Review of the literature on the statistical properties of linked datasets

ANDREW CHESHER and LARS NESHEIM

March 18th 2007

1. Introduction

- Linked datasets contain information from multiple sources: surveys, administrative databases.
  - In US and UK, social security or national insurance administrative databases of workers and longitudinal surveys of businesses.
  - In UK, English Longitudinal Survey of Aging (ELSA) and administrative and other health records.
  - In UK, Land Registry house price data, British Household Panel Survey (BHPS), and the Family Expenditure Survey.

- Linked datasets inherit properties of the source datasets.

- The linking process may modify properties.

- Questions:
  - What important statistical issues arise when linked datasets are used?
  - How do results in the statistical literature bear on these issues?

2. Five main statistical issues

1. Impact of contributing survey designs and non-response.

2. Measurement error issues arising e.g. because of imputation.

3. Impact of excluding unmatched units.

4. Impact of including erroneously matched units.

5. Consequences of linking when there are no units in common.

3. Main conclusions

- Major statistical issues fall under three headings.

  1. Survey design issues.

  2. Measurement error issues.

  3. Information loss.

- Solutions exist (and have long been known) “in principle” but,
  - implementation can be technically demanding, and:
  - either demanding of information,
  - or dependent on the veracity of assumptions.

4. Survey design issues

- Contributing surveys have complex designs - are “not representative”.

- Linking procedures bring additional design issues.

- Methods for inference with complex designs are available.

- Implementation may be difficult for many linked datasets.

- Addressing data quality issues may resolve this problem.

5. Measurement error issues

- Failure to link.

- Erroneous links.

- Imputation errors.

- Non-classical measurement error.

- Solutions require knowledge or assumptions about both datasets and about the measurement error.

6. Information loss

- Information loss may arise:
  - when unmatched records are discarded,
  - when records are linked erroneously.
- Whether there is information loss depends on the objects studied.
- Exploiting “lost information” is a current research topic.
- There are solutions for some simple cases.

- The impact of complex design in this context seems unresearched.

7. Plan of the rest of the presentation

1. Types of data linking and how the three issues arise.

2. Survey design - statistical issues, solutions and open questions.

3. Measurement error - statistical issues, solutions and open questions.

4. Review of specific literatures.

8. Types of linking: direct record linkage

- Units in common, *no errors* in identifiers.
- Design of linked data is determined by designs of contributing surveys.
- Sample inclusion probabilities (SIP) are products of SIP's for contributing surveys.
- Complex survey design issues.
- Discarding unlinked records destroys information.

9. Types of linking: probabilistic record linkage

- Units in common, *errors* in identifiers.

- Design of linked data is determined by designs of contributing surveys, the measurement error process and the linking procedure.

- If only "good links" are retained there is no measurement error issue but additional design issues.

- If "bad links" are retained there is complex measurement error.

- Linking destroys information.

10. Types of linking: statistical record linkage

- No units in common.

- Survey 1: $\{X, Y\}$, survey 2: $\{X, Z\}$, link records with “close” values of X to produce a $\{X, Y, Z\}$ data set.

- Linked data set informative about population distribution of $\{X, Y, Z\}$ only if conditional independence: $Y \perp Z | X$ holds.

- Survey design requires attention when linking - unresearched.

- Measurement error in linked dataset.

- Linking destroys information.

- Analysis is possible without linking even when conditional independence fails to hold.

11. Survey design (a)

- Variables of interest:

$$U \equiv \{X, Y, Z\}.$$

- One survey reports values of $\{X, Y\}$ the other reports values of $\{X, Z\}$.

- X: an identifier.
  - A unique identification number
  - An identifying characteristic such as employment size, location, etc.

- Y: an outcome, perhaps value added.

- Z: perhaps measures of innovative activity.

12. Survey design (b)

- Simple survey design (simple random sampling)
  - Units in *population* equally likely to appear in *sample*.
  - Sample is representative.

- Probability a random draw from the *population* falls in a set A is

$$\int_{u \in A} f(u) du \qquad \text{or} \qquad \sum_{u \in A} f(u).$$

- f is the probability density of U in the population.

13. Survey design (c)

- Complex survey design.
  - Units in the *population* are *not equally likely* to appear in a *sample*.
    - Design
    - Non-response
    - Attrition
    - Data linkage
- Define a weighting function $w(u)$

- Probability a sampled unit in set A ends up in final sample is

$$\int_{u \in A} w(u) du \qquad \text{or} \qquad \sum_{u \in A} w(u)$$

- Complex survey sample is a set of random draws from a weighted density function

$$g(u) \propto w(u) f(u).$$

- Weighting function often only depends on a few elements of $U = \{X, Y, Z\}$ and varies discretely.

14. Survey design (d): weighted analysis

- The statistical literature provides a variety of methods for inference under complex survey designs.
  - conduct *weighted* analysis, but weights must be known,
  - *maximum likelihood* methods, but sample inclusion probabilities must be known, and a detailed model specification is required.
- Unweighted analysis can be informative about the target population/density function.
- Weights, sample inclusion probabilities could be estimated.

15. Survey design (e): when to weight

- Let $c_f = C(\ f)$ be a feature of f of interest, for example an expected value, or a coefficient in a regression function.

- Recall complex survey data are regarded as random draws from $g(u) \propto w(u)f(u)$.

- If $c_f = c_g \equiv C(g)$ then unweighted analysis delivers what is required.

- Whether this happens depends on the feature of interest, the structure of f and the structure of w.

- Some analysis which requires weighting may not be much affected by it.

- Some analyses which do not *require* weighting will benefit from it.

16. Survey design of linked datasets

- The probability a unit with value u appears in the complex survey sample is

$$\int_{u \in A} g(u)du \qquad \text{or} \qquad \sum_{u \in A} g(u)$$

- Surveys contributing to a linked data set may have different weighting functions, $w_1(u)$ and $w_2(u)$.

- A unit sampled with value u is in survey 1 with probability $\propto w_1(u)$ and in survey 2 with probability $\propto w_2(u)$ and in the linked data set with probability $\propto w_1(u) \times w_2(u)$.

- Linking may introduce additional dependence on u: $w_1(u) \times w_2(u) \times l(u)$.

- Difficulties arise when this dependence cannot be characterised.

17. Measurement error (a)

- *Identification* issues are at the root of the great difficulties caused by measurement error.

- A feature of the target population is *not identified* if populations in which the feature has *different* values generate data with the *same* probability distribution.

- If *additive independent* measurement error is assumed:

$$W = U + V$$

there is, for the distribution of the observed data:

$$f_W(w) = \int f_U(w-v) f_V(v) dv$$

- Data is informative about the left hand side. Many distributions f_U and f_V can produce the same f_W. Rather like:

$$6 = 5 + 1 = 4 + 2 = 3 + 3 \cdots\cdots$$

18. Measurement error (b)

- With additive independent measurement error

$$W = U + V$$

  there is not just *inaccuracy* in estimation of means of U, but *bias* in estimation of variances of and relationships amongst elements of U.

- The literature has many solutions, all resting on assumptions that are untestable, mostly for *simple* measurement error processes and for *linear* models.

- Solutions are of limited use for many practical data linkage problems:
  - Measurement error processes for linked data are complex.
  - Much research involves complex non-linear models.
- Much research is needed - but reducing measurement error is the priority.

19. Four US linked data sets

1. Longitudinal Research Database (LRD).
   - Linked data on manufacturing establishments.

2. Longitudinal Enterprise Establishment Microdata (LEEM).
   - Linked data on all private sector establishments.

3. Pollution Abatement Cost and Expenditure (PACE) survey.
   - Linked data on manufacturing establishments.

4. Longitudinal Employer Household Database (LEHD).

    - Linked data on establishments and workers.

20. US linking processes and problems

- Complete enumeration of large establishments, sample of small establishments.

- Data imputation and measurement error more important for small establishments.

- Complexity of firm dynamics led to Company Organization Survey (COS).

- Some work using probabilistic matching based on name and address discussed in Jarmin and Miranda (2002).

21. Linkage failures: causes

- Not sampled due to survey design.
- Not in operation.
- Missing data due to non-response.
- Some units out of sampling frame due to timing of sampling (e.g. PACE and LRD).
- Identification numbers change over time due to business restructuring.

- Missing or erroneous identification numbers.

- Processing of ownership changes different in different data sources (e.g. LRD and PACE).

22. General lessons (a)

- Detailed longitudinal information about organizational structure is required to minimise linkage errors.

- Extensive cross-validation is required to minimise errors.

- As much information as possible about the linkage process, probabilities of sample selection, non-response, measurement error concerns, and any imputation procedures should be preserved in the data documentation and in the data.

23. General lessons (b)

- Collect measures of important variables like output or employment in all surveys to increase probability of correct links.

- Efforts to maximise consistency of variable definitions over time.

- Efforts to reduce misinterpretation of survey questions.

24. Studies linking business data

- Bartelsman and Doms (2000) is representative.

  Because Longitudinal Microdata sets (LMDs) provide a large number of observations, and hence lower standard errors, much of the research using LMDs has not explored data quality. This neglect has left open the questions of how much of the heterogeneity is real and how much comes from errors to variables; nor has much attention been given to the statistical properties of linked data sets.

- Nevertheless, a variety of ad hoc strategies are used.

25. Best practice (a)

- Data quality and measurement error.
  - Evaluate size, kind, and impact of measurement errors on features of interest.
  - Check for internal consistency, external consistency, and plausibility.
- Achieved sample design.
  - What is the achieved sample design?.
  - How does linking process impact achieved sample design?

26. Best practice (b)

- In sample inference.
  - Weighted analysis is appropriate except under very special circumstances.
  - Assumptions required to justify unweighted analysis can be tested.
  - If unweighted analysis is valid, then parameter estimates obtained from weighted and unweighted analyses should not be statistically significantly different.
- Out of sample extrapolation.
  - Based on theory and assumption or on outside information.

27. Studies linking firm and worker data (a)

- Abowd and Kramarz (1999) survey of literature using matched firm-employee data.
  - Discuss nearly 100 studies from 17 countries using 38 different linking systems.
  - Detailed discussion of the methods used to create the data sets and the statistical models used.
  - Only one paragraph on methodological issues: weighting and measurement error.

28. Firm and worker data (b)

- Abowd, Finer, and Kramarz (1999) use multiple imputation (see Little and Rubin, 1987) to overcome missing data.

- Abowd, Crepon, and Kramarz (2001) use a dynamic attrition model to model the missing data process.

- Abowd and Vilhuber (2003) use probabilistic matching technique to overcome erroneous identification numbers.

- Dolton, Lindeboom, and van den Berg, (1999) study variables correlated with non-response and differences in outcomes across response/non-response categories.

- Hildreth and Pudney (1999) account for probability of sample inclusion using weights.