Between- and within-pair effects in logistic regression with measurement error

Lyle Gurrin
Melbourne School of Population Health

Elizabeth Williamson
Department of Epidemiology and Preventive Medicine
The Alfred Centre, Monash University, Melbourne

Martin Hazelton
Institute of Fundamental Sciences – Statistics
Massey University, Palmerston North, New Zealand

University of California Irvine, May 2012
Outline

Models for paired data

Estimation

Shared measurement error

Simulation study

Example - Birthweight & EPO

Future work

References
Paired data

- Twin studies provide naturally matched pairs that can exploit within-pair comparisons of data to avoid confounding exposure-outcome associations by shared factors.

- Specific assumptions about shared factors allow the determination of genetic and environmental contributions to disease risk.
Between- and within-pair regression

- Model of Neuhaus & Kalbfleisch (1998): outcome $y_{ij} \sim \text{Bern}(p_{ij})$ for individual $j$ in pair $i$ where

$$
\log \left( \frac{p_{ij}}{1 - p_{ij}} \right) = \beta_0 + \beta_w \frac{1}{2} (x_{ij} - \bar{x}_i) + \beta_b \bar{x}_i \tag{1}
$$

for $j = 1, 2$ and $\bar{x}_i = (x_{i1} + x_{i2})/2$.

- We can also write

$$
\log \left( \frac{p_{ij}}{1 - p_{ij}} \right) = \beta_0 + \beta_w \frac{1}{2} (x_{ij} - x_{ik}) + \beta_b \bar{x}_i \tag{2}
$$

where $k = (3 - j)$, indicating that this model has terms for both between- and within-pair regression effects.
Estimation

- Use ordinary logistic regression (OLR) on all pairs to estimate $\beta_0$, $\beta_w$ and $\beta_b$.

- Use conditional logistic regression (CLR) within pairs on outcome-discordant pairs to estimate $\beta_w$.

- Use Neuhaus & Jewell (1990) (NJ) logistic regression between pairs on outcome-concordant pairs to estimate $\beta_b$.

There is a close empirical correspondence (Neuhaus & Kalbfleisch (1998), Ten Have et al. (1995)) between the OLR, CLR and NJ estimates, but they are not formally identical.
OLR, CLR and NJ estimators

- For a single binary exposure the CLR estimate of $\beta_w$ is the ratio of number of exposure-outcome concordant pairs ($n_c$) to exposure-outcome discordant pairs ($n_d$) among outcome-discordant pairs. For OLR there is also a contribution from the $m$ outcome-concordant exposure discordant pairs:

  \[
  \text{CLR: } \hat{\beta}_w = \log\left(\frac{n_c}{n_d}\right)
  \]

  \[
  \text{OLR: } \hat{\beta}_w = \log\left(\frac{2n_c + m}{2n_d + m}\right)
  \]

- For $\hat{\beta}_w = 0$, the OLR $\hat{\beta}_b$ has the same estimating equation as NJ $\hat{\beta}_b$ using pairs with outcomes concordant ($\overline{y}_i = 1$ or $\overline{y}_i = 0$), but for NJ outcome-discordant pairs contribute too, taking the paired-outcome value of $\overline{y}_i = 0.5$. 
Differing estimates of $\beta_b$ and $\beta_w$:

- When estimates of the between- and within-pair effects $\beta_b$ and $\beta_w$ differ (e.g. cluster effects) it raises questions regarding the interpretation of the estimate of the former, and whether it provides useful information about the latter.

- One resolution is to essentially ignore $\beta_b$ as a nuisance, and use the estimate of $\beta_w$ as effect size.

- Trouble is this seems to throw away information – we may well expect $\beta_b$ to tell us something about $\beta_w$. 
Measurement error

► Our approach: assume that the pair mean covariate $\bar{x}_i$ is measured with error, but that the within-pair difference is subject to no error. This is equivalent to

$$w_{ij} = x_{ij} + u_i$$

where $u_i$ is pair-specific measurement error.

► This model will be applicable in situations where the relative measurement of $x$ within pair is accurate, but the overall mean level of $x$ is less so.

► Even if $\beta_b = \beta_w = \beta$ then failure to account for the measurement errors leads to attenuation in the estimates of $\beta_b$, generating an apparent discrepancy with $\beta_w$. 
Proposed model

- Specifically, suppose that in truth the data are generated according to model

\[
\log \left( \frac{p_{ij}}{1 - p_{ij}} \right) = \beta_0 + \beta_w x_{ij} \quad (3)
\]

- This is equivalent to assuming that \( \beta_w = \beta_b \) in the between- and within-pair model, implying a single effect size \( \beta_w \), and individual cases are exchangeable.

- However, if we now fit the incorrect model

\[
\log \left( \frac{p_{ij}}{1 - p_{ij}} \right) = \beta_0 + \beta_w \frac{1}{2} (w_{ij} - w_{ik}) + \beta_b \bar{w}_i
\]

then \( \beta_w \) unchanged \( w_{ij} - w_{ik} = x_{ij} - x_{ik} \), but \( \beta_b \) will be attenuated because of error in \( \bar{w}_i \).
Measurement error models

➢ Note that

➢ The naive model ignoring measurement error requires between- and within-pair terms in the regression model.

➢ As the measurement error variance converges to zero we recover the true model with a single covariate and regression parameter.

➢ If this variance becomes very large then \( \beta_b \) will contain essentially no useful information about \( \beta_w \) in practice.

➢ Former situation OLR on all data is optimal, in the latter there’s no loss using CLR. Models with an intermediate level of measurement error can be regarded as lying on a continuum between these extreme cases.
Model fitting

What we now need are methods for fitting model 3 based on observations on $w_{ij}$ rather than $x_{ij}$. We consider three methods:

**CLR:** This makes use of only the errorless differences $w_{ij} - w_{ik} = x_{ij} - x_{ik}$ so should be unbiased but inefficient since it makes use of only a fraction of the data.

**NJ:** Use only outcome-concordant pairs and view each pair as a single data point, regressing the 0/1 outcome on $\bar{w}_i$ - expect this to be biased.
SIMEX method

- Adapt the SIMEX method (Cook & Stefanski 1994) with shared within-pair measurement error and apply it to the entire data set.
- Add increasing random error (shared within-pair) to covariates, estimate regression coefficients, extrapolate to mimic a lack of measurement error.
- Aim: Generate an estimate of \( \beta \) that is more efficient than estimating \( \beta_w \) from conditional or between- & within-pair logistic regression models.
Demo Graph

Between- and within-pair for binary outcomes

Gurrin, Williamson & Hazelton

Models for paired data

Estimation

Shared measurement error

Simulation study

Example - Birthweight & EPO

Future work

References

13/ 18
Simulation study

- For $\beta = 1.5, 1.0, 0.75$ and $0.50$ perform 50 simulations with 500 pairs with i.i.d. covariates $x_{ij} \sim N(0, 4^2)$.
- Add shared measurement error $u_i \sim N(0, 1^2)$ to covariates in each pair.
- Generate binary outcome from a logistic model data using between- and within-pair covariate terms.
- Fit models to data, summarize for $\beta$ using:
  - mean of estimate.
  - std dev of the estimates - empirical s.e.
  - mean of model-based s.e.
**Table: Simulation Results**

<table>
<thead>
<tr>
<th>Method</th>
<th>True value log(OR)</th>
<th>Estimate (model)</th>
<th>std error (model)</th>
<th>std error (empirical)</th>
</tr>
</thead>
<tbody>
<tr>
<td>CLR ($\beta_w$)</td>
<td>1.5</td>
<td>1.75</td>
<td>0.66</td>
<td>0.71</td>
</tr>
<tr>
<td>NJ ($\beta_b$)</td>
<td>1.5</td>
<td>0.95</td>
<td>0.23</td>
<td>0.25</td>
</tr>
<tr>
<td>SIMEX ($\beta$)</td>
<td>1.5</td>
<td>1.34</td>
<td>0.14</td>
<td>0.16</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Method</th>
<th>True value log(OR)</th>
<th>Estimate (model)</th>
<th>std error (model)</th>
<th>std error (empirical)</th>
</tr>
</thead>
<tbody>
<tr>
<td>CLR ($\beta_w$)</td>
<td>1.0</td>
<td>1.17</td>
<td>0.33</td>
<td>0.57</td>
</tr>
<tr>
<td>NJ ($\beta_b$)</td>
<td>1.0</td>
<td>0.75</td>
<td>0.16</td>
<td>0.16</td>
</tr>
<tr>
<td>SIMEX ($\beta$)</td>
<td>1.0</td>
<td>0.95</td>
<td>0.09</td>
<td>0.10</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Method</th>
<th>True value log(OR)</th>
<th>Estimate (model)</th>
<th>std error (model)</th>
<th>std error (empirical)</th>
</tr>
</thead>
<tbody>
<tr>
<td>CLR ($\beta_w$)</td>
<td>0.75</td>
<td>0.80</td>
<td>0.17</td>
<td>0.23</td>
</tr>
<tr>
<td>NJ ($\beta_b$)</td>
<td>0.75</td>
<td>0.62</td>
<td>0.12</td>
<td>0.13</td>
</tr>
<tr>
<td>SIMEX ($\beta$)</td>
<td>0.75</td>
<td>0.74</td>
<td>0.07</td>
<td>0.08</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Method</th>
<th>True value log(OR)</th>
<th>Estimate (model)</th>
<th>std error (model)</th>
<th>std error (empirical)</th>
</tr>
</thead>
<tbody>
<tr>
<td>CLR ($\beta_w$)</td>
<td>0.5</td>
<td>0.52</td>
<td>0.10</td>
<td>0.11</td>
</tr>
<tr>
<td>NJ ($\beta_b$)</td>
<td>0.5</td>
<td>0.44</td>
<td>0.08</td>
<td>0.08</td>
</tr>
<tr>
<td>SIMEX ($\beta$)</td>
<td>0.5</td>
<td>0.50</td>
<td>0.05</td>
<td>0.05</td>
</tr>
</tbody>
</table>
Example

- Association between low birthweight and cord blood erythropoietin (EPO) as a marker of hypoxic stress *in utero* and possible growth restriction (*Carlin et al.* (2005)) in 110 DZ twin pairs (220 infants).

- OLR estimate:
  - $\beta = 0.40$ (s.e.=0.12)

- Between- and within-pair regression:
  - $\hat{\beta}_b = 0.34$ (s.e.=0.13)
  - $\hat{\beta}_w = 0.62$ (s.e.=0.23)

- The SIMEX-adjusted estimate:
  - $\beta = 0.49$ (s.e.=0.13)
Future work

- Conduct a more extensive set of simulations.
- Establish a formal connection between the OLR estimates of between- and within-pair regression coefficients and the NJ and CLR estimates.
- Applying SIMEX methods separately to between-pair estimates.
- Compare SIMEX method using all data to combined (weighted average) estimates of separate between- (NJ) and within-pair (CLR) estimates.
References


