Review of the literature on the statistical properties of linked datasets

Andrew Chesher and Lars Nesheim

Centre for Microdata Methods and Practice (CeMMAP)

October 2011

Main conclusions

- Three sets of major statistical issues arise from linking datasets.
  1. Survey design issues.
  2. Measurement error issues.
  3. Information loss.
- In each category, there are solutions "in principle" but,
  - implementation can be technically demanding, and:
  - either demanding of information,
  - or dependent on the veracity of assumptions.
- This talk discusses how the 3 issues arise and discusses available solutions.

Survey design issues

- Contributing surveys have complex designs - i.e. they are "not representative".
- Linking procedures bring additional design issues.
- Methods for analysis with complex designs are available.
- Implementation may be difficult for many linked datasets.
- Addressing data quality issues may resolve this problem.

Measurement error issues

- Measurement error causes linking to fail.
- Erroneous links carry measurement error contaminated data.
- Imputation in many-to-one matching introduces measurement error.
- All analysis with measurement error is highly reliant on maintained assumptions.
- There are solutions under many specific simple assumptions.
- Many not applicable in the data linking context.
- Addressing data quality issues may ease this problem.

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.
- There are solutions for some simple cases.
- The impact of complex design in this context seems unresearched.

Plan of the rest of the presentation

1. Types of data linking and how the 3 issues arise.
2. Survey design - statistical issues and solutions.
3. Measurement error - statistical issues and solutions.
4. Review of specific literatures.
5. Recommendations.

Types of linking: direct record linkage

- Multiple (non-representative) datasets each with partial information on common observational units
-
    - 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.

Types of linking: probabilistic record linkage

- Units in common, *errors* in identifiers.
- Design of linked data is determined by designs of contributing surveys, and the measurement error process and linking procedure.
- If only “sure 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.

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.

Survey design (a)

- The statistical literature distinguishes:
  - Descriptive inference - about features of the finite population sampled.
  - Analytic inference - about features of the process generating the finite population's values.
- Much economics research conducts *analytic inference*.
- When conducting analytic inference one thinks in terms of a superpopulation of infinite extent,
  - from which the finite population is a sample of independent draws,
  - over which values of variables U are distributed with probability density function $f(u)$.

Survey design (b)

- The variables whose values are recorded are

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

- One survey reports values of $\{X, Y\}$ the other reports values of $\{X, Z\}$.
- X: an identifier, perhaps a postal address.
- Y: perhaps the market value of a house.
- Z: perhaps measures of house quality or energy efficiency ratings.
- With u denoting a value of U, the probability a random draw from the superpopulation falls in a set A is

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

Survey design (c): weighting

- In a complex survey design units in the finite population are *not equally likely* to appear in a sample.
- The probability a unit with value u is chosen in a random draw from $f(u)$ depends on u.
- Define a weighting function $w(u)$ so that the probability a unit sampled from $f(u)$ whose value u falls in a set A is chosen for the sample is

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

- The complex survey sample can be regarded as random draws from a weighted density function

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

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

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.

Survey design (e): when to weight

- Let $c_f = C(f)$ be a feature of f of interest, for example a moment, 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.

Survey design of linked datasets

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

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

- 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: $w_1(u) \times w_2(u) \times l(u)$.
- Problems arise when this dependence cannot be characterised.

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:

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

Measurement error (b)

- With additive independent measurement error

$$W = U + V$$

  there is 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 (with the current data), mostly for *simple* measurement error processes and for *linear* models.
- Solutions
-
  - What is known about measurement error? Size? Likely statistical relationship with observables?
  - Multiple measurements with independent errors.
  - Improved measurement.
  - Examine sensitivity to measurement error.

Recommendations

1. In cases of interest,
    1. Attempt to determine design of the linked datasets.
    2. Identify where measurement error arises and attempt a characterisation.
    3. Identify what additional information is required to complete these tasks.
    4. Determine the need for weighted analysis, investigate its implementation and examine sensitivity to weighting.
    5. Examine sensitivity of results to measurement error.