Record Linkage: Inference Tools

Fritz Scheuren
NORC
University Of Chicago

What's in a Name?

- Why title this talk as I did?
- Why Not "Everything is Broken"? I own the book.
- How about we survived 2016 US/Haiti elections?
- Or did we? Time will tell?

How to Start/Begin?

- Linkages and Big Data?
- Start Where We Are?
- Go Forward from There?
- Theory Reminders?
- Practical Examples?
- Getting on with Nexts?

First Linkage + Big Data

- Linkage is a key big data tool
- Pioneering work on this by Fellegi-Sunter and later others
- Certainly a great boon in our emerging “Big Data” world
- But also it has key limitations
- Hence, addressing its error properties is needed

Record Linkage Paradigm

- Typically Data Linkage Errors are of two types
- Mismatches and Nonmatches
- Minimizing and Measuring these is a Key Challenge
- To Incorporate them into an RMSE for use in inference

Linkage and Big Data?

- **Linkage has always been about making data bigger?**
- **How to prevent data errors from also getting bigger?**
- **How to measure RMSE's for "Confidence Intervals"**
- **Sorry, still lots of Unsolved Problems A little more later?**

Start Where we are?

- **Fellegi-Sunter Model is where we will start**
- **It Involves just two files to be matched**
- **Moving to “n” files greater than two can be done serially**
- **But measuring all the Interactions is hard and still under development**

Measuring True Links

- Increases as the number of files linked increases
- But data linkage uncertainty usually increases too
- How much to factor this in/widen confidence bands?
- At some point diminishing returns sets in, usually early

Adjusting for True Nonlinks

- Reweighting true links to controls can be very useful
- Multiple models of matching errors useful too
- Factor in trade-offs in data quality re linkage quality
- “What if” simulations good too? Don’t rush! Sleep on it?

Go Forward from Here?

- **Try “stuff” Be bold, Be bold, but be not too bold (Emerson)**
- **Deepen Meta/Paradata much more systematically**
- **If lifetime learning is your goal, keep a Tagebuck**
- **Alas, one of my many regrets here that I did not do enough!**

Theory Reminders?

- Read widely in emerging literature
- Ask Ivan Fellegi what he would do today?
- Everything has not changed?
- But restudy the early pioneers
- Your linkage problems may be solved elsewhere

Practical Examples?

- **Use the eventual (70 + year) Decennial Census Data Public releases as a precedent?**
- **Seek improved Meta/Para documentation practices**
- **Lots of personal regrets here!**
- **Too soon old. Too late smart!!!**

Too soon Old/Too late Smart

- All of my suggestions have been taken by me at least once!
- Data trumps (sorry) theory! But how to document data for a future use?
- Even after 50+ years I am still not smart enough. Help?

How to End/Continue?

- Let's do this deepening more together?
- Maybe use the Second edition of our book on Data Quality?
- More books by others too, a very active area now?
- Still much the same themes but better and better tools

Reflections/Next Steps

- Love to deepen contacts
- Share more/Learn more?
- My Email address is still scheuren@aol.com
- Or contact in person here and now?

Thank You