

**On latent-variable model misspecification in structural
measurement error models for binary response**

Joshua M. Tebbs*
Associate Professor
Department of Statistics
University of South Carolina

October 14, 2009

*This is collaborative research with X. Huang (Department of Statistics, [University of South Carolina](#)). This work is funded in part by the [National Institutes of Health](#) (R01-AI067373).

Huang, X. and Tebbs, J. (2009). On latent-variable model misspecification in structural measurement error models for binary response. *Biometrics*, **65**, 710-718.

OUTLINE

1. Framingham Heart Study
2. Structural measurement error models (SMEM) for binary response
3. SMEM models for pooled binary response
4. Diagnosing latent variable model misspecification
5. Analyze Framingham data
6. Discussion

1. Framingham Heart Study

- **Data set:** 1615 male subjects followed over time for the development of coronary heart disease
- **Goal:** Characterize the relationship between the risk of coronary heart disease and long-term systolic blood pressure
- **Binary response:** $Y_i = 1$, if evidence of heart disease is detected in the i th subject; $Y_i = 0$, otherwise (8-year follow up period)
- The true predictor, **long-term systolic blood pressure**, can not be observed
- Systolic blood pressure readings are collected at the 2nd and 3rd clinic visits for each subject
 - These can be viewed as **error-contaminated** versions of the true predictor

2. SMEM for individual binary response

- For the i th individual ($i = 1, 2, \dots, n$), let
 - Y_i = the binary response
 - W_i = the observed predictor value
 - X_i = the true predictor value
- We assume throughout that the n individuals are independent
- A structural measurement error model consists of three component models
- **First component:** A generalized linear model for Y_i , conditional on X_i ,

$$\text{pr}(Y_i = 1|X_i) = h(\beta_0 + \beta_1 X_i),$$

where $h(\cdot)$ is a known inverse link function

- Inference on the regression parameters $\theta = (\beta_0, \beta_1)^T$ is of central interest

2. SMEM for individual binary response

- **Second component:** The classical measurement error model

$$W_i = X_i + U_i,$$

where U_i is the **nondifferential** measurement error and $U_i \sim N(0, \sigma_U^2)$

- It follows that $W_i | X_i \sim N(X_i, \sigma_U^2)$
- Nondifferentiability implies that given X_i , Y_i is **independent** of W_i

- **Third component:** The **assumed** latent variable model

$$X \sim f_X^{(a)}(x; \boldsymbol{\tau}),$$

where $\boldsymbol{\tau}$ is a parameter vector of length t

2. SMEM for individual binary response

- The joint density of the **observed** datum, (Y_i, W_i) , is given by

$$f_{Y, W}(Y_i, W_i; \boldsymbol{\theta}, \boldsymbol{\tau}, \sigma_U^2) = \int f_{Y|X}(Y_i|x; \boldsymbol{\theta}) f_{W|X}(W_i|x; \sigma_U^2) f_X^{(a)}(x; \boldsymbol{\tau}) dx$$

- $f_{Y|X}(Y_i|x; \boldsymbol{\theta}) = h(\beta_0 + \beta_1 x)^{Y_i} \{1 - h(\beta_0 + \beta_1 x)\}^{1-Y_i}$
- $f_{W|X}(W_i|x; \sigma_U^2) = \sigma_U^{-1} \phi\{\sigma_U^{-1}(W_i - x)\}$
- $\phi(\cdot)$ denotes the $N(0, 1)$ density

- The **loglikelihood** of the observed data is

$$l(\boldsymbol{\theta}, \boldsymbol{\tau}, \sigma_U^2) = \sum_{i=1}^n \log \left\{ \int f_{Y|X}(Y_i|x; \boldsymbol{\theta}) f_{W|X}(W_i|x; \sigma_U^2) f_X^{(a)}(x; \boldsymbol{\tau}) dx \right\}$$

The choice of $f_X^{(a)}(x; \boldsymbol{\tau})$ can **affect** inference for $\boldsymbol{\theta}$!!

3. SMEM for pooled binary response

- Suppose that G **pools** are formed from the n individuals, and let n_g denote the pool size for pool g , $g = 1, 2, \dots, G$, so that $\sum_{g=1}^G n_g = n$
 - **Pooled binary response** $\rightsquigarrow Y_g^* = \max\{Y_{g1}, Y_{g2}, \dots, Y_{gn_g}\}$
 - **Observed predictor** $\rightsquigarrow \mathbf{W}_g^* = (W_{g1}, W_{g2}, \dots, W_{gn_g})^T$
 - **Latent predictor** $\rightsquigarrow \mathbf{X}_g^* = (X_{g1}, X_{g2}, \dots, X_{gn_g})^T$
- Assume as before

$$W_{gj} = X_{gj} + U_{gj},$$

where $U_{gj} \sim N(0, \sigma_U^2)$, for all g and j

- The **major difference** in the SMEM for pooled response is that now

$$\text{pr}(Y_g^* = 1 | \mathbf{X}_g^*) = 1 - \prod_{j=1}^{n_g} \{1 - h(\beta_0 + \beta_1 X_{gj})\}$$

3. SMEM for pooled binary response

- Under this **induced** primary regression model, we can derive the joint distribution of (Y_g^*, \mathbf{W}_g^*)
- This distribution is denoted by $f_{Y^*, \mathbf{W}^*}(Y_g^*, \mathbf{W}_g^*; \boldsymbol{\theta}, \boldsymbol{\tau}, \sigma_U^2)$, and is given by

$$Y_g^* \prod_{j=1}^{n_g} \int \frac{1}{\sigma_U} \phi\left(\frac{W_{gj} - x}{\sigma_U}\right) f_X^{(a)}(x; \boldsymbol{\tau}) dx$$

$$+ (1 - 2Y_g^*) \prod_{j=1}^{n_g} \int \{1 - h(\beta_0 + \beta_1 x)\} \frac{1}{\sigma_U} \phi\left(\frac{W_{gj} - x}{\sigma_U}\right) f_X^{(a)}(x; \boldsymbol{\tau}) dx$$

- Assuming that the G pools are **independent**, the loglikelihood of the observed data (based on the pooled responses) is

$$l^*(\boldsymbol{\theta}, \boldsymbol{\tau}, \sigma_U^2) = \sum_{g=1}^G \log\{f_{Y^*, \mathbf{W}^*}(Y_g^*, \mathbf{W}_g^*; \boldsymbol{\theta}, \boldsymbol{\tau}, \sigma_U^2)\}$$

3. SMEM for binary response

- We assume that σ_U^2 is **known**
 - In practice, σ_U^2 can be estimated when there are replicate W **measurements** for each subject
 - It is **straightforward** to revise our approach when this occurs (Framingham)
- For either data structure (if the first two component models in the SMEM are correct) likelihood-based inference is **consistent** if and only if
 - the latent-variable model $f_X^{(a)}(x; \boldsymbol{\tau})$ is correct **OR**
 - $\sigma_U^2 = 0$
- Neither of these is likely to hold in practice....
- Thus, it is important to **understand** how the choice of $f_X^{(a)}(x; \boldsymbol{\tau})$ affects the resulting inference on $\boldsymbol{\theta}$

4. Diagnosing latent variable model misspecification

- To assess the **impact** of latent variable model misspecification in SMEMs for binary response, we
 - use **remeasurement-based methods** to assess the robustness of target estimators
 - develop **test statistics** to “test” for robustness (under the assumed latent variable model)
- **General idea:** Compare the fits of the individual and pooled SMEMs
 - “**Large**” differences between the fits can mean something

4. Remeasurement method

- The remeasurement method involves **further contaminating** the observed covariate W
- For a chosen $\lambda > 0$, we first generate B sets of n independent $N(0, 1)$ random errors, $\{Z_{bi}, i = 1, 2, \dots, n\}_{b=1}^B$, and form λ -remeasured (dirtier) data

$$W_{bi}(\lambda) = W_i + \sqrt{\lambda}\sigma_U Z_{bi},$$

for $b = 1, 2, \dots, B$ and $i = 1, 2, \dots, n$

- We now have B new **dirtier** data sets
- By this construction, the ME variance associated with $W_{bi}(\lambda)$ is

$$(1 + \lambda)\sigma_U^2$$

- It is interesting to note that $\lambda = -1$ describes what happens in the absence of measurement error

4. Remeasurement method

- Let $\boldsymbol{\Omega} = (\boldsymbol{\theta}^T, \boldsymbol{\tau}^T)^T$ denote the $r \times 1$ vector of unknown parameters, where $r = 2 + t$
- Suppose that $\boldsymbol{\Omega}$ is to be estimated by solving the vector-valued estimating equation

$$\sum_{i=1}^n \boldsymbol{\psi}(Y_i, W_i; \boldsymbol{\theta}, \boldsymbol{\tau}, \sigma_U^2) = \mathbf{0} \quad (1)$$

in the absence of further contamination (i.e., using the **observed data**)

- For example, if the **MLE** of $\boldsymbol{\Omega}$ is desired, then

$$\sum_{i=1}^n \boldsymbol{\psi}(Y_i, W_i; \boldsymbol{\theta}, \boldsymbol{\tau}, \sigma_U^2) = (\partial/\partial\boldsymbol{\Omega})l(\boldsymbol{\theta}, \boldsymbol{\tau}, \sigma_U^2)$$

Denote the estimator that solves (1) by $\hat{\boldsymbol{\Omega}}(0) = \{\hat{\boldsymbol{\theta}}(0)^T, \hat{\boldsymbol{\tau}}(0)^T\}^T$

4. Remeasurement method

- We can construct the **same estimating equation** using the λ -remeasured data

$$\sum_{i=1}^n \psi_B \{Y_i, \widetilde{W}_i(\lambda); \boldsymbol{\theta}, \boldsymbol{\tau}, (1 + \lambda)\sigma_U^2\} = \mathbf{0}, \quad (2)$$

where

$$\psi_B \{Y_i, \widetilde{W}_i(\lambda); \boldsymbol{\theta}, \boldsymbol{\tau}, (1 + \lambda)\sigma_U^2\} = B^{-1} \sum_{b=1}^B \psi \{Y_i, W_{bi}(\lambda); \boldsymbol{\theta}, \boldsymbol{\tau}, (1 + \lambda)\sigma_U^2\}$$

and $\widetilde{W}_i(\lambda) = \{W_{bi}(\lambda)\}_{b=1}^B$

- Solving (2) yields the **same type of estimator** as $\widehat{\boldsymbol{\Omega}}(0)$ based on the λ -remeasured data; we denote this estimator by $\widehat{\boldsymbol{\Omega}}(\lambda) = \{\widehat{\boldsymbol{\theta}}(\lambda)^T, \widehat{\boldsymbol{\tau}}(\lambda)^T\}^T$

4. SIMEX plot

- Plotting one component of $\hat{\Omega}(\lambda)$ versus λ , for $\lambda \geq 0$, produces the **simulation extrapolation** (SIMEX) plot for that estimate (Cook and Stefanski, 1994)
- A **constant** (or nearly constant) SIMEX plot suggests that the considered estimator is **robust** to measurement error under the assumed model for X
- **Substantial deviation** from constancy in SIMEX plot indicates that the estimator is **sensitive** to measurement error under the assumed model for X

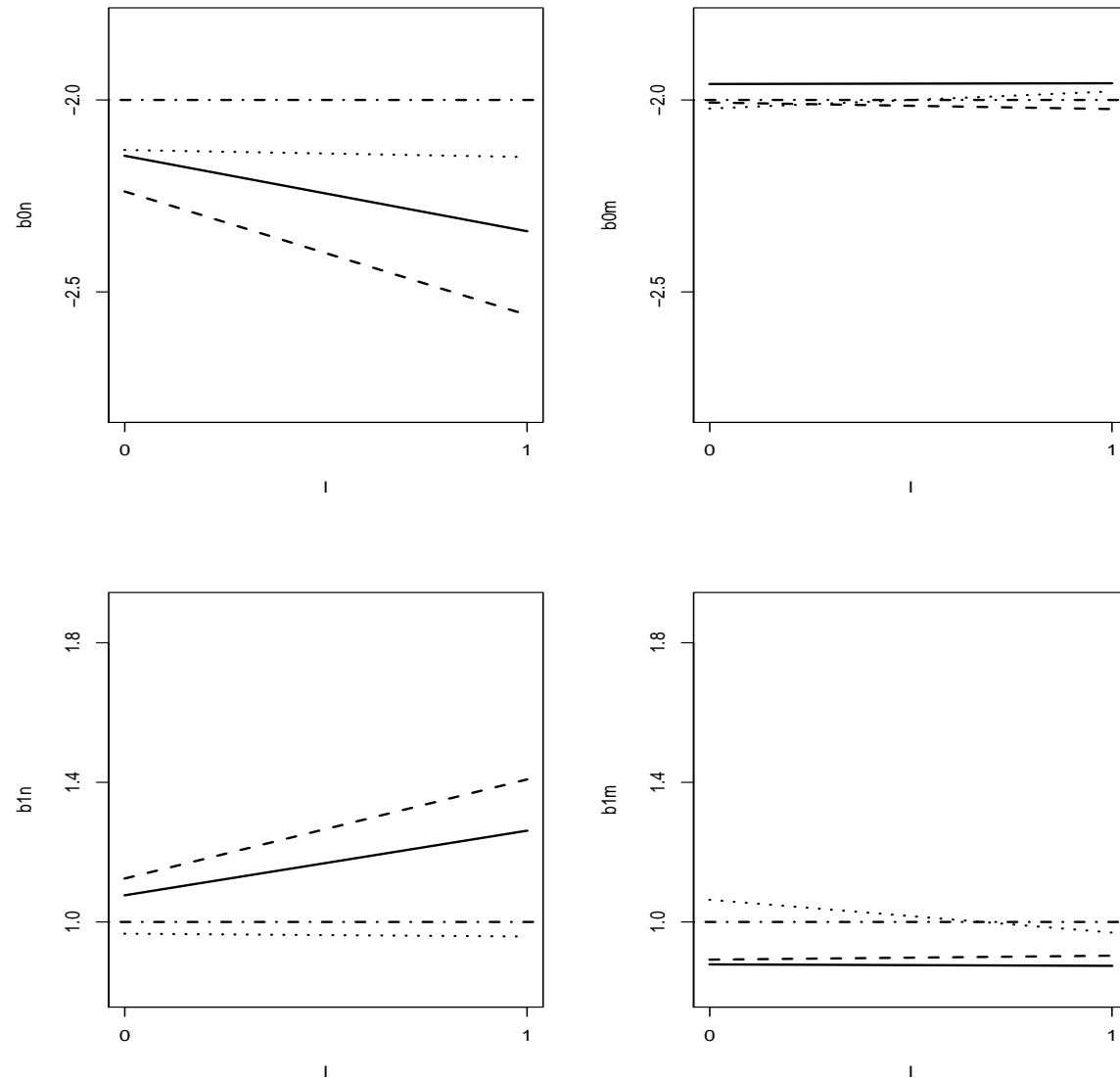


Figure 1: *SIMEX* plots for three data collection strategies. *Top*: Estimator $\hat{\beta}_0$. *Bottom*: Estimator $\hat{\beta}_1$. *Left (Right)*: Latent variable model A (B).

4. Remeasurement method: Pooled response

- The **same** data generation step applies when pooled responses are considered
- Suppose that the target estimator for Ω , based on the data with **pooled response**, solves

$$\sum_{g=1}^G \psi^*(Y_g^*, \mathbf{W}_g^*; \boldsymbol{\theta}, \boldsymbol{\tau}, \sigma_U^2) = \mathbf{0} \quad (3)$$

and, based on the λ -remeasured data, solves

$$\sum_{g=1}^G \psi_B^*\{Y_g^*, \widetilde{\mathbf{W}}_g^*(\lambda); \boldsymbol{\theta}, \boldsymbol{\tau}, (1 + \lambda)\sigma_U^2\} = \mathbf{0} \quad (4)$$

Denote by $\widehat{\Omega}^*(0) = \{\widehat{\boldsymbol{\theta}}^*(0)^T, \widehat{\boldsymbol{\tau}}^*(0)^T\}^T$ and $\widehat{\Omega}^*(\lambda) = \{\widehat{\boldsymbol{\theta}}^*(\lambda)^T, \widehat{\boldsymbol{\tau}}^*(\lambda)^T\}^T$ the solutions to (3) and (4), respectively

- **SIMEX plot** \rightsquigarrow plot of $\widehat{\Omega}^*(\lambda)$ versus λ

4. Pooling strategies

- We consider **two strategies** for pooling individuals
 - **Random** pooling
 - **Homogeneous** pooling
- In the presence of covariate measurement error, the best homogeneous composition one can hope for is to pool individuals with **similar W values**
- **Small complication:** Homogeneous pooling renders a small amount of dependence among the pools due to the ordering of the observed covariates
- Therefore, the MLE of Ω based on the homogeneous-pooling responses is not quite obtained by maximizing

$$l^*(\boldsymbol{\theta}, \boldsymbol{\tau}, \sigma_U^2) = \sum_{g=1}^G \log\{f_{Y^*, \mathbf{W}^*}(Y_g^*, \mathbf{W}_g^*; \boldsymbol{\theta}, \boldsymbol{\tau}, \sigma_U^2)\}$$

4. Pooling strategies

- **Solution:** Implement a **two-stage** estimation procedure
 - **Step 1:** Estimate τ by maximizing the loglikelihood of \mathbf{W} , given by

$$l_W(\tau, \sigma_U^2) = \sum_{i=1}^n \log \left\{ \int f_{W|X}(W_i|x; \sigma_U^2) f_X^{(a)}(x; \tau) dx \right\}$$

\rightsquigarrow we get a **consistent** estimator for τ

- **Step 2:** Estimate θ by maximizing

$$l^*(\theta, \hat{\tau}, \sigma_U^2) = \sum_{g=1}^G \log \{ f_{Y^*, \mathbf{W}^*}(Y_g^*, \mathbf{W}_g^*; \theta, \hat{\tau}, \sigma_U^2) \}$$

- We call this maximizer $\hat{\theta}$ a **pseudo-MLE**
- **Note:** This complication does not arise with random pooling; one can maximize the likelihood directly

4. Simulation evidence

- Simulation settings:

- $h(\beta_0 + \beta_1 X) = \Phi(\beta_0 + \beta_1 X)$, where $\Phi(\cdot)$ is the $N(0, 1)$ cdf
- $\theta = (\beta_0, \beta_1)^T = (-2, 1)^T$
- the **true** model for X is a two-component normal mixture

$$f_X^{(a)}(x) = 0.1f_1(x) + 0.9f_2(x),$$

where $f_1(x)$ and $f_2(x)$ are $N(2.35, 0.41)$ and $N(-0.26, 0.38)$ pdfs

- Under these settings, a random sample of size $n = 2000$ is generated from the SMEM for individual data
- **Pooled responses:** We set $n_g = 10$, for $g = 1, 2, \dots, G$, yielding $G = 200$ pools
- In the remeasurement method, we set $B = 50$ and take $\lambda = 1$

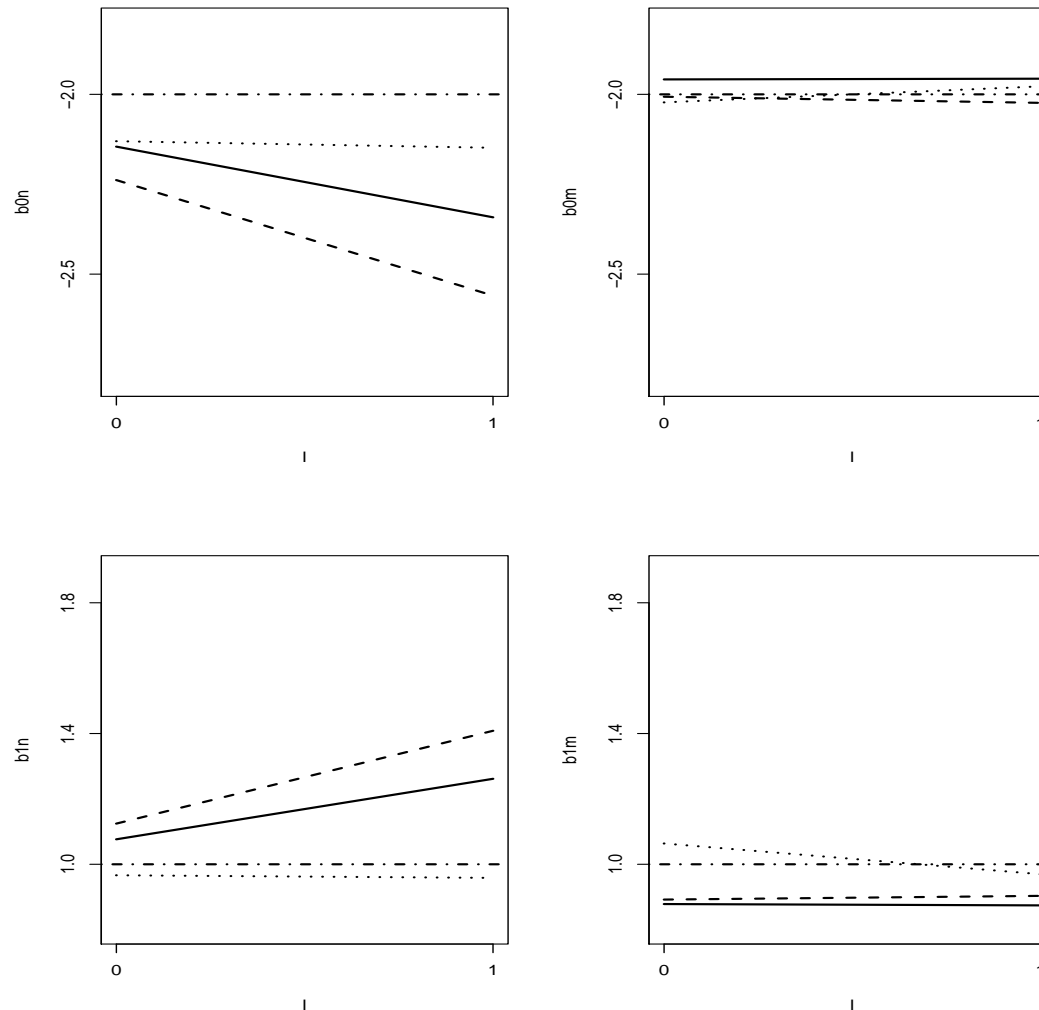


Figure 2: *SIMEX* plots. *Left*: Normal latent variable model (incorrect). *Right*: Normal mixture latent variable model (correct). *Top*: Estimator $\hat{\beta}_0$. *Bottom*: Estimator $\hat{\beta}_1$. *Solid* line: Individual data. *Dashed* line: Random pooling. *Dotted* line: Homogeneous pooling.

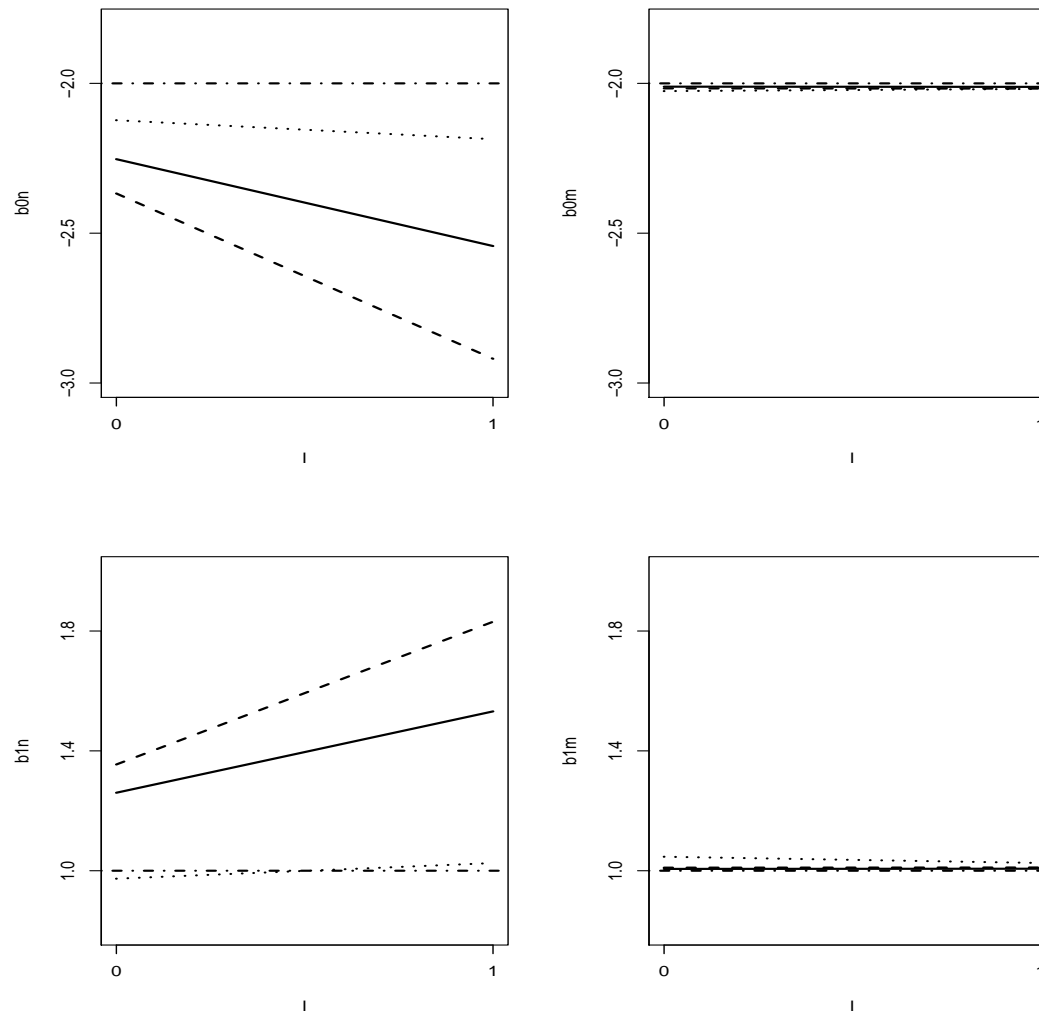


Figure 3: *SIMEX* averages. *Left*: Normal latent variable model (incorrect). *Right*: Normal mixture latent variable model (correct). *Top*: Estimator $\hat{\beta}_0$. *Bottom*: Estimator $\hat{\beta}_1$. *Solid* line: Individual data. *Dashed* line: Random pooling. *Dotted* line: Homogeneous pooling.

Table 1: **Mean regression parameter estimates** from 300 MC replications under the normal and normal mixture latent-variable assumption with $\beta_0 = -2$, $\beta_1 = 1$, and $\lambda = 0, 1$. MC MSEs are in parentheses. IND, RP, and HP represent individual response, random-pooling response, and homogeneous-pooling response, respectively.

Normal	$\widehat{\beta}_0^{(n)}(0)$	$\widehat{\beta}_0^{(n)}(1)$	$\widehat{\beta}_1^{(n)}(0)$	$\widehat{\beta}_1^{(n)}(1)$
IND	-2.25 (0.064)	-2.54 (0.295)	1.26 (0.068)	1.53 (0.283)
RP	-2.37 (0.135)	-2.92 (0.846)	1.35 (0.126)	1.83 (0.691)
HP	-2.12 (0.015)	-2.19 (0.035)	0.97 (0.001)	1.03 (0.001)
Normal mixture	$\widehat{\beta}_0^{(m)}(0)$	$\widehat{\beta}_0^{(m)}(1)$	$\widehat{\beta}_1^{(m)}(0)$	$\widehat{\beta}_1^{(m)}(1)$
IND	-2.01 (< 0.001)	-2.01 (< 0.001)	1.01 (< 0.001)	1.01 (< 0.001)
RP	-2.02 (< 0.001)	-2.02 (< 0.001)	1.01 (< 0.001)	1.01 (< 0.001)
HP	-2.03 (0.001)	-2.02 (< 0.001)	1.05 (0.002)	1.03 (0.001)

4. Robustness and other findings

- Estimates computed from HP response data can be **more robust** to measurement error than those from individual response (or to those from RP response)
 - Using HP increases the **reliability ratio**
 - The variation in the true predictor among the HP pools is usually **larger** than the variation in the true predictor based on the individuals
- Inference based on the pooled responses can actually be **more efficient** than that based on the individual response
 - Estimates computed from HP have **smaller MSE** when the pool sizes n_g are not too large
- Our findings are especially encouraging when **group testing** is used for cost concerns

4. Test statistics

- We provide a quantitative assessment of robustness in the form of a **test statistic**

$$t_1^*(\lambda) = \hat{\nu}_1^{-1} \{ \hat{\gamma}(\lambda) - \hat{\gamma}(0) \},$$

where

- γ is an element in Ω
 - $\hat{\gamma}(\cdot)$ is the target estimator for γ
 - $\hat{\nu}_1^2$ is an estimator for $\text{var}\{\hat{\gamma}(\lambda) - \hat{\gamma}(0)\}$
- The statistic $t_1^*(\lambda)$ evaluates the amount of **change** in the estimate as the ME variance increases from σ_U^2 to $(1 + \lambda)\sigma_U^2$, adjusting for background noise
 - If $t_1^*(\lambda)$ deviates significantly from zero, one may conclude that the estimator is **not robust** to ME under the assumed latent-variable model

4. Test statistics

- For $G > r$, we define

$$t_2^* = \frac{G - r}{r(G - 1)} \times \{\hat{\Omega}_R(0) - \hat{\Omega}_I(0)\}^T \hat{\Sigma}^{-1} \{\hat{\Omega}_R(0) - \hat{\Omega}_I(0)\},$$

where

- $\hat{\Omega}_{(\cdot)}(0)$ is the **MLE** of $\Omega = (\theta^T, \tau^T)^T$
- $\hat{\Sigma}$ is an estimator of the variance-covariance matrix of $\hat{\Omega}_R(0) - \hat{\Omega}_I(0)$
- **Motivation:** When the latent-variable model is correct, $\hat{\theta}_I$ and $\hat{\theta}_R$ are in close agreement, but they can differ largely under misspecification
- When the latent model is correct, $t_2^* \sim F(r, G - r)$
- **Computational advantage:** t_2^* does not depend on the remeasured data

Table 2: Rejection rates of $t_1^*(1)$ and t_2^* based on 300 MC replications. MC standard errors are in parentheses. IND, RP, and HP represent individual response, random-pooling response, and homogeneous-pooling response, respectively.

	$\widehat{\beta}_0^{(n)}$			$\widehat{\beta}_1^{(n)}$		
Normal	IND	RP	HP	IND	RP	HP
$t_1^*(1)$	1 (0)	0.45 (0.03)	0.16 (0.02)	1 (0)	0.74 (0.03)	0.12 (0.02)
t_2^*	0.87 (0.02)	—	—	—	—	—
	$\widehat{\beta}_0^{(m)}$			$\widehat{\beta}_1^{(m)}$		
Normal mixture	IND	RP	HP	IND	RP	HP
$t_1^*(1)$	0.04 (0.01)	0.01 (0.01)	0.02 (0.01)	0.04 (0.01)	0.02 (0.01)	0.07 (0.01)
t_2^*	0.03 (0.01)	—	—	—	—	—

- t_2^* has **good power** (although not as high as $t_1^*(1)$ based on individual testing)
- When the assumed latent-variable model is incorrect, the power of $t_1^*(1)$ under HP is **low** (this reinforces our robustness discovery)

5. Framingham Heart Study

- Define
 - Y = the binary indicator of the first evidence of coronary heart disease
 - X = the long-term SBP
- For each subject, two SBP readings (W_1 and W_2) collected during clinic visits can be viewed as **error-contaminated** versions of X
 - These surrogate replicates can be used to **estimate** σ_U^2 (details omitted)
- We posit the **probit model**, $\text{pr}(Y = 1|X) = \Phi(\beta_0 + \beta_1 X)$
- Normal and normal mixture latent models
- We set $n_g = 5$, for $g = 1, 2, \dots, 323$
- For the remeasurement method, we take $B = 50$ and $\lambda = 1$

Table 3: **Framingham data**. Values of $t_1^*(1)$ and t_2^* computed under the normal and normal mixture assumption. IND, RP, and HP represent individual response, random-pooling response, and homogeneous-pooling response, respectively. The numbers in parentheses in the rows for t_2^* are the p -values associated with t_2^* .

		$\widehat{\beta}_0^{(n)}$			$\widehat{\beta}_1^{(n)}$		
Normal	IND	RP	HP	IND	RP	HP	
$t_1^*(1)$	-2.87	-1.88	-1.67	2.92	1.93	1.51	
t_2^*	0.21 (0.93)	—	—	—	—	—	
		$\widehat{\beta}_0^{(m)}$			$\widehat{\beta}_1^{(m)}$		
Normal mixture	IND	RP	HP	IND	RP	HP	
$t_1^*(1)$	0.46	0.80	-1.44	-0.45	-0.79	1.31	
t_2^*	1.51 (0.16)	—	—	—	—	—	

- Under the normal model assumption, the statistic $t_1^*(1)$ based on individual response and t_2^* suggest **different** conclusions
- **Novelty**: If one prefers the simpler normal model, the pseudo MLE based on HP responses is less sensitive to measurement error

6. Discussion

- Including **multiple covariates** (error prone or not) poses no additional challenges; details given in the paper
- The statistics $t_1^*(\lambda)$ and t_2^* are motivated differently, but they share the common theme of **information reduction**
 - One **adds more noise** to the observed W to compute $t_1^*(\lambda)$
 - One **conceals the individual responses** to compute t_2^*
- **Intriguing:** One can learn more from the data by reducing information
- This general idea may be applicable to **other problems** involving binary response