Combining population and survey data

Chris Jackson
MRC Biostatistics Unit, Cambridge
chris.jackson@mrc-bsu.cam.ac.uk

With Nicky Best, Sylvia Richardson (Imperial College London)
http://www.bias-project.org.uk

NCRM Research Methods Festival
3 July, 2008

Overview

Review of statistical methods for combining population and survey data – with some applications

- Examples of population and survey data, (dis/)advantages of each
- Combining data in estimation problems:
  - Small area estimation
  - Reweighting surveys to population.
- Determining associations between variables
  - Reducing ecological bias from aggregate data using individual data.
  - Combining datasets with different sets of variables.
- Running theme: multilevel and Bayesian graphical models

Definitions

Examples of population data – nominally represents everyone

- the census area-level data (at various levels of aggregation)
- national registers of births and deaths
- case registers for specific diseases (cancer, congenital anomalies)

Definitions

Examples of population data – nominally represents everyone

- the census area-level data (at various levels of aggregation)
- national registers of births and deaths
- case registers for specific diseases (cancer, congenital anomalies)

Examples of survey data – subset of population

- Annual Population Survey, Health Survey for England
- British Household Panel Survey, Millennium Cohort Study (longitudinal)
- Samples of Anonymised Records from the Census

Characteristics of population data

Advantages of censuses and registers:

- Representative of population (nominally, although may be small selection effects / under-enumeration).
- Large – statistical power to discern small effects

Characteristics of population data

Advantages of censuses and registers:

- Representative of population (nominally, although may be small selection effects / under-enumeration).
- Large – statistical power to discern small effects

Disadvantages:

- Limited number of variables beyond basic demographics
  - especially in registers. Census better but limited information on e.g. health
- Often available only as aggregate counts / percentages over areas – loses potentially important information (see later)
- . . . access to most detailed form of data usually restricted.

Characteristics of survey data

Advantages:

- Individuals can be examined in detail on a particular research area (e.g. Health Survey for England - different health topic every year)
  - lots of variables collected

Characteristics of survey data

Advantages:

- Individuals can be examined in detail on a particular research area (e.g. Health Survey for England - different health topic every year)
  - lots of variables collected

Disadvantages:

- Small subset of the population
  - limited information on geographical variations
  - limited power to detect small effects e.g. environmental exposures
- Often selection bias by design, as well as biased non-response
  - may need to reweight to represent population

Combining population and survey data

- One dataset presents only one piece of the picture
- Make most of information from different sources
- Improves power and can alleviate particular biases:
  - selection bias, confounding, ecological bias, missing data, measurement error
- ... May require complex models to represent complex situations
  - data must be available to inform about particular biases

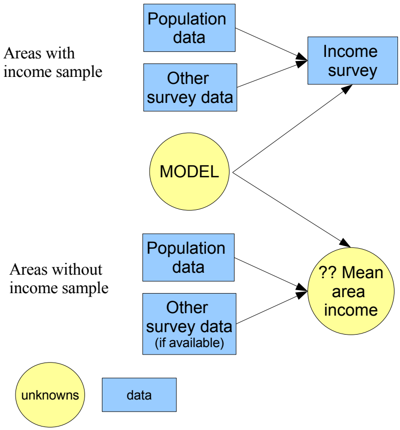
Combining population and survey data

- One dataset presents only one piece of the picture
- Make most of information from different sources
- Improves power and can alleviate particular biases:
  - selection bias, confounding, ecological bias, missing data, measurement error
- . . . May require complex models to represent complex situations
  - data must be available to inform about particular biases

Consider research questions of the form:

- estimating quantities (*"what's the mean income among (ethnic group) in (area)"*), or
- finding associations between quantities (*"how is income related to chronic illness"*)...

Small area estimation



- no population income data
- but surveyed data not available for all areas / too few responses
- Principle: fit a model to surveyed data
  - model based on association between population data / other surveyed data and income
  - may also exploit correlation between neighbouring areas (Bayesian hierarchical models, see Spatial Statistics session this afternoon)
- → use model to predict income for other areas.

Accounting for survey selection bias

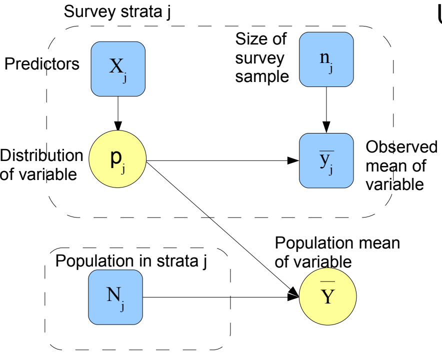
- Estimating quantity of interest for a population from survey data $y_1, \ldots, y_n$
- Mean of surveyed variable $\sum_i y_i / n$ is biased due to selection.
- Reweight survey responses to represent population: $\bar{y} = \sum_i w_i y_i / n$, where $w_i = 1/$ probability of selection.
- Weights w_i may be known from sampling design
  - but what if design not published,
  - or certain individuals decline invitation to survey,
  - or "item" non-response to some survey questions ...?

Poststratification for selection bias

Poststratification: compare survey data with population data to estimate probabilities of selection:

- Define a set of strata r (e.g. sex $\times$ age $\times$ socio-economic)
- n_r responses in each r from a population of N_r.
- $\rightarrow$ probability of selection $= n_r/N_r$ for stratum r
- Or improve precision further using a model for the variable of interest. . .

Model-based poststratification for selection bias



Using a model for the variable of interest

- Distribution p_j of variable within stratum j estimated from survey data
- modelled in terms of predictors *in addition to* strata, or smoothed,
    - exploiting correlations between similar strata / correlations through time in, e.g, annual surveys
- → compute population mean of variable of interest using p_j and population N_j

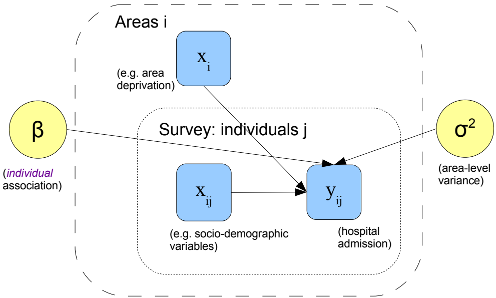
See, e.g. Gelman and Carlin, *Poststratification and weighting adjustments*, Survey Nonresponse, 2002. (with examples from US political opinion polls)

Association problems

Or *"regression"* or *"correlation"* problems.
Examples:

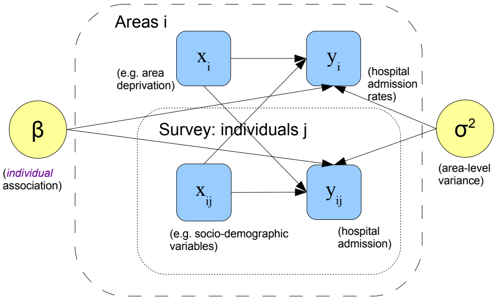
- effects of environmental exposure / socio-economic status on incidence of a disease.
- Contrasting individual and contextual effects on health:
  - do characteristics of your neighbours / area determine (e.g.) your health, as well as your own characteristics
  - needs individual and area-level data – often population and survey data

Multilevel models for individual / contextual effects



- Individual outcome: hospital admission for CVD (from Health Survey for England).
- Modelled on individual and area-level socio-demographic predictors.
- Multilevel model separates individual / area variations in outcome
  - account for correlations within areas

Multilevel models for individual / contextual effects



- What if we also have area-level outcomes as well as individual-level outcomes?
- e.g. hospital admissions data: proportion admitted to hospital for CVD.
- Estimate predictors of individual-level hospital admission simultaneously from area / individual data . . .

Ecological inference

Estimating individual-level associations from area-level averages – prone to ecological bias (= ecological fallacy):

1. association is different for area-averaged data (for non-linear individual models e.g. binary and count data)
2. can't distinguish between individual / area exposure effect

Ecological inference

Estimating individual-level associations from area-level averages – prone to ecological bias (= ecological fallacy):

1. association is different for area-averaged data (for non-linear individual models e.g. binary and count data)
2. can't distinguish between individual / area exposure effect

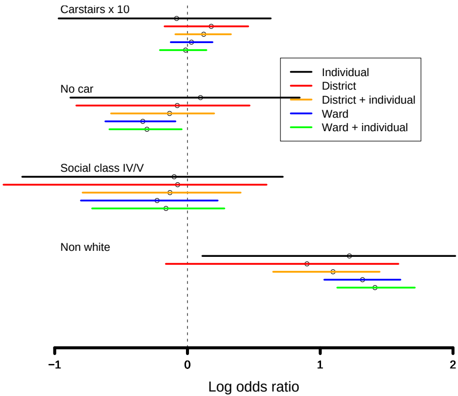
Solutions:

1. use appropriate models (see, e.g. Wakefield, J Roy Stat Soc A, 2004)
   - instead of simple regression of area outcome on area exposure
   - computes appropriate area-level risk as individual risk averaged over within-area distribution of risk factors.
   - needs an estimate of within-area exposure distribution. . .
2. incorporate individual-level data – improves power, and . . .

Using individual data to alleviate ecological bias

- Individual exposure data alone – improves estimate of within-area variability
  - (Best et al., Environmental benzene exposure and childhood leukaemia. J Roy Stat Soc A 2001)
- Individual exposure-outcome data – direct information on association of interest (*hierarchical related regression*)
  - (Jackson, Best and Richardson, Stat Med 2006, J Roy Stat Soc A 2008, also see http://www.bias-project.org.uk/research)
- Case-control data – outcome-dependent sampling – more informative than survey data when outcome is rare
  - (Haneuse and Wakefield, J Roy Stat Soc B, 2008)

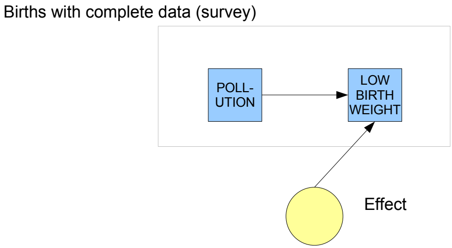
Socio-demographic predictors of CVD hospitalisation



(risk of hospital admission for CVD in 1998 for adults in London. Jackson, Best and Richardson, J Roy Stat Soc A, 2008)

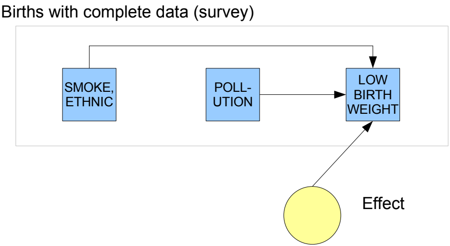
- Estimates from individual data / district data not precise
- Combining individual and aggregate data increases power.
- Using smaller areas (wards instead of districts) improves precision further.
- No additional effect of area-level deprivation (Carstairs index) on top of individual factors

Graphical model for combining mismatched data



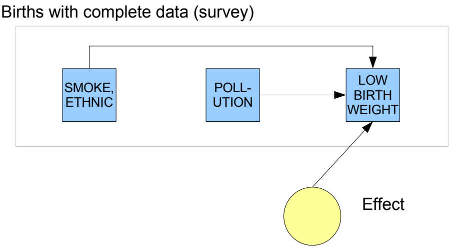
- Study of relationship of low birth weight to environmental pollution
- Millennium Cohort Study: about 20, 000 UK births in year beginning Sep 2000.

Graphical model for combining mismatched data



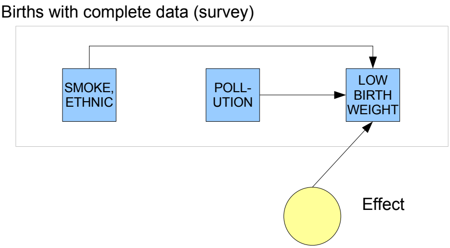
- Study of relationship of low birth weight to environmental pollution
- Millennium Cohort Study: about $20,000$ UK births in year beginning Sep 2000.
- Relationship confounded by ethnicity and smoking

Graphical model for combining mismatched data



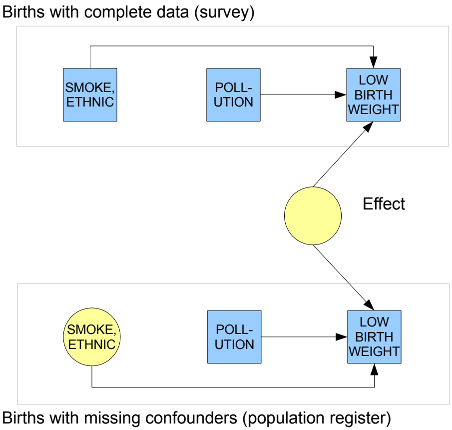
- Study of relationship of low birth weight to environmental pollution
- Millennium Cohort Study: about $20,000$ UK births in year beginning Sep 2000.
- Relationship confounded by ethnicity and smoking
- Survey may not have power to detect small association with pollution

Graphical model for combining mismatched data



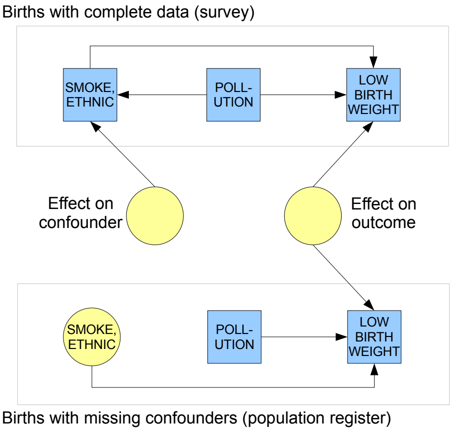
- Study of relationship of low birth weight to environmental pollution
- Millennium Cohort Study: about 20,000 UK births in year beginning Sep 2000.
- Relationship confounded by ethnicity and smoking
- Survey may not have power to detect small association with pollution
  - Incorporate population data

Graphical model for combining mismatched data



- National births register – all births in population
- Important confounders – smoking and ethnicity – not recorded

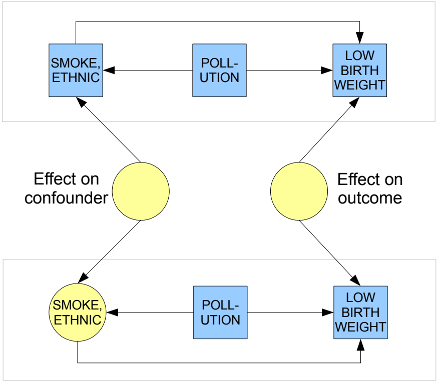
Graphical model for combining mismatched data



- Fit model for confounders using data from survey
  - based on predictors available in both datasets, including pollution.

Graphical model for combining mismatched data

Births with complete data (survey)



Births with missing confounders (population register)

- Fit model for confounders using data from survey
  - based on predictors available in both datasets, including pollution.
- Predict from that model to impute confounders in register
  - = standard multiple imputation for missing data

Graphical model for combining mismatched data

Data	Odds ratio (95% CI) for IQR of NO_2 exposure
Millennium Cohort data alone	0.94 (0.79, 1.13)
Register data (ignoring confounders)	1.15 (1.07, 1.23)
Combined data	0.98 (0.91, 1.04)

(Jackson, Best and Richardson, 2008, Submitted to *Biostatistics*. See http://www.bias-project.org.uk/research)

- Power increased by combining data
- Confounding appropriately controlled for
  - . . . but needs sufficient predictors of missing data
  - here we used area-level ethnicity from census
- Bayesian graphical model used to propagate uncertainty about imputation . . .

Bayesian graphical models

- Generalisation of multilevel models (hierarchical relationships) to any network of quantities.
- Complex system represented as global model built from smaller components
  - each representing different data source or bias
- Nice mathematical properties – network structure exploited to form joint probability distribution of unknowns
- Efficient algorithms (Markov Chain Monte Carlo) used to estimate distribution of unknowns given data – exploiting network structure $\rightarrow$ general-purpose software e.g. WinBUGS

Summary: combining population and survey data

- Observational data prone to biases (selection, confounding, missing data ...) – multiple data sources can inform and alleviate biases
- Combining suitable data can improve power – each dataset presents a small part of the picture.
- Statistical methods like graphical models useful
  - complex → need to check model assumptions
  - better to have tidier data in first place...?
- Data of different forms from the same source particularly valuable – e.g. Samples of Anonymised Records / population aggregate data from the census.