LINCE – Fernando Morelli – 06.10.2020

The proportion of missing data should not be used to guide decisions on multiple imputation

Madley-Dowd P, Hughes R, Tilling K, Heron J.
J Clin Epidemiol. 2019;110:63-73.
doi:10.1016/j.jclinepi.2019.02.016

Agenda

1

Introdução

Sobre os dados faltantes e como avalia-los

2

Simulações

CCA versus IM

3

Exemplo Real

Avon Longitudinal Study of Parents and Children

1. Introdução

Sobre os dados faltantes

Dados Faltantes

Missing data (em português, dados faltantes ou não-resposta) são um problema recorrente na epidemiologia, podendo causar redução no tamanho da amostra, viés na análise e redução da eficiência.

Mecanismos de ‘não-resposta’ de Rubin

Missing completely at random (MCAR)

Quando a probabilidade de não resposta em uma variável X é não relacionada a outras variáveis mensuráveis e ao valor de X.

Missing at random (MAR)

Quando a probabilidade de não resposta em uma variável X é relacionada a outras variáveis mensuráveis, mas não ao valor da variável não respondida.

Missing not at random (MNAR)

Quando a probabilidade de não resposta é sistematicamente relacionada aos valores que estão faltando.

Referência: Baraldi AN, Enders CK. An introduction to modern missing data analyses. *J Sch Psychol.* 2010;48(1):5-37. doi:10.1016/j.jsp.2009.10.001

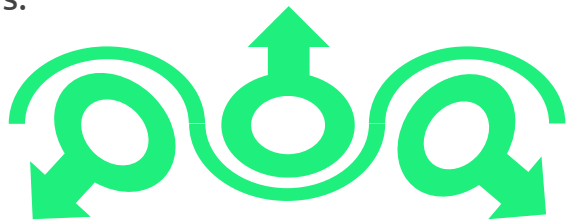
Métodos de Deleção

- Técnica mais básica
- Pode ser ‘listwise deletion’ (ou ‘complete-case analysis’) ou ‘pairwise deletion’ (ou ‘available-case analysis)
- Vantagem: produz uma base de dados completa
- Desvantagem: reduz o tamanho da amostra; supõe que os dados são MCAR

Referência: Baraldi AN, Enders CK. An introduction to modern missing data analyses. *J Sch Psychol.* 2010;48(1):5-37. doi:10.1016/j.jsp.2009.10.001

Métodos de Imputação Múltipla

2 Uma análise é conduzida em cada base de dados simulada previamente, utilizando as mesmas técnicas que seriam usadas embases normais e distintas.



1 Criação de múltiplas bases de dados, cada uma com um valor diferente imputado para a variável não-respondida. Existem diversos algoritmos para fazer essa etapa.

3 Todos os valores únicos identificados são unificados em uma estimativa e um erro únicos, usualmente através de uma simples média .

Métodos de Imputação Múltipla

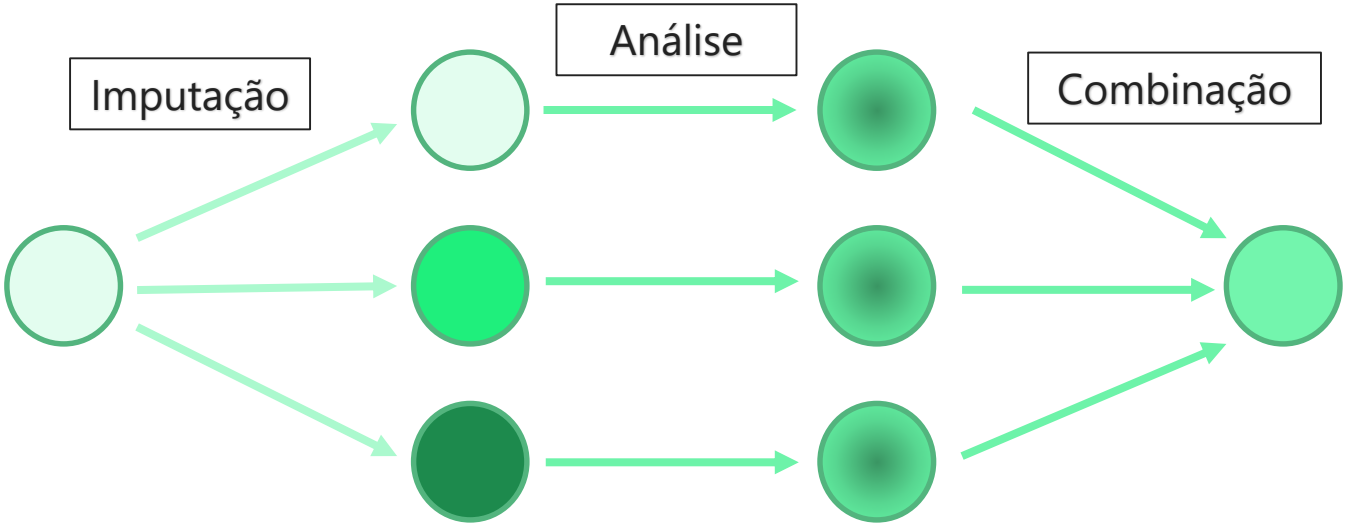


Figura adaptada de Nunes, Luciana Neves, Klück, Mariza Machado, & Fachel, Jandyra Maria Guimarães. (2010). Comparação de métodos de imputação única e múltipla usando como exemplo um modelo de risco para mortalidade cirúrgica. Revista Brasileira de Epidemiologia, 13(4), 596-606. https://doi.org/10.1590/S1415-790X2010000400005

E como podemos avaliar os dados faltantes?

Taxa de Resposta (Response Rate)

A **taxa de resposta**, também conhecida como taxa de conclusão ou taxa de retorno, é o número de pessoas que responderam à pesquisa dividida pelo número de pessoas na amostra.

Geralmente é expresso na forma de uma porcentagem.

A **proporção de dados faltantes** deriva dessa medida, e é calculada como uma proporção de dados faltantes sobre dados completos.

Fração de Dados Faltantes (Fraction of Missing Information)

A **fração de dados faltantes** é uma medida que deriva da teoria da IM, sendo calculada através de um modelo que relaciona a base de dados completa com os indicadores faltantes, criando uma razão entre a variância da IM e a variância total.

É expresso em números de 0 a 1 (quanto mais próximo de 1, maior a variabilidade).

Referência: Wagner, J. (2010). The Fraction Of Missing Information As A Tool For Monitoring The Quality Of Survey Data. The Public Opinion Quarterly, 74(2), 223-243

E como podemos avaliar os dados faltantes?

Taxa de Resposta (Response Rate)

Forças:
- Fácil de calcular
- Fácil de interpretar
- Possível calcular em “sub-amostras”

Fraquezas:
- Não ligado diretamente a viéses;
- Não especifica as variáveis
- Pode distorcer práticas de coletas

Fração de Dados Faltantes (Fraction of Missing Information)

Forças:
- Calculado com base em todos os dados da pesquisa
- Encoraja o desenvolvimento de paradados (“paradata”) robustos

Fraquezas:
- Difícil de ser calculado
- Necessita de base de dados forte
- Difícil de ser interpretado

Referência: Wagner, J. (2010). The Fraction Of Missing Information As A Tool For Monitoring The Quality Of Survey Data. The Public Opinion Quarterly, 74(2), 223-243

Pergunta introdutória

A partir de qual proporção de não-resposta é que devemos considerar fazer uma imputação múltipla?

% de dados

2. Simulações

CCA vs IM

Entendendo a Simulação

Baseado em uma coorte prospectiva onde todos os dados de base são presentes, mas alguns dados de follow-up estão faltando.

Variáveis numéricas contínuas, com média 0 e com desvio padrão de 1

Y e X: 0.6
Y e Z_1 e Z_2 :0.4
Y e $Z_3 - Z_7$:0.2
Y e $Z_8 - Z_{11}$:0.1

Cada base simulada contem 1.000 observações simuladas para as variáveis Y, X e $Z_1 - Z_{11}$

Y: variável de desfecho
X: variável de exposição
Z_1 – Z_{11}: variáveis auxiliares

Entendendo a Simulação

Missingness foi simulada como MCAR e como MAR

Cada conjunto de dados simulados foi analisado utilizando CCA e MI

Foram realizadas 1.000 imputações por modelo.

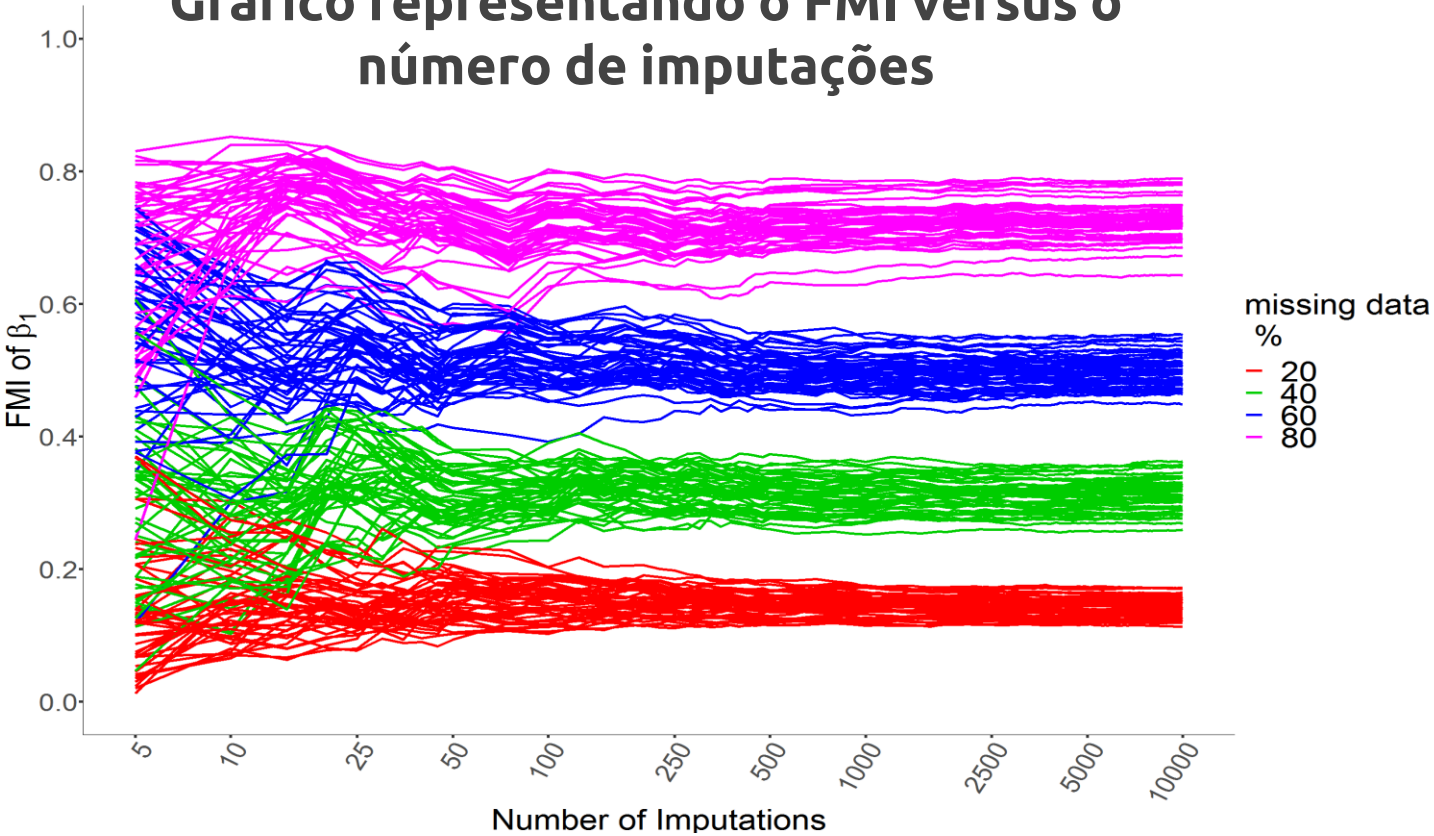
A análise dos dados se deu através de um modelo de regressão linear

Os modelos de imputação variaram entre 1 e 5, onde 1 não continha informações auxiliares (Z_1 – Z_{11}) e 5 continha todas as informações.

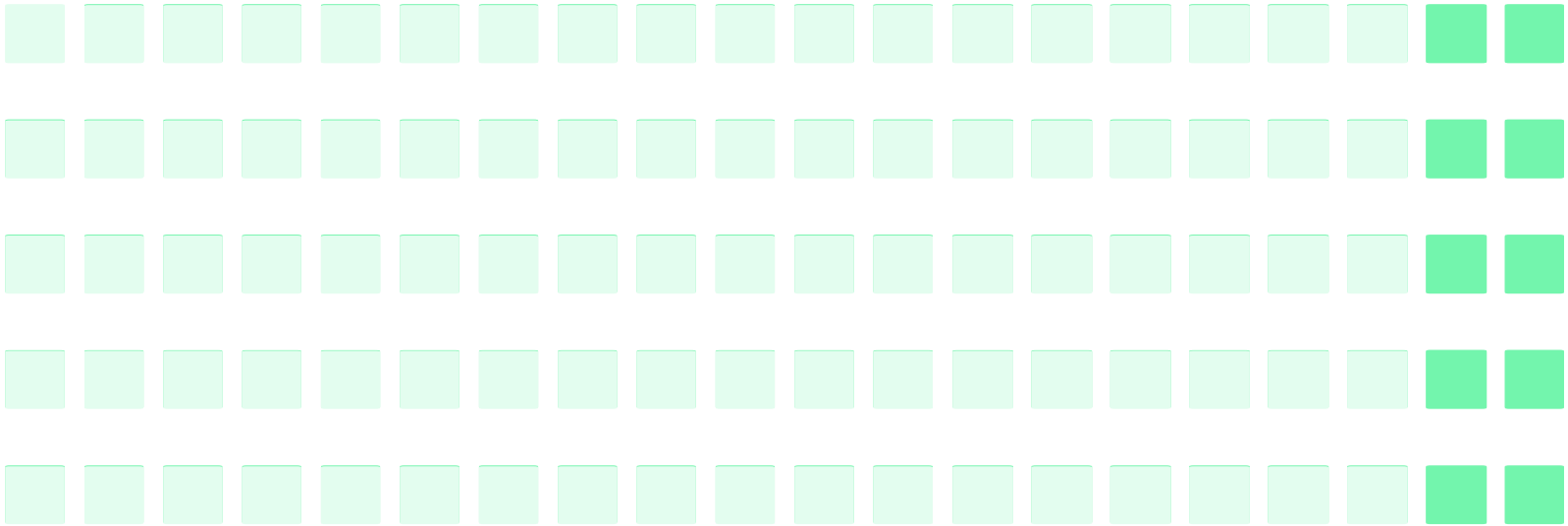
Descrição dos modelos de imputação utilizados para mecanismos MCAR e MAR

Modelo de Imputação	Variáveis Incluídas	R_Y^2
1	Y, X	0.36
2	Y, X, Z_3	0.40
3	Y, X, Z_1	0.52
4	Y, X, $Z_{1–4}$	0.76
5	Y, X, $Z_{1–11}$	0.92

Gráfico representando o FMI versus o número de imputações

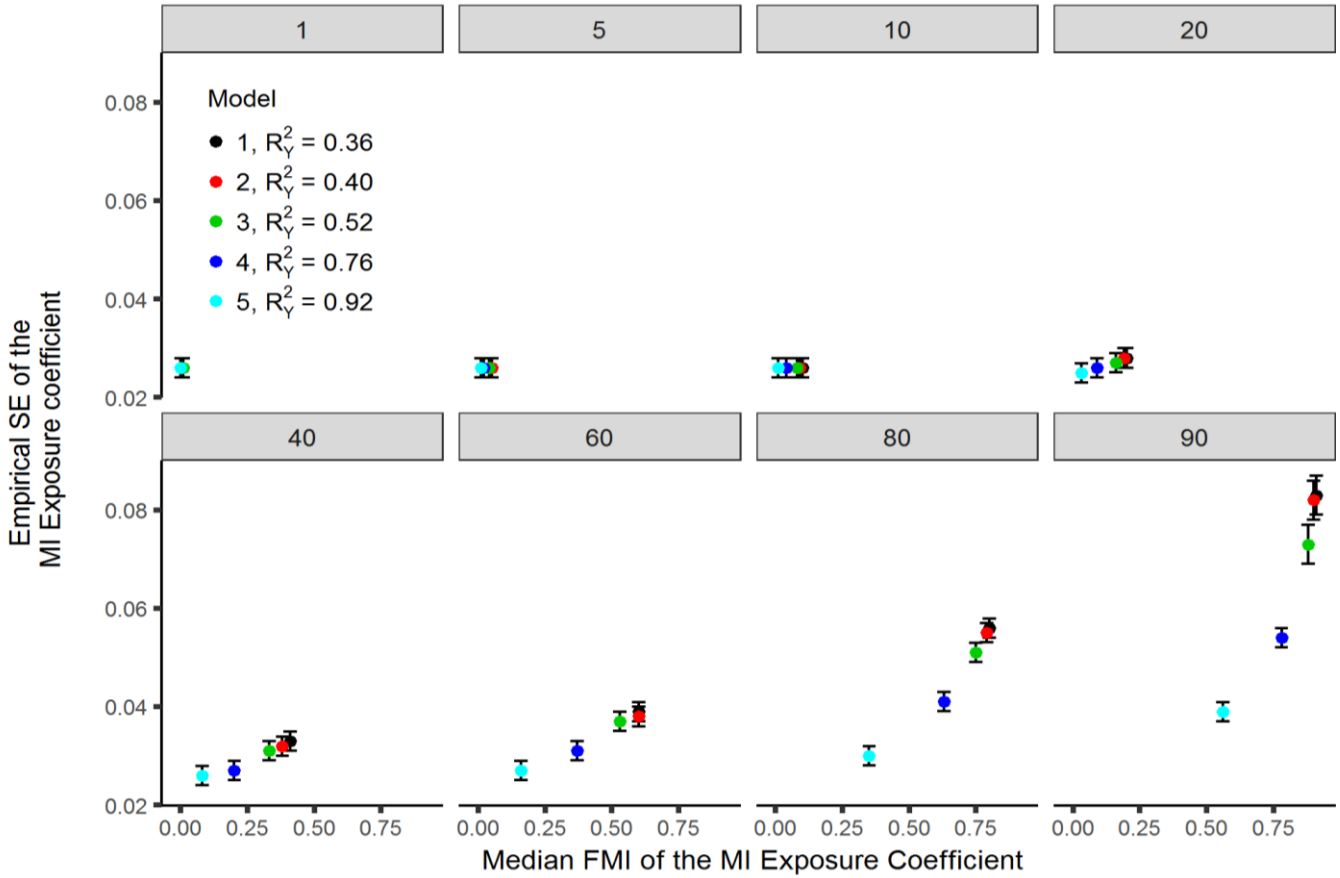


Entendendo a Simulação



% de dados

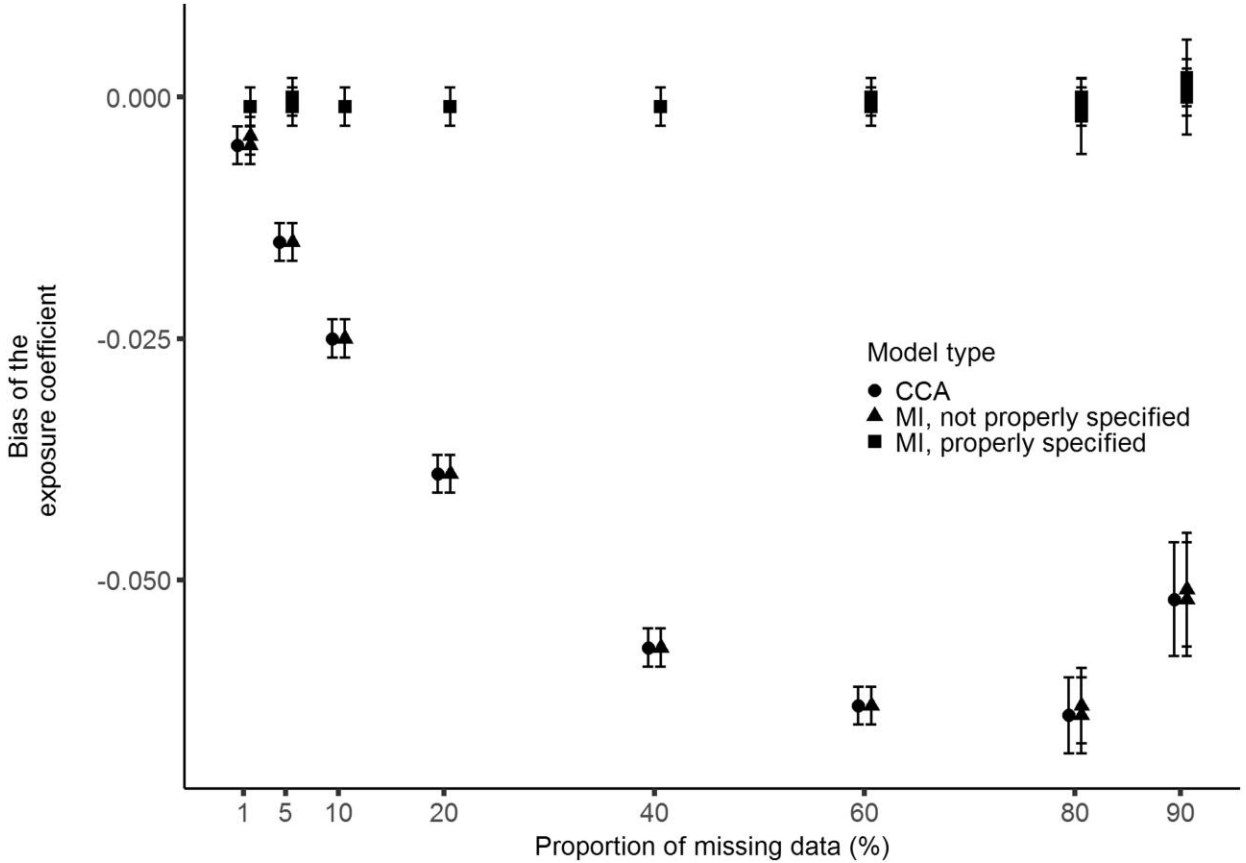
Resultados


1
5
10
20
40
60
80
90
Model
1, R_Y^2 = 0.36
2, R_Y^2 = 0.40
3, R_Y^2 = 0.52
4, R_Y^2 = 0.76
5, R_Y^2 = 0.92
Empirical SE of the MI Exposure coefficient
0.02
0.04
0.06
0.08
0.00
0.25
0.50
0.75
Median FMI of the MI Exposure Coefficient

Resultados

		% Redução no Erro Padrão comparado com CCA		% Redução de vies comparado com CCA
% Dados Faltantes	Modelo	MCAR	MAR	Viés em MAR
60	1: R2 = 0.36	−0.04%	−0.02%	0.21%
	2: R2 = 0.40	2.55%	1.68%	0.02%
	3: R2 = 0.52	5.48%	6.74%	99.77%
	4: R2 = 0.76	21.02%	18.45%	99.43%
	5: R2 = 0.92	31.59%	31.96%	98.22%
80	1: R2 = 0.36	−0.03%	−0.14%	0.00%
	2: R2 = 0.40	2.16%	1.57%	1.34%
	3: R2 = 0.52	8.18%	9.86%	96.47%
	4: R2 = 0.76	27.56%	28.21%	99.62%
	5: R2 = 0.92	45.88%	44.66%	98.77%
90	1: R2 = 0.36	0.03%	0.11%	0.04%
	2: R2 = 0.40	1.40%	2.18%	0.89%
	3: R2 = 0.52	12.44%	8.86%	99.97%
	4: R2 = 0.76	34.82%	33.76%	95.78%
	5: R2 = 0.92	53.09%	52.96%	98.73%

Resultados


0.000
-0.025
-0.050
Bias of the exposure coefficient
1
5
10
20
40
60
80
90
Proportion of missing data (%)
Model type
CCA
MI, not properly specified
MI, properly specified

3. Exemplo real

Avon Longitudinal Study of Parents and Children

Entendendo o Exemplo

Exposição: Tabagismo materno na gestação

Desfecho: QI da prole aos 15 anos

Possíveis confusores: idade materna, paridade, educação e sexo da prole

Variáveis auxiliares para a imputação múltipla: QI aos 8 anos de idade, fluência e inteligibilidade aos 9 anos de idade, um indicador binário sobre ter dificuldades de aprendizagem, relatos de professores e nota de um teste de matemática medido no 6º ano

Entendendo o Exemplo

Para categorizar os exemplos faltantes, foram identificados 4 padrões

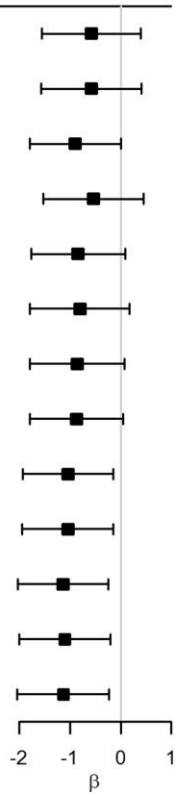
Padrão	Exposição	Desfecho	Variável Auxiliar com melhor poder de predição	N(%)
1	Completo	Missing	Missing	4,803 (40.32)
2	Completo	Completo	Completo	3,974 (33.36)
3	Completo	Missing	Completo	2,698 (22.65)
4	Completo	Completo	Missing	496 (4.16)

Resultados

Variable	Type	% Missing data	R^2 with Outcome[a]	OR for missing data in outcome[b]	95% CI[b]
IQ at age 15	Continuous	62.47			
Maternal smoking in pregnancy	Binary	0.00	0.01	2.18	1.98, 2.39
Maternal age	Categorical	0.00	0.04		
	≤ 24 years			Reference	Reference
	25–29 years			0.57	0.51, 0.64
	30–34 years			0.42	0.38, 0.47
	≥ 35 years			0.41	0.35, 0.47
Parity	Categorical	0.00	0.01		
	0			Reference	Reference
	1			1.18	1.09, 1.29
	2			1.46	1.30, 1.64
	≥ 3			2.06	1.72, 2.48
Sex	Binary	0.00	<0.01		
	Female			Reference	Reference
	Male			1.27	1.18, 1.37
Maternal education	Categorical	0.00	0.11		
	Vocational			Reference	Reference
	CSE/O level			0.91	0.80, 1.05
	A level/degree			0.45	0.39, 0.52
IQ at age 8	Continuous	44.49	0.37	0.98	0.98, 0.98
Intelligibility and fluency at age 9	Continuous	37.96	0.01	0.95	0.93, 0.97
Maths assessment score	Continuous	44.39	0.24	0.15	0.12, 0.19
Ever had learning difficulties	Binary	48.57	0.08	2.02	1.75, 2.33
Maths streaming group	Ordinal	52.76	0.20		
	Lowest			Reference	Reference
	Middle			0.58	0.50, 0.69
	Highest			0.42	0.36, 0.49
Literacy streaming group	Ordinal	55.03	0.16		
	Lowest			Reference	Reference
	Middle			0.59	0.50, 0.69
	Highest			0.39	0.33, 0.45

Resultados

Model	Exposure Coefficient	SE	FMI	% reduction in SE
CCA	-0.58	0.497		
A: No auxiliary information	-0.58	0.504	0.724	-1.39
B: IQ at age 8	-0.89	0.459	0.667	7.65
C: Intelligibility at age 9	-0.53	0.505	0.724	-1.62
D: Maths assessment score	-0.84	0.472	0.683	5.06
E: Learning Disabilities	-0.80	0.501	0.716	-0.80
F: Streaming for Maths and English	-0.86	0.476	0.686	4.34
G: IQ at age 8 and intelligibility at age 9	-0.87	0.470	0.682	5.39
H: IQ at age 8 and maths assessment score	-1.04	0.453	0.656	8.80
I: IQ, intel. and maths assessment	-1.04	0.458	0.663	7.78
J: IQ, intel., maths assessment and LD	-1.14	0.456	0.656	8.19
K: IQ, intel., maths assessment and streaming groups	-1.10	0.457	0.658	8.17
L: IQ, intel., maths assessment, LD and streaming groups	-1.13	0.461	0.664	7.17



Principais conclusões

1

A proporção de dados faltantes não deve ser usada como preditor de ganhos de eficiência com a IM (ou seja, não deve ser guia para decidir quando fazer uma IM ou um CCA)

2

A fração de dados faltantes pode ser usada para guiar a escolha de variáveis auxiliares no modelo de imputação

3

A proporção de dados faltantes fornece informações limitadas sobre a redução de vieses que podem ser feitos a partir de múltiplas imputação.

4

O aumento do número de variáveis auxiliares incluídas em um modelo de imputação nem sempre resulta em ganhos de eficiência.

Credits

This is where you give credit to the ones who are part of this project.

Did you like the resources on this template? Get them for **free** at our other websites.

Presentation template by Slidesgo

Icons by Flaticon

Images & infographics by Freepik

Author introduction slide photo created by **katemangostar** - Freepik.com

Big image slide photo created by **jcomp** - Freepik.com

Text & Image slide photo created by **rawpixel.com** - Freepik.com

Text & Image slide photo created by **Freepik**

Obrigado!