

Using Linear Mixed Models To Efficiently Allow For Covariate Measurement Error Using Replication Data

Jonathan Bartlett
Chris Frost and Bianca De Stavola

Medical Statistics Unit
London School of Hygiene and Tropical Medicine

24th August 2009

Outline

Parametric likelihood for covariate measurement error

Linear regression

Logistic regression

Conclusions

Covariate measurement error

- ▶ Ignoring classical measurement error in covariates causes bias
- ▶ e.g. effect of blood pressure on risk of coronary heart disease
- ▶ Many methods have been developed to allow for covariate measurement error in regression models
- ▶ e.g. regression calibration (RC), simulation extrapolation (SIMEX), maximum likelihood (ML) / joint modelling, semi-parametric methods (e.g. conditional score)

Internal replication setting

- ▶ We consider the situation in which the true covariate X_i cannot be observed, but replicate error-prone measurements are available $\mathbf{W}_i = (W_{i1}, W_{i2}, \dots, W_{in_i})$
- ▶ We assume the outcome Y_i is observed on all subjects
- ▶ This is referred to as internal replication data. A common scenario is where a subset of subjects have two error-prone measurements of the true covariate X_i , with the remaining having a single measurement

A parametric likelihood approach

- ▶ The parametric likelihood approach involves first specifying parametric models for:
 - ▶ An outcome model $f(Y_i|X_i)$
 - ▶ Covariate model $f(X_i)$
 - ▶ Measurement error model $f(\mathbf{W}_i|X_i)$
- ▶ Assuming the measurement error is non-differential (i.e. $f(Y_i|\mathbf{W}_i, X_i) = f(Y_i|X_i)$), these three sub-models define the joint distribution

Maximum likelihood estimation

- ▶ The models parameters can be estimated jointly by maximizing the likelihood function corresponding to the observed data

$$L = \prod_{i=1}^n f(Y_i, \mathbf{w}_i) = \prod_{i=1}^n \int f(Y_i, X_i, \mathbf{w}_i) dX_i$$

- ▶ From the general properties of MLEs, assuming the model is correctly specified, estimates are asymptotically efficient and consistent
- ▶ However, ML is not used as frequently as it might be. Simpler methods such as RC are more popular, especially in applied work

Issues with maximum likelihood for covariate error

- ▶ For many model specifications, the integral involved in L is intractable, e.g. Y_i given X_i logistic and X_i normally distributed
- ▶ Quadrature or Monte-Carlo methods have been successfully employed to approximate the integral, although these can be slow, especially when \mathbf{X}_i is multivariate and of higher dimension
- ▶ Researchers are rightly concerned about the strong parametric assumptions made by parametric ML, and sensitivity to violations of these assumptions
- ▶ Statistical packages do not routinely include commands to fit such models via ML
- ▶ Exceptions to this include NLMIXED in SAS, the user-written GLLAMM in Stata and the CME wrapper command (Rabe-Hesketh et al), and latent variable software such as Mplus

A parametric model for continuous outcomes

- ▶ Linear regression outcome model

$$Y_i = \alpha + \beta X_i + \epsilon_i$$
$$\epsilon_i \sim N(0, \sigma_\epsilon^2), \epsilon_i \perp X_i$$

- ▶ Normal covariate model

$$X_i \sim N(\mu_X, \sigma_X^2)$$

- ▶ Classical normally distributed measurement error

$$W_{ij} = X_i + U_{ij}, j = 1, \dots, n_i$$
$$U_{ij} \sim N(0, \sigma_U^2), U_{ij} \perp U_{ik}, U_{ij} \perp X_i, U_{ij} \perp \epsilon_i$$

Maximum likelihood estimation

- ▶ β can be estimated by maximizing the likelihood function corresponding to the observed data:

$$L = \prod_{i=1}^n f(Y_i, \mathbf{w}_i)$$

- ▶ The likelihood function for this model is tractable
- ▶ We could use the EM algorithm or Newton-Raphson to find ML estimates
- ▶ We show how linear mixed model commands can be used to find the MLE of β

Implied model for W_{ij} given Y_i

- ▶ The 'trick' is to consider the distribution of X_i given Y_i :

$$X_i = \gamma_0 + \gamma_Y Y_i + b_i$$

where $b_i \sim N(0, \sigma_{X|Y}^2)$

- ▶ It follows that:

$$\begin{aligned} W_{ij} &= X_i + U_{ij} \\ &= \gamma_0 + \gamma_Y Y_i + b_i + U_{ij} \end{aligned}$$

- ▶ Thus W_{ij} follows a standard linear mixed model:
 - ▶ fixed effect of Y_i
 - ▶ random subject intercepts b_i with variance $\sigma_{X|Y}^2$
 - ▶ within-subject variance σ_U^2
- ▶ This linear mixed model for W_{ij} given Y_i can be fitted with standard mixed model commands, using either ML or REML
- ▶ e.g. Stata's `xtmixed` (with data in 'long' format):

```
xtmixed w y || id:
```

Finding the MLE of β

- ▶ Recall that in linear regression:

$$\beta = \frac{\text{Cov}(Y_i, X_i)}{\text{Var}(X_i)}$$

- ▶ In terms of the parameters of the mixed model for W_{ij} given Y_i :

$$\beta = \frac{\gamma_Y \sigma_Y^2}{\sigma_{X|Y}^2 + \gamma_Y^2 \sigma_Y^2}$$

where σ_Y^2 denotes the marginal variance of Y_i .

- ▶ The MLE of β can be calculated by substituting the MLEs of the relevant parameters
- ▶ A Wald based confidence interval can also be calculated for the MLE of β

Simulations

- ▶ We compared using ML to regression calibration (RC) in simulations, with:
 - ▶ $\beta = 1$
 - ▶ $n = 5,000$ subjects, 500 of whom had $n_i = 2$ error-prone measurements of X_i , with 4,500 having a single error-prone measurement
 - ▶ $\text{Cor}(Y_i, X_i) = 0.2, 0.5, 0.8$, corresponding to weak, moderate and strong associations between Y_i and X_i
 - ▶ $\lambda = 1/3, 0.5, 2/3$, where λ is the reliability ratio of the error-prone measurements:

$$\lambda = \frac{\sigma_X^2}{\sigma_X^2 + \sigma_U^2}$$

- ▶ 10,000 simulations per scenario

Linear regression simulations with X_i normal

Scenario	Cor(Y_i, X_i)	λ	$\hat{\beta}_{RC}$	$\hat{\beta}_{ML}$
			Mean (SD)	Mean (SD)
1	0.2	2/3	1.001 (0.092)	1.001 (0.092)
2	0.2	1/2	1.004 (0.116)	1.004 (0.116)
3	0.2	1/3	1.014 (0.171)	1.014 (0.171)
4	0.5	2/3	1.000 (0.043)	1.000 (0.043)
5	0.5	1/2	1.003 (0.070)	1.003 (0.069)
6	0.5	1/3	1.017 (0.129)	1.017 (0.125)
7	0.8	2/3	1.001 (0.034)	1.001 (0.033)
8	0.8	1/2	1.004 (0.063)	1.003 (0.057)
9	0.8	1/3	1.013 (0.121)	1.011 (0.108)

- ▶ Little bias, and little efficiency benefit of ML compared to RC

Robustness, inference, and extensions

- ▶ The ML estimator for β derived under normality assumptions is consistent even if some or all of the normality assumptions do not hold
- ▶ Wald type confidence intervals for β can be calculated, based on standard errors for γ_Y , $\sigma_{X|Y}^2$ and σ_Y^2
- ▶ The approach easily extends to accommodate multivariate \mathbf{X}_i and error-free covariates \mathbf{Z}_i

Logistic regression

- ▶ The approach can also be used for binary Y_i , if we assume, as before, that:

$$X_i = \gamma_0 + \gamma_Y Y_i + b_i$$

with $b_i \sim N(0, \sigma_{X|Y}^2)$

- ▶ This implies a logistic regression model for Y_i given X_i , with log odds ratio:

$$\beta = \frac{\gamma_Y}{\sigma_{X|Y}^2}$$

- ▶ This is the normal discriminant model
- ▶ The most commonly assumed parametric model is of marginal normality for X_i , rather than normality conditional on Y_i
- ▶ The normal discriminant model is also used by the method of moment reconstruction (Freedman et al)

Logistic regression

- ▶ With replicate error-prone measurements W_{ij} , we again have the linear mixed model:

$$W_{ij} = \gamma_0 + \gamma_Y Y_i + b_i + U_{ij}$$

- ▶ Having fitted the linear mixed model by ML, the MLE of β is given by:

$$\hat{\beta} = \frac{\hat{\gamma}_Y}{\hat{\sigma}_{X|Y}}$$

- ▶ If the assumed model is correct, the MLE is consistent and asymptotically efficient
- ▶ In contrast, RC is only approximately consistent under certain assumptions

Logistic regression simulations with $X_i|Y_i$ normal

- ▶ We compared using ML to regression calibration (RC) in simulations, with:
 - ▶ $P(Y_i) = 0.1, 0.5$, representing rare and common occurrence of outcome
 - ▶ $X_i|Y_i \sim N(\gamma_0 + \gamma_Y Y_i, \sigma_{X|Y}^2)$ such that $\beta = 0.1, 1$, representing weak and moderate associations
 - ▶ $\lambda = 1/3, 0.5, 2/3$, where λ is the reliability ratio of the error-prone measurements

Logistic regression simulations with $X_i|Y_i$ normal

Scenario	$P(Y_i = 1)$	β	λ	$\hat{\beta}_{RC}$		$\hat{\beta}_{ML}$	
				Mean (SD)	RMSE	Mean (SD)	RMSE
1	0.1	0.1	2/3	0.100 (0.058)	0.058	0.100 (0.058)	0.058
2			1/2	0.101 (0.066)	0.066	0.101 (0.066)	0.066
3			1/3	0.102 (0.083)	0.083	0.102 (0.084)	0.084
4	0.5	1	2/3	0.976 (0.070)	0.074	1.003 (0.075)	0.075
5			1/2	0.966 (0.092)	0.099	1.007 (0.103)	0.104
6			1/3	0.961 (0.140)	0.145	1.019 (0.162)	0.163
7	0.5	0.1	2/3	0.100 (0.035)	0.035	0.100 (0.035)	0.035
8			1/2	0.100 (0.041)	0.041	0.100 (0.041)	0.041
9			1/3	0.101 (0.051)	0.051	0.102 (0.051)	0.051
10	0.5	1	2/3	0.940 (0.052)	0.079	1.003 (0.063)	0.063
11			1/2	0.915 (0.073)	0.113	1.007 (0.096)	0.096
12			1/3	0.897 (0.122)	0.160	1.024 (0.170)	0.172

- ▶ RC shows downward **bias** when $\beta = 1$, especially when $P(Y_i = 1) = 0.5$
- ▶ ML shows little bias, as we would expect

Logistic regression simulations with X_i marginally normal

- ▶ We also performed simulations with X_i marginally normal, with Y_i following a logistic regression given X_i
- ▶ This means X_i is not conditionally normal given Y_i , although for small $P(Y_i = 1)$ or small β , this is approximately true
- ▶ With X_i marginally normal our approach no longer gives the correctly specified MLE

Logistic regression simulations with X_i marginally normal

Scenario	$P(Y_i = 1)$	β	λ	$\hat{\beta}_{RC}$		$\hat{\beta}_{ML}^*$	
				Mean (SD)	RMSE	Mean (SD)	RMSE
1	0.1	0.1	2/3	0.100 (0.057)	0.057	0.100 (0.058)	0.058
2			1/2	0.100 (0.067)	0.067	0.101 (0.067)	0.067
3			1/3	0.102 (0.082)	0.082	0.102 (0.082)	0.082
4	0.5	0.1	2/3	0.966 (0.069)	0.077	0.982 (0.073)	0.075
5			1/2	0.954 (0.091)	0.102	0.987 (0.100)	0.101
6			1/3	0.946 (0.140)	0.150	0.997 (0.160)	0.160
7	0.1	1	2/3	0.100 (0.035)	0.035	0.100 (0.035)	0.035
8			1/2	0.101 (0.040)	0.040	0.101 (0.040)	0.040
9			1/3	0.101 (0.050)	0.050	0.101 (0.051)	0.051
10	0.5	1	2/3	0.938 (0.051)	0.080	1.000 (0.062)	0.062
11			1/2	0.911 (0.071)	0.114	1.002 (0.092)	0.092
12			1/3	0.894 (0.117)	0.158	1.020 (0.163)	0.164

* not ML for the data-generating model

- ▶ RC continues to show **bias** when $\beta = 1$
- ▶ Despite being misspecified, the misspecified MLE for β has limited bias

Inference and extensions

- ▶ Wald type confidence intervals for β can be calculated, based on standard errors for γ_Y and $\sigma_{X|Y}^2$
- ▶ Alternatively, Fieller's theorem can be used to construct confidence intervals
- ▶ The method extends to the case of multivariate \mathbf{X}_i and error-free \mathbf{Z}_i , assuming they are jointly normal given Y_i
- ▶ For non-normal \mathbf{Z}_i , one can fit the mixed model and use it to multiply impute \mathbf{X}_i

Conclusions

- ▶ Linear regression outcome models
 - ▶ Standard linear mixed models can be used to find MLEs for a particular model which include covariate measurement error, when internal replication data are available
 - ▶ The MLE for β under this model is consistent even if the normality assumptions are violated
- ▶ Logistic regression outcome models
 - ▶ The approach can be used for binary outcomes under the normal discriminant model, for which it gives the MLE of β
 - ▶ When X_i is marginally normal, our approach, despite being misspecified, had less bias than RC (c.f. moment reconstruction)
 - ▶ For multivariate X_i , our approach remains tractable, whereas assuming multivariate marginal normality for \mathbf{X}_i causes serious computational difficulties

Extension to longitudinal data

- ▶ The approach can be extended to the more general situation in which measurements follows a linear mixed model

$$\mathbf{W}_i = \mathbf{D}_i \mathbf{X}_i + \mathbf{U}_i$$

- ▶ Therefore certain joint models which include longitudinal repeated measures and a continuous or binary outcome can be fitted using our approach

References

- ▶ Bartlett JW, De Stavola BL, Frost C. Linear mixed models for replication data to efficiently allow for covariate measurement error. *Statistics in Medicine to appear*
- ▶ Freedman LS, Fainberg V, Kipnis V, Midthune D, Carroll RJ. A new method for dealing with measurement error in explanatory variables of regression models. *Biometrics* 2004; 60:172-181
- ▶ Rabe-Hesketh S, Skrondal A, Pickles A. Maximum likelihood estimation of generalized linear models with covariate measurement error. *The Stata Journal* 2003; 3:385-410