The Professor T. D. Dwivedi Memorial Lecture

April 24, 2013

"Bias Adjustment for Nonlinear Maximum Likelihood Estimators"

David Giles

(University of Victoria)

(Dwivedi Number = 2)

Based on a Research Program with

Helen Feng (UWO)

Ryan Godwin (U Manitoba)

&

Jacob Schwartz (UBC)

1. Introduction

- Widespread use of Maximum Likelihood Estimators (MLE's).
- Motivation: wanted to evaluate the first-order biases of the MLE's of the parameters of the generalized Pareto distribution.
- More generally, interested in bias in cases where likelihood equations (first-order conditions) *do not necessarily admit a closed-form solution*.
- Specifically, consider the $O(n^{-1})$ bias formula introduced by Cox and Snell (1968).
- Other options – bootstrap the bias; "preventive" methods (*e.g.*, Firth, 1993)

2. Outline

- Basic strategy.
- Definitions & notation.
- Two illustrative examples of methodology.
- New results for, gamma distribution, half-logistic distribution, & generalized Pareto distribution.
- Conclusions & related work – completed or in progress.

3. Basic Strategy (Bartlett, 1952)

$l(\theta)$ is log-likelihood for *single* parameter, θ. Assume that $l(\theta)$ is regular w.r.t.

all derivatives up to and including the third order.

If $\hat{\theta}$ is MLE, then $l'(\hat{\theta}) \equiv (\partial l / \partial \theta)_{|\theta=\hat{\theta}} = 0$, and $E[l'(\theta)] = 0$.

$$l'(\theta) + (\hat{\theta} - \theta) l''(\theta) + 0.5(\hat{\theta} - \theta)^2 l'''(\theta) \approx 0.$$

$$E[\hat{\theta} - \theta]\, E[l''(\theta)] + \text{cov.}[(\hat{\theta} - \theta), l''(\theta)] + 0.5E[(\hat{\theta} - \theta)^2]E[l'''(\theta)]$$

$$+ \text{cov.}[0.5(\hat{\theta} - \theta)^2, l'''(\theta)] \approx 0.$$

Approximate other terms to $O(n^{-1})$ and solve for approximate bias.

Note: *Don't need closed-form expression for* $\hat{\theta}$ *itself*.

4. Definitions and Notation

Let $l(\theta)$ be the log-likelihood based on a sample of n observations, with p-dimensional parameter vector, θ. Assume that $l(\theta)$ is regular with respect to all derivatives up to and including the third order.

The joint cumulants of the derivatives of $l(\theta)$ are denoted:

$$k_{ij} = E(\partial^2 l / \partial\theta_i \partial\theta_j) \quad ; \quad i, j = 1, 2, \ldots, p$$

$$k_{ijl} = E(\partial^3 l / \partial\theta_i \partial\theta_j \partial\theta_l) \quad ; \quad i, j, l = 1, 2, \ldots, p$$

$$k_{ij,l} = E[(\partial^2 l / \partial\theta_i \partial\theta_j)(\partial l / \partial\theta_l)] \quad ; \quad i, j, l = 1, 2, \ldots, p \, .$$

(*Typically, this is where some effort is needed*.)

The derivatives of the cumulants are denoted:

$$k_{ij}^{(l)} = \partial k_{ij} / \partial \theta_l \qquad ; \qquad i, j, l = 1, 2, \ldots, p.$$

Fisher's information matrix is $K = \{-k_{ij}\}$, and all of the 'k' expressions are assumed to be $O(n)$.

Cox and Snell (1968) - if the sample data are independent (but not necessarily identically distributed) the bias of the s^{th} element of the MLE of θ $(\hat{\theta})$ is:

$$Bias(\hat{\theta}_s) = \sum_{i=1}^{p}\sum_{j=1}^{p}\sum_{l=1}^{p} k^{si} k^{jl} [0.5 k_{ijl} + k_{ij,l}] + O(n^{-2}); \qquad s = 1, 2, \ldots, p.$$

Cordeiro and Klein (1994) - this bias expression also holds if the data are *non-independent*, and it can be re-written (*more conveniently*) as:

$$Bias(\hat{\theta}_s) = \sum_{i=1}^{p} k^{si} \sum_{j=1}^{p} \sum_{l=1}^{p} [k_{ij}^{(l)} - 0.5k_{ijl}] k^{jl} + O(n^{-2}); \qquad s = 1, 2, \ldots, p.$$

Let $a_{ij}^{(l)} = k_{ij}^{(l)} - (k_{ijl}/2)$, for $i, j, l = 1, 2, \ldots, p$; and define the matrices:

$$A^{(l)} = \{a_{ij}^{(l)}\}; \qquad i, j, l = 1, 2, \ldots, p$$

$$A = [A^{(1)} \mid A^{(2)} \mid \ldots\ldots. \mid A^{(p)}].$$

Cordeiro and Klein (1994) show that the bias of the MLE of θ $(\hat{\theta})$ can be re-written as:

$$Bias(\hat{\theta}) = K^{-1} A\ vec(K^{-1}) + O(n^{-2}).$$

A "bias-corrected" MLE for θ can then be obtained as:

$$\tilde{\theta} = \hat{\theta} - \hat{K}^{-1} \hat{A}\, vec(\hat{K}^{-1}),$$

where $\hat{K} = (K)\,|_{\hat{\theta}}$ and $\hat{A} = (A)\,|_{\hat{\theta}}$.

It can be shown that the bias of $\tilde{\theta}$ is $O(n^{-2})$.

5. Illustrative Results

Example 1 – exponential distribution

Suppose that X is exponentially distributed. The data are i.i.d. with

$$f(x_i) = \theta^{-1} \exp(-x_i / \theta) \;;\quad \theta > 0 \,;\, x_i > 0;\, i = 1, 2, \ldots, n,$$

$$E(X) = \theta \qquad ; \qquad l(\theta) = -n\ln(\theta) - \sum_{i=1}^{n} x_i / \theta$$

$$\partial l / \partial\theta = -n/\theta + \sum_{i=1}^{n} x_i / \theta^2 \qquad ; \qquad \partial^2 l / \partial\theta^2 = n/\theta^2 - 2\sum_{i=1}^{n} x_i / \theta^3$$

$$\partial^3 l / \partial\theta^3 = -2n/\theta^3 + 6\sum_{i=1}^{n} x_i / \theta^4$$

The MLE of θ is $\hat{\theta} = \sum_{i=1}^{n} x_i / n = \bar{x}$. So, this MLE is (exactly) unbiased.

In this example, $p = 1$; $k_{11} = -(n/\theta^2)$; $K = (n/\theta^2)$; and $K^{-1} = (\theta^2/n)$.

Further, $k_{111} = (4n/\theta^3)$; $k_{11}^{(1)} = (2n/\theta^3)$; and $a_{11} = (2n/\theta^3) - 0.5(4n/\theta^3) = 0$.

So, $A = 0$, and the Cox-Snell/Cordeiro-Klein expression for the bias is zero.

Note that not only is this result exactly correct, but it was obtained without needing to write down the MLE itself as a closed form expression.

Example 2 – normal distribution

Suppose that X is normally distributed. The data are i.i.d. with

$$f(x_i) = (2\pi\sigma^2)^{-1/2}\exp(-(x_i-\mu)^2/2\sigma^2)\ ;\ \ 0<\sigma<\infty\ ;\ -\infty<\mu<\infty;$$

$$i = 1, 2, \ldots, n$$

So,

$$l(\mu,\sigma^2) = -(n/2)\ln(2\pi) - n\ln(\sigma^2)/2 - \sum_{i=1}^{n}(x_i-\mu)^2/2\sigma^2.$$

[We know that MLE's are $\hat{\mu}=\bar{x}$ and $\hat{\sigma}^2=\sum_{i=1}^{n}(x_i-\bar{x})^2/n$, where $\hat{\mu}$ is unbiased and $Bias(\hat{\sigma}^2) = -\sigma^2/n$.]

Information matrix is $K = \begin{bmatrix} n/\sigma^2 & 0 \\ 0 & n/2\sigma^4 \end{bmatrix}$, so $vec(K^{-1}) = \begin{pmatrix} \sigma^2/n \\ 0 \\ 0 \\ 2\sigma^4/n \end{pmatrix}$.

Also,

$$A^{(1)} = \begin{bmatrix} 0 & -n/2\sigma^4 \\ -n/2\sigma^4 & 0 \end{bmatrix}, \quad A^{(2)} = \begin{bmatrix} n/2\sigma^4 & 0 \\ 0 & 0 \end{bmatrix},$$

and

$$A = \begin{bmatrix} 0 & -(n/2\sigma^4)(n/2\sigma^4) & 0 \\ -(n/2\sigma^4) & 0 & 0 & 0 \end{bmatrix}.$$

The Cox-Snell/Cordeiro-Klein expression for the bias of $\hat{\theta}$ to $O(n^{-1})$ is

$$Bias\begin{pmatrix} \hat{\mu} \\ \hat{\sigma}^2 \end{pmatrix} = K^{-1} A vec(K^{-1}) = \begin{pmatrix} 0 \\ -\sigma^2 / n \end{pmatrix},$$

Coincides with the exact biases of the MLE's, because they are $O(n^{-1})$ here.

Again, note that this result was obtained without needing to be able to write down the expressions for the MLE's themselves in closed form.

The "bias-adjusted" estimator of σ^2 is $\tilde{\sigma}^2 = \hat{\sigma}^2 - (-\hat{\sigma}^2 / n) = (n+1)\hat{\sigma}^2 / n$, and $Bias(\tilde{\sigma}^2) = -\sigma^2 / n^2$. Correcting for the $O(n^{-1})$ bias yields an estimator that is biased $O(n^{-2})$. Of course, in this particular example, we also know how to eliminate the bias in $\hat{\sigma}^2$ completely – use the estimator $n\hat{\sigma}^2 /(n-1)$.

5. Some New Results

5.1 Two-parameter gamma distribution

The p.d.f. for the gamma distribution, with shape and scale parameters α and θ is:

$$f(x) = \frac{x^{\alpha-1} e^{-x/\theta}}{\Gamma(\alpha)\theta^{\alpha}} \; ; \qquad \alpha, \theta > 0 \; ; \qquad x > 0 \; .$$

(All of following also done in terms of rate parameter, $\lambda = 1/\theta$.)

(Reliability, hydrology, signal processing, meteorology, forensics, *etc*.)

The log-likelihood function, based on a sample of n independent observations, is

$$l = (\alpha - 1)\sum_{i=1}^{n} \log(y_i) - (\sum_{i=1}^{n} y_i)/\theta - n[\log(\Gamma(\alpha)) + \alpha \log(\theta)].$$

We then have:

$$\frac{\partial l}{\partial \alpha} = \sum_{i=1}^{n} \log(y_i) - n[\Psi(\alpha) + \log(\theta)]$$

$$\frac{\partial l}{\partial \theta} = [\sum_{i=1}^{n} y_i - n\alpha\theta] / \theta^2 \qquad ,$$

where $\Psi(\alpha)$ is the usual digamma function, $\Psi(\alpha) = d \log \Gamma(\alpha) / d\alpha$.

No closed-form solution to likelihood equations.

$$Bias\ (\hat{\alpha}) = [\alpha(\Psi_{(1)}(\alpha) - \alpha\,\Psi_{(2)}(\alpha)) - 2]/[2n\{\alpha\,\Psi_{(1)}(\alpha) - 1\}^2]$$

and

$$Bias\ (\hat{\theta}) = \theta[\alpha\,\Psi_{(2)}(\alpha) + \Psi_{(1)}(\alpha)]/[2n\{\alpha\,\Psi_{(1)}(\alpha) - 1\}^2].$$

(Trigamma & tetragamma functions: $\Psi_{(i)}(\alpha) = d^i \Psi(\alpha) / d\alpha^i$; $i = 1, 2$.)

Bias($\hat{\alpha}$) and % biases of $\hat{\alpha}$ and $\hat{\theta}$, are invariant to the value of θ.

In addition, $\hat{\alpha}$ is upward-biased, and $\hat{\theta}$ is downward-biased, to $O(n^{-1})$.

Bias-adjusted estimators:

$$(\tilde{\alpha},\tilde{\theta})' = (\hat{\alpha},\hat{\theta})' - \hat{B}' \qquad ; \ \hat{B} = B\hat{i}as\begin{pmatrix}\hat{\alpha}\\ \hat{\theta}\end{pmatrix} = \hat{K}^{-1}\hat{A}\,vec(\hat{K}^{-1})$$

$$\tilde{\alpha} = \hat{\alpha} - \frac{[\hat{\alpha}(\Psi_{(1)}(\hat{\alpha}) - \hat{\alpha}\,\Psi_{(2)}(\hat{\alpha})) - 2]}{2n[\hat{\alpha}\,\Psi_{(1)}(\hat{\alpha}) - 1]^2}$$

and

$$\tilde{\theta} = \hat{\theta} - \frac{\hat{\theta}[\hat{\alpha}\,\Psi_{(2)}(\hat{\alpha}) + \Psi_{(1)}(\hat{\alpha})]}{2n[\hat{\alpha}\,\Psi_{(1)}(\hat{\alpha}) - 1]^2}.$$

Monte Carlo experiment to compare these bias-corrected estimators with bootstrap bias correction:

$$\breve{\theta} = 2\hat{\theta} - (1/N_B)[\sum_{j=1}^{N_B}\hat{\theta}_{(j)}] \ ,$$

where $\hat{\theta}_{(j)}$ is the MLE of θ obtained from the j^{th} of the N_B bootstrap samples, and similarly for α.

100,000 Monte Carlo replications and N_B = 1,000 (100 million *per* case).

Used *R* – *maxlik* package with Nelder-Mead algorithm.

Illustrative Monte Carlo Results: % Bias [%MSE]; $\alpha = \theta = 1.0$

n	$\hat{\alpha}$	$\tilde{\alpha}$	$\breve{\alpha}$	$\hat{\theta}$	$\tilde{\theta}$	$\breve{\theta}$
10	33.1554	0.1167	-21.0180	-9.3635	-1.1486	-0.4251
	[72.2336]	[29.7664]	[39.3795]	[24.7572]	[28.0324]	[28.2438]
15	20.4645	0.0127	-4.6131	-6.0828	-0.3954	-0.3030
	[27.0769]	[14.8398]	[15.0003]	[16.6527]	[18.1463]	[18.3550]
25	**11.1739**	**0.0029**	**-1.0784**	**-3.7252**	**-0.2206**	**-0.1569**
	[10.9318]	[7.5679]	[7.5833]	[10.0178]	[10.5514]	[10.5860]
50	5.2080	-0.0252	-0.1159	-1.8724	-0.0839	-0.0828
	[4.1068]	[3.4064]	[3.4412]	[5.0491]	[5.1835]	[5.2638]
100	2.4938	-0.0428	-0.0779	-0.8443	0.0599	0.0103
	[1.7757]	[1.6166]	[1.6247]	[2.5651]	[2.6011]	[2.5883]
250	0.9648	-0.0318	-0.0050	-0.3199	0.0439	-0.0037
	[0.6530]	[0.6290]	[0.6334]	[1.0182]	[1.0240]	[1.0117]

5.2 Half-logistic distribution

If $X \sim$ logistic, then $Y = |X|$ has half-logistic distribution, with p.d.f.:

$$f(y) = \frac{(2/\sigma)\exp\{-(y-\mu)/\sigma\}}{[1+\exp\{-(y-\mu)/\sigma\}]^2} \quad ; \quad y \geq \mu > 0, \ \sigma > 0$$

(Used in reliability theory – monotonically increasing hazard.)

If the location parameter is unknown, its MLE is the largest order statistic.

Let $\mu = 0$:

$$l = n\ln(2) - n\ln(\sigma) + (n\bar{y}/\sigma) - 2\sum_{i=1}^{n}\ln[1+\exp(y_i/\sigma)]$$

$$\partial l/\partial\sigma = -(n/\sigma) - (n\bar{y}/\sigma^2) + (2/\sigma^2)\sum_{i=1}^{n}[y_i\exp(y_i/\sigma)]/[1+\exp(y_i/\sigma)]$$

So the MLE for the scale parameter *cannot be expressed in closed form*.

Evaluation of joint cumulants is tedious in this case – *e.g.*, need to establish that

$$E\{[y\exp(y/\sigma)]/[1+\exp(y/\sigma)]\} = \sigma[\ln(2) + 0.5]$$

$$E\{[y^2\exp(y/\sigma)]/[1+\exp(y/\sigma)]^2\} = (\sigma^2/3)[(\pi^2/6) - 1]$$

$$E\{[y^3(\exp(y/\sigma) - \exp(2y/\sigma))]/[1+\exp(y/\sigma)]^3\} = \sigma^3[0.5 - (\pi^2/12)].$$

Then:

$$Bias(\hat{\sigma}) = K^{-1}A\,vec(K^{-1}) = -0.052567665(\sigma/n).$$

The bias is unambiguously negative, and small in relative terms.

Relative bias is invariant to σ.

Unbiased (to $O(n^{-2})$) estimator of σ is:

$$\tilde{\sigma} = (\hat{\sigma} - Bias(\hat{\sigma})) = \hat{\sigma}(n + 0.052567665) / n .$$

Monte Carlo experiment to compare analytic and bootstrap bias corrections.

250,000 Monte Carlo replications and N_B = 1,000 (250 million *per* case).

Used *R* – inversion method; *maxlik* package with Nelder-Mead algorithm.

Prefer analytic bias correction if $n < 25$.

Prefer bootstrap bias correction if $25 \le n \le 250$.

Illustrative Monte Carlo Results (invariant to σ)

n	$\% Bias(\hat{\sigma})$	$\% Bias(\tilde{\sigma})$	$\% Bias(\breve{\sigma})$	$\% MSE(\hat{\sigma})$	$\% MSE(\tilde{\sigma})$	$\% MSE(\breve{\sigma})$
10	-0.4827	0.0404	0.0988	6.9512	7.0221	7.0402
15	-0.3279	0.0214	-0.0390	4.6267	4.6581	4.6793
20	-0.2400	0.0223	0.0415	3.4784	3.4961	3.5016
25	**-0.1719**	**0.0380**	**0.0331**	**2.7966**	**2.8081**	**2.7997**
30	-0.1370	0.0380	0.0166	2.3271	2.3351	2.3342
50	-0.0811	0.0214	0.0135	1.3996	1.4025	1.4022
100	-0.0337	0.0188	-0.0073	0.6988	0.6995	0.7008
200	-0.0137	0.0126	-0.0093	0.3502	0.3504	0.3498
250	-0.0133	0.0077	0.0040	0.2808	0.2809	0.2806

5.3 Generalized Pareto distribution

Widely used in POT method for extreme value analysis. Often a relatively small number of extreme values.

$$F(y) = 1 - \left(1 + \xi y/\sigma\right)^{-1/\xi}; \quad y > 0,\ \xi \neq 0$$
$$= 1 - \exp(-y/\sigma); \qquad \xi = 0$$

$$f(y) = (1/\sigma)\left(1 + \xi y/\sigma\right)^{-1/\xi - 1}; \quad y > 0, \xi \neq 0$$
$$= (1/\sigma)\exp(-y/\sigma); \qquad \xi = 0$$

$0 < y < \infty$ if $\xi \geq 0$; and $0 < y < -\sigma/\xi$ if $\xi < 0$.

Maximum likelihood estimation of the parameters of the GPD can be very challenging in practice:

- r^{th}. integer-order moment exists if $\xi < 1/r$

- MLE for $\theta' = (\xi, \sigma)$: existence requires $\xi \geq -1$; regularity requires $\xi \geq -1/3$.

Assuming independent observations, the log-likelihood function is:

$$l(\xi,\sigma) = -n\ln(\sigma) - (1+1/\xi)\sum_{i=1}^{n}\ln(1+\xi y_i/\sigma).$$

$$\partial l/\partial\xi = \xi^{-2}\sum_{i=1}^{n}\ln(1+\xi y_i/\sigma) - (1+\xi^{-1})\sum_{i=1}^{n}[y_i/(\sigma+\xi y_i)]$$

$$\partial l/\partial\sigma = \sigma^{-1}\{-n + (1+\xi)\sum_{i=1}^{n}[y_i/(\sigma+\xi y_i)]\} \quad .$$

The likelihood equations do not admit a closed-form solution.

Monte Carlo experiment to compare analytic and bootstrap bias corrections.

Also have compared with Zhang's "likelihood moment" estimator, & quasi-Bayesian estimator of Zhang & Stephens.

50,000 Monte Carlo replications and N_B = 1,000   (50 million *per* case).

Used *R* – *evd* package and Scott Grimshaw's code.

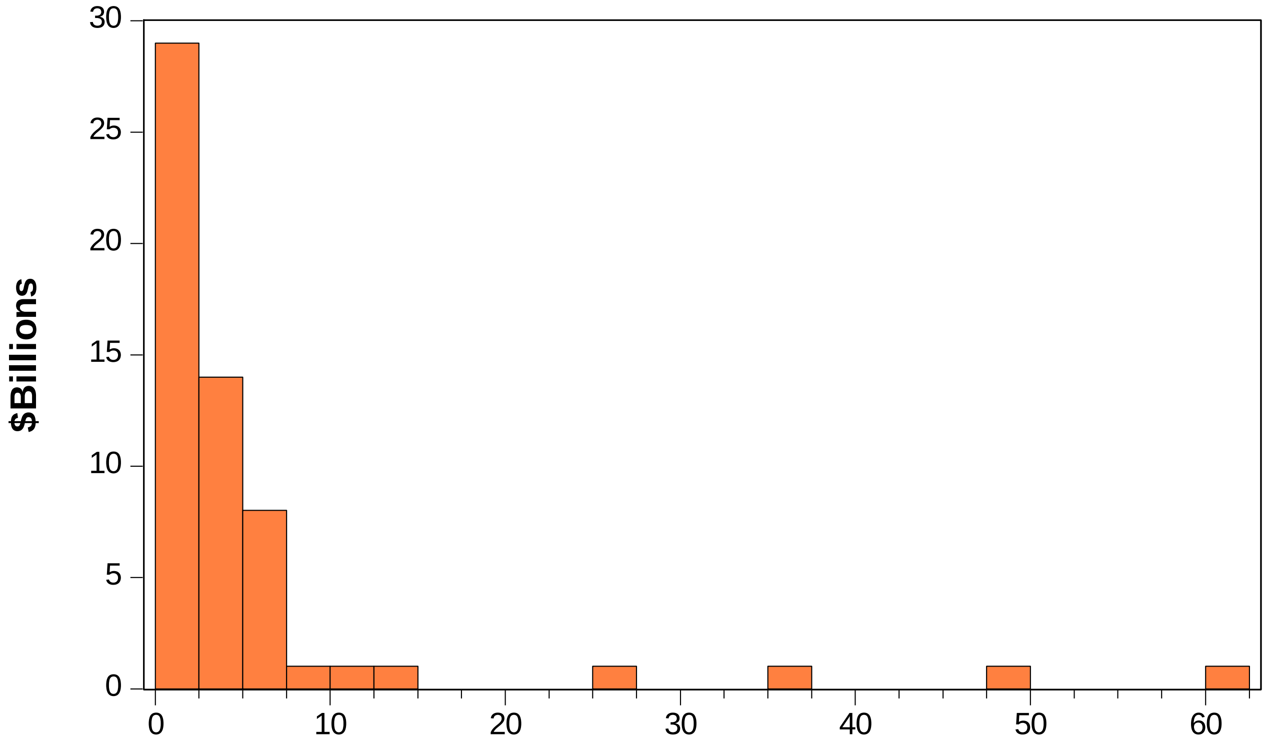
Illustrative Monte Carlo Results: $\xi = 0.5$; $\sigma = 1.0$

n	$\%Bias(\hat{\xi})$ [$\%MSE(\hat{\xi})$]	$\%Bias(\tilde{\xi})$ [$\%MSE(\tilde{\xi})$]	$\%Bias(\breve{\xi})$ [$\%MSE(\breve{\xi})$]	$\%Bias(\hat{\sigma})$ [$\%MSE(\hat{\sigma})$]	$\%Bias(\tilde{\sigma})$ [$\%MSE(\tilde{\sigma})$]	$\%Bias(\breve{\sigma})$ [$\%MSE(\breve{\sigma})$]
50	**-12.1930**	**-1.9603**	**-6.2334**	**5.9062**	**-1.7691**	**2.1070**
	[25.9748]	[24.2179]	[31.4349]	[7.1023]	[10.1278]	[7.5628]
75	-5.7610	0.3024	-2.7424	3.7248	-0.5590	1.3044
	[13.3837]	[11.4417]	[20.6066]	[4.6853]	[3.5037]	[5.1182]
100	-4.2936	0.1299	-0.3207	2.7687	-0.2444	0.1146
	[9.8939]	[8.7299]	[9.6071]	[3.4162]	[2.7323]	[3.2064]
125	-4.5675	-0.9802	0.1478	2.4156	0.0035	0.3841
	[9.5717]	[8.6061]	[10.2316]	[2.7411]	[2.2631]	[2.9058]
150	-3.5011	-0.5653	-0.2740	1.9333	-0.0105	0.1147
	[7.4976]	[6.8917]	[6.2552]	[2.2364]	[1.9205]	[2.0722]
200	-2.0973	0.0538	0.8231	1.3237	-0.0673	-0.0772
	[4.7031]	[4.4580]	[4.8872]	[1.6009]	[1.4465]	[1.6480]

WEATHER-RELATED DISASTERS IN THE U.S.

(1980 - 2003)

Weather-Related Damages Exceeding $1 Billion
(U.S.: 1980 - 2003)



Series: DAMAGE
Sample 1 58
Observations 58

Mean	6.034483
Median	2.450000
Maximum	61.60000
Minimum	1.100000
Std. Dev.	11.02268
Skewness	3.700484
Kurtosis	16.58785
Jarque-Bera	578.5599
Probability	0.000000

Maximum Likelihood Estimation of GPD

$\hat{\xi}$ (a.s.e.)	0.736	(0.223)	$\tilde{\xi}$ (b.s.e.)	0.803	(0.220)
$\hat{\sigma}$ (a.s.e.)	1.709	(0.410)	$\tilde{\sigma}$ (b.s.e.)	1.569	(0.352)

$V\hat{a}R_{0.05}$	$19.7 Billion	$V\tilde{a}R_{0.05}$	**$20.7 Billion**
$E\hat{S}_{0.05}$	$78.3 Billion	$E\tilde{S}_{0.05}$	**$109.0 Billion**

6. Conclusions & Related Work

- Analytic bias-correction using Cox-Snell bias approximation can be applied *even when we can't express MLE in closed form.*

- Can get dramatic reductions in %Bias, without increasing %MSE.

- Bootstrapping bias and then correcting often less successful for small n.

- Other results:

  Poisson *regression* model (with Helen Feng).

  ZIP model (with Jacob Schwartz)

  Nakagami distribution (with Jacob Schwartz & Ryan Godwin)

  Topp-Leone distribution

  Generalized Rayleigh distribution (with Xiao Ling)

  GPD in terms of VaR & shape parameter (with Helen Feng)