

## **USE OF MULTIVARIATE MULTILEVEL MODELS TO HANDLE MISSING DATA IN LONGITUDINAL STUDIES**

Rumana Z Omar

Department of Statistical Science and Joint UCL/H Biomedical Research Unit

University College London

email:r.omar@ucl.ac.uk

Shahed Murad

Department of Mental Health Sciences and UCLH/UCL Biomedical Research Unit

Gareth Ambler

Department of Statistical Science, UCL

Misbah Ahmed

Department of Statistical Science, UCL

Mike King

Department of Mental Health Sciences, UCL

## **Background**

In longitudinal health studies, data are often collected on multiple correlated outcomes to better capture the health of a patient.

e.g. physical & mental health outcomes in cancer patients  
different domains of quality of life scores

## **Missing data**

Missing response is a common problem, particularly for vulnerable subjects..

**Data:**

A cohort of 199 cancer patients in palliative care  
Interviewed up to five times over a 12 month period.

Outcomes: Supportive care need, (measures from Supportive Care Needs Survey) :

- Physical
- Psychological
- Health system

All three outcomes are scores, ranging from 0 to100.

Association with following explanatory variables investigated:

- Cancer network
- GHQ case
- Continuity of care score**
- Treatment phase
- Cancer type (breast, lung, colorectal)
- Time period

Assessments over all time periods for all 3 outcomes not available

## Methods of analysis:

- a) Univariate approach of fitting a separate multilevel model to each outcome using all available measurements.

## Limitations:

Not statistically efficient, ignores outcome correlations and possibility of common predictors.

Does not provide information on outcome correlations.  
important when examining whether outcomes have same aetiology.

Involves multiple significance testing as same predictors may be assessed for each outcome.

Could lead to bias & further loss of efficiency in presence of missing data

## **b) Univariate approach imputing of missing values**

Same as a) but impute missing outcome

### **Problems**

Often regression analysis is performed using crude imputation methods, e.g, last observation carried forward

Multiple Imputation not well developed for repeated measurement studies

Multiple imputation based on multilevel models have been developed more recently (Carpenter & Goldstein, 2004)

Have computational problems, particularly for large datasets with many variables of different types.

Assume multivariate normality requiring a latent variable approach to handle binary, categorical and nominal variables.

### **c) Multivariate Models**

Multivariate multilevel models have been proposed to handle multiple correlated outcomes.  
(Goldstein)

Handle all outcomes in a single modelling framework accounting for correlation between outcomes

Allow common as well as separate coefficients

Allows efficient estimation of coefficients if one of the outcomes were missing under MAR assumptions exploiting the correlations

## Multivariate models

To define the multivariate structure, 2 level model is extended to a 3<sup>rd</sup> level.

A new lower level sets the number of response variables.

Level 1: response, level 2: time period, level 3:subject.

e.g. a Bivariate model for a repeated measurement study with 1 predictor

$$Y_{1jk} = \beta_{01} Z_{1ijk} + \beta_{11} X_{jk} Z_{1ijk} + u_{0jk} Z_{1ijk}$$

$$Y_{2jk} = \beta_{02} Z_{2ijk} + \beta_{12} X_{jk} Z_{2ijk} + u_{1jk} Z_{2ijk}$$

$$\begin{bmatrix} u_{0jk} \\ u_{1jk} \end{bmatrix} \sim N(0, \Omega_u) : \Omega_u = \begin{bmatrix} \sigma_{u0}^2 & \\ \sigma_{u01}^2 & \sigma_{u1}^2 \end{bmatrix}$$

$$Z_{1ijk} = \begin{cases} 1 & \text{if response}=1 \\ 0 & \text{if response}=2 \end{cases} \quad Z_{2ijk} = \begin{cases} 1 & \text{if response}=2 \\ 0 & \text{if response}=1 \end{cases}$$

$$\begin{bmatrix} u_{0jk} \\ u_{1jk} \end{bmatrix} \sim N(0, \Omega_u) : \Omega_u = \begin{bmatrix} \sigma_{u0}^2 & \\ \sigma_{u01}^2 & \sigma_{u1}^2 \end{bmatrix}$$

## Objective:

Based on simulation studies compare bias in estimates of regression coefficients for MAR data for:

- i) Univariate multilevel models
- ii) Multivariate multilevel models
- iii) Univariate + Multiple imputation based on chained equation method using (van Buuren, implemented by Royston in Stata):

all repeated measurements of response as covariates to impute missing response.

other responses also used as covariates

Then univariate models to estimate coefficients after imputation

Random intercept models used for all 3 methods

## Simulation design:

A multivariate multilevel model was applied to the original data

*Cancer network, cancer type, treatment phase & period* have common coefficients for all outcomes

*ghqcase* and *continuity of care* have separate coefficients.

Estimates of fixed parameters were used as true values.

4 scenarios of correlation between outcomes and levels of missingness were considered.

High correlation & high missingness (HH),

High correlation & low missingness (HL),

Low correlation & high missingness (LH)

Low correlation & low missingness (LL).

Low & high correlation were approximately 0.3 & 0.5.

Low & High missingness were approximately 20% and 50%.

New sets of responses were generated using the coefficients from the real data & sampling the between and within subject residuals from a multivariate normal distribution.

MAR mechanism was induced in the data

Two MAR mechanisms considered.

- 1) Missingness depended on covariates
- 2) Missingness of 2 outcomes depended on the third outcome,  
e.g. missingness of psychological & health system need scores were dependent on physical need score

For each of the 8 scenarios, 100 data sets were simulated.

## Bias: Fixed parameters

**TABLE 1: Simulation part 1 – Physical need**

		<b>High Correlation and High Missingness (HH)</b>		<b>Low Correlation and high Missingness (LH)</b>	
		<b>Abs. Bias (95% Interval)</b>	<b>Rel. Bias (95% Interval)</b>	<b>Abs. Bias (95% Interval)</b>	<b>Rel. Bias (95% Interval)</b>
<b>MI</b>	<b>Ccare</b>	0.013 (0.007 to 0.018)	0.001 (0.001 to 0.002)	0.016 (0.010 to 0.021)	0.002 (0.001 to 0.003)
<b>Mvariate</b>		0.005 (0.000 to 0.010)	0.001 (0.000 to 0.001)	0.005 (0.001 to 0.010)	0.001 (0.000 to 0.001)
<b>Uvariate</b>		0.006 (0.001 to 0.011)	0.001 (0.000 to 0.001)	0.005 (0.001 to 0.010)	0.001 (0.000 to 0.001)
<b>MI</b>	<b>GHQ</b>	-0.117(-0.215 to -0.020)	-0.245(-0.448 to -0.041)	-0.108 (-0.218 to 0.002)	-0.225 (-0.456 to 0.005)
<b>Mvariate</b>		-0.063 (-0.149 to 0.023)	-0.131 (-0.311 to 0.049)	0.023 (-0.073 to 0.120)	0.049 (-0.153 to 0.251)
<b>Uvariate</b>		-0.049 (-0.136 to 0.039)	-0.101 (-0.284 to 0.081)	0.019 (-0.079 to 0.117)	0.040 (-0.165 to 0.245)
<b>MI</b>	<b>Per</b>	0.060 (0.023 to 0.098)	0.043 (0.016 to 0.069)	0.075 (0.039 to 0.111)	0.053 (0.028 to 0.079)
<b>Mvariate</b>		0.009 (-0.015 to 0.034)	0.007 (-0.011 to 0.024)	0.011 (-0.008 to 0.029)	0.008 (-0.005 to 0.021)
<b>Uvariate</b>		0.012 (-0.016 to 0.039)	0.008 (-0.011 to 0.028)	0.001 (-0.023 to 0.026)	0.001 (-0.017 to 0.018)

95% intervals overlap, MI interval does not include zero for GHQ.

Similar results for other outcomes, similar results for relative bias

**TABLE 2: Simulation part 2 – Psychological need**

		<u>High Correlation and High Missingness (HH)</u>		<u>Low Correlation and high Missingness (LH)</u>	
		Abs. Bias (95% Interval)	Rel Bias (95% Interval)	Abs. Bias (95% Interval)	Rel. Bias (95% Interval)
MI	Ccare	-0.011(-0.019 to -0.002)	-0.0009(-0.0015 to -0.0002)	0.014 (0.003 to 0.025)	0.0011 (0.0002 to 0.0020)
Mvariate		0.000 (-0.005 to 0.005 )	0.0000(-0.0004 to 0.0004)	0.004 (-0.002 to 0.010)	0.0003 (-0.0002 to 0.0008
Uvariate		0.058 (0.052 to 0.064)	0.0047 (0.0042 to 0.0051)	0.003 (-0.003 to 0.009)	0.0002 (-0.0002 to 0.0007)
MI	GHQ	-0.030 (-0.169 to 0.109)	-0.061 (-0.341 to 0.220)	-0.092 (-0.247 to 0.063)	-0.185 (0.456 to 0.126)
Mvariate		-0.052 (-0.162 to 0.056)	-0.105 (-0.325 to 0.116)	-0.106(-0.237 to 0.025 )	-0.214 (-0.477 to 0.050)
Uvariate		1.346 (-1.451 to -1.242)	--2.710 (-2.921 to -2.500)	-0.118 (-0.244 to 0.008)	-0.238 (-0.492 to 0.015)
MI	Per	0.059 (-0.136 to 0.019)	-0.041-0.096 to 0.013)	0.024 (-0.057 to 0.105)	0.017 (-0.040 to 0.074)
Mvariate		-0.003 (--0.016 to 0.010)	-0.002 (-0.012 to 0.007)	-0.002 (-0.041 to 0.038)	0.000 (-0.008 to 0.008)
Uvariate		0.204 (0.169 to 0.239)	0-.144 (0.120 to 0. 0.169)	0.001 (-0.023 to 0.026)	-0.001 (-0.029 to 0.027)

For high correlation, Multivariate performed best & Univariate worst for all outcomes.

## Bias – Random Parameters

For all outcomes,

HH & LH, simulation part 1:

All 3 methods underestimate  $\sigma^2_v$ .

Largest bias produced by MI with non overlapping intervals with other 2 methods.

$\sigma^2_e$  is overestimated by MI.

For both random parameters multivariate method performed best.

HH, simulation part 2:

All 3 methods underestimate  $\sigma^2_u$ .

with largest bias produced by Univariate with non overlapping intervals with other 2 methods.

*LH, simulation part 2: All methods performed similarly with small bias, but MI over estimates  $\sigma^2_e$*

## Summary

Covariate dept. missingness: Univariate performs similarly to Multivariate & MI for bias.

Response dept. missingness: For correlation of 0.5 Multivariate performed best. Univariate did not perform so well.

Multivariate model has advantage of borrowing strength through the covariance matrix.

MI takes account of correlation between outcomes by considering the non missing outcomes as covariates when imputing the missing outcome.

Multivariate models produces more efficient estimates in presence of missing data.

## Recommendations:

For multiple correlated outcomes in longitudinal studies with non-overlapping missingness, multivariate multilevel models should be used more often.

## Further work

increase correlations, different patterns of missingness, comparison with other imputation methods