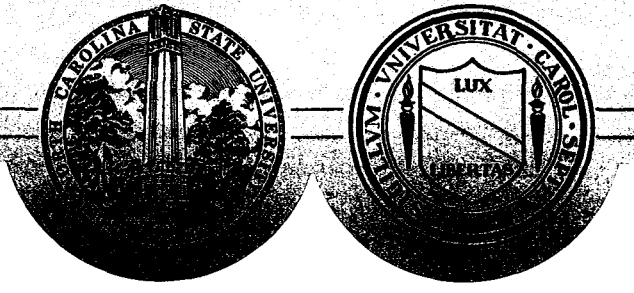
THE INSTITUTE OF STATISTICS

UNIVERSITY OF NORTH CAROLINA SYSTEM



Nonparametric Transformations for Both Sides of a Regression Model

by

Douglas W. Nychka and David Ruppert

Institute of Statistics Mimeograph Series No. 2219

April 1992

NORTH CAROLINA STATE UNIVERSITY
Raleigh, North Carolina

Nonparametric Transformations for Both Sides of a Regression Model

Douglas Nychka and David Ruppert

April 16, 1992

Abstract

One way to model heteroscedasticity and skewness of the error distribution in regression is to transform both sides (TBS) of the regression equation. If it is possible to transform the regression equation to result in normally distributed errors, then one 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 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 using nonparametric methods based on maximizing a penalized likelihood function. By expressing the likelihood in terms of the log derivative transformation we are able to derive an approximate maximum penalized likelihood estimate that has the form of a spline function. A fast algorithm for computing this estimate is introduced and applied to two data sets where a TBS model has been found to work well. The results based on estimating a spline transformation give some insight into the sensitivity of prediction intervals to the choice of the transformation. Some results are also given concerning the existence and unique of the transformation spline estimate.

1. Introduction

The presence of heteroscedasticity or non-normal errors is a typical problem encountered in regression analysis. Consider the model $Y_k = f(X_k,\beta) + e_k \quad 1 \le k \le N$ where X_k and Y_k are independent and dependent variables related by a parametric function $f(\cdot, \beta)$. The random components $\{e_k\}$, $1 \le k \le N$ are assumed to be independent, have a median of zero 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. Nonlinear 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. In this case transforming just the dependent variable destroys the functional relationship between the dependent variable and the model predictions. A natural solution to this problem is to transform both sides (TBS) of the regression equation. That is, the transformed dependent variable is modeled by the transformed prediction equation. Letting H denote a transformation, the TBS model is:

$$H(Y_k) = H\Big(f_k(\beta)\Big) + \epsilon_k.$$

where 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 independent 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 $u+\lambda_2 \geq 0$.

$$H(u,\lambda) = \frac{(u+\lambda_2)^{\lambda_1} - 1}{\lambda_1}$$

Since the choice of transformation is largely empirical it is important to consider the sensitivity of the estimated parameters to H. One problem with using parametric transformations is the difficulty in

extending H beyond a 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 create more complicated transformations based on parametric expressions, we believe 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, one can estimate the parameters by maximum likelihood. 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

$$L_p(\theta, H) = L(\theta,H) - \rho J(H), \quad \rho > 0 \ . \tag{1.3}$$

Here J(H) is a positive functional that quantifies the roughness of the transformation while ρ is the relative weight between the roughness penalty and the unconstrained log likelihood of the data. A nonparametric estimate of H is obtained by maximizing $L_p(H,\theta)$ over all H such that $J(H) < \infty$. 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 infinity H will have the form of a simple parametric function, such as a power transformation. As ρ approaches zero, H will tend to a staircase-like function with jumps at the points $\{Y_k\}$. Of course in practice one would expect the best choice for H to lie somewhere in between these two extremes.

By focusing on maximum penalized likelihood estimates (MPLE) it is natural to formulate estimates of H in terms of smoothing splines. In order to insure 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 and was applied to two non-normal data sets that have been analyzed extensively. The estimated transformations 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 several examples and these results are reported in Section 4. Section 5 discusses the existence of maximizers of the penalized likelihood for the TBS model and Section 6 draws some conclusions and suggests some possible improvements to the algorithm.

2. Maximizing the TBS Penalized Likelihood

2.1 TBS Likelihood and the Roughness Penalty

Under the assumptions discussed in the introduction the log likelihood for the observations is given by

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

As mentioned in the introduction the penalty function for H is defined through the log derivative. Set $g(u) = \log(H'(u))$. One reason for using this representation is to avoid the necessity of a constrained maximization where a monotonicity constraint is enforced on H. By expressing the penalized likelihood with respect to g one 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

$$(2.2) \qquad J(H) = \int_a^b (g'')^2 du$$

where a and b are chosen such that $\{Y_k\} \subset [a,b]$. In terms of g, this penalty quantifies the overall curvature of the function and is acknowledged to be reasonable measure of smoothness for a curve. Using

this parametrization for the transformation, the penalized likelihood is:

$$(2.3) \qquad L_P(\theta,H) = \tfrac{1}{2} \sum_{k=1}^{n} \left\{ -\left(\int_{A_k} e^{g(u)} du \right)^2 / V_k(\delta) + 2g(Y_k) - \log(V_k(\delta)) \right\} - \rho \int_a^b (g'')^2 du + C$$

where A_k is the interval with endpoints 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 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

$$(2.4) \qquad H(u) - H(u_0) = (e^{\gamma_1}/\gamma_2)\,(e^{\gamma_2 u} - e^{\gamma_2 u_0}) \ .$$

This parametric family may not seem unusual if one initially transforms both sides of the regression equation using the log and then considers H as an additional transformation. In this case u is actually in a log scale with respect to the original observations. Setting $u = \log(\omega)$, if H is given by (2.4) then $H(u) = H\big(\log(\omega)\big) = C_1 \omega^{\gamma_2} + C_2$ where C_1 and C_2 depend on $\underset{\sim}{\gamma}$ 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 the 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 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 k^{th} iteration.

Given starting estimates θ_0, $H_0(u) = u$

Do {

(1) θ_{k+1} = maximizer of $L_p(\theta, H_k)$ for $\theta \epsilon \mathbf{R}^p$

(2) H_{k+1} = maximizer of $L_p(\theta_{k+1}, H)$ for $J(H) < \infty$

}

until (convergence)

Note that 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 is maximized 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: $\Sigma \left(H(Y_k) - f(X_k)\right)^2$ subject to the variance of $\{H(Y_k)\}$ being equal to one. 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 preserve a model based upon scientific theory. The algorithm for computing these estimates is similar to that outlined above in that one alternates between optimizing with respect to the regression estimate and the transformation.

A slightly different approach to modeling 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 these authors estimate these functions 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.

In this subsection we have briefly reviewed some other approaches to nonparametric transformations. The main point is that these methods have goals that differ substantially from that of spline-TBS. Spline-TBS, ACE, AVAS, and LMS all have distinct and important roles in data analysis.

3. Estimating a nonparametric transformation

3.1 Smoothing Splines

Splines are popularly described as piecewise polynomial curves that have good approximation properties. In this work, however, it is useful to characterize a cubic smoothing spline as the solution to a variational problem. Consider the nominal smoothing problem $Z_k = g(u_k) + e_k$, for $1 \le k \le N$ 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

$$(3.1) \qquad \mathcal{L}(g) = \frac{1}{N}\sum_{k=1}^{N}\left(Z_k - g(u_k)\right)^2 w_k + \rho\int_a^b (g'')^2 du.$$

over all g such that $\int_a^b (g'')^2 du < \infty$. If there exist at least three distinct values in $\{u_k\}$ then $\hat{g}$ exists and is unique. Moreover, $\hat{g}$ will have the form of a piecewise cubic polynomial with join points at the unique values of $\{u_k\}$. The coefficients of these polynomial segments are constrained so that $\hat{g}$ will have a

continuous first and second derivative and $\hat{g}$ satisfies the "natural" boundary conditions

$$\hat{g}^{(2)}(a) = \hat{g}^{(2)}(b) = \hat{g}^{(3)}(a) = \hat{g}^{(3)}(b) = 0 .$$

(The extrapolation of $\hat{g}$ beyond [a, b] is linear.) Because of the number of constraints on the polynomial segments, the spline solution can be parametrized by the value of the function at the points $\{u_k\}$. For the moment assume that these points are unique and let $\underset{\sim}{g}^T = \{g(u_1), \ldots, g(u_N)\}$. Using this parametrization and the knowledge of the functional form of the solution, the minimization problem at (3.1) is equivalent to

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

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

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

Therefore to compute a smoothing spline one only needs to solve the linear system at (3.3). Finding the solution to (3.3) is an efficient computation. By a judicious transformation (second divided differences), R can be reduced to a symmetric, banded matrix. There are only two nonzero off-diagonal bands for cubic splines and thus the operation count to solve the system is $O(N)$. 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. Once the values of $\hat{g}$ are known it is straight forward to compute the coefficients of the cubic polynomial segments in between these knots. In this way the entire function may be recovered just from the estimate of the vector $\underset{\sim}{g}$.

An important feature of this characterization of a spline is that the abstract minimization problem given at (3.1) can be related to solving a finite dimensional problem. In fact 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: $\{g(u_k)\}$ $1 \le k \le N$. If $\left[Q(g) + \rho \int_a^b (g'')^2 du\right]$ has a minimizer over g such that $\int_a^b$

$(g'')^2 du < \infty$, 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, Yandell and Raynor (1986) and this fact will be used in characterizing the spline estimate of the transformation.

3.2 An Approximate Penalized Likelihood

It is difficult to maximize the penalized likelihood exactly due to 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 e^g and our numerical strategy is to approximate these integrals by sums. For notational convenience let $f_k \equiv f_k(\underset{\sim}{\theta}) \equiv f_k(\beta)$ and $V_k \equiv V_k(f)$. Choose a sufficiently fine mesh of points $u_1 \leq u_2 \cdots < u_N$ 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

$$(3.4) \qquad \int_{f_k}^{Y_k} e^{g(u)} du \approx \sum_{j=1}^{N} W_{kj}\, e^{g_j}$$

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

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

Note that the mesh size is arbitrary and can be chosen to achieve any desired level of accuracy for these approximations. Since g is anticipated to be a smooth function, however, we believe that the quadrature formula will be accurate even for modest mesh sizes. By the same arguments we also expect the interpolation error to be small. In fact, if the observations $\{Y_k\}$ can be included as a subset of the mesh then the interpolation error is zero.

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

$$(3.6) \qquad L_{PA}(g) = \sum_{k=1}^{n} \left\{ \left(\sum_{j=1}^{N} W_{kj} e^{g(u_j)} \right)^2 / 2V_k + \sum_{j=1}^{N} \delta_{kj} g(u_j) \right\} - \rho \int_a^b (g'')^2 du.$$

If one identifies the first expression with the nonlinear 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 one based on the integrated, squared second derivative. Thus $L_{PA}(g)$ has the form $-\left[Q(g) + \rho \int_a^b (g'')^2 du\right]$ and from the remarks in Section 3.1 the maximizer of L_{PA} for $\int_a^b (g'')^2 du < \infty$, if it exists, will be a piecewise cubic polynomial with knots at $\{u_j\}$ satisfying the usual continuity and boundary constraints. Moreover, because the functional form of the solution is known, it is enough to identify the solution at the mesh points. For this reason, it will be useful to reexpress (3.7) in vector notation. Let $\underset{\sim}{g} = \left(g(u_1), \cdots, g(u_N)\right)^T$ and $h_j = \exp\{g_j\}$ then

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

where $\Omega = W^T \mathrm{diag}(\underset{\sim}{V})^{-1} W$, $D_j = \sum_{k=1}^{N} \delta_{kj}$ and R is the roughness penalty matrix as described in Section 3.1.

At this point the advantage of the approximate likelihood is clear. The maximization over a function space has been reduced to a maximization over $\Re^N$. It straight forward to take partial derivatives of (3.7) and setting these equal to zero gives a system of nonlinear equations that are necessary for any extremal point of the approximate penalized likelihood.

$$(3.8) \qquad \frac{\partial}{\partial g_j} L_{PA}(\underset{\sim}{g}) = -h_j [\Omega \underset{\sim}{h}]_j + D_j - 2\rho [R \underset{\sim}{g}]_j \qquad 1 \le j \le N$$

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

3.3 Maximizing L_{PA} for fixed θ

Although the transformation spline is a nonlinear function of the observed data, it is possible to linearize the defining system about a previous estimate so that it resembles (3.3). This suggests an iterative

$(g'')^2 du < \infty$, 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, Yandell and Raynor (1986) and this fact will be used in characterizing the spline estimate of the transformation.

3.2 An Approximate Penalized Likelihood

It is difficult to maximize the penalized likelihood exactly due to 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 e^g and our numerical strategy is to approximate these integrals by sums. For notational convenience let $f_k \equiv f_k(\underset{\sim}{\theta}) \equiv f_k(\beta)$ and $V_k \equiv V_k(f)$. Choose a sufficiently fine mesh of points $u_1 \leq u_2 \cdots < u_N$ 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

$$(3.4) \qquad \int_{f_k}^{Y_k} e^{g(u)} du \approx \sum_{j=1}^{N} W_{kj}\, e^{g_j}$$

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

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

Note that the mesh size is arbitrary and can be chosen to achieve any desired level of accuracy for these approximations. Since g is anticipated to be a smooth function, however, we believe that the quadrature formula will be accurate even for modest mesh sizes. By the same arguments we also expect the interpolation error to be small. In fact, if the observations $\{Y_k\}$ can be included as a subset of the mesh then the interpolation error is zero.

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

$$(3.6) \qquad L_{PA}(g) = \sum_{k=1}^{n} \left\{ \left(\sum_{j=1}^{N} W_{kj} e^{g(u_j)} \right)^2 / 2V_k + \sum_{j=1}^{N} \delta_{kj} g(u_j) \right\} - \rho \int_a^b (g'')^2 du.$$

If one identifies the first expression with the nonlinear functional $-Q(g)$ then it is clear that Q will

algorithm where at each stage efficient cubic smoothing spline algorithms are used to compute a solution to an approximate, linear system.

From the discussion given above it is sufficient to solve the system of equations

(3.9) $\quad h_j[\Omega \underset{\sim}{h}]_j - D_j + 2\rho[R\underset{\sim}{g}]_j = 0 \quad 1 \le j \le N$ for

$\underset{\sim}{g} \in \Re^N$. The computational strategy is to approximate the first two terms of (3.9) so that they can be expressed as $-2(Z_j - g_j)w_j$ (c.f. (3.3)) for some choice of pseudo data, $\underset{\sim}{Z}$, and weights, $\underset{\sim}{w}$. This process involves two steps, a diagonalization of the system so that the j^{th} equation only depends on h_j and $[R\underset{\sim}{g}]_j$ and a linearization of h_j with respect to g_j. The exact of specification of these approximations is given in Appendix A. With this linearization/diagonalization of (3.9) the following algorithm is proposed for solving the system.

Determine quadrature mesh $\{u_j\}$ $1 \le j \le N$
Compute Ω and $\underset{\sim}{D}$ $\quad \Omega = W' \mathrm{diag}(V)^{-1} W$
Initialize: $g^{NEW} \equiv 0$
Do {
$\quad g^{OLD} = g^{NEW}$
$\quad$ Compute $\underset{\sim}{Z}$ and $\underset{\sim}{w}$ based on g^{OLD}, Ω, $\underset{\sim}{D}$.

$\quad$ Compute g^{NEW}, a cubic smoothing spline for $\{u_j, Z_j\}$ with weights $\{w_j\}$.
}
Until (convergence)

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

3.4 Implementation of the algorithm in S

The TBS model specification can be complicated especially when a variance function is also present. Also, it is difficult to evaluate the effect of the transformation without consulting residual plots and prediction intervals. For these reasons the top level of the algorithm and the user interface have been

implemented in the S programming environment (Becker, Chambers and Wilks 1988). The model specification is handled by creating an S function that returns the likelihood and gradient when called with particular parameter values and a transformation. The Fisher method of scoring was written in the S programming language by adapting the nonlinear regression S function distributed by Bates and Watts (1988). The algorithm for estimating the transformation was written as a FORTRAN subroutine and called from within S. In specifying the approximate likelihood the quadrature mesh was taken as the union of 150 equally spaced points over the range of Y 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 $\{Y_k\}$. To make the iterative algorithm to find the transformation more efficient, 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 rapidly enough 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: $g_{OLD} + d(g_{NEW} - g_{OLD})$ where $0 < \alpha < 1$.

4. Two application of transformation splines

4.1 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:

$$Y = \beta_1 X e^{-\beta_2 X}.$$

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 while 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 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 a unshifted power transformation with $\rho = -.2$ 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.

Figures 1 and 2 summarize our results from reanalyzing these data using a more flexible transformation model. The first row of this figure illustrates an ordinary nonlinear least squares fit to these data. Under this model $P(Y \leq f(x,\beta) + \sigma Z_\alpha \mid X = x) = \alpha$ and approximate pointwise prediction intervals can be obtained by substituting estimates for β and σ in the expression for the conditional quantile. The dashed lines are the estimated .05, .25, .75 and .95 quantiles for the conditional distribution of recruit salmon given the number of spawners. The constant separation among these limits is due to the assumption of constant variance in the regression model and no transformation. The residual plot, however, indicates some discrepancies with this homoscedastic model.

The last three rows of Figure 1 compare prediction intervals under different transformations. The second 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 H is a power transformation and when $\rho = 10^4$, the estimate is essentially a power transformation. The general trend as $\rho \rightarrow 0$ is toward a transformation that expands points sharply in the range below 5 and is nearly linear above 10. The impact 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.

One issue is whether it is necessary to include a preliminary log transformation to obtain reasonable results for the nonparametric estimate of H. Figure 2 displays the results for a TBS model without an initial log transformation. These estimated models should be compared to rows 2-4 of Figure 1 and to improve the comparison the smoothing parameters were selected to give estimates with similar degrees of smoothness. Overall the two nonparametric estimates of H yield 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 the smallest value 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 in order 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.

4.2 Clearance By the Lung of Bacteria

The defense mechanisms of the lung to inhaled bacteria can be investigated by exposing mice to bacteria under controlled conditions. A particular experiment reported in Ruppert et al. (1975) has the form of a 4×3 factorial design: the first factor being the type of exposure (Control, Ampicillin treated, Virus-infected, Virus-infected and ampicillin treated) and the second factor being the time of sacrifice after exposure (0, 4 hours and 24 hours). There were 6 replicates for each combination of the factor levels and the dependent variable is related to the number of living bacteria in the lung at time of sacrifice. Specifically, the measured response is the number of viable bacteria expressed as a percentage of the total number that entered the lung during exposure. The number of viable bacteria observed under the different experimental conditions vary over several orders of magnitude. Even to plot these data some nonlinear scaling is required and just from a practical point of view it is useful to work from logged values. One difficulty is that one response is zero and therefore before applying a power or log transformation a small shift was added to all

bacteria counts. Ruppert and Carroll analyzed these data using a SP transformation with the shift fixed at .05. In this previous work (Ruppert and Carroll 1988) several models for the median response were developed based on biological considerations. Examination of the residuals also suggested the need for a variance model for the transformed regression equation. Let Y_{itj} denote the shifted and logged clearance for the j^{th} mouse in i^{th} exposure group sacrificed at the t^{th} time point. One model suggested in Carroll and Ruppert (1988), pg. 159 and 166 is

$$\text{Median}(Y_{itj}) = \alpha_i + \beta_i(t+.25) + \Delta_i(t-4)+$$

where $(u)_+ = u$ for $u > 0$ and 0 for $u \leq 0$, and with variance function

$$V_{ijt} = \delta + (t + .25)^2$$

Figure 3 is a summary of estimated models for the lung clearance data for a range of transformations. The first plot in each row of this figure illustrates the dependence of the width of (approximate) 90% prediction intervals on the level of the prediction. The width of these intervals is not monotonically increasing with respect to the predicted value due to the effect of the variance function. The first row is the result of fitting a TBS model based on a shifted (.05) and logged transformation and the next three rows are the estimated models associated with more flexible choices for H. There is some difference in the prediction intervals' widths, between the untransformed response and the transformed models. But as the transformation becomes more flexible the prediction intervals stay nearly the same with the exception of the largest prediction values. The logged response has residuals with some dependence on the the size of the predicted value. This effect is removed with more adaptive transformations. Although the residual plots using a spline transformation differ from the fit to the logged data, they are not sensitive to the increased flexibility of the transformation. The main difference among these last three rows is that the most negative residual in the group near 0 is pulled in toward 0 as the flexibility of the transformation increases.

Initially a small shift was added to data in order to facilitate a log transformation and it is of interest

to investigate the sensitivity of these results to the size of the shift. Figure 3 displays the results of estimating a nonparametric TBS model when a shift of .02 is used rather than .05. When the smoothing parameter is large the residual plots differ slightly with respect to the most negative residual in the lower range of predicted values. But as the smoothing parameter is decreased the transformation has the flexibility to adapt to a different shift. The result is that the residuals for $\log_{10} \rho = -1$ or -1.05 are nearly the same whether a shift of either .02 or .05 is added to the original data.

5. Existence of a maximizer for a penalized 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 one might tolerate a complicated likelihood surface for the parametric portion of the model, to simplify the computation of the transformation one 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 $\mathfrak{H} = \{g: g, g' \text{ absolutely continuous and } g'' \in L^2[a,b] \}$ and let $\mathfrak{H}_0 = \{g: J(H)=0 \text{ where } g= \log(H')\}$. If either L_P or L_{PA} have maximizers in $\mathfrak{H}_0$ then a unique maximum will exist in $\mathfrak{H}$. This type of result was first suggested by Silverman (1982) for penalized likelihood problems and also appears in Cox and O'Sullivan (198?).

5.1 Uniqueness

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

Theorem 5.1

For fixed values of θ,

i) $-L_P(\theta,g)$ is a strictly convex functional of g

ii) If W is such that $\underset{\sim}{x} > 0$ implies that $W\underset{\sim}{x} > 0$ then $-L_{PA}(\theta,g)$ is a strictly convex functional of g.

Proof

For $g,\phi,\psi \in \mathcal{H}$ it is straight forward to verify that the second order Gâteaux derivative of $-L_P$ at g is

$$D^2[-L_P(g);\phi,\psi] = \sum_{k=1}^{n}\left[\int_{A_k}\phi\psi e^g\,du \int_{A_k} e^g\,du + \int_{A_k}\phi e^g\,du \int_{A_k}\psi e^g\,du\right] \Big/ V_k$$

$$+ 2\rho\int_{[a,b]}\phi''\psi''du$$

where A_k is the interval with endpoints at f_k and Y_k. Since $D^2[-L_P(g);\phi,\phi] > 0$ for all $\phi \in \mathcal{H}$ it follows that $-L_P$ must be strictly convex (Tapia and Thompson 1978, pg. 157).

Considering $-L_{PA}$, after some simplification one can show that

$$D^2[-L_{PA}(g);\phi,\phi] = 2\underset{\sim}{h}^T\Omega\Phi^2\underset{\sim}{h} + 2\underset{\sim}{h}^T\Phi\Omega\Phi\underset{\sim}{h} + 2\rho\int_{[a,b]}\phi''\phi''du$$

where $\Phi_{jj} = \phi(u_j)$, $\Phi_{jk} = 0$ for $j \neq k$, and $h_j = e^{g(u_j)}$. Recall that $\Omega = W^TW$ and it follows from the assumptions on W that $\underset{\sim}{h}^T\Omega\Phi^2\underset{\sim}{h} = (W\underset{\sim}{h})^T(W\Phi^2\underset{\sim}{h}) > 0$. Thus $D^2[-L_{PA}(g);\phi,\phi] > 0$ for all g, ϕ in $\mathcal{H}$. □

5.2 Existence

Before giving specific results for the TBS estimate we state some general criteria that guarantee the existence of a solution to minimizing a functional over a convex space (Nashed, 1971,Theorem 11.5).

Suppose that $\mathfrak{B}$ is a reflexive Banach space and ℓ is a functional that satisfies the following conditions

A) ℓ is lower semicontinuous

B) There is a $g^* \in \mathfrak{B}$ and $K < \infty$ such that $\inf \{ \ell(g) : \|g\| > K \} \geq \ell(g^*)$

then the infimum of ℓ will exist and be attained in $\mathfrak{B}$.

Theorem 5.2

For fixed θ,

i) If there exists a maximizer of L_P over $\mathfrak{H}_0$ then there will also exist a maximizer of L_P over $\mathfrak{H}$.

ii) If there exists a maximizer of L_{PA} over $\mathfrak{H}_0$ and the hypothesis in Theorem 5.1 ii) holds then there will also exist a maximizer of L_{PA} over $\mathfrak{H}$.

The proof is given in Appendix B and is based on verifying conditions A and B cited above. Recall that if the transformation is preloaded with by a log transformation then $\mathfrak{H}_0$ 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.

6. Discussion

Nonparametric estimates of the transformation in the TBS model have provided useful information in interpreting the Skeena River salmon counts and the lung clearance experiment. In the first example considering a richer family of transformations changes the shape of the prediction intervals. Rather than continuing to increase with increasing predicted value the intervals tend to have similar width for large spawning population sizes. In contrast in the lung clearance data there is little dependence on the prediction intervals on more flexible transformations. These results suggest a shifted power transformation is adequate to model the distribution of the errors. One advantage of the nonparametric approach is that the

transformation is not sensitive to the choice of the shift parameter.

The linearization of the estimating equations is admittedly *ad hoc* and unfortunately we can not give any arguments for the algorithm's convergence. We should emphasize however, that if our algorithm does converge then one does obtain an extremum of the approximate penalized likelihood. Moreover if one focuses just on the step of estimating the transformation we have identified fairly weak conditions where a unique MPLE exists. An alternative to our scheme of linearizing the system at (3.6) is to carry out Newton's Method directly. Rather than inverting a Hessian matrix at each step, however, one could solve the linear system approximately using an iterative method. This procedure might be made more efficient by reparametrizing the spline function as a linear combination of B-spline basis functions and taking advantage of these basis functions' finite support.

One advantage of considering a likelihood based approach is that it suggests a strategy for inference. To derive confidence intervals (or regions) for the parametric part of the model one might consider concentrating the likelihood by maximizing with respect to the nonparametric part. Along these lines, an approximate (1-α) confidence set for a specific parameter say θ_k might be

$$\{\theta_k: 2\left[L_p(\hat{\theta},\hat{H}) - \max_{\theta_j, j \neq k,\ H} L_p(\theta,H)\right] \leq \chi^2_\alpha(1)\}.$$

Of course the validity of such a set depends on the sampling properties of the penalized likelihood and more research is needed on this question. Cox and O'Sullivan (1990) have established some asymptotic properties for penalized likelihood estimates and we believe that by verifying several regularity condition their results also apply to our estimator. Given the flexibility of penalty methods for estimating nonparametric parts of a statistical model, we hope this approach will be applied to other problems that depend on estimating a function.

Appendix A – Approximation of the penalized likelihood equations

Given an initial value for $\underset{\sim}{g}$, say $\underset{\sim}{g}_0$, the approximation to (3.9) will be derived and the pseudo values 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.

Case 1 $\Omega_{jj} = 0$

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

Case 2 $\Omega_{jj} > 0,\ D_j = 0$

For this case (3.7) in expanded notation is

$$e^{2g_j}\Omega_{jj} + e^{g_j}\sum_{j\neq l}^{N}\Omega_{jl}e^{g_l} + 2\rho[R\underset{\sim}{g}]_j = 0 \qquad 1 \le j \le N \ . \tag{A.1}$$

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

$$e^{2g_{j0}}\big(1 + 2(g_j - g_{j0})\big)\ \Omega_{jj} + e^{g_{j0}}\sum_{j\neq l}^{N}\Omega_{jl}e^{g_{l0}} + 2\rho[R\underset{\sim}{g}]_j = 0 \qquad 1 \le j \le N \ .$$

Recall that $h_j \equiv e^{g_j}$ and collecting terms

$$-2\big(h_{j0}^2\Omega_{jj}\big)\left[\left(\frac{-1}{2h_{j0}\Omega_{jj}}[\Omega\underset{\sim}{h}_0]_j + g_{j0}\right) - g_j\right] + 2\rho[R\underset{\sim}{g}]_j = 0 \qquad 1 \le j \le N \ . \tag{A.2}$$

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

Case 3 $\Omega_{jj} > 0,\ D_j > 0$

Rearranging (3.9)

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

Following a similar strategy to that in Case 2, $\exp(2g_j)$ is replaced by $\exp(2g_{j0})$, $\exp(g_l)$ is replaced by $\exp(g_{l0})$, and the linearization: $\exp(-2g_j) \approx \exp(-2_{j0})\left(1 - 2(g_j - g_{j0})\right)$ is used. These approximations yield the system

$$e^{2g_{j0}}\left[\Omega_{jj} + e^{-g_{j0} \sum_{j \neq l}^{N} \Omega_{jl} e^{g_l}} - e^{-2g_{j0}\left(1 - 2(g_j - g_{j0})\right)} D_j\right] + 2\rho[R\underset{\sim}{g}]_j = 0 \qquad 1 \le j \le N \ .$$

Simplifying

$$(A.4) \qquad -2D_j\left[\left(-\frac{h_{j0}[\Omega\underset{\sim}{h}_0]_j}{2D_j} + 1/2 + g_{j0}\right) - g_j\right] - 2\rho[R\underset{\sim}{g}]_j = 0 \qquad 1 \le j \le N \ .$$

Based on (A.4) one would set $w_j = D_j$ and Z_k to be to the term within the parentheses.

References

Bates, D. and Watts, D. (1988), *Nonlinear Regression Analysis and its Applications*, John Wiley and Sons, New York.

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

Breiman, L. and Friedman, J. (1985), "Estimating Optimal Transformations for Multiple Regression and Correlation", JASA 80 580-597.

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

Cole, T. J. and Green P. J. (1992), "Smoothing Reference Centile Curves: the LMS Method and Penalized Likelihood", manuscript

Cox, D. D. and O'Sullivan, F. (1990), " Asymptotic Analysis of Penalized Likelihood and Related Estimators", *Annals of Statistics* 18 1676-1698.

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

Green, P. J. (1988) , "Discussion to: Fitting Smoothed Centile Curves to Reference Data by J. Cole ", *Journal of the Royal Statistical Society Series A* 151 410-411.

Hutchinson, M. F and de Hoog, F.R. (1985), "Smoothing Noisy Data with Spline Functions", *Numerische Mathematik* 47 99-106.

Nashed, M.Z. (1971), "The Role of Differentials", in *Nonlinear Functional Analysis* L.B. Rall ed., Academic Press, New York

Ricker, W.E. and Smith, H. D. (1975), "A Revised Interpretation of the Skeena River Sockeye Salmon", *J. Fish. Res. Board Can.* 32 1369-1381.

Ricker, W. E. (1954), "Stock and Recruitment", *J. Fish. Res. Board Can.* 32 559-623.

Ruppert, D., Jakab, G.J., Sylwester, D.L. and Green, G.M. (1975), "Sources of Variance in the Measurement of Intrapulmonary Killing of Bacteria", *J. Lab. Clin. Med.* 87 544-558.

O'Sullivan, F. , Yandell, B. and Raynor, W. J. (1996), "Automatic Smoothing of Regression Functions in Generalized Linear Models", JASA 81 96-104.

Silverman, B. W. (1982), " On the Estimation of a Probability Density Function by Maximum Penalized Likelihood Method", *Annals of Statistics* 10 795-810.

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

Tibshirani, R. (1988),"Estimating Transformations for Regression Via Additivity and and Variance Stabilization", JASA 83 394-405.

Acknowledgements

This work was supported by National Science Foundation Grants DMS – 8715756 and DMS – 9002791.

Figure Legends

Figure 1: Ordinary least squares estimates and TBS estimates for the Skeena River Sockeye Salmon. The solid line in the first column of plots is the estimated median response. Dashed lines correspond to the approximate conditional quantiles at .05, .25, .75, .95. Thus, the outer dashed lines give an approximate 90% prediction interval for the mean number of recruit salmon for a given number of spawners. The spline transformation has been applied to the regression model after an initial log transformation. Therefore for large rho the spline based transformation is essentially a power transformation in the original scale. The symbols on the transformation plots indicate the locations of the independent variables (|) and predicted values (+). Note that this transformation is actually the composition of the log function and the transformation obtained from the spline estimate.

Figure 2: TBS estimates of Skeena River Sockeye Salmon without an initial log transformation.

Figure 3: Analysis of bacteria lung clearance based on an initial shift of .05 and a log transformation. The plots in the first row of this panel are based on a least squares model for the shifted/logged data. Subsequent rows are the results for applying the TBS model.

Figure 4: Analysis of bacteria lung clearance based on a shift of .02 and a log transformation.

Figure 1 OLS and TBS estimates for Skeena River Sockeye Salmon
(Log used as an initial transformation)

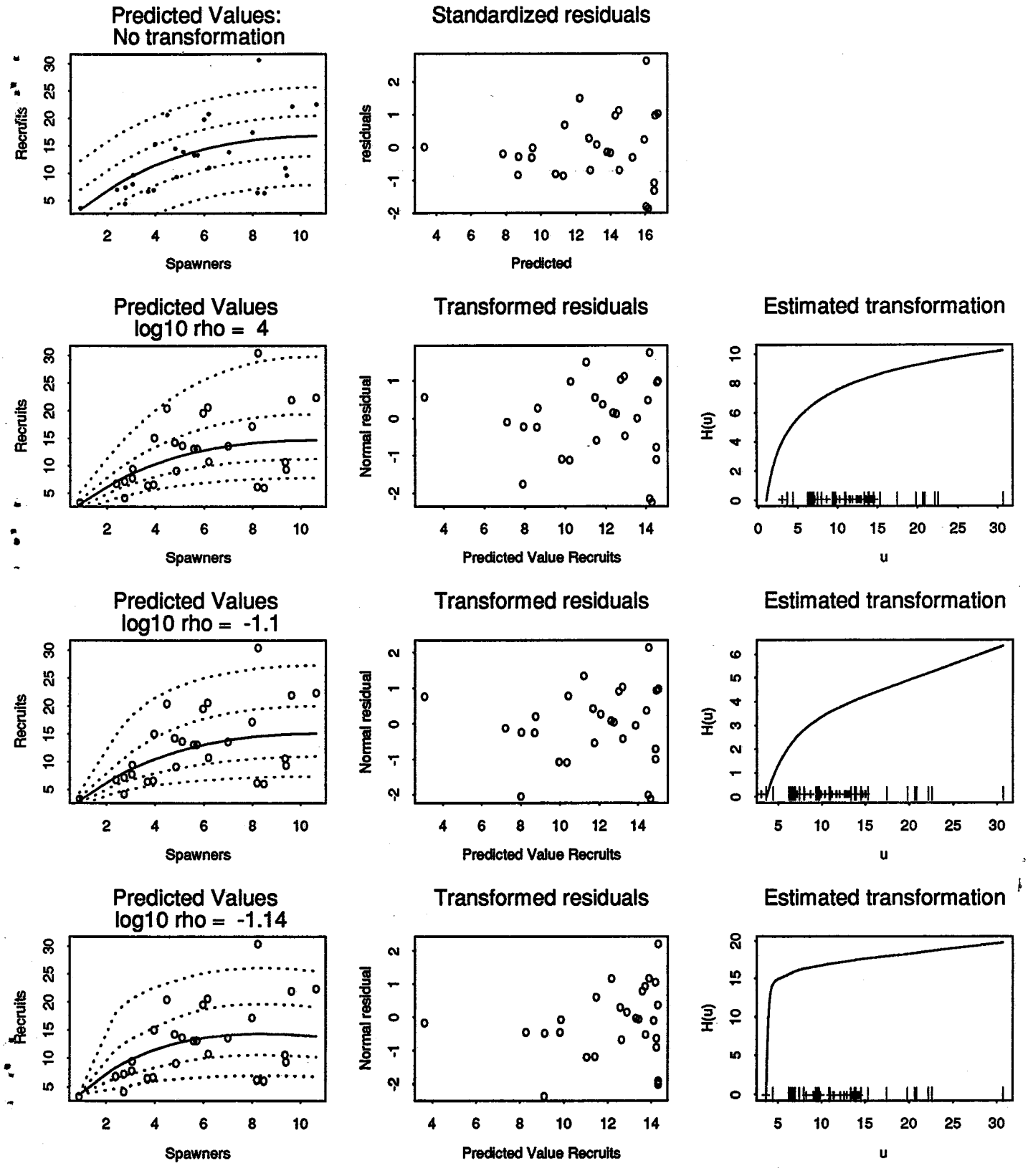


Figure 2 TBS estimates for Skeena River Sockeye Salmon
without intitial log transformation

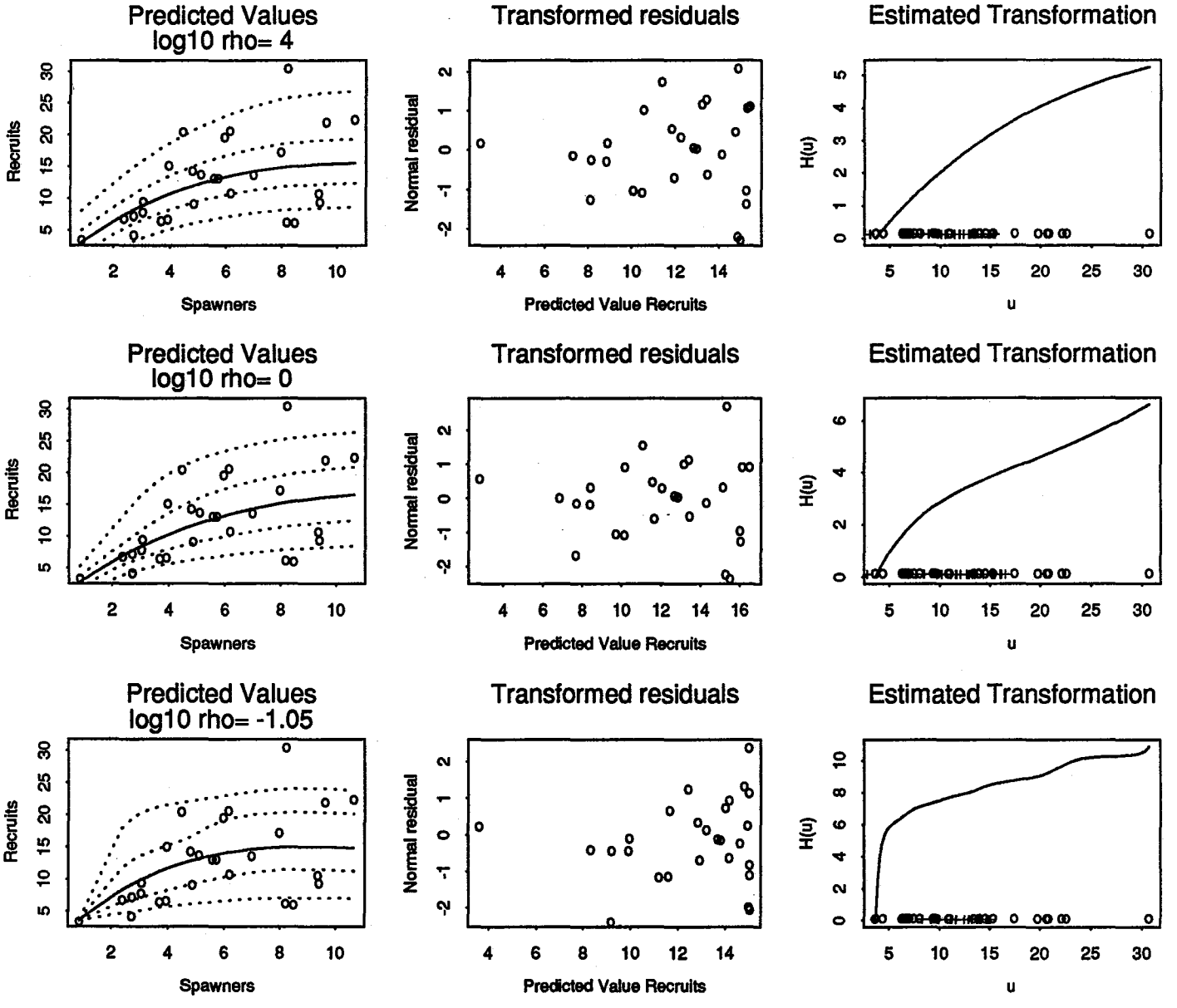


Figure 3 Lung Clearance Data: log transform and shift of .05

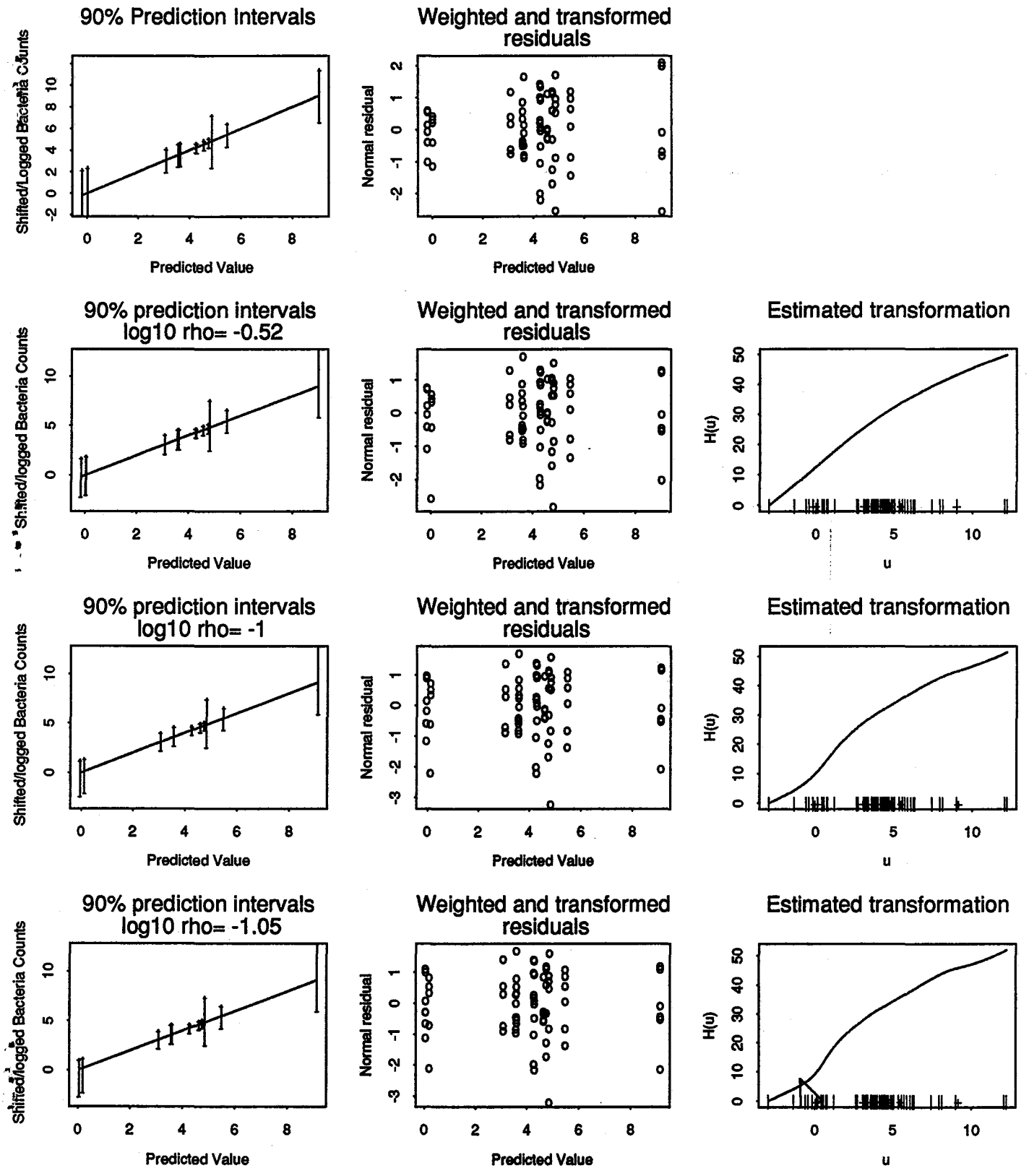


Figure 4 Lung Clearance Data: log transform and shift of .02