Analysis of data when missing values are substituted by multiple imputation

Jochen Hardt, Nanny Wermuth

GMMC Scientific Board Meeting, January 9-11, 2008

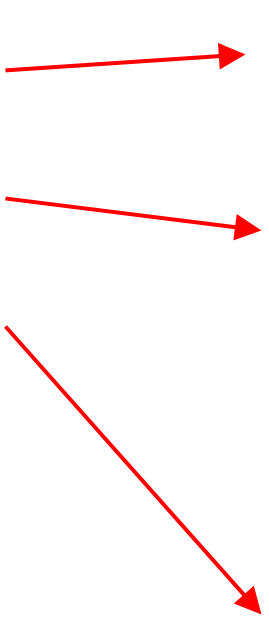
Göteborg University

Background I

- almost all sociological / medical data have missings typically in the range of .5 to 5 % in a variable
- this is particularly the case in particularly longitudinal data, where often single time points are missing
- simple substitution (means, last observation carried forward) reduces error variances
- more sophisticated methods for substitution are available, such as multiple imputation

Original dataset

#	Y	X1	X2	X3
1	0	0	0	0
2	0	0	0	1
3	1	1	0	-
4	0	0	1	0
5	0	0	1	-
6	1	0	0	0
7	1	-	-	-
8	1	1	0	1



Augmented dataset

Y	X1	X2	X3	wt
0	0	0	0	1
0	0	0	1	1
1	1	0	0	w31
1	1	0	1	w32
0	0	1	0	1
0	0	1	0	w51
0	0	1	1	w52
1	0	0	0	1
1	0	0	0	w71
1	0	0	1	w72
1	0	1	0	w73
1	0	1	1	w74
1	1	0	0	w75
1	1	0	1	w76
1	1	1	0	w77
1	1	1	1	w78
1	0	1	0	1

From: Horton, N. J & Kleinmann, K. P. (2007). Much ado about nothing: A comparison of missing data methods and software ti fit incomplete data regression models. Am Statist, 61, 79-90.

Background II

- They assume Missing At Random
- This is usually not the case in real data. Missings often are a mixture of random and non-random processes
- Previous studies exploring the effects of multiple imputation relied on substitutions of real missings
- We used datasets without missings, and set some values by random to missing - just to substitute them via multiple imputation

A first simulation was performed on a questionnaire containing 25 items with very good results, i.e. factor analysis revealed almost identical coefficients

OPEN ACCESS

Research Article

Suicide attempts and retrospective reports about parent-child relationships: evidence for the affectionless control hypothesis

Suizidversuche und retrospektive Berichte über Eltern-Kind-Beziehungen: Evidenz für die "affectionless control" Hypothese

Abstract

Objective: To compare the characteristics of recalled parent-child relationships in suicide attempters vs. non-attempters

Methods: A total of 509 patients – 249 presenting with pain at a psychosomatic clinic and 260 from the offices of general practitioners – were interviewed retrospectively about suicide attempts and parent-child relationships.

Results: The overall rate of those reporting a suicide attempt was 17%. Bivariate analyses showed associations of poor parent-child relationships

J. Hardt[1]

U.T. Egle[2]

J.G. Johnson[3]

1 Clinic for Psychosomatic Medicine and Psychotherapy,

Suicide attempts were predicted by 3 out of 20 scores (continous variables) concerning the mother:

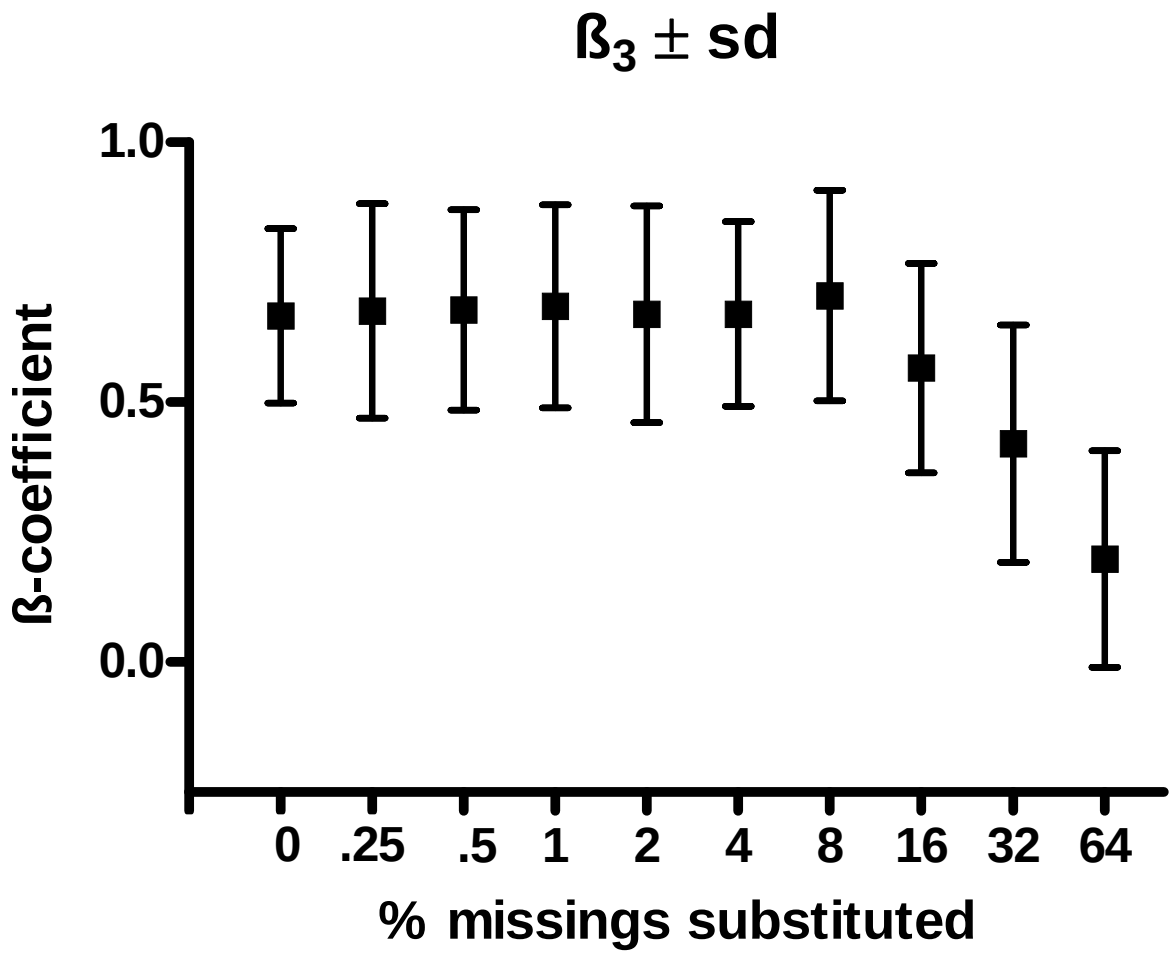
Var	ß	sd
X1: Love	-.73	.19
X2: Control	.09	.21
X3: Role Reversal	.67	.19

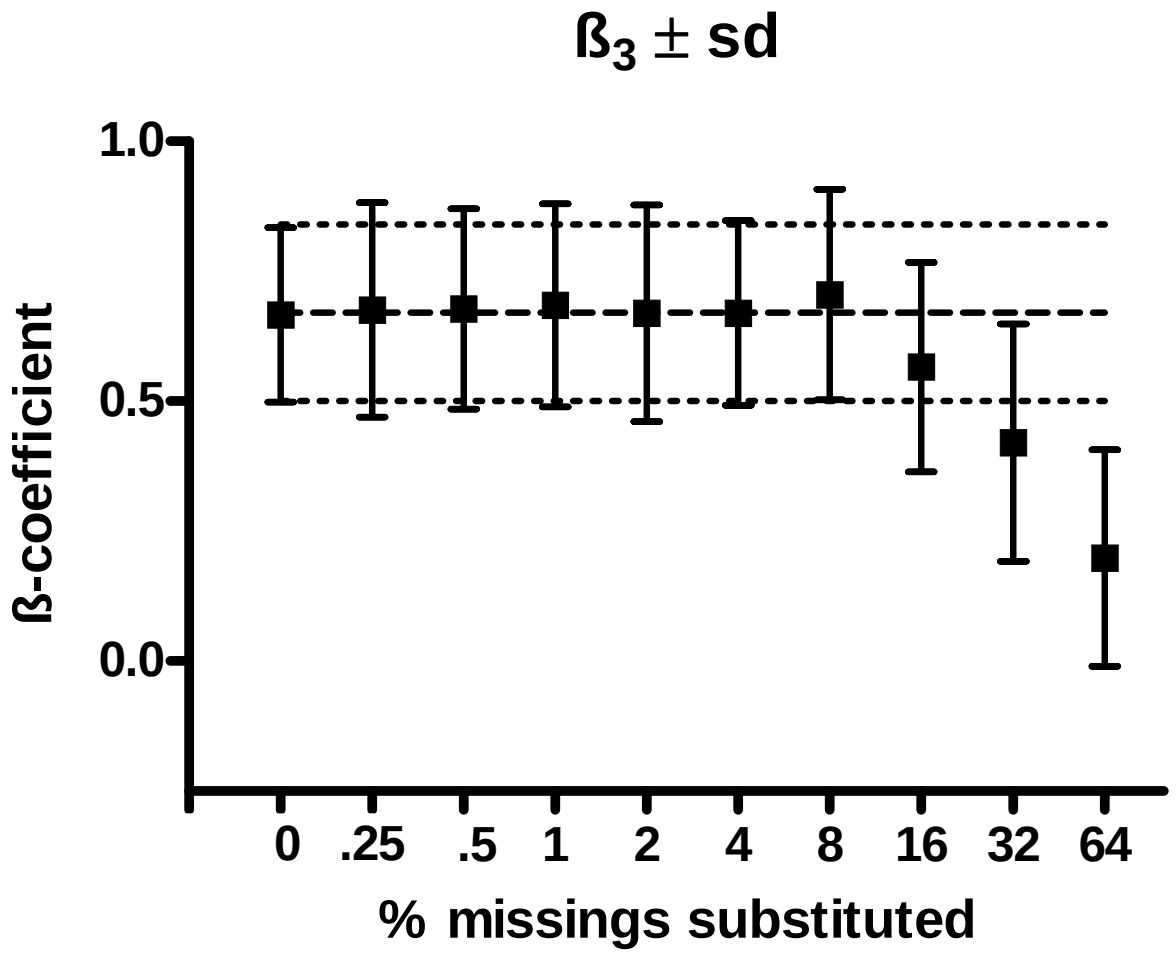
Response: Lifetime suicide attempt

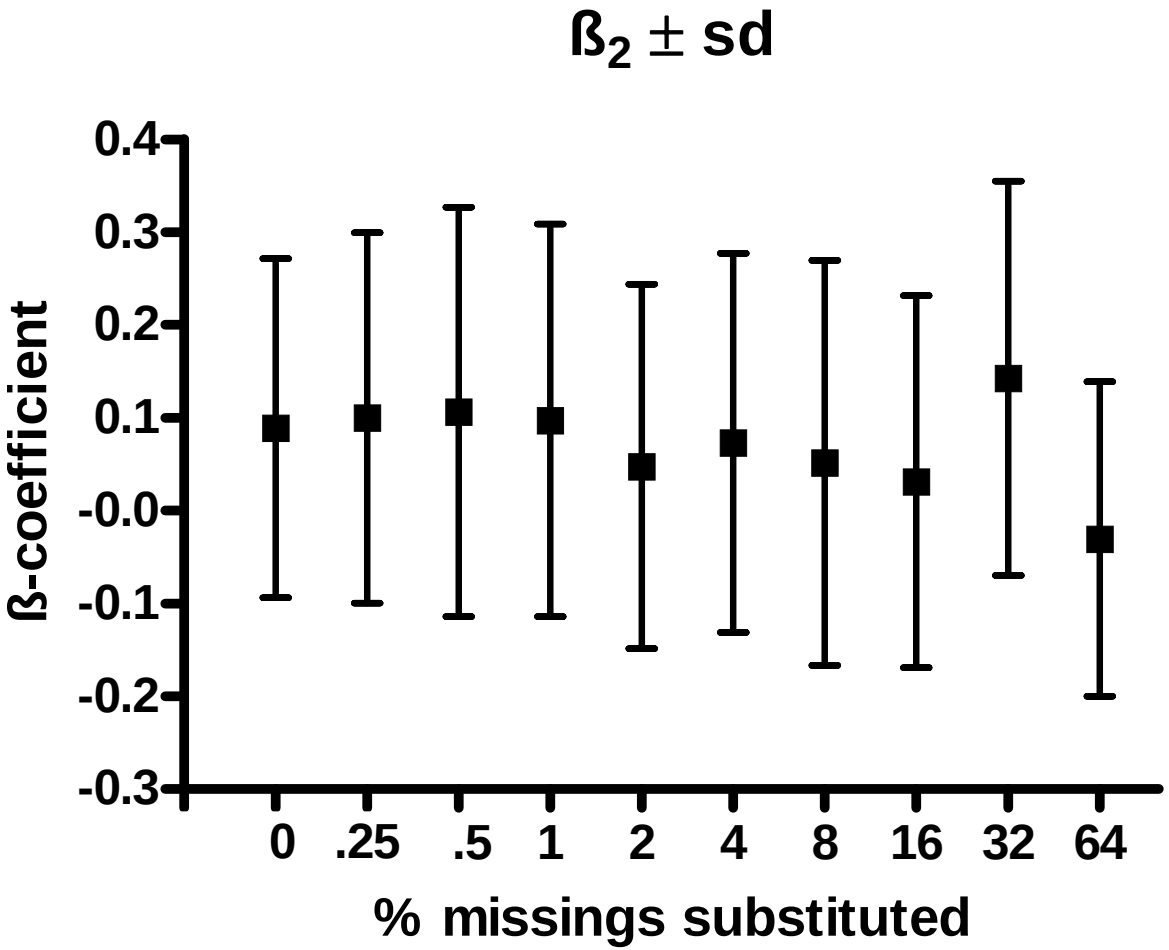
0 = no

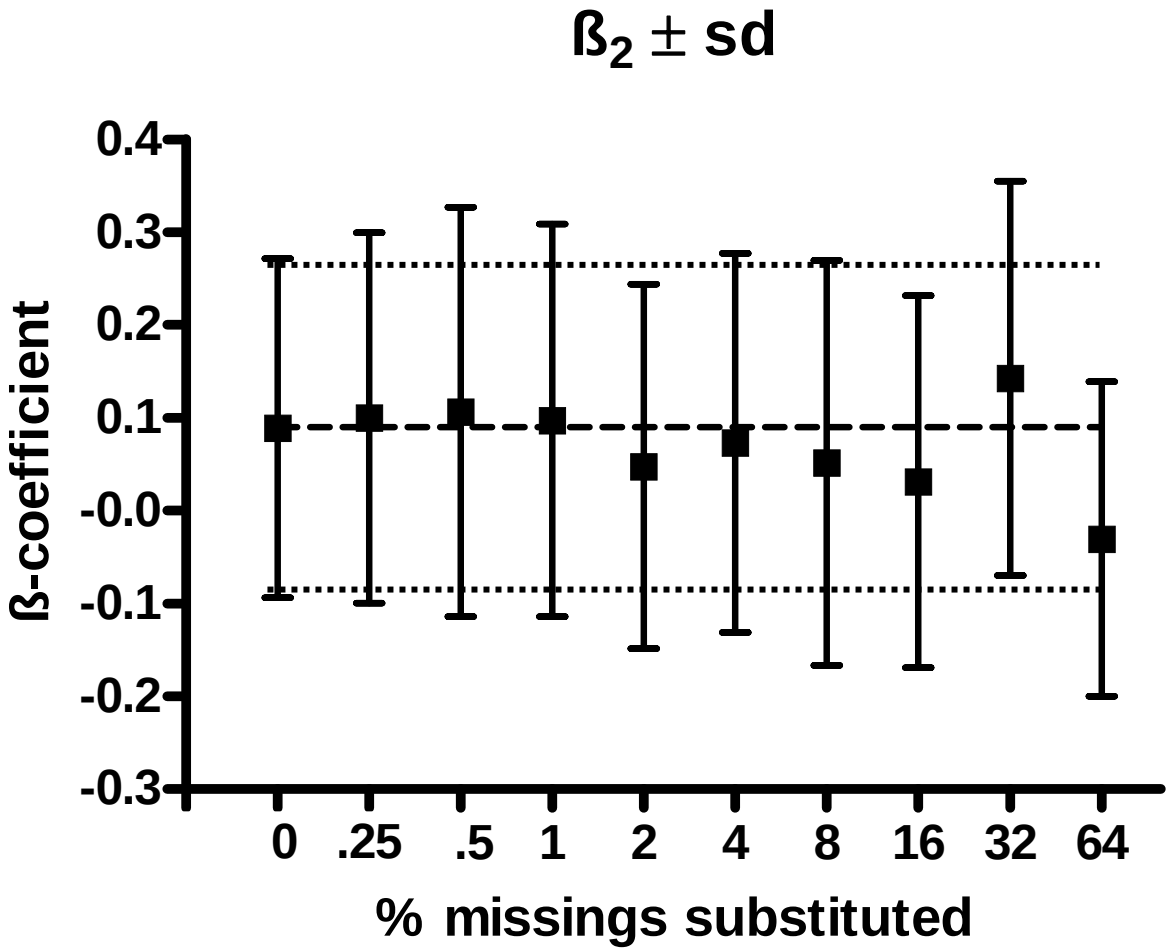
1 = yes

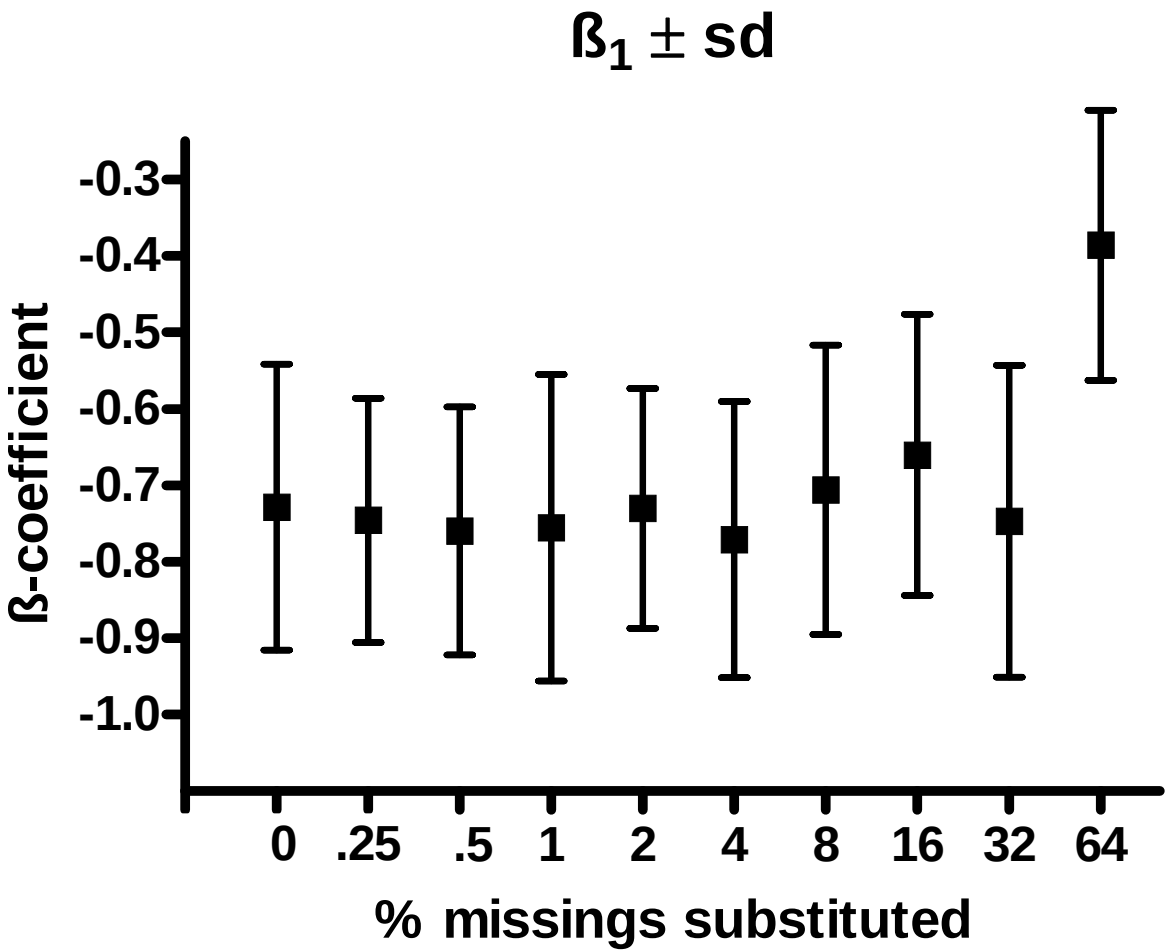
N = 505

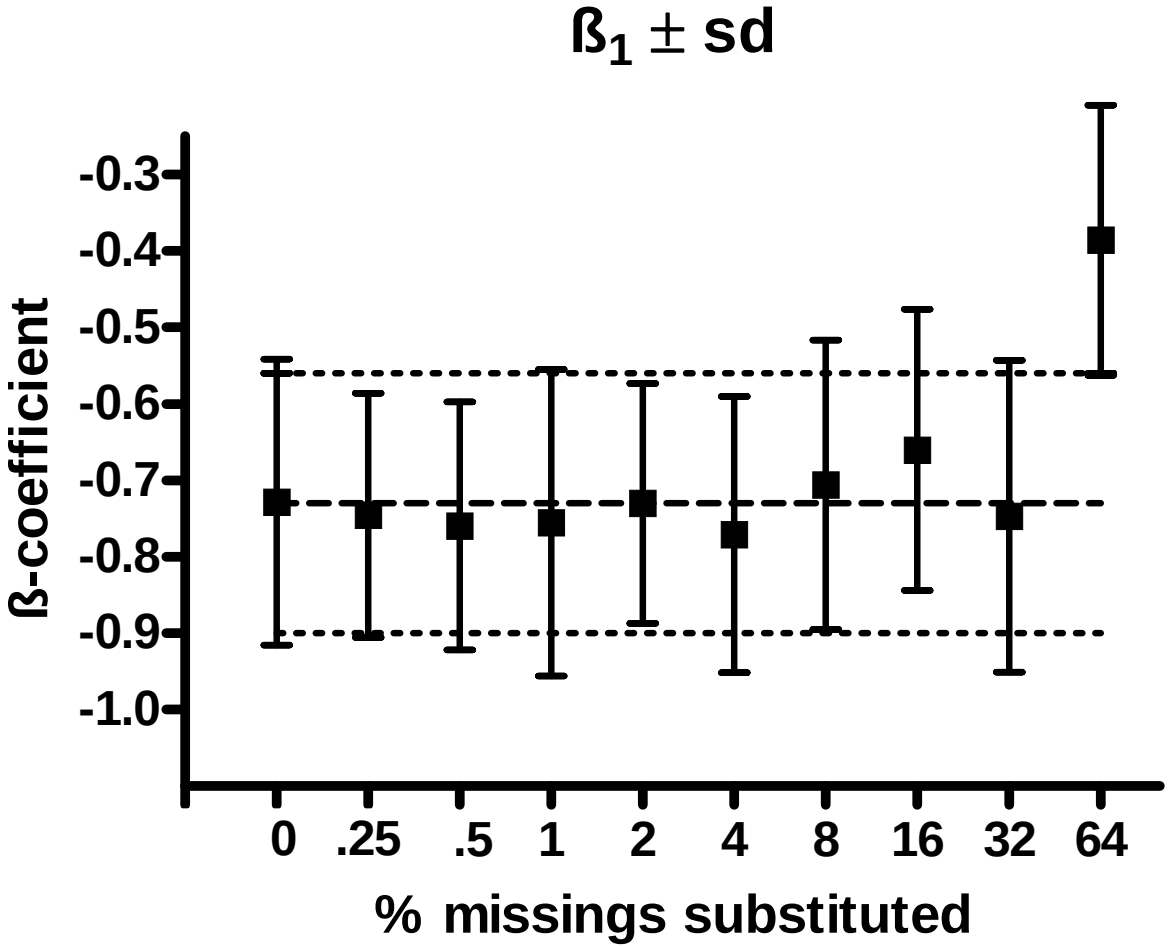

ß3 ± sd
1.0
0.5
0.0
ß-coefficient
0
.25
.5
1
2
4
8
16
32
64
% missings substituted

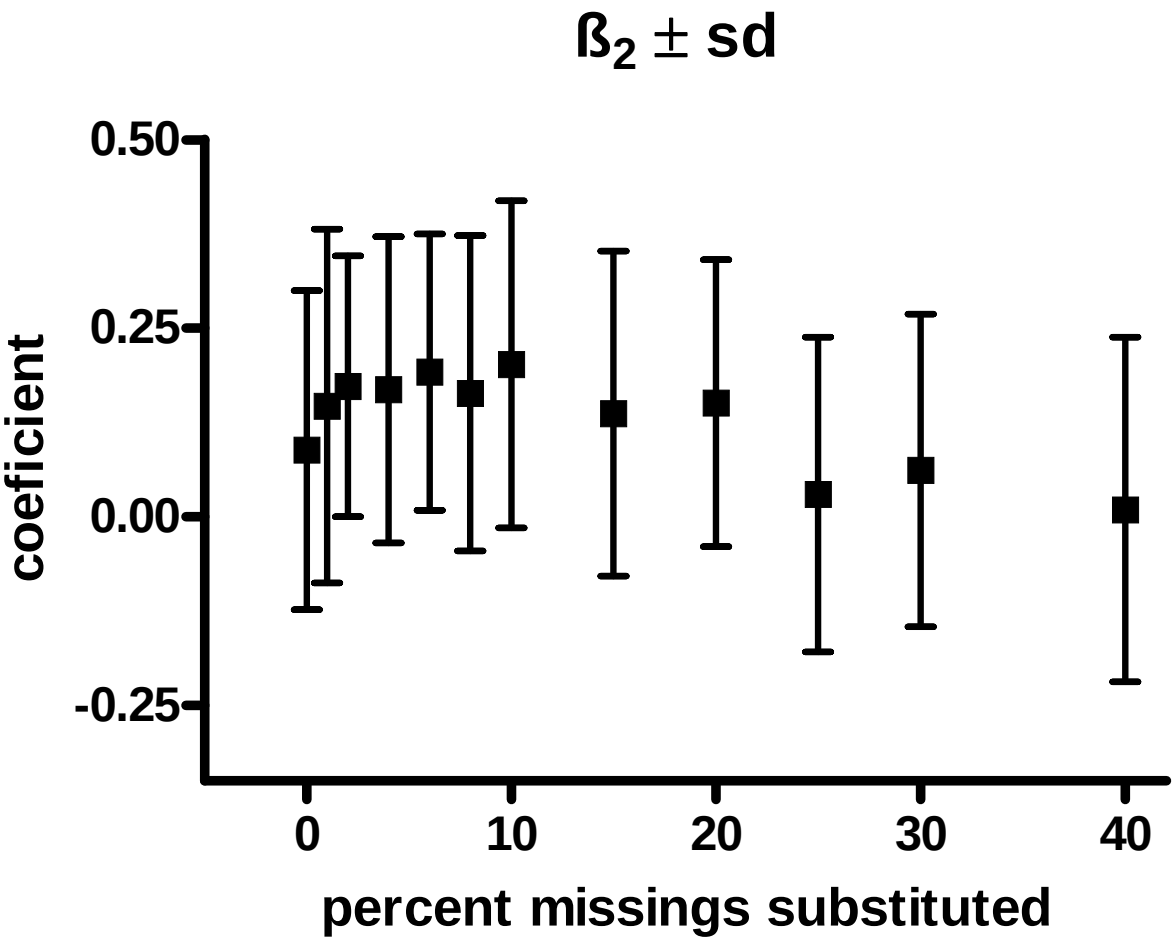

ß3 ± sd
1.0
0.5
0.0
ß-coefficient
0
.25
.5
1
2
4
8
16
32
64
% missings substituted


ß2 ± sd
0.4
0.3
0.2
0.1
-0.0
-0.1
-0.2
-0.3
ß-coefficient
0
.25
.5
1
2
4
8
16
32
64
% missings substituted


ß2 ± sd
0.4
0.3
0.2
0.1
-0.0
-0.1
-0.2
-0.3
ß-coefficient
0
.25
.5
1
2
4
8
16
32
64
% missings substituted


ß1 ± sd
-0.3
-0.4
-0.5
-0.6
-0.7
-0.8
-0.9
-1.0
ß-coefficient
0
.25
.5
1
2
4
8
16
32
64
% missings substituted


ß1 ± sd
-0.3
-0.4
-0.5
-0.6
-0.7
-0.8
-0.9
-1.0
ß-coefficient
0
.25
.5
1
2
4
8
16
32
64
% missings substituted


ß2 ± sd
0.50
0.25
0.00
-0.25
coeficient
0
10
20
30
40
percent missings substituted

Preliminary Summary

- A first result:
- substituting some single items for factor analysis does not change the results

Factors to be explored further

- is there a systematic overestimation of the coefficients in datasets when small percentages of missings are substituted?
- if so, which factors contribute to the overestimation?
- are these “jumping” coefficients simply errors in my simulation program?

Plan to procede

- Factors that should be varied
- % missing
- sample size
- substitution algorithm
- number of multiple data lines
- number of bootstraps necessary
- number of variables in the dataset
- complexity of the model

….

Thank You !