to constrain the resulting estimate to be linear in this range.

To find the transformation more efficiently, it was helpful to use the transformation in the previous iteration as the starting values. We found that nonparametric spline estimates of the transformation can be computed sufficiently rapidly for interactive data analysis and problems of convergence were only encountered for rough estimates of g (very small values of ρ). A reduced step size in updating the transformation helped in making the algorithm more stable. Following the notation from Section 3.3, this is an updated estimate of the form $g^{\mathrm{OLD}} + \alpha(g^{\mathrm{NEW}} - g^{\mathrm{OLD}})$ where $0 < \alpha < 1$.

The linearization of the estimating equations is admittedly *ad hoc* and unfortunately we cannot give any arguments for the algorithm's convergence. We emphasize, however, that if our algorithm does converge then we do obtain an extremum of the approximate penalized likelihood. Moreover if we focus just on the step of estimating the transformation we have identified fairly weak conditions in Section 5 where a unique MPLE exists. An alternative to our scheme of linearizing system (3.6) is to carry out Newton's method directly.

3.5. *Smoothing Parameter Selection*

Most nonparametric estimates of functions have a free parameter that controls the flexibility of the resulting curve estimate. In this case, it is ρ, the relative weight given to the roughness penalty over the log-likelihood. One practical issue is to choose an appropriate value for this parameter when no *a priori* information is available about the smoothness of the transformation. Because this is a non-linear estimate, previous work on smoothing parameter or bandwidth selection does not apply. One computationally intensive strategy is to estimate ρ on the basis of cross-validation. For a fixed value of ρ each data point is omitted from the likelihood and H and θ are estimated on the basis of the remaining $n - 1$ data points. The log-likelihood based on these estimates is now evaluated at the omitted data point. We do this calculation for each data point and then assemble the cross-validated log-likelihood. This quantity is now a measure of the effectiveness of using the particular value of ρ. The process is repeated for various values of ρ and the value of the smoothing parameter that maximizes this criterion is taken to be a good choice for ρ. Besides using the likelihood as a loss function, we might also consider the sum of squares of the differences between the omitted data point and its predicted value in the transformed scale.

A simpler strategy exploits the particular features of this problem. We hope that a good estimate of H will yield transformed residuals that are independent with constant variance. Thus any statistic used to test for departures from these assumptions can be used to judge the suitability of a particular value for ρ. In the example given in the next section, we examined the heteroscedasticity in the transformed residuals based on the Spearman rank correlation of the absolute residuals with the predicted values. Relatively small values of this correlation may suggest good choices for the smoothing parameter. Note that in this comparison we would not use this statistic in a formal test for departures from independence; rather the relative size of the statistic is used to identify a suitable value for ρ.

4. APPLICATION: SKEENA RIVER SALMON POPULATION

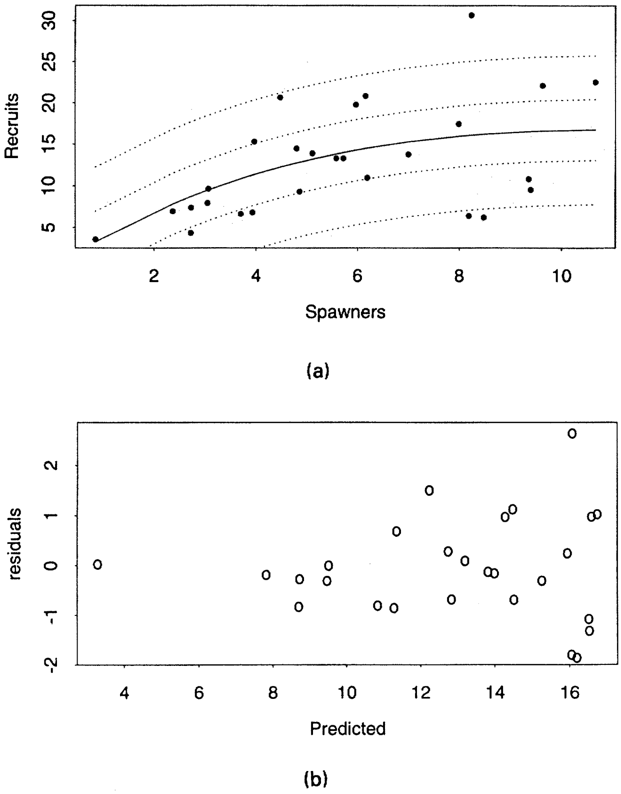
Over a period of 28 years the populations of spawning and recruit sockeye salmon were estimated for the Skeena River in British Columbia (Ricker and Smith, 1975). The intent of this study is to quantify the relationship between the number of spawning salmon and the resulting distribution of young fish that are recruited into the population. Let X_k denote the number of spawning salmon in a given year and let Y_k be the number of recruited salmon associated with the same year. A simple model proposed by Ricker (1954) to describe the relationship between these two variables is

$$Y = \beta_1 X \exp(-\beta_2 X).$$

Ricker's model is widely used for salmon stocks and appears to fit them well. This function is taken to be the parametric regression function for the median of the distribution of recruited salmon given a particular number of spawning fish. A scatterplot of these data suggests that, although the Ricker model is a reasonable choice for the median response, the variance of recruit salmon does not appear to be constant and the response is right skewed. One strategy is to use a TBS model to account for this apparent heteroscedasticity. The reader is referred to Carroll and Ruppert (1988) for more background on these data and a thorough parametric analysis. In that work it was found that an unshifted ($\lambda_2 = 0$) power transformation with $\lambda_1 = -0.2$ and a constant variance function ($V_k \equiv V > 0$) yielded residuals that were closer to being normally distributed and homoscedastic. One issue that will be addressed in this section is the sensitivity of the results to the choice of a parametric transformation.

Fig. 1 gives the results of fitting the Ricker model to the salmon count data by using non-linear least squares. Fig. 1(a) indicates the ordinary least squares fit and the implied prediction intervals. Under this model $P(Y \leqslant f(x, \beta) + \sigma Z_\alpha | X = x) = \alpha$ and approximate pointwise prediction intervals can be obtained by substituting estimates for β and σ in the expression for the conditional quantile. The dotted curves are the estimated 0.05, 0.25, 0.75 and 0.95 quantiles for the conditional distribution of recruit salmon given the number of spawners. The constant separation between these limits is due to the assumption of constant variance and no transformation. On the basis of the residual plot (Fig. 1(b)), however, it is clear that there are some discrepancies with this homoscedastic model.

Fig. 2 summarizes the results from reanalysing these data by using a more flexible transformation model. The first column of Fig. 2. compares prediction intervals



(a)

(b)

Fig. 1. Ordinary least squares estimates for the Skeena River sockeye salmon: the Ricker model has been fitted to the salmon counts (units are 100000 fish) by using non-linear least squares under the assumption of normally distributed errors with constant variance (———, estimated mean (or median) response; ········, approximate conditional quantiles at 0.05, 0.25, 0.75 and 0.95 based on additive errors); the outer dotted curves give an approximate 90% prediction interval for the mean number of recruit salmon for a given number of spawners (the residual plot suggests that this model is inadequate)

under different transformations. The first row shows the results for a simple power transformation and the subsequent rows are based on spline estimates of the transformation with increasing amounts of flexibility. In this model the log-transformation has been preloaded into the model. Because of this formulation, when $\rho = \infty$, the estimate is essentially a power transformation. (For these data the limiting properties of the smoothing parameter are achieved when ρ is of the order of 10^4 or larger.) The general trend as $\rho \rightarrow 0$ is towards a transformation that expands points sharply in the range below 5 and is nearly linear above 10. The effect of this type of transformation is prediction intervals that rapidly increase for small numbers of salmon (1–4) and are nearly constant for larger values. Note that this pattern in the prediction intervals is not as clearly captured by just a simple power transformation model. For this sequence of smoothing parameters the Spearman correlation between the absolute residuals and the predicted values were 0.445 ($\log_{10}\rho = 4$), 0.336 ($\log_{10}\rho = -1.1$) and 0.379 ($\log_{10}\rho = -1.14$). Thus the middle

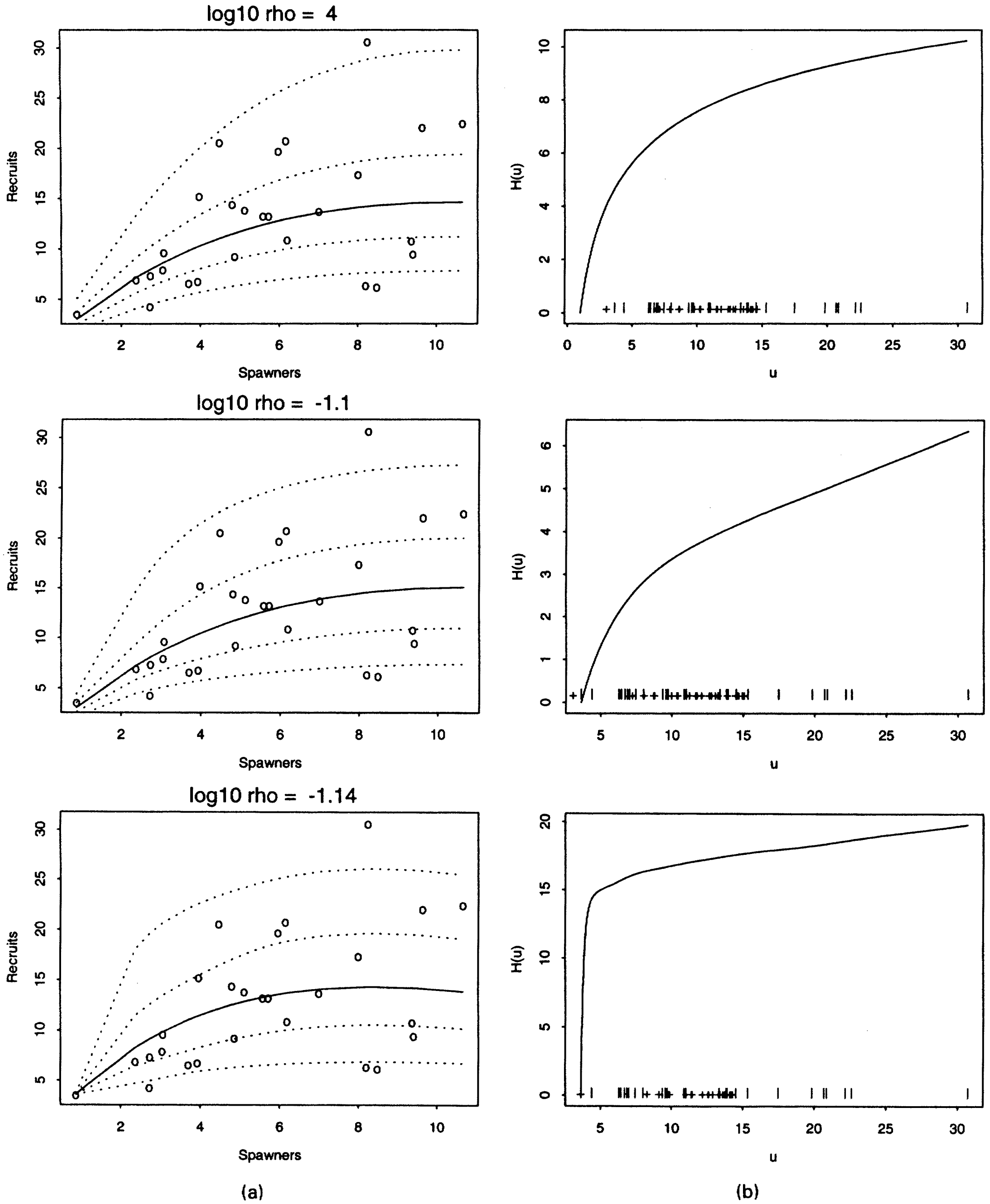


Fig. 2. Spline-based estimates of a normalizing transformation for the Skeena River sockeye salmon: (a) TBS model fitted to the salmon counts by using a spline estimate for the normalizing transformation (———, estimated median response; ·······, approximate conditional quantiles at 0.05, 0.25, 0.75, 0.95); (b) the estimated transformations applied to the regression model after an initial log-transformation (|, locations of the independent variables; +, predicted values); thus the transformations in the second column are the composition of the log-function and the transformation obtained from the spline estimate; for large ρ the spline-based transformation is essentially a power transformation in the original scale

plot, representing an intermediate amount of smoothing, may be more appropriate to reduce heteroscedasticity.

One issue is whether it is necessary to include a preliminary log-transformation to obtain reasonable results for the nonparametric estimate of H. When a preliminary log-transformation was not included the H yielded comparable prediction intervals. One departure, however, is that without an initial log-transformation the prediction intervals tend to be more symmetric about the median for large numbers of spawning fish. For small values of the smoothing parameter, the transformation in the original scale appears to have more bumps. One explanation for this structure is that, if the log-transformation is not preloaded, then to capture the sharp increase for small values this transformation must be tuned to have a large amount of flexibility. This flexibility also applies to other parts of the transformation that may be sensitive to the local influence of only a few data points.

5. EXISTENCE OF MAXIMIZER FOR PENALIZED TRANSFORM BOTH SIDES LIKELIHOOD

One basic question is whether the estimate of H based on the penalized likelihood even exists and if so whether it is unique. It is difficult to give general conditions that guarantee unique estimates of both and θ and H. However, just concentrating on step (2) in our iterative algorithm, maximization over H for fixed θ, it is possible to identify simple conditions when a maximizer will exist and be unique. From a practical point of view this is a reasonable simplification. Although we might tolerate a complicated likelihood surface for the parametric portion of the model, to simplify the computation of the transformation we would hope that the nonparametric part is well behaved.

Our analysis depends on the properties of the likelihood for transformations where $J(H) = 0$. Let $\mathscr{A} = \{g\text{: } g' \text{ absolutely continuous and } g'' \in L^2[a, b]\}$ and let $\mathscr{A}_0 = \{g \in \mathscr{A}\text{: } J(H) = 0 \text{ and } g = \log H'\}$. If either L_{P} or L_{PA} have maximizers in $\mathscr{A}_0$ then a unique maximum will exist in $\mathscr{A}$. This type of result was first suggested by Silverman (1982) for penalized likelihood problems and also appears in Cox and O'Sullivan (1990).

Uniqueness of the maximizers is implied by the strict convexity of the penalized likelihood (Tapia and Thompson (1978), p. 160).

Theorem 1. For fixed values of θ:

(a) $-L_{\mathrm{P}}(\theta, g)$ is a strictly convex functional of g;
(b) if W is such that $x > 0$ implies that $Wx > 0$ then $-L_{\mathrm{PA}}(\theta, g)$ is a strictly convex functional of g.

The proof of this theorem is given in Nychka and Ruppert (1992) and depends on showing that the second Gateux derivative is a strictly positive operator.

The next theorem is also proved in Nychka and Ruppert (1992) and gives a simple condition of existence of a transformation spline estimate.

Theorem 2. For fixed θ:

(a) if there is a maximizer of L_{P} over $\mathscr{A}_0$ then there will also be a maximizer of L_{P} over $\mathscr{A}$;

(b) if there is a maximizer of L_{PA} over $\mathscr{A}_0$ and the hypothesis in theorem 1, part (b), holds then there will also be a maximizer of L_{PA} over $\mathscr{A}$.

Recall that if the transformation is preloaded with a log-transformation then $\mathscr{A}_0$ is just the space of unshifted power transformations. From this perspective it is reasonable to require existence of this simple estimate as a condition for existence of the more general nonparametric transformation estimate.

ACKNOWLEDGEMENT

This work was supported by National Science Foundation grants DMS-8715756 and DMS-9002791.

APPENDIX A: APPROXIMATION TO PENALIZED LIKELIHOOD ESTIMATING EQUATIONS

Given an initial value for g, say g_0, the approximation to equation (3.8) will be derived and the pseudovalues and the weights used for the smoothing spline algorithm will be specified. The specification of Z_j and w_j depends on Ω_{jj} and D_j and the relevant cases are examined below.

A.1. *Case 1: $\Omega_{jj} = 0$*

If $\Omega_{jj} = 0$ set $Z_j = D_j + g_{j0}$ and $w_j = 1$.

A.2. *Case 2: $\Omega_{jj} > 0$, $D_j = 0$*

For case 2, equation (3.7) in expanded notation is

$$\exp(2g_j)\Omega_{jj} + \exp g_j \sum_{j \neq l}^{N} \Omega_{jl} \exp g_l + 2\rho[R\mathbf{g}]_j = 0, \qquad 1 \leqslant j \leqslant N. \qquad \text{(A.1)}$$

First diagonalize this system by replacing g_l by g_{l0} in the second term. The linearization is accomplished by using Taylor series expansion about g_{j0}: $\exp(2g_j) \approx \exp(2g_{j0})\{1 + 2(g_j - g_{j0})\}$ for the first term in equation (A.1). For the second term in equation (A.1) $\exp g_{j0}$ is substituted for $\exp g_j$. This series of simplifications leads to the approximate set of equations

$$\exp(2g_{j0})\{1 + 2(g_j - g_{j0})\}\Omega_{jj} + \exp g_{j0} \sum_{j \neq l}^{N} \Omega_{jl} \exp g_{l0} + 2\rho[R\mathbf{g}]_j = 0, \qquad 1 \leqslant j \leqslant N.$$

Recall that $h_j \equiv \exp g_j$ and collecting terms

$$-2h_{j0}^2\Omega_{jj}\left(-\frac{1}{2h_{j0}\Omega_{jj}}[\Omega h_0]_j + g_{j0} - g_j\right) + 2\rho[R\mathbf{g}]_j = 0, \qquad 1 \leqslant j \leqslant N. \qquad \text{(A.2)}$$

This expression is now in the form $-2w_j(Z_j - g_j) + 2\rho[R\mathbf{g}]_j = 0$ and therefore w_j and Z_j are set to the two expressions within parentheses.

A.3. *Case 3: $\Omega_{jj} > 0$, $D_j > 0$*

Rearranging equation (3.8)

$$\exp(2g_j)\left\{\Omega_{jj} + \exp(-g_j)\sum_{j\neq l}^{N}\Omega_{jl}\exp g_l - \exp(-2g_j)D_j\right\} + 2\rho[R\mathbf{g}]_j = 0, \qquad 1\leqslant j\leqslant N. \tag{A.3}$$

Following a similar strategy to that in case 2, one can derive the following linearization:

$$-2D_j\left(\frac{-h_{j0}[\Omega h_0]_j}{2D_j + \frac{1}{2} + g_{j0}} - g_j\right) - 2\rho[R\mathbf{g}]_j = 0, \qquad 1\leqslant j\leqslant N. \tag{A.4}$$

On the basis of equation (A.4) one would set $w_j = D_j$ and Z_k to be the first term within the parentheses.

REFERENCES

Becker, R., Chambers, J. and Wilks, A. (1988) *The New S Language*. Pacific Grove: Wadsworth and Brooks/Cole.

Breiman, L. and Friedman, J. (1985) Estimating optimal transformations for multiple regression and correlation. *J. Am. Statist. Ass.*, **80**, 580–597.

Carroll, R. J. and Ruppert, D. (1988) *Transformation and Weighting in Regression*. New York: Chapman and Hall.

Cole, T. J. and Green, P. J. (1992) Smoothing reference centile curves: the LMS method and penalized likelihood. *Statist. Med.,* **11**, 1305–1319.

Cox, D. D. and O'Sullivan, F. (1990) Asymptotic analysis of penalized likelihood and related estimators. *Ann. Statist.*, **18**, 1676–1698.

Eubank, R. L. (1988) *Spline Smoothing and Nonparametric Regression*. New York: Dekker.

Green, P. J. (1988) Discussion on Fitting smoothed centile curves to reference data (by T. J. Cole). *J. R. Statist. Soc.* A, **151**, 410–411.

Hutchinson, M. F. and de Hoog, F. R. (1985) Smoothing noisy data with spline functions. *Numer. Math.*, **47**, 99–106.

Nychka, D. and Ruppert, D. (1992) Nonparametric transformations for both sides of a regression equation. *Mimeo 2219*. Department of Statistics, North Carolina State University, Raleigh.

O'Sullivan, F., Yandell, B. and Raynor, W. J. (1986) Automatic smoothing of regression functions in generalized linear models. *J. Am. Statist. Ass.*, **81**, 96–104.

Ricker, W. E. (1954) Stock and recruitment. *J. Fish. Res. Bd Can.*, **32**, 559–623.

Ricker, W. E. and Smith, H. D. (1975) A revised interpretation of the Skeena River sockeye salmon. *J. Fish. Res. Bd Can.*, **32**, 1369–1381.

Silverman, B. W. (1982) On the estimation of a probability density function by maximum penalized likelihood method. *Ann. Statist.*, **10**, 795–810.

Tapia, R. A. and Thompson, J. R. (1978) *Nonparametric Probability Density Estimation*. Baltimore: Johns Hopkins University Press.

Tibshirani, R. (1988) Estimating transformations for regression via additivity and variance stabilization. *J. Am. Statist. Ass.*, **83**, 394–405.