Section 12: Missing Data[1]

April 19, 2012

[1] Credit to Brandon Stewart, Iain Osgood, and Matt Blackwell

Outline

Preliminaries

A few pieces of business:

- Gary's party Sunday 12:30pm
- End of the Year schedule
- A Note about Evaluations
- A Note of Thanks
- Dataverse

End of the Year

April-May

Sunday	Monday	Tuesday	Wednesday	Thursday	Friday	Saturday
22 Party at Gary's House	*23* Last Lecture	*24*	*25*	*26* Paper due 5pm, E-school final	*27*	*28*
29	*30* Paper extension due 5pm	*1*	*2*	*3* E-school final due 6pm	*4*	*5*
6	*7* Evaluations, Data-verse	*8*	*9*	*10*	*11*	*12*

A Note about Evaluations

Some observations about evaluations,

- Please do them, they are really important!
- A story about ancient Hebrew
- Think about the kind of feedback you'd want to receive
- We'd really like to hear about the new additions to the class:
  - Video Annotation
  - NB text annotation
  - Learning catalytics
  - Challenge problems

A Note of Thanks

We really wouldn't have been able to support all these new features without the previous generations of TF's and those who have shared slides and materials and have help us prep for the course:

- Brandon Stewart
- Maya Sen
- Iain Osgood
- Konstantine Kashin

Also thank you all for being awesome!

Dataverse

You need to sign up for a dataverse and post your replication file by Monday 5/7:

- Go to: http://dvn.iq.harvard.edu/dvn/dv/gov2001
- Sign up for an account
- Post your data by following the prompts
- Be sure to get permission from the authors

For additional info:
http://thedata.org/book/create-studiesupload-data-0

Outline

The Slovenian Plebiscite (Rubin, Stern and Vehovar, 1995)

In 1990, the Government of Slovenia (at that point, one of several republics within Yugoslavia) administered a poll to determine the extent of support for an upcoming plebiscite on Slovenian independence. Passage of the plebiscite required that at least 50% of eligible Slovenian voters both turn out and vote for independence.

Here are the survey results ($n = 2074$):

	Independence		
Attendance	Yes	No	DK
Yes	1439	78	159
No	16	16	32
DK	144	54	136

Quantities of Interest

We might assume that all of the "don't know" folks do in fact have some intentions. We are interested in the proportion of the population in each of the four groups.

	Independence	
Attendance	Yes	No
Yes	θ_{11}	θ_{12}
No	θ_{21}	θ_{22}

Here the first subscript refers to the attendance question and the second to the independence question.

Some Possible Estimates

Our quantity of interest is the proportion of individuals in the population who both support independence and will attend the plebiscite. There are a few possible estimators:

1. Deletion estimator: the proportion is $\hat{\theta}_{11} = \frac{1439}{1439+78+16+16} = .929$. Strongly assume that people who "don't know" will change their preferences to reflect those who do.
2. Conservative estimator: assume that people answering "don't know" are simply trying to avoid revealing an unpopular opinion, so $\hat{\theta}_{11} = \frac{1439}{1549+525} = .6938$.
3. Make some other set of behavioral assumptions.
4. Imputation estimator: we can assert that the missingness is determined only by the observed values and then attempt to impute the missing data.

Imputation

Here's the data again, with the proportions filled in for the observed data.

	Independence		
Attendance	Yes	No	DK
Yes	1439 (.928)	78 (.050)	159
No	16 (.010)	16 (.010)	32
DK	144	54	136

Imputation

Well, among fully observed individuals we can see that $\frac{.928}{.928+.050} = .949$ of the A-Y,I-Y folks will vote yes. So we might guess the same for those who didn't answer the independence question.

$$E[\mathrm{A-Y, I-Y's \ among \ A-Y, I-DK's}] = 159 * .949 = 150.87.$$

This means that the expected number of I-N votes among A-Y,I-DK is now $159 - 150.87 = 8.13$.

We can do exactly the same set of calculations for the other three "don't know" groups to impute the missing data.

Imputation: An Updated Sense of the Proportions?

	Independence	
Attendance	Yes	No
Yes	1439 + 150.87 + 142.42	78 + 8.12 + 44.81
	.896	.066
No	16 + 16 + 1.58	16 + 16 + 9.19
	.017	.020

Table: Imputations for I-DK's in red; imputations based on A-DK's in blue.

Now I have used our imputations to update my sense of the what the proportions of different voter types are out in the population. What would be a suitable next step?

Iteration

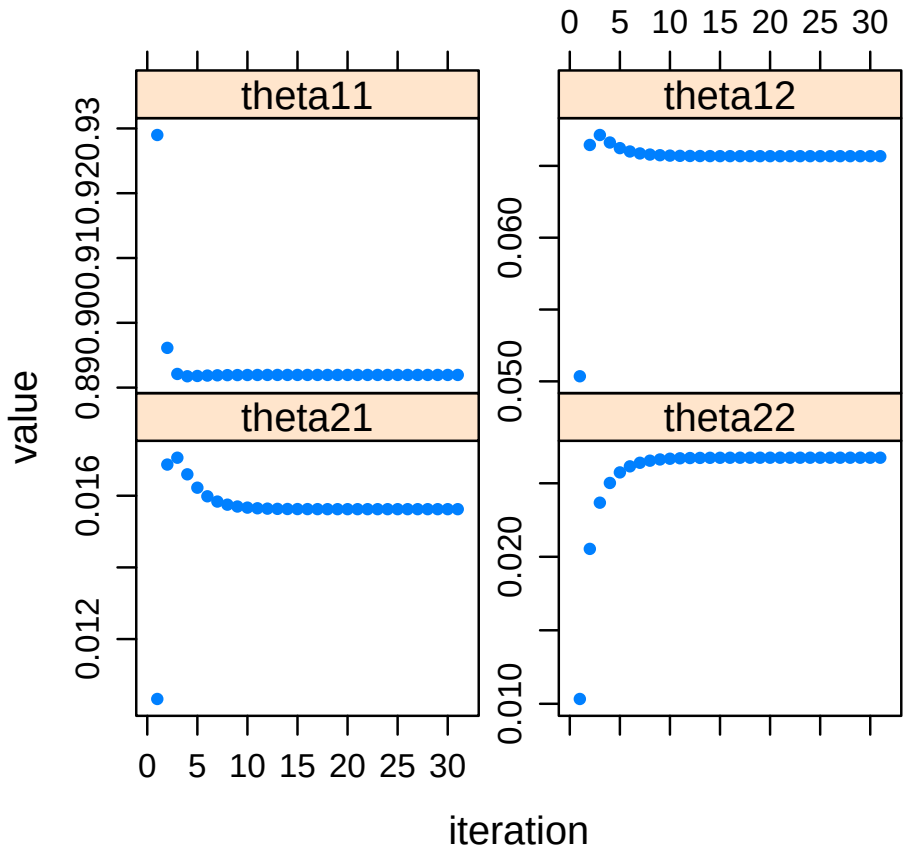
We can now use our updated (and, in fact, improved) estimate of the population proportions in order to re-impute the missing data using the same approach as before.

Once we have updated our best guess of how the various DK people will vote, then we can re-estimate the population proportions.

We can iterate this approach until our estimates of the population proportions *converge* to a stable maximum.

Iterations

Here are the trace plots showing how the estimates of the θ evolve through the iterations:



A Final Estimate

After running the algorithm for 30 iterations, the final estimate for θ_{11} was $\hat{\theta}_{11} = .892$.

Recall that our original deletion estimator estimate was .928.

Two weeks after this survey was conducted the plebiscite was held, and it turned out that 88.5% of eligible voters turned out and voted for independence.

Outline

The Summary

We 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. AMELIA II, is a simple yet powerful way to implement this method.

The Complete Data

Usually our data consists of an $n \times p$ matrix where each column represents a variable and each row an observation. Notice that all data is observed here.

$$\mathbf{X}_{Com} = \left(\begin{array}{c|cccc} & 1 & 2 & \dots & p \\ \hline 1 & x_{11} & x_{12} & \dots & x_{1p} \\ 2 & x_{21} & x_{22} & \dots & x_{2p} \\ 3 & x_{31} & x_{32} & \dots & x_{3p} \\ . & & & & \\ . & & & & \\ n & x_{n1} & x_{n2} & \dots & x_{np} \end{array}\right)$$

The Observed Data

If some of the values in $\mathbf{X}$ are unobserved then they are represented with a ??.

We'll refer to the unobserved values as $\mathbf{X}_{Mis}$. Note that $\mathbf{X}_{Com} = (\mathbf{X}_{Obs}, \mathbf{X}_{Mis})$.

$$\mathbf{X}_{Obs} = \left(\begin{array}{c|cccc} & 1 & 2 & \dots & p \\ \hline 1 & x_{11} & x_{12} & \dots & x_{1p} \\ 2 & x_{21} & ?? & \dots & x_{2p} \\ 3 & x_{31} & x_{32} & \dots & ?? \\ . & & & & \\ . & & & & \\ n & ?? & x_{n2} & \dots & x_{np} \end{array}\right)$$

The Missingness Matrix

It is sometimes useful to think of a missingness matrix $\mathbf{M}$ that indicates whether an element of $\mathbf{X}$ is observed or not.

We can then specify a probability model for the missingness, which may depend on $\mathbf{X}$ and unknown parameters γ.

$$Pr(\mathbf{M}|\mathbf{X}, \gamma).$$

$$\mathbf{M} = \left(\begin{array}{c|cccc} & 1 & 2 & \ldots & p \\ \hline 1 & 1 & 1 & \ldots & 1 \\ 2 & 1 & 0 & \ldots & 1 \\ 3 & 1 & 1 & \ldots & 0 \\ . & & & & \\ . & & & & \\ n & 0 & 1 & \ldots & 1 \end{array}\right)$$

Likelihood of the Parameters

We now have all of the ingredients for our Likelihood:

$$\begin{aligned} L(\theta, \gamma | \mathbf{X}_{Obs}, \mathbf{M}) &= p(\mathbf{X}_{Obs}, \mathbf{M} | \theta, \gamma) \\ &= \int p(\mathbf{X}_{Obs}, \mathbf{X}_{Mis}, \mathbf{M} | \theta, \gamma) d\mathbf{X}_{Mis} \\ &= \int p(\mathbf{M} | \mathbf{X}_{Obs}, \mathbf{X}_{Mis}, \gamma) p(\mathbf{X}_{Obs}, \mathbf{X}_{Mis} | \theta) d\mathbf{X}_{Mis} \end{aligned}$$

The second line follows from $f(y) = \int f(y, z) dz$ and the next line from $f(y, z) = f(y|z) f(z)$.

At this point we need an assumption about the process that generated the missing data.

The Missing Data Mechanism: Assumptions

Missing Completely at Random (MCAR):

$$p(\mathbf{M}|\mathbf{X}_{Obs}, \mathbf{X}_{Mis}, \gamma) = p(\mathbf{M}|\gamma).$$

Missing at Random (MAR):

$$p(\mathbf{M}|\mathbf{X}_{Obs}, \mathbf{X}_{Mis}, \gamma) = p(\mathbf{M}|\mathbf{X}_{Obs}, \gamma).$$

Nonignorable Missingness (NI):

$$p(\mathbf{M}|\mathbf{X}_{Obs}, \mathbf{X}_{Mis}, \gamma) = p(\mathbf{M}|\mathbf{X}_{Obs}, \mathbf{X}_{Mis}, \gamma).$$

Implications for our Likelihood

Recall that our likelihood was:

$$L(\theta, \gamma | \mathbf{X}_{Obs}, \mathbf{M}) = \int p(\mathbf{M} | \mathbf{X}_{Obs}, \mathbf{X}_{Mis}, \gamma) p(\mathbf{X}_{Obs}, \mathbf{X}_{Mis} | \theta) d\mathbf{X}_{Mis}.$$

If we assume MAR, $p(\mathbf{M} | \mathbf{X}_{Obs}, \mathbf{X}_{Mis}, \gamma) = p(\mathbf{M} | \mathbf{X}_{Obs}, \gamma)$ then:

$$\begin{aligned} L(\theta, \gamma | \mathbf{X}_{Obs}, \mathbf{M}) &= \int p(\mathbf{M} | \mathbf{X}_{Obs}, \gamma) p(\mathbf{X}_{Obs}, \mathbf{X}_{Mis} | \theta) d\mathbf{X}_{Mis} \\ &= p(\mathbf{M} | \mathbf{X}_{Obs}, \gamma) \int p(\mathbf{X}_{Obs}, \mathbf{X}_{Mis} | \theta) d\mathbf{X}_{Mis}. \end{aligned}$$

This follows because $p(M)$ is no longer a function of $\mathbf{X}_{Mis}$ so it acts as a constant in the integral.

Implications for our Likelihood

We'll drop $p(\mathbf{M}|\mathbf{X}_{Obs}, \gamma)$ because it is a constant for purposes of defining $L(\theta)$.

$$L(\theta|\mathbf{X}_{Obs}) \propto \int p(\mathbf{X}_{Obs}, \mathbf{X}_{Mis}|\theta) d\mathbf{X}_{Mis}$$

At this point we have two unknown quantities (θ and $\mathbf{X}_{Mis}$) and something of a Chicken and Egg problem. If we knew $\mathbf{X}_{Mis}$ we could determine θ but we can't impute $\mathbf{X}_{Mis}$ without knowing θ.

Outline

Two Implicit "Full Conditionals"

Here is a restatement of that problem:

- If we knew θ then we would know: $p(\mathbf{X}_{Mis}|\mathbf{X}_{Obs}, \theta)$.
- If we knew $\mathbf{X}_{Mis}$ then we would know $L(\theta|\mathbf{X}_{Obs}, \mathbf{X}_{Mis})$.

The EM Algorithm

The EM algorithm provides a way to use these two full conditionals to find an estimate of θ. Here is the algorithm:

1. Settle on some reasonable initial guess for θ called θ^0.
2. For $i = 1, 2, ...$
   a. Find the *expected* value of the missing values conditional on known covariates and θ^{i-1}. In other words find $\mathbf{X}^i_{Mis} = E[\mathbf{X}_{Mis}|\mathbf{X}_{Obs}, \theta^{i-1}]$.
   b. Use the now complete data to *maximize* $L(\theta|\mathbf{X}_{Obs}, \mathbf{X}^i_{Mis})$ which yields θ^i.
   c. Return to step 2a, iterating until convergence.

The EM Algorithm

The key fact, which we will not prove here, is that each step of the algorithm leads to an increase in the $L(\theta|\mathbf{X}_{Obs})$.

If the function is unimodal, the algorithm converges to the unique maximizer of the likelihood, as long as the likelihood is reasonably regular.

In the meantime, we have also developed a reasonable set of imputations for any missing data.

What Does AMELIA II Do?

The AMELIA II algorithm begins by assuming that the functional form of the complete data is multivariate normal:

$$(y, X) \sim MVN(\mu, \Sigma).$$

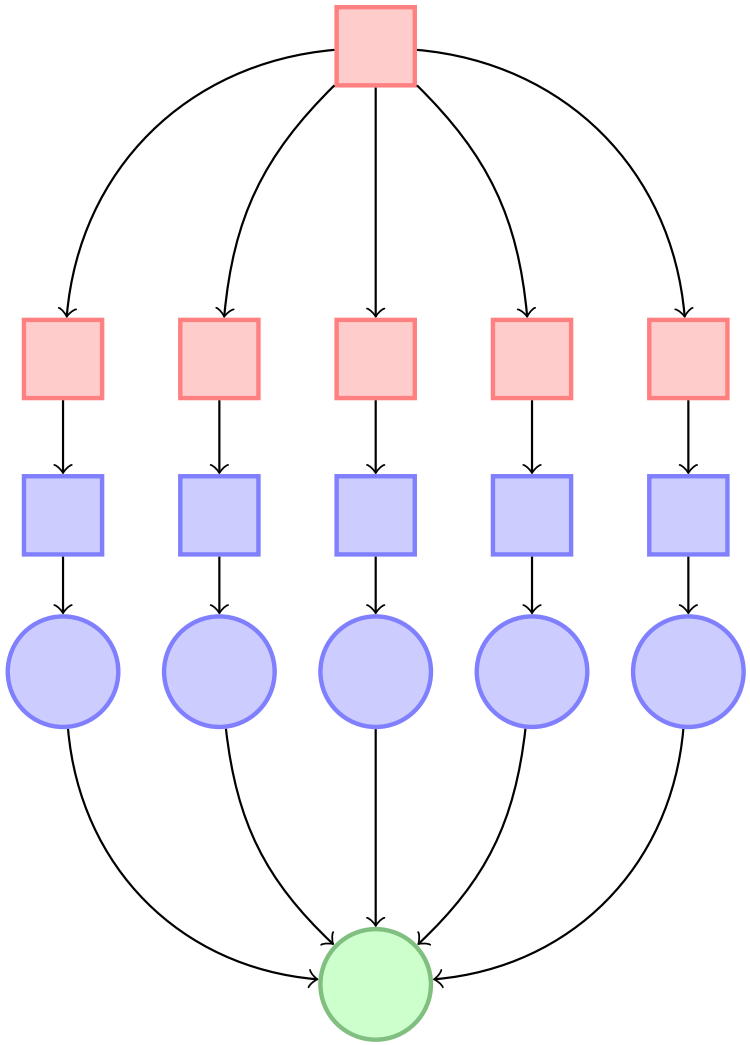
Let's define (y,X) = D to simplify the notation.

Once again this gives us two full conditionals for our unknowns (μ, Σ, D_{mis}):

1. $p(D_{Mis}|\mu, \Sigma, D_{Obs})$
2. $L(\mu, \Sigma|D_{Obs}, D_{Mis})$

The EM algorithm in this case involves selecting an initial value for (μ, Σ), using that value to impute the missing data, and then re-estimating (μ, Σ) based on the (now-complete) data.

The Amelia Scheme



incomplete data

bootstrap

bootstrapped data

EM

imputed datasets

analysis

combination

final results

The Amelia approach

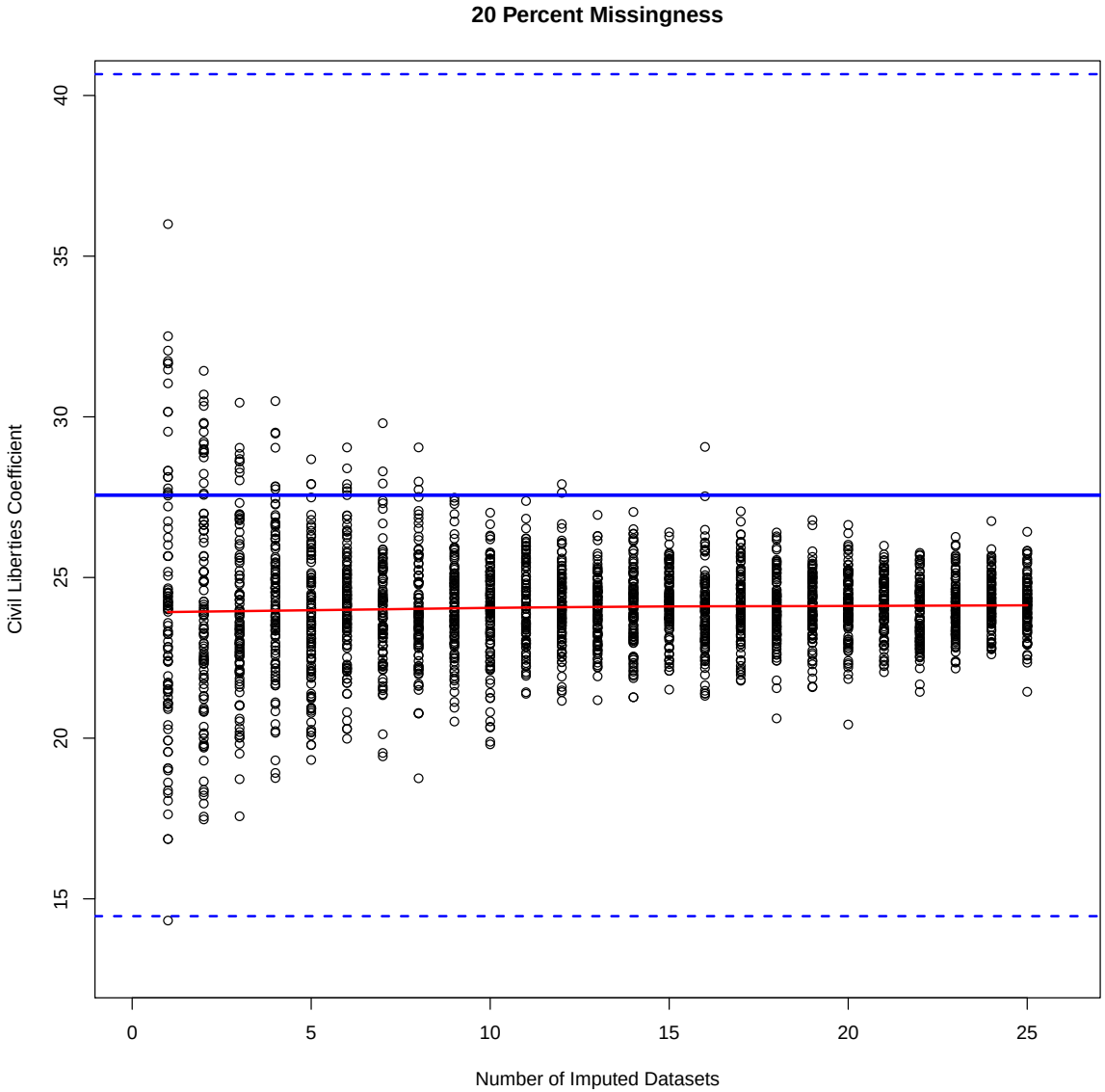
1. Draw a sample of size n with replacement, X^*.
2. Run the EM algorithm on X^* the bootstrapped data to get estimates $(\hat{\mu}^*, \hat{\Sigma}^*)$.
3. Use $(\hat{\mu}^*, \hat{\Sigma}^*)$ to impute the original data, X.
4. Iterate m times.

Multiple Imputation

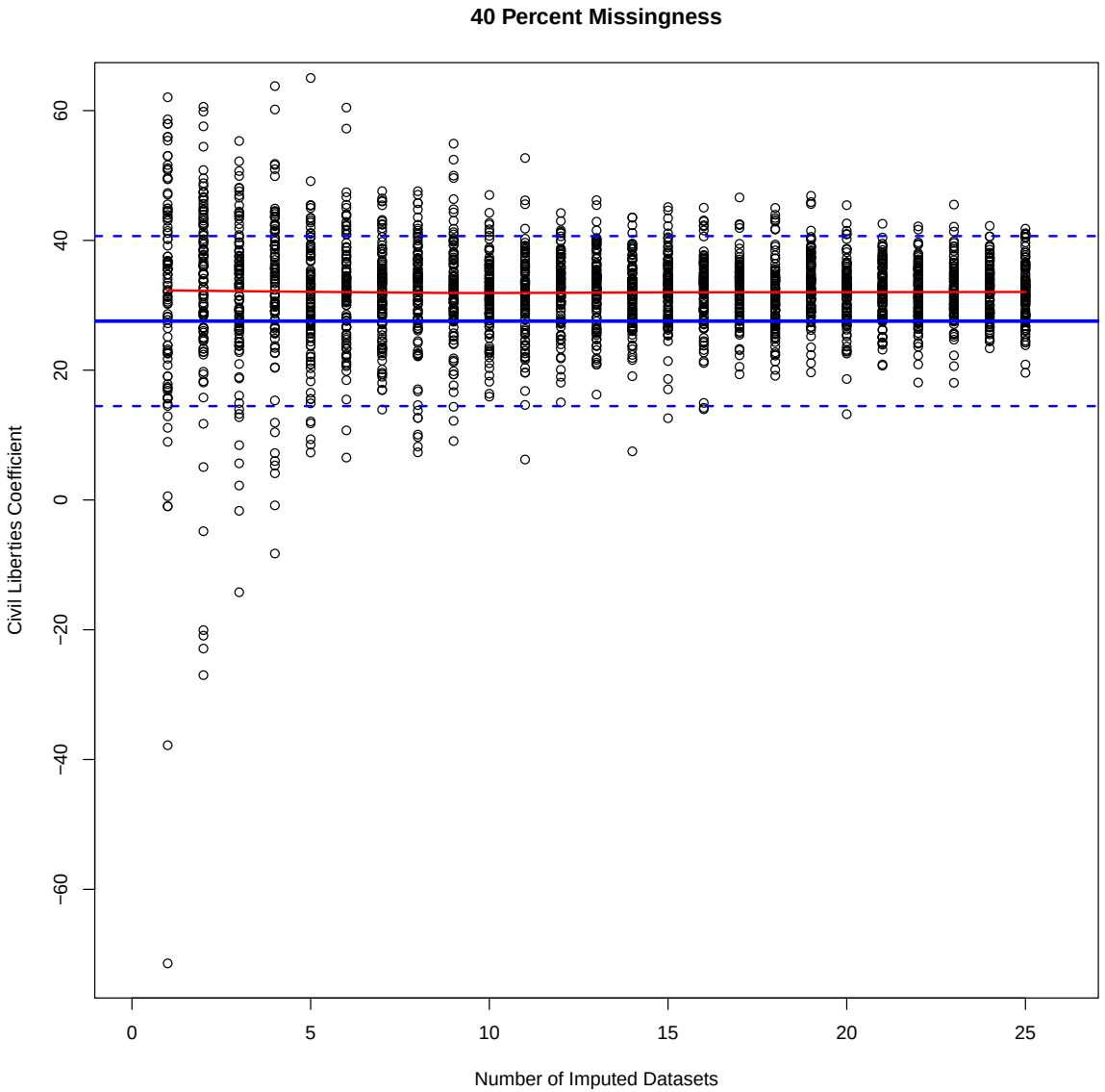
Multiple imputation creates multiple imputed datasets. It is not surprising that this would be necessary because we have a great deal of uncertainty about our imputed data which we want to integrate over when we do our usual analysis with the imputed data. What is surprising is that the number of imputed datasets

needed to account for this uncertainty is general quite small, on the order of $m = 5$ or 10.

Evidence of Needing Few Datasets



Evidence of Needing Few Datasets



Combining Estimates from the M Datasets (Rubin's Rule)

Once we have our 5 or 10 imputed datasets, we can run the model/analysis we would have run, but one each imputed dataset.

Rubin (1987) showed that the appropriate formulae for combining quantity-of-interest estimates from different datasets are as follows:

$$\bar{q} = \frac{1}{m}\sum_{j=1}^{m} q_j$$

and

$$SE(\bar{q})^2 = \frac{1}{m}\sum_{j=1}^{m} SE(q_j)^2 + S_q^2(1 + \frac{1}{m})$$

where $S_q^2 = \sum_{j=1}^{m}(q_j - \bar{q})^2/(m-1)$.

Outline

Basic Syntax

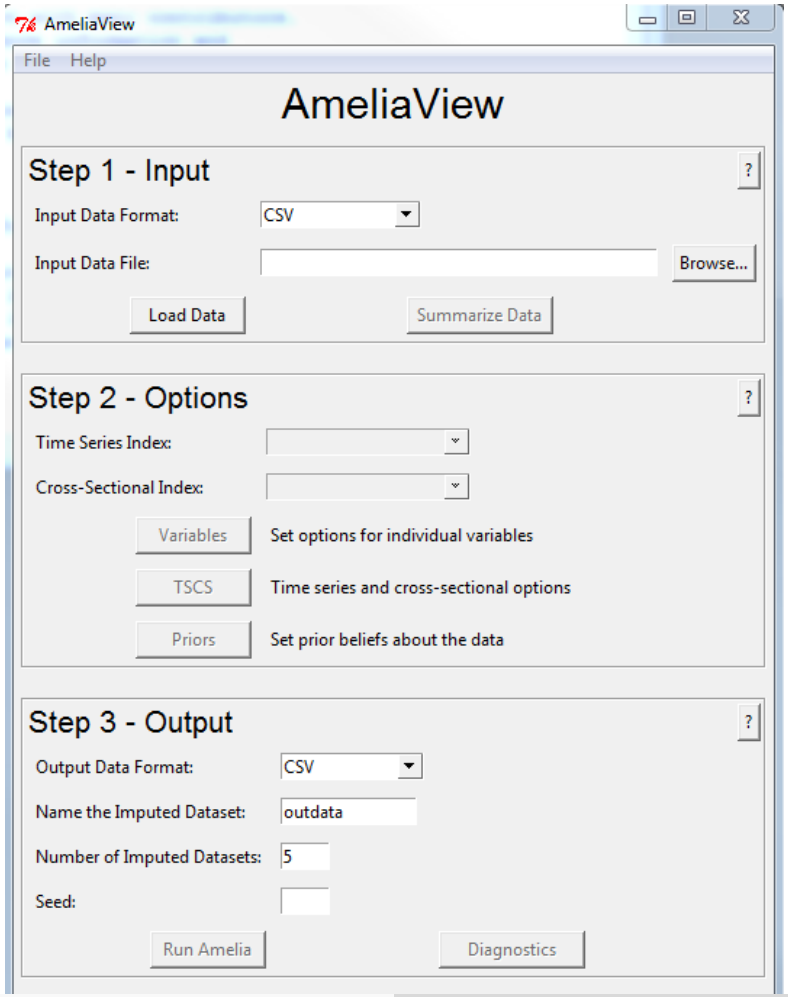
```
library(Amelia)
data(africa)

> africa[1,]
  year       country gdp_pc  infl trade civlib population
1 1972 Burkina Faso    377 -2.92 29.69    0.5    5848380

a.out <- amelia(x = africa, cs = "country", ts = "year",
  logs = "gdp_pc", m = 5)

write.amelia(obj=a.out, file.stem = "a.out")
```

a.out is a list of 5 imputed datasets, each of which can be accessed using a.out$imputations[[i]].

GUI

You can use the GUI by typing:

```
AmeliaView()
```



Basic Syntax

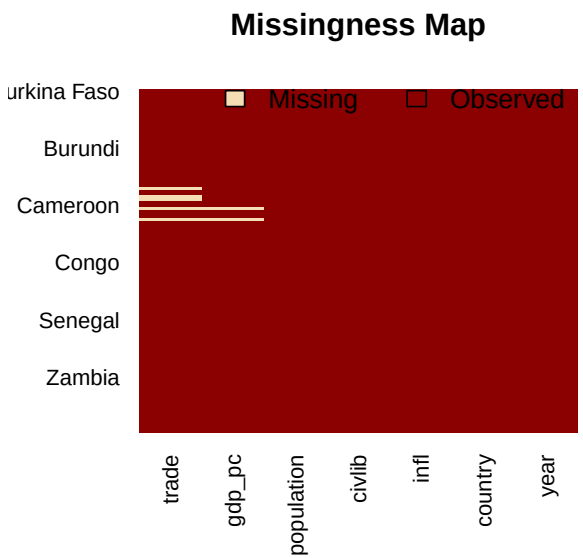
```
z.out.imp <- zelig(trade ~ log(population) + log(gdp_pc) + infl
  + civlib, data = a.out$imputations, model = "ls")
summary(z.out.imp)

Coefficients:
                   Value Std. Error t-stat   p-value
(Intercept)     112.4097   45.47779  2.472 1.345e-02
log(population) -17.8513    2.36646 -7.543 4.595e-14
log(gdp_pc)      31.3875    2.31108 13.581 1.834e-41
infl              0.2605    0.05836  4.463 8.075e-06
civlib           27.2104    6.67210  4.078 4.59$e-05
```

Zelig will automatically combine the results of the different models, but if a model you are using isn't programmed in Zelig, it isn't hard to combine your estimates using the equations I showed you before.

Diagnostics

The missingness map gives an overall sense of the shape and extent of the missingness.

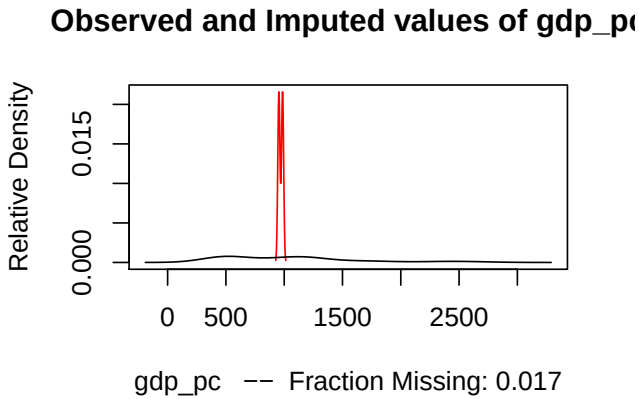
```
missmap(a.out)
```

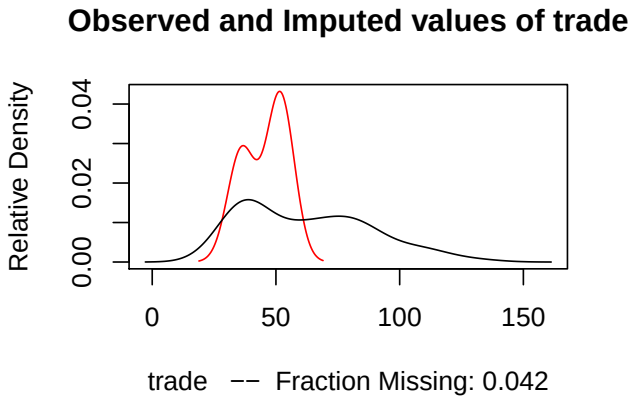


Diagnostics

Plotting the Amelia object contrasts empirical and imputed densities.

```
plot(a.out)
```

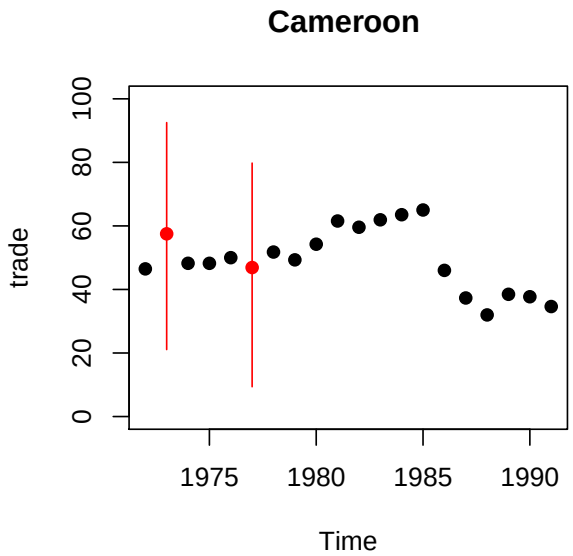




Diagnostics

Plotting Time-series cross-sectional plots

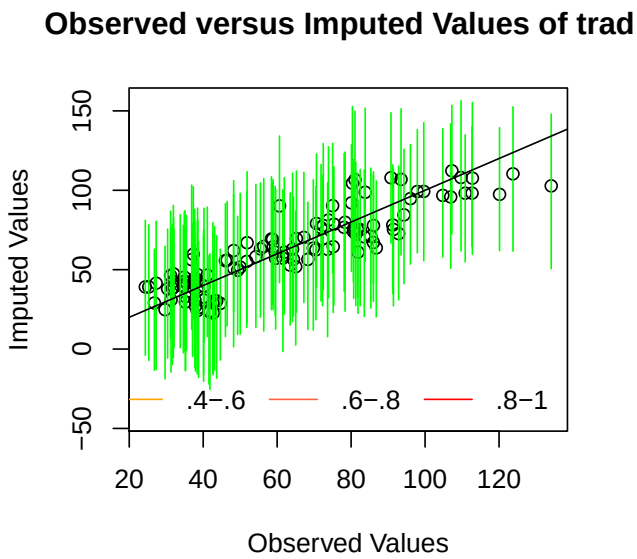
```
tscsPlot(a.out, var = "trade", cs = "Burundi")
```



Diagnostics

Overimputation for a specific variable tests the imputation model by imagining that each observation is missing and generating some imputations to check performance.

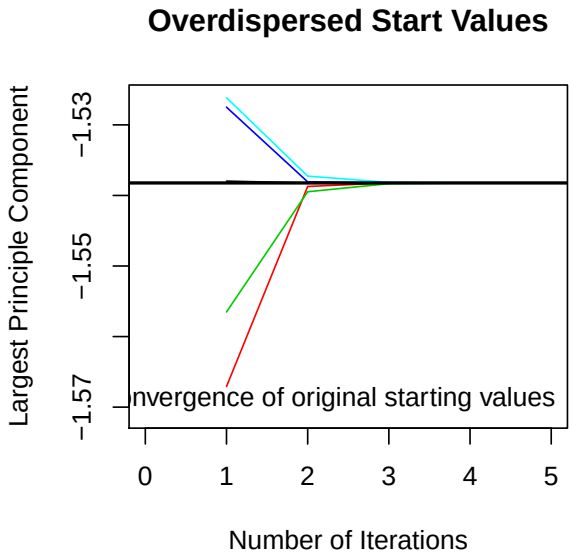
```
overimpute(a.out, var = "trade")
```



Diagnostics

The disperse function starts the algorithm at some unlikely values to check that AMELIA II hasn't found a local rather a global maximum for the likelihood of the complete data.

```
disperse(a.out, dims = 1, m = 5)
```



Considerations: Transformations

Recall that our model assumes multivariate normal data. This suggests some issues:

1. Ordinal variables: imputation is faster and more informative if ordinal variables are permitted to be continuous, but if undesirable use `ords` argument to constrain.
2. Nominal (unordered) variables: AMELIA II automatically converts into factors and imputes accordingly if a variable is passed to the `noms` argument.
3. Various transformations (logarithmic, root) are pre-programmed to make skewed distributions more normal.

Considerations: Time Series/Panel Data

1. Use the `ts` and `cs` arguments to tell AMELIA II the panel structure. It uses this information for constructing time serieses correctly.
2. `polytime` can be used to add in polynomial time trends (up to a cubic) and `intercs = TRUE` creates a separate trend for each panel.
3. Add in lags and leads if you think it will help predict the missing data. Use `lags` and `leads` with some variable.

Things to remember

1. Set the seed!
2. Include any variable in the analysis model in your imputation model.
3. Don't impute things that don't make sense.
4. Check diagnostics (and think carefully about applicability).
5. Remember transformations, polynomials and data structure.

Gary's Papers

King, Gary; James Honaker; Ann Joseph; Kenneth Scheve. 2001. "Analyzing Incomplete Political Science Data: An Alternative Algorithm for Multiple Imputation," *APSR* 95, 1 (March, 2001): 49-69.

James Honaker and Gary King. "What to do about Missing Values in Time Series Cross-Section Data," *AJPS* 54, 2 (April, 2010): 561-581

Amelia II: A Program for Missing Data

Gary's Other Interesting Work on This

Gelman, Andrew, Gary King, and Chuanhai Liu. "Not Asked and Not Answered: Multiple Imputation for Multiple Surveys." *Journal of the American Statistical Association* 93 (1999): 846-85.

Survey

Horton and Kleinman have a nice survey of methods:

Horton, Nicholas J. and Ken P. Kleinman. "Much Ado About Nothing: A Comparison of Missing Data Methods and Software to Fit Incomplete Data Regression Models" *Journal of the American Statistical Association.* Vol. 61, No 1. (2007)

http://maven.smith.edu/ nhorton/muchado.pdf

Gelman's Work

Andrew Gelman and coauthors are developing a new package for `R` called `mi`:

Su, Yu-Sung, Andrew Gelman, Jennifer Hill and Masanao Yajima. "Multiple Imputation with Diagnostics (mi) in R: Opening Windows into the Black Box" *Journal of Statistical Software*, Forthcoming.

Gelman, Andrew et. al. "Model Imputation for Model Checking: Completed-Data Plots with Missing and Latent Data" *Biometrics* 61, 74-85 (2005).

Abayomi, Kobi, Andrew Gelman and Marc Levy. "Diagnostics for multivariate imputations. *Applied Statistics.* 57, 273-291 (2008).

Andrew Gelman and Jennifer Hill. (2007). *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge University Press