The algorithmic Search for the optimal Number of Imputations

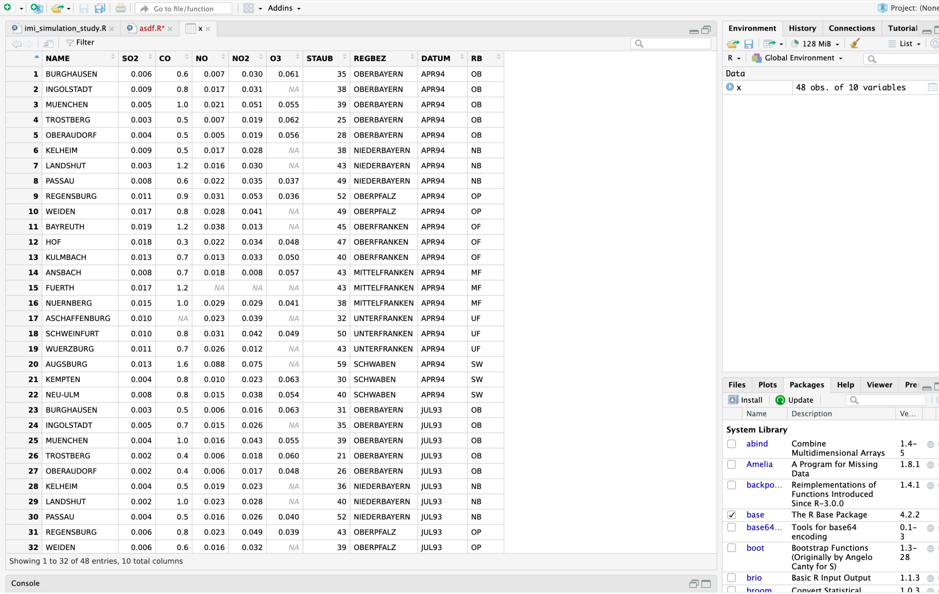
Gedeon Alexander Vogt

27.03.2023

Outline

Outline

The Problem



	NAME	SO2	CO	NO	NO2	O3	STAUB	REGBEZ	DATUM	RB
1	BURGHAUSEN	0.006	0.6	0.007	0.030	0.061	35	OBERBAYERN	APR94	OB
2	INGOLSTADT	0.009	0.8	0.017	0.031	NA	38	OBERBAYERN	APR94	OB
3	MUENCHEN	0.005	1.0	0.021	0.051	0.055	39	OBERBAYERN	APR94	OB
4	TROSTBERG	0.003	0.5	0.007	0.019	0.062	25	OBERBAYERN	APR94	OB
5	OBERAUDORF	0.004	0.5	0.005	0.019	0.056	28	OBERBAYERN	APR94	OB
6	KELHEIM	0.009	0.5	0.017	0.028	NA	38	NIEDERBAYERN	APR94	NB
7	LANDSHUT	0.003	1.2	0.016	0.030	NA	43	NIEDERBAYERN	APR94	NB
8	PASSAU	0.008	0.6	0.022	0.035	0.037	49	NIEDERBAYERN	APR94	NB
9	REGENSBURG	0.011	0.9	0.031	0.053	0.036	52	OBERPFALZ	APR94	OP
10	WEIDEN	0.017	0.8	0.028	0.041	NA	49	OBERPFALZ	APR94	OP
11	BAYREUTH	0.019	1.2	0.038	0.013	NA	45	OBERFRANKEN	APR94	OF
12	HOF	0.018	0.3	0.022	0.034	0.048	47	OBERFRANKEN	APR94	OF
13	KULMBACH	0.013	0.7	0.013	0.033	0.050	40	OBERFRANKEN	APR94	OF
14	ANSBACH	0.008	0.7	0.018	0.008	0.057	43	MITTELFRANKEN	APR94	MF
15	FUERTH	0.017	1.2	NA	NA	NA	43	MITTELFRANKEN	APR94	MF
16	NUERNBERG	0.015	1.0	0.029	0.029	0.041	38	MITTELFRANKEN	APR94	MF
17	ASCHAFFENBURG	0.010	NA	0.023	0.039	NA	32	UNTERFRANKEN	APR94	UF
18	SCHWEINFURT	0.010	0.8	0.031	0.042	0.049	50	UNTERFRANKEN	APR94	UF
19	WUERZBURG	0.011	0.7	0.026	0.012	NA	43	UNTERFRANKEN	APR94	UF
20	AUGSBURG	0.013	1.6	0.088	0.075	NA	59	SCHWABEN	APR94	SW
21	KEMPTEN	0.004	0.8	0.010	0.023	0.063	30	SCHWABEN	APR94	SW
22	NEU-ULM	0.008	0.8	0.015	0.038	0.054	40	SCHWABEN	APR94	SW
23	BURGHAUSEN	0.003	0.5	0.006	0.016	0.063	31	OBERBAYERN	JUL93	OB
24	INGOLSTADT	0.005	0.7	0.015	0.026	NA	35	OBERBAYERN	JUL93	OB
25	MUENCHEN	0.004	1.0	0.016	0.043	0.055	39	OBERBAYERN	JUL93	OB
26	TROSTBERG	0.002	0.4	0.006	0.018	0.060	21	OBERBAYERN	JUL93	OB
27	OBERAUDORF	0.002	0.4	0.006	0.017	0.048	26	OBERBAYERN	JUL93	OB
28	KELHEIM	0.004	0.5	0.019	0.023	NA	36	NIEDERBAYERN	JUL93	NB
29	LANDSHUT	0.002	1.0	0.023	0.028	NA	40	NIEDERBAYERN	JUL93	NB
30	PASSAU	0.004	0.5	0.016	0.026	0.040	52	NIEDERBAYERN	JUL93	NB
31	REGENSBURG	0.006	0.8	0.023	0.049	0.039	43	OBERPFALZ	JUL93	OP
32	WEIDEN	0.006	0.6	0.016	0.032	NA	39	OBERPFALZ	JUL93	OP

The Problem

	NAME	SO2	CO	NO	NO2	O3	STAUB	REGBEZ	DATUM	RB
1	BURGHAUSEN	0.006	0.6	0.007	0.030	0.061	35	OBERBAYERN	APR94	OB
2	INGOLSTADT	0.009	0.8	0.017	0.031	NA	38	OBERBAYERN	APR94	OB
3	MUENCHEN	0.005	1.0	0.021	0.051	0.055	39	OBERBAYERN	APR94	OB
4	TROSTBERG	0.003	0.5	0.007	0.019	0.062	25	OBERBAYERN	APR94	OB
5	OBERAUDORF	0.004	0.5	0.005	0.019	0.056	28	OBERBAYERN	APR94	OB
6	KELHEIM	0.009	0.5	0.017	0.028	NA	38	NIEDERBAYERN	APR94	NB
7	LANDSHUT	0.003	1.2	0.016	0.030	NA	43	NIEDERBAYERN	APR94	NB
8	PASSAU	0.008	0.6	0.022	0.035	0.037	49	NIEDERBAYERN	APR94	NB
9	REGENSBURG	0.011	0.9	0.031	0.053	0.036	52	OBERPFALZ	APR94	OP
10	WEIDEN	0.017	0.8	0.028	0.041	NA	49	OBERPFALZ	APR94	OP
11	BAYREUTH	0.019	1.2	0.038	0.013	NA	45	OBERFRANKEN	APR94	OF
12	HOF	0.018	0.3	0.022	0.034	0.048	47	OBERFRANKEN	APR94	OF
13	KULMBACH	0.013	0.7	0.013	0.033	0.050	40	OBERFRANKEN	APR94	OF
14	ANSBACH	0.008	0.7	0.018	0.008	0.057	43	MITTELFRANKEN	APR94	MF
15	FUERTH	0.017	1.2	NA	NA	NA	43	MITTELFRANKEN	APR94	MF
16	NUERNBERG	0.015	1.0	0.029	0.029	0.041	38	MITTELFRANKEN	APR94	MF
17	ASCHAFFENBURG	0.010	NA	0.023	0.039	NA	32	UNTERFRANKEN	APR94	UF
18	SCHWEINFURT	0.010	0.8	0.031	0.042	0.049	50	UNTERFRANKEN	APR94	UF
19	WUERZBURG	0.011	0.7	0.026	0.012	NA	43	UNTERFRANKEN	APR94	UF
20	AUGSBURG	0.013	1.6	0.088	0.075	NA	59	SCHWABEN	APR94	SW
21	KEMPTEN	0.004	0.8	0.010	0.023	0.063	30	SCHWABEN	APR94	SW
22	NEU-ULM	0.008	0.8	0.015	0.038	0.054	40	SCHWABEN	APR94	SW
23	BURGHAUSEN	0.003	0.5	0.006	0.016	0.063	31	OBERBAYERN	JUL93	OB
24	INGOLSTADT	0.005	0.7	0.015	0.026	NA	35	OBERBAYERN	JUL93	OB
25	MUENCHEN	0.004	1.0	0.016	0.043	0.055	39	OBERBAYERN	JUL93	OB
26	TROSTBERG	0.002	0.4	0.006	0.018	0.060	21	OBERBAYERN	JUL93	OB
27	OBERAUDORF	0.002	0.4	0.006	0.017	0.048	26	OBERBAYERN	JUL93	OB
28	KELHEIM	0.004	0.5	0.019	0.023	NA	36	NIEDERBAYERN	JUL93	NB
29	LANDSHUT	0.002	1.0	0.023	0.028	NA	40	NIEDERBAYERN	JUL93	NB
30	PASSAU	0.004	0.5	0.016	0.026	0.040	52	NIEDERBAYERN	JUL93	NB
31	REGENSBURG	0.006	0.8	0.023	0.049	0.039	43	OBERPFALZ	JUL93	OP
32	WEIDEN	0.006	0.6	0.016	0.032	NA	39	OBERPFALZ	JUL93	OP

Showing 1 to 32 of 48 entries, 10 total columns

Environment: Data — x: 48 obs. of 10 variables

Possible Solutions

(i) Drop rows containing NAs

Possible Solutions

(i) Drop rows containing NAs

(ii) Single imputation

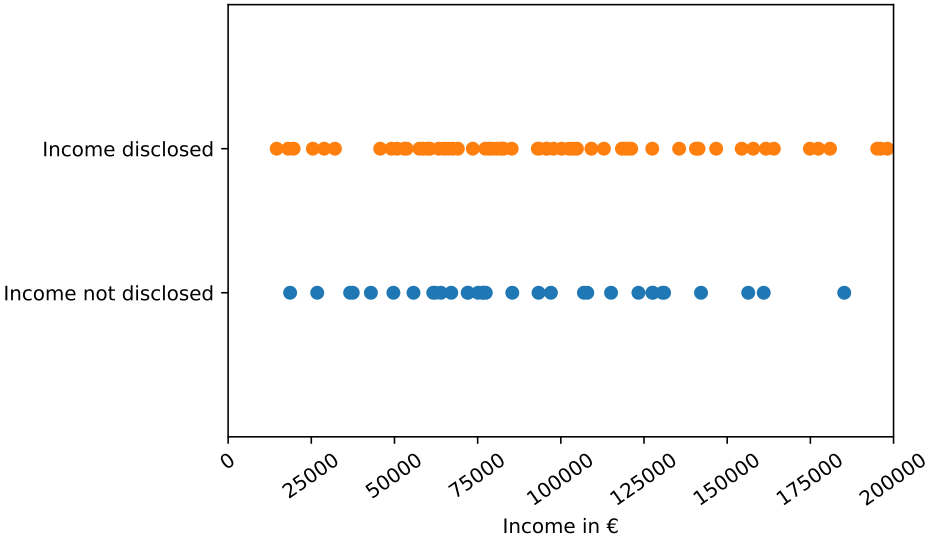
Possible Solutions

(i) Drop rows containing NAs

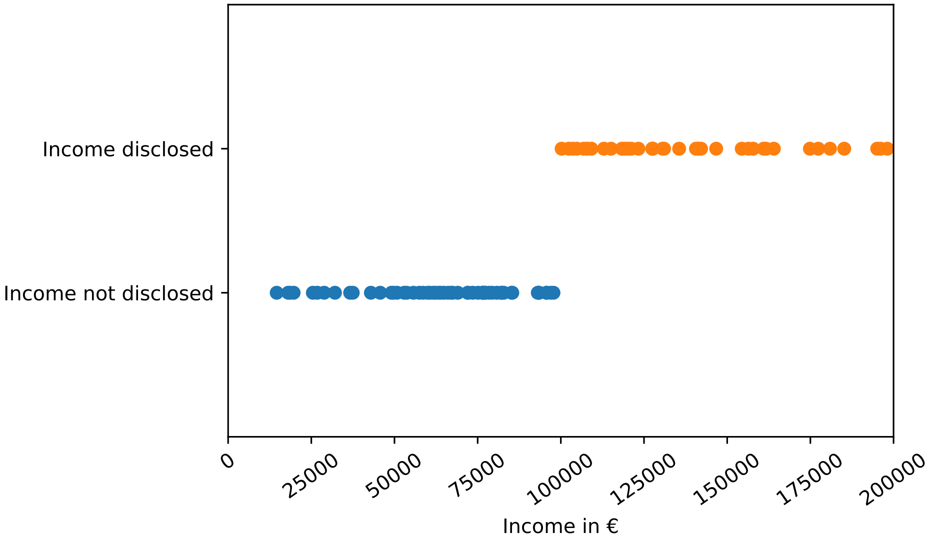
(ii) Single imputation

(iii) Multiple imputation

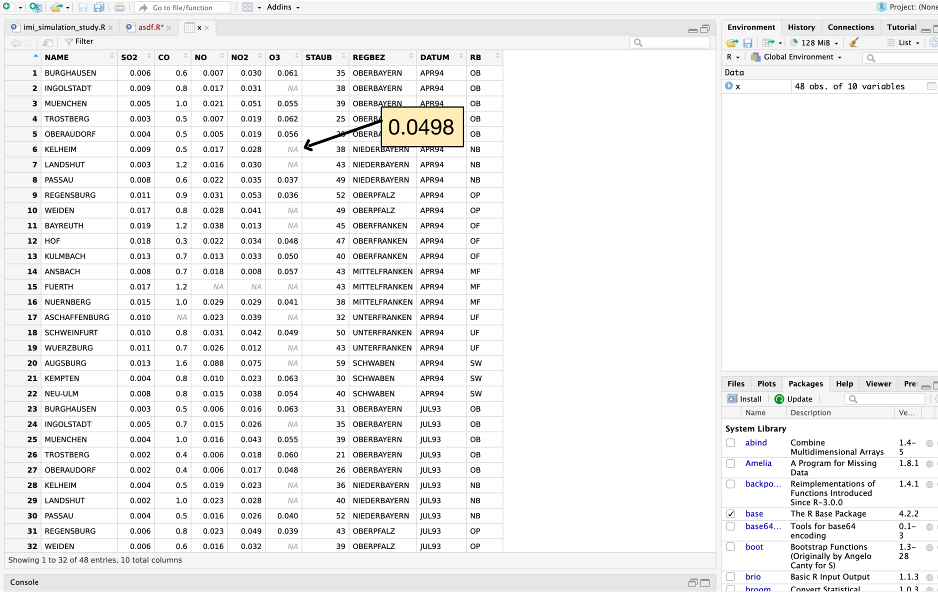
Possible Solutions – Drop Rows containing NAs



Possible Solutions – Drop Rows containing NAs



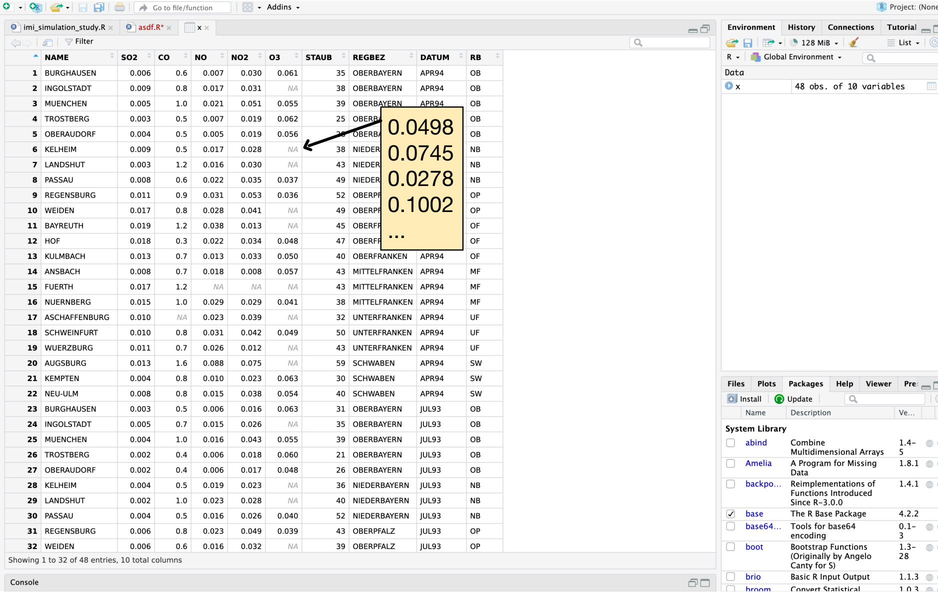
Possible Solutions – Single Imputation



	NAME	SO2	CO	NO	NO2	O3	STAUB	REGBEZ	DATUM	RB
1	BURGHAUSEN	0.006	0.6	0.007	0.030	0.061	35	OBERBAYERN	APR94	OB
2	INGOLSTADT	0.009	0.8	0.017	0.031	NA	38	OBERBAYERN	APR94	OB
3	MUENCHEN	0.005	1.0	0.021	0.051	0.055	39	OBERBAYERN	APR94	OB
4	TROSTBERG	0.003	0.5	0.007	0.019	0.062	25	OBERB		OB
5	OBERAUDORF	0.004	0.5	0.005	0.019	0.056	38	OBERBA		OB
6	KELHEIM	0.009	0.5	0.017	0.028	NA	38	NIEDERBAYERN	APR94	NB
7	LANDSHUT	0.003	1.2	0.016	0.030	NA	43	NIEDERBAYERN	APR94	NB
8	PASSAU	0.008	0.6	0.022	0.035	0.037	49	NIEDERBAYERN	APR94	NB
9	REGENSBURG	0.011	0.9	0.031	0.053	0.036	52	OBERPFALZ	APR94	OP
10	WEIDEN	0.017	0.8	0.028	0.041	NA	49	OBERPFALZ	APR94	OP
11	BAYREUTH	0.019	1.2	0.038	0.013	NA	45	OBERFRANKEN	APR94	OF
12	HOF	0.018	0.3	0.022	0.034	0.048	47	OBERFRANKEN	APR94	OF
13	KULMBACH	0.013	0.7	0.013	0.033	0.050	40	OBERFRANKEN	APR94	OF
14	ANSBACH	0.008	0.7	0.018	0.008	0.057	43	MITTELFRANKEN	APR94	MF
15	FUERTH	0.017	1.2	NA	NA	NA	43	MITTELFRANKEN	APR94	MF
16	NUERNBERG	0.015	1.0	0.029	0.029	0.041	38	MITTELFRANKEN	APR94	MF
17	ASCHAFFENBURG	0.010	NA	0.023	0.039	NA	32	UNTERFRANKEN	APR94	UF
18	SCHWEINFURT	0.010	0.8	0.031	0.042	0.049	50	UNTERFRANKEN	APR94	UF
19	WUERZBURG	0.011	0.7	0.026	0.012	NA	43	UNTERFRANKEN	APR94	UF
20	AUGSBURG	0.013	1.6	0.088	0.075	NA	59	SCHWABEN	APR94	SW
21	KEMPTEN	0.004	0.8	0.010	0.023	0.063	30	SCHWABEN	APR94	SW
22	NEU-ULM	0.008	0.8	0.015	0.038	0.054	40	SCHWABEN	APR94	SW
23	BURGHAUSEN	0.003	0.5	0.006	0.016	0.063	31	OBERBAYERN	JUL93	OB
24	INGOLSTADT	0.005	0.7	0.015	0.026	NA	35	OBERBAYERN	JUL93	OB
25	MUENCHEN	0.004	1.0	0.016	0.043	0.055	39	OBERBAYERN	JUL93	OB
26	TROSTBERG	0.002	0.4	0.006	0.018	0.060	21	OBERBAYERN	JUL93	OB
27	OBERAUDORF	0.002	0.4	0.006	0.017	0.048	26	OBERBAYERN	JUL93	OB
28	KELHEIM	0.004	0.5	0.019	0.023	NA	36	NIEDERBAYERN	JUL93	NB
29	LANDSHUT	0.002	1.0	0.023	0.028	NA	40	NIEDERBAYERN	JUL93	NB
30	PASSAU	0.004	0.5	0.016	0.026	0.040	52	NIEDERBAYERN	JUL93	NB
31	REGENSBURG	0.006	0.8	0.023	0.049	0.039	43	OBERPFALZ	JUL93	OP
32	WEIDEN	0.006	0.6	0.016	0.032	NA	39	OBERPFALZ	JUL93	OP

Showing 1 to 32 of 48 entries, 10 total columns

Possible Solutions – Multiple Imputation



	NAME	SO2	CO	NO	NO2	O3	STAUB	REGBEZ	DATUM	RB
1	BURGHAUSEN	0.006	0.6	0.007	0.030	0.061	35	OBERBAYERN	APR94	OB
2	INGOLSTADT	0.009	0.8	0.017	0.031	NA	38	OBERBAYERN	APR94	OB
3	MUENCHEN	0.005	1.0	0.021	0.051	0.055	39	OBERBAYERN	APR94	OB
4	TROSTBERG	0.003	0.5	0.007	0.019	0.062	25	OBERB		OB
5	OBERAUDORF	0.004	0.5	0.005	0.019	0.056	[illegible]	OBERB		OB
6	KELHEIM	0.009	0.5	0.017	0.028	NA	38	NIEDER		NB
7	LANDSHUT	0.003	1.2	0.016	0.030	NA	43	NIEDER		NB
8	PASSAU	0.008	0.6	0.022	0.035	0.037	49	NIEDER		NB
9	REGENSBURG	0.011	0.9	0.031	0.053	0.036	52	OBERPF		OP
10	WEIDEN	0.017	0.8	0.028	0.041	NA	49	OBERPF		OP
11	BAYREUTH	0.019	1.2	0.038	0.013	NA	45	OBERFF		OF
12	HOF	0.018	0.3	0.022	0.034	0.048	47	OBERFF		OF
13	KULMBACH	0.013	0.7	0.013	0.033	0.050	40	OBERFRANKEN	APR94	OF
14	ANSBACH	0.008	0.7	0.018	0.008	0.057	43	MITTELFRANKEN	APR94	MF
15	FUERTH	0.017	1.2	NA	NA	NA	43	MITTELFRANKEN	APR94	MF
16	NUERNBERG	0.015	1.0	0.029	0.029	0.041	38	MITTELFRANKEN	APR94	MF
17	ASCHAFFENBURG	0.010	NA	0.023	0.039	NA	32	UNTERFRANKEN	APR94	UF
18	SCHWEINFURT	0.010	0.8	0.031	0.042	0.049	50	UNTERFRANKEN	APR94	UF
19	WUERZBURG	0.011	0.7	0.026	0.012	NA	43	UNTERFRANKEN	APR94	UF
20	AUGSBURG	0.013	1.6	0.088	0.075	NA	59	SCHWABEN	APR94	SW
21	KEMPTEN	0.004	0.8	0.010	0.023	0.063	30	SCHWABEN	APR94	SW
22	NEU-ULM	0.008	0.8	0.015	0.038	0.054	40	SCHWABEN	APR94	SW
23	BURGHAUSEN	0.003	0.5	0.006	0.016	0.063	31	OBERBAYERN	JUL93	OB
24	INGOLSTADT	0.005	0.7	0.015	0.026	NA	35	OBERBAYERN	JUL93	OB
25	MUENCHEN	0.004	1.0	0.016	0.043	0.055	39	OBERBAYERN	JUL93	OB
26	TROSTBERG	0.002	0.4	0.006	0.018	0.060	21	OBERBAYERN	JUL93	OB
27	OBERAUDORF	0.002	0.4	0.006	0.017	0.048	26	OBERBAYERN	JUL93	OB
28	KELHEIM	0.004	0.5	0.019	0.023	NA	36	NIEDERBAYERN	JUL93	NB
29	LANDSHUT	0.002	1.0	0.023	0.028	NA	40	NIEDERBAYERN	JUL93	NB
30	PASSAU	0.004	0.5	0.016	0.026	0.040	52	NIEDERBAYERN	JUL93	NB
31	REGENSBURG	0.006	0.8	0.023	0.049	0.039	43	OBERPFALZ	JUL93	OP
32	WEIDEN	0.006	0.6	0.016	0.032	NA	39	OBERPFALZ	JUL93	OP

MI – Advantages & Disadvantages

+ standard complete-data procedures are applicable again

MI – Advantages & Disadvantages

+ standard complete-data procedures are applicable again
+ the 'most natural way to display [...] sensitivity' (Rubin, 1978)

MI – Advantages & Disadvantages

+ standard complete-data procedures are applicable again
+ the 'most natural way to display [...] sensitivity' (Rubin, 1978)
+ easy to implement

MI – Advantages & Disadvantages

- + standard complete-data procedures are applicable again
- + the 'most natural way to display [...] sensitivity' (Rubin, 1978)
- + easy to implement
- + the knowledge of the data collector that goes into the process of creating appropriate imputes

MI – Advantages & Disadvantages

+ standard complete-data procedures are applicable again
+ the 'most natural way to display [...] sensitivity' (Rubin, 1978)
+ easy to implement
+ the knowledge of the data collector that goes into the process of creating appropriate imputes

– increased running time

Outline

Prior- & Posterior Distribution

$$\pi(\vartheta|Y_{obs}, D) \equiv \pi(\vartheta|Y_{obs}) = constant \times \pi(\vartheta) \times f(Y_{obs}|\vartheta)$$

Expected Value & Variance

3.3 Proposition: (Approximated Expected Value and Variance)

The expected values of ϑ can be approximated as

$$E_{\vartheta}(\vartheta | Y_{obs}) \approx \bar{\vartheta},$$

where $\bar{\vartheta}$ is the corresponding MI estimator.

Expected Value & Variance

3.3 Proposition: (Approximated Expected Value and Variance)
The expected values of ϑ can be approximated as

$$E_{\vartheta}(\vartheta | Y_{obs}) \approx \bar{\vartheta},$$

where $\bar{\vartheta}$ is the corresponding MI estimator. Similarly we can approximate

$$Var_{\vartheta}(\vartheta | Y_{obs}) \approx \frac{1}{D} \sum_{d=1}^{D} Var_{\vartheta}(\vartheta | Y_{mis}^{(d)}, Y_{obs}) + \frac{1}{D-1} \sum_{d=1}^{D} [E_{\vartheta}(\vartheta | Y_{mis}^{(d)}, Y_{obs}) - \bar{\vartheta}]^2.$$

Within- and Between Variability

3.4 Definition: (Within- and Between Variability)

The summands of Prop. (3.3) can be denoted as

$$\hat{W} := \frac{1}{D} \sum_{d=1}^{D} Var_{\vartheta}(\vartheta | Y_{mis}^{(d)}, Y_{obs})$$

$$\hat{B} := \frac{1}{D-1} \sum_{d=1}^{D} [E_{\vartheta}(\vartheta | Y_{mis}^{(d)}, Y_{obs}) - \bar{\vartheta}]^2 = \frac{1}{D-1} \sum_{d=1}^{D} (\hat{\vartheta}_d - \bar{\vartheta})^2.$$

Asymptotic Distribution and finite Imputations Correction

3.5 Corollary:

Let $(\hat{\vartheta}_{(d)})_{n \in \mathbb{N}}$ be a sequence of random variables, that are i.i.d. with $\sigma^2 = Var(\hat{\vartheta}_{(1)}) = \hat{B} + \hat{W} < \infty$ and $\mu = E(\hat{\vartheta}_{(1)}) = \vartheta$. If $\hat{\vartheta}_{(d)}$ is a random vector and $\hat{\vartheta}_{(1)}, \hat{\vartheta}_{(2)}, \hat{\vartheta}_{(3)}, \ldots$ i.i.d., then for $D \longrightarrow \infty$, $\bar{\vartheta}$ is distributed as:

$$\bar{\vartheta} \overset{as}{\sim} N(\vartheta, \hat{B} + \hat{W}).$$

Asymptotic Distribution and finite Imputations Correction

3.6 Proposition: (Total Variance for finite D)

The total variance for finite D can be written as

$$\hat{V}_D := (1 + D^{-1})\,\hat{B} + \hat{W}.$$

For $D \longrightarrow \infty$ we get $\hat{V} := \hat{B} + \hat{W}$.

The algorithmic Search for the optimal Number of Imputations

Outline

The Algorithm

1. **Start.** Select an initial number of imputed datasets, D_0, $\bar{\vartheta}_{D_0} = \sum_{i=1}^{D_0} \hat{\vartheta}_i / D_0$
2. **Update.** For $D > D_0$,

$$\bar{\vartheta}_{D+1} = \frac{D\,\bar{\vartheta}_D + \hat{\vartheta}_{D+1}}{D+1}$$

3. **Distance.** Compute: $d_{D+1} = d(\bar{\vartheta}_{D+1}, \bar{\vartheta}_D)$ using an appropriate distance.
4. **Stopping rule.** $d_j < \varepsilon$ for $j = D+1, ..., D+k_0$

(Nassiri et al., 2020)

Outline

Simulation Set Up

Settings:

- n-dimensional random vector: $\mathbf{Y}_i \sim N(\mu \mathbf{1}_n, \sigma^2 I_n + \tau J_n)$
- 100 random draws
- create missing data with *mice*
- create imputed data sets with *Amelia*
- variate the following parameters: Missing data percentage, ϱ, ε, k_0

Simulation Set Up

Settings:

- n-dimensional random vector: $\mathbf{Y}_i \sim N(\mu\mathbf{1}_n, \sigma^2 I_n + \tau J_n)$
- 100 random draws
- create missing data with *mice*
- create imputed data sets with *Amelia*
- variate the following parameters: Missing data percentage, ϱ, ε, k_0

The Models:

(i) Compound-Symmetry (estimated parameters: μ, σ^2, τ)

Simulation Set Up

Settings:

- n-dimensional random vector: $\mathbf{Y}_i \sim N(\mu\mathbf{1}_n, \sigma^2 I_n + \tau J_n)$
- 100 random draws
- create missing data with *mice*
- create imputed data sets with *Amelia*
- variate the following parameters: Missing data percentage, ϱ, ε, k_0

The Models:

(i) Compound-Symmetry (estimated parameters: μ, σ^2, τ)

(ii) Logistic Regression (estimated parameters: β_1, β_2, β_3)

Simulation Results

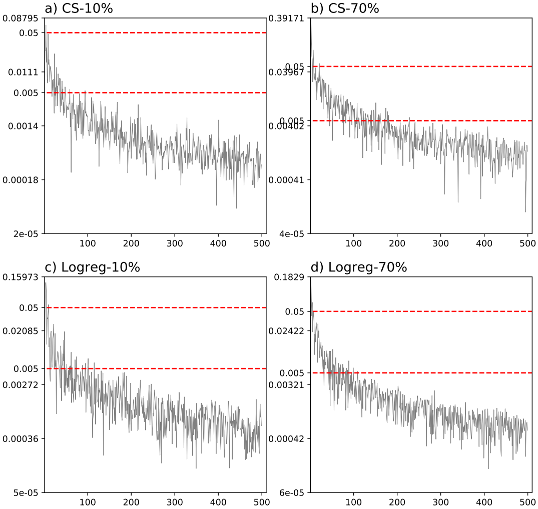


Figure: Convergence rates of the distances for CS and Logreg for $\sigma^2 = 0.25$, $\varrho = 0.1,\ k_0 = 5$ and $\beta = (0.2, -2, 0.5)^T$.

Simulation Results

Model	ϱ	ε	Mean	SD	$\mu(\beta_1)$ MAD	$\sigma^2(\beta_2)$ MAD	$\tau(\beta_3)$ MAD
			$k_0 = 1$				
CS-10%	0.1	0.005	25.55	10.11	0.02	0.02	0.01
		0.05	4.98	1.66	0.02	0.01	0.01
	0.9	0.005	12.86	4.79	0.14	0.02	0.29
		0.05	3.7	1.01	0.12	0.01	0.26
CS-70%	0.1	0.005	53.58	16.97	0.03	0.02	0.01
		0.05	10.28	3.52	0.03	0.02	0.01
	0.9	0.005	25.42	9.33	0.13	0.01	0.26
		0.05	5.72	2.49	0.12	0.02	0.25
Logreg-10%	0.1	0.005	22.57	9.01	0.11	0.14	0.06
		0.05	4.82	1.6	0.10	0.13	0.05
	0.9	0.005	22.16	9.26	0.11	0.15	0.11
		0.05	4.93	1.79	0.1	0.16	0.11
Logreg-70%	0.1	0.005	22.21	8.57	0.1	0.14	0.06
		0.05	4.91	1.46	0.1	0.15	0.05
	0.9	0.005	24.45	9.92	0.1	0.17	0.11
		0.05	4.65	1.48	0.1	0.15	0.11

(a) Validation steps: $k_0 = 1$

Model	ϱ	ε	Mean	SD	$\mu(\beta_1)$ MAD	$\sigma^2(\beta_2)$ MAD	$\tau(\beta_3)$ MAD
			$k_0 = 5$				
CS-10%	0.1	0.005	65.36	19.83	0.02	0.01	0.01
		0.05	7.73	2.67	0.02	0.01	0.01
	0.9	0.005	42.42	17.87	0.12	0.01	0.29
		0.05	5.36	2.43	0.12	0.01	0.23
CS-70%	0.1	0.005	152.23	27.11	0.02	0.01	0.01
		0.05	19.06	4.47	0.02	0.01	0.01
	0.9	0.005	91.55	26.09	0.11	0.02	0.24
		0.05	12.94	4.85	0.12	0.01	0.28
Logreg-10%	0.1	0.005	57.61	18.04	0.1	0.14	0.06
		0.05	6.52	2.42	0.1	0.12	0.06
	0.9	0.005	57.88	20.02	0.11	0.14	0.1
		0.05	7.56	3.43	0.11	0.16	0.12
Logreg-70%	0.1	0.005	60.93	17.99	0.1	0.14	0.05
		0.05	7.04	2.94	0.11	0.14	0.06
	0.9	0.005	60.7	22.02	0.11	0.16	0.11
		0.05	7.49	2.85	0.11	0.15	0.1

(b) Validation steps: $k_0 = 5$

Figure: Mean, SD and their mean absolute deviation from the true parameter (β_i corresponds to the Logreg model and μ, σ^2, τ the CS model) for selected D given $\sigma^2 = 0.25$ and different values for ε and ϱ using the Mahalanobis distance with $S = \hat{V}$.

Thank you for your attention!

References

- Hermans, L., Nassiri, V., Molenberghs, G., Kenward, M. G., Van der Elst, W., Aerts, M., Verbeke, G. (2019). *Clusters with unequal Size: Maximum Likelihood versus weighted Estimation in large Samples*. Statistica Sinica (forthcoming).
- Hosmer, D. W. (2013). *Applied Logistic Regression* (3rd ed.), John Wiley & Sons, New York.
- James, G., Witten, D., Hastie, T., Tibshirani, R. (2013). *An introduction to statistical learning* (2nd ed.), Springer, New York.
- Nassiri, V., Molenberghs, G., Verbeke, G., Barbosa-Breda, J. (2020). *Iterative Multiple Imputation: A Framework to Determine the Number of Imputed Datasets*. The American Statistician 74, 125-136.

References

- Rubin, D. B. (1978). *Multiple Imputations in Sample Surveys—A Phenomenological Bayesian Approach to Nonresponse*. Proceedings of the Survey Research Methods Section of the American Statistical Association 1, American Statistical Association, 20–34
- Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*, John Wiley & Sons, New York.
- Rubin, D. B., Little, R. (2002). *Statistical Analysis with Missing Data* (2nd ed.), John Wiley & Sons, New York.