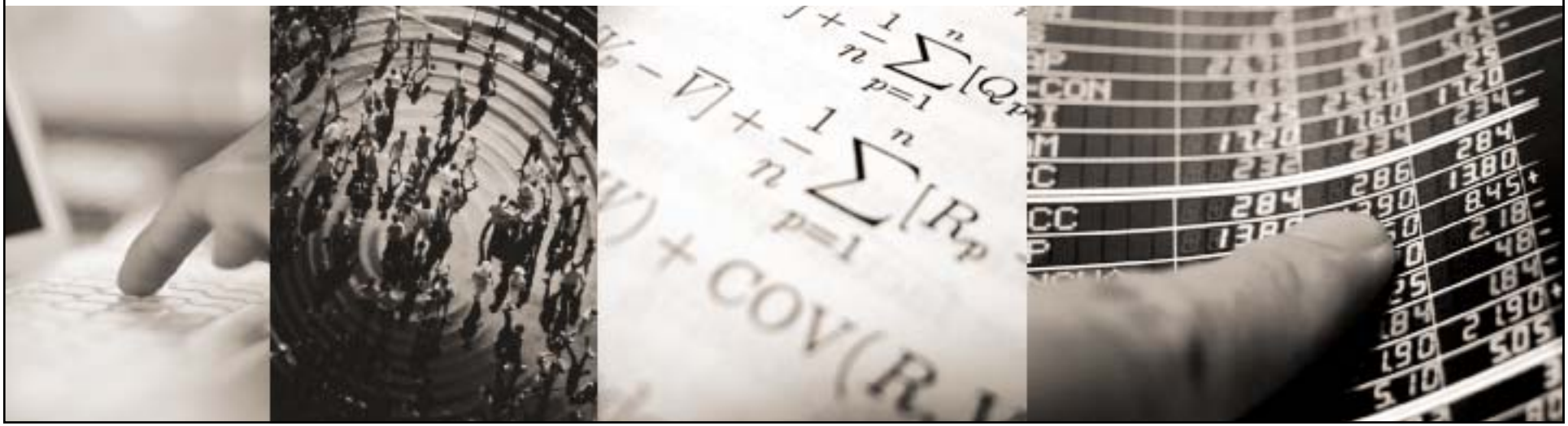
Using Linked Data

Julia Lane

NORC AT THE UNIVERSITY OF CHICAGO

A National Organization for Research and Computing



The Benefits of Linked Microdata

- Improved analysis of existing data, particularly simulation models
- Potential for new analysis from existing data (particularly admin records)
  - Information on health histories
  - Longitudinal information on earnings
  - Demand side of labor market
- Potential for linkages to new types of data becoming available on individuals (biomarkers; video; text)..access issues not addressed
- Increased access improves government's return on investment in data collection (GPRA; PART)


GOOD AFTERNOON MR. HENFENNEL! HEMORRHOID CREAM AISLE SIX!
BEEP!
TOO MUCH INFORMATION IN MY MEDICAL I.D. CHIP.
TOLES
©2004 THE WASHINGTON POST
DON'T FORGET MOUTHWASH. I FIGURED THAT ONE OUT BY MYSELF. —

The Challenges

- All data
  - Decreasing quality of public use files on households/individuals
  - Increased likelihood of reidentification => Future likelihood of no public use files
  - Particularly important for health and income data, given skewness of distribution (protection/synthetic data => reduction of information on most important populations)
- Linked data
  - Increased likelihood of reidentification
  - Admin records often received from enforcement agencies

Access Issues: Public Use Files
Example of Impact of Topcoding

Table 1: Estimated Effects of Race and Education on Log-Earnings
(estimated standard errors in parentheses)

	OLS1	OLS2	MLE	CLAD	SCLS	ICLAD
Black-White Gap						
1963	-0.355	-0.183	-0.629	-0.416	-0.444	-0.474
	(0.033)	(0.038)	(0.044)	(0.027)	(0.031)	(0.032)
1964	-0.349	-0.154	-0.674	-0.428	-0.444	-0.473
	(0.032)	(0.038)	(0.044)	(0.033)	(0.036)	(0.031)
1970	-0.262	-0.115	-0.508	-0.278	-0.302	-0.338
	(0.032)	(0.037)	(0.044)	(0.020)	(0.031)	(0.029)
1971	-0.242	-0.111	-0.486	-0.244	-0.287	-0.312
	(0.031)	(0.038)	(0.044)	(0.022)	(0.032)	(0.031)
Returns to Education						
1963	0.041	0.012	0.102	0.051	0.068	0.073
	(0.003)	(0.004)	(0.004)	(0.004)	(0.007)	(0.003)
1964	0.040	0.013	0.103	0.064	0.079	0.075
	(0.003)	(0.005)	(0.004)	(0.006)	(0.007)	(0.003)
1970	0.037	0.003	0.101	0.055	0.066	0.071
	(0.003)	(0.005)	(0.004)	(0.003)	(0.006)	(0.003)
1971	0.035	0.002	0.100	0.054	0.065	0.070
	(0.002)	(0.004)	(0.004)	(0.003)	(0.005)	(0.003)

Notes: The dependent variable is the natural logarithm of annual taxable earnings. Regressions also include a constant, and age and age-squared as explanatory variables. Observations with non-positive earnings are dropped from the analysis. The sample sizes for 1963, 1964, 1970, and 1971 are 8525, 8529, 8391, and 8275, respectively. The OLS2 specification also drops top-coded observations, leading to sample sizes of 4632, 4267, 4485, and 4163. MLE is Tobit maximum likelihood; CLAD is censored least absolute deviations; SCLS is symmetrically censored least squares; ICLAD is identically censored least absolute deviations.

Consequences of Topcoding for Decisionmaking

- Standard Censored Regression Problem
- Black/white earnings
  - Gap of .35 or .63 log points in 1963?
  - Change in gap between 1963 and 1971 .06 log points or .15 log points?
  - ⇒ Policy maker?
    - ⇒Racial earnings gap closing rapidly
    - ⇒Racial earnings gap closing slowly?
- Return to Education
  - First column: Dropped from 1% in 1963 to approximately zero in 1973?
  - Final column Consistent at 7%.
  - ⇒ Policy maker?
    - ⇒Stop investing in education?
    - ⇒Investment in education should increase?

Access Issues: Census Research Data Centers

What they are

- Researchers physically go to access data on a site controlled by NSI
- Monitored by Census Bureau Employees
- Supported by Census, NSF, host institution

Basic Approach

- Project Approval (RDC/Census Bureau/Other Data Custodian
- All projects must provide a benefit to Census Bureau programs. The benefit requirement is an explicit proposal criterion and is required by law (Title 13, Sec. 23, U.S.C.).
- Researchers using the facilities and databases at RDCs will be required to obtain Special Sworn Status from the Census Bureau.
- Disclosure penalties: $250,000, imprisonment for up to five years, or both.

Access Issues: Current Research Data Centers

- Access limited to researchers and staff authorized by the Bureau of the Census.
  - The computers within the RDCs are not linked to the outside world.
  - Researchers do not have email or world wide web access from within RDCs.
  - All analysis must be done within the RDC.
  - Researchers at the RDC may use confidential data only for the purpose for which the data are supplied; i.e., for their approved research project.
  - Researchers may not remove confidential data from RDC
  - Full Disclosure Review.

Research Data Centers: Drawbacks

- Low and declining utilization (fewer than 100 active projects) "Expensive, fragile and tenuous"
  - Length of review process
  - Cost in terms of time
  - Cost in terms of money
- Disparate use
  - Large, well endowed institutions (NY, Boston, Ann Arbor, DC, SF, LA, Chicago, NC)
  - Geographic proximity
- No remote access

Alternative Approach: Learn from other disciplines => Portfolio Approach

1. Approach
    1. NSF (cybertrust)
    2. NSF (IIS)
    3. Commercial applications (financial services)
    4. Other agencies (DOD)
2. Portfolio approach
    1. Computer protections
    2. Minimal statistical protection
    3. Legal requirements and screening
    4. Researcher training
3. Custom approach for different agencies

Potential Elements

- Multiple access modalities (driven by agency-specific needs and constraints)
- Complementary and integrated set of protections (legal; statistical; operational; educational)
- Customer driven
  - Consortium of agencies acts as hands-on advisory board guiding ongoing development of service.
- Example follows

Menu Options for Agency X (and Study Y)				
Sample Modalities	Legal Options (1,2,3,4)	Statistical (1,2,3,4,5)	Operational (1,2,3,4,5)	Educational (1,2,3,4)
Remote Access	3	1	4	2
	None	2	5	2
Onsite Access	3 w/customizations	3,5	1	None
Licensing (different levels of anonymization)	2	1	2,3	1,4

Research Access

- Remote access
  - external researchers access data via an encrypted connection with the data enclave using VPN
  - RSA Smart Card
  - Restrict user access from specific, pre-defined IP addresses
  - Citrix technology to access applications – configured so no downloads, cut and paste or print possible

Statistical Protection

- Remove obvious identifiers and replace by unique identifier
- Access limited to data requested and authorized
- Statistical techniques chosen by agency (recognising data quality issues)

Researcher Training

- Subjects
  - Basic confidentiality
  - Agency specific
  - Dataset specific
- Locations
  - Onsite
  - Webbased
  - Researcher locations e.g. NBER summer institute

Summary

Need to be proactive and develop new approaches

No "silver bullet" – use portfolio to minimize risk

Use advances in non-statistical areas – particularly cybertrust and human cyberinfrastructure => work with Super Computer Center