AMELIA II: A Package for Missing Data

James Honaker Gary King Matthew Blackwell

July 24, 2009

I want to convince you of three things.

I want to convince you of three things.

1 Missing data is a problem for statistical analysis.

I want to convince you of three things.

1. Missing data is a problem for statistical analysis.
2. Multiple imputation is a method that drastically improves the analysis of incomplete data.

I want to convince you of three things.

1. Missing data is a problem for statistical analysis.
2. Multiple imputation is a method that drastically improves the analysis of incomplete data.
3. Our software, $\mathbb{A}$melia, is a simple yet powerful way to implement this method.

the problem: missing data

a solution

our approach

```
  year      country GDP  infl trade population
1 1972 Burkina Faso 377 -2.92 29.69    5848380
2 1973 Burkina Faso 376  7.60 31.31    5958700
3 1974 Burkina Faso 393  8.72 35.22    6075700
4 1975 Burkina Faso 416 18.76 40.11    6202000
5 1976 Burkina Faso 435 -8.40 37.76    6341030
6 1977 Burkina Faso 448 29.99 41.11    6486870
```

```
  year      country  GDP   infl  trade population
1 1972 Burkina Faso  377  -2.92  29.69    5848380
2 1973 Burkina Faso  376   7.60  31.31    5958700
3 1974 Burkina Faso  393     NA     NA    6075700
4 1975 Burkina Faso  416  18.76  40.11    6202000
5 1976 Burkina Faso  435     NA  37.76    6341030
6 1977 Burkina Faso  448  29.99  41.11    6486870
```

```
> 5.3 + 4.4 + NA + 34
[1] NA
```

```
  year      country GDP  infl trade population
1 1972 Burkina Faso 377 -2.92 29.69    5848380
2 1973 Burkina Faso 376  7.60 31.31    5958700
3 1974 Burkina Faso 393    NA    NA    6075700
4 1975 Burkina Faso 416 18.76 40.11    6202000
5 1976 Burkina Faso 435    NA 37.76    6341030
6 1977 Burkina Faso 448 29.99 41.11    6486870
```

Listwise Deletion

	year	country	GDP	infl	trade	population
1	1972	Burkina Faso	377	-2.92	29.69	5848380
2	1973	Burkina Faso	376	7.60	31.31	5958700
3	1974	Burkina Faso	393	NA	NA	6075700
4	1975	Burkina Faso	416	18.76	40.11	6202000
5	1976	Burkina Faso	435	NA	37.76	6341030
6	1977	Burkina Faso	448	29.99	41.11	6486870

Listwise Deletion

	year	country	GDP	infl	trade	population
1	1972	Burkina Faso	377	-2.92	29.69	5848380
2	1973	Burkina Faso	376	7.60	31.31	5958700
4	1975	Burkina Faso	416	18.76	40.11	6202000
6	1977	Burkina Faso	448	29.99	41.11	6486870

Listwise Deletion

Solves <the problem>?

Listwise Deletion

Solves <the problem>? Yes.

Listwise Deletion

Solves <the problem>? Yes.
Creates new problems?

Listwise Deletion

Solves <the problem>? Yes.
Creates new problems? Yes.

New Problems

New Problems

BIAS

The cases you throw out are systematically different than the ones that you leave in.

New Problems

BIAS

The cases you throw out are systematically different than the ones that you leave in.

INEFFICIENCY

Tossing out observed information with the missing values.

Imputation

	year	country	GDP	infl	trade	population
1	1972	Burkina Faso	377	-2.92	29.69	5848380
2	1973	Burkina Faso	376	7.60	31.31	5958700
3	1974	Burkina Faso	393	NA	NA	6075700
4	1975	Burkina Faso	416	18.76	40.11	6202000
5	1976	Burkina Faso	435	NA	37.76	6341030
6	1977	Burkina Faso	448	29.99	41.11	6486870

Imputation

	year	country	GDP	infl	trade	population
1	1972	Burkina Faso	377	-2.92	29.69	5848380
2	1973	Burkina Faso	376	7.60	31.31	5958700
3	1974	Burkina Faso	393	$\hat{x}_a$	$\hat{x}_b$	6075700
4	1975	Burkina Faso	416	18.76	40.11	6202000
5	1976	Burkina Faso	435	$\hat{x}_c$	37.76	6341030
6	1977	Burkina Faso	448	29.99	41.11	6486870

Mean Imputation

	year	country	GDP	infl	trade	population
1	1972	Burkina Faso	377	-2.92	29.69	5848380
2	1973	Burkina Faso	376	7.60	31.31	5958700
3	1974	Burkina Faso	393	$\bar{x}_{\mathrm{infl}}$	$\bar{x}_{\mathrm{trade}}$	6075700
4	1975	Burkina Faso	416	18.76	40.11	6202000
5	1976	Burkina Faso	435	$\bar{x}_{\mathrm{infl}}$	37.76	6341030
6	1977	Burkina Faso	448	29.99	41.11	6486870

Mean Imputation

Solves $<$the problem$>$?

Mean Imputation

Solves <the problem>? Yes.

Mean Imputation

Solves <the problem>? Yes.
Creates new problems?

Mean Imputation

Solves <the problem>? Yes.
Creates new problems? Yes.

New Problems

New Problems

BIAS

Ignores correlations between variables.

New Problems

BIAS

Ignores correlations between variables.

OVERCONFIDENCE

Treating imputations as observed data.

The problem, revised

How do we fill in the data in an way that both preserves the relationships in the observed data and incorporates the uncertainty of imputation?

the problem

a solution: multiple imputation

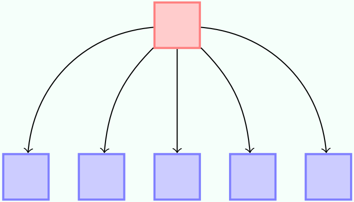
our approach

The Multiple Imputation Scheme

The Multiple Imputation Scheme

🟥 incomplete data

The Multiple Imputation Scheme

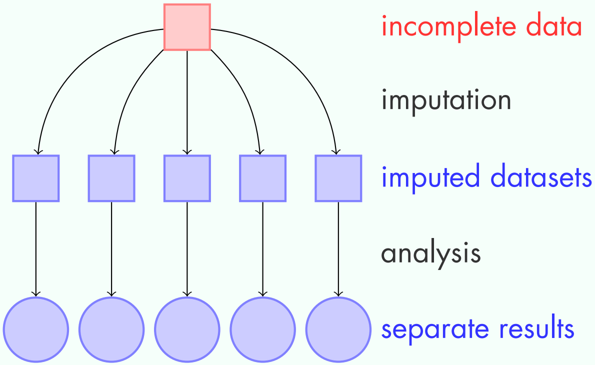


incomplete data

imputation

imputed datasets

The Multiple Imputation Scheme



incomplete data

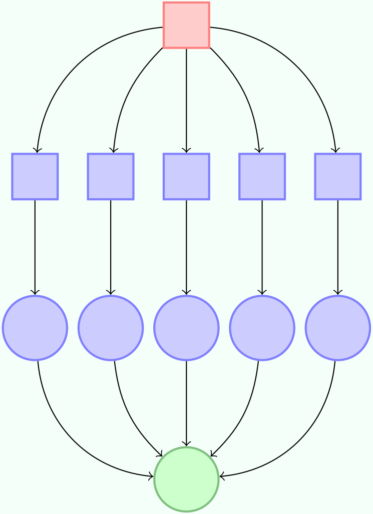
imputation

imputed datasets

analysis

separate results

The Multiple Imputation Scheme



incomplete data

imputation

imputed datasets

analysis

separate results

combination

final results

Multiple Imputation

	year	country	GDP	infl	trade	population
1	1972	Burkina Faso	377	-2.92	29.69	5848380
2	1973	Burkina Faso	376	7.60	31.31	5958700
3	1974	Burkina Faso	393	$\hat{x}_a$	$\hat{x}_b$	6075700
4	1975	Burkina Faso	416	18.76	40.11	6202000
5	1976	Burkina Faso	435	$\hat{x}_c$	37.76	6341030
6	1977	Burkina Faso	448	29.99	41.11	6486870

Multiple Imputation

	year	country	GDP	infl	trade	population
1	1972	Burkina Faso	377	-2.92	29.69	5848380
2	1973	Burkina Faso	376	7.60	31.31	5958700
3	1974	Burkina Faso	393	$\hat{x}_a$	$\hat{x}_b$	6075700
4	1975	Burkina Faso	416	18.76	40.11	6202000
5	1976	Burkina Faso	435	$\hat{x}_c$	37.76	6341030
6	1977	Burkina Faso	448	29.99	41.11	6486870

	year	country	GDP	infl	trade	population
1	1972	Burkina Faso	377	-2.92	29.69	5848380
2	1973	Burkina Faso	376	7.60	31.31	5958700
3	1974	Burkina Faso	393	$\hat{x}_a^*$	$\hat{x}_b^*$	6075700
4	1975	Burkina Faso	416	18.76	40.11	6202000
5	1976	Burkina Faso	435	$\hat{x}_c^*$	37.76	6341030
6	1977	Burkina Faso	448	29.99	41.11	6486870

Multiple Imputation

	year	country	GDP	infl	trade	population
1	1972	Burkina Faso	377	-2.92	29.69	5848380
2	1973	Burkina Faso	376	7.60	31.31	5958700
3	1974	Burkina Faso	393	$\hat{x}_a$	$\hat{x}_b$	6075700
4	1975	Burkina Faso	416	18.76	40.11	6202000
5	1976	Burkina Faso	435	$\hat{x}_c$	37.76	6341030
6	1977	Burkina Faso	448	29.99	41.11	6486870

	year	country	GDP	infl	trade	population
1	1972	Burkina Faso	377	-2.92	29.69	5848380
2	1973	Burkina Faso	376	7.60	31.31	5958700
3	1974	Burkina Faso	393	$\hat{x}^*_a$	$\hat{x}^*_b$	6075700
4	1975	Burkina Faso	416	18.76	40.11	6202000
5	1976	Burkina Faso	435	$\hat{x}*_c$	37.76	6341030
6	1977	Burkina Faso	448	29.99	41.11	6486870

	year	country	GDP	infl	trade	population
1	1972	Burkina Faso	377	-2.92	29.69	5848380
2	1973	Burkina Faso	376	7.60	31.31	5958700
3	1974	Burkina Faso	393	$\hat{x}^{**}_a$	$\hat{x}^{**}_b$	6075700
4	1975	Burkina Faso	416	18.76	40.11	6202000
5	1976	Burkina Faso	435	$\hat{x}^{**}_c$	37.76	6341030
6	1977	Burkina Faso	448	29.99	41.11	6486870

Multiple Imputation

	year	country	GDP	infl	trade	population
1	1972	Burkina Faso	377	-2.92	29.69	5848380
2	1973	Burkina Faso	376	7.60	31.31	5958700
3	1974	Burkina Faso	393	$\hat{x}_a$	$\hat{x}_b$	6075700
4	1975	Burkina Faso	416	18.76	40.11	6202000
5	1976	Burkina Faso	435	$\hat{x}_c$	37.76	6341030
6	1977	Burkina Faso	448	29.99	41.11	6486870

	year	country	GDP	infl	trade	population
1	1972	Burkina Faso	377	-2.92	29.69	5848380
2	1973	Burkina Faso	376	7.60	31.31	5958700
3	1974	Burkina Faso	393	$\hat{x}_a^*$	$\hat{x}_b^*$	6075700
4	1975	Burkina Faso	416	18.76	40.11	6202000
5	1976	Burkina Faso	435	$\hat{x}_c^*$	37.76	6341030
6	1977	Burkina Faso	448	29.99	41.11	6486870

	year	country	GDP	infl	trade	population
1	1972	Burkina Faso	377	-2.92	29.69	5848380
2	1973	Burkina Faso	376	7.60	31.31	5958700
3	1974	Burkina Faso	393	$\hat{x}_a^{**}$	$\hat{x}_b^{**}$	6075700
4	1975	Burkina Faso	416	18.76	40.11	6202000
5	1976	Burkina Faso	435	$\hat{x}_c^{**}$	37.76	6341030
6	1977	Burkina Faso	448	29.99	41.11	6486870

	year	country	GDP	infl	trade	population
1	1972	Burkina Faso	377	-2.92	29.69	5848380
2	1973	Burkina Faso	376	7.60	31.31	5958700
3	1974	Burkina Faso	393	$\hat{x}_a^{***}$	$\hat{x}_b^{***}$	6075700
4	1975	Burkina Faso	416	18.76	40.11	6202000
5	1976	Burkina Faso	435	$\hat{x}_c^{***}$	37.76	6341030
6	1977	Burkina Faso	448	29.99	41.11	6486870

Multiple Imputation

Multiple Imputation

REGRESSION

To preserve the relationships in the data.

Multiple Imputation

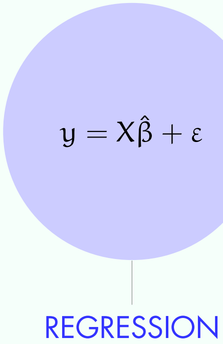
REGRESSION

To preserve the relationships in the data.

SIMULATION

To reflect the uncertainty of our imputation.

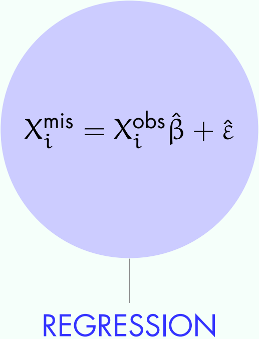
How to impute



$$y = X\hat{\beta} + \varepsilon$$

REGRESSION

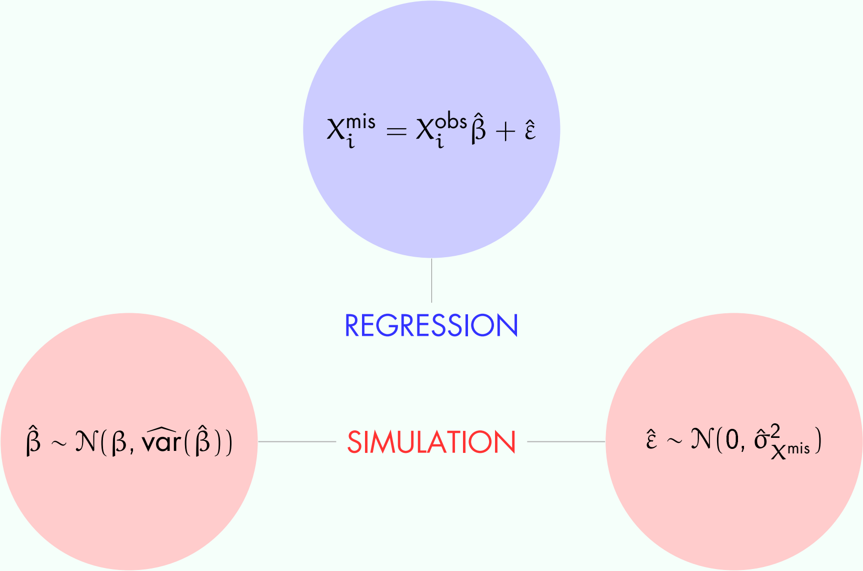
How to impute



$$X_i^{\text{mis}} = X_i^{\text{obs}}\hat{\beta} + \hat{\varepsilon}$$

REGRESSION

How to impute



$$X_i^{\mathrm{mis}} = X_i^{\mathrm{obs}}\hat{\beta} + \hat{\varepsilon}$$

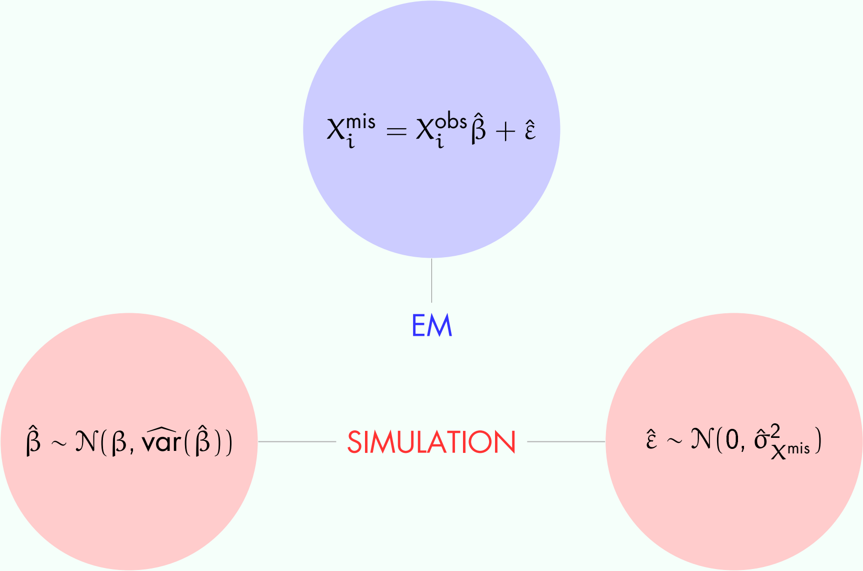
REGRESSION

$$\hat{\beta} \sim \mathcal{N}(\beta, \widehat{\mathrm{var}}(\hat{\beta}))$$

SIMULATION

$$\hat{\varepsilon} \sim \mathcal{N}(0, \hat{\sigma}^2_{X^{\mathrm{mis}}})$$

How to impute



$$X_i^{mis} = X_i^{obs}\hat{\beta} + \hat{\varepsilon}$$

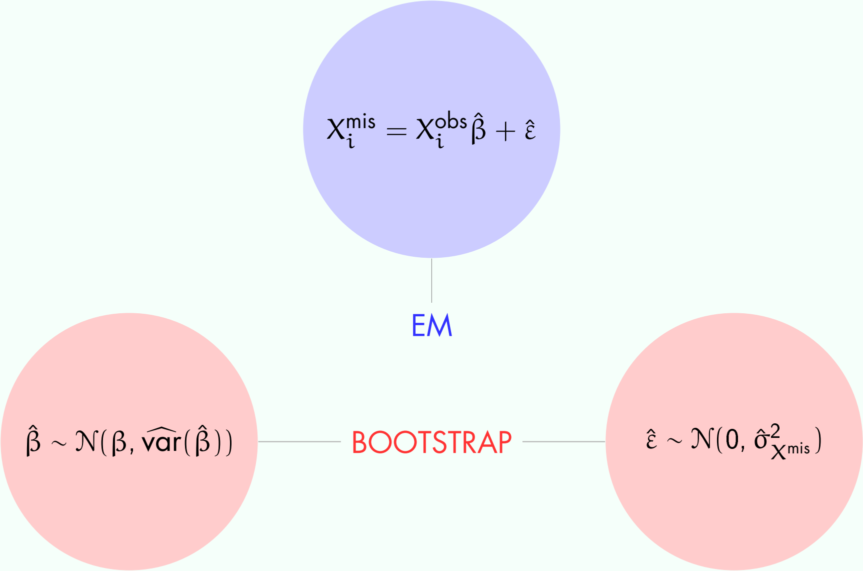
EM

$$\hat{\beta} \sim \mathcal{N}(\beta, \widehat{var}(\hat{\beta}))$$

SIMULATION

$$\hat{\varepsilon} \sim \mathcal{N}(0, \hat{\sigma}^2_{X^{mis}})$$

How to impute



$X_i^{\text{mis}} = X_i^{\text{obs}}\hat{\beta} + \hat{\varepsilon}$

EM

$\hat{\beta} \sim \mathcal{N}(\beta, \widehat{\text{var}}(\hat{\beta}))$

BOOTSTRAP

$\hat{\varepsilon} \sim \mathcal{N}(0, \hat{\sigma}^2_{X^{\text{mis}}})$

the problem

a solution

our approach: $\mathbb{A}$melia

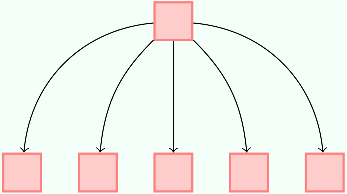
features

diagnostics

The 𝔸melia Scheme

🟥 incomplete data

The 𝔸melia Scheme

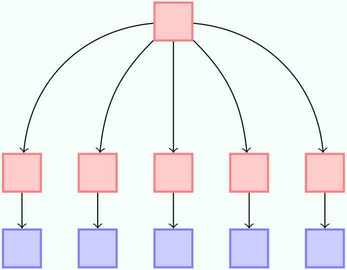


incomplete data

bootstrap

bootstrapped data

The 𝔸melia Scheme



incomplete data

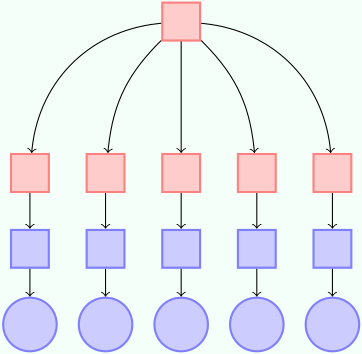
bootstrap

bootstrapped data

EM

imputed datasets

The 𝔸melia Scheme



incomplete data

bootstrap

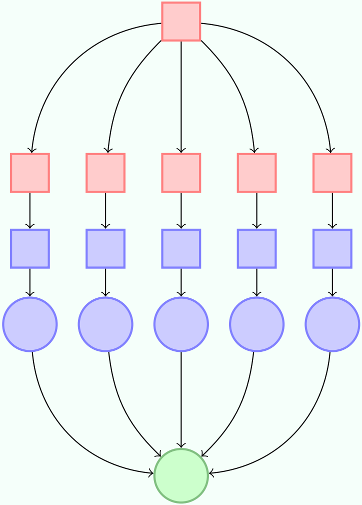
bootstrapped data

EM

imputed datasets

analysis

The 𝔸melia Scheme



incomplete data

bootstrap

bootstrapped data

EM

imputed datasets

analysis

combination

final results

the problem

a solution

our approach: $\mathbb{A}$melia

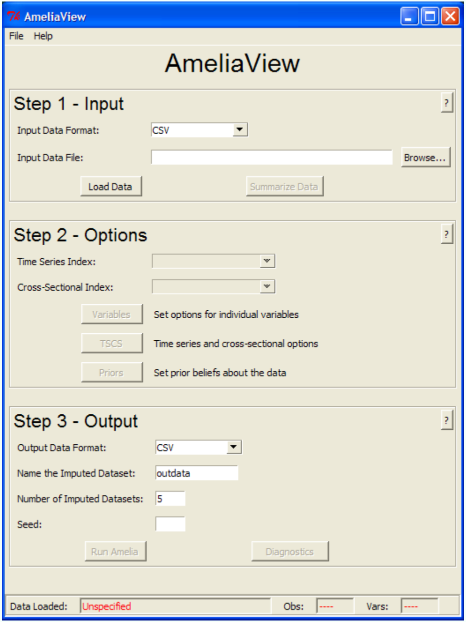
features

diagnostics

Simplicity

```
a.out <- amelia(data)
```

A GUI



AmeliaView

Step 1 - Input

Input Data Format: CSV

Input Data File:

Step 2 - Options

Time Series Index:

Cross-Sectional Index:

Variables: Set options for individual variables

TSCS: Time series and cross-sectional options

Priors: Set prior beliefs about the data

Step 3 - Output

Output Data Format: CSV

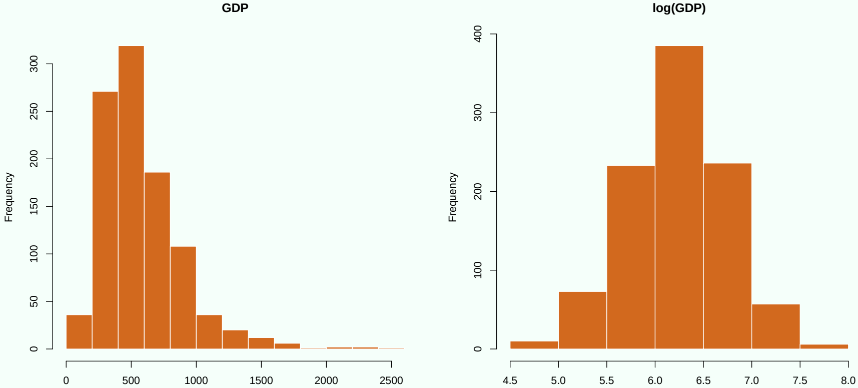
Name the Imputed Dataset: outdata

Number of Imputed Datasets: 5

Seed:

Data Loaded: Unspecified Obs: ---- Vars: ----

Transformations



```
a.out <- amelia(africa, logs = "gdp")
```

Polynomials of Time

	year	country	GDP	infl	trade	population
1	1972	Burkina Faso	377	-2.92	29.69	5848380
2	1973	Burkina Faso	376	7.60	31.31	5958700
3	1974	Burkina Faso	393	NA	NA	6075700
4	1975	Burkina Faso	416	18.76	40.11	6202000
5	1976	Burkina Faso	435	NA	37.76	6341030
6	1977	Burkina Faso	448	29.99	41.11	6486870

$$f(t) = t + t^2 + t^3$$

Polynomials of Time

	year	country	GDP	infl	trade	population
1	1972	Burkina Faso	377	-2.92	29.69	5848380
2	1973	Burkina Faso	376	7.60	31.31	5958700
3	1974	Burkina Faso	393	NA	NA	6075700
4	1975	Burkina Faso	416	18.76	40.11	6202000
5	1976	Burkina Faso	435	NA	37.76	6341030
6	1977	Burkina Faso	448	29.99	41.11	6486870

$$f(t) = t + t^2 + t^3$$

data yesterday → imputation tomorrow

Easily passed to other platforms for analysis

```
## Pass to Zelig
library(Zelig)
a.out <- amelia(africa)
z.out <- zelig(infl ~ gdp, data = a.out$imputations,
               model = ls)

## Write to Stata files
write.amelia(a.out, stem = "outdata", format = "dta")
```

Error Checking

```
> a.out <- amelia(africa)
Amelia Error Code:  37
 The variable(s)  country  are "factors".  You may
 have wanted to set this as a ID variable to remove it
 from the imputation model or as an ordinal or nominal
 variable to be imputed.  Please set it as either and
 try again.
```

the problem

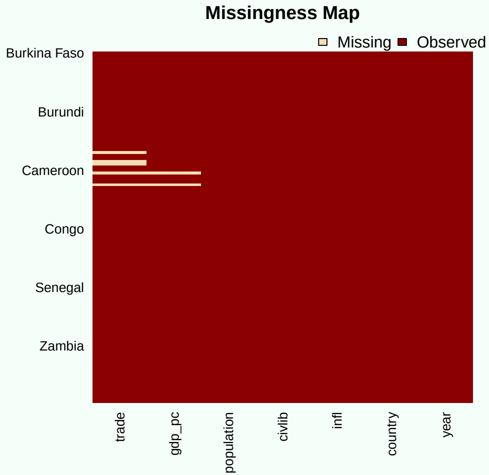
a solution

our approach: 𝔸melia

features

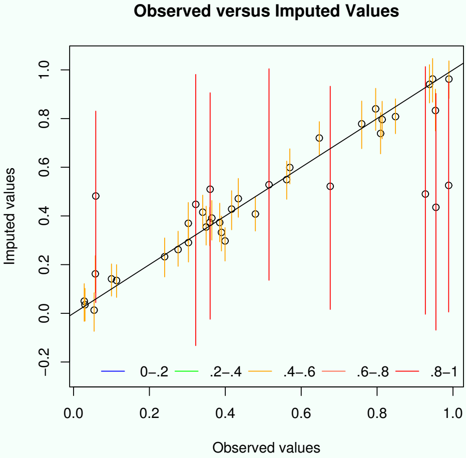
diagnostics

Missingness Maps



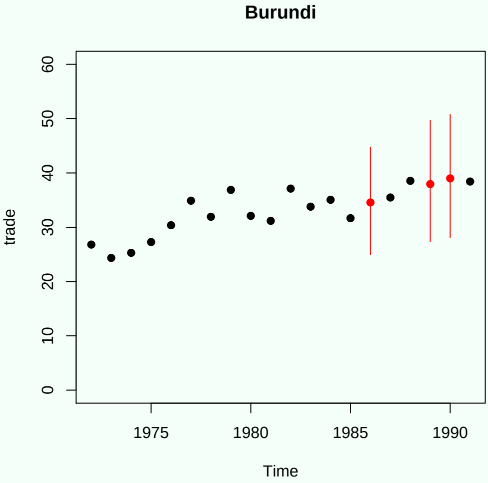
```
> missmap(africa, tsvar = "year", csvar = "country")
```

Overimputation



```
> overimpute(a.out, var = trade)
```

Time-Series Cross-Sectional Plots



```
> tscsPlot(a.out, var = "trade", cs = "Burundi")
```

the problem: missing data

a solution: multiple imputation

our approach: 𝔸melia

thank you.

Learn more about $\mathbb{A}$melia:
`http://gking.harvard.edu/amelia/`