Design and Analysis of Sample Surveys

Andrew Gelman
Department of Statistics and Department of Political Science
Columbia University

Class 3a: Missing-data imputation

Example

- Google did an internet poll a day or two before the election
- Not a simple random sample
- Expect sample to be:
  - Younger than the population
  - More affluent than the population
  - More urban than the population
  - Different than the population for many other reasons
- Without adjustment, poll would be basically useless
- Could do a *weighted analysis*
- $N = 9483$ but "sampling weights" w_i missing for 2099 people i

Whats wrong with throwing away observations?

- *Available-case* analysis
- Throwing away observations is wrong, but is it harmful?
- Given a sample, could draw a simple random subsample
  - Does not bias a (weighted) estimate of a mean
  - Does make the (weighted) estimate of a mean less precise
- Is the subsample of 7384 observations where weight w_i is observed a simple random subsample from the 9483 observations?

Why are some sampling weights missing?

- In Google's poll, w_i depends on sex, age, and region
- Other surveys adjust the weights by race, income, etc.
- So, w_i is missing if sex, age, or region is missing for respondent i
- Region has 4 levels and is missing if state is missing for i
- Sex, age, or state is missing if i refuses to answer
- Missing rates: 16% on sex, 21% on age, 1% on state
- Is missingness of w_i independent of candidate support?
  - It turns out that people who refuse to give their sex, age, or state are more likely to be young women
  - Young women are more likely than others to favor Obama
- Throwing away observations with missing w_i induces bias

Missingness on preferences for candidates

- Question was: “Who do you want to win the election?”
  - **A:** Obama / Biden, the Democrats (4152)
  - **B:** Romney / Ryan, the Republicans (3708)
  - **C:** Third party candidate / Undecided (1623)
- C is poorly worded but should C’s be considered missing?
- Also, likely-voter screening

Conclusions from polling example

- Missingness makes even “simple;; estimation complicated
  - Issues get even more complicated with regressions, etc.
- Dropping incomplete observations is common but not a fix
  - At best, your estimates are less precise
  - At worst, your estimates are biased
  - Strong implicit assumption about missingness mechanism
- Need to understand *why* observations are incomplete
  - Observations can be incomplete for multiple reasons
  - Different solutions for different missingness mechanisms
  - Data collector does not (fully) control the selection process

Missing-data imputation

- Crude imputation
- Generic imputation
- Substantive model-based imputation
- In statistics and political science: lots of work on generic imputation
  - Generic imputation can be pretty good if lots of predictors are included
  - Can handle selection bias—if the info used in the selection is included in the model
  - Discuss examples

Missing data

- Unit nonresponse: some people can't be reached or refuse to participate
- Item nonresponse: refusal to answer some questions
- Intentional missing data
    - Intentional missing *units*
    - Intentional missing *items*
- Unintentional missing data
- Structural missing data

Some practical data issues

- Partial information on nonrespondents
- Partially missing data (censoring, truncation, rounding, heaping)
- Missing-data codes (0, -999, NA, blank, . . .)

Some practical survey issues

- Callbacks
- Estimating bias arising from nonresponse
- Question wording, balancing goals of cost, accuracy, and response rate

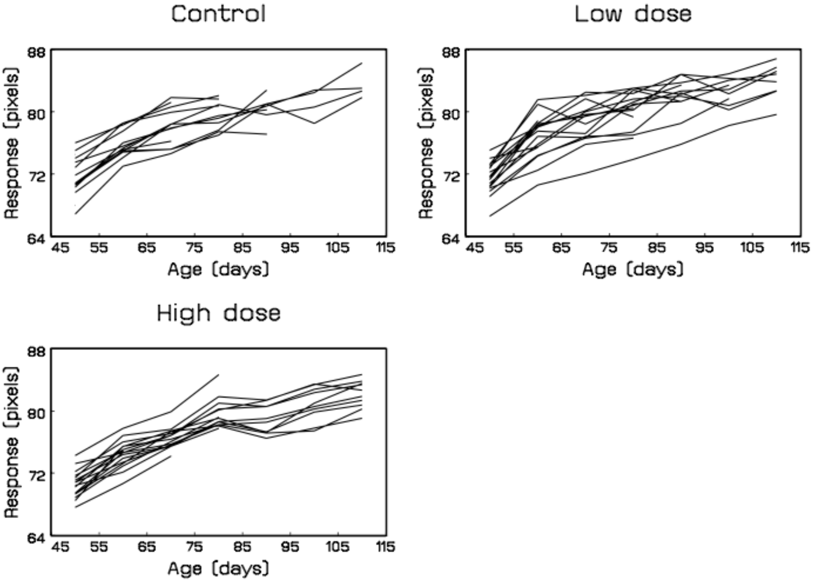
Defining the question of interest

- Which presidential candidate has the most support?
- Population is all adults in United States (250 million)
- Ignore electoral college, identifying "likely voters," etc.
- If whole population were available, question is answerable
- Answer can be estimated using sample from population

Answer under simple random sampling

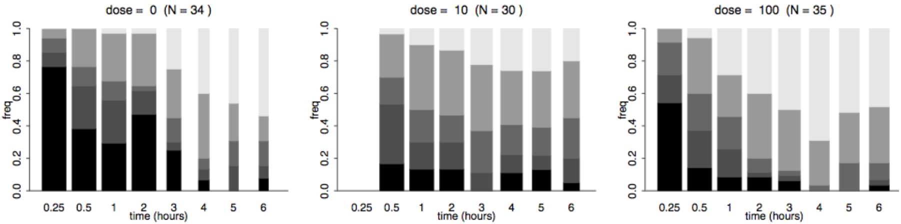
- Description of simple random sampling (SRS):
    - Put entire population in a hat (set "sampling frame")
    - Randomly draw N people from hat
    - Ask them who they support for president
- This is a (very) special case of a missing data problem:
    - N people from population are selected to be observed
    - 250,000,000N people from population are missing

Missing data in longitudinal studies

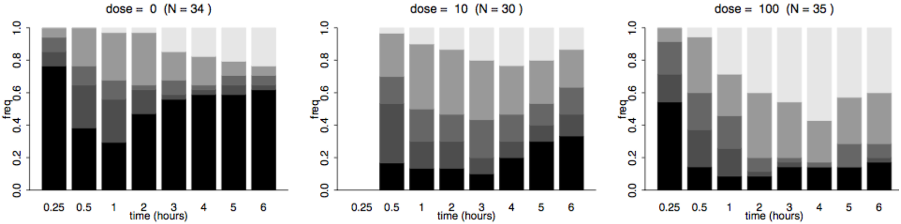


Observed and completed data

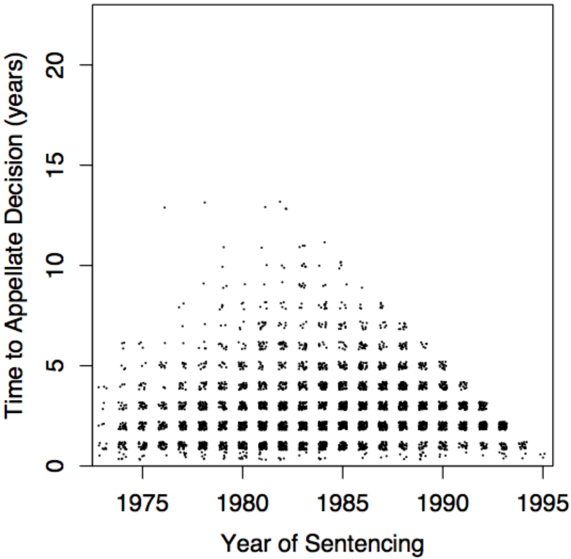
Observed data display



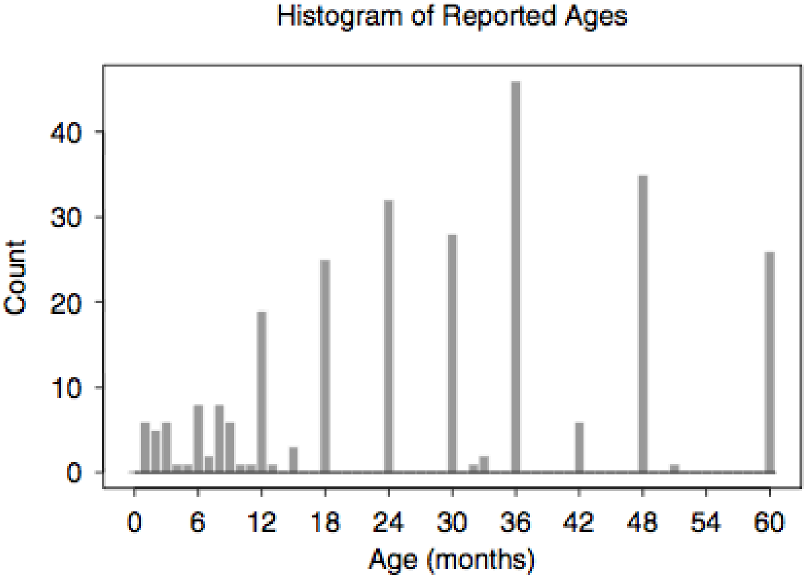
Completed data display



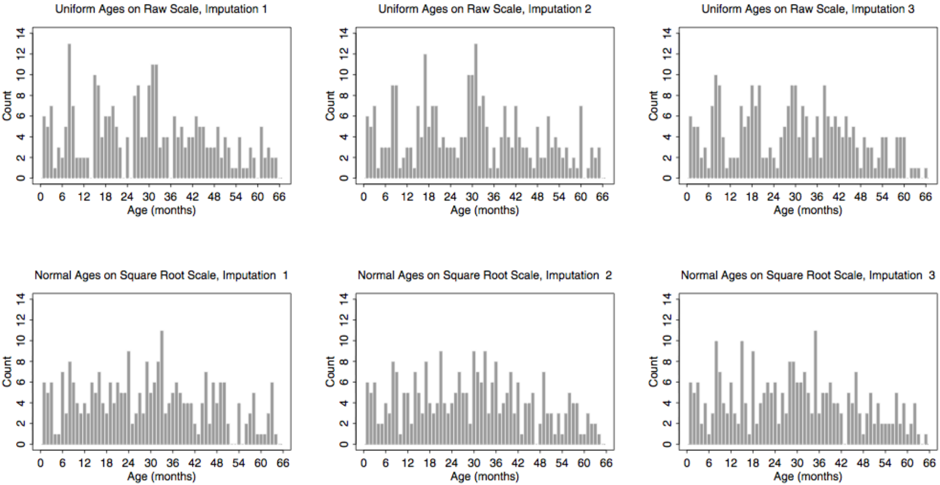
Censored data



Heaped data


Histogram of Reported Ages
Count
0
10
20
30
40
Age (months)
0
6
12
18
24
30
36
42
48
54
60

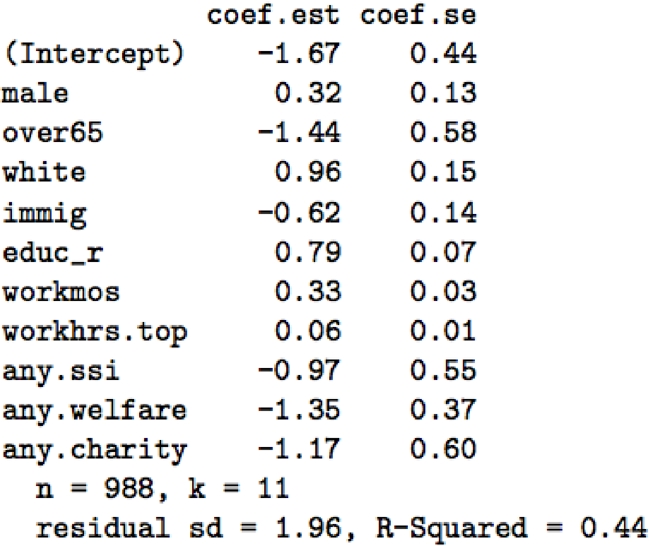
Comparing different imputations



Missing-data mechanisms

- Missing completely at random
- Missing at random (conditional on observed information)
- Not missing at random
  - Missingness depending on unobserved predictors
  - Missingness depending on the missing value itself
- Add predictors to your model until missing-at-random is a plausible assumption

Regression imputation for square root of income



```
               coef.est coef.se
(Intercept)      -1.67    0.44
male              0.32    0.13
over65           -1.44    0.58
white             0.96    0.15
immig            -0.62    0.14
educ_r            0.79    0.07
workmos           0.33    0.03
workhrs.top       0.06    0.01
any.ssi          -0.97    0.55
any.welfare      -1.35    0.37
any.charity      -1.17    0.60
  n = 988, k = 11
  residual sd = 1.96, R-Squared = 0.44
```

Regression imputation: results

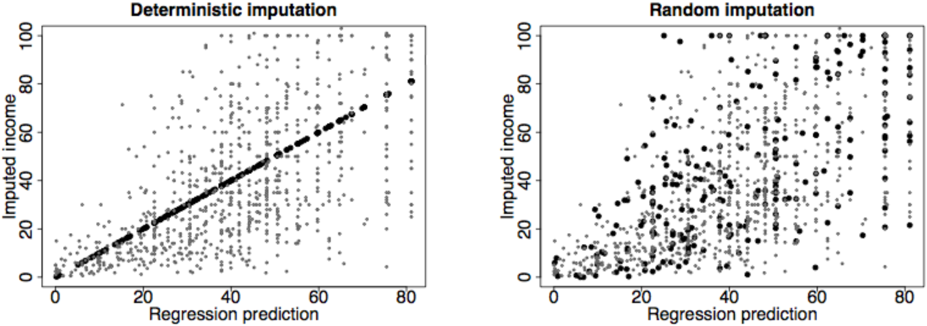
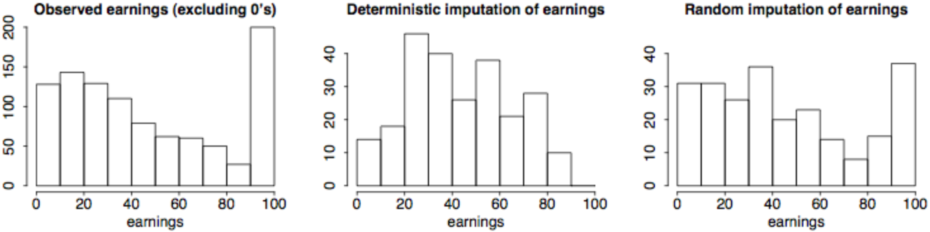


Figure 25.2 *Deterministic and random imputations for the 241 missing values of earnings in the Social Indicators Survey. The deterministic imputations are exactly at the regression predictions and ignore predictive uncertainty. In contrast, the random imputations are more variable and better capture the range of earnings in the data. See also Figure 25.1.*

Deterministic vs. random imputations



Programs for missing-data imputation

- R
  - mi (iterative regression imputation)
  - Hmisc (iterative regression imputation)
  - Amelia (multivariate normal)
- Stata
  - mi

Homework

- Due at beginning of class 5a and 6a
  1. Weighted analysis
  2. Poststratification
  3. Missing-data imputation
  4. Ratio and regression estimation