2. Presence of Missing Values

- Missingness is multivariate
- Explore joint distribution of missings, and their association with variables
- Evaluate imputation schemes

Simple Visualization Strategies

- Complete case analysis
- Assign some fixed value
  - Pro: Easy identification of missings
  - Con: Problematic for high-d views
- Assign imputed values
  - Pro: Data and all plots available
  - Con: Missingness information is lost
Advanced Visualization Strategies

- MANET: Missings Are Now Equally Treated
  (Unwin, Theus, Hofmann, et al)
  MANET incorporates missing values in barcharts,
  histograms, boxplots, dotplots, scatterplots, etc.
- XGobi
  Natural extensions of linked scatterplot paradigm

Missing Values in XGobi

- Prerequisite: Retention of case-value identity
- Optional display of missing values
- Quick reset of fixed or imputed values
- Multiple linked windows
  - One window (…) plotting the data
  - One window plotting a multivariate indicator matrix of
    missingness: that is, jittered zeroes and ones
Defining Missing Values in XGobi

- Use "." or "na" or "NA" in the data file
- Create an filename.missing file
  - Allows the investigation of censored data with missings

Exploration of Imputation Methods in XGobi

- Reads user-supplied files of sets of imputed values
- Allows rapid cycling through imputations
Variation in Work Efficiency During Exercise

What is the effect of subject characteristics on work efficiency?

- 136 human subjects
- Covariates: sex, age, body weight, %body fat, ...
- 3 walk and 4 step exercises, AM and PM
- ... 14 repeated measures for each primary response variable: mechanical work, efficiency, etc.
- Missing == failed exercise

(Data courtesy of Ming Sun, MD and George Reed, Vanderbilt University Medical Center)

Efficiency Exercise: Covariates

The points are brushed according to sex.
Mechanical Work: Repeated Measures

By default, missing values are shown as zeroes.

Aside: Parallel Coordinates Plots

- Axes are parallel instead of orthogonal
- Each case is represented by a line instead of a point
- Strength: Many variables in a single view
- Weakness: Overplotting! Interaction is essential.
Parallel Coordinates Plot of Covariates

One displays the data for men, and one for women.
Mechanical Work

For this data, something like a time series plot ...

Efficiency

Missings are not drawn in the parallel coordinates view.


**Efficiency Data: Observations**

- Using parallel coordinates plots
  - Work monotonic in exertion; not so efficiency
  - Morning and afternoon similar
  - Most “missings” on hardest step

- Using missing values plot
  - Number of misses increased with exertion
  - Correlation between missingness and % body fat
  - Correlation between number of missings and % body fat

**Efficiency data: Exploring imputation schemes**

- Univariate imputation is inadequate because correlation structure is ignored
- Multivariate imputation (courtesy of Chuan-hai Liu) is considerably better
Importing Imputed Values into XGobi

- `filename.imp`: Each column contains all imputed values
- `filename.impnames`: one imputation name to a line

`filename.dat` (Three rows from a set of data with missings)

```
467 585 na2 43 na4
463 580 86 67 12
379 na1 na3 88 12
```

`filename.imp` (Four sets of imputed values)

```
583 582 584 580 (imputed values for na1)
86 90 85 84 (imputed values for na2)
86 90 88 87 (imputed values for na3)
12 14 11 13 (imputed values for na4)
```

Quick Quiz

Are the salient features in each of these plots potentially due to coding of missing values?