Modeling surveys

Ben Klemens
U.S. Cesus Bureau

USCENSUSBUREAU
Helping You Make Informed Decisions

The objective

Minimize:

$$ = costs

$|\widehat{\mu} - \mu|$ = bias

σ = variability

Why a simulation?

- *Lots* of distributions: A new distribution for each age/sex/race/location/question
  [Or just one very complicated distribution]

Why a simulation?

- *Lots* of distributions: A new distribution for each age/sex/race/location/question
  [Or just one very complicated distribution]
- Forget Bayesian Updating or other hierarchical models
- Advises how we code the simulation—use a statistics package for the back end.

Uses

- Draw out the cost/bias and cost/variance tradeoff functions.
- Responsive design: how can initial waves advise allocations in later waves?

Uses

- Draw out the cost/bias and cost/variance tradeoff functions.
- Responsive design: how can initial waves advise allocations in later waves?
- How much can imputation replace on-the-ground surveying?
  - How robust is our answer to bad model specification?
- A test bed for imputation methods: how do our imputation methods work with small samples of different characteristics?

How a survey works

- Generate a plan for sampling from the population
  - Requires a model of the population
  - Not so controversial: use basic demographics, focus on high-variance subpopulations and the especially interesting

How a survey works

- Generate a plan for sampling from the population
  - Requires a model of the population
  - Not so controversial: use basic demographics, focus on high-variance subpopulations and the especially interesting
- Send interviewers to the field/phone banks
  - What we’ve been modeling today

How a survey works

- Generate a plan for sampling from the population
  - Requires a model of the population
  - Not so controversial: use basic demographics, focus on high-variance subpopulations and the especially interesting
- Send interviewers to the field/phone banks
  - What we've been modeling today
- Clean the collected surveys and impute missing data
  - Requires a model of the population

The imagnry slider

- 0% survey response $\Rightarrow$ results via the imputation model
- 100% survey response $\Rightarrow$ results via survey
- All surveys are somewhere in between.
  - ¿Survey-assisted modeling?
  - ¿Model-assisted surveying?

TEA

An automated survey processing system

- Data cleaning
- Editing bad values
- *Missing data imputation*
- Disclosure avoidance
- Reweighting
- *Bonus utilities*

Survey simulation: quick overview

- Generate a population
- Send interviewer agents to gather the data
- Run TEA to find the output measures $(\$\$, |\widehat{\mu} - \mu|, \sigma)$.
- Analyze the outputs.

The model—very simple

- Simple & modular $\Rightarrow$ amenable to experiments
- Give every moving part a bypass switch.

The model—very simple

- Simple & modular $\Rightarrow$ amenable to experiments
- Give every moving part a bypass switch.
- Given two surveying or imputation strategies that produce different results
  - How many moving parts can I strip out of the model before the difference disappears?
  - How many moving parts can I add before the difference disappears?

The model—very simple

- Simple & modular $\Rightarrow$ amenable to experiments
- Give every moving part a bypass switch.
- Given two surveying or imputation strategies that produce different results
  - How many moving parts can I strip out of the model before the difference disappears?
  - How many moving parts can I add before the difference disappears?
- Or, write down an objective function $f(\$\$, |\widehat{\mu} - \mu|, \sigma)$—now the entire simulation is an optimization problem.
- More technique: by interpreting $f(\cdot)$ as a likelihood function (which is valid, or valid for a transformation), we can use statistics tools out of the box.

The model—agents

- Interviewers
  - Currently uniform characteristics. They just find respondents in their neighood not yet hit, drive out, and knock on the door.

The model—agents

- Interviewers
  - Currently uniform characteristics. They just find respondents in their neighood not yet hit, drive out, and knock on the door.
- Respondents
  - Age $\times$ Sex $\times$ Race $\times$ Characteristic 1 $\times \ldots \times$ Characteristic N
  - In this presentation: three types of respondent, one question.
  - Respondents are holding a complete survey.
  - Go home three (randomly chosen) times a day; odds of finding them are a kernel density PDF with humps at those three times.
  - Location is uniform through the neighborhood.

The model—survey procedure

- For each period:
  - Allocate interviewers (the responsive design step)
  - For each interviewer (thread here):
    - interviewer picks a respondent who has not yet answered
    - interviewer drives out (costs accrue)
    - Random draw decides whether respondent is home.
    - If respondent is home, collect survey (100% accurate and complete)

The model—post-processing procedure

- Total up costs; write all gathered surveys to a text file.
- Run TEA on the text file
- Apply the imputation method specified by the user
- Report CI for each statistic
- We know the true μ, and so can report true bias and MSE.

Multiple imputation

- Posit a model for the missing data. (Normal, Hot Deck, . . . , ¿ABM?)
- Fit it using the existing data
- Make draws from the new model to fill in missing data

So imputation is a modeling problem

- If your model is right, swell.
- If you have all the data, super. Our model is irrelevant anyway.

So imputation is a modeling problem

- If your model is right, swell.
- If you have all the data, super. Our model is irrelevant anyway.
- If your model is wrong and nonresponse is high, you're screwed.
- From this perspective, the survey is insurance against a bad model.

Confidence intervals and their caveats

- The *Multiple* Part:
  - Re-draw several sets of fill-in values; re-estimate the statistic
  - The statistic's variance = within-imputation variance + across imputation variance
  - Now express a confidence interval *based on the imputation model*

Confidence intervals and their caveats

- The *Multiple* Part:
  - Re-draw several sets of fill-in values; re-estimate the statistic
  - The statistic's variance = within-imputation variance + across imputation variance
  - Now express a confidence interval *based on the imputation model*
- A CI expresses the variability of a statistic—how much we can trust a model.
- A CI is calculated using the model we are testing.

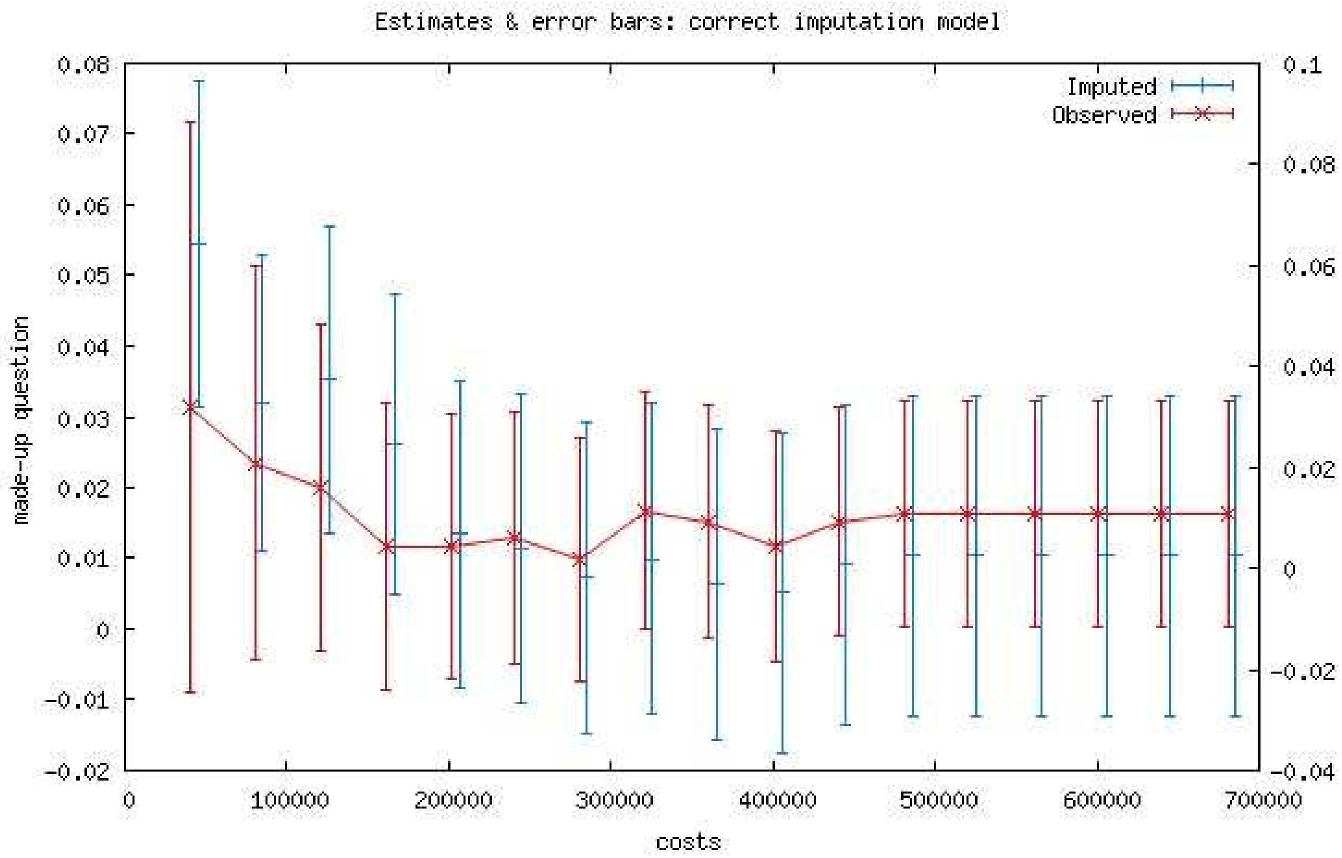
Confidence intervals and their caveats

- The *Multiple* Part:
  - Re-draw several sets of fill-in values; re-estimate the statistic
  - The statistic's variance = within-imputation variance + across imputation variance
  - Now express a confidence interval *based on the imputation model*
- A CI expresses the variability of a statistic—how much we can trust a model.
- A CI is calculated using the model we are testing.
- Supreme Court ruled that you shouldn't trust CIs too much. *Matrixx Initiatives, Inc., Et Al. V. Siracusano Et Al. (563 U. S. ___ 2011)*

An OK imputation

```
hood [Chicago]{
    q1-t1-dist: normal
    q1-t1-dist-params: 0, 1
}
impute{
    draw_count: 5
    categories: type
    models{
         q1 {
            method: normal
        }
    }
}
```

An OK imputation

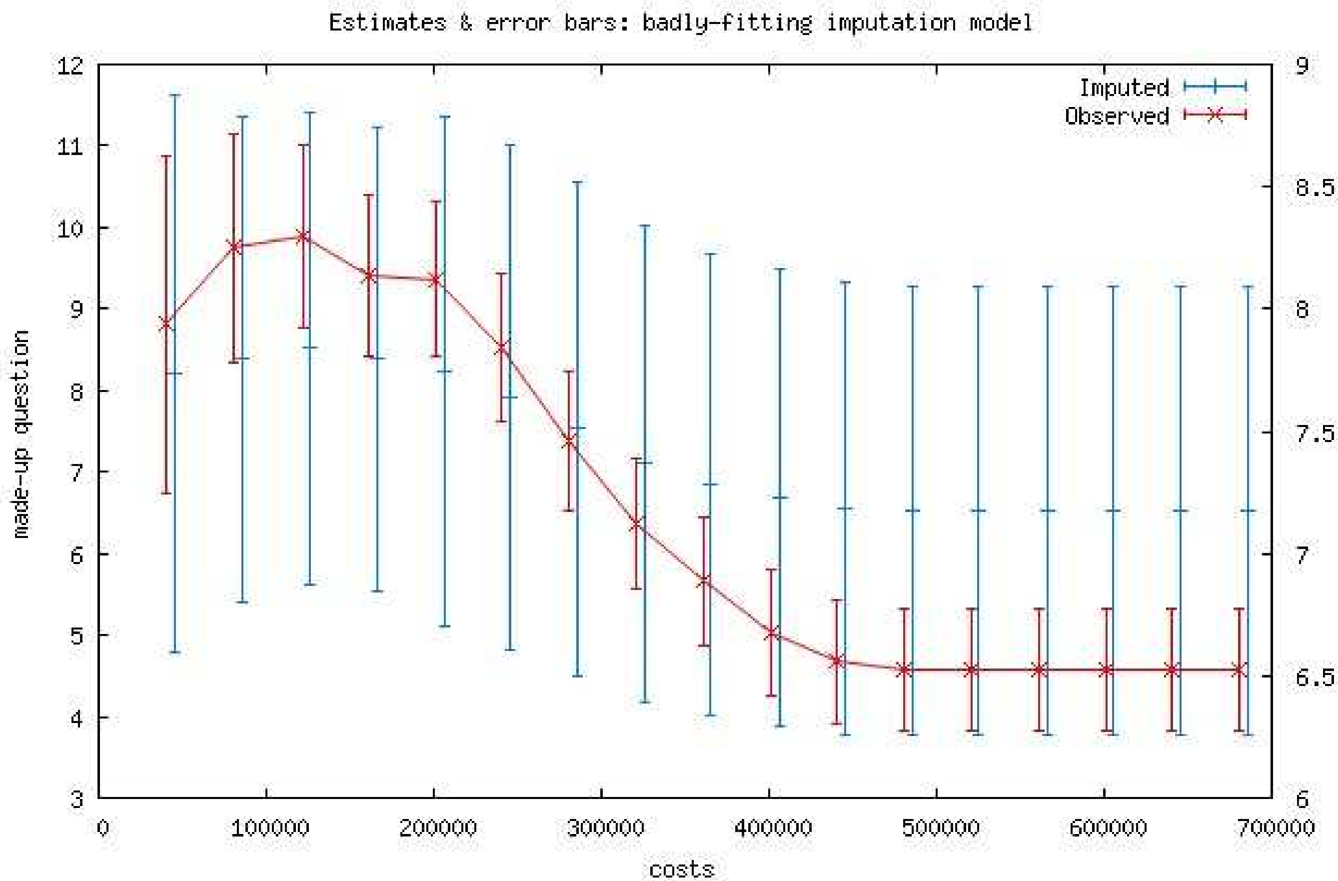


USCENSUSBUREAU
Helping You Make Informed Decisions

A bad imputation

```
hood [Chicago]{
    q1-t1-dist: lognormal
    q1-t1-dist-params: 2, 1
}
hood [Los Angeles]{
    q1-t1-dist: normal
    q1-t1-dist-params: 1, 2
}
impute{
    draw_count: 5
    categories: type
    models{
         q1 {
            method: normal
        }
    }
}
```

A bad imputation



Where to buy

- In C (~ 350 lines).
- The supporting libraries
  - GLib: commonly-used data structures, mutexes
  - GSL: vectors, matrices, RNGs, many simple distributions
  - SQLite: the database
  - Apophenia: more/complex distributions, threading, data management, db interface
  - TEA: data cleaning, editing, imputation

Conclusion

- ¡It’s a battle of the models!

Conclusion

- ¡It's a battle of the models!
- We *will* fill in some data via a model.
- Our simulations can evaluate the extent to which post-processing models can replace boots on the ground.
- Our simulations can evaluate the small-sample efficacy of the post-processing models we use.