Simultaneous treatment of unspecified heteroskedastic model error distribution and mismeasured covariates for restricted moment models

Tanya P. Garcia and Yanyuan Ma

Texas A&M University, Pennsylvania State University

August 2016

Motivating Problem

Restricted Moment Model with Measurement Error

$$Y_i = m(X_i, Z_i; \beta) + \epsilon_i, \quad E(\epsilon_i | X_i, Z_i) = 0$$
$$W_{ij} = X_i + U_{ij}, \quad j = 1, \ldots, \ell, \quad i = 1, \ldots, n.$$

1. $m(X, Z; \beta)$: linear or nonlinear.
2. Model error ϵ may depend on (X, Z) (heteroskedasticity) and its conditional distribution $p_{\epsilon|X,Z}$ is unspecified.
3. $p_{X|Z}, p_Z$ are unspecified.
4. Classical measurement error; $p_U(u; \Omega_U)$ is a general parametric distribution with Ω_U unknown. [Can be relaxed.]

Estimation challenges

1. Bias if we arbitrarily adopt models for $p_{\epsilon|X,Z}$ or $p_{X|Z}$.
2. Estimating $p_{X|Z}$ is challenging.
   - Need inverse operation such as deconvolution (Stefanski and Carroll, 1990).
   - This has a very slow rate (Carroll and Hall, 1988; Fan, 1991).

Estimation challenges

3. Estimating $p_{\epsilon|X,Z}$ is even more difficult:
   - Residuals unobtainable in measurement error models even if model parameters known.
   - Unavailable residuals makes it difficult to correctly estimate the model error's variance-covariance.
   - This is particularly problematic with heteroskedastic model error which needs a proper variance-covariance model for consistent parameter estimation.
   - Methods exist to estimate unknown variance-covariance, but they are approximate (Carroll and Wang, 2008) or involved (Delaigle and Hall, 2011).

Approach and its Features

We use a semiparametric approach and view unknown distributions $p_{\epsilon|X,Z}$ and $p_{X|Z}$ as nuisance parameters.

1. We bypass estimation of $p_{\epsilon|X,Z}$ and $p_{X|Z}$.
2. We handle model error heteroskedasticity.
3. We handle mismeasured covariates X.

Our Approach

1. Use components of variance analysis (Carroll et al., 2006) to solve for $\widehat{\Omega}_U$ in $p_U(u;\Omega_U)$ based on replicates $W_1,\ldots,W_\ell$.
2. Propose working density models η_1^* for $\eta_1 \equiv p_{X|Z}$, and $\eta_2^* = p_{\epsilon|X,Z}$ for η_2 such that $E_*(\epsilon|X,Z) = 0$.
3. Compute $f(Y,W,Z)$ such that

$$\int f(Y,W,Z)p_{W|X,Z}(W|X,Z;\Omega) = g(X,Z)\epsilon.$$

   Note: f is a function of $\widehat{\Omega}_U, \eta_1^*, \eta_2^*$.

4. Solve for $\widehat{\beta}$ from $\sum_{i=1}^n f(Y_i,W_i,Z_i;\beta,\widehat{\Omega}_U,\eta_1^*,\eta_2^*) = 0$.

Some Computational Issues

We provide one way to find $f(Y, W, Z)$. It involves solving an ill-posed problem.

But it is a "good" ill-posed problem: there is more than one solution and we only need to find one.

Problem is solved using numerical approximation which does not affect efficiency.

Selection and Impact of Working Models

Property 1. When either or both η_1^*, η_2^* are misspecified and the measurement error distribution is estimated as $p_{W|X,Z}(w|x,z;\widehat{\Omega}_U)$, the algorithm provides a consistent estimator.

- Under regularity conditions, we show estimating equation is unbiased even when η_1^*, η_2^* are misspecified.
- We are thus free to choose any working model; simple choice is Gaussian.

Selection and Impact of Working Models

Property 2. Choice of working models η_1^*, η_2^* affects efficiency.

- When $\eta_1^* = \eta_{10}$, $\eta_2^* = \eta_{20}$ [truth], we obtain optimal estimator (i.e., semiparametric efficiency bound obtained).
- Otherwise, efficiency loss is positive definite and must be evaluated on a case-by-case basis.
- From our limited empirical studies, the efficiency loss is generally small. The estimation variance is quite insensitive to the choice of the working models.

Theoretical Results

Theorem 1. Estimator from proposed method is asymptotically normal.

Theorem 2. Asymptotic efficiency of estimator does not depend on how efficiently we estimate parameters in working parametric models.

Empirical Performance of the Method

Nonlinear restricted moment model with measurement error.

$$\begin{aligned} Y_i &= \beta_2 \exp(-\beta_1 X_i^2) + \beta_3 Z_i + \epsilon_i, \\ W_{ij} &= X_i + U_{ij}, \quad U_{ij} \sim \text{Normal}(0, 2\alpha), \quad j = 1, 2. \end{aligned}$$

Variance of model error ϵ considered to be heteroskedastic and homoskedastic.

Methods Evaluated

1. Our method: Set working models η_1^*, η_2^* very different from the truth. Working model η_2^* does not account for possible heteroskedasticity.
2. Homoskedastic and Heteroskedastic sieve estimator (Hu and Schennach, 2008): Idea is to represent $\eta_1 = p_{X|Z}$, $\eta_2 = p_{\epsilon|X,Z}$ with increasingly rich parametric representations (i.e., truncated series of basis functions).
   - Homoskedastic sieve for η_2: ignores dependence between ϵ and (X, Z).
   - Heteroskedastic sieve for η_2: captures dependence between ϵ and (X, Z).

Methods Evaluated

3. Homoskedastic and Heteroskedastic Tsiatis-Ma estimator: Uses a working model $\eta_1^* = p_{X|Z}$, but requires $\eta_2^* = p_{\epsilon|X,Z}$ to be correctly specified, particularly in variance structure.
4. Naive estimator: Least squares estimator that ignores measurement error.

Simulation Results

	η_{20} ~Uniform, Homoskedastic			
	$\widehat{\beta}_1$	$\widehat{\beta}_2$	$\widehat{\beta}_3$	$\widehat{\sigma}^2_U$
		Semipar		
bias	-0.0065	-0.0080	0.0011	5.1664×10^{-5}
var	0.0030	0.0031	0.0026	1.0255×10^{-5}
$\widehat{\text{var}}$	0.0030	0.0030	0.0027	1.0255×10^{-5}
CI	0.9500	0.9390	0.9520	0.9490
		Sieve-Hom*		
bias	0.0066	0.0008	0.0021	5.1664×10^{-5}
var	0.0033	0.0030	0.0022	1.0255×10^{-5}
$\widehat{\text{var}}$	NA	NA	NA	NA
CI	NA	NA	NA	NA
		Sieve-Het*		
bias	0.5022	0.8177	0.6900	9.6823×10^{-6}
var	0.0450	0.0458	0.0795	1.0916×10^{-5}
$\widehat{\text{var}}$	NA	NA	NA	NA
CI	NA	NA	NA	NA

Simulation Results

	η_{20} ~Uniform, Homoskedastic			
	$\widehat{\beta}_1$	$\widehat{\beta}_2$	$\widehat{\beta}_3$	$\widehat{\sigma}^2_U$
		Semipar		
bias	-0.0065	-0.0080	0.0011	5.1664×10^{-5}
var	0.0030	0.0031	0.0026	1.0255×10^{-5}
$\widehat{\text{var}}$	0.0030	0.0030	0.0027	1.0255×10^{-5}
CI	0.9500	0.9390	0.9520	0.9490
		TM-Hom		
bias	0.0019	0.0019	-0.0000	-0.0002
var	0.0035	0.0033	0.0027	1.0396×10^{-5}
$\widehat{\text{var}}$	0.0035	0.0032	0.0027	9.8539×10^{-6}
CI	0.9460	0.9440	0.9470	0.9490
		TM-Het		
bias	-0.0144	-0.0185	0.0001	-0.0002
var	0.0038	0.0034	0.0032	1.0396×10^{-5}
$\widehat{\text{var}}$	0.0037	0.0032	0.0031	9.8539×10^{-6}
CI	0.9210	0.9300	0.9480	0.9490

Evaluation of Efficiency Loss from Proposed Method

Setting		$\widehat{\beta}_1$	$\widehat{\beta}_2$	$\widehat{\beta}_3$	$\widehat{\sigma}^2_U$
	bias	-0.0013	0.0009	-0.0017	-3.9892×10^{-5}
$\eta_1^* = \eta_{10}, \eta_2^* = \eta_{20}$	$\widehat{\text{var}}$	0.0044	0.0010	0.0013	9.9257×10^{-6}
	CI	0.9500	0.9500	0.9430	0.9430
	bias	-0.0001	0.0012	-0.0017	-3.9892×10^{-5}
$\eta_1^* \neq \eta_{10}, \eta_2^* = \eta_{20}$	$\widehat{\text{var}}$	0.0047	0.0010	0.0013	9.9257×10^{-6}
	CI	0.9480	0.9500	0.9430	0.9430
	bias	-0.0104	0.0007	-0.0063	-3.9892×10^{-5}
$\eta_1^* = \eta_{10}, \eta_2^* \neq \eta_{20}$	$\widehat{\text{var}}$	0.0052	0.0012	0.0018	9.9257×10^{-6}
	CI	0.9380	0.9490	0.9410	0.9430
	bias	-0.0081	0.0011	-0.0064	-3.9892×10^{-5}
$\eta_1^* \neq \eta_{10}, \eta_2^* \neq \eta_{20}$	$\widehat{\text{var}}$	0.0065	0.0013	0.0019	9.9257×10^{-6}
	CI	0.9410	0.9470	0.9430	0.9430

Acknowledgments

- We thank Yingyao Hu for providing code for the homoskedastic sieve estimator.
- Funding provided by Huntington's Disease Society of America, National Science Foundation, and NIH National Institute of Neurological Disorders and Stroke.