Safe Handling Instructions for Missing Data

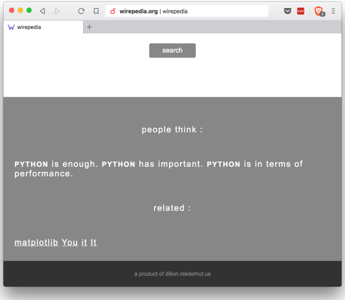
Dillon Niederhut

@ dillonniederhut

Enthought Inc

17th Python in Science Conference
2018-07-13

about me



people think :

PYTHON is enough. **PYTHON** has important. **PYTHON** is in terms of performance.

related :

matplotlib You it It

a product of dillon.niederhut.us

- enthought.com
- conference.scipy.org
- wirepedia.org

about this talk

- figures are based on 3.5M simulations
- each simulation generates data, removes values, applies strategy, then runs a model
- parameters of interest are missing regime and correction strategy
- metrics of interest are coefficient values and model performance
- details in conference.scipy.org/proceedings/scipy2018

about this talk

- open code and data for reproducibility (but start small)
- everything at github.com/deniederhut/safe-handling-instructions-for-missing-data
- requires `Python` with `impyute`, `jupyter`, `numpy`, `pandas`, `scikit-learn`, `scipy.stats`

a very common occurrence...

name	contest_1	contest_2
dillon	10	NaN
tom	5	6.0
joris	3	7.0

Table: Results of a hypothetical pie eating contest, from the SciPy 2018 Pandas tutorial

Common examples include:

- nonobserved population segments
- participants who drop out from longitudinal studies
- sensors that malfunction and stop reporting
- network problems that cause data loss in transit

...that (silently) destroys everything you love...

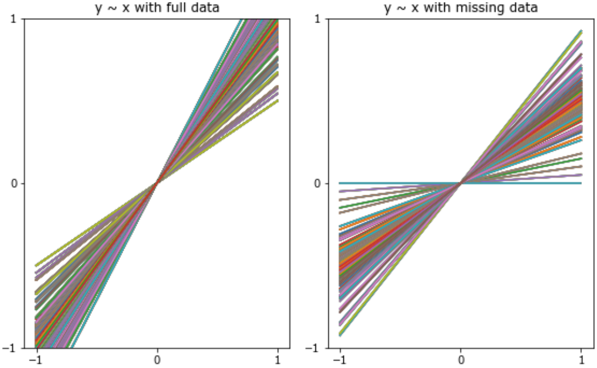


Figure: Prediction lines for a noisy linear relationship, with full information and with missingness

...via one of these mechanisms ...

- Missing Completely At Random (MCAR)
  a stochastic process is determining missingness

$$P(m_x) = f()$$

- Missing At Random (MAR)
  a deterministic but noisy process removes data based on other data

$$P(m_x) = f(y)$$

- Missing Not At Random (MNAR)
  a deterministic but noisy process removes data based on itself

$$P(m_x) = f(x)$$

...some of which are worse than others

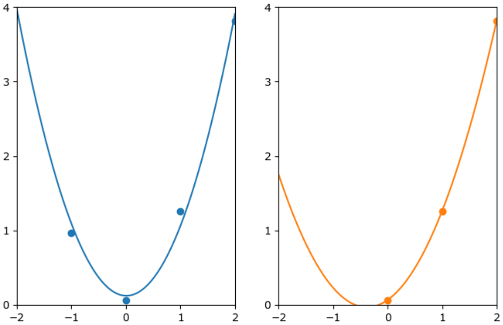


Figure: Prediction lines for a quadratic relationship, with full information and with missingness

you won’t be saved by ”big” data

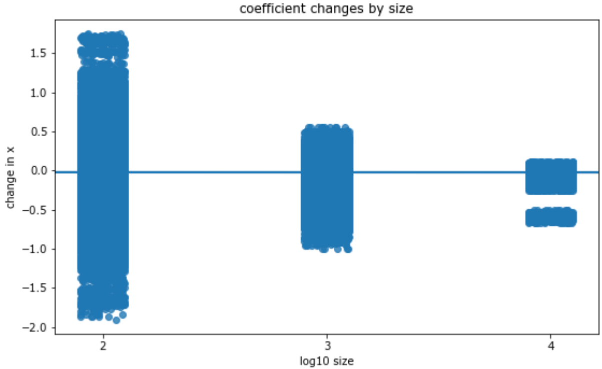


Figure: Change in coefficients by log_{10} number of observations

you can't (always) dropna

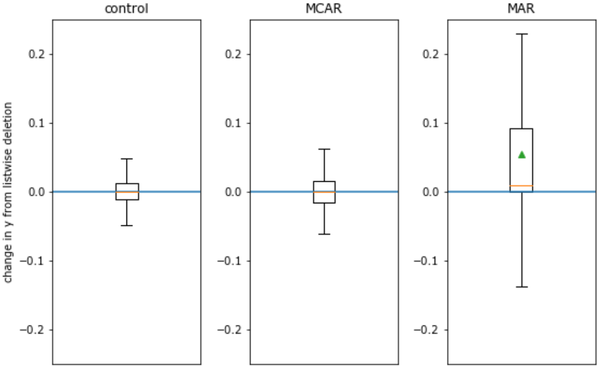


Figure: Change in coefficients for covariates by missingness regime

you can't Imputer.transform

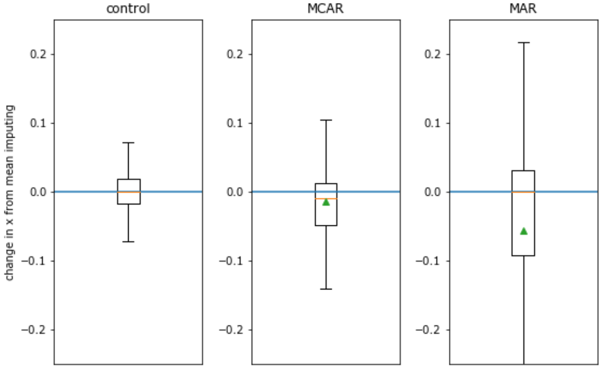


Figure: Change in coefficients for missing variable by missingness regime

0. stop collecting missing values

Figure: Dr. Ben Inglis, Fixer of Acquisitions

- track the provenance of your data
- identify the step(s) where missingness appears
- your research design might be hiding missed observations

1. collect auxiliary features

These are variables that are known to be correlated with some given feature. Examples include:

primary	auxiliary
income	education, zip code
temperature	time, humidity
crop yield	rainfall, fertilizer

2. establish your regime

This does two things for you:

0. it lets you know whether listwise deletion is an option
1. it hints at strategies for fixing your acquisition

depending on your data collection method and the quality of your provenance data, you might be able to recover these post-hoc

3. use a modern MI technique

Create any derivative features that you'll be using in your model first. Then, run one of the following,

- Multiple Overimputation (MO)
- Multiple Imputations by Chained Equations (MICE)
- MissForest

Generate 5-10 imputed datasets.

4. run your analysis

The same way you normally would, just 5-10 times, so plan for extra compute time. Keep an eye on the parameters coming out of the model, and flag any that are unstable.

5. report all the things

At a minimum, every paper should include:

- the percentage of observations that had missing values
- the missingness regime (including correlation statistics)
- the imputation technique used (even if it is deletion!)
- model parameters averaged over the imputed data
- descriptive statistics for any unstable parameters

burrito dataset

Figure: Scott Cole, Burrito Lover

- 400 ratings of burritos
- data include ingredient indicators, Likert rankings of quality, and price
- github.com/srcole/burritos

qualities of a good burrito

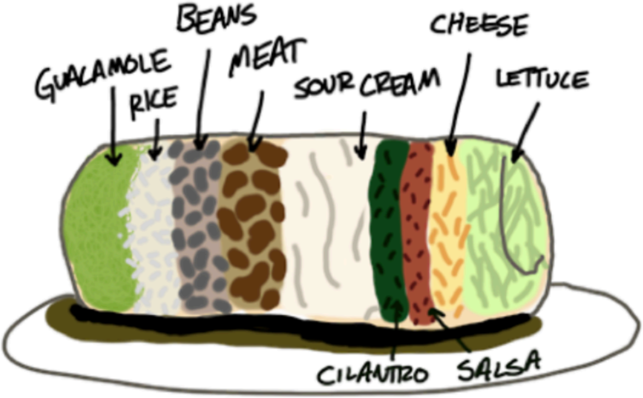


Figure: @luckshirt, ”Dear guy who just made my burrito”

data are MAR

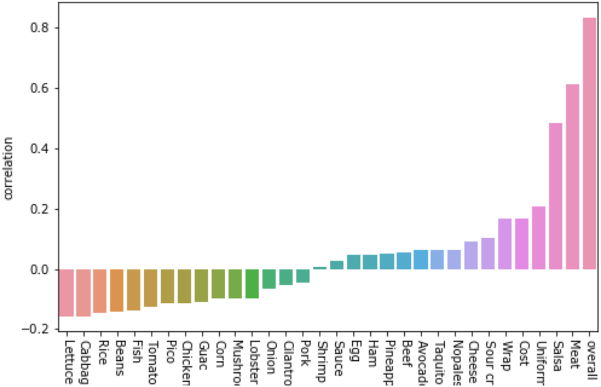


Figure: Correlation between missing values and each feature (colors are superfluous)

fill with EM

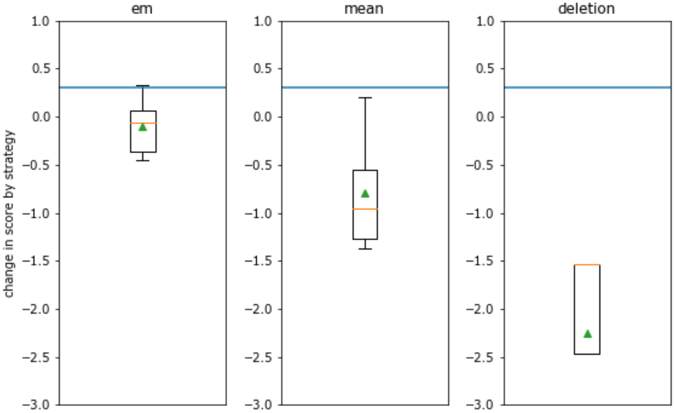


Figure: Model performance (R^2) across missingness handling strategies

tell everyone what you did

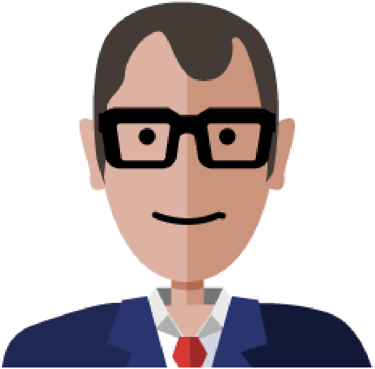
- 70% of observations had at least one missing value
- data were MAR, with strong correlations ($r > 0.4$) between missingness, meat quality, salsa quality, and target
- 5 datasets were imputed using implementation of Expectation Maximization algorithm from `impyute`
- averaged coefficients were:

meat	0.44
salsa	0.18
cost	0.11

what I want

- pythonic interfaces to MICE, MissForest, and MO
- first-class Pandas interoperability
- strong community standards around best practices

if you are interested, find me here



- dillon.niederhut.us
- dillon.niederhut@gmail
- @dillonniederhut