Power Transformations When Fitting Theoretical Models to Data

RAYMOND J. CARROLL and DAVID RUPPERT*

We investigate power transformations in nonlinear regression problems when there is a physical model for the response but little understanding of the underlying error structure. In such circumstances, and unlike the ordinary power transformation model, both the response and the model must be transformed simultaneously and in the same way. We show by an asymptotic theory and a small Monte Carlo study that for estimating the model parameters there is little cost for not knowing the correct transform a priori; this is in dramatic contrast to the results for the usual case where only the response is transformed. Possible applications of the theory are illustrated by examples.

KEY WORDS: Transformations; Box-Cox models; Theoretical models; Robustness; Nonlinear regression.

1. INTRODUCTION

Often in scientific work, an experimenter observes data y_i and $x_i' = (x_{1i} \cdots x_{pi})$ and postulates that these data follow a model

$$y_i = f(x_i, \theta_0), \quad i = 1, \ldots, N, \tag{1.1}$$

where θ_0 is a k-parameter vector. The function f may be derived, for example, from differential equations believed to govern the physical system that gave rise to the data. The deterministic model (1.1) is often inadequate since the data exhibit random variation, but whereas f was derived from theoretical considerations, there is really no firm understanding of the mechanism producing the randomness. In this case, the experimenter usually assumes that

$$y_i = f(x_i, \theta_0) + \epsilon_i, \tag{1.2}$$

where the $\{\epsilon_i\}$ are iid $N(0, \sigma_0^2)$. In those cases in which the data suggest that model (1.2) is also unsatisfactory, one might then, for example, assume that the errors are multiplicative and lognormal, so that

$$\log(y_i) = \log(f(x_i, \theta_0)) + \epsilon_i. \tag{1.3}$$

The point here is that model (1.1) is equivalent to the model

$$h(y_i) = h(f(x_i, \theta_0))$$

whenever $h(\cdot)$ is a monotonic transformation. Therefore (1.2) and (1.3) are based on the same theoretical model, but they allow variability to enter into the model in different fashions.

A more flexible approach is to take a sufficiently rich family of strictly monotonic transformations $h(y, \lambda)$, indexed by the m-vector parameter λ, and to assume that for some value λ_0,

$$h(y_i, \lambda_0) = h(f(x_i, \theta_0), \lambda_0) + \epsilon_i. \tag{1.4a}$$

Equation (1.1) could be understood to mean $Ey = f$ or $y = f$ when there is no error. We have in mind the latter meaning; the former is not possible under (1.4a). The model (1.4a) is in the spirit of Box and Cox (1964), who suggested the family of power transformations with $m = 1$ and

$$\begin{aligned} h(y, \lambda) = y^{(\lambda)} &= (y^\lambda - 1)/\lambda \quad \text{if } \lambda \neq 0 \\ &= \log(y) \qquad\quad \text{if } \lambda = 0. \end{aligned} \tag{1.4b}$$

However, as we will make clear, our proposed model (1.4) has greatly different ramifications than those usually associated with the power family. Box and Cox (1964) used their family in a study of the transformation model

$$h(y, \lambda_0) = x'\theta_0 + \epsilon. \tag{1.5}$$

Notice that here, unlike in (1.4), the regression function in (1.5) is *not* transformed. Box and Cox sought a transformation that achieves (a) a simple additive or linear model, (b) homoscedastic errors, and (c) normally distributed errors. Our model is different. Theoretical considerations already provide a regression function. We hope to transform the response *and* the regression function simultaneously to obtain homoscedasticity and normality.

There are two reasons for using model (1.4) instead of simply fitting (1.1) by least squares or some other method. First, estimation of θ_0 based on model (1.4) should be more efficient than other methods. Second, it may be necessary to estimate the entire conditional distribution of y given x; if the data clearly suggest that the distri-

* Raymond J. Carroll is Professor of Statistics and David Ruppert is Associate Professor of Statistics at the University of North Carolina, Chapel Hill, NC 27514. Research for this article was supported by the Air Force Office of Scientific Research Grant F49620-82-C-0009 and by National Science Foundation Grant MCS 8100748. Rod Reish kindly provided the authors with the menhaden data. He and Rick Deriso greatly aided our understanding of those data. The authors also thank a referee and an editor for their comments on an earlier version of this paper.

© Journal of the American Statistical Association
June 1984, Volume 79, Number 386
Theory and Methods Section

butions of $\{y_i - f(x_i, \theta_0)\}$ are not constant across i, one must go beyond standard regression methodology.

An example that motivated the research of this article is the relationship between egg production in a fish stock and subsequent recruitment into the stock. At least for some species, as egg production increases, the changes in the skewness and variance of recruitment are as large as the change in the median recruitment, and these changes in distributional shape may have important implications for management of the fishery. This example is discussed in more detail in Section 4.1.

Another possible reason for transformation is that often, for an appropriate h, $h(f(x_i, \theta))$ is a linear function of θ. Linearization was an accepted technique before the advent of nonlinear regression programs. Now, however, the statistician must decide whether to use linearization or nonlinear regression. As discussed later, our theory provides a method for deciding whether linearization is appropriate.

A natural question is, Which aspects of the data enable us to estimate λ_0? If we transform y_i by $h(\cdot, \lambda)$ and $\lambda \neq \lambda_0$, then information that $\lambda \neq \lambda_0$ is provided by both (a) nonnormality and (b) nonconstancy in i of the distribution of $h(y_i, \lambda) - h(f(x_i, \theta_0), \lambda)$. If the values of $f(x_i, \theta_0)$ are relatively constant, then (a) provides most of the information. On the other hand, if $\sigma^2 = \text{var}(\epsilon_i)$ is small, then most of the information is provided by heteroscedasticity. To see this last fact, suppose, for example, that (1.4b) holds and that we do not transform the data (i.e., we use $\lambda = 1$), but that the true value λ_0 is not 1. For each λ, let $g(\cdot, \lambda)$ be the inverse of the function $h(\cdot, \lambda)$, and define $g_y(y, \lambda) = (\partial/\partial y)\, g(y, \lambda)$. Then by (1.4) and a Taylor approximation, which is suitable if ϵ_i is small, we have

$$\begin{aligned} y_i &= g[h(f(x_i, \theta_0), \lambda_0) + \epsilon_i, \lambda_0] \\ &\approx f(x_i, \theta_0) + k_i\epsilon_i, \end{aligned}$$

where $k_i = g_y[h(f(x_i, \theta_0), \lambda_0), \lambda_0]$; therefore y_i is approximately normally distributed with mean $f(x_i, \theta_0)$ and variance $k_i^2\sigma^2$.

When analyzing data, after we have determined estimates for θ, λ, and σ, we can estimate the density of y_i (or of $[y_i - f(x_i, \theta)]$, the residual from the median). By plotting this estimated density we can check for skewness and other signs of nonnormality on the original scale. By overlaying plots for several values of x_i we can also check for heterogeneity of the distribution of the untransformed data. Instead of graphing densities, we might graph quantiles against quantiles of the normal distribution; nonnormality would then be especially easy to detect. We use such a quantile-quantile plot in Example 4.1.

When we make inferences about θ, the issue arises whether $\hat{\lambda}$ should be treated as fixed or whether we should acknowledge that it is random. For example, there are at least two approaches to estimating the variance-covariance matrix of $\hat{\theta}$. The first is invert the estimated Fisher information matrix for $(\lambda, \sigma, \theta)$. The second is to transform the model and the response by $h(\cdot, \hat{\lambda})$ and then use standard nonlinear regression methodology. The second method is not strictly correct since it treats λ as known rather than estimated. However, it is convenient and expedient since existing nonlinear least squares software can be applied. In this article we report large-sample analysis and Monte Carlo results showing that the two methods tend to give similar results. The second method usually underestimates the variability of $\hat{\theta}$, but it does give a rough approximation to this variability. In the different model (1.5) of Box and Cox (1964), the two methods can give drastically different results, and this fact has led to considerable controversy; see Bickel and Doksum (1981), Carroll and Ruppert (1981), Hinkley and Runger (1984), and Box and Cox (1982).

Another major difference between our model and that of Box and Cox (1964) is that in our model the parameter θ has physical meaning even when λ_0 is unknown; $f(x_i, \theta_0)$ is the median of y_i regardless of the value of λ_0.

2. THEORETICAL ANALYSIS

To analyze the effect of treating $\hat{\lambda}$ as fixed (and equal to λ_0), we begin by computing the information matrices for $(\lambda_0, \theta_0, \sigma_0)$ and (θ_0, σ_0), the latter case assuming that λ_0 is known. The details quickly become intractable, so we resort to the approximation $\sigma_0 \approx 0$. The following theorems are proved in Appendix A.

Theorem 1. Under general conditions, if $N \rightarrow \infty$ and then $\sigma_0 \rightarrow 0$, the limit distribution of $\hat{\theta}$ is the same whether λ_0 is known or unknown. The limit distribution of $\hat{\sigma}$ depends on whether λ_0 is known or unknown.

Theorem 1 says that the effect of having to transform the problem to get homoscedastic, normal errors is small when σ_0 is small. However, we are not concerned only, or even primarily, with small σ_0. In fact, the need for transformation will probably be greater when σ_0 is large. When σ_0 is small, $\hat{\theta}$ from the untransformed data, $\hat{\theta}_{\lambda=1}$, will have a small bias because y_i will be approximately normally distributed. Moreover, although $\hat{\theta}_{\lambda=1}$ may be inefficient in terms of variance, there may be less need for an efficient estimate if σ_0 is small. The small σ_0 asymptotics do, however, lead to major simplifications, and the Monte Carlo results presented later agree with them.

Because we are interested in all values of σ_0, we looked at a second approach. This approach is outlined in Appendix A. Basically, we construct a third estimator of θ_0 and compute its efficiency with respect to $\hat{\theta}(\lambda_0)$, the estimator of θ_0 when λ_0 is known. This gives us a bound on the efficiency of the MLE.

Theorem 2. For any λ_0, σ_0, θ_0, f, or design $\{x_i\}$, as $N \rightarrow \infty$, the asymptotic relative efficiency of the MLE $\hat{\theta}(\hat{\lambda})$ compared to that estimate $\hat{\theta}(\hat{\lambda}_0)$ with λ_0 known is at least $2/\pi$, that is,

$$\text{ARE}(\hat{\theta}(\hat{\lambda}), \hat{\theta}(\lambda_0)) \geq 2/\pi.$$

This bound is very general, and if the Monte Carlo simulation in Section 3 is any guide, the bound is conservative. It follows that the practice of transforming and then using a standard errors for $\hat{\theta}(\hat{\lambda})$ the estimates output from a nonlinear least squares package will be only moderately in error.

3. MONTE CARLO

To study $\hat{\theta}$ when N is finite and σ_0 is not necessarily small, we undertook a small simulation of the model

$$h(y_i, \lambda_0) = h(\theta_1 + \theta_2 x_i, \lambda_0) + \sigma_0 \epsilon_i, \qquad (3.1)$$

where $h(\cdot)$ is the Box and Cox (1964) power family (1.4b). In our simulations, $N = 50$, the design points $\{x_i\}$ were equally spaced on $[-1, 1]$, the errors were normally distributed with mean zero and variance one, and $\theta_1 = 7$, $\theta_2 = 2$. We considered three estimators: (a) ML estimator, λ_0 known (KNOWN), (b) ML estimator, λ_0 unknown (MLE), and (c) The ordinary least squares estimator (LSE) without any transformation.

Since it is a rather frequent practice to use least squares estimation without transformation, we included the LSE in the study. The method of computation is outlined in Appendix B. We chose three values of σ_0: $\sigma_0 = .05$, .10, and .50. We present results in Tables 1 and 2 for $\lambda_0 = 0$ (lognormal data) and $\lambda_0 = .25$. There were 600 replications of the experiment for each (λ_0, σ_0) and each estimator, all generated from a common set of random numbers. The normal random deviates were generated from the IMSL routine GGNPM. Estimation of (θ_1, θ_2) for each λ was done by the IMSL routine ZXSSQ while ZXGSN was used to estimate λ_0.

The results for the ML estimator with λ_0 unknown (denoted by MLE) are very encouraging. The mean squared errors for MLE are reasonably close to those for KNOWN, the ML estimator with λ_0 known, especially for the slope θ_2. These results agree with our small σ theory and indicate the moderate cost of not knowing λ_0. The relative efficiencies of MLE to KNOWN are always well above the lower bound of $2/\pi$. To appreciate how well MLE does compared with KNOWN (line 2 of Tables 1 and 2), see Table 5 of Bickel and Doksum (1981); in their model, which we call (1.5), they have ratios MLE(λ_0 estimated)/KNOWN(λ_0 known) always at least 1.5 and as large as 211, while ours never exceed 1.2.

Table 1. Results of the Monte Carlo Study Described in the Text. (These results are for the INTERCEPT. The median response is linear with intercept = 7 and slope = 2.)

λ =	.00			.25		
σ =	.05	.10	.50	.05	.10	.50
Bias of KNOWN	.03	.06	.56	.01	.03	.23
MSE of KNOWN	2.41	9.67	24.87	.90	3.59	9.04
Bias of MLE	.02	.04	.60	.01	.02	.19
MSE of MLE / MSE of KNOWN	1.02	1.05	1.14	1.01	1.03	1.12
MSE of MLE − MSE of KNOWN	.05	.47	3.44	.01	.09	1.09
STD. ERROR of above difference	.02	.15	.77	.01	.04	.25
Bias of LSE	.11	.40	9.48	.04	.13	2.60
MSE of MLE / MSE of LSE	.97	.90	.22	1.00	.98	.63
MSE of MLE − MSE of LSE	−.06	−1.15	−96.62	.00	−.06	−6.07
STD. ERROR of above difference	.04	.33	4.71	.01	.06	.78

NOTE: Known = ML estimate with λ known, MLE = ML estimate with λ unknown, and LSE = ordinary least squares estimate. In these calculations, the mean squared error (MSE) and STD. ERROR of difference terms are multiplied by T**2. Here $T = 10$ if $\sigma \leq .10$ and $T = 1$ if $\sigma = .50$.

Table 2. Results of the Monte Carlo Study Described in the Text. (These results are for the SLOPE. The median response is linear with intercept = 7 and slope = 2.)

λ =	.00			.25		
σ =	.05	.10	.50	.05	.10	.50
Bias of KNOWN	.01	.01	.03	.00	.01	.02
MSE of KNOWN	7.08	28.36	72.23	2.71	10.83	27.24
Bias of MLE	−.01	−.04	−.15	.00	−.02	−.16
MSE of MLE' / MSE of KNOWN	1.06	1.06	1.01	1.06	1.06	1.03
MSE of MLE − MSE of KNOWN	.41	1.57	.95	.15	.60	.72
STD. ERROR of difference	.10	.40	.67	.04	.77	.27
Bias of LSE	.05	.15	2.97	.02	.04	.50
MSE of MLE / MSE of LSE	.98		.59	1.01	1.01	.91
MSE of MLE − MSE of LSE	−.16	−1.29	−50.54	.05	.13	−2.81
STD. ERROR of above difference	.18	.80	5.10	.06	.23	.74

NOTE: Known = ML estimate with λ known, MLE = ML estimate with λ unknown, and LSE = ordinary least squares estimate. In these calculations, the mean squared error (MSE) and STD. ERROR of difference terms are multiplied by T**2. Here $T = 10$ if $\sigma \leq .10$ and $T = 1$ if $\sigma = .50$.

The other valuable point learned from Table 2 is that when we estimate the slope θ_2, the ML estimator with λ_0 unknown tends to dominate the LSE, especially for larger values of σ_0. In other words, for our model (1.4), there is real value to transformation when it is appropriate.

Finally, it should be noted that there is indeed a (moderate) cost for estimating θ_0 when λ_0 must also be estimated. The consequence of this moderate cost is that inference drawn in the "usual" way—treating $\hat{\lambda}$ as if it were preassigned—will be only moderately in error. (See Carroll and Ruppert 1981 and Carroll 1982a for details concerning the error in the usual inference for model (1.5), which tends to be moderate, on average, but which can be large for prediction at individual design points.)

4. EXAMPLES

4.1 Spawner-Recruit Data

This research was motivated by our study of the population dynamics of the Atlantic menhaden, which is, excluding shellfish, the third largest commerical U.S. fishery. The Atlantic menhaden fishery experienced a massive decline in the mid-1960's, and although there has been a slight recovery, present yields are only about half of those in the early 1960's. Our simulation study was an attempt to find strategies to reverse this decline in harvest; see Ruppert et al. (1983) for further details.

An important part of our study was the examination of the spawner-recruit (SR) relationship, in which we attempted to use the number of eggs E produced by mature menhaden (spawners) to predict the number R of juvenile menhaden recruited into the fishery (recruits). Estimates of E and R for the 21-year period 1955–1975 are given in Table 3.

An inspection of Table 3 or a plot of R against E shows that there is substantial variability. Note, for example, that 1958 has only the eighth-largest egg production, while it produced twice as many recruits as any other year. The year 1975 has the third-largest number of recruits but only the fourteenth largest egg production.

Two of the more usual ways to model the SR relationship are through the following approximations:

$$\text{(Beverton-Holt 1957)} \qquad R_i \approx (\alpha + \beta/E_i)^{-1}$$

$$\text{(Unnormalized Gamma)} \qquad R_i \approx \theta E_i^{\delta} \exp(\gamma E_i).$$

The Unnormalized Gamma (Gamma) is an extension of the Ricker (1954) equation, which allows only $\delta = 1$. Both the Beverton-Holt and the Ricker equations were derived from deterministic models. There appears to be no discussion in the fisheries literature on how these models should be interpreted for fish populations exhibiting highly variable SR relationships. The parameters are often estimated by using linearizing transformations. As stated in the Introduction, these two models can be thought of as part of a relationship driving the system, but they entail considerable variation. We wanted not only to decide upon one of the two models, but also, for our simulations, to do an adequate job of describing the nature of the variation in recruitment given egg production. The difference between the two models can have important effects on methods for managing the menhaden fishery. When, as is usual, $\gamma < 0$, the Gamma curve exhibits overcompensation; that is, eventually large egg production decreases recruitment, perhaps because of competition for food or perhaps because of a population explosion of a predator species. The Beverton-Holt model is much different, since it specifies that, except for random variation, large egg production will lead to an asymptote α^{-1} in recruitment. Since many strategies proposed for increasing the harvest depend on increasing egg production, perhaps beyond historically observed levels, the choice of the Gamma over the Beverton-Holt model could lead to a different management strategy. There has been no previous evidence for Atlantic menhaden supporting the Gamma curve, so a priori we would favor the Beverton-Holt curve, but it is obviously important for us to determine if the Beverton-Holt curve describes the present data as well as or better than the Gamma model.

Table 3. Spawner-Recruit Estimates

Year	Egg Production E^a	Recruits R^b
1955	2.42289	.85558
1956	1.77413	1.00935
1957	1.13816	.49287
1958	1.11338	2.10332
1959	1.32726	.31186
1960	1.88340	.41814
1961	2.62193	.30636
1962	1.63753	.30912
1963	.63302	.25417
1964	.33314	.29163
1965	.20943	.21642
1966	.16043	.30285
1967	.18389	.17046
1968	.23256	.24301
1969	.15267	.40457
1970	.22244	.20309
1971	.31532	.47767
1972	.33109	.37155
1973	.33011	.40746
1974	.27415	.52426
1975	.30154	.92933

[a] In units of 10^{14} eggs.
[b] In units of 10^{10} fish.

Linearization leads to the models

$$\text{(Beverton-Holt, Linear)} \quad R_i^{-1} = \alpha + \beta E_i^{-1} + \sigma_1 \epsilon_i$$

$$\text{(Gamma, Linear)} \quad \log R_i = \delta \log E_i + \theta_* + \gamma E_i + \sigma_2 \epsilon_i. \tag{4.1}$$

From the point of view of meeting the assumption that $\epsilon_1, \ldots, \epsilon_n$ are iid $N(0, 1)$, the linearized Beverton-Holt is superior; the predictions of R_i are similar for the two models, but the residuals from the linearized Gamma are less normal-looking and somewhat more heteroscedastic. Thus, if we are constrained to admitting only the linearization models (4.1), the choice for simulation studies would be the Beverton-Holt.

There is, however, no reason why the variation about the Gamma model should be best explained by forcing linearization through logarithms. As argued in the Introduction, a more flexible model for determining the structure of the model variability is through our nonlinear Box-Cox models

$$\text{(Beverton-Holt)} \quad R_i^{(\lambda_B)} = \{(\alpha + \beta E_i^{-1})^{-1}\}^{(\lambda_B)} + \sigma_B \epsilon_i$$

$$\text{(Gamma)} \quad R_i^{(\lambda_G)} = \{\theta\, E_i \exp(\gamma E_i)\}^{(\lambda_G)} + \sigma_G \epsilon_i.$$

The MLE for λ_B is $\hat{\lambda}_B = -.72$, with a 90% confidence interval of $(-1.0, -0.17)$, and $\hat{\lambda}_B$ restricted to $[-1, 1]$. The likelihood ratio test for $H_0 : \lambda_B = -1.0$ has value $\Lambda_B = .63$, indicating that the linearized Beverton-Holt model is at least reasonable. (Compare with X (1) quantiles.)

For the Gamma model, we obtained $\hat{\lambda}_G = -.71$, with a 90% confidence interval of $(-1.0, -.16)$. The likelihood ratio test for $H_0 : \lambda_G = 0$ has value $\Lambda_G = 4.61$. This indicates that linearizing the Gamma model is probably not appropriate. In fact, having transformed by the power $\hat{\lambda}_G = -.71$, the residuals are essentially as normal looking and homoscedastic as those from the linearized Beverton-Holt.

The estimated Gamma curve reaches a maximum well above historically observed levels of egg production. In fact, the fitted Gamma and Beverton-Holt curves are quite similar over the observed range. However, our simulation experiments included allowing increased egg production where overcompensation would have an effect if the Gamma curve were used in the simulation model. We decided to base our simulations on the Beverton-Holt SR relationship, because there is no real evidence for overcompensation.

As this example makes clear, nonlinear models that can be linearized should not necessarily be linearized, since transformation analysis of response and predictor function can lead to a data scale with better distributional properties. In some cases, however, such as the Beverton-Holt model given here, the transformation analysis will provide added support for linearization.

Our theory predicts that the need to estimate λ is not costly in regard to estimation of α and β, and examination of the relevant Fisher information matrices suggests that this is, in fact, the case. If we fix $\lambda = \hat{\lambda}$, and (pretending that $\lambda = \hat{\lambda}$ was known a priori) invert the information matrix for α, β, and σ, then the estimated (asymptotic) variances are .2029, 2.0361, and .0258, respectively. If we invert the information matrix for α, β, σ, and λ, then the estimated (asymptotic) variances for α, β, and σ are .2213, 2.0394, and .1674, respectively. As our theory predicted, only the variance of $\hat{\sigma}$ increased substantially.

From our data analysis, we concluded that a realistic simulation model would need to be stochastic, and it was in the development of a stochastic model that power transformations proved to be most useful. In our simulation model we used

$$R = [(\hat{\alpha} + \hat{\beta}/E)^{-\hat{\lambda}} + \hat{\sigma}\epsilon]^{1/\hat{\lambda}}, \qquad (4.2)$$

where $\hat{\alpha}$, $\hat{\beta}$, and $\hat{\sigma}$ are estimates on the $\hat{\lambda}$ scale, and ϵ is a standard normal pseudorandom number. With small probability the quantity in square brackets in (4.2) will be close to 0 or even negative, but in the model this quantity was truncated, so recruitment never exceeded twice the greatest recruitment observed in our data. In (4.2) one could use the MLE, $\hat{\lambda} = -.72$, but for simplicity, and because a likelihood ratio test indicated that $H_0: \lambda = -1.0$ was very credible, we used $\hat{\lambda} = -1.0$.

Model (4.2) with either $\hat{\lambda} = -1.0$ or $\hat{\lambda} = -.72$ is a particularly simple model that possesses these essential characteristics found in the data:

(i) Recruitment is highly variable and the variability increases with E.

(ii) Recruitment is positively skewed, and the skewness also increases with E. Therefore, except when E is small, the fishery has occasional dominant year classes.

The heteroscedasticity and variable skewness can be seen by examining the estimated distributions of recruitment with eggs set equal to the observed values during 1961 and 1969, the years with highest and lowest values of egg production, respectively, among all years for which we have data. In Figure 1, the quantiles of these estimated distributions are plotted against normal quantiles. The plots were obtained by plotting (4.2) with $\epsilon = \Phi^{-1}(i/70)$ on the horizontal axis and $\Phi^{-1}(i/70)$ on the vertical axis for $i = 2, \ldots, 68$, and interpolating these points with a spline. (Φ is the standard normal distribution function.) For the graphs, we used $\hat{\lambda} = -.72$ in (4.2), but $\hat{\lambda} = -1.0$ (the value used in simulations) would give similar plots.

With our model we were able to make a detailed simulation study of management policies for Atlantic menhaden. We found that management of a fishery with occasional, randomly occurring, dominant-year classes is a problem considerably different from managing a fishery with low variability.

In some situations, λ may be a nuisance parameter that is estimated only so that other parameters can be more efficiently estimated. However, as in this example, we may sometimes want to know the conditional distribution of the dependent variable, given the independent variables. λ then becomes a parameter equally as important as other parameters.

It is no coincidence that $\lambda_B \approx \lambda_G$. Since, for the range of E in the data, the Beverton-Holt and unnormalized Gamma curves with estimates substituted for the parameters are similar, their residuals from the estimated medians are also similar. $\hat{\lambda}$ is determined by the nonnor-

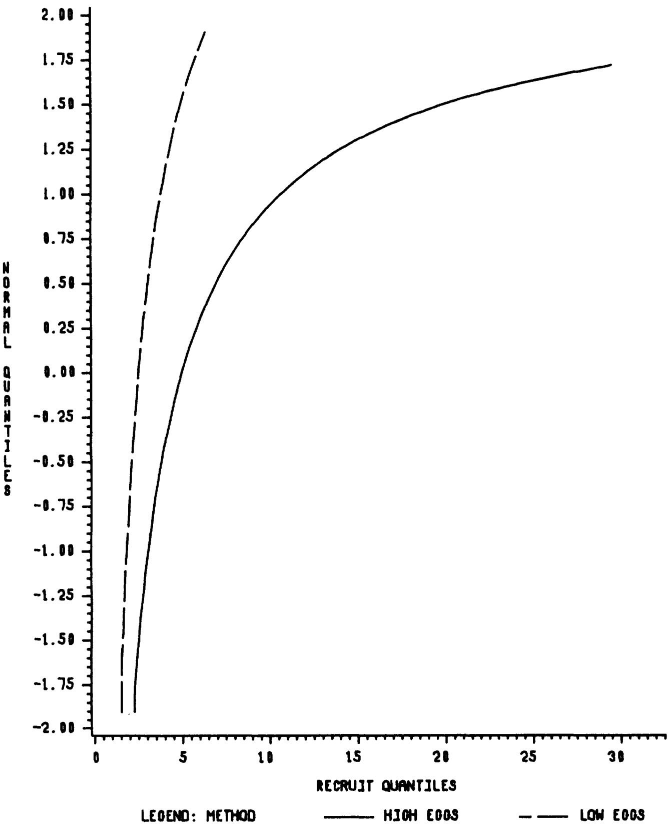

Figure 1. Estimated quantiles of recruitment plotted against standard normal quantiles. Recruitment is conditional on egg production being equal to the 1961 value (HIGH EGGS) *or the 1969 value* (LOW EGGS). *Recruitment is in units of 10^9 fish.*

mality and heterogeneity of distribution that can be detected in these residuals.

As a final note, the analysis presented here was not merely an academic exercise; it formed a part of our study of the SR relationship, which itself was only a small (albeit important) component of a large study performed under time constraints. We welcome further analyses of the data, but we hope it is clear that we do not consider the reported analysis complete. In fact, we analyzed many other models under varying assumptions. For example, the inclusion of a quadratic time trend in the linearized Beverton-Holt model substantially improved the fit to the data. However, the time trend may be due to substantial overfishing in the 1960's, and the use of the trend for predicting future recruitments does not seem warranted. Another candidate for an explanatory variable in a more complex model is recruitment lagged one year.

4.2 Chemical Reaction Data

As a second sample, consider the data of Carr (1960) on the isomerization of pentane. For that data set, one proposed model is

$$y = \frac{\theta_0\theta_2(x_2 - x_3/1.632)}{1 + \theta_1x_1 + \theta_2x_2 + \theta_3x_3}. \tag{4.3}$$

Box and Hill (1974) also list the data and discuss the model. They linearize (4.3) by taking inverses and then using a form of weighted least squares; without going into the full details, it suffices to state that their analysis suggests that $y^{(\lambda)}$ has constant variance, where $\lambda = .8$ (see also Pritchard, Downie, and Bacon 1977). We shall call the Box and Hill method power transformation (linearized) weight least squares (PTWLS).

Since the linearized model based on analyzing y^{-1} in (4.3) exhibits marked heteroscedasticity, it is interesting to see how our estimation method (based on (1.4a)–(1.4b)) performs; this method will be called PTBS for power transforming both sides. Based on Box and Hill's analysis, we should expect our PTBS to find $\lambda \approx .8$. As seen in Table 4, we estimated $\hat{\lambda} = .71$, which is definitely encouraging.

We applied PTBS to model (4.3), untransformed. See Table 4 for the results, which for θ are somewhat different from those obtained by Pritchard, Downie, and Bacon (1977), who used their algorithm DIRECT on the untransformed data. Possibly this difference is due to the presence of several local minima. When we applied unweighted nonlinear least squares to model (4.3), using Box and Hill's (1974) PTWLS solution as a starting value, other algorithms found a different solution with a smaller sum of squares than that reported by Pritchard, Downie, and Bacon (see Table 4).

Our aim in studying this example was to show that our PTBS gives reasonable results. We think our answers are perfectly sensible, and they correspond to PTWLS. For both, one obtains physically meaningful (positive) estimates of θ_1, θ_2, and θ_3, but unweighted linear least squares on the inverse scale gives negative estimates. We believe that PTWLS and PTBS can be recommended equally for this data set, although perhaps unweighted nonlinear least squares is just as effective and somewhat simpler.

A minor advantage of using the untransformed data is that on the inverse scale, Observation 6 of Box and Hill is highly influential even with power weighting (Carroll 1982b), while on the original scale no observation appears to have unusually high influence on the estimate of λ. Influence and diagnostics for inference in our model are questions that should be addressed in the future.

We used our transformation method successfully on other data sets, including the second data set mentioned by Pritchard, Downie, and Bacon.

APPENDIX A: PROOFS

Outline of Proof for Theorem 1

The likelihood analysis proceeds as follows. Define

$$z_i = dh(f_i(\theta_0), \lambda_0)/d\theta_0,$$

$$f_i(\theta) = f(x_i, \theta),\ f_i = f_i(\theta_0),$$

$$h_y(y) = h_y(y, \lambda) = dh(y, \lambda)/dy, \text{ and } h(y) = h(y, \lambda).$$

Let $h_\lambda(y)$ and $h_{\lambda\lambda}(y)$ be the gradient vector and Hessian of $h(y, \lambda)$ with respect to λ. By simple algebra we find

Table 4. Analysis of Carr's Data Using Unweighted, Least Squares, Power Transformation Weighted Least Squares (PTWLS), *and Power Transforming Both Sides* (PTBS)

Estimation Method	*Unweighted*	PTWLS	PTBS	*Unweighted*	*Unweighted*
Source	Pritchard et al.	Box and Hill	IMSL ZXSSQ[a] and ZXGSN	Pritchard et al.	BMDP3R[b]
Response Variable	y^{-1}	y^{-1}	y	y	y
λ	1	−.8	.71	1	1
Sum of Squares[c]	—	—	—	3.24397	3.23448
$\hat{\theta}_0$	16.3	40.00	39.2	35.9	35.9
$\hat{\theta}_1$	−.043	.75	.043	1.04	.071
$\hat{\theta}_2$	−.014	.35	.021	.55	.038
$\hat{\theta}_3$	−.098	1.85	.104	2.46	.167

[a] See Section 5.

[b] Same solution obtained with BMDPAR, SAS-NLIN with derivatives, and IMSL ZXSSQ.

[c] Used to compare the fits with $\lambda = 1$ and response 5 = y.

the joint information matrix of $(\theta_0, \sigma_0, \lambda_0)$ as (all summations are from 1 to N)

$$N^{-1}I = \begin{bmatrix} S/\sigma_0^2 & 0 & C_1/\sigma_0^2 \\ \cdot & 1/(2\sigma_0^4) & C_2/\sigma_0^4 \\ \cdot & \cdot & C_3/\sigma_0^2 \end{bmatrix},$$

where

$$S = N^{-1}\sum z_i z_i',$$

$$C_1 = -N^{-1}E\sum z_i[h_\lambda(y_i) - h_\lambda(f_i)]',$$

$$C_2 = -N^{-1}E\sum \epsilon_i[h_\lambda(y_i) - h_\lambda(f_i)]',$$

$$C_3 = N^{-1}E\sum\{[h_\lambda(y_i) - h_\lambda(f_i)][h_\lambda(y_i) - h_\lambda(f_i)]' + \epsilon_i[h_{\lambda\lambda}(y_i) - h_{\lambda\lambda}(f_i)] + (\partial/\partial\lambda)(\partial/\partial\lambda)'\log[h_y(y_i)]\}.$$

In general, C_1 and C_2 are not zero, and the asymptotic distribution of $(\hat\theta, \hat\sigma^2)$ when λ_0 is estimated differs from when λ_0 is known. The key question, of course, is whether C_1 and C_2 are sufficiently different from zero to seriously affect the distribution of $\hat\lambda$.

The expressions C_1, C_2, and C_3 are complex even when $f_i(\theta_0)$ has a nice form such as simple linear regression. To simplify matters sufficiently so that we can gain some insight about the difference between knowing and estimating λ_0, we follow Bickel and Doksum (1981) and others and let $\sigma_0 \to 0$.

Taylor expansions show that under mild regularity conditions $C_1 = 0(\sigma_0^2)$, $C_2 = 0(\sigma_0^2)$, and $C_3 = 0(\sigma_0^2)$ as $\sigma_0 \to 0$. Standard calculations show that when λ_0 is known,

$$N^{1/2}\text{covariance}\,[(\hat\theta - \theta_0)/\sigma_0, (\hat\sigma^2 - \sigma_0^2)/\sigma_0^2 \mid \lambda_0 \text{ known}]$$

$$\to A^{-1} = \begin{bmatrix} (\lim S)^{-1} & 0 \\ 0 & 2 \end{bmatrix}. \quad \text{(A.1)}$$

Let $D = \text{Diag}(\sigma_0, \ldots, \sigma_0, \sigma_0^2, 1, \ldots, 1)$. Then, to find this limiting covariance matrix when λ_0 is unknown, we must find the upper left $(k + 1) \times (k + 1)$ corner of

$$(\text{DID})^{-1} = \begin{bmatrix} S & 0 & C_1/\sigma_0 \\ \cdot & \frac{1}{2} & C_2/\sigma_0^2 \\ \cdot & \cdot & C_3/\sigma_0^2 \end{bmatrix}^{-1},$$

which by standard results on inverting partitioned matrices is $A^{-1} + FE^{-1}F'$, where A^{-1} is given in (A.1), $E = C_3/\sigma_0^2 - B'A\,B$, $F = A^{-1}B$, and $B' = (C_1/\sigma_0\ C_2/\sigma_0^2)$. Clearly,

$$F' = (S^{-1}C_1/\sigma_0 \quad 2C_2/\sigma_0)$$

and

$$E = C_3/\sigma_0^2 - C_1'S^{-1}C_1/\sigma_0^2 - 2C_2'C_2/\sigma_0^4.$$

To obtain simple asymptotics, we will assume that for σ_0 fixed, C_1/σ_0^2, C_2/σ_0^2, and C_3/σ_0^2 converge as $N \to \infty$, and that these, in turn, have limits D_1, D_2, and D_3, respectively, as $\sigma_0 \to 0$. We also assume that $S \to S_\infty$ (positive definite) as $N \to \infty$. If $D_3 - 2D_2'D_2$ is nonsingular, then

$$\lim_{\sigma_0 \to 0}\lim_{N\to\infty} F\,E^{-1}\,F' = \begin{bmatrix} 0 & 0 \\ 0 & W \end{bmatrix},$$

where $W = 4D_2'[D_3 - D_2'D_2]^{-1}D_2$.

Outline of Proof of Theorem 2

Let $w_1, \ldots, w_N$ be positive numbers, and let $\hat\theta_1$ be any point that minimizes the expression

$$\sum w_i \mid y_i - f_i(\hat\theta_1) \mid.$$

Under (1.4), $f_i(\theta_0)$ is the unique median of y_i, so $\hat\theta_1$ will be consistent under some regularity conditions. The asymptotic distribution of $\hat\theta_1$ can be studied using techniques in Ruppert and Carroll (1980). A particularly simple asymptotic variance matrix is obtained if $w_i = h_y(f_i(\theta_0), \lambda_0)$, that is, if w_i is proportional to the density of $[y_i - f_i(\theta_0)]$ at its median, zero. Then

$$N^{1/2}(\hat\theta_1 - \theta_0)/\sigma_0 \xrightarrow{\mathcal{L}} N(0, (\pi/2)S^{-1}).$$

Although w_i depends on θ_0 and λ_0, the methods in Carroll and Ruppert (1982) can be used to show that the same limiting distribution holds if one substitutes $\sqrt{N}$-consistent estimates for θ_0 and λ_0.

Let $V(\lambda_0)$ and $V(\hat\lambda)$ be the asymptotic variance matrices of $\hat\theta(\lambda_0)$ and $\hat\theta(\hat\lambda)$, respectively. Since $V(\lambda_0) = S^{-1}$, the asymptotic optimality of the MLE shows that

$$S^{-1} \le V(\hat\lambda) \le (\pi/2)S^{-1},$$

where the inequalities are in the sense of positive definiteness.

APPENDIX B: COMPUTATION

Let $L(\theta, \sigma, \lambda)$ denote the log-likelihood for model (1.4). We do not recommend direct maximization of this likelihood by a canned routine for maximizing a function of many parameters. Rather, we adopt the usual practice for the Box-Cox (1964) model (1.5), which reduces the problem to maximizing a function of the scalar λ. Here are the general steps we used.

Step 1. Fix an initial scale $\lambda^{(1)}$. For the simulation and second example, $\lambda^{(1)} = 1.0$, while for the first example $\lambda^{(1)}$ was chosen to satisfy (4.1).

Step 2. Obtain preliminary estimates of θ, say $\theta^{(1)}$. For the simulation and first example, these were found by least squares, while for the second example the starting values are the last column of Table 4. The value $\sigma^{(1)}$ is simply the square root of the mean squared residual.

Step 3. Now begin the maximization of the log-likelihood. At the current value of λ, find $\theta(\lambda)$, $\sigma(\lambda)$ by using a nonlinear regression algorithm, starting from $\theta^{(1)}$, $\sigma^{(1)}$. After completion, update $\theta^{(1)} = \theta(\lambda)$, $\sigma^{(1)} = \sigma(\lambda)$. Define the one-parameter function $L^*(\lambda) = L(\theta(\lambda), \sigma(\lambda), \lambda)$.

Step 4. On the interval $\lambda \in [-1.0, 1.0]$, $L^*(\lambda)$ is often concave and can be maximized by a program specifically designed to maximize a concave function of one parameter. If $L^*(\lambda)$ is not concave, use a grid search.

For Steps 3 and 4, we used the IMSL subroutines ZXSSQ and XZGSN, respectively. The latter program includes a check for convexity of $-\dot{L}^*(\lambda)$, which in the simulations was always satisifed.

[*Received November 1982. Revised October 1983.*]

REFERENCES

BEVERTON, R.J.H., and HOLT, S.J. (1957), *On the Dynamics of Exploited Fish Populations*, London: Her Majesty's Stationery Office.

BICKEL, P.J., and DOKSUM, K.A. (1981), "An Analysis of Transformations Revisited," *Journal of the American Statistical Association*, 76, 296-311.

BOX, G.E.P., and COX, D.R. (1964), "An Analysis of Transformations," *Journal of the Royal Statistical Society*, Ser. B, 26, 211-252.

——— (1982), "An Analysis of Transformations Revisited, Rebutted," *Journal of the American Statistical Association*, 77, 209-210.

BOX, G.E.P., and HILL, W.J. (1974), "Correcting Inhomogeneity of Variance With Power Transformation Weighting," *Technometrics*, 16, 385-389.

CARR, N.L. (1960), "Kinetics of Catalytic Isomerization of *n*-Pentane," *Industrial and Engineering Chemistry*, 52, 391-396.

CARROLL, R.J. (1982a), "Prediction and the Power Transformation Family When Choice of Power Is Restricted to a Finite Set," *Journal of the American Statistical Association*, 77, 908-915.

——— (1982b), "Robust Estimation in Certain Heteroscedastic Linear Models When There Are Many Parameters," *Journal of Statistical Planning and Inference*, 7, 1-12.

CARROLL, R.J., and RUPPERT, D. (1981), "Prediction and the Power Transformation Family," *Biometrika*, 68, 609-617.

——— (1982), "Robust Estimation in Heteroscedastic Linear Models," *Annals of Statistics*, 10, 429-441.

HINKLEY, D.V., and RUNGER, G. (1984), "Analysis of Transformed Data," *Journal of the American Statistical Association*, 79, 302-309.

PRITCHARD, D.J., DOWNIE, J., and BACON, D.W. (1977), "Further Consideration of Heteroscedasticity in Fitting Kinetic Models," *Technometrics*, 19, 227-236.

RICKER, W.E. (1954), "Stock and Recruitment," *Journal of Fisheries Research Board of Canada*, 11, 559-623.

RUPPERT, D., and CARROLL, R.J. (1980), "Trimmed Least Squares Estimation in the Linear Model," *Journal of the American Statistical Association*, 75, 828-838.

RUPPERT, D., REISH, R.L., DERISO, R.B., and CARROLL, R.J. (1983), "A Stochastic Population Model for Managing the Atlantic Menhaden Fishery and Assessing Managerial Risks," Mimeo Series No. 1532, Department of Statistics, University of North Carolina at Chapel Hill.