“Known unknowns”: using multiple imputation to fill in the blanks for missing data

James Stanley
Department of Public Health
University of Otago, Wellington

james.stanley@otago.ac.nz

Acknowledgments

- Cancer Trends team: Caroline Shaw, Tony Blakely, June Atkinson, Di Safarti, Matt Soeberg.

- Particular thanks to June for SAS help.

- Gordon Purdie
- Kristie Carter

- Statistics New Zealand for accommodating our technical requests!

Overview of talk

- Why might we use multiple imputation?
- How does the imputation process work?
- How can the results be interpreted?
- How does this work *in practice*? An example from the Cancer Trends project.

The missing data problem

- Data is often missing in quantitative research, especially in retrospective studies or in large scale surveys.

- The simplest solution has always been to just analyse the “complete” dataset, and ignore cases who were missing data.

Pitfalls of analysing “complete” data only

- Inefficient analysis;
- In a complex statistical model: adding extra risk factors will mean more cases will be excluded from the analysis;
- If there is systematic bias as to WHY some people have missing data, then the results of the analysis may be biased.

How missing is missing?

- There are several types of “missing” data described in the literature:
  - Missing completely at random [MCAR]
    - (no pattern as to why data are missing)
  - Missing not at random [MNAR]
    - (data are missing based on a unknown factor)
  - Missing at random [MAR]
    - (data are missing in a systematic fashion, related to values of other variables in dataset)
- The labels are confusing. What does this actually mean for analysis?
  - MCAR: “complete” data are in effect a random subsample of all the people in the study.
  - MNAR: we cannot infer any information about the missing data values from the rest of a case’s data.
  - MAR: we can make some assumptions about the missing data, conditional on the intact information.

Imputation

- So...what does imputation do?
- Very simply, we replace the missing data with new, derived values.
- These derived values are based on what is happening elsewhere in the dataset.
- Then we can analyse this imputed dataset using standard statistical methods.

(n.b. – this would be a *single* imputation process)

Multiple imputation

- Single imputation produces a single dataset. We can analyse this to produce estimates of odds ratios (OR) or rate ratios (RR) etc.
- However, interpreting these results assumes that the imputation process is “accurate”.
- Performing this multiple times allows the data to reflect the inaccuracy of the process.

Imputation: how it works

A purposefully vague overview

Imputation: filling in the blanks

- How do we fill in the blanks? There are several methods available:

1) mean or mode substitution;
2) hot deck procedures;
3) regression based procedures.

Mean or mode substitution

- For continuous numeric data (e.g. weight):
  - calculate the mean value over the complete dataset and substitute this for the missing data;
- For categorical data (e.g., ethnicity):
  - Calculate the modal value (most common) over the complete dataset and substitute this for missing data.
- Highly likely to introduce bias to results.
- Therefore not recommended.

Hot deck imputation

- Step 1: identify cases with missing data:

ID #	Disease	Sex	Age grp	Ethnicity	NZDep
342678	1	F	25-44	NZ Euro	[missing]

- Step 2: Find cases who match on non-missing elements:

ID #	Disease	Sex	Age grp	Ethnicity	NZDep
795134	1	F	25-44	NZ Euro	2
146845	1	F	25-44	NZ Euro	5
986458	1	F	25-44	NZ Euro	4

- Choose one of these records at random, and “borrow” the NZDep value for person #342678...

Hot deck imputation

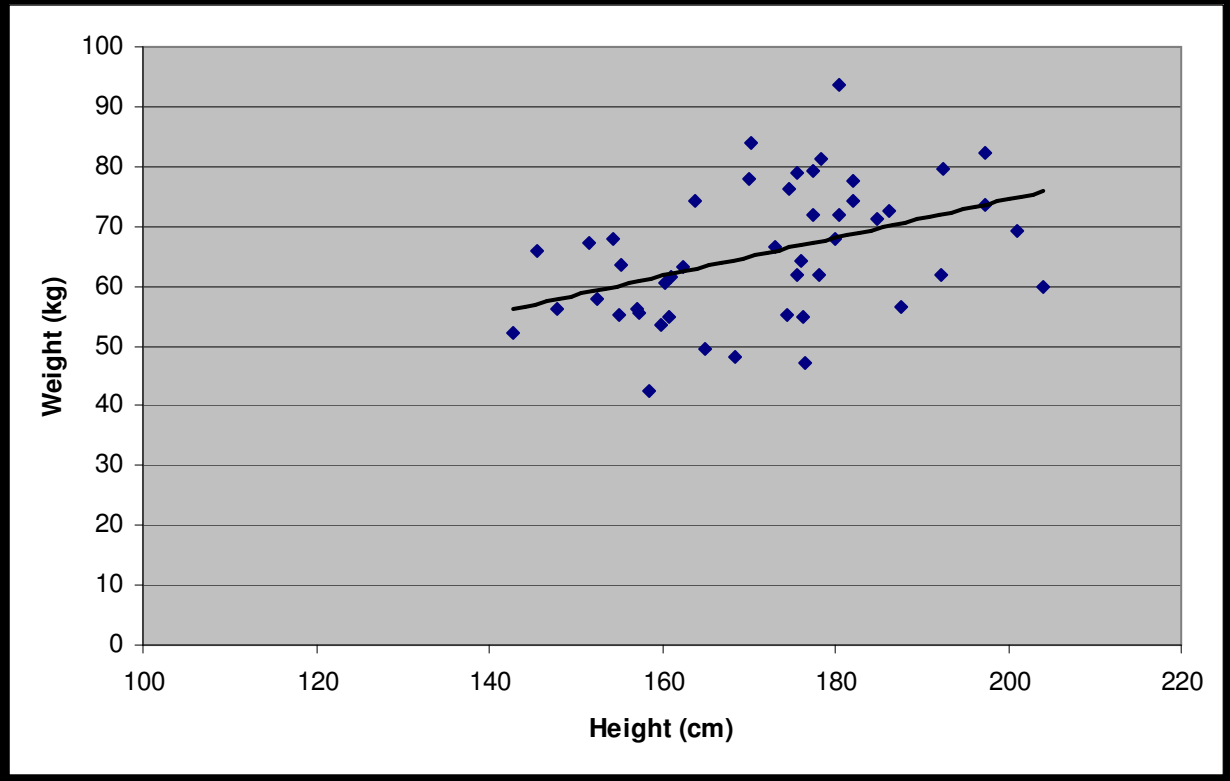
- Imputed data are from “real” observations, and therefore reflect plausible values.
- However, procedure is difficult to implement...
- ...especially when data are missing on more than one variable...
- ...or if the analysis model is so complex that few/no matches exist between those missing data and those with complete data.

Regression based imputation

- [As used for Cancer Trends]
- For those people who have complete data, run (e.g.) a cumulative logistic regression model to predict NZDep values – given predictor variables age, sex, ethnicity, & personal income.
- Apply this regression model to the person missing NZDep information: what value of NZDep is most likely for someone with that age, sex, ethnicity, & personal income?

Regression based imputation: a simple example

- Based on the complete data we have observed:



W = 10.5 + 0.33 x H

We can use this regression line for these data to predict what the weight would be for someone who was 172 cm tall.

W = 10.5 + 0.32 x 172

W = 65.54 kg

This predicted weight of 65.54 kg would then be used as the imputed weight value for this person.

Complications

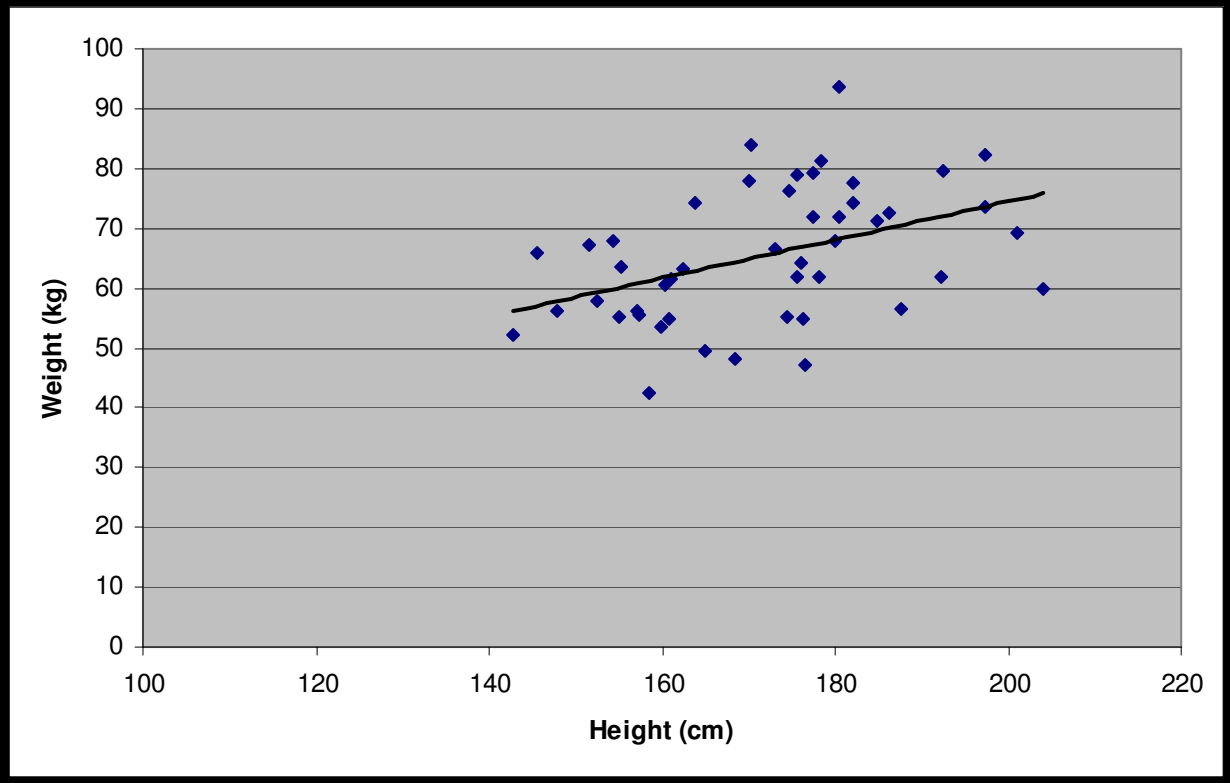
- If we just take the imputed value from the regression line, we run into problems.

1) the predicted value presumes the imputation model used was “accurate”...

2) we would get the same imputed value of 67.26 kg for each of our multiple imputation datasets;

3) When it comes time to calculate a confidence interval from our analyses, using the “best fit” prediction value will underestimate the width of our confidence interval (i.e. our analysis will look more accurate than it truly is)

Regression based imputation: a simple example



Our regression of weight on height has some residual error.

(the space between the individual data points and the line)

When we calculated the predictive regression equation, we got error terms associated with both the intercept and slope parameters.

Dealing with estimation accuracy

- Results of regression analysis:

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	10.518	17.262		.609	.545
	height	.320	.100	.430	3.198	.003

a. Dependent Variable: weight

- We use random draws from the confidence intervals of both intercept & slope to allow the data to reflect the model's accuracy.

e.g., slope to use = coefficient + random standard normal deviate x std. error of slope

Moving from single to multiple imputations

- So far we have just created one dataset.

- We can analyse this dataset just like any other

(e.g. Poisson regression to identify rate ratios associated with several risk factors for a disease.)

RR (Female to male) = 1.84, 95% CI (1.64, 2.06)

single to multiple...

- We could repeat the imputation procedure multiple times...

- Because of the random elements in the imputation procedure, across these datasets there will be:

1) identical values across datasets for originally “intact” data points;
2) different values across datasets for originally “missing” data points.

How many times?

- Rubin (1987) shows that the efficiency of imputation depends on both the proportion of missing data, and the number of imputed datasets.

	γ (proportion missing data)			
m	0.1	0.2	0.3	0.5
2	95	91	87	80
3	97	94	91	86
5	98	96	94	91
10	99	98	97	95

- For Cancer Trends, chose 3 imputations.

[Formula for table from Rubin, 1987]

Analysis of multiply imputed data

- Simply perform analysis on each dataset as per standard analysis

  (e.g. Poisson regression; logistic regression).

- We then end up with estimates for each *parameter* in each set of results:

  Dataset 1: RR (Female to male) = 1.84, 95% CI (1.64, 2.06)
  Dataset 2: RR (Female to male) = 1.77, 95% CI (1.58, 1.97)
  Dataset 3: RR (Female to male) = 1.93, 95% CI (1.73, 2.16)

- How do we combine results across these?

Variability between imputed datasets

- We can now see how our estimates vary over the multiple imputations we have run.
- The consistency of the results across datasets should be reflected in the confidence interval around our combined result.
- The calculation of this confidence interval has to take into account both this variability between datasets, and also the estimation accuracy *within* each dataset.

Dataset 1: RR (Female to male) = 1.84, 95% CI (1.64, 2.06)
Dataset 2: RR (Female to male) = 1.77, 95% CI (1.58, 1.97)
Dataset 3: RR (Female to male) = 1.93, 95% CI (1.73, 2.16)

(n.b. if estimates are RR/OR calculated on log scale, perform calculations in this log scale before converting back to RR/OR)

- Calculate average pooled estimate across the three analyses
- Calculate the mean variance for the three estimates;
- Combine this with the variability *across* datasets;
- Gives pooled variance estimate:

$$\text{Total var} = \frac{1}{m}\sum_{j=1}^{m} s_j^2 + (1+\frac{1}{m})(\frac{1}{m-1}\sum_{j=1}^{m}(\hat{P}_j - \overline{P})^2$$

mean variance

weighted between-imputations variance

A real example: linear regression of left ventricular mass (in g)

TABLE 7. Comparison of results from three imputations of left ventricular mass

Covariate	Imputation 1		Imputation 2		Imputation 3	
	Coefficient (SE*)	*p* value	Coefficient (SE)	*p* value	Coefficient (SE)	*p* value
Age, years	0.24 (0.12)	0.041	0.45 (0.12)	<0.0001	0.45 (0.12)	<0.0001
Male sex	21.6 (1.5)	<0.0001	21.7 (1.4)	<0.0001	20.8 (1.4)	<0.0001
Current smoker	1.9 (1.9)	0.32	2.2 (1.9)	0.23	3.9 (1.9)	0.04
Weight, pounds†	0.56 (0.02)	<0.0001	0.56 (0.02)	<0.0001	0.58 (0.02)	<0.0001
History of myocardial infarction	8.7 (2.1)	<0.0001	9.3 (2.1)	<0.0001	10.4 (2.1)	<0.0001
History of congestive heart failure	32.2 (3.1)	<0.0001	28.0 (3.0)	<0.0001	28.0 (3.0)	<0.0001
Diastolic blood pressure, mmHg	−0.29 (0.06)	<0.0001	−0.264 (0.07)	<0.0001	−0.245 (0.06)	<0.0001
Systolic blood pressure, mmHg	0.29 (0.04)	<0.0001	0.266 (0.04)	<0.0001	0.253 (0.03)	<0.0001
History of hypertension	7.0 (1.3)	<0.0001	6.3 (1.3)	<0.0001	5.9 (1.3)	<0.0001
Total cholesterol, mg/dl	−0.038 (0.02)	0.019	−0.021 (0.02)	0.18	−0.035 (0.02)	0.03
High density lipoprotein cholesterol, mg/dl	−0.11 (0.04)	0.011	−0.12 (0.04)	0.005	−0.13 (0.04)	0.002
Major electrocardiogram abnormality	19.3 (1.4)	<0.0001	17.9 (1.4)	<0.0001	16.8 (1.4)	<0.0001
Minor electrocardiogram abnormality	8.2 (1.2)	<0.0001	7.3 (1.2)	<0.0001	8.0 (1.2)	<0.0001

* SE, standard error.
† One pound = 0.45 kg.

Arnold & Kronmal, 2003.

A real example: linear regression of left ventricular mass (in g)

TABLE 7. Comparison of results from three imputations of left ventricular mass

Covariate	Imputation 1		Imputation 2		Imputation 3	
	Coefficient (SE*)	*p* value	Coefficient (SE)	*p* value	Coefficient (SE)	*p* value
Age, years	0.24 (0.12)	0.041	0.45 (0.12)	<0.0001	0.45 (0.12)	<0.0001
Male sex	21.6 (1.5)	<0.0001	21.7 (1.4)	<0.0001	20.8 (1.4)	<0.0001
Current smoker	1.9 (1.9)	0.32	2.2 (1.9)	0.23	3.9 (1.9)	0.04

Arnold & Kronmal, 2003.

- The effect of sex was relatively constant over all three imputed datasets.
- Current smoking status appeared to have a different impact on left ventricular mass in the different imputed datasets.

A real example: linear regression of left ventricular mass (in g)

TABLE 7. Comparison of results from three imputations of left ventricular mass

Covariate	Imputation 1		Imputation 2		Imputation 3	
	Coefficient (SE*)	*p* value	Coefficient (SE)	*p* value	Coefficient (SE)	*p* value
Age, years	0.24 (0.12)	0.041	0.45 (0.12)	<0.0001	0.45 (0.12)	<0.0001
Male sex	21.6 (1.5)	<0.0001	21.7 (1.4)	<0.0001	20.8 (1.4)	<0.0001
Current smoker	1.9 (1.9)	0.32	2.2 (1.9)	0.23	3.9 (1.9)	0.04

TABLE 6. Comparison of complete case and multiple imputation model results for left ventricular mass

Covariate	Complete case (n = 3,177)			Combined imputation (n = 5,201)		
	Coefficient (SE*)	95% CI*	*p* value	Coefficient (SE)	95% CI	*p* value
Age, years	0.30 (0.15)	0.01, 0.60	0.044	0.38 (0.18)	–0.07, 0.83	0.08
Male sex	19.9 (1.8)	16.4, 23.4	<0.0001	21.4 (1.5)	18.3, 24.4	<0.0001
Current smoker	4.2 (2.3)	–0.34, 8.8	0.07	2.7 (2.3)	–2.0, 7.4	0.25

Arnold & Kronmal, 2003.

Practicalities

The multiple imputation procedure for Cancer Trends

Background

- The Cancer Trends project:

> "aims to determine trends in cancer incidence and survival in New Zealand from 1981 onwards, by ethnic group and socio-economic status. *CancerTrends* uses anonymously and probabilistically linked cancer registrations and census records of all people who developed cancer from 1981 to 2004 in New Zealand."

http://www.otago.ac.nz/wsmhs/academic/dph/research/HIRP/cancertrends/CancerIndex.html

Census data and cancer trends

- There are a total of 5 census datasets being used in the CT study (from 1981 to 2001).
- Total number of records in 2001 in excess of 4,000,000.
- Of these, approx. 2.8m were NZ resident adults.

2001 Census and missing data

- 2.8 million NZ adults.
- Variables of interest:
  - age, sex, qualifications, personal income*

78.8% had data for all these variables.

15.1% were missing data on 1 variable only.

6.1% were missing data on 2+ variables.

*used to derive total equivalised household income

Imputing values

- Step 1: deciding on models to predict the to-be-imputed variables.

e.g. for personal income:

cumulative logistic regression model
(14 income categories)

Predictors:

sex, age group, ethnicity, school quals, tertiary quals, NZDep, TA [nb. all categorical]

Imputing values in SAS

- Step 2: imputing values
- ...is simple, if imputing continuous (e.g., weight) or ordered categorical variable (e.g. income).
- Can use PROC MI – surprisingly straightforward.
- If you have a non-ordered categorical variable (e.g. ethnicity) – then welcome to the (exciting) world of writing macros...

Multiple values to impute

78.8% had data for all these variables.

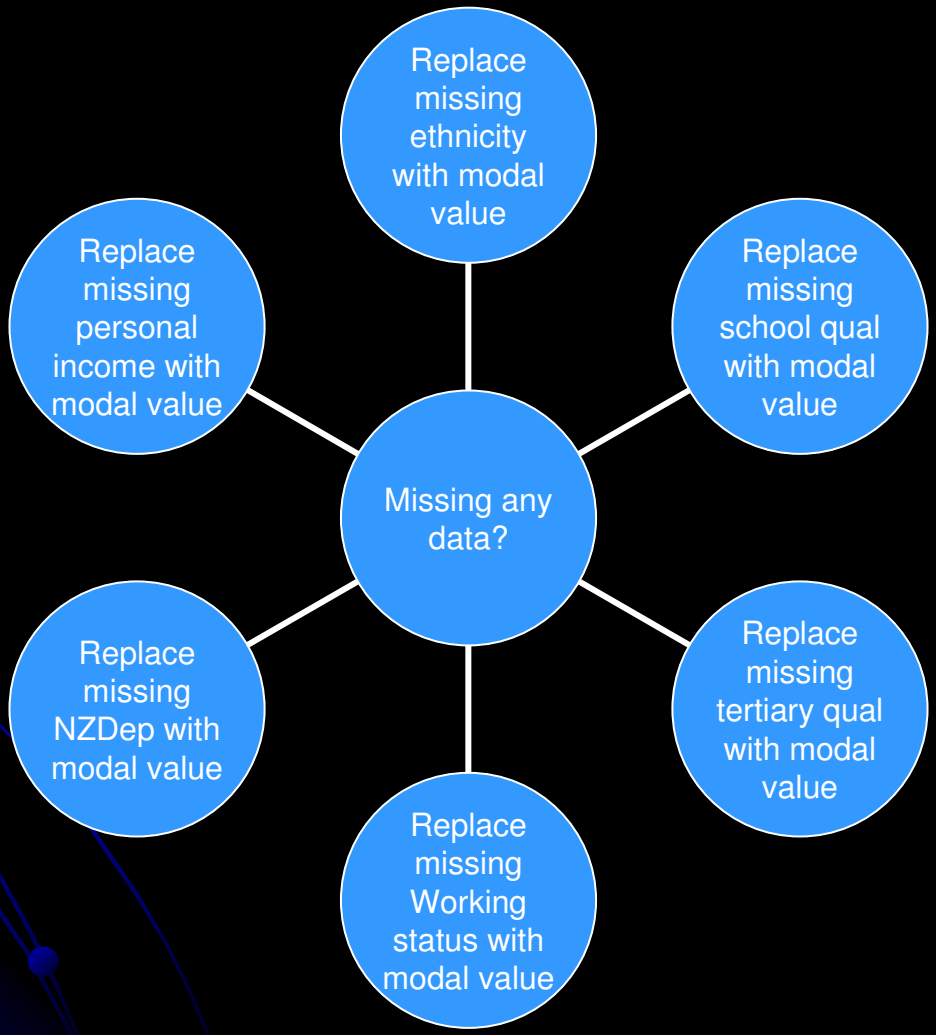
15.1% were missing data on 1 variable only.

6.1% were missing data on 2+ variables.

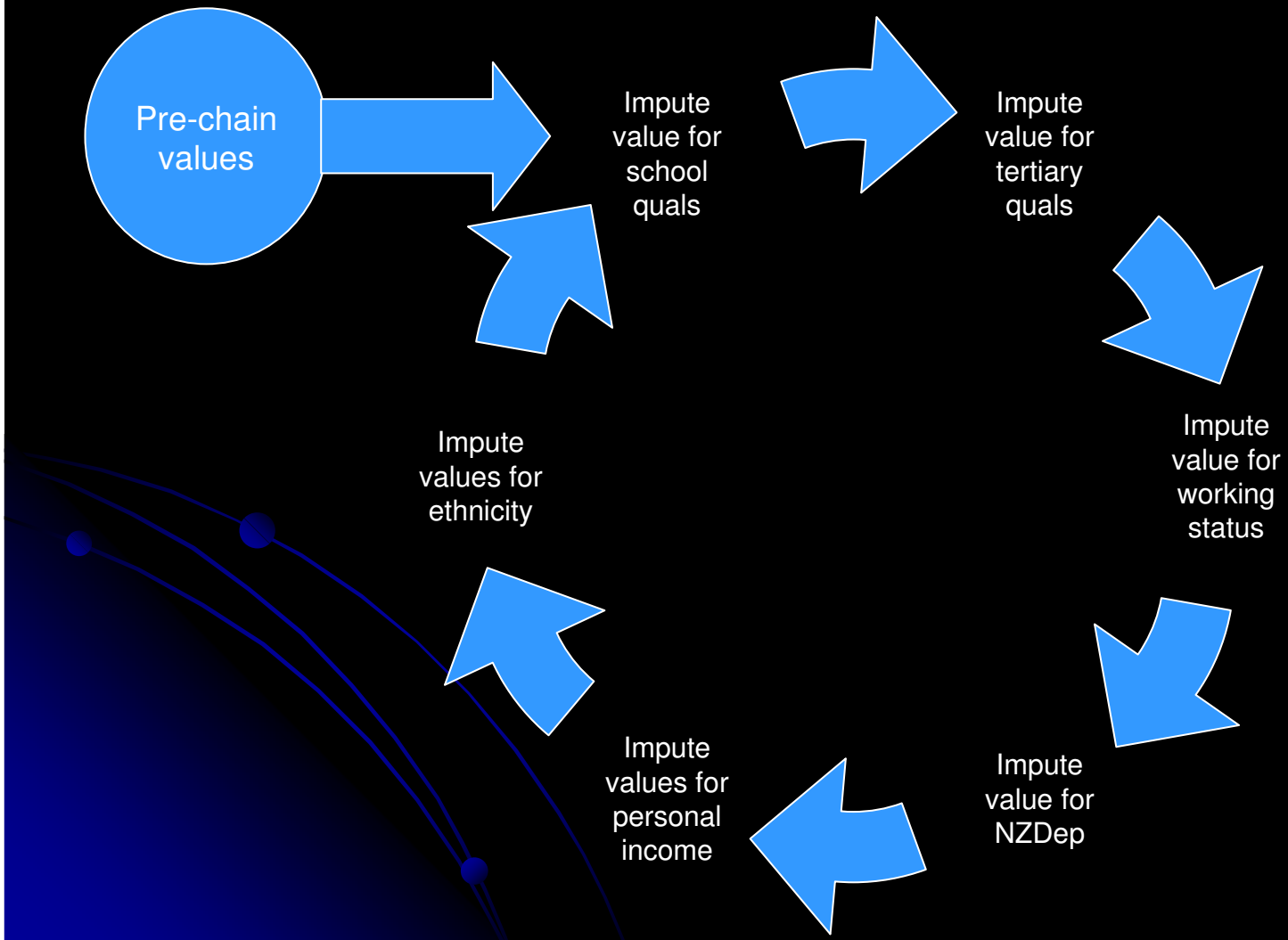
- What about people missing data on multiple variables?

- In these cases, our imputation model might need to predict a variable (e.g. income) on the basis of other variables which might have missing values...
- How do we deal with this?
- Answer: chained equations...

Pre-chain setup: starting values



The chain


Pre-chain values
Impute value for school quals
Impute value for tertiary quals
Impute value for working status
Impute value for NZDep
Impute values for personal income
Impute values for ethnicity

Rinse and repeat

- For a single census year, creating three imputed datasets (variables prev. noted) takes about 42 hrs. This was with ten iterations (10 loops around the chain).

[Now using just five iterations: 21 hours].

- Bulk of time is in the logistic regression modelling for predicting ethnicity (complex model, even when restricted to five possible ethnicity groups).

Analysing the data

- Step 3: Performing analyses & combining results across imputed datasets.
- Working on this now!
- Again, SAS has developed procedures to combine these analyses. Relatively simple to do by macro as well...

[and someone in the Cancer Trends team can tell you the results later]

References and further reading

- Arnold AM and Kronmal RA (2003). Multiple Imputation of Baseline Data in the Cardiovascular Health Study. American Journal of Epidemiology, 157, 74-84.

- Donders AR, van der Heijden GJ, Stijnen T, Moons KG. (1996). Review: a gentle introduction to imputation of missing values. *J Clin Epidemiol., 59*,1087-91.

- http://www.stat.psu.edu/~jls/mifaq.html

- Rubin (1987). *Multiple Imputation for Nonresponse in Surveys*. J. Wiley & Sons, New York.