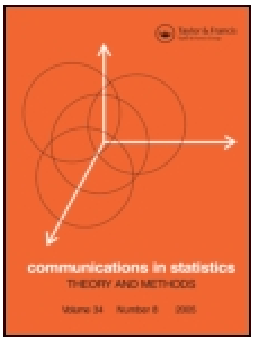

Taylor & Francis
Taylor & Francis Group

Communications in Statistics - Theory and Methods

ISSN: 0361-0926 (Print) 1532-415X (Online) Journal homepage: http://www.tandfonline.com/loi/lsta20

Minimum anderson-darling estimation

Dennis D Boos

To cite this article: Dennis D Boos (1982) Minimum anderson-darling estimation, Communications in Statistics - Theory and Methods, 11:24, 2747-2774, DOI: 10.1080/03610928208828420

To link to this article: http://dx.doi.org/10.1080/03610928208828420

COMMUN. STATIST.-THEOR. METH., 11(24), 2747-2774 (1982)

MINIMUM ANDERSON-DARLING ESTIMATION

Dennis D. Boos

Department of Statistics
North Carolina State University
Raleigh, North Carolina 27650 USA

Key Words and Phrases: Anderson-Darling distance; influence curves; asymptotic relative efficiency; goodness of fit.

ABSTRACT

The Anderson-Darling distance $d(F_n,F_\theta)$ between the empirical distribution function F_n and the hypothesized model F_θ is minimized to produce estimators with good robustness and efficiency properties. In addition, the residual distance $d_{min} = \min_\theta d(F_n,F_\theta)$ is available to assess model validity. The approximate distribution-free character of nd_{min} is studied using Monte Carlo methods.

1. INTRODUCTION

Recent papers by Parr and Schucany (1980), Parr and DeWet (1981), Millar (1981), and Boos (1981) have shown that the minimization of a weighted Cramér-von Mises distance between the hypothesized model distribution function F_θ and the empirical distribution function F_n results in estimators $\hat{\theta}$ which are consistent, asymptotically normal, and robust and/or efficient if the weight function is chosen appropriately. Other relevant literature is listed in Parr (1981). In this paper I focus on the Anderson-Darling distance

$$d_{F_n}(\theta) = \int_{-\infty}^{\infty} [F_n(x) - F_\theta(x)]^2 [F_\theta(x)(1-F_\theta(x))]^{-1} f_\theta(x)dx \qquad (1.1)$$

because the weight function $w_\theta = [F_\theta(1-F_\theta)]^{-1} f_\theta$ allows a nice balance between robustness and efficiency in a variety of models, and $nd_{F_n}(\hat{\theta})$ has a manageable null distribution.

The basic approach is to minimize (1.1) using the well-known computing formula

$$d_{F_n}(\theta) = -n^{-2} \sum_{i=1}^{n} (2i-1)[\ell n\, F_\theta(X_{(i)}) + \ell n(1-F_\theta(X_{(n+1-i)}))] - 1 , \qquad (1.2)$$

where $X_{(1)} \leq \cdots \leq X_{(n)}$ is the ordered sample. The values $\hat{\theta}$ which minimize $d_{F_n}(\theta)$ are the parameter estimates and $nd_{min} = nd_{F_n}(\hat{\theta})$ is a suitable statistic for testing model validity. Section 3 explores the distribution of this composite goodness-of-fit statistic including Monte Carlo results for a number of models. These latter results tend to confirm a conjecture of Boos (1981) that the null distribution of nd_{min} is *approximately* distribution-free. Thus, only one set of tabled critical values is required for fitting different models F_θ. Also, selection of the best model from a set of competing models is accomplished by choosing the one with the smallest d_{min}. Since smooth monotone transformations of the data such as $Y_i = \ell n\, X_i$ with the corresponding model change $F_\theta(x) \rightarrow F_\theta(\exp(x))$ leave the results unchanged, examples focus on location-scale families and several three-parameter generalizations.

Section 2 describes the asymptotic properties of $\hat{\theta}$ with supporting lemmas contained in the Appendix. For a wide range of location-scale families, the estimators have bounded influence curves and high asymptotic efficiencies. However, in the Box-Cox three-parameter normal model, the asymptotic efficiency for the skewness parameter λ is only .667. Results are also obtained for the Burr II and generalized gamma distributions. The paper concludes with a computational example and a summary.

2. MINIMUM ANDERSON-DARLING ESTIMATORS

Consider a random sample $X_1,\ldots,X_n$ from a parametric family having distribution function $F_{\theta_0}(x)$, $\theta_0 \in \Theta$, and let $F_n(x) = n^{-1}\,\Sigma I(X_i \le x)$ be the usual empirical distribution function. Minimum Anderson-Darling estimators (MADE's) $\hat\theta$ are found by minimizing the distance (1.1) between F_n and F_θ. If Θ is a compact set and F_θ is continuous in θ, then existence of at least one minimizing value $\hat\theta$ is guaranteed since $d_{F_n}(\theta)$ is continuous in θ. There may be more than one such $\hat\theta$, but Lemma 1 of the Appendix gives conditions on F_θ so that all $\hat\theta \xrightarrow{wp1} \theta_0$ as $n \to \infty$. If F_θ has partial derivatives with respect to θ_i, $\theta = (\theta_1,\ldots,\theta_k)^T$, and $\hat\theta$ lies in the interior of Θ, then $\hat\theta$ is also a solution of the set of equations

$$
\begin{aligned}
\tfrac{1}{2} d_{F_n}{}^{i}(\theta) = &\int_{-\infty}^{\infty} [F_n(x) - F_\theta(x)][-F_\theta{}^{i}(x)]\, w_\theta(x)dx \\
&+ \tfrac{1}{2}\int_{-\infty}^{\infty} [F_n(x) - F_\theta(x)]^2\, w_\theta{}^{i}(x)\, dx = 0,\ i = 1,\ldots,k,
\end{aligned}
\tag{2.1}
$$

where $w_\theta = [F_\theta(1-F_\theta)]^{-1} f_\theta$ and $q_\theta{}^{i}(x)$ is defined to be $\partial/\partial\theta_i\, q_\theta(x)$ for any function $q_\theta(x)$. Let $q_\theta{}^{ij}(x)$ be the analogously defined second derivatives. Lemma 2 of the Appendix gives conditions for the usual Taylor expansion of (2.1) to hold so that

$$
\hat\theta - \theta_0 = n^{-1} \sum_{i=1}^{n} IC_{\theta_0}(X_i) + o_p(n^{-\frac{1}{2}}), \tag{2.2}
$$

where

$$IC_{\theta_0}(y) = -\Delta^{-1} \int_{-\infty}^{\infty} [I(y \le x) - F_{\theta_0}(x)][-F_{\theta_0}^{\ 1}(x), \ldots, -F_{\theta_0}^{\ k}(x)]^T w_{\theta_0}(x)dx \tag{2.3}$$

is the influence curve of $\hat{\theta}$ and Δ is the matrix of probability limits of second derivatives, i.e., $\frac{1}{2} d_{F_n}^{\ ij}(\hat{\theta}) \xrightarrow{p} \Delta_{ij}$. Thus, if C is the covariance matrix of $\Delta \cdot IC_{\theta_0}(X_1)$, then the asymptotic covariance matrix of $n^{\frac{1}{2}}(\hat{\theta}-\theta_0)$ is $\Delta^{-1}C\Delta^{-1}$.

Whenever $\int_{-\infty}^{\infty} [-F_{\theta_0}^{\ i}(t)]w_{\theta_0}(t)dt$ exists, define

$$Q_i(x) = \int_{-\infty}^{x} [-F_{\theta_0}^{\ i}(t)]w_{\theta_0}(t)dt .$$

Then integrating (2.3) by parts we see that

$$IC_{\theta_0}(y) = \Delta^{-1}\left[Q_1(y) - \int_{-\infty}^{\infty} Q_1(x)dF_{\theta_0}(x), \ldots, Q_k(y) - \int_{-\infty}^{\infty} Q_k(x)dF_{\theta_0}(x)\right]^T$$

is bounded. Note that these results hold when the data is from the true model. When the model is incorrectly specified, the second term of (2.1) has a contribution to (2.2) which may cause the influence curve to be unbounded. For example, suppose that we assume a normal location model $F_\theta(x) = \Phi(x-\theta)$, but the data is truly Cauchy centered at 0 , $F(x) = \frac{1}{2} + \pi^{-1} \tan^{-1}x$. Then, by expanding (2.1) or using (3.7) of Boos (1981), we find that the influence curve of $\hat{\theta}$ at F is proportional to

$$\int [I(y \le x) - F(x)][-\phi(x)w(x) + [F(x) - \Phi(x)]w'(x)]dx , \tag{2.4}$$

where $w(x) = [\Phi(x)(1-\Phi(x))]^{-1}\phi(x)$. The $-\phi(x)w(x)$ term in (2.4) is the usual "correct model" part of the influence curve arising from the first term of (2.1). The $[F(x)-\Phi(x)]w'(x)$ term in (2.4) comes from the second term of (2.1) and would be zero if F were really normal. This latter term can be rewritten using integration by parts to get

$$\begin{aligned} &-\int w(x)[F(x) - \Phi(x)]d[I(y\le x) - F(x)] \\ &-\int w(x)[I(y\le x)- F(x)]d[F(x) - \phi(x)] . \end{aligned} \tag{2.5}$$

Now, since $w(x) \sim |x|$ for $x \to \pm\infty$, the first integral of (2.5) is bounded in y. However, as $y \to -\infty$, the second integral blows up since $\int w(x)dF(x) = \infty$. In such misspecified cases, I expect the goodness-of-fit statistic nd_{min} to be sensitive to outliers and flag their presence.

A. Location-Scale Families

Let $F_\theta(x) = F((x-\mu)/\sigma)$ with $\theta = (\mu,\sigma)$ and let F have density f. Then at $\theta = (\mu_0,\sigma_0)$

$$\Delta = \frac{1}{\sigma_0^2}\begin{bmatrix} \int f^2(x)w(x)dx & \int f^2(x)xw(x)dx \\ \int f^2(x)xw(x)dx & \int f^2(x)x^2w(x)dx \end{bmatrix} \tag{2.6}$$

and

$$C = \frac{1}{\sigma_0^2}\begin{bmatrix} \iint r(x,y)dxdy & \iint r(x,y)xdxdy \\ \iint r(x,y)ydxdy & \iint r(x,y)xydxdy \end{bmatrix},$$

where $w(x) = [F(x)(1-F(x))]^{-1}f(x)$ and $r(x,y) = [\min(F(x), F(y)) - F(x)F(y)]f(x)w(x)f(y)w(y)$.

1) Symmetric densities, $f(x) = f(-x)$. Here $\Delta_{12} = \Delta_{21} = C_{12} = C_{21} = 0$ and

$$\Delta^{-1}C\Delta^{-1} = \begin{pmatrix} C_{11}/\Delta_{11}^{2} & 0 \\ 0 & C_{22}/\Delta_{22}^{2} \end{pmatrix} .$$

The following parent distributions satisfy the conditions of Lemmas 1 and 2 of the Appendix: t distribution with ν degrees of freedom; logistic, $f(x) = \exp(-x)(1+\exp(-x))^{-2}$; hyperbolic secant, $f(x) = \pi^{-1}\text{sech}(x)$; and f = standard normal. The double exponential $f(x) = \frac{1}{2}\exp(-|x|)$ is not smooth enough to satisfy the conditions of Lemma 2. However, an alternative argument presented in Boos (1981) justifies (2.2). For each of these families, the influence curves for location and scale are bounded with

$$Q_1(y) = \frac{1}{\sigma_0}\int_{-\infty}^{(y-\mu_0)/\sigma_0} f(x)w(x)dx \qquad Q_2(y) = \frac{1}{\sigma_0}\int_{-\infty}^{(y-\mu_0)/\sigma_0} f(x)xw(x)dx ,$$

and

$$IC_{\mu_0}(y) = \Delta_{11}^{-1}\left(Q_1(y) - \int_{-\infty}^{\infty} Q_1(y)dF\left(\frac{y-\mu_0}{\sigma_0}\right)\right)$$

$$IC_{\sigma_0}(y) = \Delta_{22}^{-1}\left(Q_2(y) - \int_{-\infty}^{\infty} Q_2(y)dF\left(\frac{y-\mu_0}{\sigma_0}\right)\right) .$$

Figure 1 displays IC_μ and IC_σ for the standard normal.

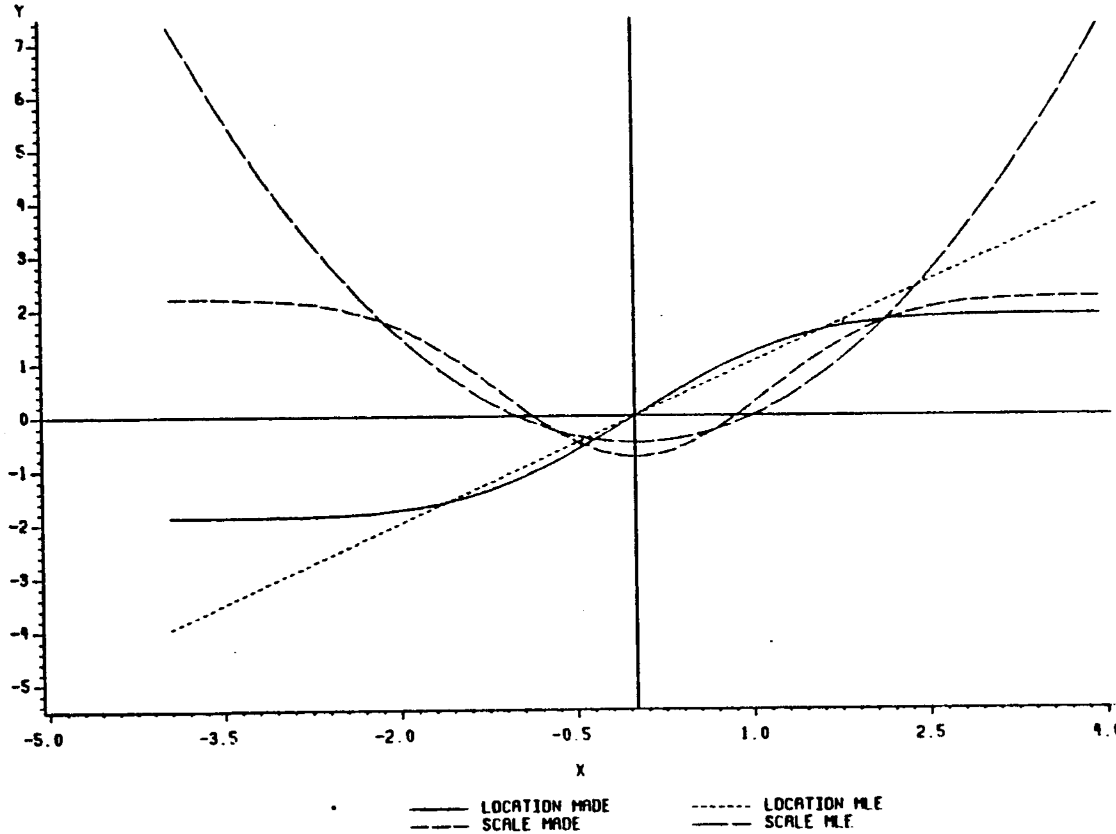

Fig. 1. Influence Curves for Normal Location and Scale.

Table I lists the asymptotic variances $E_{\theta_0}[C_{\theta_0}(X_1)]^2$ and asymptotic relative efficiencies (ARE's) with respect to the maximum likelihood estimates (MLE's). Numerical methods were employed for most of the calculations. The ARE's are quite high over a wide range of densities, and the three situations with ARE's less than .9 can be explained in terms of the influence curves. The influence curve of the Cauchy location MLE redescends to zero as $x \to \pm\infty$ so that the bounded (but nondecreasing) influence curve of the MADE is still "too large." The influence curve of the double exponential location MLE is proportional to sign(x) and the smooth MADE influence curve apparently does not mimic the jump at zero closely enough. The influence curve of the normal scale MLE is proportional to x^2 and the bounded influence curve of the MADE estimator is apparently not "large enough" in this case.

2) Extreme value distribution, $F(x) = \exp(-\exp(-x))$. F is skewed with shape parameters $\sqrt{\beta_1} = 1.14$ and $\beta_2 = 5.4$. Fitting a Weibull distribution $P(Y \le y) = 1 - \exp(-(y/d)^c)$ is the same as

TABLE I

MADE Asymptotic Variances and Relative Efficiencies for Symmetric Location-Scale Families.

	β_2*	Location		Scale	
		Asymptotic Variance	ARE	Asymptotic Variance	ARE
Normal	3	1.035	.966	.589	.849
Logistic	4.2	3.000	1.000	.758	.923
Sech	5.0	2.020	.990	.871	.931
DEXP	6	1.262	.792	1.096	.912
t_5	9	1.341	.994	.831	.963
t_3	∞	1.557	.964	1.009	.991
t_1 (Cauchy)	-	2.736	.731	2.077	.963

*$\beta_2 = E(X-\mu)^4/(E(X-\mu)^2)^2$.

fitting $X = -\ell n\ Y$ to an extreme value distribution $F((x-\mu)/\sigma)$ where $\mu = -\ell n\ d$ and $\sigma = c^{-1}$. The information matrix and inverse for $(\mu,\sigma) = (0,1)$ are

$$I = \begin{bmatrix} 1.000 & -.423 \\ -.423 & 1.824 \end{bmatrix} \qquad I^{-1} = \begin{bmatrix} 1.109 & .257 \\ .257 & .608 \end{bmatrix}.$$

For the MADE's

$$\Delta = \begin{bmatrix} .404 & .023 \\ .023 & .252 \end{bmatrix} \quad C = \begin{bmatrix} .192 & .039 \\ .039 & .049 \end{bmatrix} \quad \Delta^{-1}C\Delta^{-1} = \begin{bmatrix} 1.145 & .231 \\ .231 & .717 \end{bmatrix}.$$

Thus, when location and scale are both estimated, the ARE's are .968 and .848, respectively. Figure 2 displays the influence curves for both MADE's and MLE's. The MLE influence curves really explode for $x < 0$. The example of Section 4 shows the practical consequences of letting one observation in a sample get very large in absolute value.

The Appendix gives some of the details for applying Lemmas 1 and 2 to the extreme value distribution. One may note that these lemmas will not apply to nonregular location-scale families such as F = uniform or F = exponential.

B. Three-Parameter Models

1) Let $F_\theta(x) = P(-\ell n\ Y \le x)$ where Y has a generalized gamma distribution (see Stacy (1962)). Then $F_\theta(x) = F_\lambda((x-\mu)/\sigma)$ where $F_\lambda(x) = 1 - (\Gamma(\lambda))^{-1} \int_0^{\exp(-x)} t^{\lambda-1}e^{-t}dt$. At $(\mu,\sigma,\lambda) = (0,1,1)$, $F_\theta(x) = \exp(-\exp(-x))$ and the information matrix is the same as in the previous example augmented by $I_{13} = I_{31} = -1$, $I_{23} = I_{32} = -.577$, and $I_{33} = 1.645$. Likewise, for the MADE of (μ,σ,λ), Δ and C are augmented by $\Delta_{13} = \Delta_{31} = -.508$, $\Delta_{23} = \Delta_{32} = -.187$, $\Delta_{33} = .740$, $C_{13} = C_{31} = -.258$, $C_{23} = C_{32} = -.078$, and $C_{33} = .363$. Then

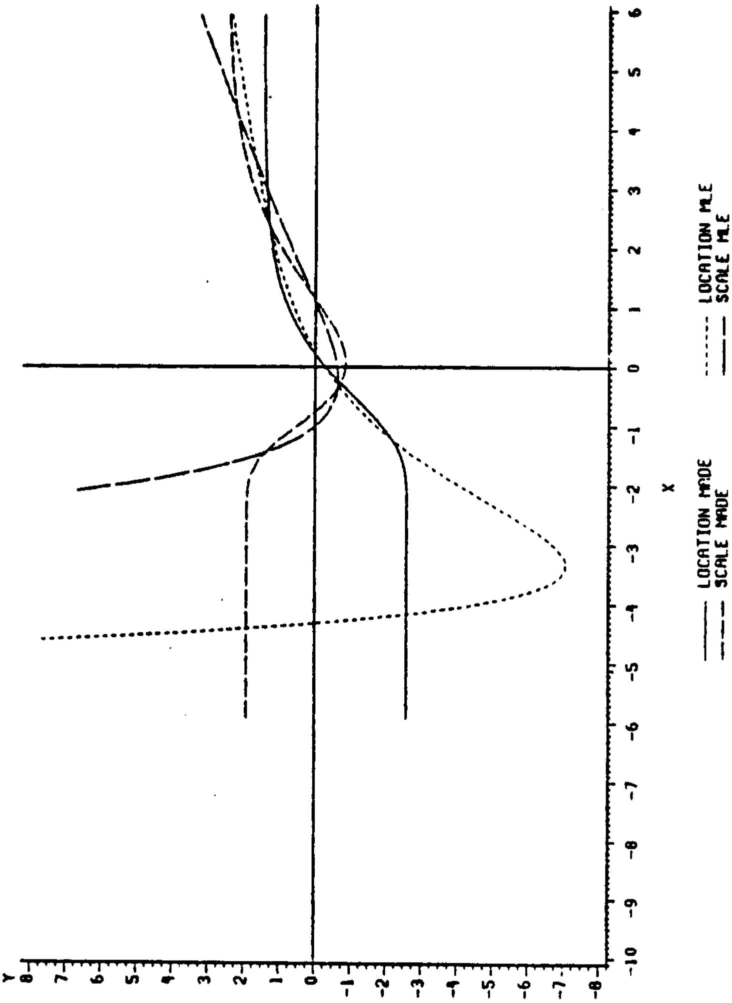

Fig. 2. Influence Curves for Extreme Value Location and Scale.

$$I^{-1} = \begin{pmatrix} 43.8 & 20.9 & 34.0 \\ 20.9 & 10.6 & 16.4 \\ 34.0 & 16.4 & 27.0 \end{pmatrix} \qquad \Delta^{-1}C\Delta^{-1} = \begin{pmatrix} 58.3 & 28.5 & 45.8 \\ 28.5 & 14.7 & 22.6 \\ 45.8 & 22.6 & 36.6 \end{pmatrix}$$

Thus, the ARE's for estimating (μ,σ,λ) at $(0,1,1)$ are .751, .723, and .738, respectively.

2) The Burr II distribution is $F_\lambda(y) = [1 + \exp(-y)]^{-\lambda}$. With scale and location parameters the model is $F_\theta(x) = F_\lambda((x-\mu)/\sigma)$, $\theta = (\mu,\sigma,\lambda)$. At $\lambda = 1$, F_λ is just the logistic distribution function $[1+\exp(-x)]^{-1}$ and for $(\mu,\sigma,\lambda) = (0,1,1)$ we have

$$I = \begin{pmatrix} .333 & 0 & .5 \\ 0 & 1.430 & -.5 \\ .5 & -.5 & 1.0 \end{pmatrix} \qquad I^{-1} = \begin{pmatrix} 32.9 & -7.0 & -20.0 \\ -7.0 & 2.3 & 4.7 \\ -20.0 & 4.7 & 13.3 \end{pmatrix}$$

$$\Delta = \begin{pmatrix} .167 & 0 & .250 \\ 0 & .215 & -.072 \\ .250 & -.072 & .404 \end{pmatrix} \qquad C = \begin{pmatrix} .083 & 0 & .125 \\ 0 & .035 & -.012 \\ .125 & -.012 & .192 \end{pmatrix}$$

$$\Delta^{-1}C\Delta^{-1} = \begin{pmatrix} 39.7 & -8.4 & -24.5 \\ -8.4 & 2.7 & 5.6 \\ -24.5 & 5.6 & 16.3 \end{pmatrix} .$$

The ARE's of the MADE's are then .829, .869, and .815, respectively.

3) Box-Cox model. Box and Cox (1964) suggested the transformation

$$\begin{aligned} X^{(\lambda)} &= \frac{Y^\lambda - 1}{\lambda} \qquad \lambda \neq 0 \\ &= \log Y \qquad \lambda = 0 \end{aligned}$$

to achieve more "normal" and homogeneous errors in linear models. They proposed that λ be estimated from the data using a normal likelihood. Other approaches are found in Hinkley (1975), Carroll (1980), and Bickel and Doksum (1981).

Here we assume that the $Y^{(\lambda)}$ are IID normally distributed random variables with mean μ and variance σ^2. Since normality can only hold when $\lambda = 0$, we restrict ourselves to $Y \sim$ lognormal and fit Y to the model $F_\theta(x) = \Phi((x^{(\lambda)}-\mu)/\sigma)$, $x > 0$.

The inverse of the information matrix of (μ,σ,λ) at $\lambda = 0$ is given by

$$I^{-1} = \begin{pmatrix} \frac{7}{6}\sigma^2 + \frac{1}{3}\mu^2 + \frac{1}{6}\frac{\mu^4}{\sigma^2} & \frac{\mu}{3\sigma}(\sigma^2+\mu^2) & \frac{1}{3}\left(1 + \frac{\mu^2}{\sigma^2}\right) \\ \frac{\mu}{3\sigma}(\sigma^2 + \mu^2) & \frac{\sigma^2}{2} + \frac{2}{3}\mu^2 & \frac{2}{3}\frac{\mu}{\sigma} \\ \frac{1}{3}\left(1 + \frac{\mu^2}{\sigma^2}\right) & \frac{2}{3}\frac{\mu}{\sigma} & \frac{2}{3\sigma^2} \end{pmatrix} .$$

I^{-1} differs from that found in Hinkley (1975) because he considers $\tau = \sigma^2$ as the scale parameter. For the MADE's

$$\Delta^{-1}C\Delta^{-1} = \begin{pmatrix} 1.213\sigma^2+.457\mu^2 + .250\frac{\mu^4}{\sigma^2} & \frac{\mu}{\sigma}(.457\sigma^2+.499\mu^2) & .457+.499\frac{\mu^2}{\sigma^2} \\ \frac{\mu}{\sigma}(.457\sigma^2+.499\mu^2) & .589\sigma^2+.999\mu^2 & .999\frac{\mu}{\sigma} \\ .457+.499\frac{\mu^2}{\sigma^2} & .999\frac{\mu}{\sigma} & \frac{.999}{\sigma^2} \end{pmatrix} .$$

When $\mu = 0$, the ARE's for location and scale are the same as in the $\lambda = 0$ known case, .966 and .849, respectively. For arbitrary (μ,σ), .667 is the ARE for estimating λ and also the lower bound of the ARE's for estimating μ and σ. These low efficiencies can be partly explained by the fact that the MLE influence curves are cubic polynomials in log y whereas the MADE's have bounded influence curves. This result is not just an artifact of the normal distribution. If $Y^{(\lambda)}$ is double exponential and we fit $Y^{(\lambda)}$ to a double exponential distribution, then at $\lambda = 0$ the MLE's have influence curves which are quadratic in log y.

3. GOODNESS OF FIT

Consider the composite goodness-of-fit hypothesis H_0: distribution function of the data $= F_\theta$, θ unknown but F_θ a member of some specified parametric family. The minimized distance $d_{min} = d_{F_n}(\hat{\theta})$ is a natural statistic for testing H_0 although its null percentiles are much smaller than those of $d_{F_n}(\theta)$ when θ is specified. A similar result holds if $\hat{\theta}$ is replaced by other estimators and Stephens (1974, 1976, 1977, 1979) has published tables of the null distribution of $nd_{F_n}(\hat{\theta}_{MLE})$. Unfortunately, each parametric family requires a different table. In contrast, Boos (1981, Sec. 4) considered location-scale models and conjectured that the null limiting distribution of nd_{min} could be reasonably approximated by the distribution of $A_2^2 = \sum_{i=3}^{\infty} Z_i^2/i(i+1)$ for a range of symmetric distributions, where the Z_i are IID standard normal random variables. The Monte Carlo results of this section support that conjecture and suggest that the approximation is also valid for more than just symmetric location-scale models. The general use of $A_k^2 = \sum_{i=k+1}^{\infty} Z_i^2/i(i+1)$ for the case of k estimated parameters can be motivated by analogy with the chi-squared goodness-of-fit statistic, where typically the degrees of freedom are reduced by the number of estimated parameters. Here, the limiting null distribution of $nd_{F_n}(\theta)$ when no parameters are estimated is $A_0^2 = \sum_{i=1}^{\infty} Z_i^2/i(i+1)$ and estimation of parameters results in the approximate loss of degrees of freedom corresponding to $Z_1^2/2$, $Z_2^2/6$, etc.

More formally, Pollard (1980) gives general conditions for

$$\min_{\theta} \quad n\int_{-\infty}^{\infty} [F_n(y)-F_\theta(y)]^2 w_\theta(y)dy \xrightarrow{d} \min_{x=(x_1,\ldots,x_k)} \int_0^1 [U(t)-x\cdot h(t)]^2 v(t)dt \; , \tag{3.1}$$

where $U(t)$ is the Brownian Bridge on $C[0,1]$, $h(t) = (F_{\theta_0}{}^1(F_{\theta_0}{}^{-1}(t)),\ldots,F_{\theta_0}{}^k(F_{\theta_0}{}^{-1}(t)))^T$ and $v(t) = w_{\theta_0}(F_{\theta_0}{}^{-1}(t))/f_{\theta_0}(F_{\theta_0}{}^{-1}(t))$. Let $B(U) = \int U(t)h(t)v(t)dt$. Then the right-hand side of (3.1) is

$$\int U^2(t)v(t)dt + \min_x [x^T\Delta x - 2x^T B(U)]$$

$$= \int U^2(t)v(t)dt - B(U)^T\Delta^{-1}B(U) \; . \tag{3.2}$$

Note that $\Delta^{-1}B(U)$ is another representation of the limiting normal distribution of $n^{\frac{1}{2}}(\hat{\theta}-\theta_0)$. Boos (1981, Theorem 4.1) established (3.2) for location-scale families.

For the Anderson-Darling statistic $v(t) = [t(1-t)]^{-1}$ and following Anderson and Darling (1952) we have $\int U^2(t)v(t)dt = \sum_{i=1}^{\infty} Z_i{}^2/i(i+1)$, where the Z_i are IID standard normal random variables defined in terms of $U(t)$. When estimating location for the logistic distribution, $B(U)^T\Delta^{-1}B(U) = Z_1{}^2/2$ and thus $nd_{min} \xrightarrow{d} A_1{}^2 = \sum_2^{\infty} Z_i{}^2/i(i+1)$. For other parent distributions it is reasonable to expect that $B(U)^T\Delta^{-1}B(U) \doteq Z_1{}^2/2$ so that $A_1{}^2$ is an approximation to the limiting distribution of nd_{min} . The mean of $B(U)^T\Delta^{-1}B(U)$ is just C_{11}/Δ_{11} and thus (3.2) has mean $1 - C_{11}/\Delta_{11}$ whereas $A_1{}^2$ has mean $\frac{1}{2}$. Table II lists the ratio of means $2(1 - C_{11}/\Delta_{11})$ for the distributions of Table I and the extreme value distribution. When location and scale are both estimated, then I expect A_2^2 to be a reasonable approximation to the limiting distribution of nd_{min}. In the case of symmetric densities, $E(3.2) = 1 - C_{11}/\Delta_{11} - C_{22}/\Delta_{22}$, and for asymmetric densities, $E(3.2) = 1 - (C_{11}\Delta_{22} + C_{22}\Delta_{11} - 2C_{12}\Delta_{12})/(\Delta_{11}\Delta_{22} - \Delta_{12}^2)$. The ratio of these means to $E(A_2^2) = 1/3$ is found in column two of Table II. In general, $E(3.2) = 1 - Tr(\Delta^{-1}C)$, and the ratio of means for the three-parameter models discussed in Section 2 is found in column three of Table II. The results below suggest that dividing nd_{min} by the appropriate ratio of means helps the approximation to A_k^2.

Monte Carlo methods were employed to investigate the distribution of nd_{min} in the following situations: (a) μ unknown, normal and extreme value; (b) (μ,σ) unknown, normal, extreme

TABLE II

Ratios of Asymptotic Means, $E(3.2)/E(A_k^2)$

Parameters estimated	μ	(μ,σ)	(μ,σ,λ)
Normal	1.0054	1.0309	1.0776
Logistic	1.0000	1.0113	1.0426
Extreme Value	1.0509	1.0376	1.0705
Sech	1.0020	1.0116	
DEXP	1.0250	1.0480	
t_5	1.0006	1.0075	
t_3	1.0065	1.0116	
t_1 (Cauchy)	1.0800	1.1284	

Note: Entries are the ratio of $E(3.2)$ to $E(A_1^2) = 1/2$, $E(A_2^2) = 1/3$, and $E(A_3^2) = 1/4$, respectively. Column 3 corresponds to the Box-Cox model at $\lambda = 0$, the Burr II at $\lambda = 1$, and the generalized gamma at $\lambda = 1$.

value, double exponential (DEXP), t_5 and t_1; (c) (μ,σ,λ) unknown, Box-Cox model at $\lambda = 0$, Burr II at $\lambda = 1$. In cases (a) and (b) 1000 Monte Carlo replications were used and in case (c) 500 were used. The samples were generated by the SUPER DUPER random number generator of Marsaglia, Maclaren, and Bray (1964) and the t family generator of Kinderman and Monahan (1980). The inverse distribution function transformation was applied to uniform random variables to get logistic, DEXP, Cauchy, and Burr samples. The IMSL minimization routine ZXMIN was used to find d_{min} after starting with simple initial estimates, in most cases the median and interquartile range. For the Box-Cox model $\lambda = .1$ was used as a start and for the Burr II $\lambda = 1$ was used.

TABLE III

Empirical Levels of Goodness-of-Fit Tests Using the Upper Percentiles of A_1^2. Location Estimated.

	α	.25	.10	.05	.025	.01	D^2
			Normal				
n = 20	nd_{min}	.232	.103	.050	.025	.006	6.2
n = 20	$nd_{min}/1.0054$	.229	.101	.049	.025	.006	6.6
n = 50	nd_{min}	.262	.114	.062	.031	.012	3.2
n = 50	$nd_{min}/1.0054$	.262	.113	.058	.028	.011	2.1
			Extreme Value				
n = 20	nd_{min}	.280	.115	.058	.032	.010	7.0
n = 20	$nd_{min}/1.0509$	.247	.096	.050	.026	.008	1.4
n = 50	nd_{min}	.282	.123	.063	.031	.015	8.4
n = 50	$nd_{min}/1.0509$	.253	.104	.051	.027	.011	.4
Percentiles of A_1^2		.614	.857	1.046	1.240	1.505	

Note: Entries are the proportion of test statistics exceeding the appropriate percentile of A_1^2 in N = 1000 Monte Carlo replications.

TABLE IV

Empirical Levels of Goodness-of-Fit Tests Using the Percentiles of A_2^2. Location and Scale Estimated.

	α	.25	.10	.05	.025	.01	D^2
			Normal				
n = 20	nd_{min}	.269	.111	.068	.043	.017	14.9
n = 20	$nd_{min}/1.0309$	.249	.101	.062	.033	.014	5.8
n = 50	nd_{min}	.255	.094	.048	.027	.012	2.2
n = 50	$nd_{min}/1.0309$	.229	.087	.046	.025	.011	3.4
			Extreme Value				
n = 20	nd_{min}	.253	.116	.056	.027	.008	5.2
n = 20	$nd_{min}/1.0376$	.227	.098	.046	.020	.005	6.3
n = 50	nd_{min}	.291	.118	.061	.027	.010	10.3
n = 50	$nd_{min}/1.0376$	.259	.102	.050	.024	.008	1.0
			DEXP				
n = 20	nd_{min}	.267	.123	.064	.026	.010	9.1
n = 20	$nd_{min}/1.048$	.238	.099	.050	.021	.009	2.4
n = 50	nd_{min}	.314	.115	.063	.032	.019	31.4
n = 50	$nd_{min}/1.048$	.269	.104	.051	.026	.014	4.4
			t_5				
n = 20	nd_{min}	.221	.097	.040	.012	.005	13.7
n = 20	$nd_{min}/1.0075$	.217	.096	.038	.012	.005	15.1
n = 50	nd_{min}	.238	.089	.042	.019	.009	2.3
n = 50	$nd_{min}/1.0075$	.232	.083	.041	.018	.009	4.4
			Cauchy				
n = 20	nd_{min}	.300	.146	.076	.052	.032	61.5
n = 20	$nd_{min}/1.1284$	.227	.098	.054	.038	.021	20.0
n = 50	nd_{min}	.307	.166	.097	.061	.029	72.3
n = 50	$nd_{min}/1.1284$	.237	.116	.066	.038	.019	15.4
Percentiles of A_2^2		.403	.533	.631	.730	.863	

TABLE V

Empirical Levels of Goodness-of-Fit Tests Using the Upper Percentiles of A_3^2. Location, Scale, and Skewness Estimated.

	α	.25	.10	.05	.025	.01	D^2
			Burr II				
n = 50	nd_{min}	.258	.102	.058	.034	.016	2.6
n = 50	$nd_{min}/1.0426$	.226	.090	.050	.028	.012	2.5
			Box-Cox Model				
n = 50	nd_{min}	.330	.154	.074	.042	.018	22.4
n = 50	$nd_{min}/1.0776$	.268	.110	.448	.032	.008	6.4
Percentiles of A_3^2		.299	.382	.443	.504	.585	

Note: Entries are the proportion of test statistics exceeding the appropriate percentile of A_3^2 in N = 500 Monte Carlo replications.

In Tables III-V are listed the proportion of times nd_{min} and $nd_{min} \div$ (ratio of asymptotic means from Table II) exceeded the appropriate percentiles of A_k^2. The last column of each table is $D^2 = N(\hat{\alpha}-\alpha)^T\Sigma^{-1}(\hat{\alpha}-\alpha)$, where $\hat{\alpha}^T$ is the row of estimates, $\alpha^T = (.25, .1, .05, .025, .01)$, and $[\Sigma]_{ij} = (1-\alpha_i)\alpha_j$ $i \leq j$. This statistic may be compared to $\chi^2_{.05} = 11.1$, the 95th percentile of a chi-squared distribution with 5 degrees of freedom. D^2 indicates that the mean adjustment is important for the normal, extreme value, DEXP, Cauchy (t_1), and the Box-Cox model. The approximation appears especially good for $\alpha = .10$ and $\alpha = .05$. However, one would *not* need the mean adjustment for most informal testing situations.

Table VI compares the Monte Carlo estimates of the percentiles of nd_{min} with those obtained by Stephens (1974, 1977, 1979) for $nd_{F_n}(\hat{\theta}_{MLE})$. In the logistic location situation Stephens' results are listed since the MADE and the MLE are the same for this case.

TABLE VI

Comparison of Upper Percentiles of $nd_{F_n}(\hat{\theta}_{MLE})$ and nd_{min} at $n = 50$.

	Percentile	.75	.90	.95	.975	.99
	Location Estimated					
Normal	$nd_{min}/1.0054$	.63	.89	1.09	1.29	1.56
	$nd_{F_n}(\hat{\theta}_{MLE})$	-	.91	1.11	1.30	1.57
Logistic	$nd_{F_n}(\hat{\theta}_{MLE})$	.61	.85	1.04	1.24	1.50
Extreme Value	$nd_{min}/1.0509$	.62	.86	1.07	1.27	1.53
	$nd_{F_n}(\hat{\theta}_{MLE})$	.73	1.06	1.31	1.58	1.95
A_1^2		.614	.857	1.046	1.240	1.505
	Location and Scale Estimated					
Normal	$nd_{min}/1.0309$	.39	.51	.62	.71	.84
	$nd_{F_n}(\hat{\theta}_{MLE})$	-	.61	.74	.86	1.02
Logistic	$nd_{min}/1.0113$	.41	.53	.61	.70	.81
	$nd_{F_n}(\hat{\theta}_{MLE})$	.42	.56	.66	.77	.90
Extreme Value	$nd_{min}/1.0376$	.39	.53	.63	.73	.85
	$nd_{F_n}(\hat{\theta}_{MLE})$	.46	.62	.74	.85	1.01
A_2^2		.403	.533	.631	.730	.863

Note: The maximum likelihood results are taken from Stephens (1974, 1977, 1979).

The 95th, 97.5th and 99th percentiles for nd_{min} were estimated using the extreme value theory of Weissman (1978). For the first half of the table the estimated standard errors are .014, .024, .035, .046, and .064, respectively, and for the second half they are .007, .014, .016, .022, and .030, respectively. Differences between the percentiles of nd_{min} can be explained in large part by random sampling for $N = 1000$ Monte Carlo replications, whereas the Stephens' percentiles are based on $N = 5000$ and $N = 10000$ replications and tend to reflect the differences in distributions, It is interesting that the Stephens' percentiles for the normal and extreme value are very close for the two-parameter case but not for the one-parameter case.

The approach presented here assumes that one will estimate in a specific order: location, location and scale, or location, scale, and skewness. It is also assumed that when categorizing the type of parameter, all one-sided distributions such as the exponential have been transformed by a log transformation to have support on $(-\infty,\infty)$. The usual gamma has location and skewness parameters, but no scale parameter. It would be possible to use $Z_2^2/6 + \sum_4^\infty Z_j^2/j(j+1)$ as an approximating distribution in this case, but I have not tabled these percentiles. A fairly quick method of finding these percentiles is given in Solomon and Stephens (1977). The (untransformed) location-scale exponential model $F_\theta(x) = 1 - \exp((x-\mu)/\sigma)$, $x \geq \mu$, also does not fit into the above scheme. In this case, an expansion of $d_{F_n}'(\theta)$ similar to Lemma 2 of the Appendix suggests that $n^{\frac{1}{2}}\log n\,(\hat{\mu}-\mu)$ is asymptotically normal and that the limiting distribution of nd_{min} is A_1^2 rather than A_2^2. A brief Monte Carlo study supports this rate of convergence for $\hat{\mu}$ but indicates that the limiting distribution of nd_{min} is reached very slowly. I suggest the following alternative approach. Conditional on $X_{(1)} = x_{(1)}$, the remaining observations which exceed $x_{(1)}$ are IID with density $f(y) = \sigma^{-1}\exp\{-(y-x_{(1)})/\sigma\} \times I(y > x_{(1)})$ (see, e.g., Quesenberry, 1975). Thus, these remaining observations transformed to $-\ln(X_i - x_{(1)})$ can be fit to a location extreme value distribution.

The non-null asymptotic distribution of nd_{min} in location-scale families and some examples of asymptotic power are given in Boos (1981).

4. NUMERICAL EXAMPLE

Hahn and Shapiro (1967, p. 300-302) analyzed 20 failure times and showed that a one-parameter exponential did not fit the data but that a two-parameter lognormal was quite adequate. Table VII lists the original data in hours and also the transformed data. First, I fitted the transformed data to a location extreme value distribution. This corresponds to fitting a scale exponential to the original data. The value of $nd_{min}/1.0509 = 2.28$ is much larger than the 99th percentile of A_1^2 indicating agreement with Hahn and Shapiro. Next, I fitted a location-scale extreme value and location-scale normal to the transformed data. The values $nd_{min}/1.0376 = .18$ and $nd_{min}/1.0309 = .23$ are both less than the 75th percentile of A_2^2 and indicate a slight preference for the extreme value (Weibull) over the normal (lognormal).

TABLE VII

Time to Failure of Guidance Systems*

Hours	1	4	5	6	15	20	40	40	60	93
$-\ln$ (hours)	0.0	-1.39	-1.61	-1.79	-2.71	-3.00	-3.69	-3.69	-4.09	-4.53
Hours	95	106	125	151	200	268	459	827	840	1089
$-\ln$ (hours)	-4.55	-4.66	-4.82	-5.01	-5.30	-5.59	-6.13	-6.72	-6.73	-6.99

*From Hahn and Shipiro (1967), p. 300.

In order to see how the MADE's react to outliers I changed the 11th ordered observation 95 by several multiples of ten. The changes in parameter estimates and nd_{min} are recorded in Table VIII. In either model one can see that the MLE's react more strongly to changes than the MADE's.

5. SUMMARY

Minimization of the Anderson-Darling statistic produces estimators with good robustness and efficiency properties and the residual distance nd_{min} is available to judge model validity or select from competing models. A correction for the asymptotic mean

TABLE VIII

Parameter Estimates and Goodness-of-Fit Statistics for the Data of Table VII

Value of 11th Observation		Normal MADE			Normal MLE		Extreme Value MADE			Extreme Value MLE	
Hours	$-\ln$ (hours)	$\hat{\mu}$	$\hat{\sigma}$	$nd_{min}/1.0309$	$\bar{X}$	$\hat{\sigma}$	$\hat{\mu}$	$\hat{\sigma}$	$nd_{min}/1.0376$	$\hat{\mu}$	$\hat{\sigma}$
95000	-11.46	-4.40	2.34	.24	-4.50	2.47	-5.46	2.38	.76	-5.81	2.90
9500	- 9.16	-4.39	2.25	.14	-4.38	2.18	-5.33	2.03	.28	-5.47	2.14
95*	- 4.55	-4.22	2.00	.23	-4.15	1.89	-5.05	1.71	.18	-5.05	1.59
.095	2.35	-3.97	2.38	.30	-3.81	2.36	-4.92	1.96	.12	-4.89	1.83
.0095	4.66	-3.95	2.47	.45	-3.69	2.69	-4.92	1.99	.15	-4.85	1.91
.00095	6.96	-3.94	2.56	.63	-3.58	3.06	-4.91	2.01	.19	-4.82	1.99

*Original value of 11th observation.

improves the approximation of nd_{min} by $\sum_k^{\infty} Z_i^2/i(i+1)$, especially for α-levels near .10 to .05 .

BIBLIOGRAPHY

Anderson, T. W., and Darling, D. A. (1952), "Asymptotic Theory of Certain 'Goodness of Fit' Criteria Based on Stochastic Processes," *Annals of Mathematical Statistics*, 23, 193-212.

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

Boos, D. D. (1981), "Minimum Distance Estimators for Location and Goodness of Fit," *Journal of the American Statistical Association*, 76, 663-670.

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

Carroll, R. J. (1980), "A Robust Method for Testing Transformations to Achieve Approximate Normality," *Journal of the Royal Statistical Society*, Ser. B, 42, 71-78.

Hahn, G. J., and Shapiro, S. S. (1967), *Statistical Models in Engineering*, New York: John Wiley.

Hinkley, D. V. (1975), "On Power Transformations to Symmetry," *Biometrika*, 62, 101-112.

Kinderman, A. J., and Monahan, J. F. (1980), "New Methods for Generating Student's t and Gamma Variables," *Computing*, 25, 369-377.

Marsaglia, G., Maclaren, M. D., and Bray, T. A. (1964), "A Fast Procedure for Generating Normal Random Variables," *Communications of the ACM*, 7, 4-10.

Millar, P. W. (1981), "Robust Estimation via Minimum Distance Methods," *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 55, 73-89.

Parr, W. C., and Schucany, W. R. (1980), "Minimum Distance and Robust Estimation," *Journal of the American Statistical Association*, 75, 616-624.

Parr, W. C. (1981), "Minimum Distance Estimation: A Bibliography," *Communications in Statistics*, A10, 1205-1224.

Parr, W. C., and DeWet, T. (1981), "On Minimum Cramér-von Mises-Norm Parameter Estimation," *Communications in Statistics*, A10, 1149-1166.

Pollard, D. (1980), "The Minimum Distance Method of Testing," *Metrika*, 27, 43-70.

Quesenberry, C. P. (1975), "Transforming Samples from Truncation Parameter Distributions to Uniformity," *Communications in Statistics*, 4, 1149-1155.

Solomon, H., and Stephens, M. A. (1978), "Approximations to Density Functions using Pearson Curves," *Journal of the American Statistical Association*, 73, 153-160.

Stacy, E. W. (1962), "A Generalization of the Gamma Distribution," *Annals of Mathematical Statistics*, 28, 1187-1192.

Stephens, M. A. (1974), "EDF Statistics for Goodness of Fit and Some Comparisons," *Journal of the American Statistical Association*, 69, 730-737.

Stephens, M. A. (1976), "Asymptotic Results for Goodness of Fit Statistics with Unknown Parameters," *Annals of Statistics*, 4, 357-369.

Stephens, M. A. (1977), "Goodness of Fit for the Extreme Value Distribution," *Biometrika*, 64, 583-588.

Stephens, M. A. (1979), "Tests of Fit for the Logistic Distribution Based on the Empirical Distribution Function," *Biometrika*, 66, 591-595.

Weissman, I. (1978), "Estimation of Parameters and Large Quantiles Based on the k Largest Observations," *Journal of the American Statistical Association*, 73, 812-815.

Wellner, J. A. (1977), "A Glivenko-Cantelli Theorem and Strong Law of Large Numbers for Functions of Order Statistics," *Annals of Statistics*, 5, 473-480.

APPENDIX

Assumption 1. $X_1, \ldots, X_n$ are IID real-valued random variables, each having distribution function $F(x) = F_{\theta_0}(x)$, $\theta_0 \in \Theta$, $\Theta \subset R_k$. Let $w_\theta(x)$ be a nonnegative weight function which is positive on at least part of the support of F_θ and define

$$d_{F_n}(\theta) = \int_{-\infty}^{\infty} [F_n(x) - F_\theta(x)]^2 w_\theta(x)dx ,$$

where $F_n(x) = n^{-1} \sum I(X_i \le x)$ is the empirical distribution function. When it exists define $\hat{\theta}_n$ by $\min_{\theta\in\Theta} d_{F_n}(\theta) = d_{F_n}(\hat{\theta}_n)$.

The following lemma generalizes Theorem 2.1 (c) of Boos (1981).

LEMMA 1. *Under Assumption 1, if Θ is compact, $d_F(\theta)$ is continuous in θ, $\hat{\theta}_n$ exists, and there exist constants A_1 and A_2 such that*

$$\sup_{\theta\in\Theta} \int_{-\infty}^{\infty} [F(x)(1-F(x))]^{2(1-\varepsilon)} w_\theta(x)dx \le A_1$$

$$\sup_{\theta\in\Theta} \int_{-\infty}^{\infty} [F(x)(1-F(x))]^{1-\varepsilon} |F(x) - F_\theta(x)| w_\theta(x)dx \le A_2 \tag{1}$$

for some $0<\varepsilon<1$, then $\hat{\theta}_n \xrightarrow{wp1} \theta_0$ as $n \to \infty$.

PROOF. Let $||F_n-F||_a = \sup_{-\infty<x<\infty} |F_n(x)-F(x)|/[F(x)(1-F(x))]^a$. Then

$$\begin{aligned} |d_F(\hat{\theta}_n) - d_F(\theta_0)| &\le 2 \sup_{\theta\in\Theta} |d_{F_n}(\theta) - d_F(\theta)| \\ &\le 2[A_1(||F_n-F||_{1-\varepsilon})^2 + 2A_2||F_n-F||_{1-\varepsilon}] \qquad (2) \\ &\xrightarrow{wp1} 0 \text{ as } n \to \infty . \end{aligned}$$

The convergence result for $||F_n-F||_{1-\varepsilon}$ is found in Wellner (1977). Since $d_F(\theta_0) = 0$ and $d_F(\theta) \ne 0$ for $\theta \ne \theta_0$, a subsequence of $\hat{\theta}_n$ which did not converge to θ_0 would contradict (2).

EXAMPLE. Consider the Anderson-Darling weight function $w_\theta(x) = f_\theta(x)/[F_\theta(x)(1-F_\theta(x))]$ and the extreme value location-scale model $F_\theta(x) = \exp(-\exp(-(x-\mu)/\sigma))$, $\Theta = [C_1,C_2] \times [C_3,C_4]$ with $-\infty < C_1 < C_2 < \infty$, $0 < C_3 < C_4 < \infty$. Without loss of generality let $(\mu_0,\sigma_0) = (0,1)$ so that $F(x) = \exp(-\exp(-x))$.

Continuity of $d_F(\theta)$ can easily be shown and $\hat{\theta}_n$ exists since $d_{F_n}(\theta)$ is also continuous. The first integral of (1) can be written as $\int_0^1 B_\theta(t)dt$ where $B_\theta(t) = [F(F_\theta^{-1}(t)) \times (1-F(F_\theta^{-1}(t))]^{2(1-\epsilon)}/t(1-t)$. It suffices to verify $\sup_\theta \int_0^{\delta_1} B_\theta(t)dt \leq A_1^*$ and $\sup_\theta \int_{\delta_2}^1 B_\theta(t)dt \leq A_1^{**}$ since $\sup_\theta B_\theta(t)$ is bounded on $[\delta_1,\delta_2]$, $0 < \delta_1 < \delta_2 < 1$. Now for some positive constants C_5 and C_6 we have

$$\sup_{\theta\epsilon\Theta} \int_0^{\delta_1} B_\theta(t)dt \leq C_5 \sup_{\theta\epsilon\Theta} \int_0^{\delta_1} t^{-1}F(F_\theta^{-1}(t))dt$$

$$\leq C_5 \int_0^{\delta_1} t^{-1}e^{-[-\ln t]^{C_3} e^{-C_2}} dt$$

$$= C_6 \int_{(-\ln\delta_1)^{C_3}e^{-C_2}}^{\infty} v^{(\frac{1}{C_3} - 1)} e^{-v}dv < \infty .$$

The right tail, $\sup_\theta \int_{\delta_2}^1 B_\theta(t)dt$, and the second integral of (1) are handled similarly.

The next lemma formalizes the typical Taylor expansion used to prove asymptotic normality of $\hat{\theta}_n$. For notational simplicity let $d_n(\theta) = d_{F_n}(\theta)$ and $d_n'(\theta) = (d_n^1(\theta),\ldots,d_n^k(\theta))^T$ where $d_n^i(\theta) = \partial/\partial\theta_i\ d_n(\theta)$. Also, let $d_n''(\theta)$ be the matrix of second partial derivatives with elements $d_n^{ij}(\theta) = \partial^2/\partial\theta_i\partial\theta_j\ d_n(\theta)$.

LEMMA 2. Under Assumption 1, if $\hat{\theta}_n$ exists and $\hat{\theta}_n \xrightarrow{p} \theta_0$ where $\theta_0 \epsilon$ interior of Θ and

(a) $d_n^i(\theta)$, $d_n^{ij}(\theta)$, $i,j = 1,\ldots,k$, all exist in some open neighborhood of θ_0 ;

(b) $n^{1/2}[\frac{1}{2}d_n'(\theta_0)] \xrightarrow{d}$ multivariate normal $(0,C)$;

(c) $\frac{1}{2}d_n''(\theta^*) \xrightarrow{p} \Delta$, where $|\theta^*-\theta_0|_e \leq |\hat{\theta}_n-\theta_0|_e$ and Δ has

rank k ($|\cdot|_e$ is Euclidean distance) ;
then $n^{\frac{1}{2}}(\hat{\theta}_n-\theta_0) \xrightarrow{d}$ multivariate normal $(0,\Delta^{-1}C\Delta^{-1})$.

PROOF. By Taylor expansion about $\hat{\theta}_n$

$$d_n'(\theta_0) = d_n'(\hat{\theta}_n) + d_n''(\theta^*) \cdot (\theta_0-\hat{\theta}_n) , \qquad (3)$$

where $|\theta^*-\theta_0|_e \leq |\hat{\theta}_n-\theta_0|_e$. But $d_n'(\hat{\theta}_n) = 0$ as long as $\hat{\theta}_n$ belongs to the interior of Θ. A Slutsky argument using (b) and (c) finishes the proof.

EXAMPLE (continued). The existence of the derivatives $d_n'(\theta)$ and $d_n''(\theta)$ can be verified using the dominated convergence theorem and monotonicity properties of F_θ, f_θ , and w_θ . Recall that $(\mu_0,\sigma_0) = (0,1)$ and $F(x) = \exp(-\exp(-x))$. Then $w_{\theta_0}'(x) = -(w'(x), w(x) + xw'(x))^T$ with $w'(x) = w(x)[f(x)[1 - F(x)]^{-1}-1]$ and

$$\tfrac{1}{2}d_n'(\theta_0) = \begin{pmatrix} \int [F_n(x)-F(x)]f^2(x)/[F(x)(1-F(x))]dx \\ \quad - \tfrac{1}{2}\int [F_n(x)-F(x)]^2 w'(x)dx \\ \int [F_n(x)-F(x)]xf^2(x)/[F(x)(1-F(x))]dx \\ \quad - \tfrac{1}{2}\int [F_n(x)-F(x)]^2[w(x) + xw'(x)]dx \end{pmatrix} . \qquad (4)$$

The first terms of (4) are $n^{-1}\sum IC_{\mu_0}(X_i)$ and $n^{-1}\sum IC_{\sigma_0}(X_i)$ and the second terms are $O_p(n^{-1})$. For example, $|w'(x)| \leq C_7 w(x)$ and

$$\begin{aligned} P(\tfrac{1}{2}\int [F_n(x)-F(x)]^2|w'(x)|dx > C_8 n^{-1}) \\ \leq E\, \tfrac{1}{2}C_7\int [F_n(x)-F(x)]^2 w(x)dx/C_8 n^{-1} \\ = \tfrac{1}{2}C_7/C_8 , \end{aligned}$$

and C_8 can be chosen arbitrarily large. Thus, the Central Limit Theorem and Slutsky's Theorem yield Lemma 2 (b). The typical element of $\frac{1}{2}d_n''(\theta)$ consists of four terms,

$$\begin{aligned} \tfrac{1}{2}d_n^{ij}(\theta) &= \int (-F_\theta^{i}(x))(-F_\theta^{j}(x))w_\theta(x)dx \\ &+ \int [F_n(x)-F_\theta(x)](-F_\theta^{ij}(x))w_\theta(x)dx \\ &+ 2\int [F_n(x)-F_\theta(x)](-F_\theta^{i}(x))w_\theta^{j}(x)dx \\ &+ \tfrac{1}{2}\int [F_n(x)-F_\theta(x)]^2 w_\theta^{ij}(x)dx \ . \end{aligned} \tag{5}$$

Since $-F_\theta^{1}(x) = \sigma^{-1}f((x-\mu)/\sigma)$ and $-F_\theta^{2}(x) = [(x-\mu)/\sigma^2]f((x-\mu)/\sigma)$, the first term of (5) evaluated at any θ^* lying between $\hat{\theta}_n$ and θ_0 converges in probability to Δ_{ij} given by (2.6). All the other terms converge in probability to 0. For example,

$$\begin{aligned} 2\int &|F_n(x)-F_{\theta^*}(x)|\,|-F_{\theta^*}^{1}(x)w_{\theta^*}^{1}(x)|dx \\ &\le C_9 \int [|F_n(x)-F(x)|+|F(x)-F_{\theta^*}(x)|]f_{\theta^*}(x)w_{\theta^*}(x)dx \\ &\le C_9 \, [\sup_x|F_n(x)-F(x)|+\sup_x|F(x)-F_{\theta^*}(x)|]\int f_{\theta^*}(x)w_{\theta^*}(x)dx \\ &\xrightarrow{p} 0 \ . \end{aligned}$$

Received by Editorial Board member July, 1981; Revised July, 1982.

Recommended by William R. Schucany, Southern Methodist University, Dallas, TX

Refereed by William C. Parr, Texas A&M University, College Station, TX