A multiple imputation approach to remove residual confounding through coarse data models

Robert Grant
Chen, Gilbert & Daling (1999)

- Maternal risk factors for Down's syndrome
- Smoking appears to be protective! OR=0.80 (95% CI 0.68-0.95)
- Then adjusted for age (dichotomised at 35) it is not significant but still on the protective side: OR=0.89 (95% CI 0.73-1.10)
- Finally, adjusted for the precise age in years, it disappears: OR=1.00 (95% CI 0.82-1.24)
Residual confounding

- Chen, Gilbert & Daling's 2nd analysis was residually confounded.
- Age was the confounder, and it was measured coarsely.
- We only know a region within which it might lie for each of the mothers in the study.
- Heitjan & Rubin investigated MLE methods for coarse data, but with covariates we generally don't care about the marginal distribution.
Incomplete data

- This is a bit like missing data (note the presence of Don Rubin)
- In fact missing data is a special case of extreme coarseness.
- We can try some missing data methods on these confounding variables, but we might not know which observations are coarsened.
- For example, digit preference in number of cigarettes smoked per day.
How to make a medical scare story

Can we say anything about the true values?

Artificial data based on statistics of Oda & Kawai. Diab Care (2009); 32(9): e113.
Consider two “current smokers”
Ingredients

- Assumption about form of the conditional distribution of confounder's true values (hopefully informed by evidence)
- Any other correlates in the data, leading to conditional distribution
- Assumption about coarsening mechanism (hopefully informed by evidence)
Procedure

- Find the parameters of the conditional distribution of the true confounder, and if necessary the coarsening mechanism
- Plug these into the conditional distribution of the true values given all known data
  - (under your assumptions...)
- Multiply impute from this
  - or do it all in one by MCMC / HMC
- Analyse the substantive model as normal and combine by Rubin's rules
Heitjan-Rubin and its extension

\[ L(\theta, \gamma \mid X^*) = f(X^* \mid \theta, \gamma) = \int_X f(X, X^* \mid \theta, \gamma) \, dX \]

\[ = \int_X f(X^* \mid X, \gamma)f(X \mid \theta) \, dX \]  \hspace{1cm} (1)

\[ f(C \mid C^*, X, Y, \alpha, \gamma) = \frac{f(C^* \mid C, X, Y, \gamma)f(C \mid X, Y, \alpha)}{\int_C f(C^* \mid C, X, Y, \gamma)f(C \mid X, Y, \alpha) \, dC} \]  \hspace{1cm} (2)
Example: Heaped Poisson

\[ f(C \mid X, Y, \alpha) = \frac{(e^{-\lambda_i} \lambda_i^{c_i})}{c_i!} \text{, where } \lambda_i = e^{\alpha_0 + \alpha_1 x_i + \alpha_2 y_i} \]

\[ f(C^* \mid C, X, Y, \gamma) = \pi_i g_i I(c_i^* = 5 \mid c_i / 5) \left( 1 - \pi_i \right) (1 - g_i) I(c_i^* = c_i) \]

\[ \pi_i = \expit(\gamma_0 + \gamma_1 x_i + \gamma_2 y_i) \]

\[
\frac{e^{-\lambda_i} \lambda_i^{c_i}}{c_i!} \pi_i g_i I(c_i^* = 5 \mid c_i / 5) \left( 1 - \pi_i \right) (1 - g_i) I(c_i^* = c_i) \times \] \[
\left( \pi_i \sum_{j=0}^{4} \frac{e^{-\lambda_i} \lambda_i^{c_i + j}}{(c_i^* + j)!} \right)^{g_i} \left( \frac{e^{-\lambda_i} \lambda_i^{c_i^*}}{c_i^*!} \right) \left( 1 - \pi_i \right)^{(1 - g_i)}
\]
Example: Heaped Poisson

```plaintext
program define resconf_poisson
  version 11
  args lnfj loglambda logitpi
tempvar pi
tempvar lambda
qui gen double `lambda'=exp(`loglambda')
qui gen double `pi'=exp(`logitpi')/(1+exp(`logitpi'))
qui replace `lnfj'=((1-g)*(-1*`lambda') + (cstar*`loglambda') + ln(1-`pi') - lnfactorial(cstar)) + ///
  g*(-1*`lambda') + ln(((`lambda'^(cstar+1))/round(exp(lnfactorial(cstar+1))) * `pi') + ///
  ((`lambda'^(cstar+2))/round(exp(lnfactorial(cstar+2))) * `pi') + ///
  ((`lambda'^(cstar+3))/round(exp(lnfactorial(cstar+3))) * `pi') + ///
  ((`lambda'^(cstar+4))/round(exp(lnfactorial(cstar+4))) * `pi'))
end
```
Interval-censored normal

- A special case because most stats software has Tobit-esque regressions for this kind of data
- Get the predicted value and the SE
- Impute truncated normal by rejection sampling
Interval-censored with overlap...

- Useful sensitivity analysis
- Allows some misclassification
- Linear overlap makes integration simpler
Whitehall II attrition

- Real-life example based on Mein et al (2012)
- Gender difference in non-response at phase 2 of the study, adjusted for age
- Occupational grade (3 levels) is confounder – a coarse proxy for socio-economic status
- Looks like grade and sex are more strongly correlated than SES (including other predictors)

<table>
<thead>
<tr>
<th>Model</th>
<th>Beta</th>
<th>95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Confounded</td>
<td>0.262</td>
<td>0.161 to 0.362</td>
</tr>
<tr>
<td>Residually confounded</td>
<td>-0.012</td>
<td>-0.127 to 0.102</td>
</tr>
<tr>
<td>Imputed x80 (intreg)</td>
<td>0.028</td>
<td>-0.086 to 0.143</td>
</tr>
</tbody>
</table>
What's next?

- Robustness to mispecification
- Collection of likelihood functions for various common coarsening mechanisms and forms of conditional distribution
- Application to clustered coarsening such as coding habits of data collectors
References