SIMEX, and other methods for the adjustment of measurement error

Implications of ME

- It leads to a loss of power
- It bias the estimates of the regression coefficients

e.g. simple linear regression with classical additive ME in the predictor.

$$X^* = X + U$$

$$Y = \beta_0 + \beta_1 X^* + \epsilon$$

$$\hat{\beta}_1^* = \beta_1 \left(\frac{\sigma_X^2}{\sigma_X^2 + \sigma_U^2} \right)$$

- Often size and direction of the bias are unpredictable

"*Measurement error is, to borrow a metaphor, a gremlin hiding in the details of our research that can contaminate the entire set of estimated regression parameters*" (Nugent, et al. 2000, p.60).

Methods for the Adjustment of ME

- Methods that require additional variables:

  Latent variable estimation

  Instrumental variables

- Methods that require at least a subsample of replicated or validation data:

  Regression calibration

- Methods that only require knowledge (assumptions) about the behavior of the ME:

  Maximum likelihood based methods

  Simulation-extrapolation (SIMEX)

Methods that Require Additional Variables

- Latent Variable Models

These methods use a set of variables that are thought to be formed by a latent trait that they share in common and an idiosyncratic error term.

The true variable is estimated exploiting the common variance of the set of variables used in the model.

The main problem the availability of additional variables reflecting the common trait.

Also, something worth considering, the latent variable is an estimate and therefore it is also prone to measurement error too.

- Instrumental Variables

IV can be applied to measurement error problems by: 1) regressing the observed variable on the instrument; 2) the estimates of that regression are used to impute a new variable; 3) that new variable is replaced by the observed variable in the outcome model.

The main problem is the difficulty of finding an IV.

In addition it is often impossible to check the assumptions that the instrument is based on.

Methods that Require Replicated or Validation Data

- Regression Calibration

This method postulates the use of the best approximation of the unobserved variable, given the observed information available.

Implementation: 1) the true data is defined from replicated or validation data; 2) a calibration model is specified where the true variable is regressed on the observed variable and the rest of covariates from the outcome model; 3) like in IV the model estimates are used to impute a new variable; 4) this new variable is used as an approximation of the true variable.

Again the problem is that replicated or validation data is often unavailable.

In addition, the effectiveness of the method depends on how well the calibration function is estimated, and it has been proven to be inadequate in highly non-linear models.

Methods that Require Distributional Assumptions

- Maximum Likelihood-Based Methods

The most flexible solution: It can be used in the context of many different outcome and measurement models and it can operate without additional data.

Usually the likelihood function is formed by different building blocks

$$f(X^*,Y,Z) = \int f(X^*|Y,X,Z)f(Y|X,Z)f(X|Z)dx$$

Problems: to integrate the likelihood function over the unobserved variable sometimes require iterative methods, it requires specification of an exposure model, it is sensitive to misspecifications, often identifiability is not achieved.

The Bayesian approach can help to achieve identifiability, thanks to the use of priors, which allow learning from the past.

The change of statistical paradigm can be impractical

Methods that Require Distributional Assumptions

- Simulation-Extrapolation (SIMEX)

Probably the simplest solution, regardless of how complex the outcome model is.

No additional data is required.

No need to specify the true variable.

Disadvantages:

It requires good knowledge of the reliability of the observed variable.

It is computationally intensive.

So far only developed for simple measurement models.

Only capable of making approximate adjustments: The quality of the adjustment depends on the precision of the estimate of the measurement error variance and on the choice of extrapolation function to be used.

The Logic of SIMEX

"The key idea underlying SIMEX is the fact that the effect of measurement error on an estimator can be determined experimentally via simulation" (Carroll, 2006, p. 98).

First, in the simulation step, additional datasets are generated by simulating independent measurement error terms with variance σ_u^2. These variances are multiplied by a positive and increasing factor $(1+\lambda_t)$ and added to the observed variable *X**.

$$X_{ti}^* = X_i^* + (1+\lambda_t)\sigma_u^2$$

Second, the outcome model is estimated using each of the new datasets that were simulated and their biased estimates, $\hat{\beta}_{1t}^*$, are saved.

Steps 1 and 2 are repeated a large number of times and the estimates for each level of lambda are averaged.

At this stage we can pair the $\hat{\beta}_{1t}^*$ and the λ_t, and consider the former as a function of the latter

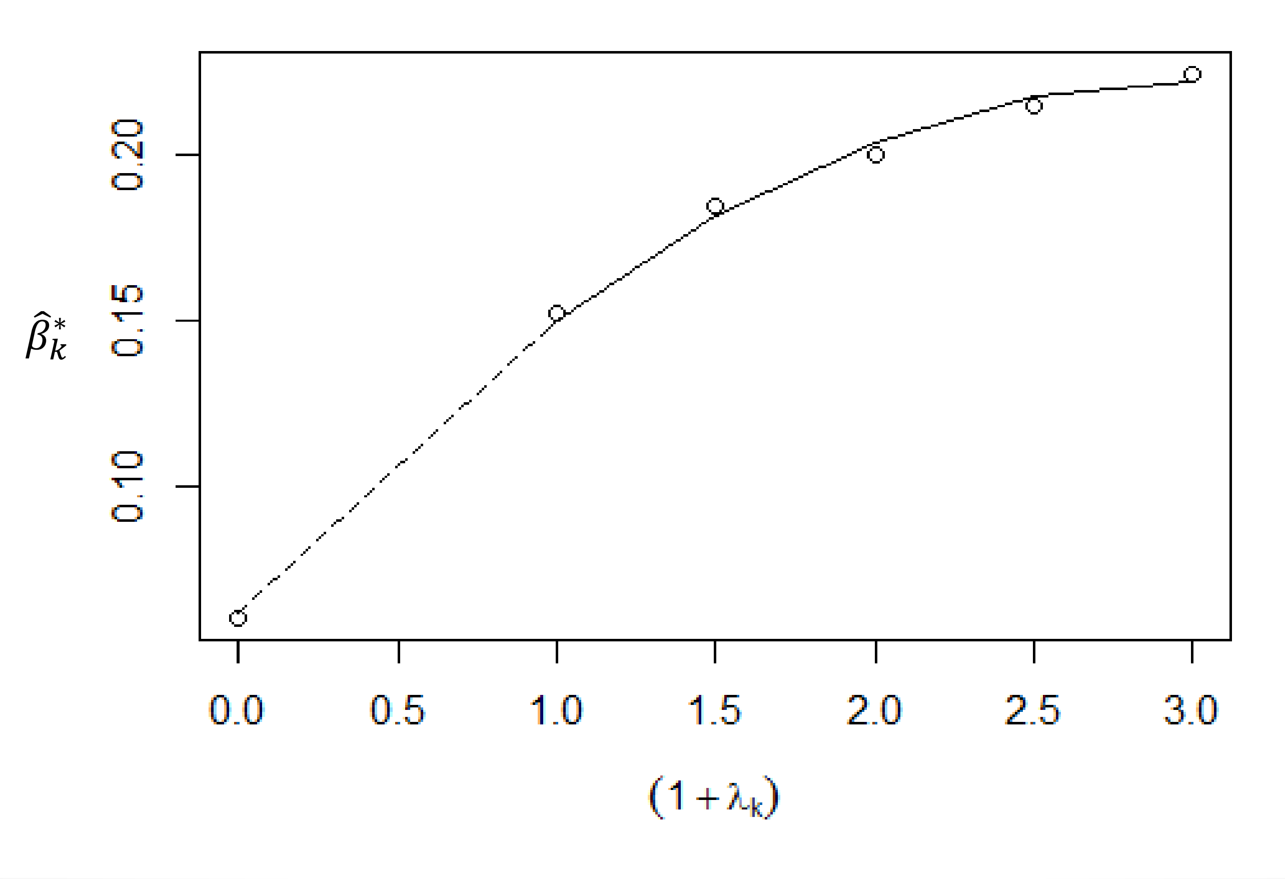
$$G(\lambda_t) = \hat{\beta}_{1t}^* = \beta_1\sigma_x^2/(\sigma_x^2 + (1+\lambda_t)\sigma_u^2),$$

The third step will be to estimate this function, for that the $\hat{\beta}_{1t}^*$ are regressed on λ_t ;

$$\hat{\beta}_{1t}^* = \gamma_1 + \gamma_2\lambda_t + \gamma_3{\lambda_t}^2 + \epsilon$$

The extrapolation step: $\lambda_t$is replaced by -1 in the above model to obtain the SIMEX estimate.

The SIMEX Algorithm



Standard Errors Adjustment in SIMEX

- Bootstrap and Jackknife methods.
- The measurement error Jackknife (Stefanski and Cook, 1995).
- Asymptotic covariance estimation methods (Carroll et al., 1996).

Possible Extensions

“SIMEX is ideally suited to problems with additive measurement error, and more generally to any problem in which the measurement error generating process can be imitated on a computer via Monte Carlo methods” (Carroll, p.97, 2006).

- SIMEX for classical multiplicative errors

$$X^* = X \cdot U$$

$$X_t^* = exp\{log(X^*) + \sqrt{\lambda_t} log(u)\}$$

- Misclassification-SIMEX

$$\theta_{X^*|X} = P(X^* = x^* | X = x)$$

$$\Theta = \begin{pmatrix} \theta_{1|1} & \theta_{1|0} \\ \theta_{0|1} & \theta_{0|0} \end{pmatrix}$$

$$X^*(\lambda_t) = \Theta^{\lambda}\ applied\ to\ X$$

- SIMEX for inverse-Gamma distributed Berkson errors.

SIMEX for Duration Data

- Classical additive measurement error in the response also generates a bias in the coefficients of duration models.
- First applications using simulated errors seem robust.
- Help! Has any of you used duration data that you suspect is prone to measurement error?

Ideally a situation where all observations start from the same state (no misclassification) and the measurement error affects the duration randomly.

Example: Study of drug addicts reporting the time it took them to relapse after they have been rehabilitated.

Conclusion

- SIMEX only produces partial adjustments and is computationally intensive, but it is widely applicable, robust, and simple to implement.

Doesn't matter how complex your outcome model is.

No additional data is required; an estimate of the reliability of the error-prone variable suffices. If we are unaware of that reliability, at least sensitivity analyses can be run.

Originally developed for cases of classical additive measurement error, but with the potential to be extended to other types of measurement error.

If applied with certain care it does not make things worse.

No need to do any programming or complex modelling: SIMEX package available in STATA and R.

- If a ratio practicality/effectiveness existed, SIMEX would surely be the best of the methods to account for measurement error.