Errors-in-variables problems in statistical analysis of random objects

Kyunghee Han (Supervisor) Max Nguyen Tom Hammons Xinyu Qin Jiuhang Qin

University of Illinois at Chicago

MSCS Undergraduate Research Laboratory

Motivation

When we have n random samples $(W_1,Y_1),\ldots,(W_n,Y_n)$ where W_i's are the independent variables or predictors and Y_i's are the dependent variables or response then as counter intuitive as it may be, but the measurement errors of the predictor variables can induces bias in the parameter estimates. So in the most optimal case, we would want all of W_i to be error free, but that may not be the case and one of the most general method for measurement error bias-correction method in the sense that the bias due to measurement error in almost any estimator of almost any parameter is readily estimated and (approximately) corrected, is the SIMEX or Simulation and Extrapolation method and we are going to check its nature in the context of (global) Fréchet regression.

Simulation and Extrapolation (SIMEX)

For simplicity and flexibility but this works more generally, given the random samples $(W_1,Y_1),\ldots,(W_n,Y_n)$ where predictors are subjected to additive measurement errors, that is $W_i = X_i + U_i$ where X_i is the true error free predictors and that $U_i \overset{\text{iid}}{\sim} \mathcal{N}(0,\sigma_u^2)$, the SIMEX algorithm is applied as follows:

- With any strictly increasing sequence of real numbers $1 = \zeta_1 < \zeta_2 < \cdots < \zeta_M$ then we simulate by adding additional independent measurement errors with variance $\zeta_m \sigma_u^2$ are generated and added to the original predictor data thereby creating data sets with successively larger measurement error variances. For the m-th data set, the total measurement error variance is $(1+\zeta_m)\sigma_u^2$. For our assumption of measurement error model, the i-th predictor of the m-th data sets is then inductively taken on the form $W_{m,i}(\zeta_m) = W_i + \sqrt{\zeta_m}\, U_{m,i}$ where $U_{m,i}$ are computer generated *pseudo errors* in which $U_{m,i} \overset{\text{iid}}{\sim} \mathcal{N}(0,\sigma_u^2)$.
- Next, obtain estimates from the generated datasets.
- Repeat the simulation and estimation steps for all the ζ_m values.
- Extrapolate using regression techniques to the ideal case of no measurement error, that is, the case where the variance vanishes or $\zeta = -1$.

Why SIMEX?

"The key idea underlying SIMEX is the fact that the effect of measurement error on an estimator can be determined experimentally via simulation."[2]

Global Fréchet Regression

At the theoretical level, given a metric space (Ω, d) with a random process $(X,Y) \sim F$ in which X taking values in $\mathbb{R}^m$ along with Y has values in Ω and F is their joint distribution then assume that we also know their means, (co)variances and conditional distributions, the notion of mean and variances for Y is defined in [1]. Given that $X = x$, the weight function at x is the map $s(\cdot,x) : \mathbb{R}^m \to \mathbb{R}$ such that

$$s(z,x) = 1 + (z-\mu)^T \Sigma^{-1}(x-\mu)$$

where μ and Σ are the mean and variance of X, respectively then the conditional Fréchet regression at $X = x$ is

$$M(\cdot,x) = E(s(X,x)d^2(Y,\cdot))$$

which leads to the global Fréchet regression model

$$m_\oplus(x) = \operatorname{argmin}_{\omega\in\Omega} M(\omega,x)$$

In practice, given independent random samples $(X_i,Y_i) \sim F$ with $i = 1,\ldots,n$ where X_i taking values in $\mathbb{R}^m$ and Y_i has values in Ω again, then there are empirical values $\overline{X} = n^{-1}\sum_{i=1}^n X_i$ for μ and $\hat{\Sigma} = n^{-1}\sum_{i=1}^n (X_i - \overline{X})(X_i - \overline{X})^T$ for Σ which leads to the empirical weight

$$s_{in}(x) = 1 + (X_i - \overline{X})^T \hat{\Sigma}^{-1}(x - \overline{X})$$

and subsequently the empirical estimates for the conditional regression

$$M_n(\cdot,x) = n^{-1}\sum_{i=1}^n s_{in}(x) d^2(Y,\cdot)$$

which means the Fréchet regression model become

$$\hat{m}_\oplus(x) = \operatorname{argmin}_{\omega\in\Omega} M_n(\omega,x)$$

Wasserstein space and its Fréchet regression model

Our metric space of choice is the 2-Wasserstein space or the space with the underlying set Ω of probability distributions on $\mathbb{R}$ with the (2-)Wasserstein metric d_W given by

$$d_W(G_1,G_2) = \sqrt{\int_0^1 \left(G_1^{-1}(t) - G_2^{-1}(t)\right)^2 dt}$$

where G^{-1} is the quantile function of the distribution G. Given that $X = x$, upon taking Fréchet derivative, the global Fréchet regression model for this space then becomes

$$\hat{m}_\oplus(x) = \hat{Q}_x^{-1}$$

in which with the empirical weights $s_{in}(x)$ of X_i,

$$\hat{Q}_x(t) = \frac{\sum_{i=1}^n s_{in}(x) Y_i^{-1}(t)}{\sum_{i=1}^n s_{in}(x)}$$

In other words, this is just the distribution with quantiles being weighted average of the quantiles of the Y_i's

Testing the validity of SIMEX

Note that for the Fréchet regression model for the (2-)Wasserstein space is entirely determined by the weight values (and quantiles), to show that SIMEX was able to recover the weight values on the quantile distribution of the contaminated predictors to the true predictor values.

We will test this by using real data, namely the grade distribution for all 13 UIC's Calculus III sessions in Fall 2022. We will take the number of people still stay in class after the add drop period as our predictor. Assuming that these numbers are error free and that distributions are continuous, we will add extra measurement errors to this and running the SIMEX algorithm and compare the induced quantile values.

The graphs for a few parameters tested are given on the other side.

Conclusion and difficulties

From the graphs, it suggested that SIMEX was able to approximate the true values. Although there are a few difficulties confirming its validity namely

- Our data sets is small so continuity approximation might is not optimal.
- Since the data sets is small, it is hard to perturb to large values without statistical problems, namely that normal distributions are distributed across the whole real line so for large value perturbation might gives negative values for data sets.

What can be improved?

Facing all of these challenges, there are a few things to check to improve our findings:

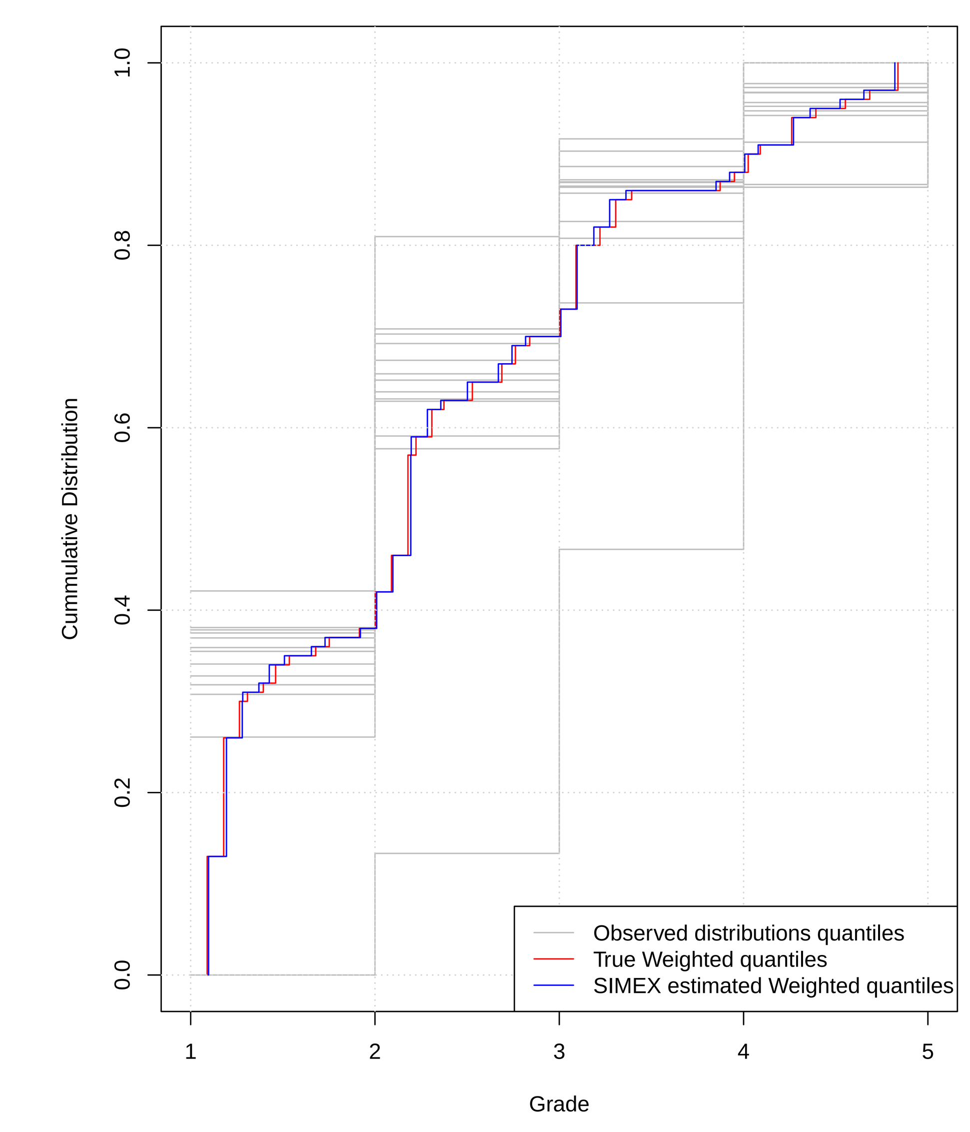
- Obtain and use a larger data sets or generate our own data.
- Trying a wider range of error distributions.
- Find better predictor values.

References

[1] Alexander Petersen, Hans-Georg Muller, *Fréchet regression for random objects with Euclidean predictors*, The Annals of Statistics, 2019 .

[2] Raymond J. Carroll et.al, *Measurement Error in Nonlinear Models. A Modern Perspective*, 2006 .

Weighted quantiles given that $X = 32$ and errors are iid $\mathcal{N}(0,10^2)$



Weighted quantiles given that $X = 46$ and errors are iid $\mathcal{N}(0,4.5^2)$