Imputacja brakujących danych binarnych w modelu autologistycznym [1]

Marta Zalewska

Warszawski Uniwesytet Medyczny

Statystyka Matematyczna
Wisła, grudzień 2009

[1] Współautorzy: Wojciech Niemiro, UMK Toruń i UW, Bolesław Samoliński, WUM

Praca częściowo finansowana przez Grant No. N N206 356036.

Plan

Plan

Plan

Plan

Imputacja Bayesowska i Próbnik Gibbsa

Dane:

$$X = (X_{\text{obs}}, X_{\text{mis}}) \sim p_\beta.$$

- X_{obs} – obserwowane dane
- X_{mis} – brakujące dane
- $\beta \sim \pi(\cdot)$ – parametr o rozkładzie *a priori*

Próbnik Gibbsa. Powtarzamy dwa kroki:

- generujemy β z rozkładu $\pi(\beta|X_{\text{obs}}, X_{\text{mis}})$,
- generujemy X_{mis} z rozkładu $\pi(X_{\text{mis}}|X_{\text{obs}}, \beta)$.

Metoda MCMC: łańcuch Markowa zbieżny do $\pi(X_{\text{mis}}, \beta|X_{\text{obs}})$.

Imputacja Bayesowska i Próbnik Gibbsa

Dane:

$$X = (X_{\text{obs}}, X_{\text{mis}}) \sim p_\beta.$$

- X_{obs} – obserwowane dane
- X_{mis} – brakujące dane
- $\beta \sim \pi(\cdot)$ – parametr o rozkładzie *a priori*

Próbnik Gibbsa. Powtarzamy dwa kroki:

- generujemy β z rozkładu $\pi(\beta|X_{\text{obs}}, X_{\text{mis}})$,
- generujemy X_{mis} z rozkładu $\pi(X_{\text{mis}}|X_{\text{obs}}, \beta)$.

Metoda MCMC: łańcuch Markowa zbieżny do $\pi(X_{\text{mis}}, \beta|X_{\text{obs}})$.

Model

Rozkład autologistyczny na $\mathcal{X} = \{0,1\}^d$:

$$p_\beta(x) := \frac{1}{Z(\beta)} \exp\left(\sum_{i,j=1}^{d} \beta_{ij} x_i x_j\right),$$

$x \sim \mathrm{AL}(\beta)$. Parametry: macierz symetryczna $\beta = (\beta_{ij})$.
Pełne rozkłady warunkowe:

$$p_\beta(x_i = 1 \mid x_{-i}) = \frac{\exp\left(\beta_{ii} + \sum_{j \neq i} x_j \beta_{ij}\right)}{1 + \exp\left(\beta_{ii} + \sum_{j \neq i} x_j \beta_{ij}\right)},$$

gdzie $x_{-i} = (x_j, j \neq i)$.
Rozkład taki sam jak w standardowym modelu regresji logistycznej.

Model

Rozkład autologistyczny na $\mathcal{X} = \{0,1\}^d$:

$$p_\beta(x) := \frac{1}{Z(\beta)} \exp\left(\sum_{i,j=1}^{d} \beta_{ij} x_i x_j\right),$$

$x \sim \mathrm{AL}(\beta)$. Parametry: macierz symetryczna $\beta = (\beta_{ij})$.
Pełne rozkłady warunkowe:

$$p_\beta(x_i = 1 \mid x_{-i}) = \frac{\exp\left(\beta_{ii} + \sum_{j \neq i} x_j \beta_{ij}\right)}{1 + \exp\left(\beta_{ii} + \sum_{j \neq i} x_j \beta_{ij}\right)},$$

gdzie $x_{-i} = (x_j, j \neq i)$.
Rozkład taki sam jak w standardowym modelu regresji logistycznej.

Symulacja i estymacja

Wniosek:

- symulowanie x jest łatwe przy pomocy próbnika Gibbsa,
- parametry β można estymować standardowymi metodami GLM.

Próbka: $x(1), \ldots, x(n)$ i.i.d. $\sim \mathrm{AL}(\beta)$.

Symulacja i estymacja

Wniosek:

- symulowanie x jest łatwe przy pomocy próbnika Gibbsa,
- parametry β można estymować standardowymi metodami GLM.

Próbka: $x(1), \ldots, x(n)$ i.i.d. $\sim \mathrm{AL}(\beta)$.

Maksimum pseudowiarogodności

Cząstkowe wiarogodności:

$$L_i(\beta|x) = L_i(\beta_i|x) := \log p_\beta(x_i \mid x_{-i}).$$

Pseudowiarogodność:

$$L_{\mathrm{ps}}(\beta|X) = \sum_{k=1}^{n} \sum_{i=1}^{d} L_i(\beta|x(k)).$$

$X = (x(1), \ldots, x(n))$ – próbka.

Maksimum wiarogodności

Estymator największej wiarogodności via MCMC: Geyer i Thopmpson (1992, JRSS).

Rodzina wykładnicza:

$$p_\beta(x) = \frac{1}{Z(\beta)} \mathrm{e}^{\beta^T T(x)},$$

gdzie $T(x)$ – wektor statystyk dostatecznych.

$$Z(\beta) = \sum_{x^*} \mathrm{e}^{\beta^T T(x^*)} = \sum_{x^*} \mathrm{e}^{(\beta-\beta^*)^T T(x^*)} p_{\beta^*}(x^*) Z(\beta^*)$$
$$= \mathbb{E}\mathrm{e}^{(\beta-\beta^*)^T T(x^*)} \times Z(\beta^*).$$

gdzie $x^* \sim p_{\beta^*}$.

Aproksymacja wiarogodności: $x^*(1), \ldots, x^*(n^*) \sim p_{\beta^*}$,

$$L_{\mathrm{MCMC}}(\beta|x) = \beta^T T(x) - \log \sum_{k=1}^{n^*} \mathrm{e}^{(\beta-\beta^*)^T T(x^*(k))} + \text{const.}$$

Maksimum wiarogodności

Estymator największej wiarogodności via MCMC: Geyer i Thopmpson (1992, JRSS).

Rodzina wykładnicza:

$$p_\beta(x) = \frac{1}{Z(\beta)} \mathrm{e}^{\beta^T T(x)},$$

gdzie $T(x)$ – wektor statystyk dostatecznych.

$$\begin{aligned} Z(\beta) &= \sum_{x^*} \mathrm{e}^{\beta^T T(x^*)} = \sum_{x^*} \mathrm{e}^{(\beta-\beta^*)^T T(x^*)} p_{\beta^*}(x^*) Z(\beta^*) \\ &= \mathbb{E}\mathrm{e}^{(\beta-\beta^*)^T T(x^*)} \times Z(\beta^*). \end{aligned}$$

gdzie $x^* \sim p_{\beta^*}$.

Aproksymacja wiarogodności: $x^*(1), \ldots, x^*(n^*) \sim p_{\beta^*}$,

$$L_{\mathrm{MCMC}}(\beta|x) = \beta^T T(x) - \log \sum_{k=1}^{n^*} \mathrm{e}^{(\beta-\beta^*)^T T(x^*(k))} + \text{const.}$$

Próbkowanie z rozkładu *a posteriori*

Wybieramy $\beta^* = \hat{\beta}_{\mathrm{ps}}$, $n^* = n$.

Estymator „jednokrokowy” Newtona-Raphsona

$$\hat{\beta} = \hat{\beta}_{\mathrm{ps}} - \nabla^2 L_{\mathrm{MCMC}}(\hat{\beta}_{\mathrm{ps}})^{-1} \nabla L_{\mathrm{MCMC}}(\hat{\beta}_{\mathrm{ps}}),$$

gdzie

- $\hat{\beta}_{\mathrm{ps}}$ – estymator największej pseudowiarogodności,
- L_{MCMC} – aproksymacja wiarogodności na podstawie „sztucznej próbki” $X^* = (x^*(1), \ldots, x^*(n))$ wygenerowanej z rozkładu $p_{\hat{\beta}_{\mathrm{ps}}}$.

$\hat{\beta}$ ma w przybliżeniu rozkład $\pi(\beta|X)$.

Próbkowanie z rozkładu *a posteriori*

Wybieramy $\beta^* = \hat{\beta}_{\mathrm{ps}}$, $n^* = n$.

Estymator „jednokrokowy” Newtona-Raphsona

$$\hat{\beta} = \hat{\beta}_{\mathrm{ps}} - \nabla^2 L_{\mathrm{MCMC}}(\hat{\beta}_{\mathrm{ps}})^{-1} \nabla L_{\mathrm{MCMC}}(\hat{\beta}_{\mathrm{ps}}),$$

gdzie

- $\hat{\beta}_{\mathrm{ps}}$ – estymator największej pseudowiarogodności,
- L_{MCMC} – aproksymacja wiarogodności na podstawie „sztucznej próbki” $X^* = (x^*(1), \ldots, x^*(n))$ wygenerowanej z rozkładu $p_{\hat{\beta}_{\mathrm{ps}}}$.

$\hat{\beta}$ ma w przybliżeniu rozkład $\pi(\beta|X)$.

Metodologia

Dane: ECAP: (Choroby alergiczne w Polsce, 2006-2008). 18617 jednostek (przypadków, respondentów) i 1225 zmiennych (większość binarnych, ale również liczbowe).

Do naszych eksperymentów wybraliśmy małą podmacierz: $n = 2962$ jednostek, $d = 6$ zmiennych, **bez brakujących danych**.

- generujemy sztucznie „braki”,
- imputujemy i estymujemy,
- porównujemy i sprawdzamy.

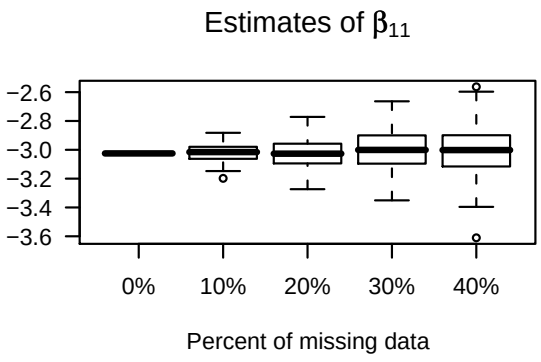
Metodologia

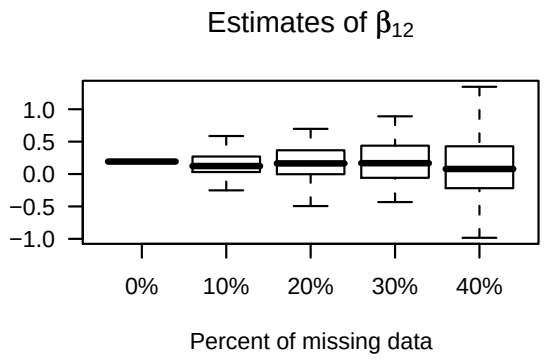
Dane: ECAP: (Choroby alergiczne w Polsce, 2006-2008). 18617 jednostek (przypadków, respondentów) i 1225 zmiennych (większość binarnych, ale również liczbowe).

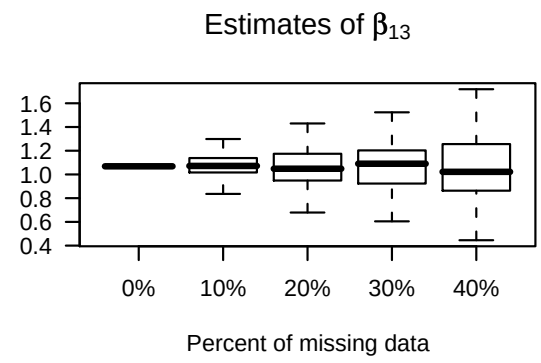
Do naszych eksperymentów wybraliśmy małą podmacierz: $n = 2962$ jednostek, $d = 6$ zmiennych, **bez brakujących danych**.

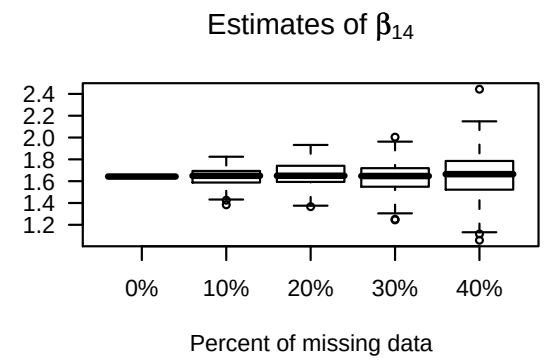
- generujemy sztucznie „braki”,
- imputujemy i estymujemy,
- porównujemy i sprawdzamy.

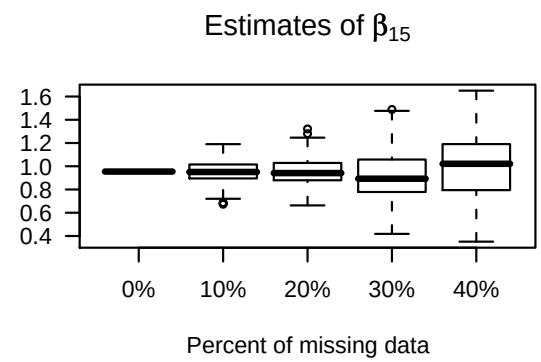
Estymatory

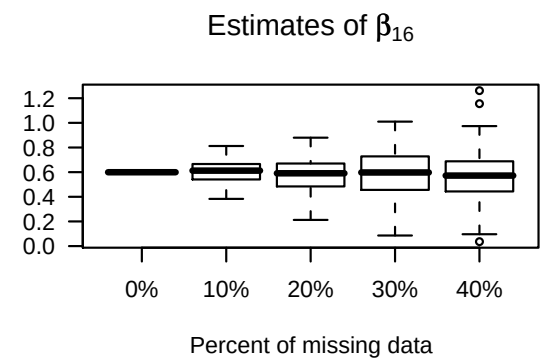

Estimates of β_{11}
−2.6
−2.8
−3.0
−3.2
−3.4
−3.6
0%
10%
20%
30%
40%
Percent of missing data


Estimates of β_{12}
1.0
0.5
0.0
−0.5
−1.0
0%
10%
20%
30%
40%
Percent of missing data

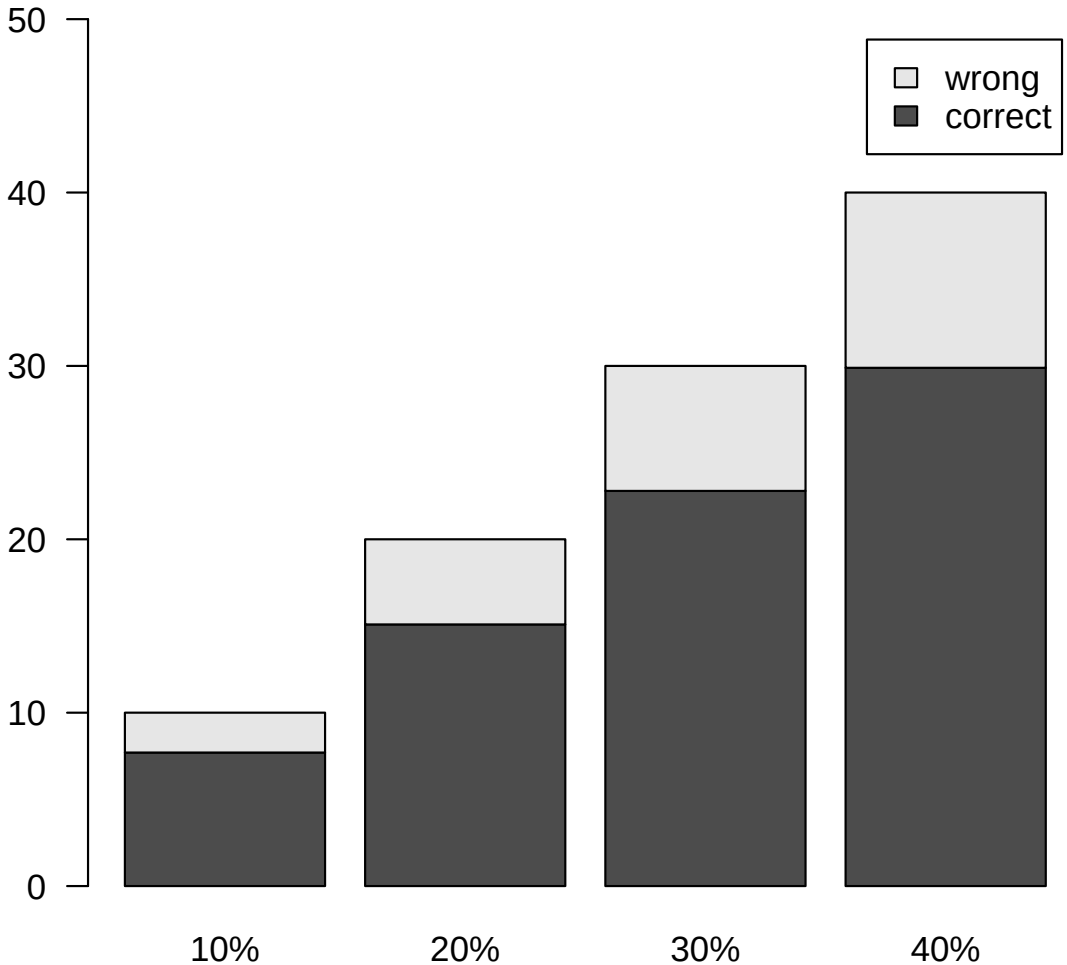

Estimates of β_{13}
1.6
1.4
1.2
1.0
0.8
0.6
0.4
0%
10%
20%
30%
40%
Percent of missing data


Estimates of β_{14}
2.4
2.2
2.0
1.8
1.6
1.4
1.2
0%
10%
20%
30%
40%
Percent of missing data


Estimates of β_{15}
1.6
1.4
1.2
1.0
0.8
0.6
0.4
0%
10%
20%
30%
40%
Percent of missing data


Estimates of β_{16}
1.2
1.0
0.8
0.6
0.4
0.2
0.0
0%
10%
20%
30%
40%
Percent of missing data

Rezultaty imputacji



Percent of missings correctly and wrongly imputed