ESTIMATION METHOD 11: Cumulative Distribution Function and Variance for Proportion of a Resource; Simulation-Extrapolation Method

1 Scope and Application

This report describes an estimation procedure called Simulation-Extrapolation (Cook and Stefanski, 1994) used to estimate a population cumulative distribution when sample units are measured with error. Estimates obtained when the measurement error is ignored are biased and may be misleading. The Simulation-Extrapolation(SIMEX) method reduces the bias induced by measurement error by establishing a relationship between measurement error-induced bias and the variance of the error. Extrapolating this relationship back to the case of no measurement error, an estimator with smaller bias is produced. The method assumes that the variance of the measurement error in the observed sample is known or at least well estimated. A variance estimator of the SIMEX estimator is also described.

2 Statistical Estimation Overview

Let $\mathbf{U} = \{U_1, U_2, \cdots, U_N\}$ be the true (unobserved) data subject to measurement error and $\mathbf{X} = \{X_1, X_2, \cdots, X_N\}$ denote the observed data where X_i is a measure of U_i. A functional measurement error model with additive independent normal error is assumed. That is, $X_i = U_i + \sigma Z_i$, for 1, , n, where $\{Z_i\}_{i=1}^{n}$ are mutually independent, independent of random sampling, and identically distributed standard normal random variables. Hence, the measurement errors in the observed sample have mean zero and variance σ^2.

The estimand F(t) is the cumulative distribution of some population of interest. Let $\hat{F}_U(t) = T(U)$, where T(U) is a function of the data $\{U_i\}_{i=1}^{n}$, denote the unbiased estimator of F(t) that would be calculated in the absence of measurement error and let $T_{Var}(U)$ be an unbiased estimator of the variance of $\hat{F}_U(t)$. The naive estimator $\hat{F}_X(t) = T(X)$ obtained when the measurement error is ignored is biased for F(t). Estimates of F(t) with even greater bias can be obtained by adding additional measurement error in known increments to *X*. From these estimates, a relationship may be established between the bias induced by the measurement error and the variance of the added measurement error.

The first step of the procedure consists of computing a large number *B* of pseudo data sets $\{\{X^*_{b,i}(\lambda)\}_{i=1}^{n}\}_{b=1}^{B}$ for different values of λ, where $X^*_{b,i}(\lambda) = X_i + \sigma\sqrt{\lambda}Z^*_{b,i}$ for i = 1, ⋯, n and b = 1, ⋯, B, and $\{\{Z^*_{b,i}(\lambda)\}_{i=1}^{n}\}_{b=1}^{B}$ are mutually independent, independent of the data $\{X_i\}_{i=1}^{n}$, and identically distributed standard normal pseudo-random variables. For fixed λ, the measurement error variance of the additional errors $\{\{\sigma\sqrt{\lambda}Z^*_{b,i}(\lambda)\}_{i=1}^{n}\}_{b=1}^{B}$ is $\sigma^2\lambda$. Therefore, the total measurement error in $X^*_{b,i}(\lambda)$ for $1 \le i \le$ n and $1 \le$ b $\le$ B has variance $\lambda(\sigma^2 + 1)$. The estimates $\hat{F}_{X,\lambda,b}(t) = T(\{X^*_{b,i}(\lambda)\}_{i=1}^{n})$ are then calculated for b = 1, ⋯, B. The average of these estimates is used to estimate the expectation of $\hat{F}_{X,\lambda,b}(t)$ with respect to the distribution of the pseudo-random variates $\{Z^*_{b,i}\}_{i=1}^{n}$. This is the simulation step of the SIMEX method.

Next the expectation, $\hat{F}_{X,\lambda}(t) = E\left\{\hat{F}_{X,\lambda,b}(t) \mid \{X_i\}_{i=1}^{n}\right\}$, is modeled by a quadratic function in λ. That is, $\hat{F}_{X,\lambda}(t)$ is estimated by $\hat{F}_{X,\lambda}(t) \approx \beta_0 + \beta_1\lambda + \beta_2\lambda^2$. Estimates of the model parameters β_0, β_1, and β2 are obtained by least squares estimation of $\hat{F}_{X,\lambda}(t)$ on λ. Note that the error in the data $\{X_i + \sigma\sqrt{\lambda}Z^*_{b,i}\}_{i=1}^{n}$ has variance $\lambda(\sigma^2 + 1)$, hence, taking λ = -1 corresponds to the case of no measurement error. The SIMEX estimator is obtained by extrapolating the assumed model back to λ = -1 (i.e., the case of no measurement error). The resulting estimator is given by $\hat{F}_{SIMEX}(t) \approx \hat{\beta}_0 - \hat{\beta}_1 + \hat{\beta}_2$. Different parameter estimates must be calculated for each value of t.

For additional information on the SIMEX estimator see Cook and Stefanski (1994).

3 Conditions Under Which This Method Applies

- Data observed with error.
- Additive independent and normally distributed measurement errors with mean zero and common variance σ^2.

4 Required Elements

4.1. Input Data

$\{X_1, \cdots, X_N\}$	=	probability sample of size *n*, where X is a measured value of U_i and $\{U_i\}_{i=1}^{n}$ is the true (unobserved) sample subject to measurement error.
σ^2	=	measurement error variance.
$V\hat{a}r(\hat{\sigma}^2)$	=	variance estimate of $\hat{\sigma}^2$ when σ^2 is estimated; zero otherwise.

5 Formulas and Definitions

Let t denote a fixed argument in the following definitions and formulas. Define:

- $\hat{F}_U(t)$ = estimator of the population cumulative distribution (CD) in the absence of measurement error.
- $\hat{F}_{X,\lambda,b}(t)$ = estimator based on the data $\{X_i + \sigma\sqrt{\lambda}Z^*_{b,i}\}_{i=1}^{n}$, where $\{X_i\}_{i=1}^{n}$ is the observed sample, $\{Z_i\}_{i=1}^{n}$ are standard normal pseudo-random variables, σ^2 is the measurement error variance, and λ > 0 is a constant.
- $\hat{\tau}_b^2(\lambda)$ = estimator of the variance of $\hat{F}_{X,\lambda,b}(t)$.
- $\hat{F}_{X,\lambda}(t)$ = estimator of the expectation of $\hat{F}_{X,\lambda,b}(t)$ with respect to the distribution of the pseudo-random errors $\{Z^*_{b,i}\}_{i=1}^{n}$.
- $\hat{\tau}^2(\lambda)$ = estimator of the expectation of $\hat{\tau}_b^2(\lambda)$ with respect to the distribution of $\{Z^*_{b,i}\}_{i=1}^{n}$ only.

- $s_\Delta^2(\lambda)$ = estimator of $\mathrm{Var}\left\{\hat{F}_{X,\lambda,b}(t) - \hat{F}_{X,\lambda}(t) \mid \{X_i\}_{i=1}^n\right\}$.

- $\hat{F}_{X,\epsilon,\lambda,b}(t)$ = estimator based on the data $\left\{X_i + \sqrt{(\sigma^2+\epsilon)\lambda}\, Z_{b,i}^*\right\}_{i=1}^n$, where $\{X_i\}_{i=1}^n$ is the observed sample, $\{Z_{b,i}^*\}_{i=1}^n$ are standard normal pseudo-random variables, σ^2 is the measurement error variance, and $\epsilon > 0$ ($\epsilon \approx 0$) and $\lambda > 0$ are constants.

- $\hat{F}_{X,\epsilon,\lambda}(t)$ = estimator of the expectation of $\hat{F}_{X,\epsilon,\lambda,b}(t)$ with respect to the distribution of $\{Z_{b,i}^*\}_{i=1}^n$ only.

- $\hat{f}_{X,\lambda}(t)$ = estimator of the derivative of $\hat{F}_{X,\lambda}(t)$ with respect to the measurement error variance σ^2.

- $\hat{F}_{SIMEX}(t)$ = SIMEX estimator.

- $\hat{Var}(\hat{F}_{SIMEX}(t))$ = variance estimator of the SIMEX estimator.

- $\hat{L}(t)$ = lower 100(1)% confidence limit for $\hat{F}_{SIMEX}(t)$.

- $\hat{U}(t)$ = upper 100(1)% confidence limit for $\hat{F}_{SIMEX}(t)$.

The formulas for the above definitions are as follows:

- $\hat{F}_U(t) = T(\{U_i\}_{i=1}^n)$

- $\hat{F}_{X,\lambda,b}(t) = T(\{X_i + \sigma\sqrt{\lambda}\, Z_{b,i}^*\}_{i=1}^n)$

- $\hat{\tau}_b^2(\lambda) = T_{Var}(\{X_i + \sigma\sqrt{\lambda}\, Z_{b,i}^*\}_{i=1}^n)$

- $\hat{F}_{X,\lambda}(t) = \dfrac{1}{B}\sum_{b=1}^{B} \hat{F}_{X,\lambda,b}(t)$

- $\hat{\tau}^2(\lambda) = \dfrac{1}{B}\sum_{b=1}^{B} \hat{\tau}_b^2(\lambda)$

- $s_\Delta^2(\lambda) = \dfrac{1}{B-1}\sum_{b=1}^{B}\left(\hat{F}_{X,\lambda,b}(t) - \hat{F}_{X,\lambda}(t)\right)^2$

- $\hat{F}_{X,\epsilon,\lambda,b}(t) = T(\{X_i + \sqrt{(\sigma^2+\epsilon)\lambda}\, Z_{b,i}^*\}_{i=1}^n)$

- $\hat{F}_{X,\epsilon,\lambda}(t) = \dfrac{1}{B-1}\sum_{b=1}^{B}\left(\hat{F}_{X,\lambda,b}(t) - \hat{F}_{X,\lambda}(t)\right)^2$

- $\hat{f}_{X,\lambda}(t) = \frac{1}{\epsilon}\left(\hat{F}_{X,\epsilon,\lambda}(t) - \hat{F}_{X,\lambda}(t)\right)$

- $\hat{F}_{SIMEX}(t) = V^{T}\Gamma$

If σ^2 is known,

- $\hat{Var}(\hat{F}_{SIMEX}(t)) = V^{T}\eta$

If σ^2 is estimated,

- $\hat{Var}(\hat{F}_{SIMEX}(t)) = V^{T}\eta + \hat{Var}(\hat{\sigma}^2)\cdot\{V^{T}\gamma\}^2$
- $\hat{L}(t) = \hat{F}_{SIMEX}(t) - z_{1-\alpha/2}\cdot\sqrt{\hat{Var}(\hat{F}_{SIMEX}(t))}$
- $\hat{U}(t) = \hat{F}_{SIMEX}(t) + z_{1-\alpha/2}\cdot\sqrt{\hat{Var}(\hat{F}_{SIMEX}(t))}$

where

$U = \{U_i\}_{i=1}^{n}$ = (true) unobserved data values,

$X = \{X_i\}_{i=1}^{n}$ = sample observed with error,

$\{\{Z^{*}_{b,i}(\lambda)\}_{i=1}^{n}\}_{b=1}^{B}$ = independent and identically distributed standard normal pseudo-random variables,

σ^2 = variance of measurement error,

$\hat{Var}(\hat{\sigma}^2)$ = variance estimate of $\hat{\sigma}^2$ when σ^2 is estimated,

$T(\{Y_i\}_{i=1}^{n})$ = a function of the data $\{Y_i\}_{i=1}^{n}$ that is unbiased for the population cumulative distribution from which the data was sampled,

$T_{Var}(\{Y_i\}_{i=1}^{n})$ = a function of the data $\{Y_i\}_{i=1}^{n}$ that is unbiased for $Var\{T(\{Y_i\}_{i=1}^{n})\}$,

$V^{T} = (1, -1, 1)(D^{T}D)^{-1}D^{T}$,

$$D = \begin{bmatrix} 1 & \lambda_1 & \lambda_1^2 \\ \vdots & \vdots & \vdots \\ 1 & \lambda_m & \lambda_m^2 \end{bmatrix}$$

$\Gamma = [\hat{F}_{X,\lambda_1}(t), \hat{F}_{X,\lambda_2}(t), \cdots, \hat{F}_{X,\lambda_m}(t)]^{T}$,

$\eta = [\hat{\tau}^2(\lambda_1) - s_\Delta^2(\lambda_1), \hat{\tau}^2(\lambda_2) - s_\Delta^2(\lambda_2), \cdots, \hat{\tau}^2(\lambda_m) - s_\Delta^2(\lambda_m)]^T$

$\gamma = [\hat{f}_{X,\lambda_1}(t), \hat{f}_{X,\lambda_2}(t), \cdots, \hat{f}_{X,\lambda_m}(t)]^T$

$z_{1-\alpha/2}$ = $100(1-\alpha/2)th$ percentile in the standard Normal distribution.

6 Procedure

6.1 Generate a sequence of k grid points, $t_1 < t_2 < \cdots < t_k$, spanning the range of the observed data $\{X_i\}_{i=1}^n$.

For example, suppose $\min\{X_1, X_2, \cdots, X_n\} = 0$ and $\max\{X_1, X_2, \cdots, X_n\} = 25$. We could let $k = 51$ and define $t_1 = 0,\ t_2 = 0.5,\ t_3 = 1.0,\ t_4 = 1.5,\ \cdots,\ t_{50} = 24.5,\ t_{51} = 25.0$.

6.2 Generate a sequence of values $0 < \lambda_1 < \lambda_2 < \cdots < \lambda_m$.

See Section 8.1 for more information.

6.3 For each grid point t_h, $h = 1, \cdots, k$,

6.3.1 For each λ_j, $j = 1, \cdots, m$,

6.3.1.1 For $b = 1, \cdots, B$,

6.3.1.1.1 Generate n standard normal pseudo-random variates $\{Z_{b,i}^*\}_{i=1}^n$.

6.3.1.1.2 Calculate the pseudo-data set $\{X_{b,i}^*(\lambda_j)\}_{i=1}^n$.

$$X_{b,i}^*(\lambda_j) = X_i + \sigma\sqrt{\lambda_j}Z_{b,i}^* \text{ for } i = 1, \cdots, n$$

6.3.1.1.3 Calculate $\hat{F}_{X,\lambda_j,b}(t_h)$.

$$\hat{F}_{X,\lambda_j,b}(t_h) = T(\{X_{b,i}^*(\lambda_j)\}_{i=1}^n)$$

6.3.1.1.4 Calculate $\hat{\tau}_b^2(\lambda_j)$.

$$\hat{\tau}_b^2(\lambda_j) = T_{Var}(\{X_{b,i}^*(\lambda_j)\}_{i=1}^n)$$

6.3.1.1.5 If $V\hat{a}r(\hat{\sigma}^2) > 0$,

6.3.1.1.5.1 Calculate the data set $\{X_{b,\epsilon,i}^*\}_{i=1}^n$.

$$\{X_{b,\epsilon,i}^*\}_{i=1}^n = X_i + \sqrt{(\sigma^2+\epsilon)\lambda_j}Z_{b,i}^* \text{ for } i = 1, \cdots, n$$

6.3.1.1.5.2 Calculate $\hat{F}_{X,\,\epsilon,\,\lambda_j,\,b}(t_h)$.

$$\hat{F}_{X,\,\epsilon,\,\lambda_j,\,b}(t_h) = T(\{X^{*}_{\epsilon,i}(\lambda_j)\}^{n}_{i=1})$$

6.3.1.2 Calculate $\hat{F}_{X,\,\lambda_j}(t_h)$.

$$\hat{F}_{X,\,\lambda_j}(t_h) = \frac{1}{B}\sum_{b=1}^{B}\hat{F}_{X,\,\lambda_j,\,b}(t_h)$$

6.3.1.3 Calculate $\hat{\tau}^2(\lambda_j)$.

$$\hat{\tau}^2(\lambda_j) = \frac{1}{B}\sum_{b=1}^{B}\hat{\tau}^2_b(\lambda_j)$$

6.3.1.4 Calculate $s^2_{\Delta}(\lambda_j)$

$$s^2_{\Delta}(\lambda_j) = \frac{1}{B-1}\sum_{b=1}^{B}\left(\hat{F}_{X,\,\lambda_j,\,b}(t_h) - \hat{F}_{X,\,\lambda_j}(t_h)\right)^2$$

6.3.1.5 If $\hat{Var}(\hat{\sigma}^2) > 0$,

6.3.1.5.1 Calculate $\hat{F}_{X,\,\epsilon,\,\lambda_j}(t_h)$.

$$\hat{F}_{X,\,\epsilon,\,\lambda_j}(t_h) = \frac{1}{B}\sum_{b=1}^{B}\hat{F}_{X,\,\epsilon,\,\lambda_j,\,b}(t_h)$$

6.3.1.5.2 Calculate $\hat{f}_{X,\,\lambda_j}(t_h)$.

$$\hat{f}_{X,\,\lambda_j}(t_h) = \frac{1}{\epsilon}\left(\hat{F}_{X,\,\epsilon,\,\lambda_j}(t_h) - \hat{F}_{X,\,\lambda_j}(t_h)\right)$$

6.3.2 Calculate $\hat{F}_{SIMEX}(t_h)$.

$$\hat{F}_{SIMEX}(t_h) = V^{T}\Gamma$$

where

$$V^{T} = (1, -1, 1)(D^{T}D)^{-1}D^{T} \text{ such that } D = \begin{bmatrix} 1 & \lambda_1 & \lambda_1^2 \\ \vdots & \vdots & \vdots \\ 1 & \lambda_m & \lambda_m^2 \end{bmatrix}$$

$$\Gamma = [\hat{F}_{X,\lambda_1}(t), \hat{F}_{X,\lambda_2}(t), \cdots, \hat{F}_{X,\lambda_m}(t)]^T$$

6.3.3 Calculate $\hat{Var}(\hat{F}_{SIMEX}(t_h))$.

If σ^2 is known, i.e., $\hat{Var}(\hat{\sigma}^2) = 0$,

$$\hat{Var}(\hat{F}_{SIMEX}(t_h)) = V^T \eta$$

where

V^T is defined above

$$\eta = [\hat{\tau}^2(\lambda_1) - s^2_\Delta(\lambda_1), \hat{\tau}^2(\lambda_2) - s^2_\Delta(\lambda_2), \cdots, \hat{\tau}^2(\lambda_m) - s^2_\Delta(\lambda_m)]^T$$

If σ^2 is estimated, i.e., $\hat{Var}(\hat{\sigma}^2) > 0$,

$$\hat{Var}(\hat{F}_{SIMEX}(t_h)) = V^T \eta + \hat{Var}(\hat{\sigma}^2) \cdot \{V^T \gamma\}^2$$

where

V^T and η are defined above

$$\gamma = [\hat{f}_{X,\lambda_1}(t_h), \hat{f}_{X,\lambda_2}(t_h), \cdots, \hat{f}_{X,\lambda_m}(t_h)]^T$$

6.3.4 Calculate approximate 100(1 - α)% confidence limits, $\hat{L}(t_h)$ and $\hat{U}(t_h)$.

$$\hat{L}(t_h) = \hat{F}_{SIMEX}(t_h) - z_{1-\alpha/2} \cdot \sqrt{\hat{Var}(\hat{F}_{SIMEX}(t_h))}$$

$$\hat{U}(t_h) = \hat{F}_{SIMEX}(t_h) + z_{1-\alpha/2} \cdot \sqrt{\hat{Var}(\hat{F}_{SIMEX}(t_h))}$$

where $z_{1-\alpha/2}$ is the 100(1 - α/2)*th* percentile in the standard Normal distribution.

6.4 Apply isotonic regression to $\{\hat{F}_{SIMEX}(t_1), \cdots, \hat{F}_{SIMEX}(t_k)\}$ on $\{t_1, \cdots, t_k\}$

While$_1$ $\{\hat{F}_{SIMEX}(t_h)\}_{h=1}^{k}$ is NOT non-decreasing

Let $i = 1$ and $j = 2$

While$_2$ $j < n$

While$_3$ $\hat{F}_{SIMEX}(t_{j-1}) > \hat{F}_{SIMEX}(t_j)$

Let $j = j + 1$

End of While$_3$

For $h = 1, \cdots, j - 1$,

$$\hat{F}_{SIMEX}(t_h) = \frac{1}{j - i} \sum_{q = i}^{j - 1} \hat{F}_{SIMEX}(t_q)$$

End of For

Let $i = j$ and $j = j + 1$

End of While$_2$

End of While$_1$

6.5 Restrict range of $\{\hat{F}_{SIMEX}(t_h)\}_{h=1}^{k}$ to [0, 1].

Set $h = 1$

While $\hat{F}_{SIMEX}(t_h) < 0$

$\hat{F}_{SIMEX}(t_h) = 0$

$h = h + 1$

End of While

Set $h = k$

While $\hat{F}_{SIMEX}(t_h) > 1$

$\hat{F}_{SIMEX}(t_h) = 1$

$h = h - 1$

End of While

(Note, isotonic regression simply ensures that $\hat{F}_{SIMEX}$ is a non-decreasing function on $[t_1, t_k]$. This is accomplished by averaging function values on intervals where $\hat{F}_{SIMEX}$ is decreasing. The procedure is repeated until $\hat{F}_{SIMEX}$ is non-decreasing. For more information on isotonic regression see Barlow et al (1972).

6.6 Apply isotonic regression to $\{\hat{L}(t_1), \cdots, \hat{L}(t_k)\}$ on $\{t_1, ,\cdots, t_k\}$ and restrict range of $\{\hat{U}(t_h)\}_{h=1}^{k}$ to [0, 1].

See Sections 6.4 and 6.5 above.

6.7 Apply isotonic regression to $\{\hat{U}(t_1), \cdots, \hat{U}(t_k)\}$ on $\{t_1, \cdots, t_k\}$ and restrict range of $\{\hat{L}(t_h)\}_{h=1}^{k}$ to [0, 1].

See Sections 6.4 and 6.5 above.

7 Associated Methods

A similar procedure of estimating the cumulative distribution of a finite population in the presence of measurement error is described in Estimation Method 9: The Parametric Jackknife Estimator. This method assumes a particular sampling model that allows the expectation of the sample cumulative distributions to be obtained analytically, rather than by simulation as in the SIMEX method.

8 Notes

The procedure outlined in Section 6 requires specification of $0 < \lambda_1 < \cdots < \lambda_m$. Cook and Stefanski (1994) propose taking equally spaced values over the interval [0.05, 2.00]. They also suggest using $m \geq 5$, although the exact number of values is not critical.

The algorithm in Section 6 is designed for calculating estimates of the cumulative proportion. A slight variation of this algorithm would allow for estimating the cumulative total. In this case we assume that $\hat{F}_U(t) = T(U)$ is an unbiased estimator of the cumulative total. The algorithm is modified by changing the upper bound of the SIMEX estimate and the confidence limits from one to the population size, if the population is finite, or ∞, if the population is infinite. This modification is required in Sections 6.5 through 6.7.

9 References

Barlow, R. E., Bartholamu, D. J., Brenner, J. M., and Brunk, H. D.(1972), *Statistical Inference under Order Restrictions*, New York: John Wiley & Sons.

Cook, J. R. and Stefanski, L. A. (1994), Simulation-Extrapolation Estimation in Parametric Measurement Error Models, Journal of the American Statistical Association, 89, 1314-1328.