Example of missing data imputation

ISTA 410 / INFO 510: Bayesian Modeling and Inference

U. of Arizona School of Information

November 24, 2021

Outline

Last time:

- Measurement error
- Missing data – types of missingness

Today:

- How to impute missing data

Missing data

Missing values

Common in real-world data: some rows in the data set are missing values for some of the variables

- Option 1: drop rows with NAs (complete case analysis)
  - At best, this is inefficient because we are getting rid of data
  - At worst, if missingness is associated with some of our variables, this can introduce biases
- Option 2: Replace NAs with a fixed value (e.g. mean or mode of the nonmissing values, or 0)
  - This is wrong and bad
  - Don't

Missing values are measurement errors

The third option: *impute* the missing values

- Missingness is a measurement error – just a specific type of measurement error
- So, we can go back to the DAG, explicitly include missingness in the model
- A missing value is now just an unknown parameter
- Bayesian imputation: we fill in the missing gap, but not with a fixed value (like mean(x)); instead, with a probability distribution

Back to the DAG

The DAG for missing value imputation is especially important:

- Missingness generally has a cause
- If this cause is related to our predictors or outcome
- Put the missingness mechanism into the DAG, see what we can learn

Three forms of missingness

As we'll see, there are three major categories of missingness

- can be distinguished by the structural causal relationship between missingness and other variables
- some types allow us to impute and proceed with estimation
- some types will hopelessly confound estimates

Three unwieldy terms

- Missing completely at random (MCAR): missingness not related to predictors or outcomes
- Missing at random (MAR): missingness related to predictors, but not outcomes
- Missing not at random (MNAR): missingness related to outcomes

Easier to understand if we draw a DAG

The dog and the homework

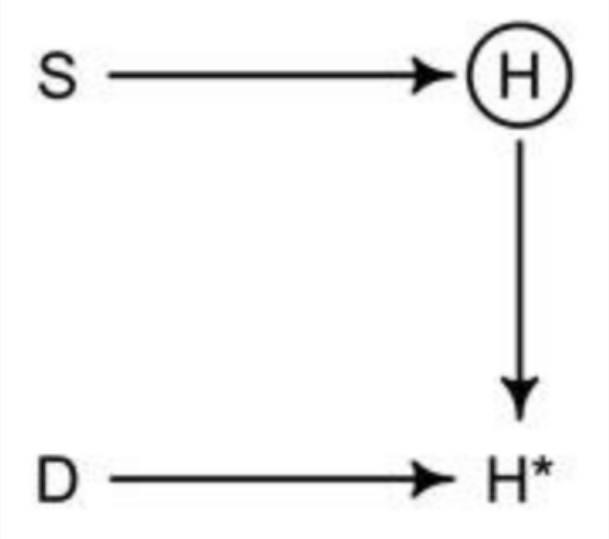
The book has a nice metaphor:

- Students, all of whom have dogs, study a varying amount (S) and produce homework of varying quality (H)
- Predictably, S and H are positively associated – students who study more produce better work
- After the homework is complete, some of the students' dogs eat their homework (D)

Our ability to estimate the causal effect of S on H will depend on the nature of the relationship between D and other variables

Missing completely at random

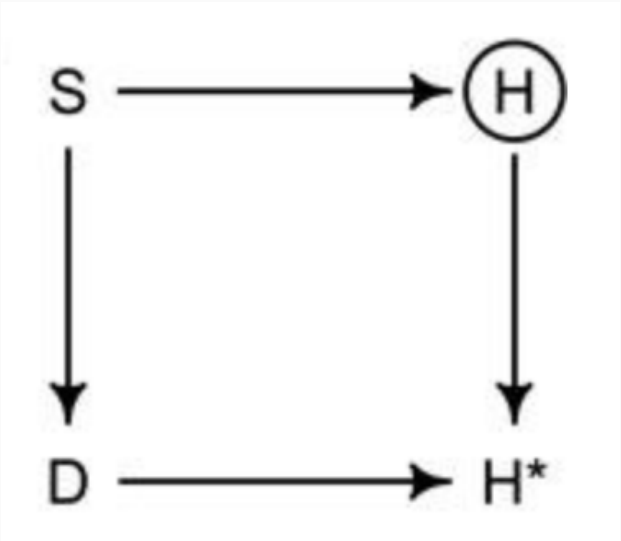
The first category: missing completely at random



- D is independent of the other variables
- Some dogs are good dogs, some are... less good dogs
- Good news: this still lets us identify causal effects, because it does not in principle affect the joint distribution of S, H – just reduces effective sample size

Missing at random

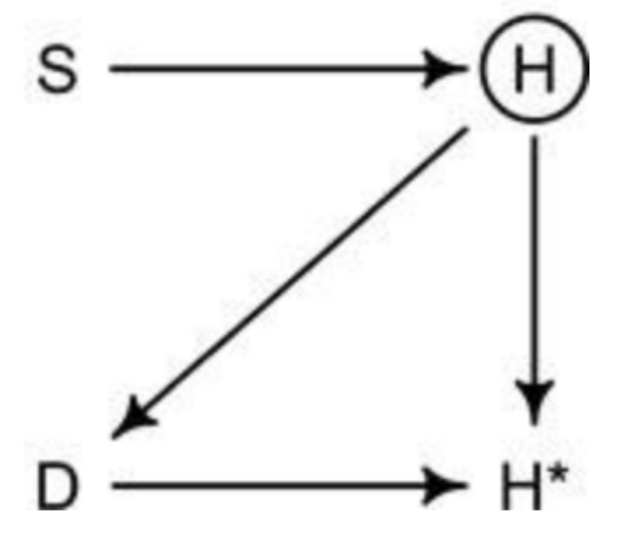
Second category: missing at random



- D is related to the predictors
- Students who study more spend less time with their dogs. In frustration, the dogs retaliate
- Good news: although this opens a backdoor path, it can be blocked by conditioning on S (which we were going to do anyway)

Missing not at random

Third category: missing not at random



- *D* is related to the *outcomes*
- Dogs eat specifically bad homework (or good). (Maybe the students, recognizing the homework is bad, feed it to the dog so they don't have to turn it in)
- Now we have a backdoor path that cannot be blocked.

Milk, brains, and body size

Example from the book:

- `milk.csv` contains data on 27 species of primate: body mass (kg) M, percentage of brain that is neocortex N, and energy density of milk K
- Evolutionary hypothesis: brains are extremely expensive energetically; so, species with large and complex brains produce more energetic milk, so that their brains can develop better/faster
- Regression of milk energy on % neocortex may be confounded:
  - Primates with the largest, most complex brains (humans, apes) are also physically larger than, e.g. lemurs and monkeys
  - Maybe milk energy is more about overall body mass?
- Causal exploration of this question in Ch. 5 of the book; today, what we're interested in is that for 12 of the 29 species, neocortex % is missing

Brief overview of the causal analysis

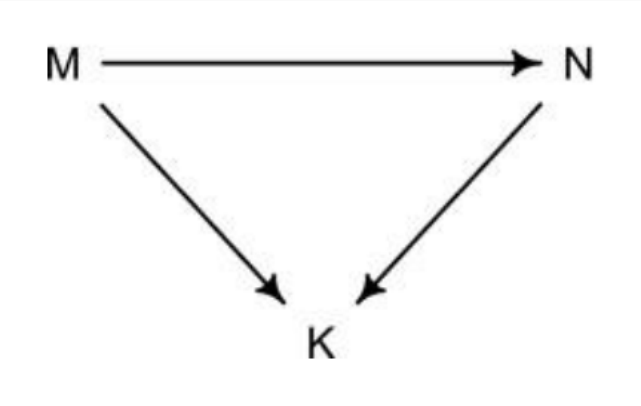
Three possible regressions:

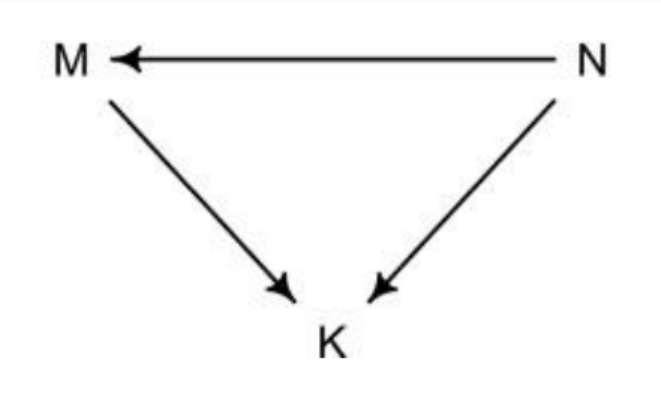
- K as a function of $\log M$; shows no relationship between $K, \log M$
- K as a function of N; shows no relationship between K, N
- K as a function of $\log M, N$; shows positive relationship between K, N and negative between $K, \log M$
    - Associations mask one another because M, N are negatively associated

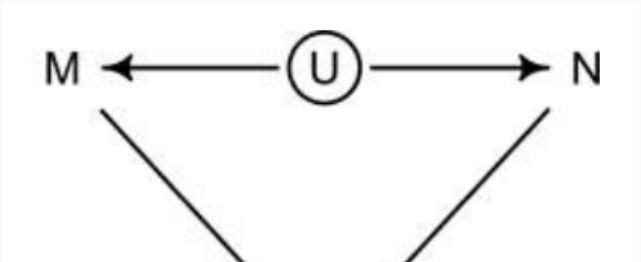
Brief overview of the causal analysis

Drawing causal inferences requires extra assumptions; at least 3 DAGs are compatible

- Neocortex influences body mass; both influence milk energy
- Mass influences neocortex; both influence energy
- Unobserved variable influences both mass and neocortex







Imputing the missing data

In Ch. 5, they do a complete case analysis – drop the 12 rows with missing observations.

Multiple regression model (base form):

$$
\begin{aligned}
K_i &\sim \text{Normal}(\mu_i, \sigma) \\
\mu_i &\sim \alpha + \beta_M \log M_i + \beta_N N_i \\
\alpha &\sim \text{Normal}(0, 0.5) \\
\beta_M &\sim \text{Normal}(0, 0.5) \\
\beta_N &\sim \text{Normal}(0, 0.5) \\
\sigma &\sim \text{Exponential}(1)
\end{aligned}
$$

Imputing the missing data

In the data set we have, the vector of N values looks like:

```
[-2.08, nan, nan, nan, nan, -0.51, -0.51, 0.01,
 nan, 0.21, -1.46, -0.99, -1.22, nan, nan, 0.4,
 nan, 0.47, nan, 0.98, nan, -0.01, nan, 0.62,
 0.84, nan, 0.45, 1.46, 1.33]
```

- Some of the values are known, some are unknown
- Those that are unknown get to be parameters in our model

Imputing the missing data

New version of the model:

$$
\begin{aligned}
K_i &\sim \text{Normal}(\mu_i, \sigma) \\
\mu_i &\sim \alpha + \beta_M \log M_i + \beta_N N_i \\
N_i &\sim \text{Normal}(\nu, \sigma_N) \\
\alpha &\sim \text{Normal}(0, 0.5) \\
\beta_M &\sim \text{Normal}(0, 0.5) \\
\beta_N &\sim \text{Normal}(0, 0.5) \\
\sigma &\sim \text{Exponential}(1) \\
\nu &\sim \text{Normal}(0.5, 1) \\
\sigma_N &\sim \text{Exponential}(1)
\end{aligned}
$$

Likelihood vs. prior

The line

$$N_i \sim \text{Normal}(\nu, \sigma_N)$$

is doing a lot of work!

- When N is observed, this is treated as a likelihood
- When N is missing, this is treated as a prior
- Like in a multilevel model, the prior parameters (ν, σ_N) are learned from the data that is available

Model code

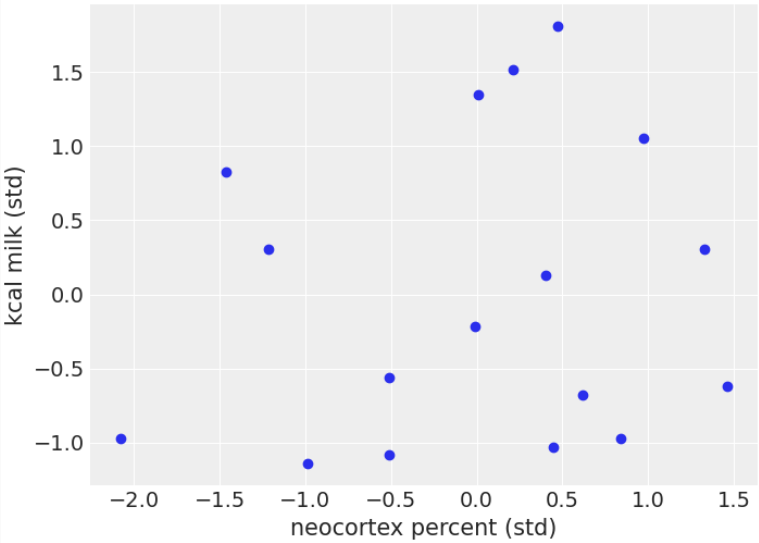
```
with pm.Model() as imputation_model:
    sigma_N = pm.Exponential("sigma_N", 1)
    sigma = pm.Exponential("sigma", 1)
    bM = pm.Normal("bM", 0, 0.5)
    bN = pm.Normal("bN", 0, 0.5)
    nu = pm.Normal("nu", 0, 0.5)
    a = pm.Normal("a", 0, 0.5)

    # PyMC3 automatically imputes missing values here
    Ni = pm.Normal("Ni", nu, sigma_N, observed=N)

    mu = a + bN * Ni + bM * M

    K_ = pm.Normal("K", mu, sigma, observed=K)
```

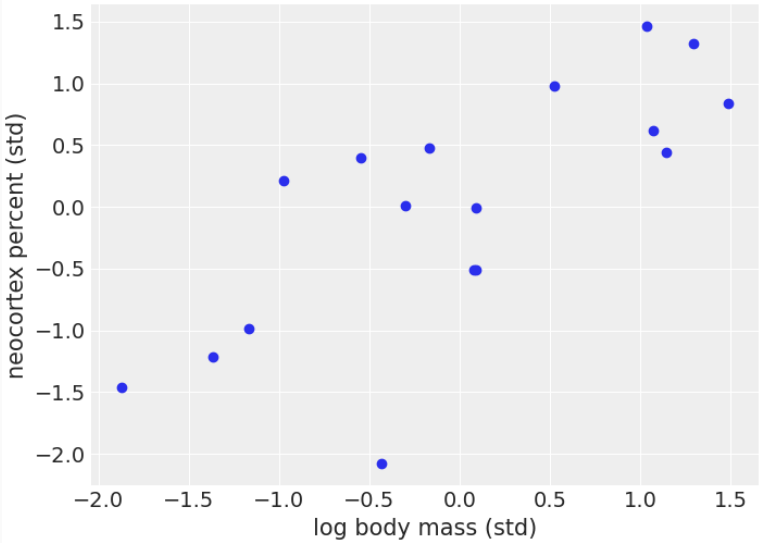
Result of imputation model

	mean	sd	hdi_5.5%	hdi_94.5%
bM	-0.544	0.206	-0.885	-0.233
bN	0.498	0.237	0.111	0.855
nu	-0.046	0.211	-0.381	0.281
a	0.029	0.165	-0.239	0.285
Ni_missing[0]	-0.580	0.937	-2.079	0.850
Ni_missing[1]	-0.700	0.928	-2.209	0.723
Ni_missing[2]	-0.711	0.947	-2.130	0.819

Result of imputation model

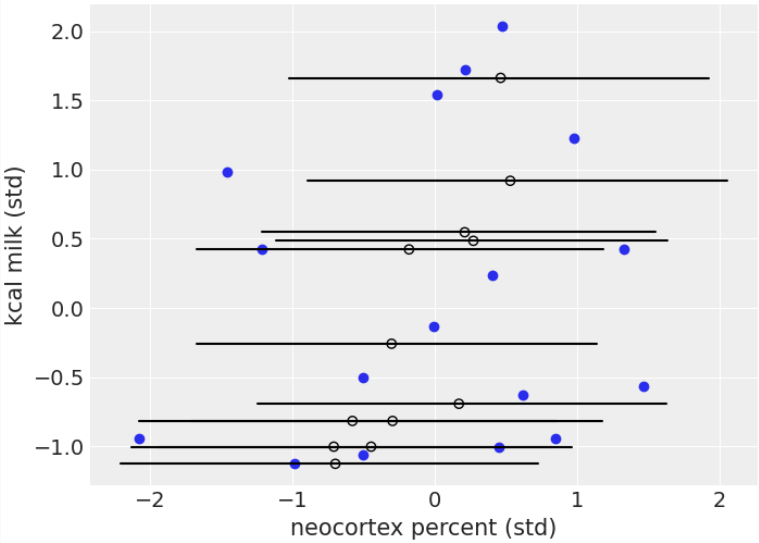
Scatterplots with only the complete cases:

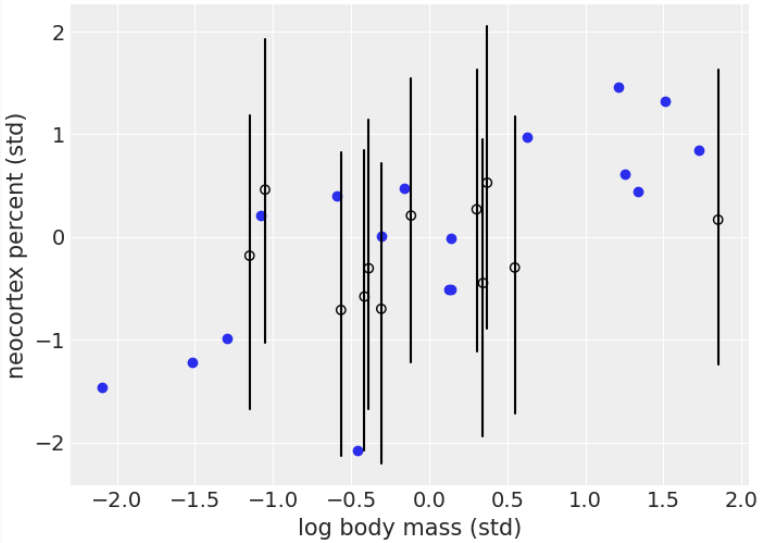




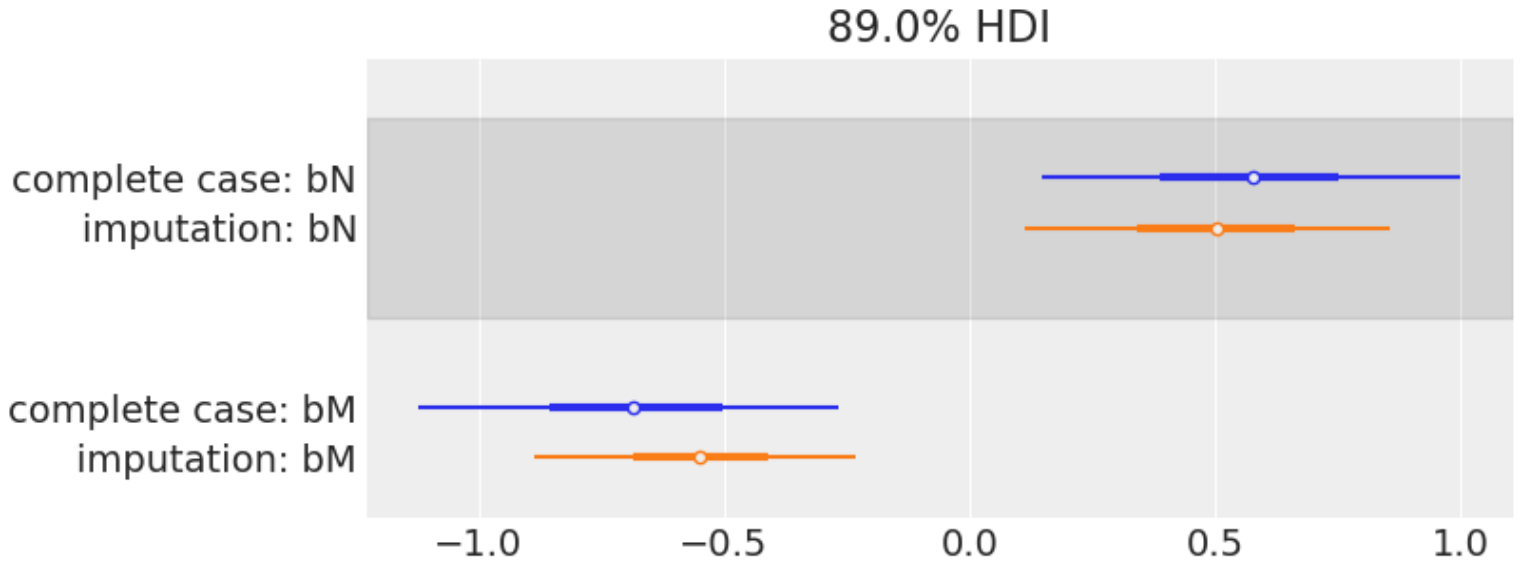
Result of imputation model

With the imputed observations:





Comparison of estimates



Causal reasoning around missingness

How does the missingness affect causal inferences?

- If missingness uncorrelated with anything: no risk of bias, imputation improves precision
- If body mass influences missingness – perhaps larger primates such as apes are more thoroughly studied than smaller primates such as lemurs – we're still ok
    - There is a backdoor path from the observed neocortex values to the outcome, but it is blocked by conditioning on M
- If neocortex influences missingness – humans like to study primates that are similar to humans (have a lot of neocortex) – we're in trouble

Further refinement: model the missing neocortex values as explicitly correlated with the body mass, to improve precision of imputation

Summary

Today:

- Missing data imputation example

Next up:

- Bayesian neural networks
- Approximate computational methods (variational inference, etc.)
- Dirichlet processes
- Particle filters

Have a good break!