I dati imputati: innocenti o colpevoli (update)

Massimiliano Pastore

PSICOSTAT meeting:
26 Novembre 2021

Outline

Introduzione

Data imputation: quello che ho capito

- In statistics, **imputation** is the process of replacing missing data with substituted values.
- By far, the most common means of dealing with missing data is **listwise deletion** which is when all cases with a missing value are deleted.
- **Single imputation** (e.g. mean substitution or regression) does not take into account the uncertainty in the imputations.
- **Multiple Imputation** (Rubin, 1987) has become a generally accepted way to handle statistical analysis of incomplete data (Koller-Meinfelder, 2009).

---

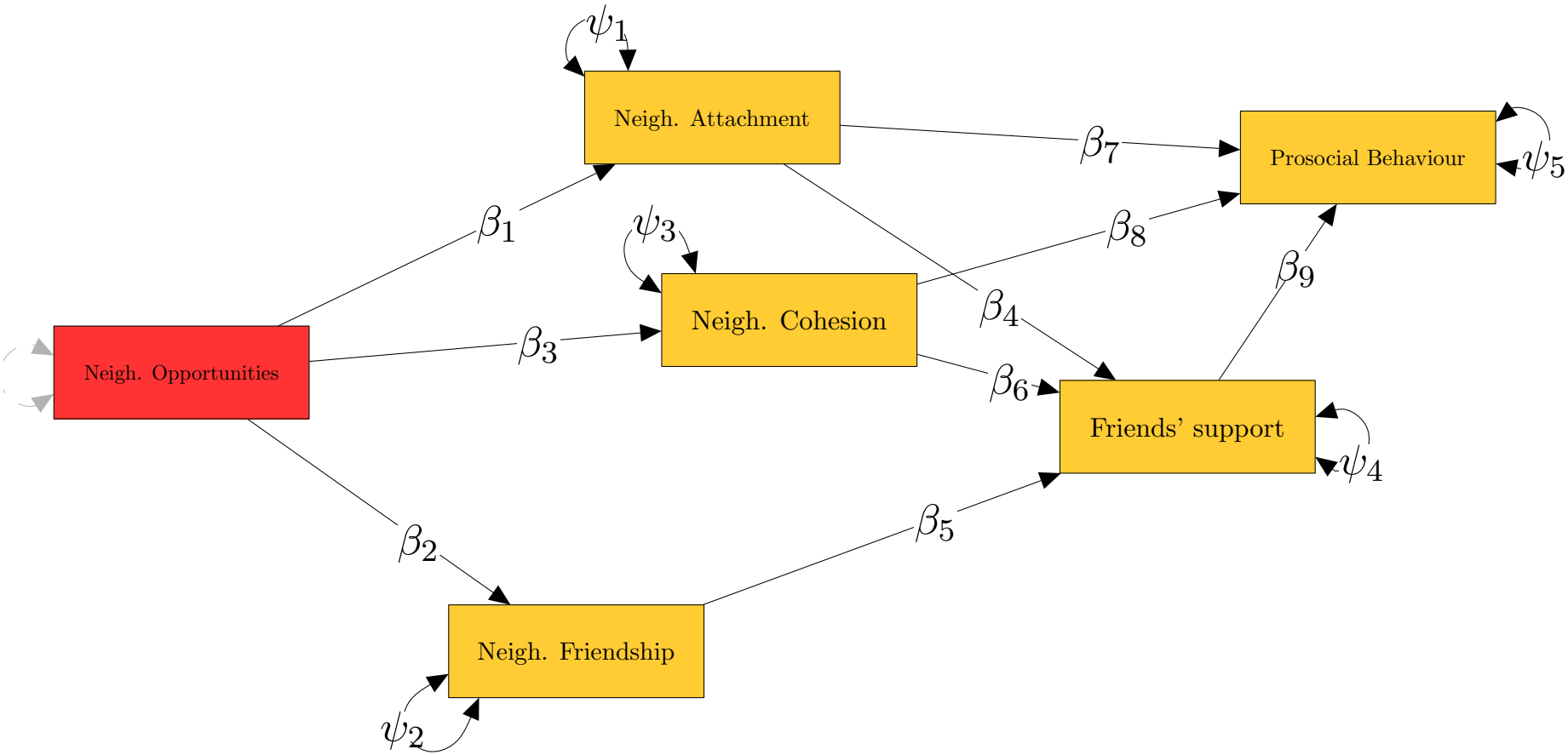
https://en.wikipedia.org/wiki/Imputation_(statistics)

Rubin, D.B. (1987). *Multiple imputation for survey nonresponse*. New York: Wiley.

Koller-Meinfelder, F. (2009). *Analysis of Incomplete Survey Data-Multiple Imputation via Bayesian Bootstrap Predictive Mean Matching* (Unpublished doctoral dissertation).

Qualche esempio pratico

Nel 2010



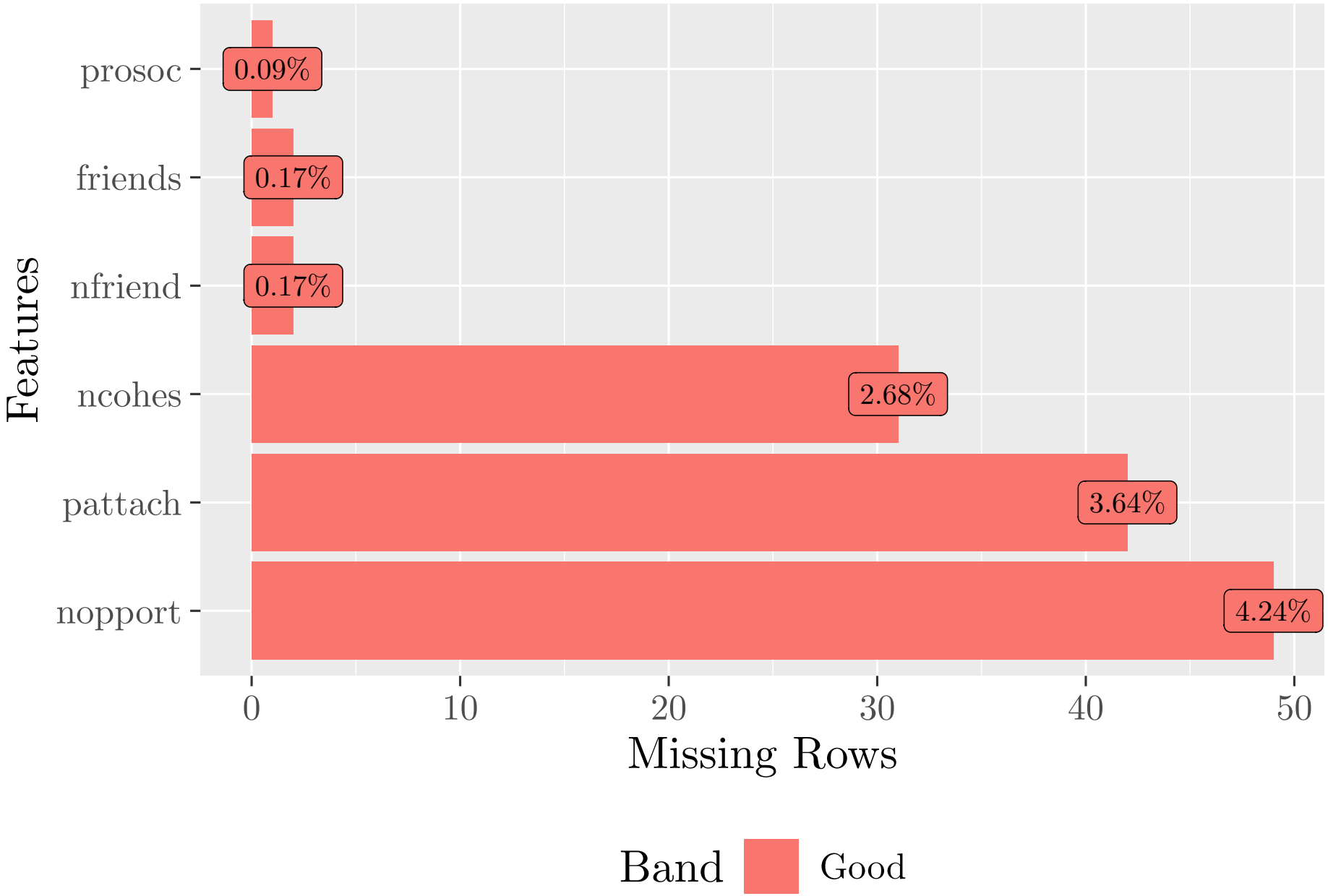
Lenzi, M., Vieno, A., Perkins, D. D., Pastore, M., Santinello, M., & Mazzardis, S. (2012). Perceived neighborhood social resources as determinants of prosocial behavior in early adolescence. *American journal of community psychology*, 50(1-2), 37-49.

Nel 2010

On pages 8-9, regarding the missing data, please describe the extent of the "excessive" missing data both for the 50 participants dropped from analyses as well as for the variables in the dataset. It would also be beneficial to explain why missing data procedures were not followed in the SEM modeling (such as multiple imputation or maximum likelihood estimation procedures) for missing observations, or preferably, to analyze with these procedures in order to provide the most unbiased parameter estimates.

Nota: il campione è composto da 1155 soggetti!!

Data Inspection



Data Imputation

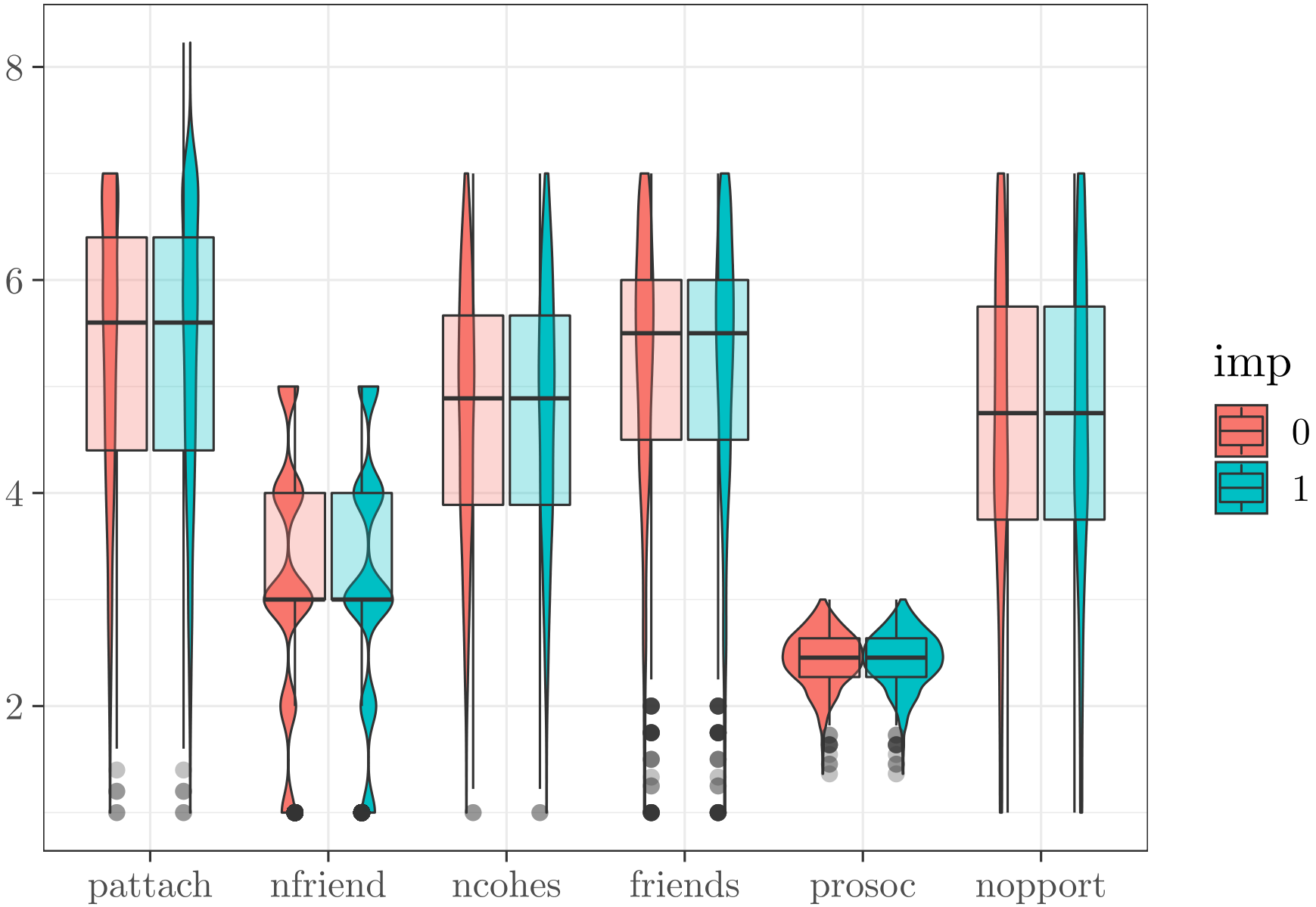
- Per l'imputazione dei dati mancanti abbiamo utilizzato un metodo *k-nearest neighbor* grazie al pacchetto `robCompositions` (Templ, Hron & Filzmoser, 2011).
- In pratica basta scrivere:

```
> Z <- impKNNa( data )
```

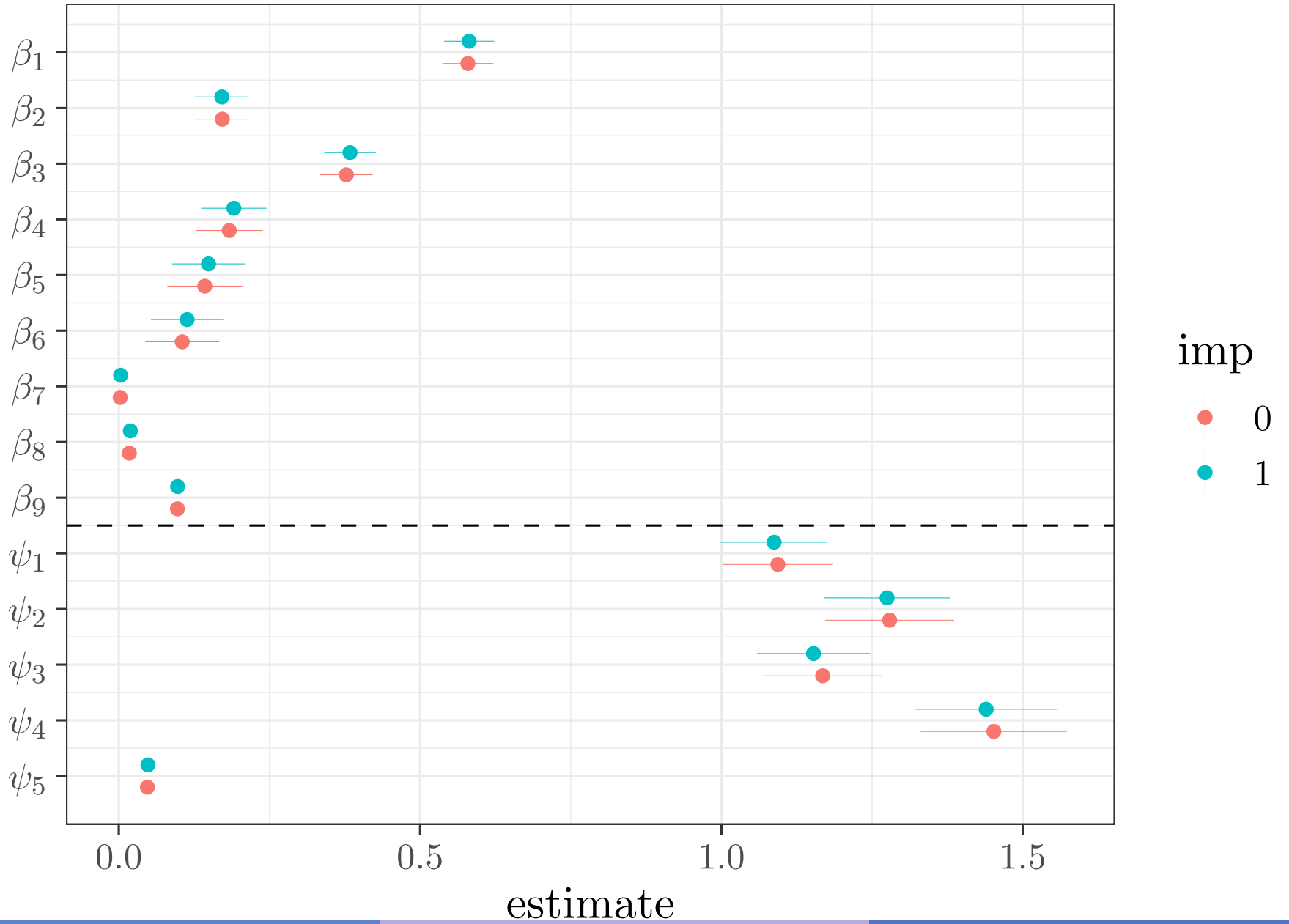
in cui `data` è il dataset con tutte le variabili implicate nel modello e tutti i soggetti.

Templ, M., Hron, K., Filzmoser, P. (2011). `robCompositions`: an R-package for robust statistical analysis of compositional data. In V. Pawlowsky-Glahn and A. Buccianti, editors, *Compositional Data Analysis. Theory and Applications*, pp. 341-355, John Wiley & Sons, Chichester (UK) ..

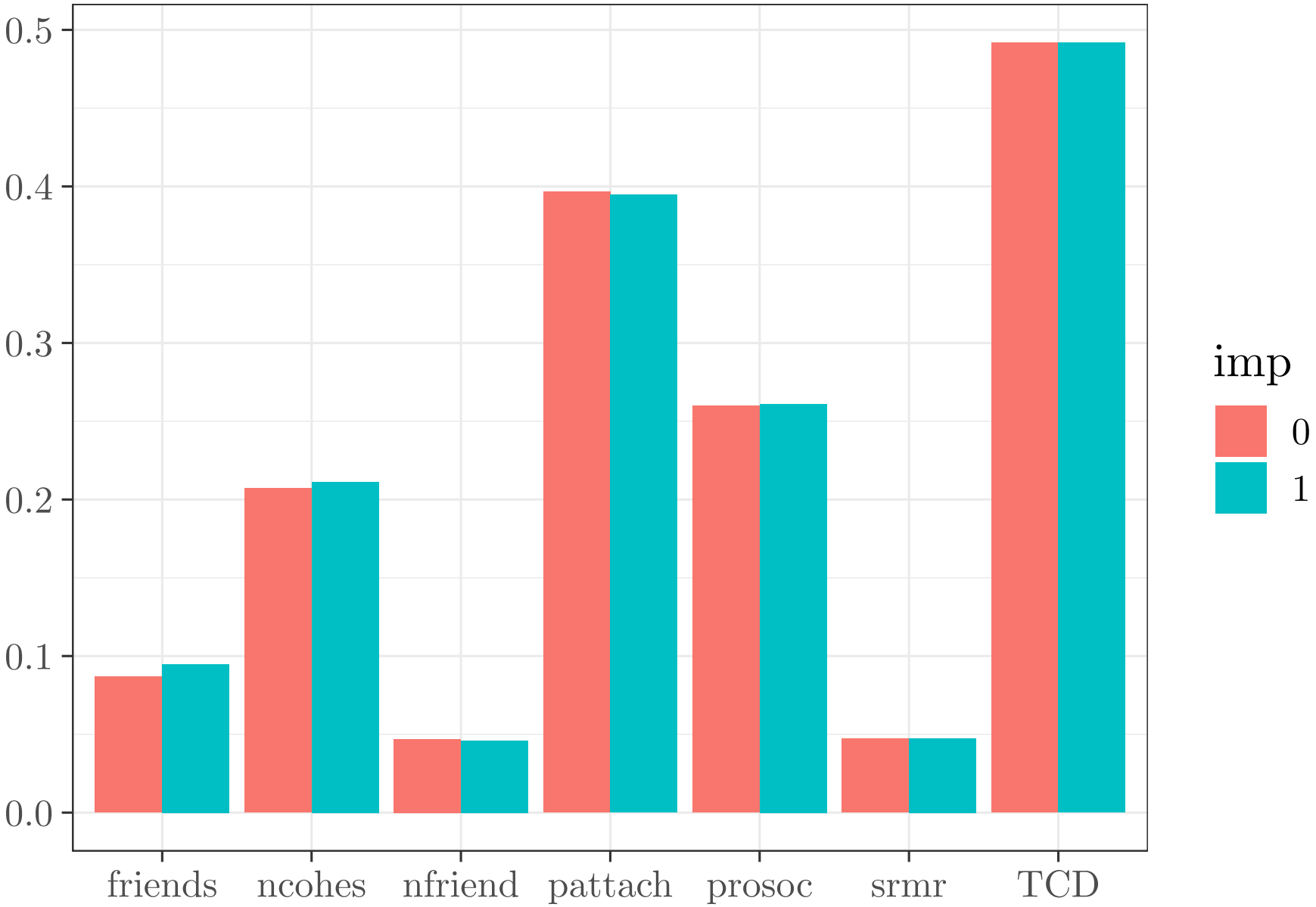
Imputation results: data



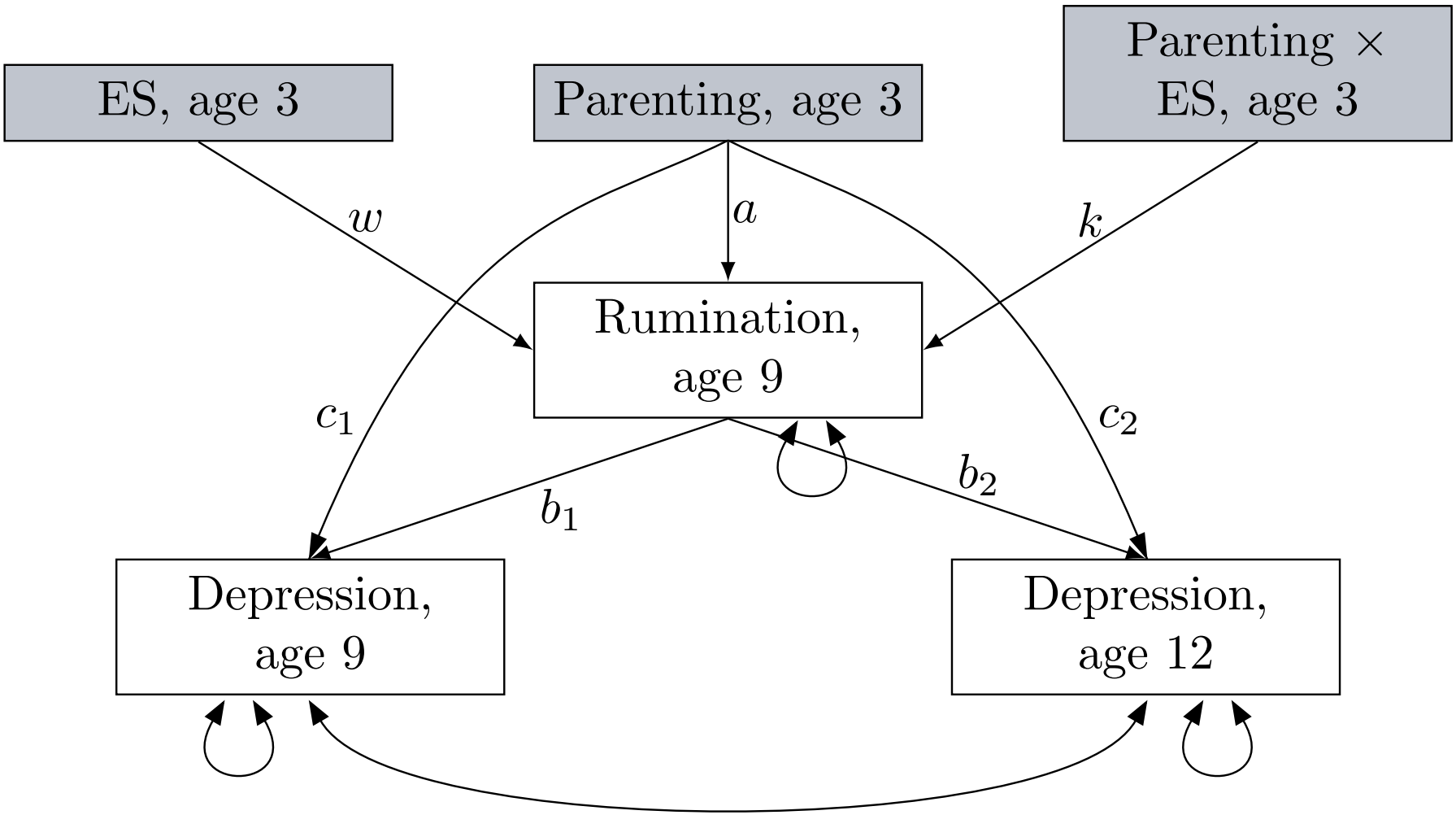
Imputation results: parameters



Imputation results: fit indices (R^2)



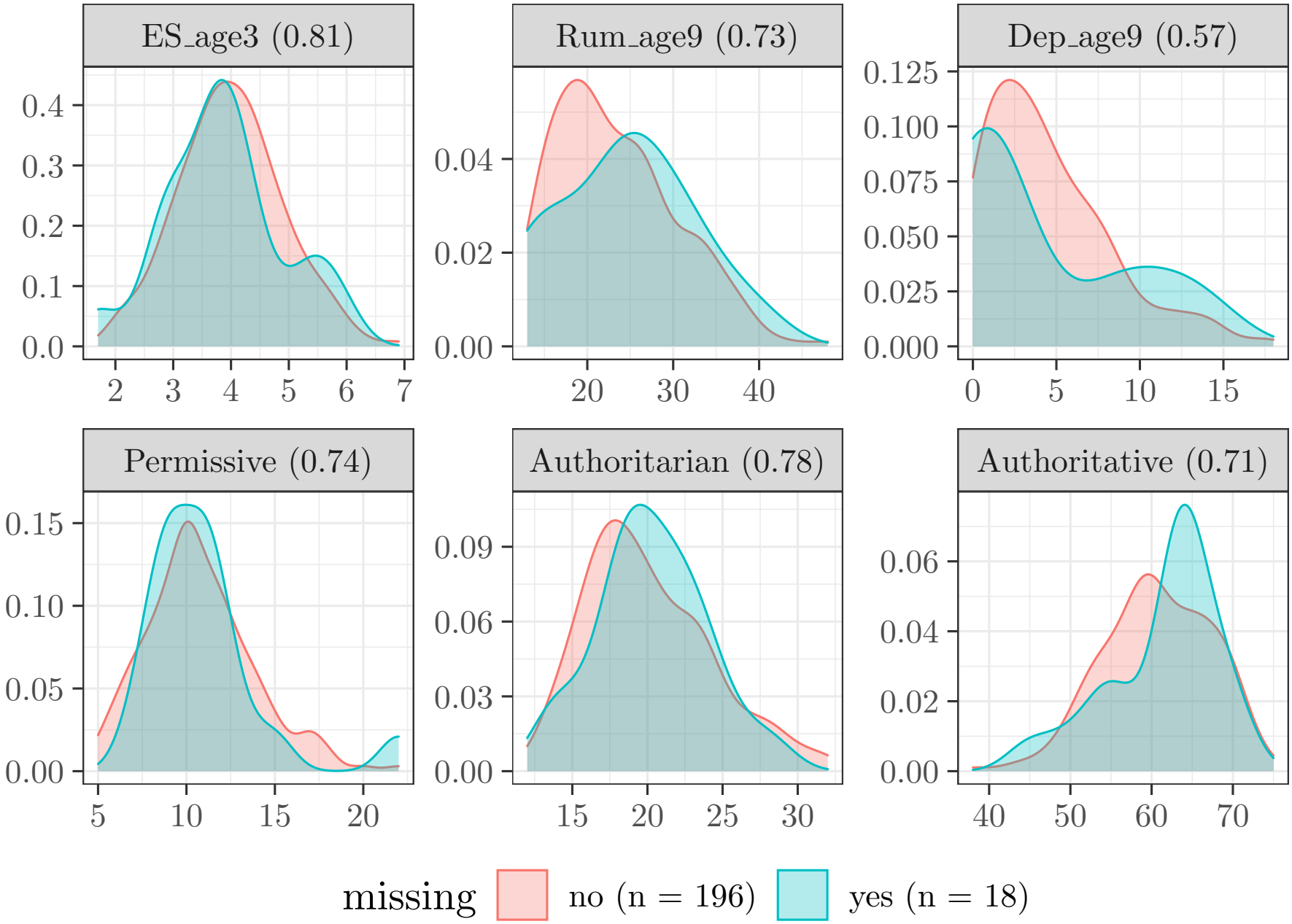
Nel 2020



Lionetti, F., Klein, D.N., Pastore, M. Aron, E. N., Aron, A., Pluess, M. (2021). The role of environmental sensitivity in the development of rumination and depressive symptoms in childhood: a longitudinal study. *European Child & Adolescent Psychiatry*, https://doi.org/10.1007/s00787-021-01830-6

- Il campione si compone di 214 soggetti.
- Per 18 di essi mancano i dati nella variabile *Depression, age 12*.

	vars	n	mean	sd	min	max	range	se
ES_age3	1	214	3.99	0.91	2	7	5	0.06
Rum_age9	2	214	23.50	7.08	13	48	35	0.48
Dep_age9	3	214	4.50	3.87	0	18	18	0.26
Dep_age12	4	196	4.55	5.39	0	28	28	0.39
Permissive	5	214	10.77	3.06	5	22	17	0.21
Authoritarian	6	214	20.09	4.13	12	32	20	0.28
Authoritative	7	214	60.75	6.65	38	75	37	0.45


ES_age3 (0.81)
Rum_age9 (0.73)
Dep_age9 (0.57)
Permissive (0.74)
Authoritarian (0.78)
Authoritative (0.71)
missing
no (n = 196)
yes (n = 18)

```
Model Info:
 function:     stan_glm
 family:       binomial [logit]
 formula:      is.na ~ ES_age3 + Rum_age9 + Dep_age9 +
   Permissive + Authoritarian + Authoritative
 algorithm:    sampling
 sample:       4000 (posterior sample size)
 priors:       see help('prior_summary')
 observations: 214
 predictors:   7

Estimates:
                mean   sd    5%    95%
(Intercept)   -3.40   3.21 -8.65  1.75
ES_age3       -0.25   0.29 -0.73  0.21
Rum_age9       0.04   0.04 -0.03  0.10
Dep_age9      -0.07   0.08 -0.20  0.05
Permissive     0.02   0.09 -0.13  0.16
Authoritarian  0.01   0.07 -0.11  0.11
Authoritative  0.02   0.04 -0.05  0.08
```

Data Imputation

- Per l'imputazione dei dati mancanti abbiamo utilizzato una procedura di *Multiple Imputation through Bayesian Bootstrap Predictive Mean Matching* grazie al pacchetto `BaBooN` (Meinfelder & Schnapp, 2015).
- In pratica basta scrivere:

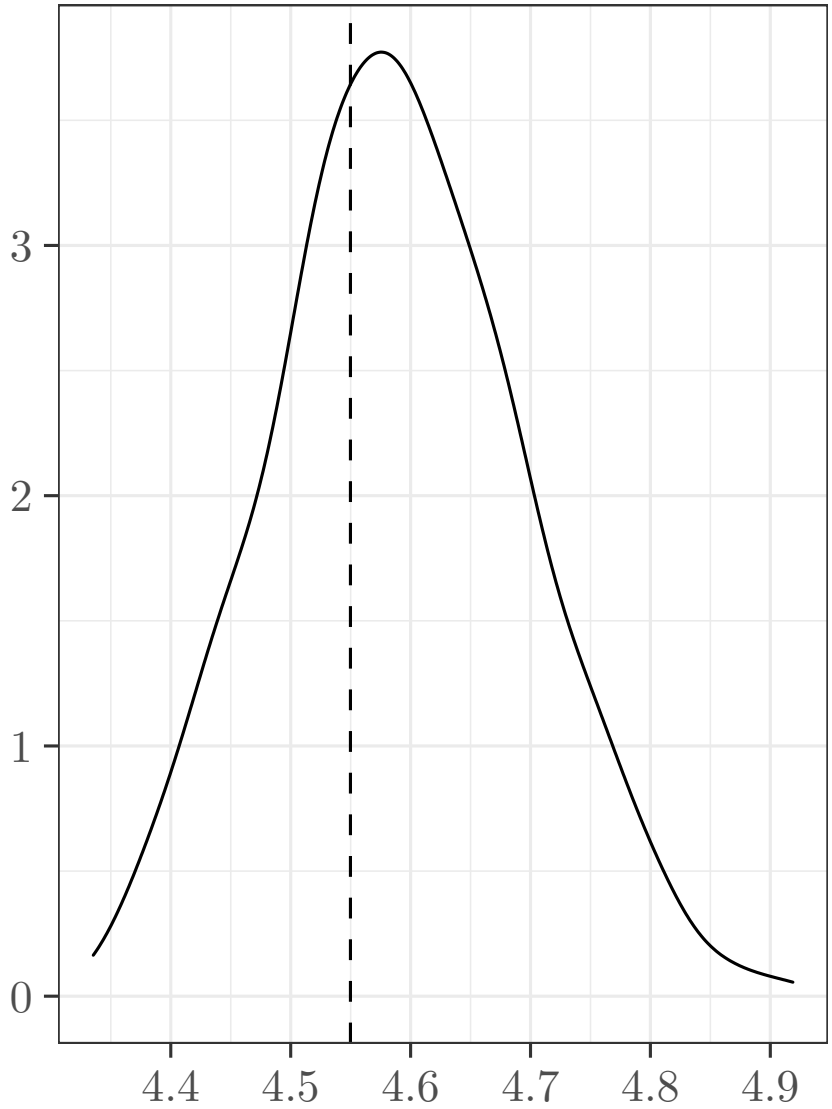
```
> BBPMM( data, M = M )
```

  in cui `data` è il dataset contenente la variabile contenente i missing e le variabili coinvolte nel processo e `M` il numero di imputazioni desiderate.
- Per valutare il funzionamento della procedura abbiamo ripetuto l'imputazione per 500 volte, ricalcolato la media della variabile `Dep_age12` e confrontato le distribuzioni dei valori ottenuti con quelli osservati.

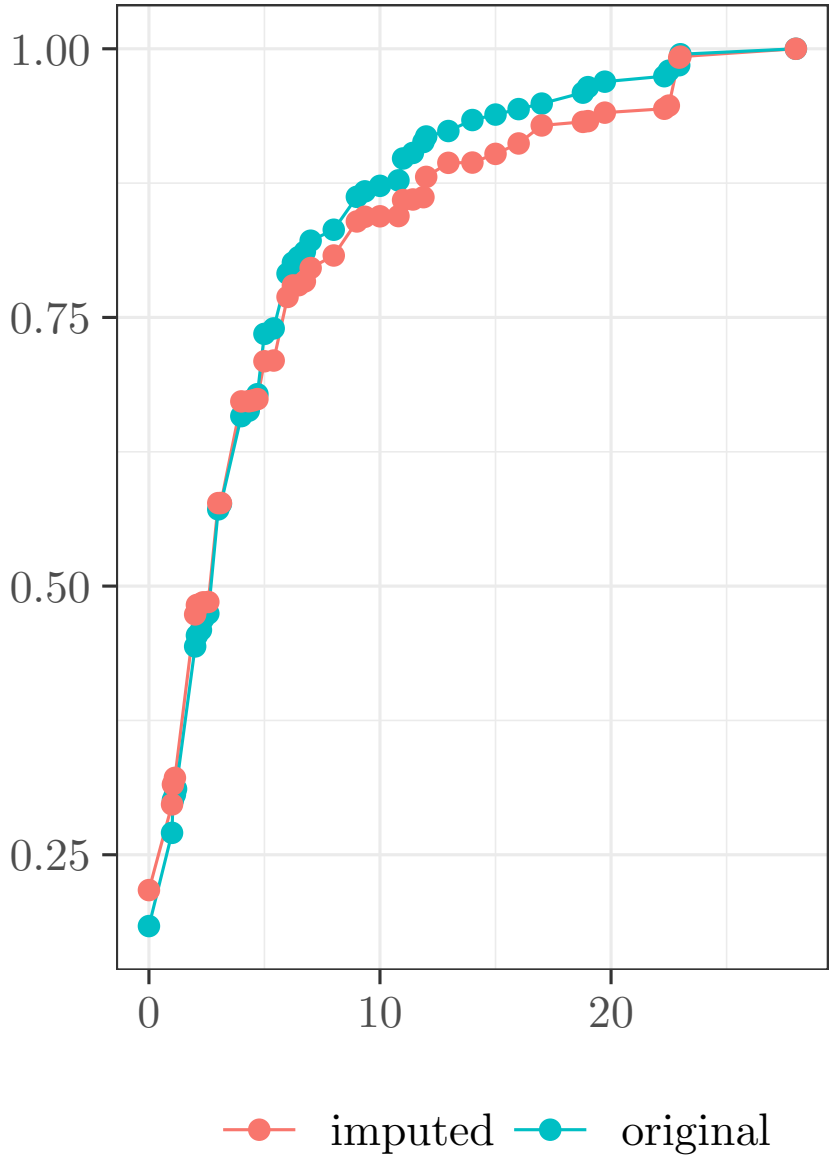
Meinfelder, F., & Schnapp, T. (2015). `BaBooN`: *Bayesian Bootstrap Predictive Mean Matching - Multiple and Single Imputation for Discrete Data*. Retrieved from `https://CRAN.R-project.org/package=BaBooN` (R package version 0.2-0).

Check on imputed data

[A] means



[B] cumulative



Sensitivity analysis

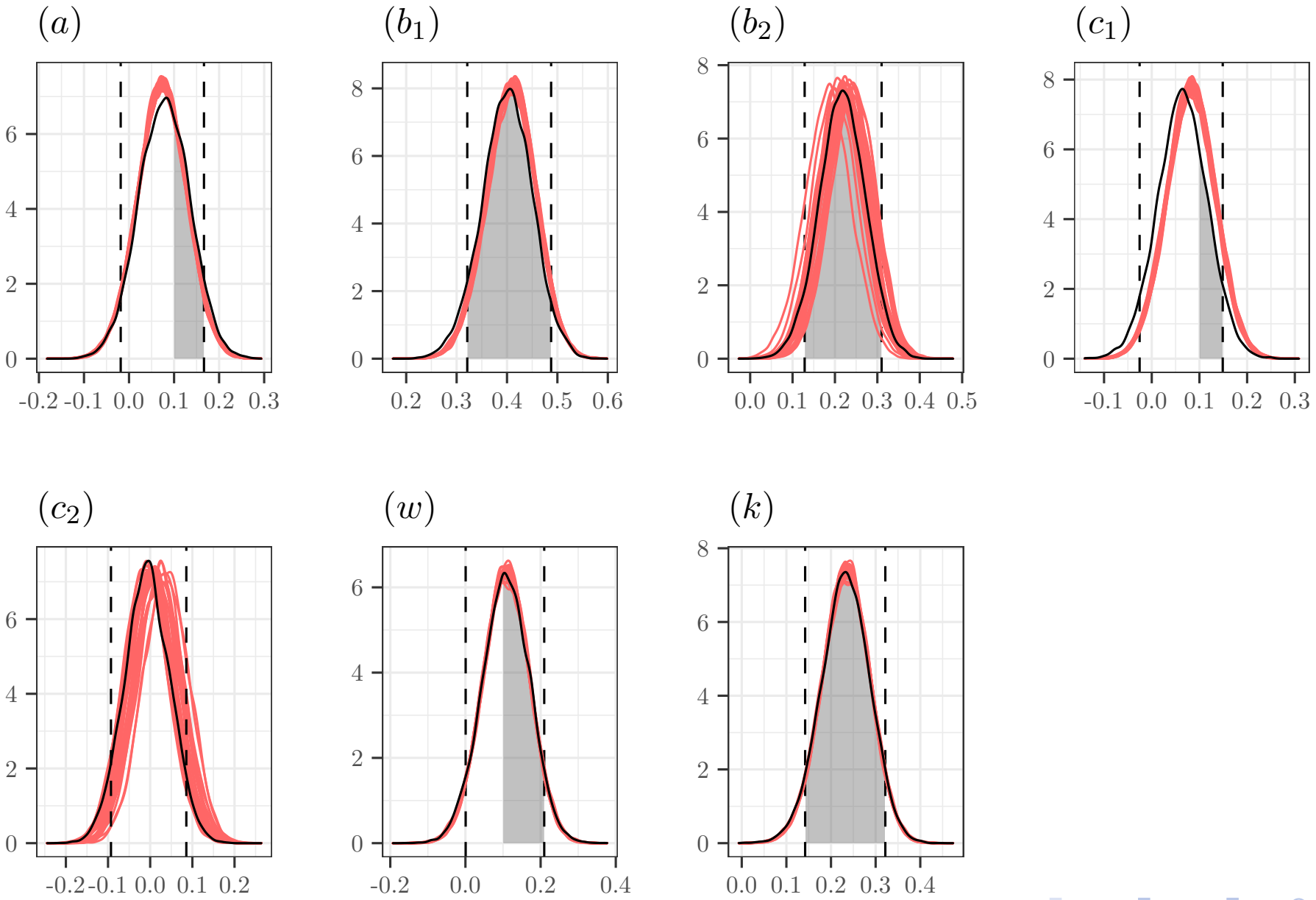
- A questo punto abbiamo selezionato 25 imputazioni e ristimato per altrettante volte le distribuzioni a posteriori dei parametri dei tre modelli con le stesse prior informative usate con i dati effettivi:

$$
\begin{aligned}
a &\sim \text{Normal}(0.1, 0.1) \\
b_1 &\sim \text{Normal}(0.5, 0.1) \\
b_2 &\sim \text{Normal}(0.35, 0.1) \\
c_1 &\sim \text{Normal}(\pm 0.1, 0.1) \\
c_2 &\sim \text{Normal}(\pm 0.05, 0.1) \\
w &\sim \text{Normal}(0, 0.2) \\
k &\sim \text{Normal}(0.3, 0.1)
\end{aligned}
$$

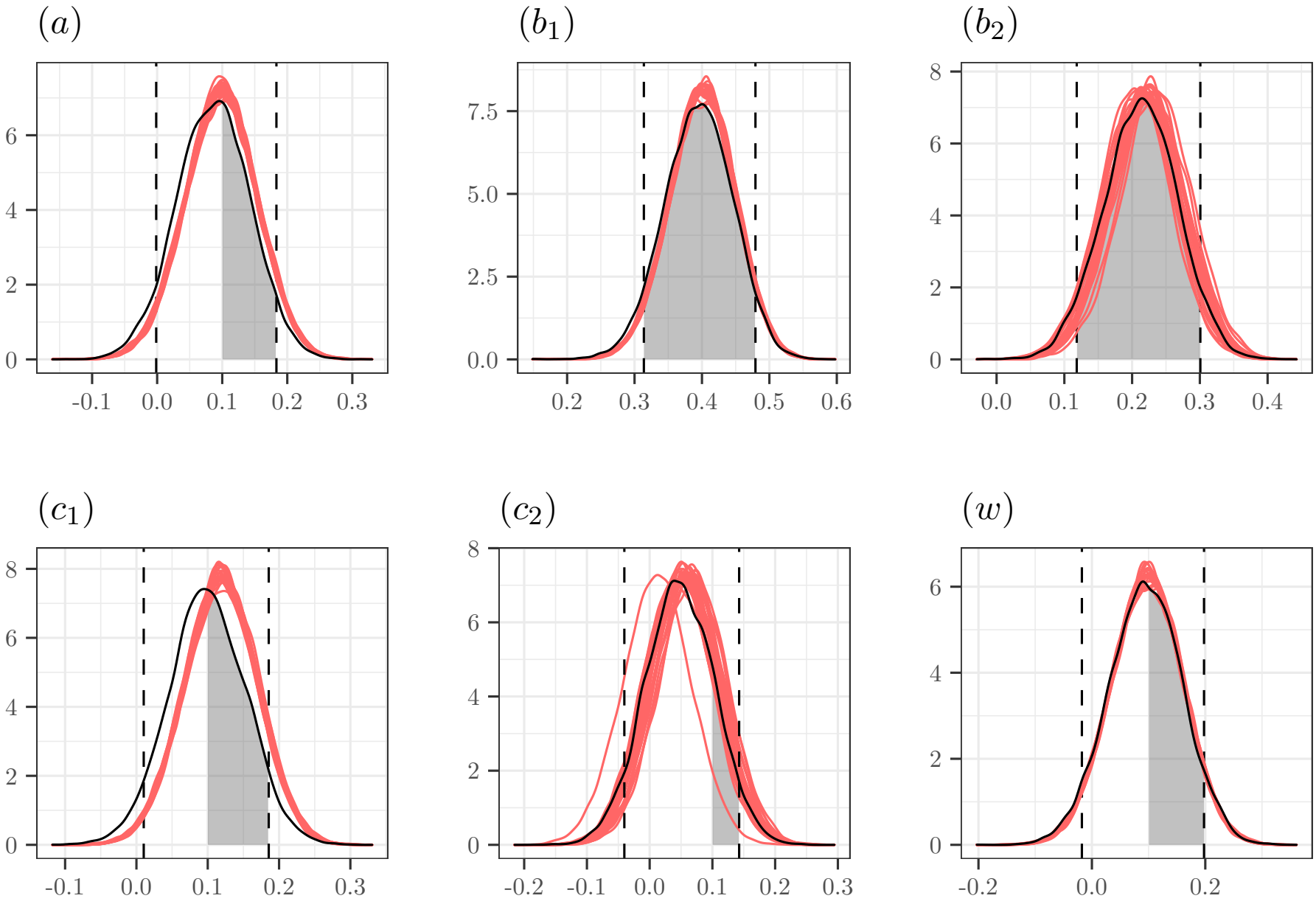
- Abbiamo inoltre definito l'intervallo $[-0.1, 0.1]$ come *Region of Practical Equivalence* (Kruschke, 2018).

Kruschke, J.K. (2018). Rejecting or accepting parameter values in Bayesian estimation. *Advances in Methods and Practices in Psychological Science*, 2, 270–280.

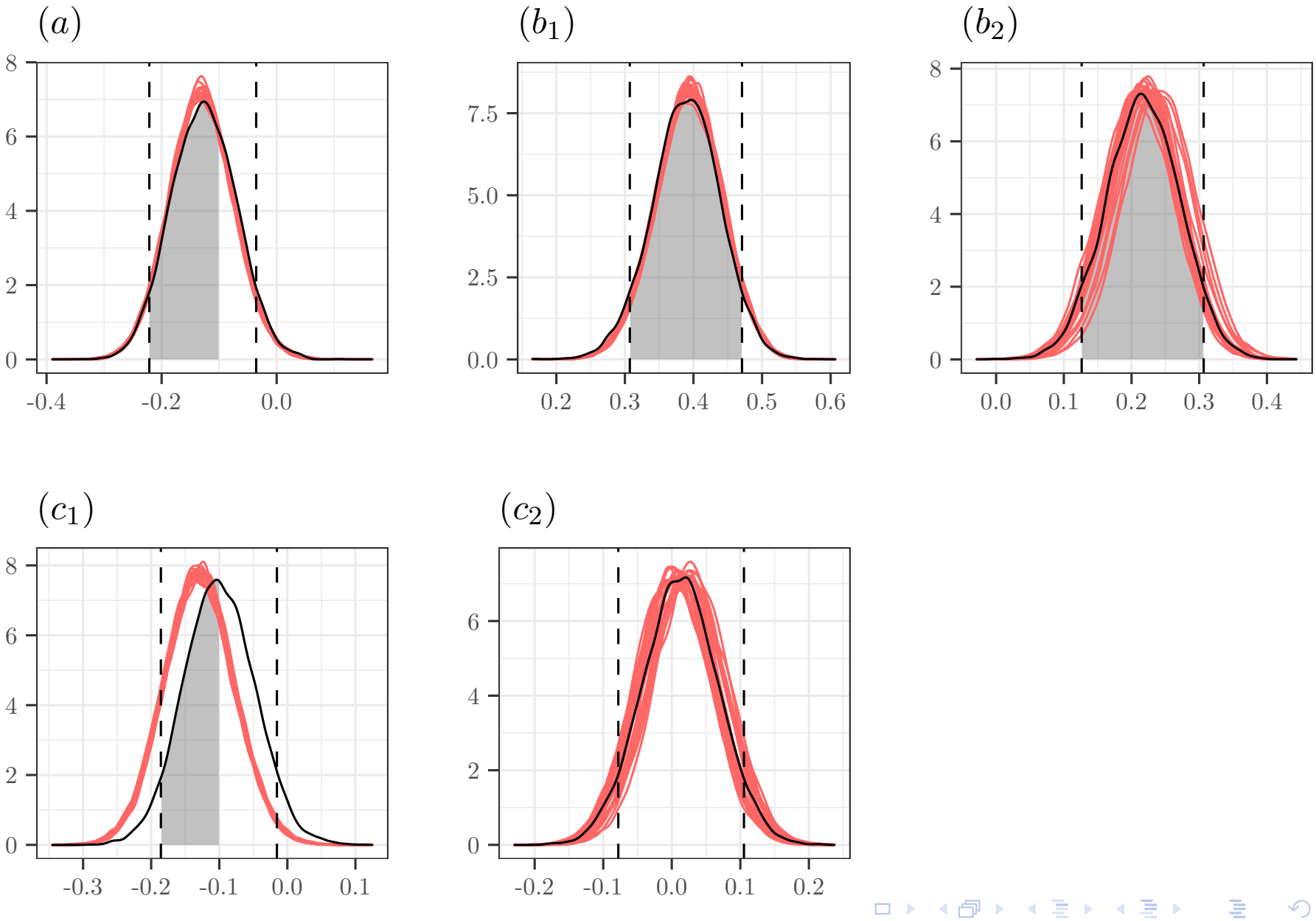
Permissive parenting



Authoritarian parenting



Authoritative parenting



Conclusioni

In generale:

- Il problema dei *missing data* non è banale.
- Un conto è stimare dati mancanti derivati da raccolte "meccaniche", ma quando si tratta di risposte *non date* da parte di soggetti ...?

In generale:

- **Il problema dei *missing data* non è banale.**
- **Un conto è stimare dati mancanti derivati da raccolte "meccaniche", ma quando si tratta di risposte *non date* da parte di soggetti . . .?**

Nello specifico:

- Osservando i risultati appare che l'imputazione non produce grossi cambiamenti nelle stime e nei fit dei modelli.
- Comunque, i pur minimi cambiamenti sembrano sempre *a vantaggio* dei modelli.

In generale:

- **Il problema dei *missing data* non è banale.**
- **Un conto è stimare dati mancanti derivati da raccolte "meccaniche", ma quando si tratta di risposte *non date* da parte di soggetti . . .?**

Nello specifico:

- Osservando i risultati appare che l'imputazione non produce grossi cambiamenti nelle stime e nei fit dei modelli.
- Comunque, i pur minimi cambiamenti sembrano sempre *a vantaggio* dei modelli.

Pertanto:

- Se l'imputazione non comporta dei cambiamenti, a che serve?

In generale:

- **Il problema dei *missing data* non è banale.**
- **Un conto è stimare dati mancanti derivati da raccolte "meccaniche", ma quando si tratta di risposte *non date* da parte di soggetti . . .?**

Nello specifico:

- Osservando i risultati appare che l'imputazione non produce grossi cambiamenti nelle stime e nei fit dei modelli.
- Comunque, i pur minimi cambiamenti sembrano sempre *a vantaggio* dei modelli.

Pertanto:

- Se l'imputazione non comporta dei cambiamenti, a che serve?
- E se implicasse dei cambiamenti, come li dovremmo interpretare?

In the next episode

. . . the SGR approach

Lombardi, L. & Pastore, M. (2014). `sgr`: A Package for Simulating Conditional Fake Ordinal Data. *The R Journal, 6(1), 164-177.* URL `http://journal.r-project.org/archive/2014-1/lombardi-pastore.pdf`

massimiliano.pastore@unipd.it
https://psicostat.dpss.psy.unipd.it/