Chapter 2
Transformation and Weighting

By David Ruppert

About the Author. David Ruppert is Andrew Schulz Jr. Professor of Engineering, School of Operations Research and Information Engineering, and Professor of Statistical Science, Cornell University. He received a PhD in Statistics and Probability from Michigan State University in 1977. He was Assistant and then Associate Professor of Statistics at the University of North Carolina, Chapel Hill, from 1977 to 1987 during which time his office was next to Ray's and they collaborated intensely. He is a Fellow of the ASA and IMS and received the Wilcoxon Prize in 1986 jointly with Ray. He has had 28 PhD students and three of them, Len Stefanski, David Giltinan, and Doug Simpson, were jointly advised with Ray. Professor Ruppert has written five books of which three, *Transformation and Weighting in Regression*, *Measurement Error in Nonlinear Models* (first and second editions), and *Semiparametric Regression* were coauthored with Ray. He and Ray have coauthored 37 papers.

Selected Papers on Transformation and Weighting

[TW-1]-[70] Carroll, R. J. and Ruppert, D. (1981). Prediction and the power transformation family. *Biometrika*, 68, 609–616.

[TW-2]-[57] Carroll, R. J. and Ruppert, D. (1982). A comparison between maximum likelihood and generalized least squares in a heteroscedastic linear model. *Journal of the American Statistical Association*, 77, 878–882.

[TW-3]-[150] Carroll, R. J. and Ruppert, D. (1984). Power transformations when fitting theoretical models to data. *Journal of the American Statistical Association*, 79, 321–328.

[TW-4]-[422] Davidian, M. and Carroll, R. J. (1987). Variance function estimation. *Journal of the American Statistical Association*, 82, 1079–1092.

By the early 1980s, regression with homoscedastic errors was well understood, but methodology for handling heteroscedastic noise was just being developed. There were two general approaches. In the first, studied by Carroll and Ruppert (1981 [TW-1], 1984 [TW-3]), the response is transformed to homoscedasticity. In the second, studied by Carroll and Ruppert (1982 [TW-2]) and Davidian and Carroll (1987 [TW-4]), one uses a variance function that specifies the conditional variance of the response given the covariates. Transformation has the added feature that it can also reduce skewness of the errors, but transformation is useful only when the conditional variance is of a special form and, in particular, is a function of the conditional mean; this is a common occurrence, but there are many applications where it does not occur. Transformation and variance functions can be combined into a very general methodology as described briefly below.

There are two important reasons for modeling the conditional variance. The first is that the regression parameters can be more precisely estimated if one weights by the reciprocals of the conditional variances. The second is that prediction and

M. Davidian et al. (eds.), *The Work of Raymond J. Carroll: The Impact and Influence of a Statistician*, DOI 10.1007/978-3-319-05801-6_2,
© Springer International Publishing Switzerland 2014

calibration intervals can be grossly inaccurate (true coverage probabilities far from nominal values) if one ignores the heteroscedasticity. As Davidian and Carroll (1987 [TW-4]) note, the second reason may be more important. A weighted analysis is significantly more efficient than an unweighted one only when there is substantial heteroscedasticity, but even a small amount of heteroscedasticity, say the conditional standard deviation varying by a factor of two, can cause prediction and calibration intervals to be seriously in error.

Transformation and the Box–Cox Controversy

Carroll and Ruppert (1981 [TW-1]) find a middle ground in a somewhat acrimonious controversy about the use of the Box–Cox transformation model in practice. Although the transformation of variables, e.g., replacing a variable by its logarithm, has had a long history in statistics, estimation of transformation parameters was not put on a firm theoretical footing until Box and Cox (1964). Their model is

$$y_i^{(\lambda)} = x_i\beta + \sigma\varepsilon_i, \tag{2.1}$$

where y_i is a nonnegative response for the ith case, x_i is a vector of predictors, β is a vector of regression coefficients, σ is the residual standard deviation, and $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_N$ are i.i.d. $N(0,1)$, or more generally i.i.d. F for some known F. Here,

$$\begin{aligned} y^{(\lambda)} &= (y^\lambda - 1)/\lambda, \quad && \lambda \neq 0, \\ &= \log(y), && \lambda = 0, \end{aligned} \tag{2.2}$$

embeds the log transformation smoothly into the power transformation family. Model (2.1) states that, after transformation by an unknown parameter λ, the response follows a homoscedastic, Gaussian linear model.

The controversy was over whether inference about β should be conditional on the value of λ or not. Box and Cox (1982) recommend the conditional approach so that once λ is estimated, λ is treated as if it were known and equal to its estimator $\hat{\lambda}$. Bickel and Doksum (1981) disagree and study the sampling variability of β when λ is treated as unknown. Because the value of β is highly dependent on that of λ, the estimators $\hat{\beta}$ and $\hat{\lambda}$ are highly correlated, and the standard deviations of the components of $\hat{\beta}$ are much larger when λ is estimated compared to when λ is treated as known. In summary, Box and Cox argue that uncertainty about λ should be ignored when making inference about β, while Bickel and Doksum argue that this uncertainty should be acknowledged and has a large effect, so that inference about β is unstable.

Neither of these viewpoints seems entirely satisfactory. In a rebuttal to Bickel and Doksum, Box and Cox (1982) ask "how can it be sensible scientifically to state a conclusion as number measured on an unknown scale?" This is a reasonable question. On the other hand, there are few if any other estimation problems where ignoring the uncertainty in nuisance parameters is recommended in practice. Certainly, there must be some cost due to estimation of λ.

Carroll and Ruppert (1981 [TW-1]) study the problem of prediction about y on the original scale. That is, they study $f(\hat{\lambda}, x_0\beta^*)$ where $f(\lambda, \cdot)$ is the inverse of $y^{(\lambda)}$ so that $f(\lambda, y^{(\lambda)}) \equiv y$, x_0 is a value where prediction is to be made, and β^* is an estimator of β. Working on the original scale circumvents Box and Cox's objection to conclusions stated on an unknown scale.

Carroll and Ruppert (1981 [TW-1]) show that the high correlation between $\hat{\lambda}$ and β^* has effects that are similar to the effects of multicollinearity in multiple regression. Both have small, but non-ignorable, effects on prediction. Carroll and Ruppert first look at the case of simple linear regression with $\lambda = 0$ and prove a general result showing that the cost (inflation of the mean squared error) due to estimating λ cannot exceed 50 % and often is much smaller, e.g., at most 8 % in the balanced two-sample problem. Then, they look at the general case where the dimension of β is p and extend the model so that $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_N$ are i.i.d. F for some known F. Asymptotic results are messy in the general case but simplify if one uses small-σ asymptotics where $\sigma \to 0$ as $n \to \infty$. Small-σ asymptotics were also used by Bickel and Doksum. In the small-σ case, the cost of estimating λ is $1/p$, exactly the same as the effect of adding an additional covariate in linear regression. In their last section, they look at the problem of predicting the mean response and show that the cost of adding r additional nuisance parameters when there are q parameters in the model is bounded by r/q.

Estimation of the mean on the original scale was studied further by Taylor (1986). Taylor (1988) studied the related problem of estimating event probabilities using binary regression where the link function contains an unknown parameter. Taylor, Siqueira, and Weiss (1996) propose a general framework that includes the Box–Cox model and binary regression with link parameters as special cases. In all three papers, it was found that the cost of estimating the unknown nuisance parameters is small but not ignorable.

Weighting in Regression

Carroll and Ruppert (1982 [TW-2]) address the question of whether one should use the generalized least squares estimator (GLSE) or the normal-theory maximum likelihood estimator (MLE) when fitting heteroscedastic models. The weighted least-squares estimator weights each squared residual by the reciprocal of its conditional variance, but is generally not available since the conditional variances typically are unknown. The GLSE replaces the unknown conditional variances by estimators. The MLE maximizes the likelihood under the working assumption that the errors are normally distributed. Of course, it is only a true maximum likelihood estimator when that assumption holds. The Carroll and Ruppert model is

$$Y_i = x_i^T \beta + \varepsilon_i \{f(x_i, \beta, \theta)\}^{-1/2}, \tag{2.3}$$

where Y_i is the response, x_i is a vector of covariates, β contains the regression coefficients, $\varepsilon_1, \ldots, \varepsilon_N$ are i.i.d. with variance σ^2, f is unknown function that models the heteroscedasticity, and θ is a vector of parameters that specify the conditional variance of Y_i given x_i. A typical example of f is

$$\{f(x_i,\beta,\theta)\}^{-1/2} = (x_i^T\beta)^\alpha, \tag{2.4}$$

so that the conditional variance is proportional to a power of the conditional mean. (In this model, it is usually assumed that $x_i^T\beta$ is positive.)

There are two sources of information about β, the conditional mean $x_i^T\beta$ and the conditional standard deviation $[f(x_i,\beta,\theta)]^{-1/2}$. Maximum likelihood uses both sources and has the smallest possible asymptotic covariance matrix if the errors are Gaussian as the MLE assumes. The GLSE uses only the first source and, in general, is not fully efficient. However, variance models are often only approximations. Carroll and Ruppert (1982 [TW-2]) show that even a minor misspecification of the heteroscedasticity can degrade the performance of the MLE but has little effect on the GLSE.

More precisely, Carroll and Ruppert (1982 [TW-2]) assume that

$$Y_i = x_i^T\beta + \varepsilon_i[G_N(x_i,\beta,\theta)]^{-1/2}, \tag{2.5}$$

where N is the sample size, $G_N(x_i,\beta,\theta) = f(x_i,\beta,\theta)\{1+2BN^{-1/2}h(x_i,\beta,\theta)\}$, and $N^{-1}\sum_{i=1}^N h^2(x_i,\beta,\theta) \to \mu$ for some $0<\mu<\infty$. Thus, $2BN^{-1/2}h(x_i,\beta,\theta)$ represents the misspecification of the conditional standard deviation and, since it decays to 0 at rate $N^{-1/2}$, the model misspecification is too small to be detected with certainty even in the limit as $N\to\infty$. More formally, the true model is contiguous to the assumed model.

The asymptotic distribution of the GLSE assuming model (2.3) is the same under the models (2.3) and (2.5), so that the GLSE is not affected by contiguous misspecification. The asymptotic distribution of the MLE assuming model (2.3) has the same (fully efficient) asymptotic variance under models (2.3) and (2.5), but there is a bias under (2.5). Whether the MLE or the GLSE has the smaller asymptotic mean squared error (MSE) depends on the amount of model misspecification as determined by $B, h(x_1,\beta,\theta),\ldots,h(x_N,\beta,\theta)$, and how much information about β is contained in the conditional standard deviations. The latter is determined by $w_1,\ldots,w_N$ where, with $\hat\beta_M$ the MLE, we have

$$N^{1/2}(\hat\beta_M-\beta) = N^{-1/2}\sum_{i=1}^N\{v_i\varepsilon_i + w_i(\varepsilon_i^2-1)\} + o_P(1), \tag{2.6}$$

so that, roughly speaking, w_i, $i=1,\ldots,N$, determine how the second source of information about β is used and $v_1,\ldots,v_N$ do the same for the first source.

In summary, the asymptotic distribution of the GLSE is robust to misspecification of the conditional standard deviation, but this is not true of the MLE. If there is no misspecification, then the MLE has the smallest asymptotic mean squared error (MSE), but under misspecification either the MLE or the GLSE may have the smallest MSE.

Carroll and Ruppert (1982 [TW-2]) also discuss robustness to outliers. For the GLSE, (2.6) holds with $w_i\equiv 0$ so the GLSE depends linearly, not quadratically, on $\varepsilon_1,\ldots,\varepsilon_N$. Although neither the GLSE nor the MLE is robust to outliers, the

MLE is more seriously affected by outliers because it depends quadratically upon the errors. A robust M-estimator called ROBUST WEIGHTED is also considered in the paper and, in a Monte Carlo study, is the best performing estimator, even when the heteroscedasticity is correctly specified and the errors are normally distributed; in this case, it is tied with the MLE.

Carroll and Ruppert (1984 [TW-3]) propose a model that is at first glance superficially similar to, but ultimately rather different from, the Box–Cox (1964) transformation model. The Carroll–Ruppert model starts with a theoretical model

$$y_i = f(x_i, \theta_0),\ i = 1, \ldots, N, \tag{2.7}$$

relating a response y_i to a covariate vector x_i. Here f is a known function that might have been derived from scientific theory, e.g., pharmacokinetics, and θ_0 is an unknown parameter vector. Model (2.7) will not hold exactly and in many cases there will be substantial variation of y_i about $f(x_i, \theta_0)$.

To estimate θ_0, one can expand (2.7) to the nonlinear regression model

$$y_i = f(x_i, \theta_0) + \varepsilon_i,\ i = 1, \ldots, N, \tag{2.8}$$

where $\varepsilon_1, \ldots, \varepsilon_N$ are i.i.d. errors and typically are assumed to be normally distributed. Carroll and Ruppert noted that (2.7) is equivalent to $h(y_i) = h\{f(x_i, \theta_0)\}$, for all i, where h is any invertible transformation. However, the noise model

$$h(y_i) = h\{f(x_i, \theta_0)\} + \varepsilon_i, \tag{2.9}$$

with $\varepsilon_1, \ldots, \varepsilon_N$ i.i.d. Gaussian, can hold for at most one h. Therefore, there is no compelling reason to assume (2.8). Instead, Carroll and Ruppert (1984 [TW-3]) argue that (2.9) holds for some h in a parametric family of transformations, e.g., (2.2). As an example, if there are multiplicative lognormal errors so that $y_i = f(x_i, \theta_0)\exp(\varepsilon_i)$ where $\varepsilon_1, \ldots, \varepsilon_N$ are i.i.d. normal, then (2.9) holds with $h(y) = \log(y)$.

Model (2.9) seeks a transformation h that induces additive, homoscedastic, and Gaussian errors. The Box–Cox transformation also has these goals, but the Box–Cox transformation model has a third goal, inducing a simple linear model. For example, $x_i\beta$ in (2.1) might be a no-interaction model and then one seeks a λ so that this no-interaction model holds; Box and Cox (1964) provide such an example. In other examples, $x_i = (1\ w_i)$ for a scalar covariate w_i and one seeks λ so that $E(y_i|w_i)$ is linear in w_i. In contrast, model (2.9) does *not* seek to simplify the regression model. Instead, it preserves the regression model by applying h to both y_i and $f(x_i, \theta_0)$. In practice, h will be monotonic and then (2.9) implies that the median of y_i is $f(x_i, \theta_0)$; this is the sense in which the model is preserved. Stated differently, the Carroll–Ruppert method is used when y_i already fits the regression model while the Box–Cox method is used when y_i must be transformed to fit the regression model.

Because $f(x_i, \theta_0)$ is the median of y_i, the problem of stating conclusions on an unknown scale is avoided. Conclusions can be stated about y_i itself. Therefore,

the controversy discussed previously about inference for the Box–Cox model is avoided. Using small-σ asymptotics, Carroll and Ruppert show that the limit distribution of $\hat{\theta}$ is the same when the transformation parameter is unknown as when it is known. A more general result that does not use small-σ asymptotics is that the cost of not knowing the transformation parameter is at most $\pi/2 = 1.57$. This bound should be contrasted with the huge costs that Bickel and Doksum found for the Box–Cox model. Moreover, Carroll and Ruppert's Monte Carlo study shows that this bound is usually quite conservative.

Davidian and Carroll (1987 [TW-4]) provide a comprehensive study of variance function estimation and compare the many variance function estimators that have been proposed. They use the model

$$EY_i = \mu_i = f(x_i, \beta); \quad \mathrm{var}(Y_i) = \sigma^2 g^2(z_i, \beta, \theta), \tag{2.10}$$

where Y_i is a response, x_i is a vector of covariates in the regression function f, z_i is the vector of covariates in the variance function g^2, β is a vector of regression parameters, θ is a vector of variance parameters, and $\varepsilon_1, \ldots, \varepsilon_N$ are i.i.d.. Typically, β is estimated by ordinary least squares and fixed. The residuals from this preliminary estimator of β can be used to estimate θ. For example, the squared residuals are estimators of g^2 though they are biased unless one corrects for the loss of degrees of freedom. Often, $\log(g)$ is linear in θ, and then it is tempting to use the logarithms of the absolute residuals as the responses, though Davidian and Carroll note that residuals near zero induce outliers when this is done. If the data come in groups where x_i and z_i are constant, then the sample variances of these groups are unbiased estimators of g^2 and can be used as the responses in a regression model with g^2 as the regression function.

Combining Transformation and Weighting

Transformation and weighting need to be combined in some applications. A generalization of (2.9) discussed in Chapter 5 of Carroll and Ruppert (1988) is

$$h(y_i) = h\{f(x_i, \theta_0)\} + \sigma g(z_i, \beta, \theta)\varepsilon_i. \tag{2.11}$$

One application of this model is to fitting the Michaelis–Menten equation of enzyme kinetics. A number of methods for estimating the Michaelis–Menten parameters have been proposed. Ruppert, Carroll, and Cressie (1989) show that all of these are special cases of a general transformation/weighting model, so each is efficient only for a certain error structure, that is, for particular values of the transformation and variance parameters. By using the general model, one can adapt to the error structure and obtain more accurate estimators.

References

Other publications by Ray Carroll cited in this chapter.

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

Ruppert, D., Carroll, R. J., and Cressie, N. (1989). A transformation/weighting model for estimating Michaelis-Menten parameters. *Biometrics*, 45, 637–656.

Publications by other authors cited in this chapter.

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 transformation (with discussion). *Journal of the Royal Statistical Society, Series B*, 35, 473–479.

Box, G. E. P. and Cox, D. R. (1982). An analysis of transformation revised, rebutted. *Journal of the American Statistical Association*, 77, 209–210.

Taylor, J. M. G. (1986). The retransformed mean after fitting a power transformation, *Journal of the American Statistical Association*. **81**, 114–118.

Taylor, J. M. G. (1988). The cost of generalized logistic regression. *Journal of the American Statistical Association*, **83**, 1078–1083.

Taylor, J. M. G., Siqueira, A. L., and Weiss, R. T. (1996). The cost of adding parameters to a model. *Journal of the Royal Statistical Society, Series B*, 58, 593–607.

Biometrika (1981), 68, 3, *pp.* 609–15
Printed in Great Britain

On prediction and the power transformation family

BY R. J. CARROLL AND DAVID RUPPERT
Department of Statistics, University of North Carolina, Chapel Hill

SUMMARY

The power transformation family is often used for transforming to a normal linear model. The variance of the regression parameter estimators can be much larger when the transformation parameter is unknown and must be estimated, compared to when the transformation parameter is known. We consider prediction of future untransformed observations when the data can be transformed to a linear model. When the transformation must be estimated, the prediction error is not much larger than when the parameter is known.

Some key words: Asymptotic distribution; Box–Cox family; Maximum likelihood estimation; Monte-Carlo simulation; Prediction of conditional median; Robustness.

1. INTRODUCTION

The power transformation family studied by Box & Cox (1964) takes the following form: for some unknown λ and $i = 1, \ldots, n$,

$$y_i^{(\lambda)} = x_i\beta + \sigma\varepsilon_i, \quad x_i = (1, c_{i2}, \ldots, c_{ip}), \quad \beta' = (\beta_0, \ldots, \beta_{p-1}). \tag{1·1}$$

Here σ is the standard deviation; the ε_i are independently and identically distributed with mean zero, variance one and distribution F, and

$$y^{(\lambda)} = \begin{cases} (y^\lambda - 1)/\lambda & (\lambda \neq 0), \\ \log y & (\lambda = 0). \end{cases}$$

Box & Cox propose maximum likelihood estimates for λ and β when F is the normal distribution. There are numerous alternative methods as well as proposals for testing hypotheses of the form H_0: $\lambda = \lambda_0$ (Hinkley, 1975; Andrews, 1971; Atkinson, 1973; Carroll, 1980). Carroll studied the testing problem via Monte-Carlo; by allowing F to be nonnormal he approximated a problem with outliers and found that the chance of mistakenly rejecting the null hypothesis can be very high indeed.

Bickel & Doksum (1981) develop an asymptotic theory for estimation. For technical reasons they assume that the design vectors $x_1, x_2, \ldots$ are independent and identically distributed according to G. If the maximum likelihood estimate of the regression parameter is $\hat{\beta}$ when λ is known, and $\beta^* = \hat{\beta}(\hat{\lambda})$ when λ is unknown and estimated by $\hat{\lambda}$, they compute the asymptotic distributions of $n^{\frac{1}{2}}(\hat{\beta}-\beta)/\sigma$ and $n^{\frac{1}{2}}(\beta^*-\beta)/\sigma$ as $n \to \infty$ and $\sigma \to 0$. These distributions, which are given in the Appendix, are different, and as regards variances

the cost of not knowing λ and estimating it . . . is generally severe. . . . The problem is that β^* and $\hat{\lambda}$ are highly correlated.

Their theoretical and Monte Carlo work indicate that $\hat{\lambda}$ and β^* are highly variable and highly correlated, and as discussed in §2, the problem is similar in nature to that of multicollinearity. An example of the variability of β^* is given in the next section.

These results are somewhat controversial. One point of discussion concerns the scale on which inference is to be made: i.e. should one make unconditional inference about the regression parameter in the correct but unknown scale, as in Bickel & Doksum's theory, or a conditional inference for an appropriately defined 'regression parameter' in an estimated scale?

In order to eliminate such problems, we will study the cost of estimating λ when one wants to make inferences in the original scale of the observations. In the multicollinearity problem, reasonably good prediction is still possible if new vectors x arrive independently with the distribution G. Motivated by this fact, we focus our attention specifically on prediction, but we also discuss the two-sample problem and a somewhat more general estimation theory. Using Bickel & Doksum's asymptotic theory and Monte Carlo, we find that for prediction as well as other problems in the original scale there is a cost due to estimating λ, but it is generally not severe.

2. PREDICTING THE CONDITIONAL MEDIAN IN REGRESSION

2·1. *The general case*

Our model specifically includes an intercept, i.e. $x_i = (1, c_i)$; by suitable rescaling we assume the c_i have mean zero and identity covariance. From the sample we calculate $\hat{\lambda}$ and β^*, and we are given a new vector $x_0 = (1, c_0)$, which is independent of the other x's but still has the same distribution G. This formulation is simple but hardly necessary; the design vectors x_i could satisfy the usual regression assumptions, and x_0 can be thought of as chosen according to the design. Our predicted value in the transformed scale would be $x_0\beta^*$, so a natural predictor is $f(\hat{\lambda}, x_0\beta^*)$ where

$$f(\lambda, \theta) = \begin{cases} (1+\lambda\theta)^{1/\lambda} & (\lambda \neq 0), \\ e^{\theta} & (\lambda = 0). \end{cases}$$

Notice that if F has median equal to 0, then $f(\lambda, x_0\beta)$ is the median of the conditional distribution of y given x_0, even though it is not necessarily the conditional expectation. Calculation of conditional expectations would require the use of numerical integration and that F be known or an estimator of F be available. See §3 for further discussion.

A Taylor expansion shows that

$$f(\hat{\lambda}, x_0\beta^*) - f(\lambda, x_0\beta)/g(\lambda, x_0\beta) \simeq x_0(\beta^* - \beta) + h(\lambda, x_0\beta)(\hat{\lambda} - \lambda) \tag{2·1}$$

where

$$g(\lambda, \theta) = f(\lambda, \theta)/(1+\lambda\theta), \quad h(\lambda, \theta) = \theta/\lambda - \{(1+\lambda\theta)\log(1+\lambda\theta)\}/\lambda^2.$$

Estimates $\hat{\lambda}$ and β^* are unstable and highly correlated, and expansion (2·1) shows that our problem as presently formulated is quite similar to a prediction problem in regression when there is multicollinearity.

2·2. *Case* 1

We now assume that F is a normal distribution, $\lambda = 0$, $\sigma = 1$, and the model is simple linear regression with slope β_1 and intercept β_0.

For this special case, likelihood calculations (Hinkley, 1975) can be made. Here the correct scale is the log scale and $E(c_i) = 0$, $E(c_i^2) = 1$, $E(c_i^3) = \mu_3$ and $E(c_i^4) = \mu_4$. Lengthy likelihood analysis shows

$$n \operatorname{cov}(\hat{\lambda}, \beta_0^*, \beta_1^*) \to \Sigma_0,$$

where

$$\Sigma_0 = 2\gamma^{-1} \begin{bmatrix} 1 & -c & \beta_0 \beta_1^* \\ -c & \frac{1}{2}\gamma + c^2 & -c\beta_0 \beta_1^* \\ \beta_0 \beta_1^* & -c\beta_0 \beta_1^* & \frac{1}{2}\gamma + \beta_0^2 \beta_1^{*2} \end{bmatrix}$$

and where

$$c = -\tfrac{1}{2}(1 + \beta_0^2 + \beta_1^2), \quad \beta_1^* = \beta_1 + \tfrac{1}{2}\beta_1^2 \mu_3/\beta_0, \quad \gamma = 3 + 4\beta_1^2 + \beta_1^4(\mu_4 - \mu_3^2 - 1).$$

Note that if λ were not estimated we would have had Σ_0 as the identity matrix, and in the next section we give an example which demonstrates the multicollinearity.

THEOREM 1. *Let* MSE (λ, x_0) *be the mean squared error for estimating the conditional median of Y given x_0 and λ known, while* MSE $(\hat{\lambda}, x_0)$ *is the same quantity but with λ unknown. Then*

$$E_G\left(\| x_0 \|^2 \frac{\text{MSE}(\hat{\lambda}, x_0)}{\text{MSE}(\lambda, x_0)} \right) \Big/ E(\| x_0 \|^2) \to H(\beta_1), \tag{2·2}$$

where

$$H(\beta_1) = 1 + \tfrac{1}{2}\{1 + \beta_1^4(\mu_4 - 1 - \mu_3^2)\}\{6 + 8\beta_1^2 + \beta_1^4(\mu_4 - 1 - \mu_3^2)\}^{-1}.$$

Note that $\mu_4 - 1 - \mu_3^2 = E\{(c_1^2 - \mu_3 c_1 - 1)^2\} \geqslant 0$. The quantity (2·2) is a modified form of the average cost for prediction when λ is estimated. If one prefers to assume the design vectors are constants, then one might think of (2·2) as an average over the design. In either case the results are encouraging:

(i) there is a cost due to estimating λ, but it cannot exceed 50%;

(ii) for the balanced two-sample problem, $c_i = \pm 1$ with probability $\frac{1}{2}$, the cost is at most 8% and decreases to zero as $\beta_1 \to \infty$.

2·3. *Case 2: Symmetric errors*

We now allow λ and the number of regression parameters, p, to be arbitrary, but we assume that F is symmetric about zero.

Here we use the asymptotic theory of Bickel & Doksum, in which $n \to \infty$ and $\sigma \to 0$ simultaneously; see the Appendix for details. We report results only for the simplest case of an orthogonal design in which

$$n^{-1} \sum_{i=1}^{n} x_i' x_i \to I.$$

It then follows that $(\lambda, \hat{\beta}^*)$ is asymptotically normally distributed with mean (λ, β) and covariance $\sigma\Sigma_1/n$, where

$$\Sigma_1 = e^{-1} \begin{bmatrix} 1 & -D \\ -D' & eI + D'D \end{bmatrix},$$

and

$$x = (1, x_2, \ldots, x_p) = (x_1, \ldots, x_p), \quad H(a, \lambda) = \lambda^{-1} a - \lambda^{-2}(1+\lambda a)\log(1+\lambda a),$$

$$D = E\{H(x\beta, \lambda)x\}, \quad e = E[\{H(x\beta, \lambda)\}^2] - \sum_{j=1}^{p} [E\{x_j H(x\beta, \lambda)\}]^2.$$

It is interesting that in the case of simple linear regression $\lambda = 0$, Σ_1 is different from but of the same form as Σ_0. More precisely, c is replaced by $c_* = c + \frac{1}{2}$ and $\frac{1}{2}\gamma$ by $e = \beta_1^4(\mu_4 - \mu_3^2 - 1)/4$.

THEOREM 2. *As $N \to \infty$ and $\sigma \to 0$ for any λ,*

$$E_G\left\{\|x_0\|^2 \frac{\mathrm{MSE}(\hat{\lambda}, x_0)}{\mathrm{MSE}(\lambda, x_0)}\right\} \Big/ E_G(\|x_0\|^2) \to 1 + 1/p,$$

where p is the dimension of the vector β.

The small σ asymptotics of Bickel & Doksum tell us that there is a positive but bounded cost due to estimating λ, with the cost decreasing as p increases. Note that Theorem 2 and Theorem 1 agree for simple linear regression, $\lambda = 0$, $\mu_4 - 1 - \mu_3^2 > 0$ and $\beta_1 \to \infty$.

Bickel & Doksum and Carroll also simultaneously introduced robust estimates of (λ, β) based on the ideas of Huber (1977). One can use Bickel & Doksum's small σ asymptotics to show that (i) the cost in robust estimation for estimating λ is still $1/p$ and (ii) Bickel & Doksum's and Carroll's methods have better robustness properties than does maximum likelihood.

We conducted a small Monte Carlo study to check small sample performance and to investigate the results of Theorems 1 and 2. The observations were generated according to $(1 + \beta_0 + \beta_1 c_i + \varepsilon_i)^{1/\lambda}$ for $\lambda = -1$, and $\exp(\beta_0 + \beta_1 c_i + \varepsilon_i)$ for $\lambda = 0$. Here $n = 20$, the ε_i are standard normal, $\beta_0 = 5$, $\beta_1 = 1$ and the c_i centred at zero, equally spaced, satisfy $\Sigma c_i^2 = n$ and range from $-1{\cdot}65$ to $1{\cdot}65$. Then $\mu_4 = 1{\cdot}79$ and $H(\beta_1) = 1{\cdot}06$, so that Theorems 1 and 2 lead us to expect very little cost due to estimating λ. There were 600 repetitions of the experiment. Likelihood calculations show that

$$\Sigma_0 = \begin{bmatrix} 0{\cdot}27 & 3{\cdot}65 & 1{\cdot}35 \\ \cdot & 50{\cdot}28 & 18{\cdot}25 \\ \cdot & \cdot & 7{\cdot}76 \end{bmatrix}$$

with correlation matrix

$$\begin{bmatrix} 1 & 0{\cdot}99 & 0{\cdot}93 \\ \cdot & 1 & 0{\cdot}92 \\ \cdot & \cdot & 1 \end{bmatrix},$$

which illustrates the multicollinearity quite well, for if λ were known then $n^{\frac{1}{2}}(\hat{\beta}_0 - \beta_0)$ and $n^{\frac{1}{2}}(\hat{\beta}_1 - \beta_0)$ would be uncorrelated with common variance 1.

In rows 1 to 4 of Table 1, we provide an analysis of the estimates β_0^* and β_1^* in the case that λ is estimated. The estimates are biased and have much larger mean squared errors than the estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ obtained for the case that λ is known.

The remaining rows of Table 1 give the results for the prediction problem. The last row corresponds to Theorems 1 and 2, although the actual mean squared errors are computed. It appears that, on the average, our asymptotic calculations are reasonable, and there is

Table 1. *Monte-Carlo results for the model* $y_i = \beta_0 + \beta_1 c_i + \sigma\varepsilon_i$, $\beta_0 = 5$, *and* $\beta_1 = 1$; E_K *and* E_U *denote expectation when* λ *is known and unknown, respectively*

	$\lambda = -1{\cdot}0$	$\lambda = 0{\cdot}0$
$\lvert E_U(\hat\beta_0) - \beta_0 \rvert$	0·44	0·60
$\lvert E_U(\hat\beta_1) - \beta_1 \rvert$	0·20	0·26
$[E_U\{(\hat\beta_0-\beta_0)^2\}/E_K\{(\hat\beta_0-\beta_0)^2\}]^{\frac{1}{2}}$	12·9	9·6
$[E_U\{(\hat\beta_1-\beta_1)^2\}/E_K\{(\hat\beta_1-\beta_1)^2\}]^{\frac{1}{2}}$	4·0	4·0
$\dfrac{E_U[\{f(\hat\lambda,\hat\beta_0)-f(\lambda,\beta_0)\}^2]}{E_K[\{f(\lambda,\hat\beta_0)-f(\lambda,\beta_0)\}^2]}$	—	1·35 1·27* 2·27†
$\dfrac{E_U[\{f(\hat\lambda,\hat\beta_0-1{\cdot}65\hat\beta_1)-f(\lambda,\beta_0-1{\cdot}65\beta_1)\}^2]}{E_K[\{f(\lambda,\hat\beta_0-1{\cdot}65\hat\beta_1)-f(\lambda,\beta_0-1{\cdot}65\beta_1)\}^2]}$	—	1·08 1·01* 2·00†
$\dfrac{E_U[\{f(\hat\lambda,\hat\beta_0+\hat\beta_1 c_0)-f(\lambda,\beta_0+\beta_1 c_0)\}^2]}{E_K[\{f(\lambda,\hat\beta_0+\hat\beta_1 c_0)-f(\lambda,\beta_0+\beta_1 c_0)\}^2]}$	1·02	1·06

* The value predicted by a likelihood analysis using Σ_0.
† The value predicted by the small σ analysis using Σ_1.
For the last entry, c_0 is randomly chosen from the design.

only a small cost involved in estimating λ for prediction. To read rows 5 and 6, we note that to this point we have defined the cost of estimating λ as an average over the distribution of the new value x_0. It is also of interest to study the costs conditional on a given value of x_0. For Case 1 when $x_0 = (1, c_0)$ and $\lambda = 0$ we find that

$$\frac{\mathrm{MSE}\,(\hat\lambda, x_0)}{\mathrm{MSE}\,(\lambda = 0, x_0)} \to \Upsilon_0(c_0, \beta),$$

while for Case 2 this limit is $\Upsilon_1(c_0, \beta)$, where

$$\Upsilon_j(c_0, \beta) = a\,\Sigma_j a^{\mathrm{T}} \quad (j = 1, 2), \quad a = [-\tfrac{1}{2}(\beta_0 + \beta_1 c_0)^2, 1, c_0).$$

Rows 5 and 6 of Table 1 give the ratios of the mean squared errors at two points, the centre and an extreme of the design. As expected from Theorems 1 and 2, there is only a slight cost due to estimating λ, and the small σ asymptotics of Bickel & Doksum are somewhat conservative.

3. Prediction of the conditional mean

The estimator in § 2 is the median of the conditional distribution of y given x_0. Our focus in this section is on estimating the conditional mean of y given x_0.

We sketch a general result which indicates that the cost of extra nuisance parameters, such as λ, is not large. We assume a regression model with (Y_i, X_i) having a joint density $g(y, x \mid \theta_0)$. As in normal theory regression we assume

$$g(y, x \mid \theta_0) = g_1(y \mid x, \theta_0)\, g_2(x).$$

Letting $L_n(\theta)$ denote the log likelihood, we make the usual assumptions:

$$\begin{gathered} E\{L_n'(\theta_0)\} = 0, \\ E\{L_n'(\theta_0)\, L_n'(\theta_0)^{\mathrm{T}}\} = -E\{L_n''(\theta_0)\} = I(\theta_0), \\ n^{\frac{1}{2}}(\theta_n - \theta_0) \to N_q\{0, I^{-1}(\theta_0)\}, \end{gathered} \tag{3·1}$$

where θ_n is the maximum likelihood estimate, q is the dimension of the parameter θ_0 and the prime denotes differentiation with respect to θ at $\theta = \theta_0$. We are given a new value x_0 and wish to predict $E(Y \mid x_0)$; the natural estimate, which usually is only computable numerically, is

$$\hat{E}(Y \mid x_0) = \int y g_1(y \mid x_0, \theta_n)\, dy.$$

Taylor expansion shows that

$$\begin{aligned} A_n(\theta_0, x_0) &= n^{\frac{1}{2}}\{(\hat{E}(Y \mid x_0) - E(Y \mid x_0)\} \\ &\simeq \int \{y - E(y \mid x_0)\} \left\{\frac{d}{d\theta} \log g_1(y \mid x_0, \theta_0)\right\} n^{\frac{1}{2}}(\theta_n - \theta_0)\, g_1(y \mid x_0, \theta_0)\, dy \\ &= \int \{y - E(y \mid x_0)\} \left[\frac{d}{d\theta} \log g(y, x_0 \mid \theta_0)\right] n^{\frac{1}{2}}(\theta_n - \theta_0)\, g_1(y \mid x_0, \theta_0)\, dy. \end{aligned} \tag{3·2}$$

An overall measure of the accuracy of the prediction is $E\{A_n^2(\theta_0, x_0)\}$; (3·1) and (3·2) and Schwarz's inequality show that for a sample $\mathscr{S}$

$$E\{A_n^2(\theta_0, x_0) \mid \mathscr{S}\} \leqslant \operatorname{var}\{y - E(y \mid x_0)\}\, n^{\frac{1}{2}}(\theta_n - \theta_0)^{\mathrm{T}} I(\theta_0)\, n^{\frac{1}{2}}(\theta_n - \theta_0).$$

Since $n^{\frac{1}{2}}(\theta_n - \theta_0)^{\mathrm{T}} I(\theta_0)\, n^{\frac{1}{2}}(\theta_n - \theta_0)$ converges in distribution to a chi-squared variable with q degrees of freedom, this suggests that

$$E\{A_n^2(\theta_0, x_0)\} \leqslant q \operatorname{var}\{y - E(y \mid x_0)\}. \tag{3·3}$$

Equation (3·3) shows that in prediction with q parameters the average squared prediction error is bounded, and this bound increases in relative magnitude by r/q when r additional nuisance parameters are added. A similar result holds for the two-sample problem.

Example. Consider the transformation model (1·1) but take $\lambda = 1$; this means one uses the Box–Cox model when transformation is unnecessary. If there are p regression parameters, then $q = p + 1$ when $\lambda = 1$ is known and

$$E\{A_n^2(\theta_0, x_0)\} = \operatorname{var}\{y - E(y \mid x_0)\}\, p.$$

When one estimates λ, (3·3) shows that

$$E\{A_n^2(\theta_0, x_0)\} \leqslant \operatorname{var}\{y - E(y \mid x_0)\}\, (p + 2).$$

Thus, the relative cost of estimating λ is at most $2/p$, which agrees qualitatively with Theorem 2.

We thank Professors Bickel and Doksum for providing a copy of their paper and the referee for his helpful comments.

APPENDIX

Some asymptotics

Suppose that the distribution function F is symmetric. In the theory of Bickel & Doksum (1981), it is assumed that $\sigma = r\eta$ where $r = r(n)$ is a known sequence tending to zero and η is unknown and fixed. Define

$$A = (x_1, \ldots, x_n)^{\mathrm{T}}, \quad P = A(A^{\mathrm{T}} A)^{-1} A^{\mathrm{T}}, \quad Q = (A^{\mathrm{T}} A)^{-1} A^{\mathrm{T}} d^{\mathrm{T}}, \quad d = (d_1, \ldots, d_n).$$

$$d_i = \{\lambda^{-2}(v_i - 1) - v_i \log|v_i|\}, \quad v_i = 1 + \lambda x_i \beta, \quad e = dd^{\mathrm{T}} - dPd^{\mathrm{T}}.$$

Assuming that e converges to a positive limit, they prove after very detailed calculations that $n^{\frac{1}{2}}\{(\hat{\lambda}-\lambda)/\sigma, (\beta^*-\beta)/\sigma, (\hat{\eta}-\eta)/\eta\}$ is asymptotically normally distributed with mean zero and covariance

$$\lim_{n\to\infty} e^{-1}\begin{bmatrix} 1 & -Q & 0 \\ -Q^{\mathrm{T}} & (n^{-1}A^{\mathrm{T}}A)^{-1}e+QQ^{\mathrm{T}} & 0 \\ 0 & 0 & \frac{1}{2}e \end{bmatrix}.$$

Hence when λ is estimated one adds to the covariance of β^* the term $\lim(QQ^{\mathrm{T}}e^{-1})$, which is positive-semidefinite and, as the example shows, can often be much larger than the covariance of $\hat{\beta}$ when λ is known. It is this extra term which causes the instability of the regression estimate β^* when λ is estimated.

References

Andrews, D. F. (1971). A note on the selection of data transformations. *Biometrika* **58**, 249–54.

Atkinson, A. C. (1973). Testing transformations to normality. *J. R. Statist. Soc.* B **35**, 473–9.

Bickel, P. J. & Doksum, K. A. (1981). An analysis of transformations revisited. *J. Am. Statist. Assoc.* **76**, 296–311.

Box, G. E. P. & Cox, D. R. (1964). An analysis of transformations (with discussion). *J. R. Statist. Soc.* B **26**, 211–52.

Carroll, R. J. (1980). A robust method for testing transformations to achieve approximate normality. *J. R. Statist. Soc.* B **42**, 71–8.

Hinkley, D. V. (1975). On power transformations to symmetry. *Biometrika* **62**, 101–11.

Huber, P. J. (1977). *Robust Statistical Procedures*. Philadelphia: Society of Industrial and Applied Mathematics.

[*Received February* 1980. *Revised May* 1981]

A Comparison Between Maximum Likelihood and Generalized Least Squares in a Heteroscedastic Linear Model

R.J. CARROLL and DAVID RUPPERT*

We consider a linear model with normally distributed but heteroscedastic errors. When the error variances are functionally related to the regression parameter, one can use either maximum likelihood or generalized least squares to estimate the regression parameter. We show that likelihood is more sensitive to small misspecifications in the functional relationship between the error variances and the regression parameter.

KEY WORDS: Linear models; Heteroscedasticity; Contiguity; Robustness; Weighted least squares; Maximum likelihood.

1. INTRODUCTION

There has been considerable recent interest in the heteroscedastic linear model, which we write as

$$Y_i = x'_i\beta + \epsilon_i[f(x_i, \beta, \theta)]^{-1/2}, \tag{1.1}$$

where $\beta(p \times 1)$ is the regression coefficient, $\{x_i(p \times 1)\}$ are the design vectors, $\{\epsilon_i\}$ are independent and identically distributed with distribution function F, and the function $f(x_i, \beta, \theta)$ expresses the possible heteroscedasticity. Bickel (1978) considers various tests of the hypothesis of homoscedasticity, that is, tests of

$$H_0: f(x_i, \beta, \theta) \equiv \text{constant}. \tag{1.2}$$

His work has been extended by Carroll and Ruppert (1981), and the tests have been shown to be locally most powerful by Hammerstrom (1981). Other recent papers are Jobson and Fuller (1980), Carroll and Ruppert (1982), Box and Hill (1974), and Fuller and Rao (1978).

Box and Hill (1974), Carroll and Ruppert (1982), and Jobson and Fuller (1980) suggest various forms of generalized weighted least squares estimates (GLSE) of β. Basically, the suggestion is to obtain preliminary estimates $(\hat{\beta}_p, \hat{\theta})$ of (β, θ), estimate variances by $[f(x_i, \hat{\beta}_p, \hat{\theta})]^{-1}$, and then perform ordinary weighted least squares. Carroll and Ruppert (1982) emphasize robustness and develop methods that are robust against outliers and nonnormal distributions F; they prove that generalized M estimates of β, which include GLSE estimates as special cases, are just as good asymptotically as if the weights were really known. The same phenomenon has been found in other models of heteroscedasticity; see Williams (1975) for a review.

Jobson and Fuller (1980) suggest using the information about β in the function f to improve the GLSE. They state that their method is asymptotically equivalent to the MLE for β obtained by setting up the normal likelihood based on (1.1) and maximizing it; this likelihood is

$$\tfrac{1}{2}\sum_{i=1}^{N} \log(f(x_i, \beta, \theta)) - \tfrac{1}{2}\sum_{i=1}^{N} (Y_i - x'_i\beta)^2 f(x_i, \beta, \theta). \tag{1.3}$$

They have a very interesting result that suggests that as long as (1.1) is correct and F is normal the MLE will be preferred to the GLSE.

In this heteroscedasticity problem, we have an additional robustness consideration. Besides the usual goal (Huber 1981) of protecting ourselves against outliers and nonnormal error distributions, we also must protect ourselves against slight misspecifications in the functional relationship between $\text{var}(Y_i)$ and (x_i, β, θ). Since this functional relationship expressed in (1.1) through f is typically at best an approximation, and since our primary interest is estimating β, we would prefer not to estimate β by a statistic that is adversely affected by slight misspecification of f.

In this note, we assume that the error distribution F is actually normal. We study the robustness of GLSE and MLE to small specification errors in f using simple contiguity techniques. We show that small mistakes in specifying f can easily make GLSE preferable to the MLE.

2. A CONTIGUOUS MODEL

We consider small deviations from (1.1) in the form of

$$Y_i = x'_i\beta + [g_N(x_i, \beta, \theta)]^{-1/2}\epsilon_i, \tag{2.1}$$

where for a scalar B and arbitrary unknown function h,

$$g_N(x_i, \beta, \theta) = f(x_i, \beta, \theta)\{1 + 2BN^{-1/2}h(x_i, \beta, \theta)\} \tag{2.2}$$

* R.J. Carroll is Associate Professor and David Ruppert is Assistant Professor, Department of Statistics, University of North Carolina, Chapel Hill, NC 27514. R.J. Carroll was supported by the U.S. Air Force Office of Scientific Research Contract AFOSR-80-0080. Part of his research was completed at the Universitat Heidelberg, with support from the Deutsche Forschunggemeinschaft. David Ruppert was supported by NSF Grant MCS78-01240. The authors wish to thank the editor and associate editor for many helpful comments.

© Journal of the American Statistical Association
December 1982, Volume 77, Number 380
Theory and Methods Section

$$N^{-1}\sum_{i=1}^{N} h^2(x_i, \beta, \theta) \to \mu \quad (0 < \mu < \infty)$$

$\{\epsilon_i\}$ are iid standard normal.

One should note that the model (2.1) is very close to the assumed model (1.1). Thus the model (2.1) fits our needs because the variance misspecification error is very small and decreases for larger sample sizes. An estimate of β that is robust against specification errors should have the same asymptotic properties under both models (1.1) and (2.1). Thus the question at hand is to study the sensitivity of the MLE and GLSE when (1.1) is assumed but (2.1) is true. If l_1 denotes the log-likelihood for (1.1), and l_2 is the log-likelihood for (2.1), it is quite simple to show that, when (1.1) is true, to order $o_p(1)$,

$$l_* = l_2 - l_1$$
$$\doteq -B^2\mu - \sum_{i=1}^{N} (\epsilon_i^2 - 1)Bh(\underline{x}_i, \beta, \underline{\theta})N^{-1/2}, \quad (2.3)$$

so that by the Central Limit Theorem,

$$\mathcal{L}(l_*) \xrightarrow{\mathcal{L}} N(-B^2\mu, 2B^2\mu) \quad \text{when model (1.1) holds,} \quad (2.4)$$

where $N(a, b)$ is the normal distribution with mean a and variance b. From Corollary 1.2 of Hájek and Sĭdák (1967, p. 204), this means that model (2.1) is contiguous to model (1.1).

3. LIMIT DISTRIBUTIONS FOR GLSE

Suppose that for some positive definite matrix S,

$$N^{-1}\sum_{i=1}^{N} x'_i x_i f(x_i, \beta, \theta) \to S. \quad (3.1)$$

Then, assuming normal errors and smoothness conditions on f, Carroll and Ruppert (1982) (as well as Jobson and Fuller 1980) show that when model (1.1) is true, the GLSE $\hat{\beta}_G$ satisfies

$$N^{1/2}(\beta_G - \beta) - N^{-1/2}\sum_{i=1}^{N} S^{-1}x'_i f^{1/2}(x_i, \beta, \theta)\epsilon_i \xrightarrow{P} 9, \quad (3.2)$$

$$N^{1/2}(\hat{\beta}_G - \beta) \xrightarrow{\mathcal{L}} N(0, S^{-1}). \quad (3.3)$$

A formal proof is possible as long as f is smooth, $\{f(x_i, \beta, \theta)\}$ is bounded away from ∞ uniformly in i, and $(\hat{\beta}_p, \hat{\theta})$ satisfy

$$N^{1/2}(\hat{\beta}_p - \beta) = 0_p(1)$$

and

$$N^{1/2}(\hat{\theta} - \theta) = 0_p(1). \quad (3.4)$$

Carroll and Ruppert (1982) and Jobson and Fuller (1980) verify (3.4) in the normal case under certain technical conditions.

Now, since $\{\epsilon_i\}$ are normal random variables, one uses (2.3) and (3.2) to show that $l_* = l_2 - l_1$ and $N^{1/2}(\hat{\beta}_G - \beta)$ are asymptotically independent, so that by LeCam's third lemma (Hájek and Sĭdák 1967, p. 208),

$$\mathcal{L}(N^{1/2}(\hat{\beta}_G - \beta)) \to N(0, S^{-1}), \quad (3.5)$$

and this *under either model* (1.1) *or* (2.1). This means that GLSE is robust against small specification errors of the variance function f. This encouraging result suggests that one will not go too wrong with GLSE as long as model (1.1) is reasonable. These results are easily extended to the robust estimates introduced by Carroll and Ruppert (1982).

4. LIMIT DISTRIBUTION FOR THE MLE

While GLSE is robust against minor errors in specifying the function f in model (1.1), the same cannot be said for the MLE. Denote this MLE by $\hat{\beta}_M$. Jobson and Fuller (1980) show that for a particular covariance matrix Σ, if the MLE is computed assuming (1.1), then under (1.1),

$$N^{1/2}(\hat{\beta}_M - \beta) \xrightarrow{\mathcal{L}} N(0, \Sigma). \quad (4.1)$$

The result of particular interest is that Σ is no larger than S^{-1} (see 3.1) and (3.3) in the sense that $S^{-1} - \Sigma$ is positive semi-definite under the model (1.1). In addition to (4.1), from (2.3) and the proof of Theorem 2 in Jobson and Fuller (1980), $N^{1/2}(\hat{\beta}_M - \beta)$ and l_* are jointly asymptotically normal with mean $(0, -B^2\mu)$, marginal variances $(\Sigma, 2B^2\mu)$, and covariances Bq computed below, that is,

$$(N^{1/2}(\hat{\beta}_M - \beta)', l_*) \xrightarrow{\mathcal{L}} N\left((0, -B^2\mu), \begin{bmatrix} \Sigma & Bq \\ q'B' & 2B^2\mu \end{bmatrix}\right). \quad (4.2)$$

We now indicate why it is true that the only cases in which the MLE can be expected to be robust against variance specification errors is when $S^{-1} = \Sigma$ and the MLE is asymptotically equivalent to GLSE. To see this, first consider model (1.1) to hold. Jobson and Fuller (1980) show that $\hat{\beta}_M$ is essentially a linear function of $\{\epsilon_i\}$ and $\{\epsilon_i^2 - 1\}$, that is, for vectors $\{v_i\}$ and $\{w_i\}$,

$$N^{1/2}(\hat{\beta}_M - \beta) = N^{-1/2}\sum_{i=1}^{N} \{v_i\epsilon_i + w_i(\epsilon_i^2 - 1)\} + o_p(1). \quad (4.3)$$

If we have $w_i \equiv 0$ $(i = 1, \ldots, N)$, then from (3.2), (4.3), and Gauss-Markov, we have that

$$N^{1/2}(\hat{\beta}_m - \hat{\beta}_G) \xrightarrow{P} 0,$$

and the estimates have the same limit distribution. Thus the only way for $\hat{\beta}_M$ to improve on $\hat{\beta}_G$ under (1.1) is for the $\{w_i\}$ to be nonzero. In this case, however, we can perform contiguity calculations based on (2.3) and (4.3), thus showing that under model (2.1), for the MLE com-

puted assuming (1.1),

$$N^{1/2}(\hat{\beta}_M - \beta) \xrightarrow{\mathcal{L}} N(-2Bq, \Sigma), \tag{4.4}$$

$$q = \lim_{N\to\infty} N^{-1} \sum_{i=1}^{N} w_i h(x_i, \beta, \theta).$$

Of course, q will be nonzero in general if the $\{w_i\}$ are.

These results have important consequences for efficiency. Suppose we wish to estimate the linear combination $\alpha'\beta$. Then, under model (1.1),

$$N \text{ MSE } (\alpha'\hat{\beta}_G) \to \alpha' S^{-1}\alpha$$

$$N \text{ MSE } (\alpha'\hat{\beta}_M) \to \alpha \Sigma \alpha \le \alpha' S^{-1}\alpha. \tag{4.5}$$

However, under the model (2.1), when the GLSE and MLE are computed assuming (1.1),

$$N \text{ MSE } (\alpha'\hat{\beta}_G) \to \alpha' S^{-1}\alpha \quad \text{(no change)}$$

$$N \text{ MSE } (\alpha'\hat{\beta}_M) \to \alpha'\Sigma \alpha + 4B^2(\alpha' q)^2, \tag{4.6}$$

and of course $\alpha'\hat{\beta}_M$ will be a rather poor estimate if α is not orthogonal to q and B is large.

5. MONTE CARLO SPECIFICATIONS

We performed a small Monte Carlo study to illustrate the results given in the previous section, as well as to determine the effect of nonnormality; these are the two aspects of robustness discussed in this note, distributional robustness in heteroscedastic models as well as robustness against misspecification of the form of the variance function. All of the results are based on the following model ($\sigma_i = [f(x_i, \beta, \theta)]^{-1/2}$ in the previous notation):

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \sigma_i \epsilon_i \quad (i = 1, \ldots, N).$$

Here $N = 40$ and the design $\{(x_{i1}, x_{i2})\}$ is as given in a similar experiment performed by Jobson and Fuller (1980). What varies in our experiments is the form of $\{\sigma_i\}$ and the distribution of the errors $\{\epsilon_i\}$. However, all weighted estimates were computed assuming the following model for variances:

$$\sigma_i^2 = \alpha_1 + \alpha_2 \tau_i^2, \tau_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2}. \tag{5.1}$$

In context, (5.1) acts like (1.1) of the text. In all the experiments, we took $(\beta_0, \beta_1, \beta_2) = (10, -4, 2)$, as is done by Jobson and Fuller (1980). Normal random numbers were generated by the IMSL routine GGNPM. Contaminated normal random numbers were generated by first finding a normal deviate Z, and then multiplying Z by 3.0 if a uniform (0, 1) random number generated by IMSL's GGUBS exceeded .90. The starting seed was 325017, and the experiments were repeated 800 times.

The estimators we used included first of all the ordinary least squares estimate (LSE). We also attempted to study the estimator JLS of Jobson and Fuller (1980), which is a one-step version of the MLE; see their paper for details. Their estimate worked well at the normal error model and for their choice $(\alpha_1, \alpha_2) = (300.0, .2)$, but it was very bad at nonnormal distributions or even when the heteroscedasticity was severe. Consequently, the estimator JLS* studied here is a modification of Jobson and Fuller's. Basically, JLS* is JLS if both $\hat{\alpha}_{1\text{JLS}} \ge 0$ and $\hat{\alpha}_{2\text{JLS}} \ge 0$, where $\hat{\alpha}_{1\text{JLS}}, \hat{\alpha}_{2\text{JLS}})$ are the estimates of (α_1, α_2) using JLS. However, if either $\hat{\alpha}_{1\text{JLS}} < 0$ or $\hat{\alpha}_{2\text{JLS}} < 0$, we estimated (α_1, α_2) as in Equation (5.1) of Jobson and Fuller. The modified estimator JLS* appeared to us to be very much better than JLS in overall performance.

We also defined a GLSE called GLSE and a weighted robust estimate ROBUST WEIGHTED (Carroll and Ruppert 1982). In extensive trial and error work, we found that in small samples, the choice of method of estimating the weights has a very big effect on GLSE, although asymptotically there is no effect as long as consistent estimates are available; ROBUST WEIGHTED seems almost insensitive to the choice of weighting method even in small samples. Details will be reported in a future paper. We finally settled on the following somewhat complicated method.

First, for any function Ψ, define

$$\xi(\Psi) = (2\pi)^{-1/2} \int \Psi^2(v) \exp(-v^2/2) dv.$$

In general, Huber's Proposal 2 simultaneously solves

$$\Sigma\Psi((Y_i - X_i'\beta)/\sigma)\{X_i/\sigma\} = 0$$

and

$$\Sigma\Psi^2((Y_i - X_i'\beta)/\sigma) = (N - p)\,\xi(\Psi), \tag{5.4}$$

where p = dimension of β. Now define

$$\Psi_k(x) = \min(k, |x|) \text{ sign } (x).$$

The LSE solves (5.4) using $k = \infty$. A general algorithm for defining weighted estimates is based on k. Essentially, what we do is estimate α_2 robustly and α_1 consistently. The estimates of α_2 will also be consistent, although robust estimates of α_1 are apparently not feasible (see Carroll 1979, Sec. 3 for theoretical details). Estimating α_2 by any of the standard methods is not robust and results in poor overall performance of GLSE. For any given k, the algorithm we used is as follows.

1. Let $\hat{\beta}$ solve (5.4) using Ψ_2.
2. Define $\dot{r}_i = (Y_i - X_i'\hat{\beta})^2$, and P as in Jobson and Fuller (1980).
3. Form predicted values $t_i = x'_i\beta$.
4. Define H as an $(N \times 2)$ matrix, the first column of which consists of ones, the second the t_i^2.
5. Solve (5.4) for the regression model

$$E\dot{r} = PH\begin{pmatrix}\alpha_1\\ \alpha_2\end{pmatrix}$$

using Ψ_2 (this is much like (5.1) of Jobson and Fuller 1980). Define $Z_i = \dot{r}_i - \max(\hat{\alpha}_2, 0)t_i^2$.

6. Define $\hat{\alpha}_1 = N^{-1} \sum Z_i$.
7. Compute $\hat{\sigma}_i^2 = \max(\hat{\alpha}_1, 0) + \max(\hat{\alpha}_2, 0)\, t_i^2$.
8. Solve (5.4) using Ψ_2 with Y_i and X_i replaced by $Y_i/\hat{\sigma}_i$ and $X_i/\hat{\sigma}_i$. Call this estimate $\hat{\beta}$.

9. Repeat (2)–(8), except in Step (8) use Ψ_k instead of Ψ_2.

Our estimate GLSE uses $k = \infty$, while ROBUST WEIGHTED uses $k = 2$. The estimate HUBER was the Huber Proposal 2 computed in Step 1.

Finally, the mean squared error (MSE) of an estimator, as well as the standard error of this MSE, were calculated by the following simple device. Denote by WLS the weighted least squares estimate based on the weights σ_i^{-2}. This is not a real statistic since the weights are unknown in practice. Then, for JLS* as an example,

$$\begin{aligned} \text{MSE (JLS*)} &= E\{\hat{\beta}(\text{JLS*}) - \beta\}^2 \\ &= E[\{\hat{\beta}(\text{JLS*}) - \beta\}^2 - \{\hat{\beta}(\text{WLS}) - \beta\}^2] \\ &\quad + \text{MSE (WLS)}. \end{aligned} \tag{5.5}$$

The second term on the right side of (5.5) is known exactly; the first term and its standard error are calculated by the Monte Carlo experiment. Because of the correlation between JLS* and WLS, this method produces better estimates of MSE(JLS*) than would the usual direct Monte Carlo calculation.

6. MONTE CARLO RESULTS

The first part of the study concerns the effect of nonnormality on the estimates and is reported in Table 1. In constructing this table, the assumed model (5.1) was actually true, with $(\alpha_1, \alpha_2) = (300, .2)$ as in Jobson and Fuller's work. For each estimator, the first line is the ratio of its MSE with that of WLS (the weighted estimator with known weights). Note that the Carroll-Ruppert ROBUST WEIGHTED is the best; it is quite competitive at the normal model and the clear winner at the contaminated normal model; this is in agreement with theory. Note too that, qualitatively at least, JLS* suffers the worst in the switch from normal to contaminated normal.

The benefit of using our modification JLS* to Jobson and Fuller's JLS is dramatic here. Ordered as in Table 1, the MSE ratio values for JLS are 1.22, 1.27, 1.25, 2.72, 7.27, and 13.27.

Table 2 is designed to cover the problem of specification robustness discussed theoretically in Sections 1–4. Designing a Monte Carlo experiment that illustrated the theory was quite difficult because the theory is a local theory. We finally used heteroscedastic models that had fairly large inequalities in variances. The assumed model was (5.1), but with $\alpha_1 = 100$, $\alpha_2 = .20$. For the left side of Table 2, we do calculations when (5.1) is in fact true, the errors are normally distributed, and $\alpha_1 = 100$, $\alpha_2 = .20$. In the right side of Table 2 the model corresponding to (2.1) and (2.2) has

$$\sigma_i^2 = \alpha_1 \exp(2\alpha_2^2 \mid \tau_i \mid),$$

$$\tau_i = EY_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2},$$

$$\alpha_1 = 100,$$

and

$$\alpha_2 = .128. \tag{6.1}$$

The choice of $\alpha_2 = .128$ in Table 2 reflects a model whose variance behavior is close to that of (5.1) with $\alpha_1 = 100$, $\alpha_2 = .20$; the ratio of (6.1) to (5.1) over the range of the mean value is between .95 and 1.15. Further, the

Table 1. Distributional Robustness When Model (5.1) Is Assumed and Is True, $\alpha_1 = 300$ and $\alpha_2 = .20$

	Standard Normal Errors			Contaminated Normal Errors		
	β_0	β_1	β_2	β_0	β_1	β_2
LSE	1.32	1.27	1.27	1.33	1.26	1.24
	.05	.04	.04	.06	.05	.05
	−.11	.03	−.01	.14	.00	.01
JLS*	1.18	1.17	1.12	1.25	1.21	1.21
	.04	.04	.04	.06	.06	.10
	−.35	.05	−.03	−.26	.05	−.02
GLSE	1.16	1.18	1.16	1.16	1.14	1.11
	.03	.04	.03	.05	.04	.04
	.16	.02	−.02	.39	.00	.00
Huber	1.27	1.27	1.27	0.96	0.94	.099
	.04	.04	.04	.04	.04	.05
	.07	.01	−.01	.32	−.01	.00
Robust Weighted	1.16	1.18	1.16	0.88	0.89	0.92
	.03	.03	.03	.04	.04	.05
	.23	.01	−.02	−.02	−.01	−.01
Actual MSE of WLS	196.9	1.08	.57	354.4	1.94	1.02

NOTE: The first row is the MSE ratio (MSE of indicated estimator/MSE of WLS), the second its standard error, and the third is the observed Monte Carlo bias.

Table 2. Specification Robustness When (5.1) Is Assumed. Small Specification Error

	Model (5.1) Is True $\alpha_1 = 100$, $\alpha_2 = .20$ (correct model)			Model (6.1) Is True $\alpha_1 = 100$, $\alpha_2 = .128$ (misspecified model)		
	β_0	β_1	β_2	β_0	β_1	β_2
LSE	1.62	1.59	1.60	1.63	1.63	1.58
	.07	.06	.06	.07	.07	.06
	−.04	.02	−.01	−.04	.02	−.01
JLS*	1.54	1.51	1.39	1.62	1.62	1.54
	.12	.12	.09	.15	.16	.16
	−.19	.04	−.03	−.22	.04	−.03
GLSE	1.27	1.29	1.28	1.28	1.31	1.27
	.05	.05	.05	.05	.05	.05
	.07	.02	−.02	.09	.02	−.02
Huber	1.62	1.59	1.60	1.63	1.63	1.58
	.07	.06	.06	.07	.07	.06
	−.04	.02	−.01	−.04	.02	−.01
Robust Weighted	1.27	1.29	1.28	1.27	1.29	1.26
	.05	.04	.04	.05	.05	.04
	.18	.01	−.02	.21	.01	−.01
Actual MSE of WLS	116.1	.72	.35	120.2	.74	.35

NOTE: The first row is the MSE ratio (MSE of indicated estimator/MSE of WLS), the second its standard error, and the third is the observed Monte Carlo bias.

difference in actual MSE for WLS (with known weights) is negligible, as can be seen in the last row of Table 2. As predicted by our theory, the only estimate that seems at all affected by the error in specifying the variances (assuming (5.1) when (6.1) is true) is JLS*, the MLE approximation. Note, too, how well our estimate ROBUST WEIGHTED performs; it is quite good at the normal error model and, as seen in Table 1, is superior for the contaminated normal model.

We also analyzed the case for (6.1) that $\alpha_2 = .131$, which illustrates a specification error resulting in overweighting the points with largest variance. This hardly affected WLS. Once again, the worst overall performance was turned in by JLS*; this MLE approximation was the weighted method most affected by the specification error.

One might ask how well our theory predicts the specific numbers in Table 2. As this section shows, the theory is at least qualitatively correct in predicting that the MLE approximation JLS* would be most sensitive to variance function misspecification, while GLSE and ROBUST WEIGHTED would be only slightly affected (see (4.6)). On the other hand, the result (4.5) that indicates that JLS* should be better than GLSE when the variances are correctly specified was not borne out. Another instance, which is really not too bad, in which asymptotic theory and Monte Carlo theory do not closely agree is that the GLSE had about 30 percent higher mean squared error than WLS. Because the approximation given by (4.5) is not close to the Monte Carlo results, evaluating the excess mean squared error $4B^2(\alpha' q)^2$ in (4.6) due to variance misspecification in unlikely to be too accurate. If

$$\sigma_i^2 = 1/f(x_i, \beta, \theta)$$

and

$$\delta_i^2 = 1/g_N(x_i, \beta, \theta),$$

then from (2.2) we have approximately

$$2Bh(x_i, \beta, \theta) \doteq N^{1/2}(\sigma_i^2/\delta_i^2 - 1). \tag{6.2}$$

Using the theory of Jobson and Fuller, we can evaluate the terms $\{w_i\}$ of (4.3), which then enables us from (6.2) to approximate $2Bq$ of (4.4). When this is done, we are able to predict that in going from the correct model to the misspecified model in Table 2, the MSE for JLS* should increase by (4.7 percent, 5.8 percent, 4.7 percent) for estimating $(\beta_0, \beta_1, \beta_2)$. The actual Monte Carlo increases were (8.9 percent, 10.3 percent, 10.8 percent). In other words, in this example, the effect of variance function misspecification on the MLE approximation JLS* was more than that predicted by the asymptotic theory.

7. DISCUSSION

The theoretical work and the small Monte Carlo study presented here indicate that the maximum likelihood estimate (or approximations to it) in a heteroscedastic model is sensitive both to the normal error assumption and to small errors in specifying a functional form for the variances. Generalized least squares estimates are sensitive to the normal error assumption but, at least theoretically, are robust against small variance specification errors; a particular GLSE was constructed that, in a limited Monte Carlo study, had these properties in small samples. The robust weighted estimators of Carroll and Ruppert (1982) had the best theoretical and empirical robustness behavior, while at the same time giving up only very little when all assumptions about the variances and error distributions are true. For homoscedastic regression models, estimators with bounded influence functions have been defined and studied (Krasker and Welsch 1982). We did not consider the question but believe it is possible to develop bounded influence weighted estimators with appealing properties for heteroscedastic situations.

[*Received April 1981. Revised June 1982.*]

REFERENCES

BICKEL, P.J. (1978), "Using Residuals Robustly I: Tests for Heteroscedasticity, Nonlinearity," *Annals of Statistics*, 6, 266–291.

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

CARROLL, R.J. (1979), "Estimating Variances of Robust Estimators When the Errors are Asymmetric," *Journal of the American Statistical Association*, 74, 674–679.

CARROLL, R.J., and RUPPERT, D. (1981), "On Robust Tests for Heteroscedasticity," *Annals of Statistics*, 9, 205–209.

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

FULLER, W.A., and RAO, J.N.K. (1978), "Estimation for a Linear Regression Model With Unknown Diagonal Covariance Matrix," *Annals of Statistics*, 6, 1149–1158.

HÁJEK, J., and SÍDÁK, Z. (1967), *Theory of Rank Tests*, New York: Academic Press.

HAMMERSTROM, T. (1981), "Asymptotically Optimal Tests for Heteroscedasticity in the General Linear Model," *Annals of Statistics*, 9, 368–380.

HUBER, P.J. (1981), *Robust Statistics*, New York: John Wiley.

JOBSON, J.D., and FULLER, W.A. (1980), "Least Squares Estimation When the Covariance Matrix and Parameter Vector are Functionally Related," *Journal of the American Statistical Association*, 75, 176–181.

KRASKER, W.S., and WELSCH, R.E. (1982), "Efficient Bounded Influence Regression Estimation," *Journal of the American Statistical Association*, 77, 595–604.

WILLIAMS, J.S. (1975), "Lower Bounds on Convergence Rates of Weighted Least Squares to Best Linear Unbiased Estimators," in *A Survey of Statistical Design and Linear Models*, ed. J.N. Srivastava, Amsterdam: North-Holland.

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 $x_i^t = (x_{1i} \cdots x_{pi})$ 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

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

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

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

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 $\sigma^2 = \mathrm{var}(\epsilon_i)$ 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 $N \to \infty$ and then $\sigma_0 \to 0$, 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 $N \to \infty$, 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,

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

This bound is very general, and if the Monte Carlo sim-

ulation 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.

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.)

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 $\sigma = .50$.

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. fish-

ery. 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:

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

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

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

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

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

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-

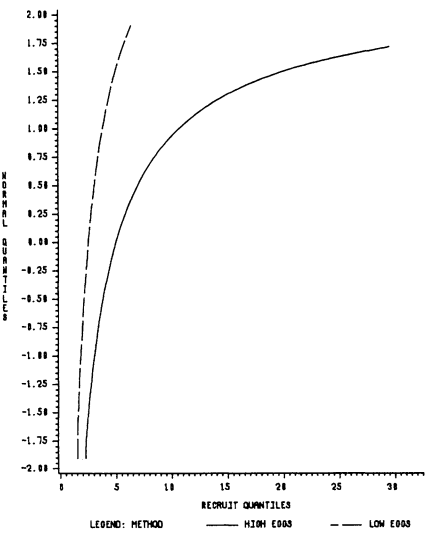

Figure 1. Estimated quantiles of recruitment plotted against standard normal quantiles. Recruitment is conditional on egg production being equal to the 1961 value (HIGH EGGS) *or the 1969 value* (LOW EGGS). *Recruitment is in units of* 10^9 fish.

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

$$y = \frac{\theta_0\theta_2(x_2 - x_3/1.632)}{1 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_3}. \tag{4.3}$$

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

$$\begin{aligned} S &= N^{-1}\sum z_i z_i', \\ C_1 &= -N^{-1}E\sum z_i[h_\lambda(y_i) - h_\lambda(f_i)]', \\ C_2 &= -N^{-1}E\sum \epsilon_i[h_\lambda(y_i) - h_\lambda(f_i)]', \\ C_3 &= N^{-1}E\sum\{[h_\lambda(y_i) - h_\lambda(f_i)][h_\lambda(y_i) - h_\lambda(f_i)]' \\ &\quad + \epsilon_i[h_{\lambda\lambda}(y_i) - h_{\lambda\lambda}(f_i)] \\ &\quad + (\partial/\partial\lambda)(\partial/\partial\lambda)'\log[h_y(y_i)]\}. \end{aligned}$$

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

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

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 $-L^*(\lambda)$, 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.

Variance Function Estimation

M. DAVIDIAN and R. J. CARROLL*

Heteroscedastic regression models are used in fields including economics, engineering, and the biological and physical sciences. Often, the heteroscedasticity is modeled as a function of the covariates or the regression and other structural parameters. Standard asymptotic theory implies that how one estimates the variance function, in particular the structural parameters, has no effect on the first-order properties of the regression parameter estimates; there is evidence, however, both in practice and higher-order theory to suggest that how one estimates the variance function does matter. Further, in some settings, estimation of the variance function is of independent interest or plays an important role in estimation of other quantities. In this article, we study variance function estimation in a unified way, focusing on common methods proposed in the statistical and other literature, to make both general observations and compare different estimation schemes. We show that there are significant differences in both efficiency and robustness for many common methods.

We develop a general theory for variance function estimation, focusing on estimation of the structural parameters and including most methods in common use in our development. The general qualitative conclusions are these. First, most variance function estimation procedures can be looked upon as regressions with "responses" being transformations of absolute residuals from a preliminary fit or sample standard deviations from replicates at a design point. Our conclusion is that the former is typically more efficient, but not uniformly so. Second, for variance function estimates based on transformations of absolute residuals, we show that efficiency is a monotone function of the efficiency of the fit from which the residuals are formed, at least for symmetric errors. Our conclusion is that one should iterate so that residuals are based on generalized least squares. Finally, robustness issues are of even more importance here than in estimation of a regression function for the mean. The loss of efficiency of the standard method away from the normal distribution is much more rapid than in the regression problem.

As an example of the type of model and estimation methods we consider, for observation-covariate pairs (Y_i, x_i), one may model the variance as proportional to a power of the mean response, for example,

$$E(Y_i) = f(x_i, \beta), \qquad \mathrm{var}(Y_i) = \sigma f(x_i, \beta)^\theta,$$
$$f(x_i, \beta) > 0,$$

where $f(x_i, \beta)$ is the possibly nonlinear mean function and θ is the structural parameter of interest. "Regression methods" for estimation of θ and σ based on residuals $r_i = Y_i - f(x_i, \hat{\beta}_*)$ for some regression fit $\hat{\beta}_*$ involve minimizing a sum of squares where some function T of the $|r_i|$ plays the role of the "responses" and an appropriate function of the variance plays the role of the "regression function." For example, if $T(x) = x^2$, the responses would be r_i^2, and the form of the regression function would be suggested by the aproximate fact $E(r_i^2) \approx \sigma^2 f(x_i, \hat{\beta}_*)^{2\theta}$. One could weight the sum of squares appropriately by considering the approximate variance of r_i^2. For the case of replication at each x_i, some methods suggest replacing the r_i in the function T by sample standard deviations at each x_i. Other functions T, such as $T(x) = x$ or $\log x$, have also been proposed.

KEY WORDS: Asymptotic efficiency; Heteroscedasticity; Regression; Variance estimation.

1. INTRODUCTION

Consider a heteroscedastic regression model for observable data Y:

$$EY_i = \mu_i = f(x_i, \beta); \qquad \mathrm{var}(Y_i) = \sigma^2 g^2(z_i, \beta, \theta). \quad (1.1)$$

Here, $\{x_i\}$ are the design vectors, $\beta(p \times 1)$ is the regression parameter, f is the mean response function, and the variance function g expresses the heteroscedasticity, where $\{z_i\}$ are known vectors, possibly the $\{x_i\}$, σ is an unknown scale parameter, and $\theta(r \times 1)$ is an unknown parameter. For example, the variance may be modeled as proportional to a power of the mean:

$$g(z_i, \beta, \theta) = f(x_i, \beta)^\theta, \qquad f(x_i, \beta) > 0. \quad (1.2)$$

One might also model the variance as quadratic in the predictors, that is,

$$\sigma g(z_i, \beta, \theta) = 1 + \theta_1 x_i + \theta_2 x_i^2,$$

or by an expanded power of the mean model, that is,

$$\sigma^2 g^2(z_i, \beta, \theta) = \theta_1 + \theta_2 f(x_i, \beta)^{\theta_3}. \quad (1.3)$$

Box and Meyer (1986) used

$$g(z_i, \beta, \theta) = \exp(z_i'\theta).$$

An important feature of (1.1) is that no assumption about the distribution of the $\{Y_i\}$ has been made other than that of the form of the first two moments. Models that may be regarded as special cases of (1.1) are used in diverse fields, including radioimmunoassay, econometrics, pharmacokinetic modeling, enzyme kinetics, and chemical kinetics, among others. The usual emphasis is on estimation of β with estimation of the variances as an adjunct.

The most common method for estimating β is generalized least squares, in which one estimates $g(z_i, \beta, \theta)$ by using an estimate of θ and a preliminary estimate of β and then performs weighted least squares; see, for example, Carroll and Ruppert (1982a) and Box and Hill (1974). This might be iterated, with the preliminary estimate replaced by the current estimate of β, a new estimate of θ obtained, and the process repeated. Standard asymptotic theory as in Carroll and Ruppert (1982a) or Jobson and Fuller (1980) shows that as long as the preliminary estimators for the parameters of the variance function are consistent, all estimators of β obtained in this way will be asymptotically equivalent to the weighted least squares estimator with known weights.

There is evidence that for finite samples, the better one's estimate of θ, the better one's final estimate of β. Williams (1975) stated that "both analytic and empirical studies . . . indicate that . . . the ordering of efficiency (of estimates of β) . . . in small samples is in accordance with the ordering by efficiency (of estimates of θ)" (p. 563). Rothenberg (1984) showed via second-order calculations that if g does not depend on β, when the data are normally distributed the covariance matrix of the generalized least squares estimator of β is an increasing function of the covariance matrix of the estimator of θ.

* M. Davidian is Assistant Professor, Department of Statistics, North Carolina State University, Raleigh, NC 27695-8203. R. J. Carroll is Professor, Department of Statistics, University of North Carolina, Chapel Hill, NC 27514. This work was supported by Air Force Office of Scientific Research Grant F-49620-85-C-0144.

© 1987 American Statistical Association
Journal of the American Statistical Association
December 1987, Vol. 82, No. 400, Theory and Methods

Second-order asymptotics provide only a weak justification for studying the properties of variance function estimates. Instead, our thesis is that estimation of the structural variance parameter θ is of independent interest. In many engineering applications, an important goal is to estimate the error made in predicting a new observation; this can be obtained from the variance function once a suitable estimate of θ is available. In chemical and biological assay problems, issues of prediction and calibration arise. In such problems, the estimator of θ plays a central role. As motivation for the study of variance function estimation, in Section 2 we discuss the problem of calibration and prediction in the case of heteroscedasticity. For a formal investigation of how the statistical properties of prediction intervals and calibration constructs, such as the minimal detectable concentration, are highly dependent on how one estimates θ, see Davidian, Carroll, and Smith (1987). In off-line quality control, the emphasis is not only on the mean response but also on its variability; Box and Meyer (1986) stated that "one distinctive feature of Japanese quality control improvement techniques is the use of statistical experimental design to study the effect of a number of factors on variance as well as the mean" (p. 19). The goal is to adjust the levels of a set of experimental factors to bring the mean of the responses to some target value while minimizing standard deviation; the problem involves simultaneous consideration of both mean and variability, where the latter may be a function of the mean (see Box 1986; Box and Ramirez 1986). These authors advocated methods based on data transformations to account for the heteroscedasticity in separating the factors into those affecting dispersion but not location, those affecting location but not dispersion, and those affecting neither. Similarly, one might employ effective estimation of variance functions in this application. We briefly discuss the relationship between variance function estimation and one type of data transformation in Section 3.

It should be evident from this brief review that, far from being only a nuisance parameter, the structural variance parameter θ can be an important part of a statistical analysis. The foregoing discussion suggests the need for a unified investigation of estimation of variance functions, in particular, estimation of the structural parameter θ. Previous work in the literature tends to treat various special cases of (1.1) as different models with their own estimation methods. The intent of this article is to study parametric variance function estimation in a unified way. Nonparametric variance function estimation has also been studied (see, e.g., Carroll 1982); we will confine our study to the parametric setting.

Parametric variance function estimation may be thought of as a type of regression problem in which we try to understand variance as a function of known or estimable quantities and in which θ plays the part of a "regression" parameter. The major insight that allows for a unified study is that the absolute residuals from the current fit to the mean or the sample standard deviations from replicates are basic building blocks for analysis. At the graphical level, this means that transformations of the absolute residuals and sample standard deviations can be used to gain insight into the structure of the variability and to suggest parametric models. For estimation, a major contribution is to point out that most of the methods proposed in the literature are (possibly weighted) regressions of transformations of the basic building blocks on their expected values. Many exceptions to this are dealt with in this article as well.

Our study yields these major qualitative conclusions. As stated here, they apply strictly only to symmetric error distributions, but they are fairly definitive, and one is unlikely to be too successful ignoring them in practice.

1. Robustness plays a great role in the efficiency of variance function estimation, probably even greater than in estimation of a mean function. For example, if the variance does not depend on the mean response, the standard method will be normal theory maximum likelihood, as in Box and Meyer (1986). A weighted analysis of absolute residuals yields an estimator only 12% less efficient at the normal model, which rapidly gains efficiency over maximum likelihood for progressively heavier tailed distributions. This slope of improvement is much larger than is typical for estimation of the mean response. For a standard contaminated normal model for which the best robust estimators have efficiency 125% with respect to least squares, the absolute residual estimator of the variance function has efficiency 200%.

2. We obtain implications for fit to the means upon which the residuals are based. It has been our experience that unweighted least squares residuals yield unstable estimates of the variance function when the variances depend on the mean. This is confirmed in our study, in the sense that the asymptotic efficiency of the variance function estimators is an increasing function of the efficiency of the current fit to the means. Thus we suggest the use of iterative weighted fitting, so the variance function estimate is based on generalized least squares residuals. As far as we can tell, this part of our article is one of the first formal justifications for iteration in a generalized least squares context.

3. It is standard in many applied fields to take m replicates at each design point, where usually $m \leq 4$. Rather than using (transformations of) absolute residuals for estimating variance function parameters, one might use the sample standard deviations. We develop an asymptotic theory from which we obtain the efficiency of this substitution. The effect is typically, although not always, a loss of efficiency, at least when there are $m \leq 4$ replicates. The clearest results occur when the variance does not depend on the mean. Normal theory maximum likelihood is a weighted regression of squared residuals; the corresponding method would be a weighted regression based on sample variances. Using the latter entails a loss of efficiency, no matter what the underlying distribution. For normally distributed data, the efficiency is $(m - 1)/m$, thus being only 50% for duplicates. For other methods, using the replicate standard deviations can be more efficient. This is particularly true of a method due to Harvey (1976), which is based on the logarithm of absolute residuals. A small absolute residual, which seems always to occur in

practice, can wreak havoc with this method. This is consistent with our influence function calculations, so we suggest some trimming of the smallest absolute residuals before applying Harvey's method.

4. Our results indicate that the precision of estimates of θ is approximately independent of σ. In addition, in the power of the mean model (1.2), the efficiency of a regression estimator improves as the relative range of values of the mean response increases; efficiency depends on the spread of the logarithms of means, not their actual values. This helps explain why in assays, estimating variances is typically much harder than estimating means.

In Section 2 we discuss the prediction and calibration problems as a motivating example of a situation in which variance function estimation is of key importance. In Section 3 we describe a number of methods for estimation of θ. We do not discuss robust methods (see Giltinan, Carroll, and Ruppert 1986). In Section 4 we present an asymptotic theory for a general estimator of θ whose construction encompasses the methods of Section 3. Section 5 contains examples of specific applications of our theory and a discussion of the implications of our formulation. Sketches of proofs are presented in Appendix A.

2. AN EXAMPLE: THE ROLE OF VARIANCE ESTIMATION IN PREDICTION AND CALIBRATION PROBLEMS

One example in which heterogeneity of variation occurs is in calibration experiments in the physical and biological sciences, in which one fits a model such as (1.1) to a sample $\{Y_i, x_i\}$ $(i = 1, \ldots, N)$. The $\{x_i\}$ may be concentrations of a substance and the $\{Y_i\}$ may be counts or intensity levels that vary with concentration. The interest lies in using the estimated regression to make inference about a pair $\{Y_0, x_0\}$, which is independent of the original data set. One may wish to obtain point and interval predictors for Y_0 in the case in which x_0 is known (prediction) or estimate x_0 in the case in which Y_0 only is known (calibration) (see Rosenblatt and Speigelman 1981). As a motivating example for considering estimation of variance functions as an independent problem, we describe the primary role of form and estimation of the variance function in construction of prediction/calibration intervals in the case of heteroscedasticity.

Throughout this discussion assume that $x_i \equiv z_i$ so that we may write the variance function as $g(x_i, \beta, \theta)$, and assume that the data are approximately normally distributed. Given x_0, the standard point estimate of the response Y_0 is $f(x_0, \hat{\beta})$, where $\hat{\beta}$ is some estimate for β. For any consistent estimator $\hat{\beta}$ of β, under (1.1) the variance in the error made by the prediction is, for large sample sizes, $\mathrm{var}\{Y_0 - f(x_0, \hat{\beta})\} \approx \sigma^2 g^2(x_0, \beta, \theta)$, so the error in prediction is determined mainly by the variance function $\sigma^2 g^2(x_0, \beta, \theta)$ and not the original data set itself. An approximate $(1 - \alpha)$ 100% confidence interval for Y_0 is $I(x_0) = \{$all Y in the interval $f(x_0, \hat{\beta}) \pm t_{1-\alpha/2}^{N-p}\, \hat{\sigma} g(x_0, \hat{\beta}, \hat{\theta})\}$; here $t_{1-\alpha/2}^{N-p}$ is the $(1 - \alpha/2)$ percentage point of the t distribution with $(N - p)$ degrees of freedom and $\hat{\sigma}$ and $\hat{\beta}$ are estimates. If the parameters are estimated by a weighted analysis, such as generalized least squares assuming (1.1), all estimates are consistent and the prediction interval becomes

$$I(x_0) \approx \{\text{all } Y \text{ in the interval } f(x_0, \beta) \pm t_{1-\alpha/2}^{N-p} \sigma g(x_0, \beta, \theta)\}. \quad (2.1)$$

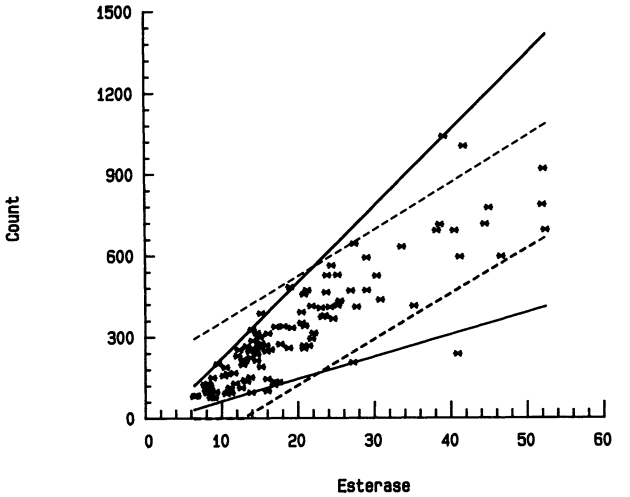

Figure 1. Approximate Form of Prediction Intervals for a Linear Mean Response Function Based on Unweighted (ignoring heteroscedasticity) and Weighted [as in (1.1)] Regression Fits. Esterase assay 95% prediction limits: dashed line—unweighted, solid line—weighted.

If one were to ignore the heterogeneity, the interval would be given by $I_U(x_0) = \{\text{all } Y \text{ in the interval } f(x_0, \hat{\beta}) \pm t_{1-\alpha/2}^{N-p}\hat{\sigma}\}$. For an unweighted analysis, however, σ^2 would be estimated by the unweighted mean squared error $\hat{\sigma}_U^2 \approx \sigma^2 N^{-1} \sum g^2(x_i, \beta, \theta) = \sigma^2 g_N^2$ for large N. Thus the unweighted prediction interval satisfies

$$I_U(x_0) \approx \{\text{all } Y \text{ in the interval } f(x_0, \beta) \pm t_{1-\alpha/2}^{N-p}\sigma g_N\}. \quad (2.2)$$

Comparing (2.1) and (2.2), we see that where the variability is small, the unweighted interval will be too long and hence pessimistic, and conversely where the variance is large. Figure 1 illustrates this phenomenon for the results of an assay for the concentration of an enzyme esterase, where the responses are binding counts in the simple situation of an approximately linear mean response function where variability increases with mean response.

The situation is the same for calibration. For simplicity in discussing calibration, assume that $f(x, \beta)$ is strictly increasing or decreasing in x. Given Y_0, the usual estimate of x_0 is that value satisfying $Y_0 = f(x, \hat{\beta})$. The common confidence interval for x_0 is the set of all x values for which Y_0 falls in the prediction interval $I(x)$; this interval is actually a $(1 - \alpha)$ 100% confidence interval for the unknown x_0. Since the confidence interval for x_0 is thus an inversion of the intervals in Figure 1, again, the effect of not weighting is intervals that are too long for x_0 when the variance is small and the opposite when the variance is large. We are not familiar with any extensive investigation of calibration confidence intervals for heteroscedastic models, but see Watters, Carroll, and Spiegelman (1987).

The key point of this discussion is that when heterogeneity of variance is present, how well one models and estimates the variances will have substantial impact on prediction and calibration based on the estimated mean response, since the form of the intervals depends on the form of the variance function. Some theoretical work has been done verifying the implications of this discussion; for an investigation of how the statistical properties of estimators for calibration quantities depend on those of the estimator θ, see Davidian et al. (1987) and Carroll (1987).

3. ESTIMATION OF θ

We now discuss the form and motivation for several estimators of θ in (1.1). In what follows, let $\hat{\beta}_*$ be a preliminary estimator for β. This could be unweighted least squares or the current estimate in an iterative reweighted least squares calculation. Let $\varepsilon_i = \{Y_i - f(x_i, \beta)\}/\{\sigma g(z_i, \beta, \theta)\}$ denote the errors so that $E\varepsilon_i = 0$ and $E\varepsilon_i^2 = 1$, and denote the residuals by $r_i = Y_i - f(x_i, \hat{\beta}_*)$. We consider some methods requiring $m_i \geq 2$ replicates at each of M design points; for simplicity, we consider only the case of equal replication $m_i \equiv m$ and write in obvious fashion $\{Y_{ij}\}$ $(j = 1, \ldots, m)$ to denote the m observations at x_i where appropriate, so that $N = Mm$ is the total number of observations. In this case, let $\bar{Y}_{i\cdot}$ and s_i denote the sample mean and standard deviation at x_i. For consistency of exposition, however, we denote the sum over all observations as $\sum_{i=1}^N$ instead of $\sum_{i=1}^M \sum_{j=1}^m$. When we speak of replacing absolute residuals $\{|r_i|\}$ by sample deviations $\{s_i\}$ in the case of replication, $|r_i|$ or s_i appears m times in the sum.

3.1 Regression Methods

3.1.1. Pseudolikelihood. Given $\hat{\beta}_*$, the pseudolikelihood estimator maximizes the normal log-likelihood $l(\hat{\beta}_*, \theta, \sigma)$, where

$$l(\beta, \theta, \sigma) = -N \log \sigma - \sum_{i=1}^N \log\{g(z_i, \beta, \theta)\} - (2\sigma^2)^{-1} \sum_{i=1}^N \{Y_i - f(x_i, \beta)\}^2/g^2(z_i, \beta, \theta) \quad (3.1)$$

(see Carroll and Ruppert 1982a). Here the term "pseudolikelihood" is used as in Gong and Samaniego (1981). Generalizations of pseudolikelihood for robust estimation have been studied by Carroll and Ruppert (1982a) and Giltinan et al. (1986).

3.1.2. Least Squares on Squared Residuals. Besides pseudolikelihood, other methods using squared residuals have been proposed. The motivation for these methods is that the squared residuals have approximate expectation $\sigma^2 g^2(z_i, \beta, \theta)$ (see Amemiya 1977; Jobson and Fuller 1980). This suggests a nonlinear regression problem in which the "responses" are $\{r_i^2\}$ and the "regression function" is $\sigma^2 g^2(z_i, \hat{\beta}_*, \theta)$. The estimator $\hat{\theta}_{SR}$ minimizes in θ and σ,

$$\sum_{i=1}^N \{r_i^2 - \sigma^2 g^2(z_i, \hat{\beta}_*, \theta)\}^2.$$

For normal data the squared residuals have approximate variance $\sigma^4 g^4(z_i, \beta, \theta)$; in the spirit of generalized least squares, this suggests the weighted estimator that minimizes in θ and σ,

$$\sum_{i=1}^N \{r_i^2 - \sigma^2 g^2(z_i, \hat{\beta}_*, \theta)\}^2/g^4(z_i, \hat{\beta}_*, \hat{\theta}_*), \quad (3.2)$$

where $\hat{\theta}_*$ is a preliminary estimator for θ, $\hat{\theta}_{SR}$, for example. Full iteration, when it converges, would be equivalent to pseudolikelihood.

3.1.3. Accounting for the Effect of Leverage. One objection to methods such as pseudolikelihood and least squares based on squared residuals is that no compensation is made for the loss of degrees of freedom associated with preliminary estimation of β. For example, the effect of applying pseudolikelihood directly seems to be a bias depending on p/N. For settings such as fractional factorials, where p is large relative to N, this bias could be substantial.

Bayesian ideas have been used to account for loss of degrees of freedom (see Harville 1977; Patterson and Thompson 1971). When g does not depend on β, the restricted maximum likelihood approach of Patterson and Thompson suggests in our setting one estimate θ from the mode of the marginal posterior density for θ assuming normal data and a prior for the parameters proportional

to σ^{-1}. When g depends on β, one may extend the Bayesian arguments and use a linear approximation as in Box and Hill (1974) and Beal and Sheiner (1987) to define a restricted maximum likelihood estimator.

Let Q be the $N \times p$ matrix with ith row $f_\beta(x_i, \beta)^t/g(z_i, \beta, \theta)$, where $f_\beta(x_i, \beta) = \partial/\partial\beta\{f(x_i, \beta)\}$, and let $H = Q(Q^tQ)^{-1}Q^t$ be the "hat" matrix with diagonal element $h_{ii} = h_{ii}(\beta, \theta)$; the values $\{h_{ii}\}$ are the leverage values. It turns out that the restricted maximum likelihood estimator is equivalent to an estimator obtained by modifying pseudolikelihood to account for the effect of leverage. This characterization, although not unexpected, is new; we derive this estimator and its equivalence to a modification of pseudolikelihood in Appendix B.

The least squares approach using squared residuals can also be modified to show the effect of leverage. Jobson and Fuller (1980) essentially noted that for nearly normally distributed data we have the approximations

$$Er_i^2 \approx \sigma^2(1 - h_{ii})g^2(z_i, \beta, \theta),$$

$$\text{var } r_i^2 \approx 2\sigma^4(1 - h_{ii})^2g^4(z_i, \beta, \theta).$$

To exploit these approximations modify (3.2) to minimize in θ and σ,

$$\sum_{i=1}^{N} \{r_i^2 - \sigma^2(1 - \hat{h}_{ii})g^2(z_i, \hat{\beta}_*, \theta)\}^2 \div \{(1 - \hat{h}_{ii})^2g^4(z_i, \hat{\beta}_*, \hat{\theta}_*)\}, \quad (3.3)$$

where $\hat{h}_{ii} = h_{ii}(\hat{\beta}_*, \hat{\theta}_*)$ and $\hat{\theta}_*$ is a preliminary estimator for θ. An asymptotically equivalent variation of this estimator in which one sets the derivatives of (3.3) with respect to θ and σ equal to 0 and then replaces $\hat{\theta}_*$ by θ can be seen to be equivalent to pseudolikelihood in which one replaces standardized residuals by studentized residuals. Although this estimator also takes into account the effect of leverage, it is different from restricted maximum likelihood.

3.1.4. Least Squares on Absolute Residuals. Squared residuals are skewed and long-tailed, which has led many authors to propose using absolute residuals to estimate θ (see Glejser 1969; Theil 1971). Assume that

$$E|Y_i - f(x_i, \beta)| = \eta g(z_i, \beta, \theta),$$

which is satisfied if the errors $\{\varepsilon_i\}$ are iid. Mimicking the least squares approach based on squared residuals, one obtains the estimator $\hat{\theta}_{AR}$ by minimizing in η and θ,

$$\sum_{i=1}^{N} \{|r_i| - \eta g(z_i, \hat{\beta}_*, \theta)\}^2.$$

In analogy to (3.2), the weighted version is obtained by minimizing

$$\sum_{i=1}^{N} \{|r_i| - \eta g(z_i, \hat{\beta}_*, \theta)\}^2/g^2(z_i, \hat{\beta}_*, \hat{\theta}_*),$$

where $\hat{\theta}_*$ is a preliminary estimator for θ, probably $\hat{\theta}_{AR}$. As for least squares estimation based on squared residuals, one presumably could modify this approach to account for the effect of leverage.

3.1.5. Logarithm Method. The suggestion of Harvey (1976) is to exploit the fact that the logarithm of the absolute residuals has approximate expectation $\log\{\sigma g(z_i, \beta, \theta)\}$. Estimate θ by ordinary least squares regression of $\log|r_i|$ on $\log\{\sigma g(z_i, \hat{\beta}_*, \theta)\}$, since if the errors are iid, the regression should be approximately homoscedastic. If one of the residuals is near 0, the regression could be adversely affected by a large "outlier"; hence in practice one might wish to delete a few of the smallest absolute residuals, perhaps trimming the smallest few percent.

3.2 Other Methods

Besides squares and logarithms of absolute residuals, other transformations could be used. For example, the square root and $\frac{2}{3}$ root would typically be more normally distributed than the absolute residuals themselves. Such transformations appear to be useful, although they have not been used much to our knowledge. Our asymptotic theory applies to such transformations.

In a parametric model such as (1.1), joint maximum likelihood estimation is possible, where we use the term maximum likelihood to mean normal theory maximum likelihood. When the variance function does not depend on β, it can be easily shown that maximum likelihood is asymptotically equivalent to weighted least squares methods based on squared residuals. In the situation in which the variance function depends on β this is not the case. In this setting, it has been observed by Carroll and Ruppert (1982b) and McCullagh (1983) that, although maximum likelihood estimators enjoy asymptotic optimality when the model and distributional assumptions are correct, the maximum likelihood estimator of β can suffer problems under departures from these assumptions. This suggests that joint maximum likelihood estimation should not be applied blindly in practice. The theory of the next section shows the asymptotic equivalence of maximum likelihood with other methods in a simplifying special case. Based on this theory, we tend to prefer weighted regression methods even when the data are approximately normal for reasons of relative computational simplicity.

Although we have chosen to describe the methods of Section 3.1 as "regression methods," asymptotically equivalent versions of such methods may be derived by considering maximum likelihood assuming some underlying distribution. For example, the form of the weighted squared residuals method is that of normal theory maximum likelihood with β known and $\hat{\theta}_*$ replaced by θ (pseudolikelihood); the form of the weighted absolute residual method is that of maximum likelihood assuming β known and $\hat{\theta}_*$ replaced by θ under the double exponential distribution. Thus what we term a regression method may be viewed as an approximation to maximum likelihood assuming a particular distribution. We feel that the regression interpretation is a much more appealing and natural motivation, since no particular distribution need be considered

Table 1. Description of Some Methods for Variance Function Estimation

Maximum likelihood	Normal theory maximum likelihood in β, σ, θ.
Pseudolikelihood	Normal theory maximum likelihood when β is set to current value. When iterated, equivalent to maximum likelihood if the variance does not depend on β.
Weighted squared residuals	Regress squared residuals on the variance, function, weight inversely with squared current variance estimate.
Weighted absolute residuals	Regress absolute residuals on the standard deviation function, weight inversely with current variance estimate.
Logarithm method	Regress logarithm of absolute residuals on log of standard deviation function. Be wary of near-zero residuals.
Restricted maximum likelihood	Pseudolikelihood corrected for leverage. Maximizes marginal posterior for noninformative prior.
All of the preceding except restricted maximum likelihood have analogs formed by replacing absolute residuals by sample standard deviations in the case of replication. The following are based on the mean function or design being fully or partially unknown and are often used in assays.	
Rodbard and Frazier	Regress log sample standard deviation on log sample mean, where the variance function depends on β only through the means.
Modified maximum likelihood	Modified functional maximum likelihood [Eq. (2.5)], where variance function depends on β only through means.
Sadler and Smith	Same as modified maximum likelihood, but means estimated by sample means.

to obtain the form of the estimators, only the mean-variance relationship.

Another joint estimation method is the extended quasi-likelihood of Nelder and Pregibon (1987) also described in McCullagh and Nelder (1983). This estimator is based on assuming a class of distributions "nearly" containing skewed distributions, such as the Poisson and gamma. Although it may be viewed as iteration between estimation of θ and σ and generalized least squares for β, technically this scheme does not fit in the general framework of the next section: an asymptotic theory was developed elsewhere (see Davidian and Carroll, in press). A related formulation was given by Efron (1986).

Methods requiring replicates at each design point have been proposed in the assay literature. These methods do not depend on the postulated form of the regression function; one reason that this may be advantageous is that in many assays, along with observed pairs (Y_{ij}, x_i), there will also be pairs in which only Y_{ij} is observed. A popular and widely used method is that of Rodbard and Frazier (1975). If we assume that

$$g(z_i, \beta, \theta) = g(\mu_i, z_i, \theta), \tag{3.4}$$

as in, for example, (1.2) or (1.3), the method is identical to the logarithm method previously discussed except that one replaces $|r_i|$ by the sample standard deviation s_i and $f(x_i, \hat{\beta}_*)$ in the "regression" function by the sample mean $\bar{Y}_{i\cdot}$. As a motivation for this and the method of Harvey, consider that under (1.2) θ is simply the slope parameter for a simple linear regression.

As an alternative, under the assumption of independence and (3.4), the modified maximum likelihood method of Raab (1981) estimates θ by joint maximization in the $(M + r + 1)$ parameters $\sigma^2, \theta, \mu_1, \ldots, \mu_M$ of the "modified" normal likelihood

$$\prod_{i=1}^{M} \{2\pi\sigma^2 g^2(\mu_i, z_i, \theta)\}^{(m-1)/2} \times \exp\left[-\sum_{j=1}^{m} (Y_{ij} - \mu_i)^2/\{2\sigma^2 g^2(\mu_i, z_i, \theta)\}\right]. \tag{3.5}$$

The modification serves to make the estimator of σ unbiased. The idea here is to improve upon the regression method of Rodbard by appealing to a maximum likelihood approach that, despite a parameter space increasing as the number of design points, is postulated to have reasonable properties. A related method is that in which θ and σ are estimated by maximizing (3.5) with μ_i replaced by $\bar{Y}_{i\cdot}$, the motivation being computational ease and evidence that this estimator may not be too different from that of Raab in practice (see Sadler and Smith 1985).

Table 1 contains a summary of some of the common methods for variance function estimation and their formulations.

4. AN ASYMPTOTIC THEORY OF VARIANCE FUNCTION ESTIMATION

In this section we construct an asymptotic theory for a general class of regression-type estimators for θ. Since our major interest lies in obtaining general insights, we do not state technical assumptions or details. In what follows, in the case of replication $N \to \infty$ in such a way that m remains fixed. The reader uninterested in this development may wish to review the definition of the form of the estimators in the first two paragraphs of Section 4.1 and then skip to Section 5, where conclusions and implications of the theory are presented.

4.1 Methods Based on Transformations of Absolute Residuals

Write $d_i(\beta) = |Y_i - f(x_i, \beta)|$. Let T be a smooth function and define M_i by

$$M_i(\eta, \theta, \beta) = E[T\{d_i(\beta)\}],$$

where η is a scale parameter that is usually a function of σ only. We consider estimation of the more general parameter η instead of σ itself for ease of exposition, and since σ is estimated jointly with θ in regression methods, our theory focuses on expansions for η and θ jointly. If $\hat{\eta}_*$, $\hat{\theta}_*$, and $\hat{\beta}_*$ are any preliminary estimators for η, θ, and β, define $\hat{\eta}$ and θ to be the solutions of

$$N^{-1/2} \sum_{i=1}^{N} H_i(\eta, \theta, \hat{\beta}_*)\{T\{d_i(\hat{\beta}_*)\} - M_i(\eta, \theta, \hat{\beta}_*)\} \div V_i(\hat{\eta}, \hat{\theta}_*, \hat{\beta}_*) = 0, \tag{4.1}$$

where $V_i(\eta, \theta, \beta)$ is a smooth function and H_i is a smooth function that for the estimators of Section 3 is the partial derivative of M_i with respect to (η, θ). In what follows, we suppress the arguments of the functions M_i, V_i, etcetera when they are evaluated at the true values η, θ, and β. Specific examples are considered in the next section.

The class of estimators solving (4.1) includes directly or includes an asymptotically equivalent version of the estimators of Section 3.1. For methods that account for the effect of leverage, M_i, V_i, and H_i will depend on the h_{ii}. In this case we need the additional assumption that if $h = \max\{h_{ii}\}$, then $N^{1/2}h$ converges to 0.

Theorem 4.1. Let $\hat{\eta}_*$, $\hat{\theta}_*$, and $\hat{\beta}_*$ be $N^{1/2}$ consistent for estimating η, θ, and β. Let $\dot{T}$ be the derivative of T, and define

$$C_i = H_i[T\{d_i(\beta)\} - M_i]/V_i,$$

$$B_{1,N} = N^{-1}\sum_{i=1}^{N} H_i H_i'/V_i,$$

$$B_{2,N} = -N^{-1}\sum_{i=1}^{N} (H_i/V_i)\partial/\partial\beta\{M_i(\eta, \theta, \beta)\},$$

$$B_{3,N} = -N^{-1}\sum_{i=1}^{N} (H_i/V_i) f_\beta(x_i, \beta) E[\dot{T}\{d_i(\beta)\}\mathrm{sign}(\varepsilon_i)].$$

Then, under regularity conditions as $N \to \infty$,

$$B_{1,N}N^{1/2}\begin{bmatrix}\hat{\eta} - \eta \\ \hat{\theta} - \theta\end{bmatrix} = N^{-1/2}\sum_{i=1}^{N} C_i + (B_{2,N} + B_{3,N})N^{1/2}(\hat{\beta}_* - \beta) + o_p(1). \quad (4.2)$$

We may immediately make some general observations about the estimator $\hat{\theta}$ solving (4.1). Note that if the variance function does not depend on β, then M_i does not depend on β and hence $B_{2,N} \equiv 0$. For the estimators of Section 2.1, $\dot{T}$ is an odd function. Thus, if the errors $\{\varepsilon_i\}$ are symmetrically distributed, $E[\dot{H}_i\{d_i(\beta)\}\mathrm{sign}(\varepsilon_i)] = 0$ and hence $B_{3,N} \equiv 0$.

Corollary 4.1(a). Suppose that the variance function does not depend on β and the errors are symmetrically distributed. Then the asymptotic distributions of the regression estimators of Section 3.1 do not depend on the method used to obtain $\hat{\beta}_*$. If both of these conditions do not hold simultaneously, then the asymptotic distributions will depend in general on the method of estimating β.

The implication is that in the situation for which the variance function does not depend on β and the data are approximately symmetrically distributed, for large sample sizes the preliminary estimator for β will play little role in determining the properties of $\hat{\beta}$. Note also from (4.2) that for weighted methods, the effect of the preliminary estimator of θ is asymptotically negligible regardless of the underlying distributions.

The preliminary estimator $\hat{\beta}_*$ might be the unweighted least squares estimator, a generalized least squares estimator, or some robust estimator. See, for example, Huber (1981) and Giltinan et al. (1986) for examples of robust estimators for β. For some vectors $\{v_{N,i}\}$, these estimators admit an asymptotic expansion of the form

$$N^{1/2}(\hat{\beta}_* - \beta) = N^{-1/2}\sum_{i=1}^{N} \Psi(v_{N,i}, \varepsilon_i) + o_p(1). \quad (4.3)$$

Here Ψ is odd in the argument ε. In case the variance function depends on β, $B_{2,N} \neq 0$ in general; however, if the errors are symmetrically distributed and $\hat{\beta}_*$ has expansion of form (4.3), then the two terms on the right side of (4.2) are asymptotically independent. The following is then immediate.

Corollary 4.1(b). Suppose that the errors are symmetrically distributed and that $\hat{\beta}_*$ has an asymptotic expansion of the form (4.3). Then for the estimators of Section 3.1, the asymptotic covariance matrix of $\hat{\theta}$ is a monotone nondecreasing function of the asymptotic covariance matrix of $\hat{\beta}_*$.

By the Gauss–Markov theorem and the results of Jobson and Fuller (1980) and Carroll and Ruppert (1982a), the implication of Corollary 4.1(b) is that using unweighted least squares estimates of β will result in inefficient estimates of θ. This phenomenon is exhibited in small samples in a Monte Carlo study of Davidian et al. (1987). If one starts from the unweighted least squares estimate, one ought to iterate the process of estimating θ—use the current value $\hat{\beta}_*$ to estimate θ from (4.1), use these $\hat{\beta}_*$ and $\hat{\theta}$ to obtain an updated $\hat{\beta}_*$ by generalized least squares, and repeat the process $\mathcal{C} - 1$ more times. It is clear that the asymptotic distribution of $\hat{\theta}$ will be the same for $\mathcal{C} \geq 2$ with larger asymptotic covariance for $\mathcal{C} = 1$, so in principle one ought to iterate this process at least twice. See Carroll, Wu, and Ruppert (1987) for more on iterating generalized least squares.

4.2 Methods Based on Sample Standard Deviations

Assume replication, and as before let $\{s_i\}$ be the sample standard deviations at each x_i, which themselves have been proposed as estimators of the variance in generalized least squares estimation of β. This can be disastrous (see Jacquez, Mather, and Crawford 1968). When replication exists, however, practitioners feel comfortable with the notion that the $\{s_i\}$ may be used as a basis for estimating variances; thus one might reasonably seek to estimate θ by replacing $d_i(\hat{\beta}_*)$ by s_i in (4.1).

The following result is almost immediate from the proof of Theorem 4.1 in Appendix A.

Theorem 4.2. If $d_i(\hat{\beta}_*)$ is replaced by s_i in (4.1), then under the conditions of Theorem 4.1 the resulting estimator for θ satisfies (4.2) with $B_{3,N} \equiv 0$ and the redefinitions

$$C_i = (H_i/V_i)\{T(s_i) - M_i\}, \quad (4.4a)$$

$$M_i = E\{T(s_i)\} = M_i(\eta, \theta, \beta). \quad (4.4b)$$

If the errors are symmetrically distributed, then, from (4.2) and Theorem 4.2, whether one is better off using absolute residuals or sample standard deviations in the

methods of Section 3.1 depends only on the differences between the expected values and variances of $T\{d_i(\beta)\}$ and $T(s_i)$. In Section 5 we exhibit such comparisons explicitly and show that absolute residuals can be preferred to sample standard deviations in situations of practical importance.

4.3 Methods Not Depending on the Regression Function

We assume throughout this discussion that the variance function has form (3.4) and replication is available. From Section 3.1 we see that the "regression function" part of the estimating equations depends on $f(x_i, \hat{\beta}_*)$, so in the general equation (4.1) M_i, V_i, and H_i all depend on $f(x_i, \hat{\beta}_*)$. In some settings, one may not postulate a form for the μ_i for estimating θ; the method of Rodbard and Frazier (1975), for example, uses s_i in place of $d_i(\hat{\beta}_*)$ as in Section 4.2 and replaces $f(x_i, \hat{\beta}_*)$ by the sample mean $\overline{Y}_{i\cdot}$. We now consider the effect of replacing predicted values by sample means for the general class (4.1).

The presence of the sample means in the variance function in (4.1) requires more complicated and restrictive assumptions than the usual large sample asymptotics applied heretofore. The method of Rodbard and Frazier and the general method (4.1) with sample means are nonlinear errors-in-variables problems as studied by Wolter and Fuller (1982) and Stefanski and Carroll (1985). Standard asymptotics for these problems correspond to letting σ go to 0 at rate $N^{-1/2}$. In Section 4.4 we discuss the practical implications of σ being small; for now, we state the following result.

Theorem 4.3. Suppose that we replace $f(x_i, \hat{\beta}_*)$ by $\overline{Y}_{i\cdot}$ in M_i, V_i, and H_i in (4.1) and adopt the assumptions of Theorems 4.1 and 4.2. Further, suppose that as $N \to \infty$, $\sigma \to 0$ simultaneously and

(i) $N^{1/2}\sigma \to \lambda$, $0 \le \lambda < \infty$;
(ii) $N^{1/2} \sum_{i=1}^{N} C_i$ has a nontrivial asymptotic normal limit distribution;
(iii) The $\{\varepsilon_i\}$ are symmetric and iid;
(iv) $\{|\overline{Y}_{i\cdot} - \mu_i|/\sigma\}^2$ has uniformly bounded k moments, some $k > 2$.

Then the results of Theorems 4.1 and 4.2 hold with $B_{2,N} = B_{3,N} \equiv 0$.

This result shows that under certain restrictive assumptions, one may replace predicted values by sample means under replication; it is important to realize, however, that the assumption of small σ is not generally valid and hence the use of sample means may be disadvantageous in situations where these asymptotics do not apply. Further, relaxation of Assumption (iii) will result in an asymptotic bias in the asymptotic distribution of the estimator not present for estimators based on residuals regardless of the assumption of symmetry (see App. A).

The estimator of Raab (1981) discussed in Section 3.2 is also a functional nonlinear error-in-variables estimator, complicated by a parameter space with size of order N. Sadler and Smith (1985) observed that the Raab estimator is often indistinguishable from their estimator with μ_i replaced by $\overline{Y}_{i\cdot}$ in (3.5); such an estimator is contained in the general class (4.1). Davidian (1986) showed that under the asymptotics of Theorem 4.3 and additional regularity conditions the two estimators are asymptotically equivalent in an important special case. We may thus consider the result of Theorem 4.3 relevant to this estimator.

4.4 Small σ Asymptotics

In Section 4.3 technical considerations forced us to pursue an asymptotic theory in which σ is small. It turns out that in some situations of practical importance these asymptotics are relevant. In particular, in assay data we have observed values for σ that are quite small relative to the means. Such asymptotics are used in the study of data transformations in regression. It is thus worthwhile to consider the effect of small σ on the results of Sections 4.1 and 4.2 and to comment on some other implications of letting $\sigma \to 0$.

In the situation of Theorem 4.1, if the errors are symmetrically distributed, then for the estimators of Section 3.1, if $\sigma \to 0$ as $N \to \infty$, then there is no effect for estimating the regression parameter β. In the situation of Theorem 4.2, the errors need not even be symmetrically distributed. The major insight provided by these results is that in certain practical situations in which σ is small, the choice of $\hat{\beta}_*$ may not be too important even if the variance function depends on β.

Small σ asymptotics may be used to provide insight into the behavior of other estimators for θ that do not fit into the general framework of (4.1). It can be shown that the extended quasilikelihood estimator need not necessarily be consistent for fixed σ, but if one adopts the asymptotics of the previous section, this estimator is asymptotically equivalent to regression estimators based on squared residuals as long as the errors are symmetrically distributed. Otherwise, an asymptotic bias may result, which may have implications for inference for θ. For discussion see Davidian and Carroll (in press).

The small σ assumption also provides an illustration of the relationship between variance function estimation and data transformations. Let $l(y, \varphi) = (y^\varphi - 1)/\varphi$, and consider the model

$$E\{l(Y_i, \varphi)\} = l(f(x_i, \beta), \varphi\}, \qquad \mathrm{var}\{l(Y_i, \varphi)\} = \sigma; \tag{4.5}$$

such "transform both sides" models are proposed and motivated by Carroll and Ruppert (1984). For $\sigma \approx 0$, $E(Y_i) \approx f(x_i, \beta)$ and $\mathrm{var}(Y_i) \approx \sigma f(x_i, \beta)^{(1-\varphi)}$, so in (1.2) we have $\theta \approx 1 - \varphi$. Thus, when the small σ assumption is relevant, (4.5) and (1.1), (1.2) represent approximately the same model.

5. APPLICATIONS AND FURTHER RESULTS

In Section 4 we constructed an asymptotic theory for and stated some general characteristics of regression-type estimators of θ. In this section we use the theory to exhibit the specific forms for the various estimators of Section 3

and compare and contrast their properties. In our investigation we rely on the simplifying assumptions implied by the theory of Section 4, in particular the small σ asymptotic approach in which $\sigma \to 0$ and $N \to \infty$. Throughout, define $v(i, \beta, \theta) = \log g(z_i, \beta, \theta)$, let $v_\theta(i, \beta, \theta)$ be the column vector of partial derivatives of v with respect to θ, let $\xi(\beta, \theta)$ be the covariance matrix of $v_\theta(i, \beta, \theta)$, and let $\tau(i, \beta, \theta) = \{1, v_\theta^t(i, \beta, \theta)\}^t$. For simplicity, assume that the errors $\{\varepsilon_i\}$ are iid with kurtosis κ; $\kappa = 0$ for normality.

5.1 Maximum Likelihood, Pseudolikelihood, Restricted Maximum Likelihood, and Weighted Squared Residuals

Writing $\eta = \log \sigma$, we have $T(x) = x^2$, $M_i = \exp(2\eta)g^2(z_i, \beta, \theta)$, $V_i = M_i^2$, $H_i^t = \partial M_i/\partial(\eta, \theta^t)^t$, and $E[\dot{T}\{d_i(\beta)\}\text{sign}(\varepsilon_i)] = 2E[Y_i - f(x_i, \beta)] = 0$, so $B_{3,N} \equiv 0$ regardless of the underlying distributions. If $h \to 0$ such that $N^{1/2}h \to 0$ for methods accounting for the effect of leverage, then all of these methods admit an expansion of the form (4.2) with $B_{3,N} = 0$. The expansion will be different depending on whether $\hat{\beta}_*$ is a generalized least squares estimator for β or full maximum likelihood, since the maximum likelihood estimator has an expansion quadratic in the errors and that of the generalized least squares estimator is linear in the $\{\varepsilon_i\}$ (see Carroll and Ruppert 1982b). The implication is that regression methods based on iterated weighted squared residuals and full maximum likelihood are different in general asymptotically. Regardless of the underlying distributions, for fixed σ, Davidian (1986) showed that the asymptotic covariance matrix of the former methods increases without bound as a function of σ whereas that of maximum likelihood remains bounded for all σ. Further, a simple comparison of the two covariances reveals that under reasonable conditions maximum likelihood has smaller asymptotic covariance as long as $\kappa \le 2$. Although these facts may suggest a preference for full maximum likelihood even away from normality, the computational and model robustness considerations mentioned earlier may make this preference tenuous. Generalized least squares and maximum likelihood estimators for β both satisfy $\hat{\beta}_* - \beta = O_p(\sigma N^{-1/2})$, so if $\sigma \to 0$ or g does not depend on β, then $\hat{\theta}$ is asymptotically normally distributed with mean θ and covariance matrix

$$(2 + \kappa)\{4N\xi(\beta, \theta)\}^{-1}. \tag{5.1}$$

As mentioned in Section 3, under the small σ asymptotics of Theorem 3.3, the extended quasilikelihood estimator of θ is asymptotically equivalent to the estimators here with asymptotic covariance matrix (5.1). Thus, if g does not depend on β or $\sigma \to 0$, pseudolikelihood, weighted squared residuals, restricted maximum likelihood, maximum likelihood and, if $\sigma \to 0$, extended quasilikelihood, are all asymptotically equivalent. In addition, all of these estimators have influence functions that are linear in the squared errors, indicating substantial nonrobustness.

We may also observe that these methods are preferable to unweighted regression on squared residuals. Write (5.1) as

$$(\tfrac{1}{2} + \kappa/4)(WV^{-1}W)^{-1}, \tag{5.2}$$

where V is the $N \times N$ diagonal matrix with elements V_i and W is the $N \times p$ matrix with ith row H_i^t. For the unweighted estimator based on squared residuals, calculations similar to those above show that the asymptotic covariance matrix when either g does depend on β or $\sigma \to 0$ is given by

$$(\tfrac{1}{2} + \kappa/4)(W^tW)^{-1}(W^tVW)(W^tW)^{-1}. \tag{5.3}$$

The comparison between (5.2) and (5.3) is simply that of the Gauss–Markov theorem, so (5.2) is no larger than (5.3).

5.2 Logarithms of Absolute Residuals and the Effect of Inliers

We do not consider deletion of the few smallest absolute residuals. Here $T(x) = \log x$, so $\dot{T}(x) = x^{-1}$. Letting $\eta = \log \sigma$ and assuming iid errors we have $M_i = \eta + v(i, \beta, \theta) + E \log|\varepsilon|$, $V_i \equiv 1$, and $H_i = \tau(i, \beta, \theta)$. Under the assumption of symmetry of the errors, with g not depending on β or $\sigma \to 0$, tedious algebra shows that $\hat{\theta}$ is asymptotically normally distributed with mean θ and covariance matrix

$$\text{var}\{\log(|\varepsilon|^2)\}\{4N\xi(\beta, \theta)\}^{-1}. \tag{5.4}$$

The influence function for this estimator is linear in the logarithm of the absolute errors. This indicates nonrobustness more for inliers than for outliers, which at the very least is an unusual phenomenon. If the errors are not symmetric, then there will be an additional effect due to estimating β not present for the methods of Section 5.1, even if g does not depend on β.

5.3 Weighted Absolute Residuals

Assume that the errors are iid, and let $\exp(\eta) = \sigma E|\varepsilon|$. Consider the weighted estimator. We have $T(x) = x$, $\dot{T}(x) = 1$, $M_i = \exp(\eta)g(z_i, \beta, \theta)$, and $V_i = M_i^2$. Thus, if the errors are symmetrically distributed and either g does not depend on β or $\sigma \to 0$, $\hat{\theta}$ is asymptotically normally distributed with mean θ and covariance matrix

$$\{\delta/(1 - \delta)\}\{N\xi(\beta, \theta)\}^{-1}, \tag{5.5}$$

where $\delta = \text{var}|\varepsilon|$. The influence function for this estimator is linear in the absolute errors. By an argument similar to that at the end of Section 5.1, we may conclude that when the effect of $\hat{\beta}_*$ is negligible one should use a weighted estimator and iterate the method.

5.4 General Transformations

One may also consider other power transformations of absolute residuals. If $\lambda \neq 0$ is the power of absolute residuals on which the regression is based, then define η by $\exp(\lambda\eta) = \sigma^\lambda E(|\varepsilon|^\lambda)$ and $T(x) = x^\lambda$. Then $M_i = \exp(\lambda\eta)g^\lambda(z_i, \beta, \theta)$, $V_i = M_i^2$. Straightforward calculations show that if the errors are symmetric and either g does not depend on β or $\sigma \to 0$, then $\hat{\theta}$ is asymptotically normally distributed with mean θ and asymptotic covariance matrix

$$[\text{var}(|\varepsilon|^\lambda)/\{E(|\varepsilon|^\lambda)\}^2]\{\lambda^2 N\xi(\beta, \theta)\}^{-1}, \tag{5.6}$$

with influence function linear in $|\varepsilon|^\lambda$. Thus (5.6) yields (5.1)

when $\lambda = 2$ and (5.5) when $\lambda = 1$. For square root transformations, for example, $\lambda = \frac{1}{2}$, and from (5.1) and (5.6), the asymptotic relative efficiency of the square root transformation relative to pseudolikelihood under normal errors is .693; from (5.5), the efficiency relative to weighted absolute residuals is .791.

At this point it is worthwhile to mention that under the simplifying assumptions of our discussion, the precision of general regression estimators does not depend on σ, since a general expression such as (5.6) is independent of η. Thus how well we estimate θ in many practical cases will be approximately independent of σ. Furthermore, when the power of the mean model for variance (1.2) holds, $v_\theta(i, \beta, \theta) = \log \mu_i$, so $\xi(\beta, \theta)$ is the limiting variance of the $\{\log \mu_i\}$. From the general expression (5.6), the precision with which one can estimate θ depends only on the relative spread of the mean responses, not their actual sizes, and clearly this spread must be fairly substantial so that the spread of the logarithms of the means will be so as well. The implications are that for (1.2), the design will play an important role in efficiency of estimation of θ, and in some practical situations we may not be able to estimate θ well no matter which estimator we employ.

5.5 Comparison of Methods Based on Residuals

We assume that the errors are symmetric and iid and that either g does not depend on β or σ is small. By (5.1), (5.4), and (5.5), the asymptotic relative efficiency of the three methods depends only on the distribution of the errors. For normal errors, using absolute residuals results in a 12% loss in efficiency, whereas for standard double exponential errors there is a 25% gain in efficiency for using absolute residuals. For normal errors, the logarithm method represents a 59% loss of efficiency with respect to pseudolikelihood.

In Table 2 we present asymptotic relative efficiencies for various contaminated normal distributions. The asymptotic efficiency of the weighted absolute residual method to pseudolikelihood is the same as the asymptotic relative efficiency of the mean absolute deviation with respect to the sample variance for a single sample (see Huber 1981, p. 3); the first column of the table is thus identical to that of Huber. The table shows that, although at normality neither the absolute residuals nor the logarithm methods are efficient, a very slight fraction of "bad" observations is enough to offset the superiority of squared residuals in a dramatic fashion. For example, just 2 bad observations in 1,000 negate the superiority of squared residuals. If 1% or 5% of the data are "bad," absolute residuals and the logarithm method, respectively, show substantial gains over squared residuals. The implication is that, although it is commonly perceived that methods based on squared residuals are to be preferred in general, these methods can be highly nonrobust. Our formulation includes this result for maximum likelihood, showing its inadequacy under slight departures from the assumed distributional structure. We also include asymptotic relative efficiencies for appropriately weighted residual methods based on square, cube, and $\frac{2}{3}$ roots to pseudolikelihood using (5.6) and observe that these methods also exhibit comparative robustness to contamination.

Table 2. Asymptotic Relative Efficiency of Appropriately Weighted Regression Methods Based on a Function T of Absolute Residuals and the Method Based on Logarithms of Absolute Residuals With Respect to Appropriately Weighted Regression Methods Based on Squared Residuals for Underlying Contaminated Normal Error Distributions With Distribution Function $F(x) = (1 - \alpha)\Phi(x) + \alpha\Phi(x/3)$

Contamination fraction α	$T(x)$				
	x	$x^{2/3}$	$x^{1/2}$	$x^{1/3}$	$\log x$
.000	.876	.772	.693	.606	.405
.001	.948	.841	.756	.662	.440
.002	1.016	.906	.816	.715	.480
.010	1.439	1.334	1.216	1.075	.720
.050	2.035	2.100	1.996	1.823	1.220

5.6 Methods Based on Sample Standard Deviations

Assume that $m \geq 2$ replicate observations are available at each design point. In practice, m is usually small (see Raab 1981). We compare using absolute residuals with using sample standard deviations in the estimators of Section 3.1. One advantage of sample standard deviations over absolute residuals is that, because they do not use the mean function, they will be robust to misspecification of the model for the mean response; absolute residuals will not. We assume that one is fairly confident in the postulated form of the model, thus viewing methods based on sample standard deviations as not taking full advantage of the information available. For simplicity, assume that the errors are iid and symmetrically distributed and that either g does not depend on β or σ is small. If the errors are not symmetric and σ is not small or the variance depends on β, using sample standard deviations presumably will be more efficient than in the discussion that follows. This issue deserves further attention.

Let s_m^2 be the sample variance of m errors $\{\varepsilon_1, \ldots, \varepsilon_m\}$. It is easily shown by calculations analogous to those of Section 5.1 that replacing absolute residuals by sample standard deviations has the effect of changing the asymptotic covariance matrices (5.1), (5.4), and (5.5) to

$$\text{Pseudolikelihood: } \{(2 + \kappa) + 2/(m - 1)\}\{4N\xi(\beta, \theta)\}^{-1};$$

$$\text{Logarithm method: } m \operatorname{var}\{\log(s_m^2)\}\{4N\xi(\beta, \theta)\}^{-1};$$

Weighted absolute residuals:

$$\{m\delta_*/(1 - \delta_*)\}\{N\xi(\beta, \theta)\}^{-1}, \quad (5.7)$$

where $\delta_* = \operatorname{var}(s_m)$. Table 3 compares the asymptotic relative efficiencies of using sample standard deviations with using transformations of absolute residuals for various values of m when the errors are standard normal. The values in the table for $T(x) = x^2$ and x indicate that if the data are approximately normally distributed, using sample standard deviations can entail a loss in efficiency with respect to using residuals if m is small. For substantial replication ($m \geq 10$), using sample standard deviations pro-

Table 3. Asymptotic Relative Efficiency of Regression Methods Based on a Function T of Sample Standard Deviations Relative to Using Regression Methods Based on a Function T of Absolute Residuals under Normality for T(x) (weighted methods)

	T(x)		
m	x^2	log x	x
2	.500	.500	.500
3	.667	1.000	.696
4	.750	1.320	.801
⋮	⋮	⋮	⋮
9	.889	1.932	.986
10	.900	1.984	1.001
∞	1.000	2.467	1.142

duces a slight edge in efficiency with respect to weighted absolute residuals for $T(x) = x$.

The second column of Table 3 shows that, for the logarithm method, using sample standard deviations surpasses using residuals in terms of efficiency except when $m = 2$ and is more than twice as efficient for large m. In its raw form, $\log|r_i|$ is very unstable because, at least occasionally, $|r_i| \approx 0$, producing a wild "outlier" in the regression. The effect of using sample standard deviations is to decrease the possibility of such inliers; the sample standard deviations will likely be more uniform, especially as m increases. The implication is that the logarithm method should not be based on residuals unless remedial measures are taken. The suggestion to trim a few of the smallest absolute residuals before using this method is clearly supported by the theory; presumably, such trimming would reduce or negate the theoretical superiority of using sample standard deviations.

Table 4 contains the asymptotic relative efficiencies of weighted squared sample standard deviations and logarithms of these to weighted squared residuals under normality of the errors. The first column is the efficiency of Raab's method to pseudolikelihood, and the second column is the efficiency of the Rodbard and Frazier method to pseudolikelihood. The results of the table imply that using the Raab and Rodbard and Frazier methods, which are popular in the analysis of radioimmunoassay data, can entail a loss of efficiency when compared with methods based on weighted squared residuals. Davidian (1986) showed that the Rodbard and Frazier estimator can have a slight edge in efficiency over the weighted squared residuals methods for some highly contaminated normal distributions. From (5.7), the squared residual methods will be more efficient than Raab's method in the limit. Also note that the entries for $T(x) = x$ and $\log x$ in Table 3 for $m = \infty$ are the reciprocals of the first row of Table 2 and that the entries for the last row of Table 4 are 1.0; thus if both N and m grow large all the methods yield the same results.

Table 4. Asymptotic Relative Efficiency of Regression Methods Based on a Function T of Sample Standard Deviations Relative to Regression Methods Based on Weighted Squared Residuals Under Normal Errors

	T(x)	
m	x^2	log x
2	.500	.203
3	.667	.405
4	.750	.535
⋮	⋮	⋮
9	.889	.783
10	.900	.804
∞	1.000	1.000

Table 4 also addresses the open question as to whether Raab's method is asymptotically more efficient than the Rodbard and Frazier method for normally distributed data. The answer is a general yes, thus agreeing with the Monte Carlo evidence available when the variance is a power of the mean. The results of this section suggest that, in the case of assay data containing pairs for which only Y_{ij} is observed, an estimator for θ combining estimation based on residuals for the observations for which x_i is known and on standard deviations otherwise in an appropriately weighted fashion would offer some improvement over the methods currently employed (see Davidian et al. 1987).

6. DISCUSSION

In Section 4 we constructed a general theory of regression-type estimation for θ in the heteroscedastic model (1.1). This theory includes as special cases common methods described in Section 3 and allows for the regression to be based on absolute residuals from the current regression fit as well as sample standard deviations in the event of replication at each design point. Under various restrictions such as symmetry or small σ, when the variance function g does not depend on β, we showed in Sections 4 and 5 that we can draw general conclusions about this class of estimators as well as make comparisons among the various methods.

When employing methods based on residuals, one should weight the residuals appropriately and iterate the process. There can be large relative differences among the methods in terms of efficiency. Under symmetry of the errors, squared residuals are preferable for approximately normally distributed data, but this preference is tenuous, since these can be highly nonrobust under only slight departures from normality; methods based on logarithms or the absolute residuals themselves exhibit relatively more robust behavior. For the small amount of replication found in practice, using sample standard deviations rather than residuals can entail a loss in efficiency if estimation is based on the squares of these quantities or the quantities themselves. For the logarithm method based on residuals, trimming the smallest few absolute residuals is essential, since for normal data using sample standard deviations is almost always more efficient than using residuals, even for a small number of replicates. Popular methods in applications such as radioimmunoassay based on sample means and sample standard deviations can be less efficient than methods based on weighted squared residuals. In some instances, the precision with which we can estimate θ depends on the relative range of values of the mean responses, not their actual values, so immediate implications for design are suggested.

Efficient variance function estimation in heteroscedastic regression analysis is an important problem in its own right. There are important differences in estimators for variance when it is modeled parametrically.

APPENDIX A: PROOFS OF MAJOR RESULTS

We now present sketches of the proofs of the theorems of Section 4. Our exposition is brief and nonrigorous, as our goal is to provide general insights. In what follows, we assume that

$$N^{1/2}\begin{bmatrix}\hat{\eta} - \eta \\ \hat{\theta} - \theta\end{bmatrix} = O_p(1); \tag{A.1}$$

under sufficient regularity conditions it is possible to prove (A.1). Such a proof would be long, detailed, and essentially noninformative; see Carroll and Ruppert (1982a) for a proof of $N^{1/2}$ consistency in a special case.

Sketch of the Proof of Theorem 4.1. From (4.1), a Taylor series, the fact that $E[T\{d_i(\beta)\}] = M_i$ and laws of large numbers, we have

$$0 = N^{-1/2}\sum_{i=1}^{N}(H_i/V_i)[T\{d_i(\hat{\beta}_*)\} - M_i(\hat{\eta}, \hat{\theta}, \hat{\beta}_*)] + o_p(1). \tag{A.2}$$

By the arguments of Ruppert and Carroll (1980) or Carroll and Ruppert (1982a),

$$\begin{aligned} N^{-1/2}\sum_{i=1}^{N}(H_i/V_i)[T\{d_i(\hat{\beta}_*)\} - T\{d_i(\beta)\}] \\ = N^{-1/2}\sum_{i=1}^{N}(H_i/V_i)\dot{T}\{d_i(\beta)\}\{d_i(\hat{\beta}_*) - d_i(\beta)\} + o_p(1) \\ = B_{3,N}N^{1/2}(\hat{\beta}_* - \beta) + o_p(1). \end{aligned} \tag{A.3}$$

Applying this result to (A.2) along with a Taylor series in M_i gives

$$\begin{aligned} 0 = N^{-1/2}\sum_{i=1}^{N}C_i + (B_{2,N} + B_{3,N})N^{1/2}(\hat{\beta}_* - \beta) \\ - B_{1,N}N^{1/2}\begin{bmatrix}\hat{\eta} - \eta \\ \hat{\theta} - \theta\end{bmatrix} + o_p(1), \end{aligned}$$

which is (4.2).

Theorem 4.2 follows by a similar argument; in this case the representation (A.3) is unnecessary.

Sketch of the Proof of Theorem 4.3. We consider Theorem 4.2; the proof for Theorem 4.1 is similar. Recall here that (3.4) holds. In the following, all derivatives are with respect to the mean μ_i and the definitions of C_i and M_i are as in (4.4).

Assumption (iv) implies that

$$N^{1/2}\max_{1\le i\le N}|\overline{Y}_i - \mu_i| \xrightarrow{p} 0,$$

so a Taylor series in η, θ, and $\overline{Y}_i$ gives

$$\begin{aligned} B_{1,N}N^{1/2}\begin{bmatrix}\hat{\eta} - \eta \\ \hat{\theta} - \theta\end{bmatrix} \\ = N^{-1/2}\sum_{i=1}^{N}C_i - N^{-1/2}\sum_{i=1}^{N}(\dot{M}_iH_i/V_i)(\overline{Y}_i - \mu_i) \\ + N^{-1/2}\sum_{i=1}^{N}\{(\dot{H}_i/V_i) - (\dot{V}_i/V_i)\}(\overline{Y}_i - \mu_i) + o_p(1). \end{aligned} \tag{A.4}$$

Since $\overline{Y}_i - \mu_i = \sigma g(\mu_i, z_i, \theta)\bar{\varepsilon}_i \approx \lambda N^{-1/2}g(\mu_i, z_i, \theta)\bar{\varepsilon}_i$, where $\bar{\varepsilon}_i$ is the mean of the errors at x_i, we can write the last two terms on the right side of (A.4) as

$$\lambda N^{-1}\sum_{i=1}^{N}\bar{\varepsilon}_i(q_{i,1} + q_{i,2}C_i) \tag{A.5}$$

for constants $\{q_{i,j}\}$. Since $\bar{\varepsilon}_i$ has mean 0, (A.5) converges in probability to 0 if $E(\bar{\varepsilon}_iC_i) = 0$, which holds under the assumption of symmetry. Thus (A.5) converges to 0, which from (A.4) completes the proof. Note that if we drop the assumption of symmetry, from (A.5) the asymptotic normal distribution of $N^{1/2}(\hat{\theta} - \theta)$ will have mean

$$p\text{-}\lim_{N\to\infty}\{\lambda B_{1,N}^{-1}N^{-1}\sum_{i=1}^{N}(\bar{\varepsilon}_iC_iq_{i,2})\}.$$

APPENDIX B: CHARACTERIZATION OF RESTRICTED MAXIMUM LIKELIHOOD

Let $\hat{\beta}_*$ be a generalized least squares estimator for β. Assume first that g does not depend on β. Let the prior distribution for the parameters $\pi(\beta, \theta, \sigma)$ be proportional to σ^{-1}. The marginal posterior for θ is hard to compute in closed form for nonlinear regression. Following Box and Hill (1974) and Beal and Sheiner (1987), we have the linear approximation

$$f(x_i, \beta) \approx f(x_i, \hat{\beta}_*) + f_\beta(x_i, \hat{\beta}_*)'(\beta - \hat{\beta}_*).$$

Replacing $f(x_i, \beta)$ by its linear expansion, the marginal posterior for θ is proportional to

$$p(\theta) = \frac{\left\{\prod_{i=1}^{N}g_i^2(\theta)\right\}^{-1/2}}{\hat{\sigma}_G^{(N-p)}(\theta)\{\text{Det } S_G(\theta)\}^{1/2}}, \tag{B.1}$$

where

$$\hat{\sigma}_G^2(\theta) = (N - p)^{-1}\sum_{i=1}^{N}r_i^2/g_i^2(\theta),$$

$$S_G(\theta) = N^{-1}\sum_{i=1}^{N}f_\beta(x_i, \hat{\beta}_*)f_\beta(x_i, \hat{\beta}_*)'/g_i^2(\theta),$$

and where Det A = determinant of A. If the variances depend on β, we extend the Bayesian arguments by replacing $g_i(\theta)$ by $g(z_i, \hat{\beta}_*, \theta)$.

Let H be the hat matrix H evaluated at $\hat{\beta}_*$ and let $h_{ii} = h_{ii}(\hat{\beta}_*, \theta)$. From (3.1), pseudolikelihood solves in (θ, σ)

$$\begin{aligned} \sum_{i=1}^{N}[r_i^2/\{\sigma^2g^2(z_i, \hat{\beta}_*, \theta)\}]\begin{bmatrix}1 \\ v_\theta(i, \hat{\beta}_*, \theta)\end{bmatrix} \\ = \sum_{i=1}^{N}\begin{bmatrix}1 \\ v_\theta(i, \hat{\beta}_*, \theta)\end{bmatrix}. \end{aligned} \tag{B.2}$$

Since H is idempotent, the left side of (B.2) has approximate expectation

$$\sum_{i=1}^{N}\begin{bmatrix}1 - p/N \\ v_\theta(i, \hat{\beta}_*, \theta)(1 - h_{ii})\end{bmatrix}. \tag{B.3}$$

To modify pseudolikelihood to account for loss of degrees of freedom, equate the left side of (B.2) to (B.3). From matrix computations as in Nel (1980), this can be shown to be equivalent to restricted maximum likelihood.

[*Received July 1986. Revised April 1987.*]

REFERENCES

Amemiya, T. (1977), "A Note on a Heteroscedastic Model," *Journal of Econometrics*, 6, 365–370. [See also *Corrigenda*, Vol. 8, p. 265.]

Beal, S. L., and Sheiner, L. B. (1987), "Heteroscedastic Nonlinear Regression With Pharmokinetic Type Data," preprint.

http://www.springer.com/978-3-319-05800-9

The Work of Raymond J. Carroll
The Impact and Influence of a Statistician
Davidian, M.; Lin, X.; Morris, J.S.; Stefanski, L.A... (Eds.)
2014, XX, 579 p. 16 illus., 14 illus. in color., Hardcover
ISBN: 978-3-319-05800-9