FACULDADE DE CIÊNCIAS
UNIVERSIDADE DE LISBOA

Seminários em Bioestatística

Seleção de variáveis na presença de valores omissos

Uma aplicação na modelação do Índice de Massa Corporal nos imigrantes africanos e brasileiros residentes em Lisboa e Setúbal

Beatriz Goulão
Orientadoras: Patrícia Bermudez e Valeska Andreozzi
Mestrado em Bioestatístca

Introdução – Dados omissos

- Muito comuns, em particular em estudos que envolvem pessoas como unidades amostrais (clínicos e epidemiológicos) (Molenberghs & Kenward, 2007)
- Métodos habitualmente usados (Casos completos por defeito) não são satisfatórios (Gelman et al, 2007):
  - Se os valores omissos diferem grandemente dos valores observados – viés nos resultados
  - Se variáveis incluídas no modelo têm muitos dados omissos poderão ter de ser excluídas
  - Ineficiência das estimativas obtidas (Harrell,2001)
  - Perda de potência estatística (Little and Rubin, 2002)

Objetivo

- Comparar diferentes abordagens para o tratamento de valores omissos na seleção de variáveis associadas ao índice de massa corporal (IMC) dos imigrantes africanos e brasileiros residentes em Lisboa e Setúbal

Introdução – Tipos dados omissos

- MNAR (**Missing Not At Random**)
  - Está relacionado com fatores não observados
- MAR (**Missing at Random**)
  - A probabilidade de que um valor seja omisso depende de valores de variáveis que foram, de facto, medidas.
- MCAR (**Missing Completely At Random**)
  - Não se encontra associado a características ou respostas dos sujeitos, incluindo o valor omisso, se este fosse conhecido. A probabilidade de não-resposta é a mesma.

Little and Rubin, 1987
Harrell, 2001

Introdução

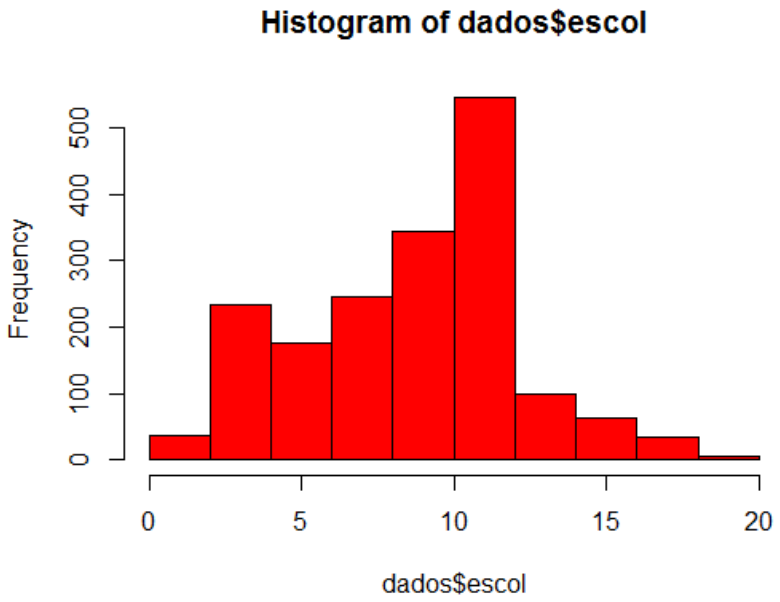
- **Imputação simples (IS)**
  - Métodos dedutivos (respostas anteriores)
  - Métodos determínisticos (pela média; por métodos regressivos)
  - Métodos estocásticos (Hot-Deck; Associação flexível; métodos regressivos com efeitos aleatórios)
- **Imputação múltipla (IM)**

Introdução - **Imputação pela média ou mediana não condicional**

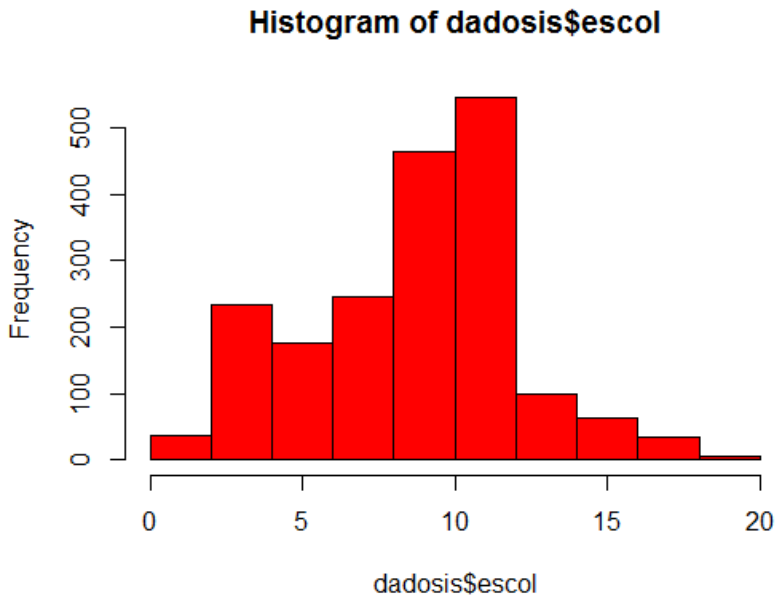
- Substitui-se o valor omisso pela mediana ou média dos valores observados da mesma variável nos restantes sujeitos. O investigador não usa informação acerca do sujeito para o qual a imputação é gerada.
- Características: Simples. Produz estimativas da variância e covariância enviesadas e subestimadas. Incorreta representação da distribuição da amostra.

Introdução – Imputação Simples

Histograma casos completos



Histograma dados IS mediana



Introdução - **Imputação por hot-deck**

- Imputação por emparelhamento. Para uma unidade com um valor omisso y, encontra-se uma unidade com valores similares de X nos dados observados e usa-se o valor de y correspondente (Gelman, 2007).
- Características: nunca encontramos valores "impossíveis", ao contrário da imputação por regressão (Kazemi, 2005). Pode distorcer correlações e covariâncias.

Introdução - **Imputação por hot-deck**

- ID 3: Imputação a partir do ID5 – 11 anos escol
- ID 7: Imputação a partir do ID 8 – 7 anos escol
- ID 10: Imputação a partir do ID9 – 9 anos escol

ID	Sexo	Grupo idade	Est Civil	Origem	Escol
1	F	2	S	A	13
2	F	3	C	B	12
3	F	2	C	B	-
4	F	3	C	A	4
5	M	2	C	B	11
6	F	3	S	B	12
7	M	1	S	B	-
8	M	1	S	A	7
9	F	4	V	A	9
10	F	4	D	A	-

Introdução - **Imputação múltipla**

- A IM consiste em **três passos**:
- São obtidos *m* bancos de dados completos através de técnicas adequadas de imputação;
- Separadamente, os *m* bancos são analisados por um método estatístico tradicional, como se realmente fossem conjuntos completos de dados;
- Os *m* resultados encontrados no passo 2 são combinados para obter a chamada inferência da imputação repetida.

Rubin, 1970
Nunes L, 2009

Projeto

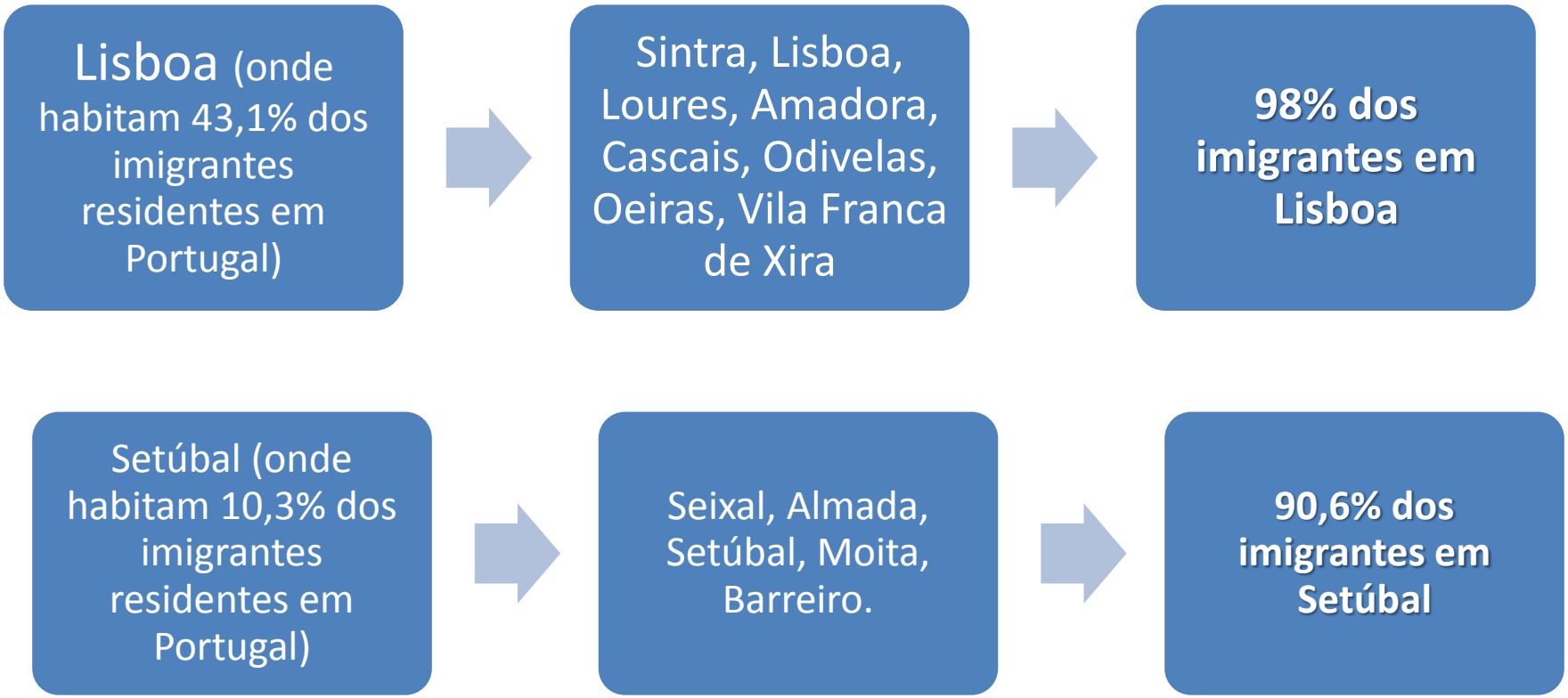
- Projeto **Acesso aos Cuidados de Saúde e Nível de Saúde das Comunidades Imigrantes Africana e Brasileira em Portugal** (SAIMI) - Unidade de Epidemiologia, IMP, FMUL
- Objetivos principais:
  - Caracterização do **estado de saúde das comunidades imigrantes**
  - **Comparação** do estado de saúde desta população **com o da população Portuguesa em geral**

Desenho do estudo

- Desenho de estudo: transversal analítico; recolha de dados através de questionário ministrado face-a-face no domicílio.
- Critérios inclusão
  - Imigrantes 1ª geração residentes em Lisboa e Setúbal e que tenham:
    - nascido num país PALOP e que tenham vindo para Portugal após 1980;
    - ou nascido no Brasil e que se considerem na situação de imigração desde 1995.
    - Entre os 18 e 64 anos
    - Peso e altura auto-relatados
- Tamanho amostral: 1904 individuos

Amostra

Critério seleção: Maior proporção de imigrantes



Amostragem

Dados recolhidos em 2007

Amostragem espacial simples por clusters

20 clusters (50 x 50 m^2)

Todos os sujeitos elegíveis/ cluster entrevistados

Equipas de 2 entrevistadores da comunidade ou mistas

97,9% adesão

Projeto

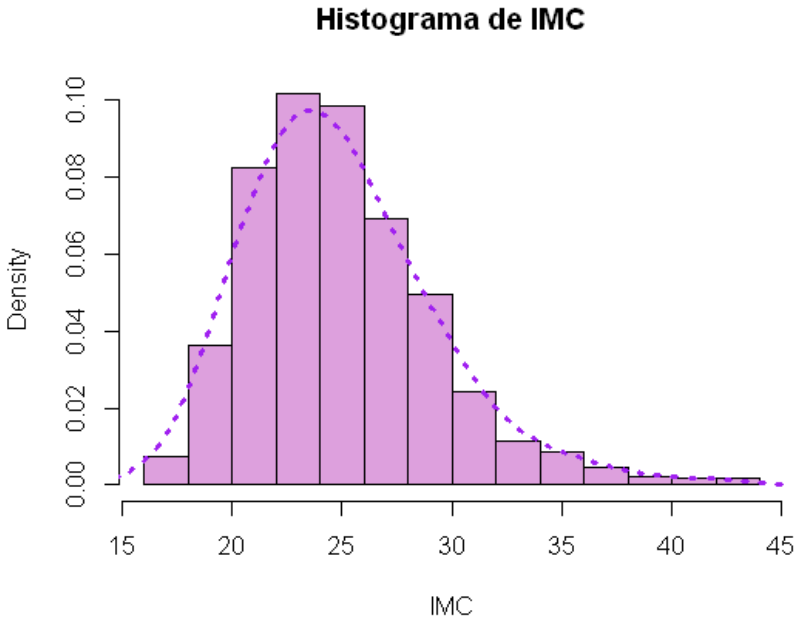
- Instrumento - Adaptação do 4º INS
- Para efeitos do presente estudo, foram usadas as seguintes secções de questões:
- *Caracterização sócio-demográfica;*
- *Trajectória imigratória;*
- *Informações gerais de saúde (auto-avaliação do estado de saúde, dados antropométricos e outros);*
- *Consumo de alimentos e bebidas;*
- *Actividade física.*

Resultados

Dados

- Variáveis resposta: **IMC (peso/altura^2)** (contínua)
- Variáveis explicativas:
  - **Sexo (**0 = Feminino; 1 = Masculino**)**
  - **Idade (**Anos**)**
  - **Estado civil (**0 = solteiro; 1 = casado; 2 = outro**)**
  - **Escolaridade (**Anos**)**
  - **Origem (**0 = africana; 1 = brasileira**)**
  - **Anos residência (**Anos**)**
  - **Nº refeições principais (**0 = menos de 3 refeições; 1 = 3 refeições**)**
  - **Nº refeições intermédias (Snack)**

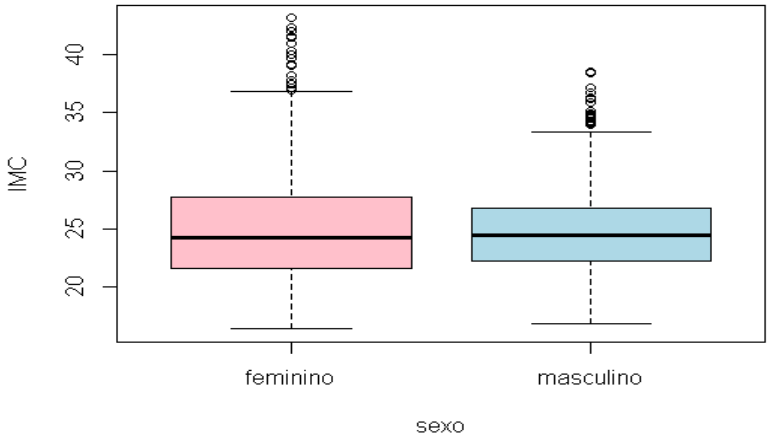
IMC - Distribuição

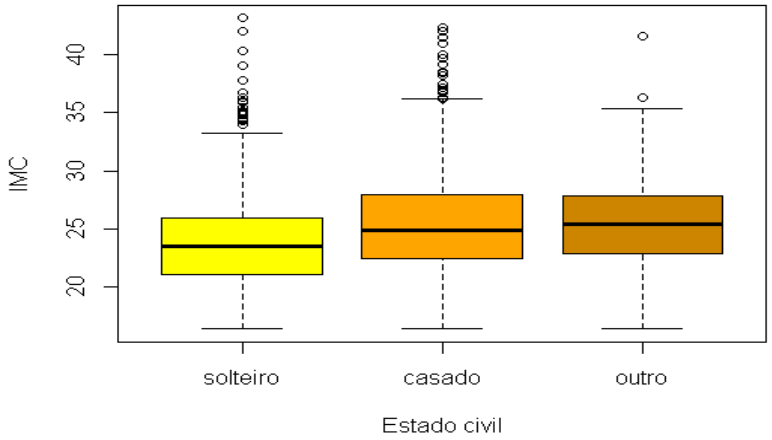


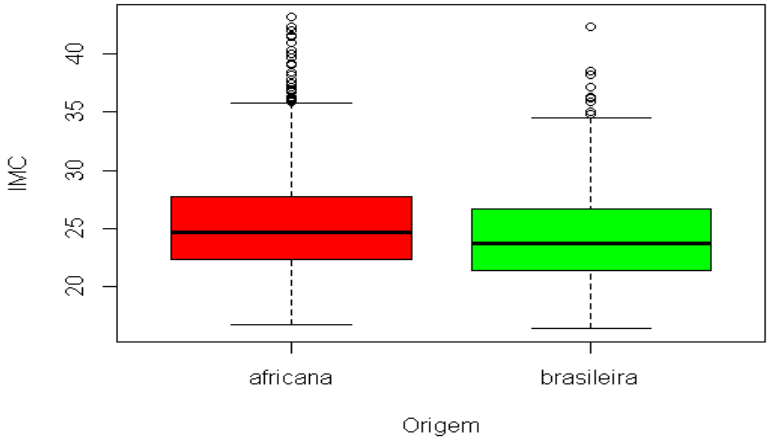
Resultados

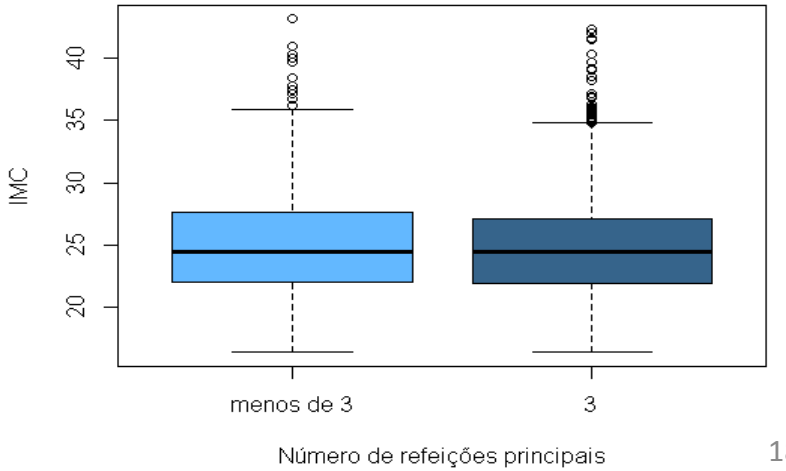
Variável	N = 1904
Sexo	Feminino: 1063 (54,0%) Masculino: 906 (46,0%)
IMC	24,47 ± 4,46 kg/m2
Idade	35,1 ± 10,95 anos
Estado Civil	Solteiro: 751 (38,1%) Casado: 1102 (60,0%) Outro: 130 (6,6%)
Escolaridade	9,21 ± 3,53 anos
Origem	Africanos: 1080 (56,7%) Brasileiros: 705 (37,3%)
Anos residência	9,83 ± 8,14 anos
Nº refeições principais	<3: 539 (27,4 %) 3: 1432 (72,3 %)
Nº refeições intermédias	1,22 ± 1,13 refeições

Resultados – Dados completos

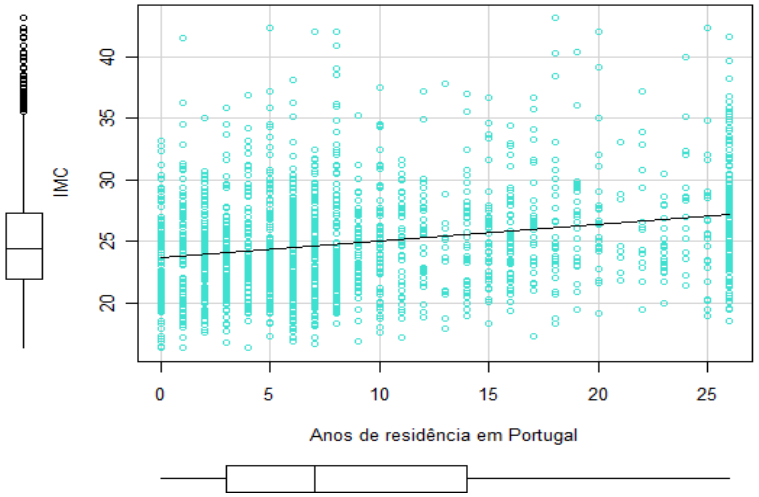


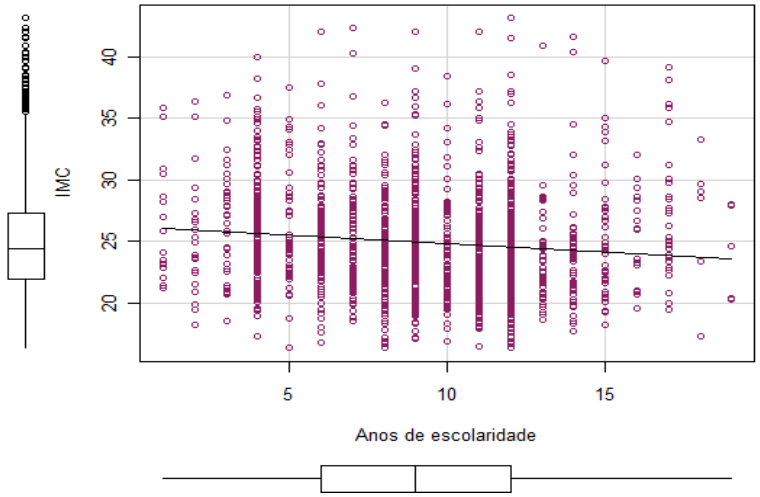


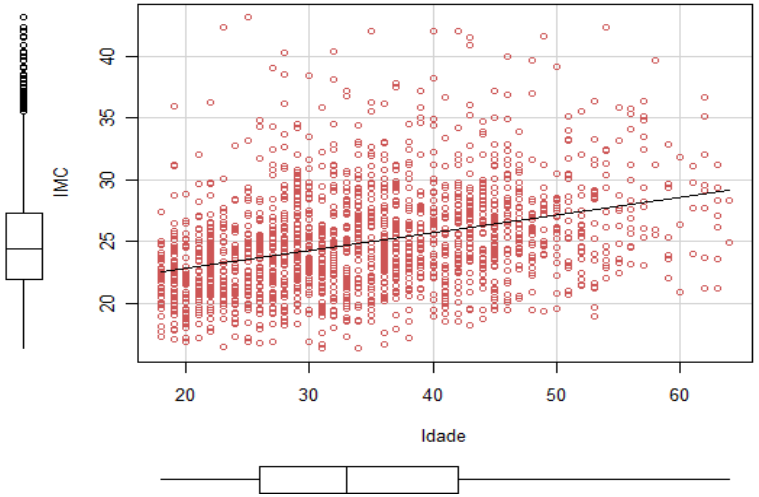


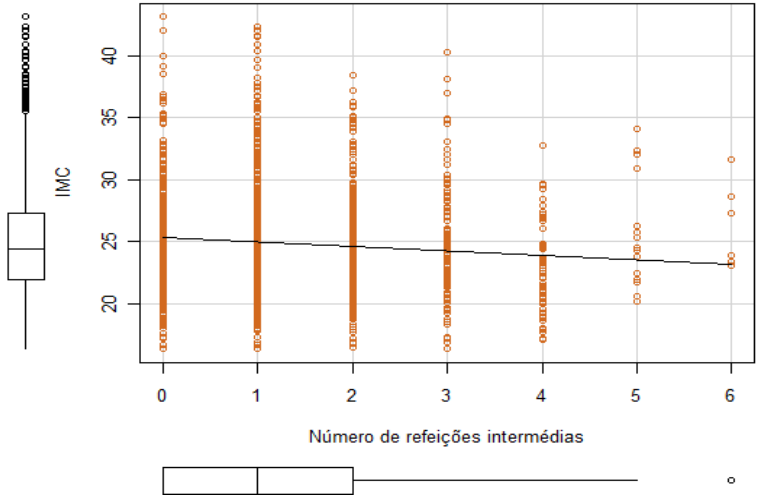


Resultados – Base completa









Resultados – Modelo Linear Generalizado

Modelo inicial : $E(IMC) = \beta_0 + \beta_{sx} * sx + \beta_{idade} * idade + \beta_{estcivil} * estcivil + \beta_{escol} * escol + \beta_{origem} * origem + \beta_{anos} * anos + \beta_{refeições} * refeicoes + \beta_{snack} * snack$

Covariável	Estimativa	(H0: $\beta_x = 0$) p-value
intercept	20.72	≈ 0
Sx (masculino)	-0.086	0.643
Idade	0.111	≈ 0
Estado civil (casado)	0.603	0.004
Estado civil (outro)	0.061	0.887
anos	0.066	≈ 0
Refeições (3)	-0.367	0.081
Snack	-0.199	0.014

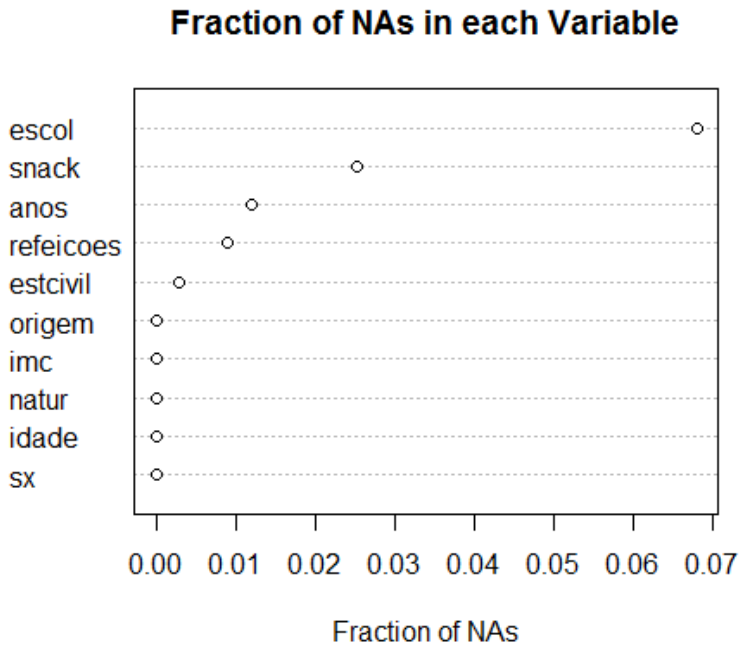
N = 1785

Função ligação: Identidade

Deviance 40.981 com 1774 graus de liberdade

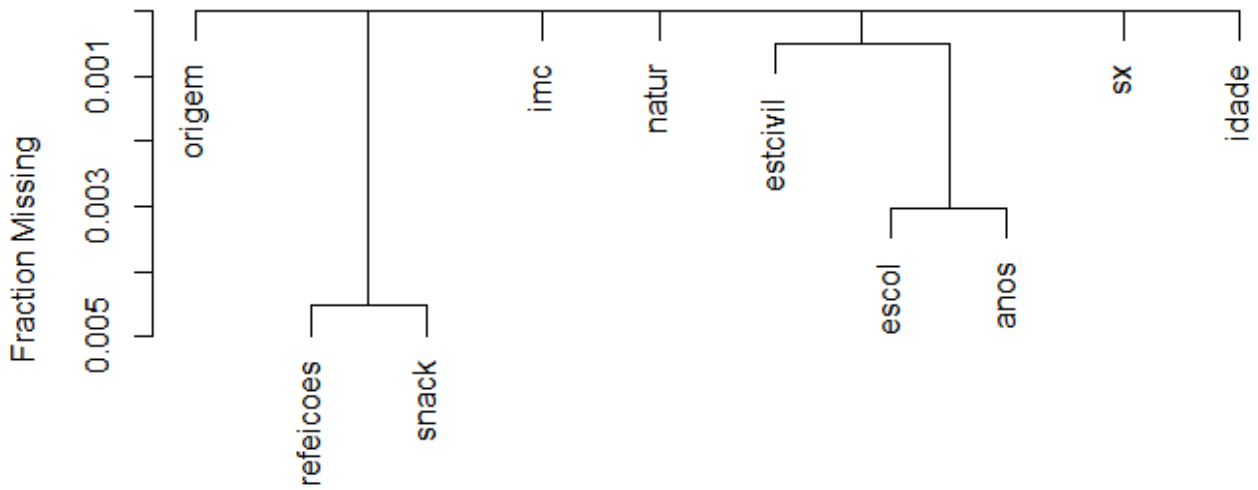
Resultados – exploração dados omissos

- Mesmo quando o investigador não pretende tratar os dados omissos, antes de descartar os sujeitos em causa, deve no mínimo estudar a tendência para os valores da variável Y estarem omissos (Harrell,2001).



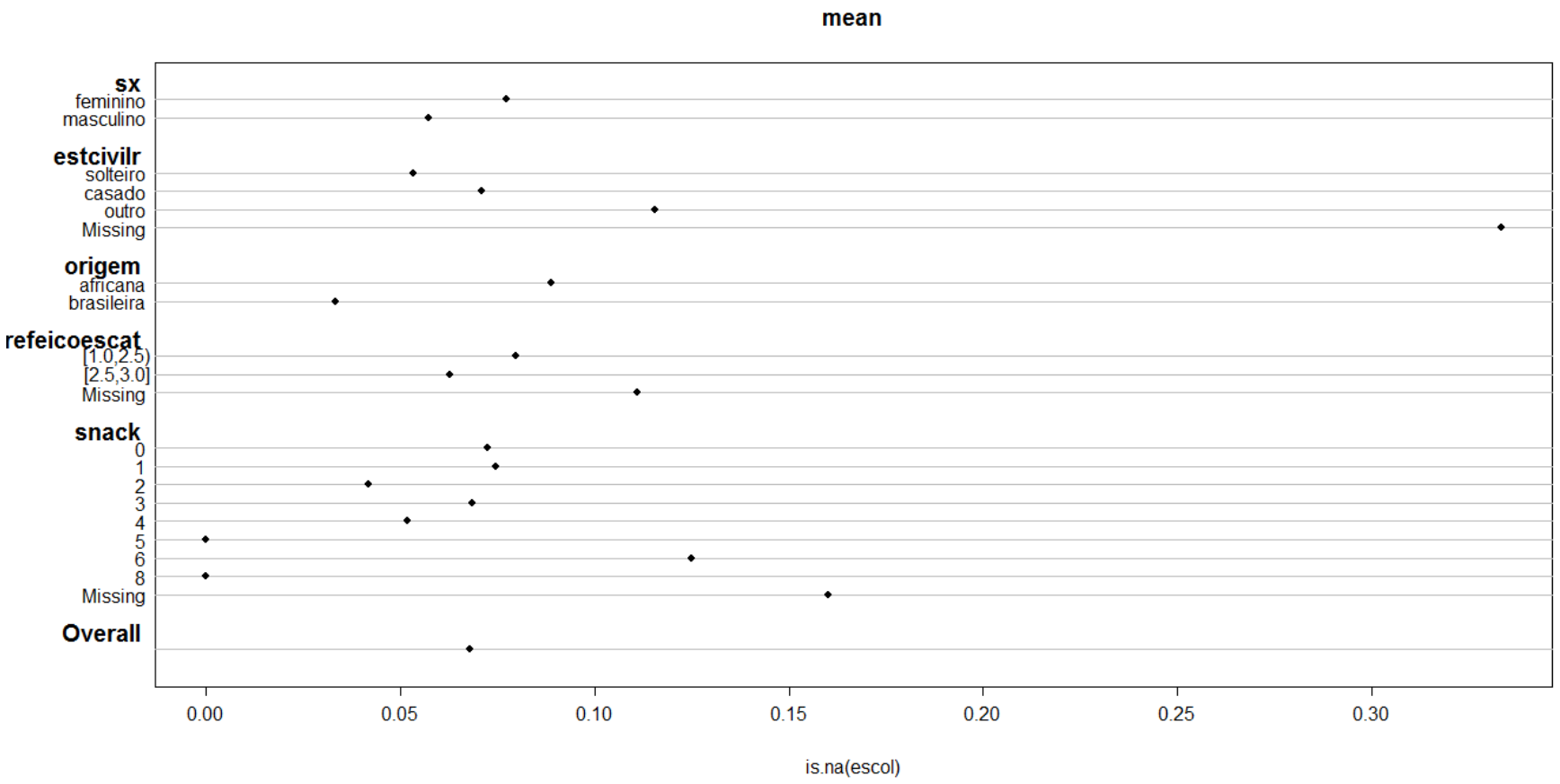
Resultados – exploração dados omissos

Análise hierárquica dos dados omissos



Resultados – exploração dados omissos

Descrição univariada da proporção de sujeitos com dados omissos na variável escolaridade



Resultados – exploração dados omissos

Regressão logística com is.NA(escolaridade) como variável resposta

Modelo inicial: $\text{logit}(P[\text{is.na(escolaridade)}] = \beta_0 + \beta_{sx} * sx + \beta_{idade} * idade + \beta_{estcivil} * estcivil + \beta_{IMC} * IMC + \beta_{origem} * origem + \beta_{anos} * anos + \beta_{refeições} * refeicoes + \beta_{snack} * snack$

Variável	Estimativa do coeficiente	Exp(coeficiente)	(H0: βx = 0) p-value
Sexo(masc)	-0,520	0,594	0,0131 *
Idade	0,0667	1,069	≈ 0 ***
Origem(Br)	-0,559	0,572	0,0786 .
Anos	0,323	1,381	0,0284*

Resultados – MLG com IS

Modelo inicial: $E(IMC) = \beta_0 + \beta_{sx} * sx + \beta_{idade} * idade + \beta_{estcivil} * estcivil + \beta_{Escolaridade} * Escolaridade + \beta_{natur} * natur + \beta_{anos} * anos + \beta_{refeições} * refeicoes + \beta_{snack} * snack$

Covariável	Estimativa	($H0: \beta_x = 0$) p-value
Idade	0.102	≈ 0 ***
Sexo(masculino)	-0.227	0.233
Estado civil (casado)	0.511	0.017*
Estado civil (outro)	0.264	0.544
anos	0.073	≈ 0 ***
Snack	-0.212	0.0102*

N = 1904; Função ligação: Identidade

Próximos passos

- Aplicação do *Hot-Deck*
- Aplicação da Imputação Múltipla
- Aplicação de métodos de regressão para identificar o efeito na seleção de variáveis explicativas

FACULDADE DE CIÊNCIAS
UNIVERSIDADE DE LISBOA

Seminários em Bioestatística

Seleção de variáveis na presença de valores omissos

Uma aplicação na modelação do Índice de Massa Corporal nos imigrantes africanos e brasileiros residentes em Lisboa e Setúbal

Beatriz Goulão
Orientadoras: Patrícia Bermudez e Valeska Andreozzi
Mestrado em Bioestatístca