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

Motivation
When we have \( n \) random samples \((W_1, Y_1), \ldots, (W_n, Y_n)\) where \( W_i \) are the independent variables or predictors and \( Y_i \) 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 \( U_i \sim \mathcal{N}(0, \sigma^2) \), the SIMEX algorithm is applied as follows:

- With any strictly increasing sequence of real numbers \( 1 = \xi_1 < \xi_2 < \cdots \xi_T \) then we simulate by adding additional independent measurement errors with variance \( \xi_T \sigma^2 \) 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 \( (1 + \xi_m) \sigma^2 \). For our assumption of measurement error model, the \( i \)-th predictor of the \( m \)-th data sets is then inductively taken on the form \( W_{m,i} = X_i + U_{m,i} \) where \( U_{m,i} \) are computer generated pseudo errors in which \( U_{m,i} \sim \mathcal{N}(0, \xi_m \sigma^2) \).
- Next, obtain estimates from the generated datasets.
- Repeat the simulation and estimation steps for all the \( \xi_m \) values.
- Extrapolate using regression techniques to the ideal case of no measurement error, that is, the case where the variance vanishes or \( \xi = 1 \).

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 \((\Omega, \rho)\) with a random process \((X, Y)\sim \mathcal{F}\) in which \(X\) taking values in \( \mathbb{R}^n\) and \(Y\) 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

\[ w(x, \cdot) = \int_\Omega \|x - y\| \rho(dx, dy) \]

for simplicity and flexibility but this works more generally, given independent random samples \((X_i, Y_i)\sim \mathcal{F}\) with \(i = 1, \ldots, n\) where \(Y_i\) has values in \(\Omega\) and \(X_i\) has values in \(\mathbb{R}^n\) again, then there are empirical values \(\bar{X}_i\) and \(\bar{Y}_i\) for \(\Sigma_1 \bar{X}_i \Sigma_2^{-1} \Sigma_1\) which leads to the empirical weight function

\[ w(x, \cdot) = \sum_{i=1}^n \rho(\bar{X}_i - x, \Sigma_1^{-1} \bar{Y}_i - y) \]

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.

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 that our data sets is small so continuity approximation might not be optimal.

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