Multiple Imputation and Simulation Methods

Brian D. Ripley

Professor of Applied Statistics
University of Oxford

Assumptions

Need certain irreducible assumptions about the mechanism whereby the data become missing.

Missing at Random (MAR)

Whether the data which was actually missing was missed does not depend on what the missed values were, given the observed data.

Ignorable

The parameters of the ‘missing’ mechanism do not depend on those of the data-generating mechanism.

Observed at Random (OAR)

The pattern of missing data does not depend on the actual data.

Likelihood and Bayesian methods need the first two to justify ignoring the ‘missing’ mechanism.

Frequentist methods need OAR.

Example – Weighing machine

Ten items weighed to find common weight θ. Scale sometimes fails to read. Believe observations are $X_i \sim N(\theta, 1)$.

(a) Fails 10% of the time, irrespective of θ and X_i.

(b) Fails if $X_i < 27$.

(c) Fails with probability $\theta/(1+\theta)$.

Then we have

Case	MAR	ignorable	OAR
(a)	✔	✔	✔
(b)	✘[1]	✔	✘
(c)	✔	✘	✔

[1] If any data are actually missing.

Prediction

Suppose we want to predict y from explanatory variables x. Do so by predictive density $p(y \mid x)$. Suppose $x = (x_O, x_U)$ where only x_O is observed in this example. Given *MAR* and *ignorable*, the predictive density is

$$p(y \mid x_O) = E\big[p(y \mid x_O, X_U) \;\big|\; X_U\big].$$

That is, we average over the missing values X_U drawn from the conditional distribution $p(x_U \mid x_O)$.

Do this by simulation, drawing M samples to complete the data, and averaging over the predictions. (Note: densities.)

This is called *multiple imputation*. In a sample survey all variables are categorical, so find the other questionnaires that agree on x_O, and select randomly from them—‘hot-deck’ imputation.

If X_U is discrete and takes relatively few values, alternatively use frequencies for all possible values rather than simulate.

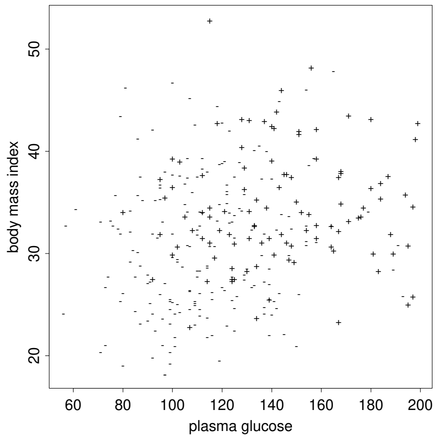
Diabetes in Pima Indians

A population of 768 women who were at least 21 years old, of Pima Indian heritage and living near Phoenix, Arizona, was tested for diabetes according to World Health Organization criteria. The reported variables are

number of pregnancies
plasma glucose concentration (5 missing)
diastolic blood pressure (mm Hg) (35)
triceps skin fold thickness (mm) (227)
serum insulin (μU/ml) (374)
body mass index (weight in kg/(height in m)^2) (11)
diabetes pedigree function
age in years

Of the 768 records, 376 were incomplete. Most of our illustrations omit serum insulin and use the 532 complete records on the remaining variables. These were randomly split into a training set of size 200 and a test set of size 332. Methods which can deal with missing values were given 100 of the incomplete cases as part of the training set.

Consider logistic discrimination fitted to complete observations of plasma glucose and body mass index:



The presence or absence of diabetes against plasma glucose and body mass index for the training set of the Pima Indians data.

To predict when one is missing (both never are) need to model the conditionals in this figure (ignoring symbols): joint normal or non-parametric?

Note: this will *not* be the same as plugging in the mean, mode, ...

Modelling conditionals

Need to find $p(x_U \mid x_O)$ from the training data. A common special case is *monotone* missing data, where the variables can be ordered (possibly in groups) so that if x_i is missing, so is $x_j, j > i$.

Then we can model $p(x_2 \mid x_1)$ from complete (x_1, x_2) examples, $p(x_3 \mid x_1, x_2)$ from complete (x_1, x_2, x_3) examples, and so on. Then is $x_k, k > m$ are missing, we have

$$p(x_U \mid x_O) = \prod_{i<m} p(x_i \mid x_j, j < i).$$

Can model the conditionals any way we like. Monotonicity reduces drastically the number of patterns of missing data.

More generally, may have no complete data at all. Can use *data augmentation* methods (Tanner & Wong, 1987, but known earlier).

Slovenia, 1990

Opinion poll of 2074 adults about a plebiscite. To count, had to attend and vote for independence, say proportion θ.

		Independence		
Secession	Attendance	Yes	No	NA
Yes	Yes	1191	8	21
	No	8	0	4
	NA	107	3	9
No	Yes	158	68	29
	No	7	14	3
	NA	18	43	31
NA	Yes	90	2	109
	No	1	2	25
	NA	19	8	96

Third question ('secession from Yugoslavia') may help resolve the NAs ("don't knowks"), but 96 people admit to no opinion at all. Complete data give $\widehat{\theta} = 1439/1549 = 0.929$, but if NA is really 'no', $\widehat{\theta} = 1439/2074 = 0.694$.

Is this MAR and ignorable? Assume NAs are concealed responses.

Using EM on just attendance and independence gives $\widehat{\theta} = 0.892$ with approximate 95% prediction interval $[0.873, 0.908]$; adding the secession question gives $\widehat{\theta} = 0.883$ with 95% interval $[0.863, 0.900]$.

Data augmentation reassigns the partial NAs to the 8 complete cells. How this is done depends on the model for the 2^3 table of complete data. Data augmentation iterates:

- Sample the missing data M times from the current parameter distribution.
- Fit using the M complete datasets and take current parameter distribution as the average over these M estimates.

The plebiscite gave $\theta = 0.885$, with 5.0% voting NO, and 6.5% not attending. The model under-estimated non-attendance at 4.3%.

Multiple Imputation for Neural Networks and Classification Trees

Brian D. Ripley

Professor of Applied Statistics
University of Oxford

Training with Missing Values

Want to model $p(y \mid x)$ for complete x. Training set $\mathcal{T}$ might have

- Partially missing x's.
- Partially observed or completely missing y's.

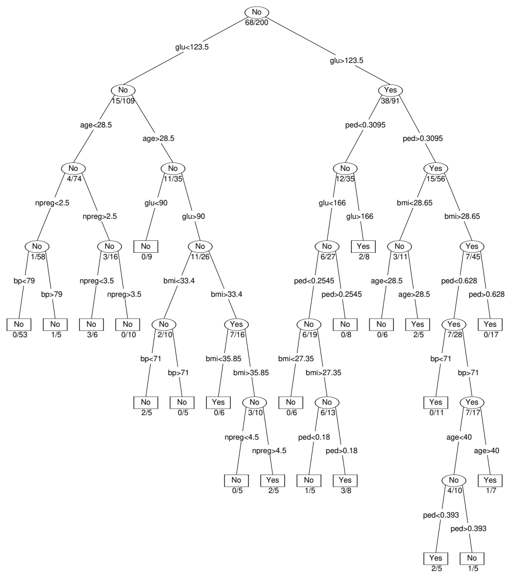
Normally only consider censored y's, but could be incompletely categorized cases in a classification problem.

Slovenia poll is of this form, if prediction had been 'vote for independence'.

Not-quite-correct shortcut is to complete each missing x ($m = 10$?) times, and include all these cases in the training with weights $1/m$.

[Incorrect, as we should average $p(y \mid x_O, X_U)$, and this (usually) averages $\log p(y \mid x_O, X_U)$ in a likelihood.]

Classification Trees

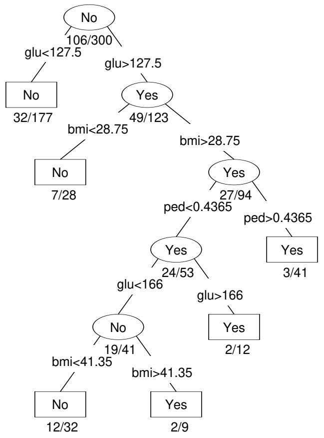


The classification tree grown for the complete examples of the Pima Indians diabetes data.

What do we do about missing examples? All attributes are discretized, so use *fractional cases*.

In theory we should use $p(x_j \mid x_O)$ for the variable x_j at the split, but Quinlan used the distribution of x_j amongst cases reaching that node (which keeps the procedure recursive).

Quinlan was approximate: can find $p(y \mid x_O)$ exactly and use it in tree construction:



The classification tree grown and pruned for the Pima Indians diabetes data based on a training set with missing data.

The (deviance-reduction) criterion used here is biased towards nearly-complete attributes.

Neural Networks

Usual to use the *diagnostic paradigm* and only model $p(y \mid x)$, not $p(x)$. Need aspects of $p(x)$ for handling missing data (and also for outliers).

Consider the diabetes data. We know y in all cases. The classification tree suggests the most important explanatory variables are plasma glucose, body mass index and pedigree. In our training set of size 300, glucose was missing twice and bmi is missing 5 times, whereas the pedigree was complete.

This is a classification problem, so we need to work with $\mathsf{Pr}(\text{diabetic} \mid x_O)$. Options:

- Train a separate neural network for each pattern of x_O, possibly 3 patterns here if only these variables are used, 8 for all the variables but only 4 are at all common.
- Train one network using multiply imputed data. In the likelihood we would average $P = \mathsf{Pr}(\text{diabetic} \mid x_O, X_U)$ over X_U (and not average the term in the likelihood, $\log P$ or $\log(1-P)$, but may be similar.)

The second approach has the advantage that whereas only three variables seem to be good at predicting diabetes, other variables may be good at the predicting the missing X_U. Thus we can decompose the task into smaller pieces. It has automatic consistency properties. (First approach could predict 'diabetic' ix x_3 was missing, but 'not diabetic' for any known value of x_3.)

How precisely do we fill in the missing values here? May as well use the complete data to find the distributions of glu | rest and bmi | rest. These are continuous variates, so could use a regression network, or one that predicts $N(\mu(x), \sigma^2(x))$ or even a mixture of normals. For categorical variables we would use a classification network.

For the two missing values of plasma glucose (with bmi = 24.7, 32) we predict means of 96 and 118, with standard error of 28. For the five missing values of bmi we predict means of 63.6, 50.5, 51.5, 53.1 and 59.2 with a standard error of about 6. In short, the missing values are not very predictable, and they do affect the probabilities P, so EM imputation would be misleading.