Graphical Estimators from Probability Plots with Censored Data

Anupap (Paul) Somboonsavatdee

Joint work with

Vijay Nair
Ananda Sen

Department of Statistics
University of Michigan – Ann Arbor

JRC 2006 – Knoxville TN

Overview

- Introduction and Example
- Graphical Estimators
- Large-Sample Results
- Small-Sample Simulation Results
- Summary

Introduction

• Probability plots popular tools for assessing distributional assumptions
  -- especially useful in (log) location-scale models

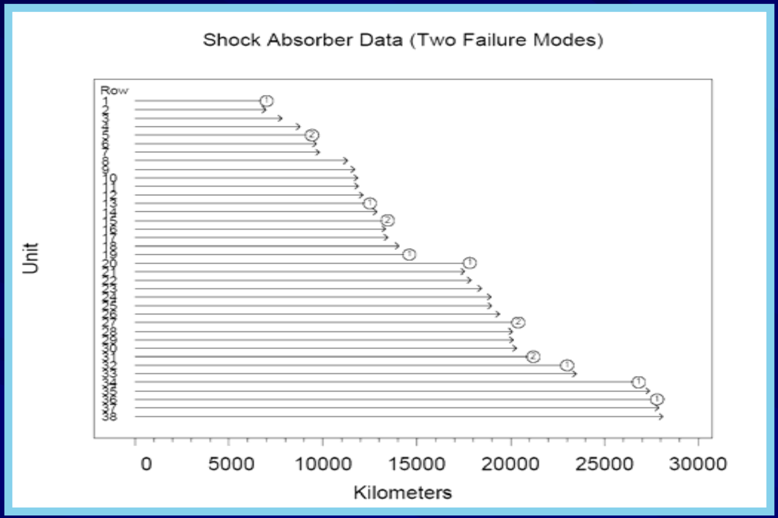
• If model is reasonable, plot looks linear

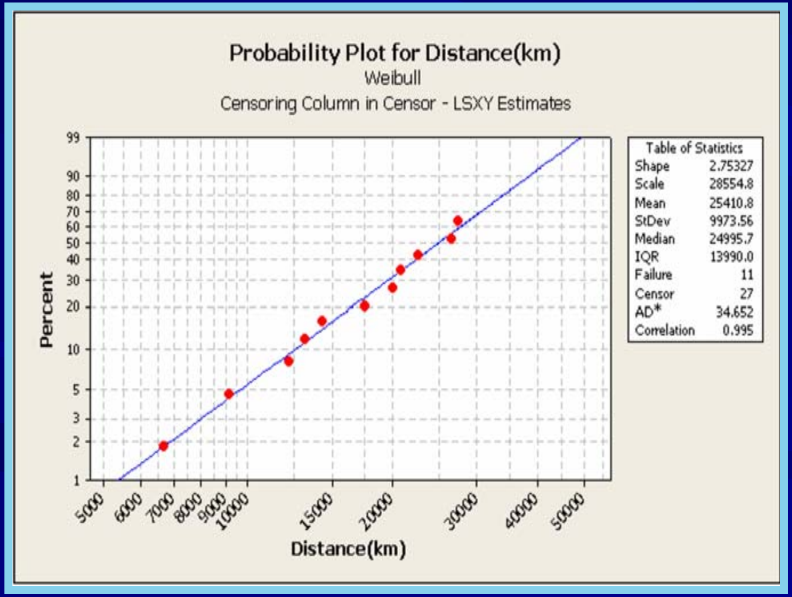
• Graphical method of estimation – fit a line to the Q-Q plot and use the fitted intercept and slope to estimate location and scale parameters
  → estimates of quantiles (design life), hazard rate, CDF, etc.

• Easy, non-iterative method of estimation with censored data
  as compared to MLE

• In common use among reliability engineers
  (Default estimation method in Minitab version 14.0)

An Example





Shock Absorber Data

(Meeker and Escobar, 1998)

← TTF of vehicle shock absorbers (in kilometers)
Two failure modes: M1 and M2

Goal: Estimate CDF of mode M1, mode M2, and overall shock absorber

Competing risks → Random right censoring

← Weibull probability plot with censored data
Model looks reasonable → fit LS line to plot and estimate location-scale parameters

← Minitab output for one failure mode

Graphical Estimators

- For uncensored case, Q-Q plot $(X_{iN},\ Y_{iN})$, $1 \le i \le N$, where $X_{iN} = F_0^{-1}\left(\frac{i-.5}{N}\right)$

Let $\bar{Y} = \frac{1}{N}\sum_{i=1}^{N} Y_{iN}$, $S_Y^2 = \frac{1}{N}\sum_{i=1}^{N}(Y_{iN}-\bar{Y})^2$, and $\bar{X}$ and S_X^2 be defined similarly.

- Graphical least-squares (LS) estimators are

$$\hat{\sigma} = \frac{\frac{1}{N}\sum_{i=1}^{N}(X_{iN}-\bar{X})(Y_{iN}-\bar{Y})}{S_X^2}$$

and $\hat{\mu} = \bar{Y} - \hat{\sigma}\,\bar{X}.$

Multiply Right Censored Data

- Plot ($X_{i\,M}$, Y_{iM}), $1 \leq i \leq M$, where $X_{i\,M} = F_0^{-1}(p_{iM})$, where p_{iM} is "plotting position" from Kaplan-Meier CDF estimator, and Y_{iM}'s are the ordered uncensored observations.

- Can minimize weighted LS with weights $\{w_{iM}\}$ s.t. $\sum_{i-1}^{M} w_{iM} = 1$ which gives WLS estimators of location and scale.

- Ordinary least-squares (OLS) estimators -- $w_{iM} = 1/M$.

Properties of Graphical Estimators

- Easy to compute (non-iterative)
- Standard errors?
  - Incorrect use of output from regression packages
  - Minitab Version 14.0 – uses expressions for MLEs
- Consistent? Asymptotically normal?
- How efficient compared to MLEs?
- How to compute standard errors in practical situations?

Large-Sample Results

Assume a random right-censoring frameowrk

Fix $T < \infty$ such that $(1-F(T))(1-G(T)) > 0$.

Define $W_N(t) = w_{iM}/\Delta_{iM}$ for $\hat{F}_N(Y_{i-1,M}) < t \leq \hat{F}_N(Y_{iM}) \wedge \hat{F}_N(T)$, and $= 0$ for $\hat{F}_N(Y_{MM}) < t \leq 1$ for WLS estimators defined earlier.

where $\Delta_{iM} = \hat{F}_N(Y_{iM}) - \hat{F}_N(Y_{i-1,\,M})$ and G is the CDF of the censoring distribution.

Conditions:

1. The location-scale family $F_0(y)$ has a density that is continuous and strictly positive on $(-\infty,\infty)$
2. The CDF $F_0(y)$ has an absolute moment of order $\nu \geq 4$.
3. There exists a function W(t) that is continuous a.e. for $t \in (0,1)$ with $|W(t)| \leq L$ for some constant L, and $W_N(t) \xrightarrow{a.s.} W(t)$.

Consistency and Asymptotic Normality of Graphical Estimators

Under Conditions 1-3, $N^{1/2}[(\hat{\mu}_w-\mu),(\hat{\sigma}_w-\sigma)]^T$ has a limiting bivariate normal distribution with mean $\mathbf{0}$ and covariance matrix V with elements

$$V_{11}=\sigma^2(\lambda_{11}+m_1^2\lambda_{22}/4 - m_1\lambda_{12}), \qquad V_{12}=\sigma^2(\lambda_{12}+m_2\lambda_{22}/2)/2, \qquad V_{22}=\sigma^2\lambda_{22}/4$$

where

$$\lambda_{ij}=\int_0^{F(T)}\int_0^{F(T)}(1-s)(1-t)B^*(s\wedge t)W(s)W(t)d\tilde{H}_i(s)d\tilde{H}_j(t)$$

with

$$s\wedge t=\min(s,t), \qquad \tilde{H}_1(t)=F_0^{-1}(t), \qquad \tilde{H}_2(t)=(F_0^{-1}(t)-m_1)^2/m_2$$

$$m_1=\int_0^{F(T)}W(t)F_0^{-1}(t)dt, \qquad m_1=\int_0^{F(T)}W(t)[F_0^{-1}(t)-m_1]^2dt, \qquad B^*(t)=\int_{-\infty}^{F^{-1}(t)}\frac{dF(u)}{(1-F(u))^2(1-G(u))}$$

Asymptotic relative efficiency (ARE) of Graphical Estimators vs MLEs

Underlying Distribution		ARE: Location, Censoring Distribution: Same	ARE: Location, Censoring Distribution: Uniform	ARE: Scale, Censoring Distribution: Same	ARE: Scale, Censoring Distribution: Uniform
Weibull	$\theta = 0$	0.949	0.949	0.551	0.551
	0.25	0.840	0.802	0.552	0.555
	0.50	0.667	0.560	0.554	0.547
	0.75	0.520	0.434	0.555	0.543
Log-normal	$\theta = 0$	1.000	1.000	1.000	1.000
	0.25	0.997	0.994	0.993	0.970
	0.50	0.998	0.955	0.997	0.916
	0.75	1.000	0.862	0.999	0.855
Log-logistic	$\theta = 0$	0.912	0.913	0.911	0.909
	0.25	0.967	0.991	0.854	0.765
	0.50	0.961	0.847	0.819	0.646
	0.75	0.939	0.580	0.814	0.587

← Low

← High

← Intermediate

Impact of censoring Distributions?

Asymptotic relative efficiency (ARE) of Graphical Estimators vs MLEs for Quantiles

Censoring Settings		ARE $X_{0.10}$	$X_{0.25}$	$X_{0.50}$	$X_{0.75}$	$X_{0.90}$
Weib./Weib.	$\theta = 0$	0.685	0.792	0.951	0.869	0.709
	0.25	0.680	0.780	0.883	0.753	0.634
	0.50	0.668	0.732	0.726	0.612	0.556
	0.75	0.621	0.606	0.544	0.507	0.500
Weib./Unif.	$\theta = 0$	0.685	0.792	0.951	0.869	0.709
	0.25	0.681	0.774	0.855	0.716	0.609
	0.50	0.647	0.678	0.617	0.521	0.493
	0.75	0.570	0.514	0.446	0.433	0.441
LNorm./LNorm.	$\theta = 0$	1.000	1.000	1.000	1.000	1.000
	0.25	0.995	0.997	0.996	0.994	0.993
	0.50	0.999	0.999	0.998	0.997	0.996
	0.75	0.999	1.000	1.000	1.000	1.000
LNorm./Unif.	$\theta = 0$	1.000	1.000	1.000	1.000	1.000
	0.25	0.987	0.994	0.994	0.984	0.977
	0.50	0.955	0.965	0.955	0.931	0.920
	0.75	0.889	0.885	0.862	0.847	0.844
Llogis./Llogis.	$\theta = 0$	0.912	0.912	0.913	0.912	0.911
	0.25	0.887	0.926	0.967	0.945	0.910
	0.50	0.873	0.927	0.961	0.911	0.870
	0.75	0.875	0.927	0.939	0.892	0.859
Llogis./Unif.	$\theta = 0$	0.912	0.912	0.913	0.912	0.911
	0.25	0.846	0.924	0.991	0.918	0.845
	0.50	0.767	0.854	0.847	0.724	0.670
	0.75	0.654	0.640	0.580	0.555	0.554

← Low

← High

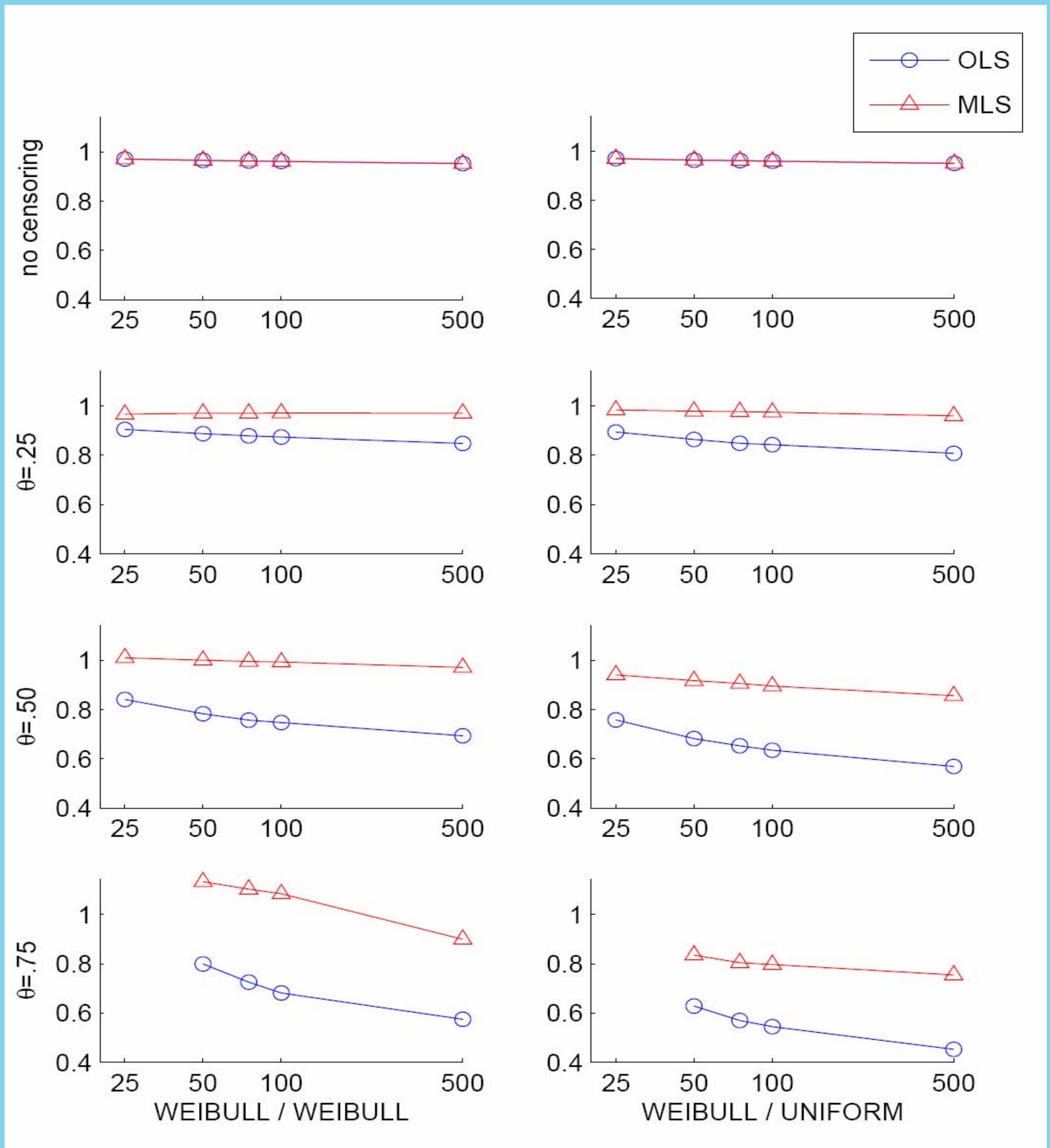
← Intermediate

Finite-Sample Results from Simulation Study

All results are based on simulation size of 10,000.

- Estimators - OLS, MLS, and MLE
- μ=0 and σ=1

- Censoring Settings
  - Sample size -- N = 25, 50, 75, 100, 500
  - Censoring rate -- θ = 0, .25, .50 , and .75
  - Distributions -- Weibull, log-normal, and log-logistic
  - Multiple right censoring -- F and G from the same family , G is Uniform.
    - When F and G are from the same family scale (σ) parameters are held the same while the location (μ) parameters are determined by censoring rate (θ).



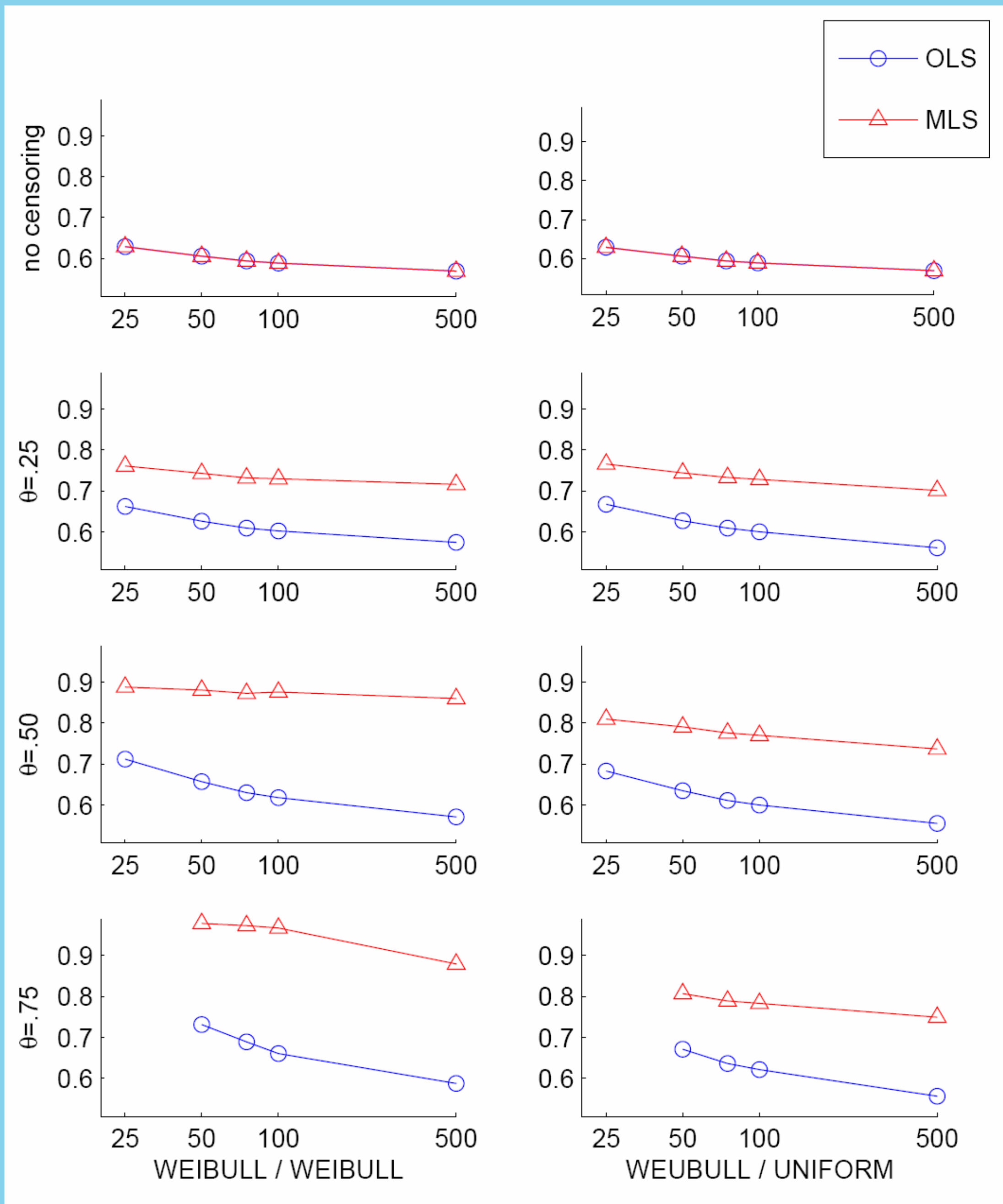
Relative Efficiencies of Location Estimators

Weibull TTF distribution

OLS and MLS
vs MLE

RE decrease with sample size
OLS is inefficient
Worse with uniform cens.

MLS quite efficient!

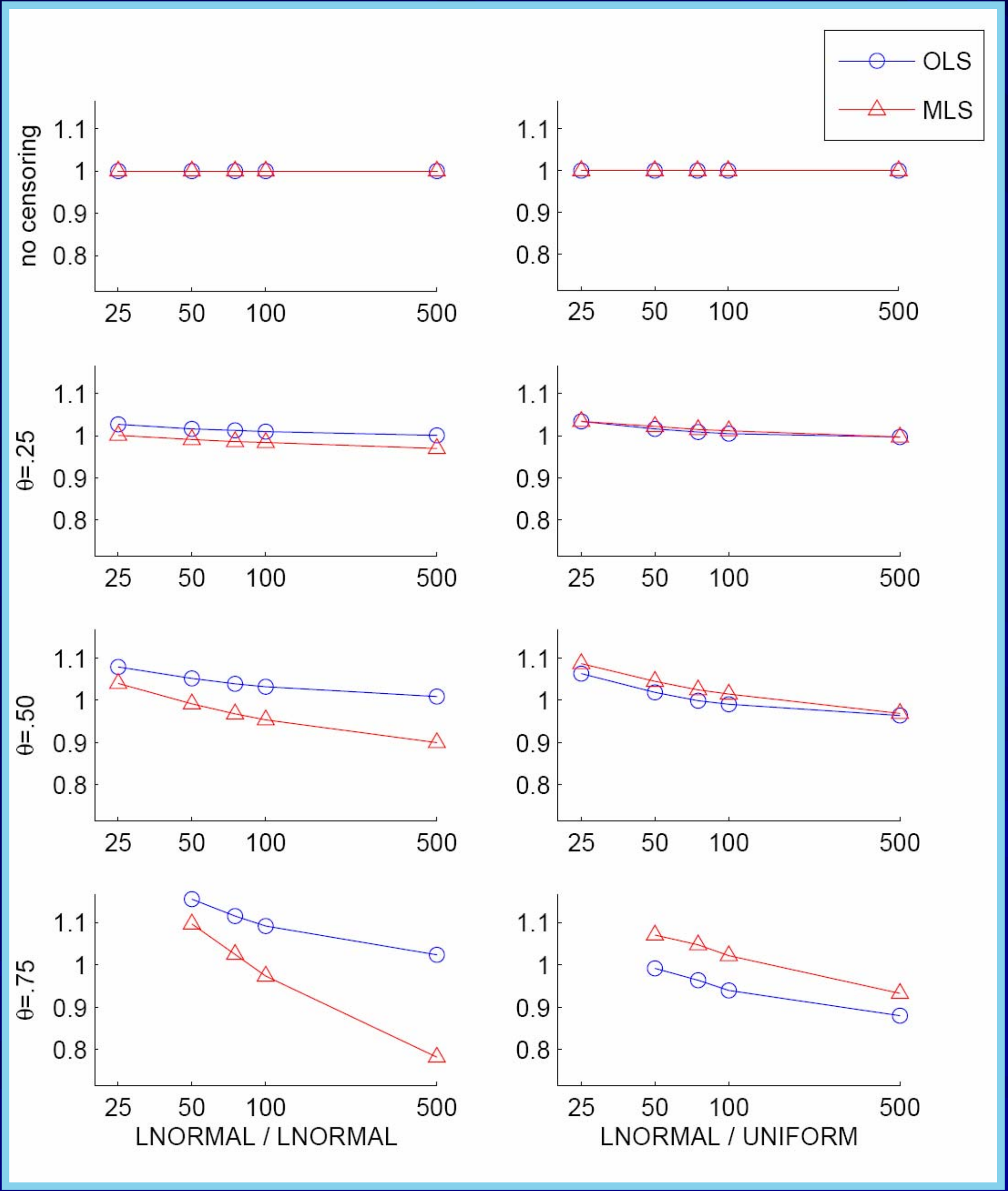


Relative Efficiencies of Scale Estimators

Weibull TTF distribution

OLS and MLS
vs MLE

RE of OLS is quite poor
Worse than location



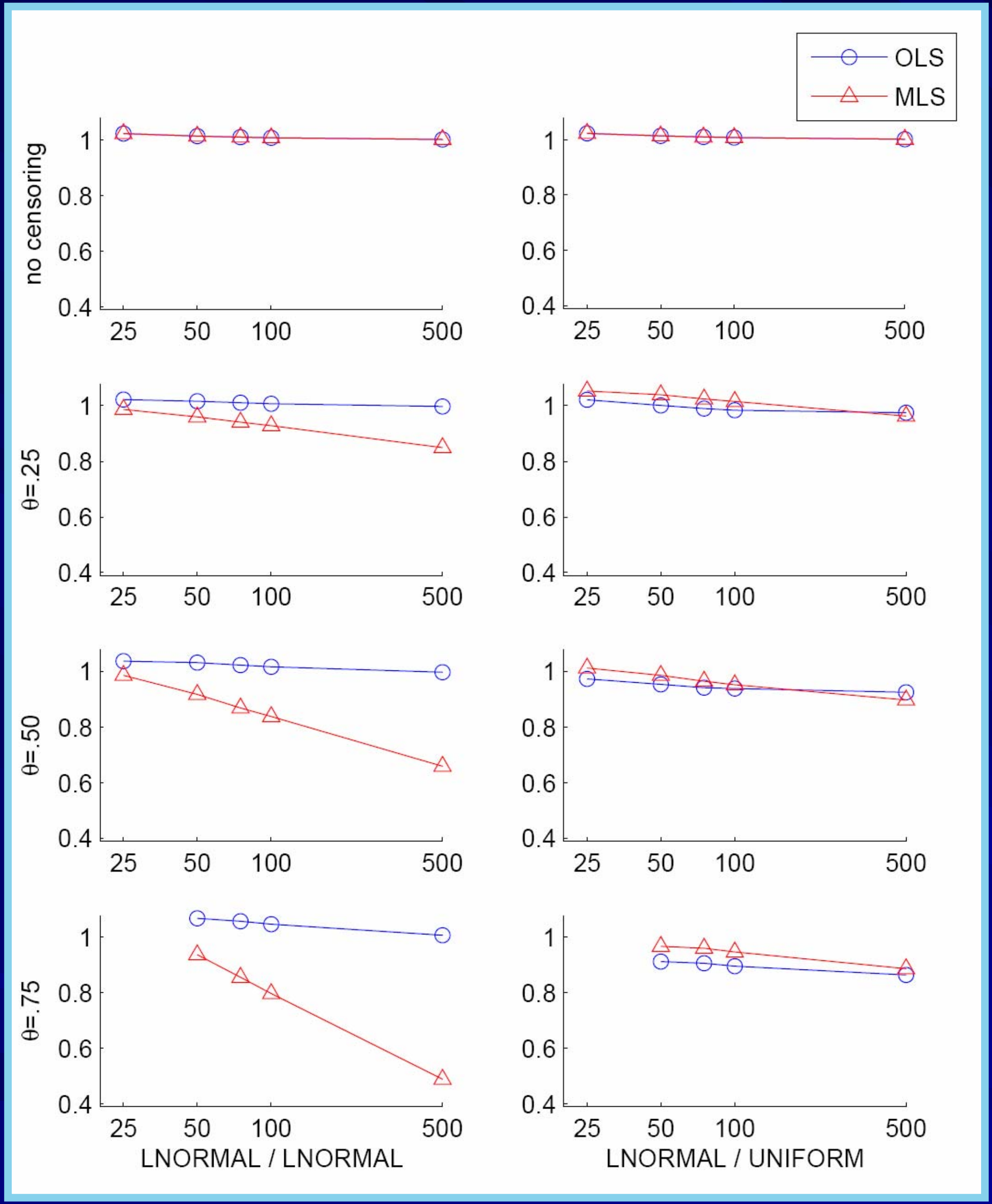
Relative Efficiencies of Location Estimators

Log-normal TTF distribution

OLS and MLS
vs MLE

RE decrease with sample size.
OLS is very efficient.
Worse with uniform cens.

MLS is more efficient than OLS
with uniform cens.
Worse with log-normal cens.



Relative Efficiencies of
Scale Estimators

Log-normal TTF distribution

OLS and MLS
vs MLE

RE of OLS is high
But a little smaller than location

Overall conclusions

- Graphical estimators are consistent and asymptotically Gaussian.
  Variances and covariances are complicated. Must be computed numerically.
  Bootstrap is a practical alternative.

- No consistent pattern for the magnitude of the biases across various choices of failure time distributions and censoring schemes.

- Relative efficiencies of the graphical estimators suggest that they do well for log-normal failure-time distributions, reasonably for log-logistic distributions, and poorly for Weibull distributions.

- For small-samples, the MLS is quite efficient (and does much better than the OLS) for Weibull!

Summary

- We are not recommending the use of these graphical estimators over MLEs

- Rather, our goal is to:
  - Shed light on the properties of these quick-and-easy estimators that are popular among some practitioners

  - Provide ways of computing standard errors and point out pitfalls in using standard outputs from regression packages!

Questions