Addressing Missing Data in the Development of a Risk Prediction Model for Childhood Obesity

Mafalda Oliveira

Faculdade de Ciências da Universidade do Porto

Mestrado em Estatística Computacional e Análise de Dados

Supervision: Rita Gaio, Susana Santos

U.PORTO
FACULDADE DE CIÊNCIAS UNIVERSIDADE DO PORTO

July 8, 2024

ISPUP
INSTITUTO DE SAÚDE PÚBLICA DA UNIVERSIDADE DO PORTO

Main Goal: To predict childhood obesity (yes/no) from longitudinal data, using an *exposome* approach.

INSTITUTO DE SAÚDE PÚBLICA DA UNIVERSIDADE DO PORTO

GERAÇÃO 21

Generation XXI cohort

- 8647 newborns and their mothers, 55 variables.
- Dataset under analysis: 4246 observations, 4 time-points (pregnancy/infancy, 4-, 7-, 10- y.o.).
- Pregnancy Var.: mother's age, working status, marital status, years of education, income, parity, smoking habits during pregnancy, gestational hypertension and diabetes, BMI, weight gain, number of gestational weeks, age of first solid food, first solid food, breastfeeding, mode of delivery.
- Newborn Var.: sex, BMI z-score, sedentary time, active play, sleep duration, sports activity, calories consumption, soft drinks consumption, soup/vegetables/fruit.
- Outcome: BMI z-score categorized as follows: normal weight or overweight/obesity.

Problems:

- Missing values (7%, 4246×55) - Which imputation method?
- *Many* Correlated Predictors - Which variable selection method?
- How to perform variable selection on multiply imputed datasets?
- How to model a cross-sectional response with longitudinal predictors?

Imputation Methods

- Deletion Methods
- Maximum Likelihood with Expectation-Maximization Algorithm
- Single Imputation
- Multiple Imputation

Multiple Imputation with Chained Equations (MICE)

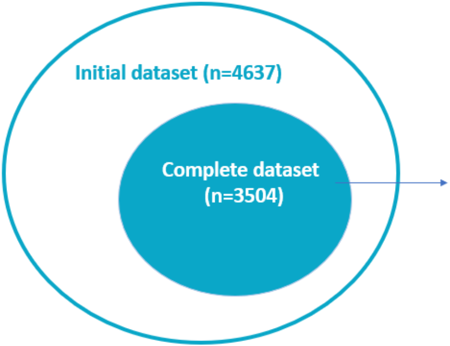
1. Single imputation is performed for every missing value in the dataset.
2. For the vector representing one particular variable, say x_j, the values imputed in step 1 are set back to miss.
3. The observed values in x_j are regressed on the remaining variables of the imputation model.
4. The missing values in x_j are replaced by the model's predictions.
5. Steps 2-4 are repeated for every variable. This completes a cycle.
6. Perform several cycles.

Which Imputation Method: Simulation Study

Simulation Study on the Pregnancy Dataset

- 6 Imputation Methods: Mean/Mode Imputation, Hot Deck Imputation, Random Imputation, MICE with 3 imputation models
- Models for MICE:
  - MICE with predictive mean matching, logistic binomial regression model, logistic multinomial regression model.
  - MICE with a linear model, logistic binomial regression model, logistic multinomial regression model
  - MICE with random forests.

Which Imputation Method: Simulation Study



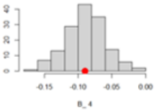
150 times

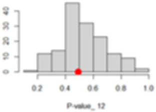
1. Logistic Regression;
2. Random generation of missing values (5.39%);
3. Application of each of the 6 imputation methods;
4. Logistic regression in each new imputed dataset.

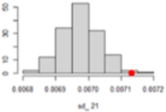
Which Imputation Method: Simulation Study

Table 1: Sample quantiles of the original estimates.

	MICE-Lm	MICE-Pmm	MICE-Rf	Hot Deck	Mean	Random
Coefficients						
average	0.47	0.49	0.52	0.49	0.51	0.55
median	0.48	0.49	0.53	0.48	0.49	0.53
sd	0.11	0.05	0.14	0.21	0.16	0.24
min	0.23	0.40	0.23	0.00	0.13	0.06
max	0.78	0.59	0.79	1.00	0.84	0.98
P-values						
average	0.44	0.40	0.36	0.52	0.45	0.42
median	0.42	0.40	0.38	0.52	0.45	0.42
sd	0.16	0.15	0.19	0.18	0.19	0.21
min	0.04	0.02	0.01	0.13	0.04	0.03
max	0.82	0.85	0.89	1.00	0.9	0.97
Standard Errors						
average	0.02	0.02	0.04	0.92	0.33	0.73
median	0.00	0.00	0.02	0.99	0.073	0.70
sd	0.05	0.04	0.05	0.23	0.41	0.13
min	0.00	0.00	0.00	0.59	0.00	0.35
max	0.15	0.13	0.16	1.00	1.00	1.00





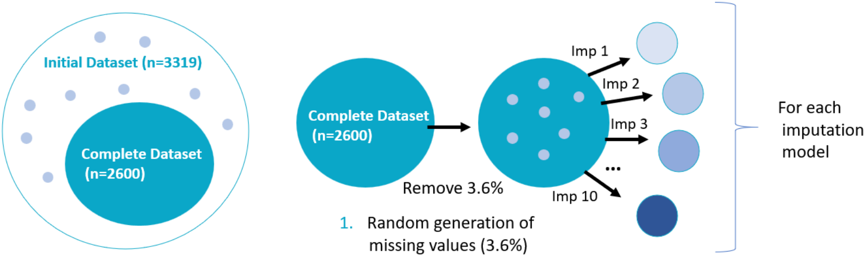


Which Imputation Method: Simulation Study

Table 1: Median (min-max) of the relative increase in variance (RIV), the fraction of missing information (FMI), and relative efficiency (RE), for each model.

	RIV	**FMI**	**RE**
Predictive Mean Matching	0.067 (0.053-0.106)	0.064 (0.051-0.098)	0.994 (0.990-0.995)
Linear Model	0.064 (0.007-0.087)	0.061 (0.007-0.082)	0.994 (0.991-0.999)
Random Forest	0.063 (0.023-0.135)	0.060 (0.024-0.122)	0.994 (0.988-0.998)

Agreement Among Imputed Values: Simulation Study



1. Random generation of missing values (3.6%)

2. MICE with 3 imputation models (m=10)

3. Concordance measures among the 10 imputed values for each variable in each model; → **Fleiss'kappa** **ICC**
4. Comparison of the imputed values with the original values. → **Bias, MSE, MAPE** **Proportion of true class in the imputed classes**

Agreement Among Imputed Values: Simulation Study

Table 2: Summary of the agreement measurements for each variable and each imputation method.

Method	ICC			Fleiss' kappa					
	Age	G. Weeks	BMI	Inc.	Educ	Work	Marital	Parity	Smoking
Linear	0.982	0.996	0.984	0.239	0.218	0.174	0.062	0.280	0.059
Model	N. Weight	N. Weight		Hip.	Diab.	Vaginal	Ceasearen	Weight G.	Sex
	0.999	0.999		0.611	0.668	0.930	0.923	0.099	0.027
	Age	G. Weeks	BMI	Inc.	Educ	Work	Marital	Parity	Smoking
Random	0.237	0.438	0.110	0.245	0.099	0.117	0.291	0.256	0.014
Forest	N. Weight	N. Weight		Hip.	Diab.	Vaginal	Ceasearen	Weight G.	Sex
	0.488	0.406		0.692	0.532	0.983	0.978	0.100	0.078
	Age	G. Weeks	BMI	Inc.	Educ	Work	Marital	Parity	Smoking
Predictive	0.298	0.579	0.227	0.232	0.134	0.185	0.231	0.292	0.036
Mean	N. Weight	N. Weight		Hip.	Diab.	Vaginal	Ceasearen	Weight G.	Sex
Matching	0.701	0.659		0.598	0.695	0.942	0.955	0.076	0.014

Comparison with the Original Observed Value: Simulation Study

Percentage of matching: linear model - 73%, random forest - 76%, predictive mean matching - 71%

Table 3: Median values for the performance measures, for each variable and each imputation model.

	Bias			MSE			MAPE		
Variable	**Pmm**	**Lm**	**Rf**	**Pmm**	**Lm**	**Rf**	**Pmm**	**Lm**	**Rf**
Mother's Age	0.13	0.10	0.05	1.07	0.40	1.16	1.44	0.93	1.44
Mother's BMI	-0.04	-0.01	0.01	1.06	0.25	0.93	1.56	0.90	1.47
Gestational weeks	-0.08	-0.12	-0.08	0.71	0.26	0.77	1.25	0.98	1.19
Newborn's Weight	-0.02	0.02	0.05	0.40	0.13	0.62	0.92	0.57	1.05
Newborn's Length	0.03	-0.14	-0.12	0.52	0.15	0.70	0.93	0.62	1.05

Conclusions - Imputation Methods

1. Among the six tested imputation methods, MICE with predictive mean matching and logistic regression exhibited regression coefficient values that were closest to those obtained from the complete dataset;
2. The relative efficiency of the imputation procedure remained consistent across the three imputation models considered for MICE;
3. MICE with linear models and logistic regression resulted in the imputed datasets with the highest concordance measures for the continuous variables.

Modelling a Cross-Sectional Response with Longitudinal Predictors

1. Two-step Regression:
   1. a linear regression of each predictor against time
   2. a logistic fixed-effects regression of obesity on the estimated random effects

$$X_{ij} = \beta_0 + \beta_1 t_{ij} + e_{ij},\; e_{ij} \sim N(0, \sigma^2)$$
$$logit(P(Y_i = 1|\hat{\beta}_0, \hat{\beta}_1)) = \beta_0 + \beta_1 \hat{\beta}_0 + \beta_2 \hat{\beta}_1$$

2. Penalized Regression Models.
3. Finite Mixture of Regressions

Variable Selection - Penalized Regression

$$\hat{\theta} = \arg min_{\theta}\{-\frac{1}{n}\sum_{i=1}^{n} logL(\theta|Y_i, X_i) + \lambda P_{\alpha}(\beta)\}$$

Examples of penalty functions $P_{\alpha}(\beta)$:

1. Ridge: $P_{\alpha}(\beta) = \sum_{j=1}^{p} \beta_j^2$
2. LASSO: $P_{\alpha}(\beta) = \sum_{j=1}^{p} |\beta_j|$
3. ENET: $P_{\alpha}(\beta) = \alpha \sum_{j=1}^{p} |\beta_j| + (1-\alpha) \sum_{j=1}^{p} \beta_j^2$

where, L is the likelihood function, β is the $p \times 1$ vector of regression coefficients, $\theta = (\beta_0, \beta)$ is the total vector of the regression parameters which includes the intercept, and λ is the shrinkage parameter.

Finite Mixture of Regressions

A statistical model that assumes that a population is composed of multiple latent subpopulations or components, each following a different probability distribution.
For K components, it is defined by:

$$h(y|x,\phi) = \sum_{k=1}^{K} \pi_k f(y|x,\theta_k), \qquad \pi_k > 0 \qquad \sum_{k=1}^{K} \pi_k = 1$$

y is the dependent variable with conditional density h, x is the vector of predictors, π_k is the prior probability of component k, θ_k is the vector of the component-specific parameters for the density function f, and $\phi = (\pi_1, ...\pi_k, \theta_1^T, ..., \theta_K^T)^T$ is the total vector of parameters.

The Followed Procedure

1. Divide the dataset in a train set (70 %) and a test set (30 %):
2. Run MICE in each set; obtain 10 complete datasets;
3. Build 4 static models, one for each time-point (pregnancy/infancy, 4 y.o., 7 y.o., 10 y.o.);
4. Build a dynamic model with all exposures.

Pregnancy/Infancy Static Model

The model includes 23 predictors including the BMI of the child at 6 months, 1 year, and 2 years old, which are correlated.

1. Model 1: two-step regression;
2. Model 2: a penalized regression model with ENET penalization.

Results - Model 1

Table 4: Odds Ratios and p-values for the final model.

Exposures	**Odds-Ratio (95 % CI)**	**p-value**
Age	0.98 (0.96, 0.99)	0.008
Household Income		
Low	Ref	Ref
Middle	0.87 (0.70, 1.07)	0.19
High	0.77 (0.61, 0.96)	0.02
Smoking Habits during Pregnancy		
Never smoked	Ref	Ref
Smoked	1.40 (1.14, 1.72)	0.001
Hypertensive complications		
Yes	1.64 (1.25, 2.14)	< 0.001
No	Ref	Ref
Pre-conceptual BMI		
Underweight/normal	Ref	Ref
Overweight/obese	2.35 (1.98, 2.80)	< 0.001
First solid food		
Cereal porridge	0.84 (0.70, 0.99)	0.04
Fruit	0.79 (0.52, 1.19)	0.27
Soup	Ref	Ref
Intercept of the BMI Regression	1.48 (1.38, 1.58)	< 0.001
Slope of the BMI Regression	2.02 (1.74, 2.35)	< 0.001

Model's Performance

Table 5: Performance measures for each model.

Models	Specificity	Sensitivity	AUC	PPA	NPA	Prediction Error
Pregnancy model	0.71	0.61	0.70	0.53	0.77	33 %
4-Year Model	0.86	0.61	0.76	0.71	0.80	23 %
7-Year Model	0.86	0.67	0.82	0.70	0.84	20%
10-Year Model	0.88	0.81	0.87	0.70	0.93	17 %

Dynamic Model

The dynamic model was fitted considering **all predictors at once**, using 2 approaches:

1. a penalized regression model with ENET penalization;
2. a finite mixture of penalized regressions.

ENET Model - Results

- $\hat{\lambda} \approx 0.009$, $\hat{\alpha} \approx 0.1$;
- 27 variables selected (from the initial 55);
- The child's z-score BMI was the predictor with the largest effect on obesity at age 13, and its effect increases as the time-point increases;

$\widehat{OR}$ (Normal, Obese)=1.082 at age 4
$\widehat{OR}$ (Normal, Obese)=1.464 at age 10

Model's Performance

Table 6: Performance measures for each model.

Models	Specificity	Sensitivity	AUC	PPA	NPA	Prediction Error
10-Year Model	0.88	0.81	0.87	0.70	0.93	17 %
Dynamic Model ENET	0.86	0.81	0.88	0.72	0.91	17 %

Finite Mixture of Regressions - Results

- Finite mixture of penalized regressions with LASSO regularization;
- Outcome: BMI at 13;
- 2 components identified (BIC): component 1 has 2187 observations, component 2 has 796;
- $\hat{\pi}_1 = 0.6, \quad \hat{\pi}_2 = 0.4$;
- The mean (standard deviation) of the posterior probabilities from Component 1 was 0.75 (0.12) and 0.83 (0.17) for Component 2.

Model's Performance

Table 7: Performance measures for the dynamic model.

Model	MSE	MAE	MAPE
Finite Mixture Model	2.60	1.15	5.6 %

Conclusions - Modelling Process

1. Our findings identified strong associations between several variables collected during pregnancy and childhood, and obesity-13;
2. The Child's BMI measured at each follow-up period was systematically the most important predictor for obesity;
3. The 10 y.o.-model had the best prediction ability;
4. The dynamic finite mixture of regressions presented the highest accuracy.

Thank you for your attention!