Informative Model Specification Tests

Xianzheng Huang

University of South Carolina

June 08, 2011

Outline

- **Research goal**
  - — To disentangle multiple sources of model misspecification.
- **Existing methods**
- **Proposed methods**
  - — For measurement error models
  - — For mixed effects models
- **Summary**

Latent variable models

- Structural measurement error models:

$$\begin{cases} P(Y_i = 1|X_i, \boldsymbol{\beta}) = \Phi(\beta_0 + \beta_1 X_i), \\ W_i = X_i + U_i. \end{cases}$$

- Linear mixed models (LMM):

$$Y_{ij} = \beta_0 + b_{i0} + (\beta_1 + b_{i1})X_{ij} + \epsilon_{ij}.$$

Latent variable models

- Structural measurement error models:

$$\begin{cases} P(Y_i = 1|X_i, \boldsymbol{\beta}) = \Phi(\beta_0 + \beta_1 X_i), \\ W_i = X_i + U_i. \end{cases}$$

- Linear mixed models (LMM):

$$Y_{ij} = \beta_0 + b_{i0} + (\beta_1 + b_{i1})X_{ij} + \epsilon_{ij}.$$

Concerns:

- Model assumptions for latent variables can be inappropriate.
- Other model assumptions can be violated.

Existing Methods

- **Remeasurement method** (Huang,Stefanski,and Davidian, 2006).
- **Score-type tests** (Ma, Hart, Janicki, Carroll, 2011).
- **Estimate random-effects distributions** (Lange and Ryan, 1989; Ritz, 2004; Waagepetersen, 2006).
- **Compare two estimators** (Hausman, 1978; Tchetgen and Coull, 2006; White, 1981, 1982; Litière, 2007).
- **Variance component tests** (Self and Liang 1987; Stram and Lee, 1994; Crainiceanu and Ruppert, 2004; Saville and Herring, 2009).

Limitations:

- Some are like (overall) goodness-of-fit tests.
- Most of them assume only one source of misspecification.
- Many test statistics do not have familiar null distribution.

Coarsened Data

Notations: For $i = 1, \ldots, m$,

- $\mathbf{Y}_i$: the ith observed datum in the raw data.
- $\mathbf{Y}_i^*$: the ith datum in the coarsened data.
- User-designed coarsening mechanism: $f_{\mathbf{Y}_i^*|\mathbf{Y}_i}(\mathbf{Y}_i^*|\mathbf{Y}_i; \boldsymbol{\lambda})$.

Coarsened Data

Notations: For $i = 1, \ldots, m$,

- $\mathbf{Y}_i$: the ith observed datum in the raw data.
- $\mathbf{Y}_i^*$: the ith datum in the coarsened data.
- User-designed coarsening mechanism: $f_{\mathbf{Y}_i^*|\mathbf{Y}_i}(\mathbf{Y}_i^*|\mathbf{Y}_i; \boldsymbol{\lambda})$.

Examples of coarsened data:

— censored data,

— missing data,

— grouped data, etc.

Coarsened Data

Notations: For $i = 1, \ldots, m$,

- $\mathbf{Y}_i$: the ith observed datum in the raw data.
- $\mathbf{Y}_i^*$: the ith datum in the coarsened data.
- User-designed coarsening mechanism: $f_{\mathbf{Y}_i^* | \mathbf{Y}_i}(\mathbf{Y}_i^* | \mathbf{Y}_i; \boldsymbol{\lambda})$.

Examples of coarsened data:

— censored data,

— missing data,

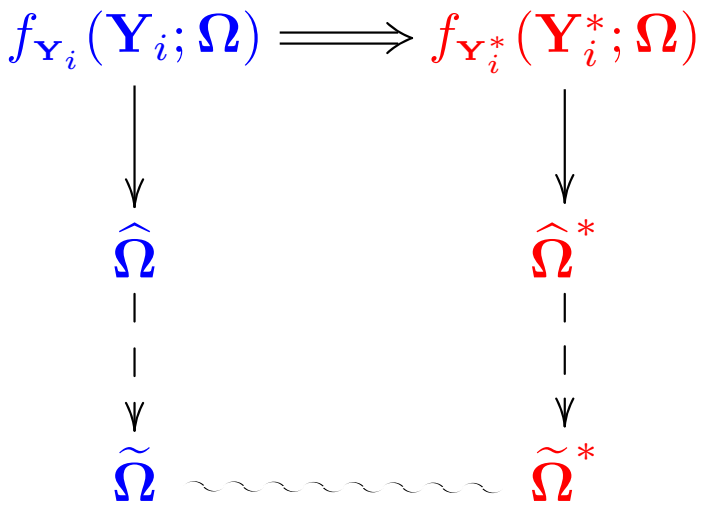
— grouped data, etc.

Generalization: $\mathbf{Y}_i^*$ contains less information than $\mathbf{Y}_i$.

Implementation of the proposed methods:

- Step 1: Design a coarsening mechanism. Generate coarsened data $\{\mathbf{Y}_i^*, i = 1, \ldots, m\}$.

- Step 2: Draw likelihood-based inference twice:



- Step 3: Inspect the discrepancy between $\widehat{\mathbf{\Omega}}$ and $\widehat{\mathbf{\Omega}}^*$ elementwise.

How to assess the discrepancy between $\widehat{\mathbf{\Omega}}$ and $\widehat{\mathbf{\Omega}}^*$?

A series of t tests will do!

Consider testing $H_0 : \tilde{\theta}^* = \tilde{\theta}$, where θ is an element in $\mathbf{\Omega}$.

A test statistic is given by

$$t_{1\theta} = \frac{\widehat{\theta}^* - \widehat{\theta}}{\widehat{\nu}},$$

where $\widehat{\nu} = \widehat{\text{std}}(\widehat{\theta}^* - \widehat{\theta})$.

Illustration in Measurement Error Models

Consider a probit linear model:

$$\begin{cases} P(Y_i = 1|X_i, \boldsymbol{\beta}) = \Phi(\beta_0 + \beta_1 X_i), \\ W_i = X_i + U_i. \end{cases}$$

where $U_i \sim N(0, \sigma_u^2)$.

Theoretical motivation of Remeasurement Method (RM):

When X-model is misspecified,

$$\widehat{\boldsymbol{\beta}} \xrightarrow{p} \boldsymbol{\beta}(\sigma_u) \neq \boldsymbol{\beta}^{(0)}.$$

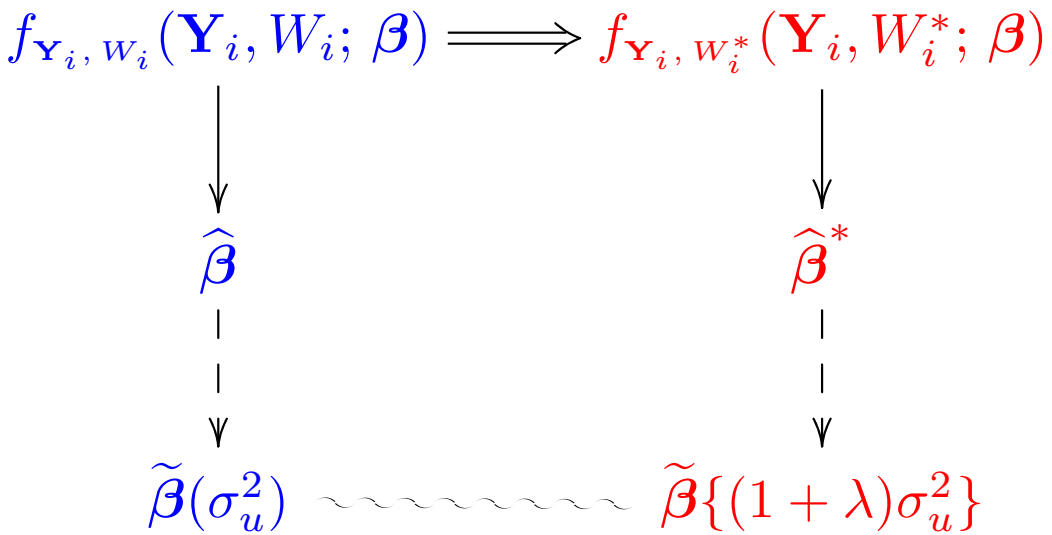
Implementation of Remeasurement Method (RM):

- Step 1: Generate "*coarsened data*":

$$W_i^* = W_i + \sqrt{\lambda}\sigma_u Z_i,$$

where $Z_i \sim N(0,1)$.

- Step 2:



- **Limitation (or a good feature) of RM:**

  Robust to link function misspecification.

- **Limitation (or a good feature) of RM:**

Robust to link function misspecification.

- **How to detect link misspecification?**

New solution: Reclassification Method (RC).

Consider coarsening Y_i, e.g., define "*coarsened data*":

$$P(Y_i^* = Y_i | W_i; \tau) = \Phi(W_i + \tau).$$

• **Limitation (or a good feature) of RM:**

Robust to link function misspecification.

• **How to detect link misspecification?**

New solution: Reclassification Method (RC).

Consider coarsening Y_i, e.g., define "*coarsened data*":

$$P(Y_i^* = Y_i | W_i; \tau) = \Phi(W_i + \tau).$$

• **Theoretical motivation of Reclassification Method (RC):**

When the link is misspecified,

$$\widehat{\boldsymbol{\beta}} \xrightarrow{p} \boldsymbol{\beta}(\tau) \neq \boldsymbol{\beta}^{(0)}.$$

Implementation of RM-RC method:

- Step 1: Generate "*coarsened data*":

$$\begin{aligned} P(Y_i^* = Y_i | W_i; \tau) &= \Phi(W_i + \tau), \\ W_i^* &= W_i + \sqrt{\lambda}\sigma_u Z_i. \end{aligned}$$

- Step 2:

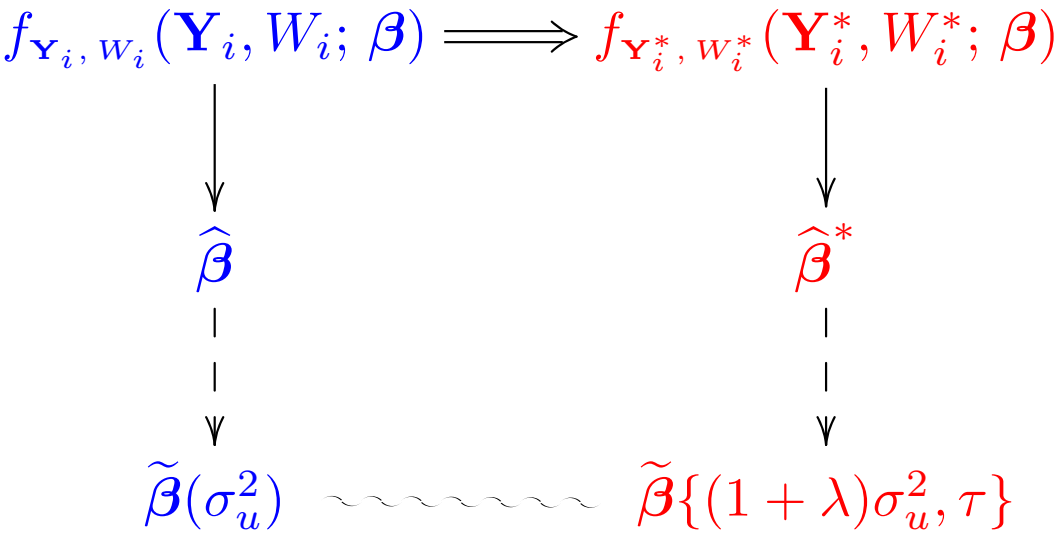


Illustration in LMM: Tests for Random Slopes

- For $i = 1, \ldots, m$, $j = 1, \ldots, n$,

$$M_2: \quad Y_{ij} = \beta_0 + \beta_1 X_{ij} + \epsilon_{ij}, \qquad \text{(assumed model)}$$
$$M_3: \quad Y_{ij} = \beta_0 + (\beta_1 + b_{i1}) X_{ij} + \epsilon_{ij}, \qquad \text{(true mode)}$$

where

— $b_{i1} \sim$ [some distribution]$(0, \sigma^2_{b1})$,

— $\epsilon_{ij} \sim N(0, \sigma^2_\epsilon)$,

— X_{ij} are covariate values.

Under M_2, $\boldsymbol{\Omega} = (\boldsymbol{\beta}^T, \ \sigma^2_\epsilon)^T$.

Illustration in LMM: Tests for Random Slopes

- For $i = 1, \ldots, m$, $j = 1, \ldots, n$,

$$M_2: \quad Y_{ij} = \beta_0 + \beta_1 X_{ij} + \epsilon_{ij}, \qquad \text{(assumed model)}$$
$$M_3: \quad Y_{ij} = \beta_0 + (\beta_1 + b_{i1}) X_{ij} + \epsilon_{ij}, \qquad \text{(true mode)}$$

where

— $b_{i1} \sim$ [some distribution]$(0, \sigma^2_{b1})$,

— $\epsilon_{ij} \sim N(0, \sigma^2_\epsilon)$,

— X_{ij} are covariate values.

Under M_2, $\boldsymbol{\Omega} = (\boldsymbol{\beta}^T, \ \sigma^2_\epsilon)^T$.

- One way to create coarsened data:

Create missing data in $\mathbf{Y}_i = (Y_{i1}, \ldots, Y_{in})$.

• **Start from the simplest missingness mechanism:**

Let $\Delta_i = I(Y_{in}$ is missing), and

$$P(\Delta_i = 1 | \mathbf{Y}_i) = \lambda \in (0, 1).$$

—— **M**issing **C**ompletely **A**t **R**andom (**MCAR**)

• **Start from the simplest missingness mechanism:**

Let $\Delta_i = I(Y_{in}$ is missing), and

$$P(\Delta_i = 1|\mathbf{Y}_i) = \lambda \in (0,\, 1).$$

—— **M**issing **C**ompletely **A**t **R**andom (**MCAR**)

• One can show that $\widetilde{\boldsymbol{\beta}} = \widetilde{\boldsymbol{\beta}}^* = \boldsymbol{\beta}^{(0)}$, and

$$\tilde{\sigma}_\epsilon^2 = \sigma_\epsilon^{2(0)} + \frac{\sigma_{b1}^{2(0)}}{n} \sum_{j=1}^{n} E(X_{1j}^2), \tag{1}$$

$$\tilde{\sigma}_\epsilon^{2*} = \sigma_\epsilon^{2(0)} + \frac{\sigma_{b1}^{2(0)}}{n-\lambda} \left\{ \sum_{j=1}^{n} E(X_{1j}^2) - \lambda E(X_{1n}^2) \right\}. \tag{2}$$

• **Start from the simplest missingness mechanism:**

Let $\Delta_i = I(Y_{in}$ is missing), and

$$P(\Delta_i = 1|\mathbf{Y}_i) = \lambda \in (0,\, 1).$$

—— **M**issing **C**ompletely **A**t **R**andom (**MCAR**)

• One can show that $\widetilde{\boldsymbol{\beta}} = \widetilde{\boldsymbol{\beta}}^* = \boldsymbol{\beta}^{(0)}$, and

$$\tilde{\sigma}_\epsilon^2 = \sigma_\epsilon^{2(0)} + \frac{\sigma_{b1}^{2(0)}}{n} \sum_{j=1}^{n} E(X_{1j}^2), \tag{1}$$

$$\tilde{\sigma}_\epsilon^{2*} = \sigma_\epsilon^{2(0)} + \frac{\sigma_{b1}^{2(0)}}{n-\lambda} \left\{ \sum_{j=1}^{n} E(X_{1j}^2) - \lambda E(X_{1n}^2) \right\}. \tag{2}$$

Comparing (1) and (2) gives

$$\tilde{\sigma}_\epsilon^{2*} = \tilde{\sigma}_\epsilon^2 \Longleftrightarrow E(X_{1n}^2) = \frac{1}{n-1} \sum_{j=1}^{n-1} E(X_{1j}^2). \tag{3}$$

Implications:

- One can use Equation (1)

$$\tilde{\sigma}_{\epsilon}^{2} = \sigma_{\epsilon}^{2(0)} + \frac{\sigma_{b1}^{2(0)}}{n} \sum_{j=1}^{n} E(X_{1j}^{2})$$

to identify which covariate needs a random slope.

Implications:

- One can use Equation (1)

$$\tilde{\sigma}_{\epsilon}^{2} = \sigma_{\epsilon}^{2(0)} + \frac{\sigma_{b1}^{2(0)}}{n} \sum_{j=1}^{n} E(X_{1j}^{2})$$

to identify which covariate needs a random slope.

- Equations (1) and (2) allow one to identify σ_{b1}^{2}:

$$\hat{\sigma}_{b1}^{2} = (\hat{\sigma}_{\epsilon}^{2} - \hat{\sigma}_{\epsilon}^{2*}) \frac{n(n-\lambda)}{(n-1)\lambda} \left\{ E(X_{1n}^{2}) - \frac{1}{n-1} \sum_{j=1}^{n} E(X_{1j}^{2}) \right\}^{-1} .$$

General strategy to identify the "unidentifiables":

- Using the raw observed data, one has MLE

$$\hat{\gamma} \xrightarrow{p} \alpha^{(0)} + \gamma^{(0)}.$$

— One cannot identify α and γ.

General strategy to identify the "unidentifiables":

- Using the raw observed data, one has MLE

$$\hat{\gamma} \xrightarrow{p} \alpha^{(0)} + \gamma^{(0)}.$$

— One cannot identify α and γ.

- Suppose one generates a coarsened data set, and obtains MLE

$$\hat{\gamma}^* \xrightarrow{p} \alpha^{(0)} + 2\gamma^{(0)}.$$

— Now one can identify α and γ by combining two MLEs!

General strategy to identify the "unidentifiables":

• Using the raw observed data, one has MLE

$$\hat{\gamma} \xrightarrow{p} \alpha^{(0)} + \gamma^{(0)}.$$

— One cannot identify α and γ.

• Suppose one generates a coarsened data set, and obtains MLE

$$\hat{\gamma}^* \xrightarrow{p} \alpha^{(0)} + 2\gamma^{(0)}.$$

— Now one can identify α and γ by combining two MLEs!

• This phenomenon also explains why using coarsened data can test some (traditionally) "untestable" model assumptions.

- **Recap:** Using **MCAR** can reveal random slopes.

- **Unresolved:** If b_{i1} is needed, is $b_{i1} \sim$ Normal?

- **Solution:** Generate coarsened data according to

$$P(\Delta_i = 1|Y_i) = \Phi(\lambda_0 + \lambda_1 Y_{in}),$$

where $\lambda_1 \neq 0$.
—— **N**ot **M**issing **A**t **R**andom (**NMAR**)

One can show that $\widetilde{\mathbf{\Omega}}^* = \widetilde{\mathbf{\Omega}}$ if and only if

$$\lim_{m\to\infty} m^{-1} \sum_{i=1}^{m} E_0 \left(\Delta_i \frac{\partial l_{\text{mis},i}}{\partial \mathbf{\Omega}} \Big|_{\widetilde{\mathbf{\Omega}}} \right)$$

$$= \lim_{m\to\infty} m^{-1} \sum_{i=1}^{m} E_0(\Delta_i) \frac{E\left(\Delta_i \frac{\partial l_{\text{mis},i}}{\partial \mathbf{\Omega}} \Big|_{\widetilde{\mathbf{\Omega}}}; \widetilde{\mathbf{\Omega}} \right)}{E(\Delta_i; \widetilde{\mathbf{\Omega}})}. \tag{4}$$

A sufficient condition for (4) is

$$f_Y^{(0)}(Y_{\text{mis},i}|X_i; \mathbf{\Omega}^{(0)}) = f_Y(Y_{\text{mis},i}|X_i; \widetilde{\mathbf{\Omega}}). \tag{5}$$

Implications:

- If b_{i1} is not normal, then Equation (5) does not hold.

$$f_Y^{(0)}(Y_{\text{mis},i}|X_i;\boldsymbol{\Omega}^{(0)}) = f_Y(Y_{\text{mis},i}|X_i;\widetilde{\boldsymbol{\Omega}}).$$

Thus $\widetilde{\boldsymbol{\Omega}}^* \neq \widetilde{\boldsymbol{\Omega}}$.

Implications:

- If b_{i1} is not normal, then Equation (5) does not hold.

$$f_Y^{(0)}(Y_{\text{mis},i}|X_i;\boldsymbol{\Omega}^{(0)}) = f_Y(Y_{\text{mis},i}|X_i;\widetilde{\boldsymbol{\Omega}}).$$

Thus $\widetilde{\boldsymbol{\Omega}}^* \neq \widetilde{\boldsymbol{\Omega}}$.

- If $b_{i1} \sim N(0, \sigma_{b1}^2)$, then (assume fixed design points)

(C1) If

$$X_{1n}^2 = \frac{1}{n-1}\sum_{j=1}^{n-1} X_{1j}^2, \tag{6}$$

then Equation (5) holds and thus $\widetilde{\boldsymbol{\Omega}}^* = \widetilde{\boldsymbol{\Omega}}$.

(C2) If (6) does not hold then Equation (5) does not hold either and thus $\widetilde{\boldsymbol{\Omega}}^* \neq \widetilde{\boldsymbol{\Omega}}$.

Summary

- Dare to "mess up" the raw data!
- Dare to "mess up" the raw data repeatedly in different ways.
- Properties of MLE in the presence of different sources of model misspecification interacting with different coarsening mechanisms.